ESTADO DEL ARTE DEL RECONOCIMIENTO DE VOZ ARTIFICIAL

JESÚS DAVID MENA RIVERA

JUAN CAMILO ROJAS CORTÉS

UNIVERSIDAD TECNOLÓGICA DE PEREIRA

FACULTAD DE INGENIERÍAS

PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

PEREIRA

2021

ESTADO DEL ARTE DEL RECONOCIMIENTO DE VOZ ARTIFICIAL

JESÚS DAVID MENA RIVERA

JUAN CAMILO ROJAS CORTÉS

Trabajo de grado presentado como requisito para optar al título de Ingeniero de Sistemas y Computación

Directora
IVONNE CASTAÑO OSORIO

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS
PROGRAMA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

PEREIRA
2021

AGRADECIMIENTOS

Yo Juan Camilo Rojas, quiero agradecer y dedicar este trabajo a mis padres y a mi familia en general, quienes han sido fundamentales en mi formación universitaria con su apoyo constante. Agradezco también a cada persona que, aún sin tener un vínculo sanguíneo, estuvo dispuesta a brindarme apoyo moral y/o económico en ciertas etapas difíciles de mi formación. Por último, pero con una importancia inmensa, agradezco a mi esposa Stephany Colorado, quien fue determinante en el tramo final de mi formación universitaria, pues sin su apoyo constante y amor, la finalización de mis estudios y la realización de este trabajo no hubiese sido posible.

Por mi parte, Jesús David Mena, quiero dedicar el presente proyecto en primer lugar a mis hermanos Walter Mena R. y Sharick Mena R., en segundo lugar, a todas las personas que formaron parte de mi vida en la universidad durante este proceso de formación, y en último lugar a las personas que conocí en el espacio universitario El Tortazo.

Finalmente es de nuestro interés agradecer a todas las personas que de alguna forma ayudaron con la culminación de la presente monografía. Agradecemos puntualmente a nuestra directora Ivonne Castaño O., al director del programa Carlos Augusto Meneses Escobar y a cada docente de la Universidad Tecnológica de Pereira que con sus conocimientos y pedagogía contribuyó de alguna forma a nuestra formación y a la realización del presente proyecto.

Resumen

La siguiente monografía condensa información relevante que nos brinda un entendimiento de qué es y cómo se han ido desarrollando los sistemas de reconocimiento automático del habla (SRAH), los cuales se encargan de interpretar señales de audio emitidas por un usuario.

Los múltiples avances matemáticos dentro del campo de la inteligencia artificial han permitido llevar a cabo reconocimientos con porcentajes de aciertos cada vez más altos. Técnicas como Bancos de Filtros, Codificación Predictiva Lineal, Modelos Ocultos de Markov, Redes Neuronales Artificiales y Lógica Difusa permiten que las señales de audio previamente procesadas sean clasificadas como palabras del lenguaje humano.

Grandes empresas como Google, Amazon y Apple ya han hecho uso de estos sistemas para el desarrollo de sus asistentes de voz como lo son Google Assistant, Alexa y Siri respectivamente.

Es así como la sociedad actualmente se encamina a un futuro tecnológico en donde las órdenes por comandos de voz serán cada vez más frecuentes, proporcionando ayuda en áreas como lo son la medicina, o incluso la domótica con el fin de facilitar tareas del hogar.

Abstract

The present monograph condenses relevant information that gives us an understanding of what automatic speech recognition systems (ARMS) are and how they have been developed, which are responsible for interpreting audio signals emitted by a user.

The multiple mathematical advances within the field of artificial intelligence have made it possible to carry out recognitions with increasingly high percentages of correct answers. Techniques such as Filter Banks, Linear Predictive Coding, Hidden Markov Models, Artificial Neural Networks and Fuzzy Logic allow previously processed audio signals to be classified as human language words.

Large companies such as Google, Amazon and Apple have already made use of these systems for the development of their voice assistants such as Google Assistant, Alexa and Siri respectively.

This is how society is currently heading towards a technological future where orders by voice commands will be increasingly frequent, providing help in areas such as medicine, or even home automation in order to facilitate household tasks.

Contenido

Pág.

Lista de Figuras

Pág.

Introducción

Con el nacimiento de las computadoras y los estudios en el campo de la inteligencia artificial ha sido de interés para el ser humano el ser capaz de comunicarse con las máquinas mediante medios como lo son la voz. Es así como el desarrollo de los sistemas de reconocimiento automático del habla (SRAH) comienzan a investigarse y desarrollarse.

Un Sistema de Reconocimiento Automático del Habla o sistema de reconocimiento de voz es un sistema automático que tiene la capacidad de interpretar y gestionar la voz emitida por un individuo. Esta señal pasa por un proceso de digitalización con el objetivo de obtener elementos de medición o muestras, permitiendo que el sistema interprete su comportamiento y que implemente procesos de tratamiento enfocados al reconocimiento [1]. Para lograr el objetivo de reconocer patrones en la señal de voz digitalizada se han implementado sistemas que, entre otras herramientas algorítmicas, pueden utilizar bancos de filtros, Codificación Predictiva Lineal, Modelos Ocultos de Markov, Redes Neuronales Artificiales, Lógica Difusa y Sistemas de Reconocimiento Híbrido [1]

La variedad de asistentes virtuales en el mercado como lo son Siri, Cortana, Alexa o Google Assistant son evidencia de que tan lejos hemos llegado en el campo del reconocimiento de voz, pero estas no son las únicas aplicaciones que los SRAH nos brinda.

En la presente monografía se hace un estado del arte en el que se reseñan las principales técnicas utilizadas por los Sistemas Automáticos de Reconocimiento de Voz. Se hará un repaso histórico del desarrollo de algoritmos y herramientas que ayudan a este propósito y además se hará una visión teórica de los elementos algorítmicos y matemáticos que posibilitan el funcionamiento

de este tipo de sistemas. Finalizando con un conjunto de aplicaciones que involucren el desarrollo o implementación de sistemas de reconocimiento de voz.

Planteamiento del problema

Los sistemas de reconocimiento de voz nacieron, en un principio, por la creciente necesidad de contar con sistemas a los que se les pueda controlar de manera no física y que sirvieran como ayuda para un porcentaje cada vez mayor de personas discapacitadas en el mundo [2]. Además de lo mencionado anteriormente, durante las últimas décadas se ha estudiado con insistencia la posibilidad de crear interfaces entre el ser humano y la máquina que sean controladas por la voz, sustituyendo así ciertas aplicaciones tradicionales basadas en teclados, paneles y otros dispositivos similares. Además, esta línea de investigación ha avanzado notablemente en los últimos años, ya que los reconocedores actuales manejan vocabularios cada vez más amplios, tienen tasas de error menores a la hora de hacer reconocimiento y tiempos de procesamiento pequeños comparados con los primeros desarrollos hechos en los años 90, principalmente porque se están usando algoritmos más eficientes, equipos de cómputo más potentes y económicos, y sobre todo porque se ha aumentado la complejidad de estos sistemas, empleando modelos más sofisticados, refinados y precisos [3].

Hacer algoritmos de reconocimiento de voz presenta una serie de dificultades y retos para los programadores. El principal desafío lo plantea la extracción de características de la voz, ya que este tópico puede plantear una serie de problemas a solucionar. Por ejemplo, se ha descubierto que la misma persona no es capaz de pronunciar dos veces una palabra de la misma forma. Esto puede deberse al estado de ánimo, la salud, la fuerza de pronunciación, el tiempo, la entonación, entre otras características [4]. Sin embargo, a pesar de las dificultades del proceso, se ha desarrollado una serie de algoritmos que determinan un nivel de coincidencia entre las pronunciaciones y así realizan un reconocimiento eficaz. Entre los algoritmos mencionados es posible listar aquellos que

usan herramientas matemáticas como los Dinamic Time Warping (DTM), modelos ocultos de Markov (HMM), redes neuronales, entre otros modelos. Además, en la actualidad existen una serie de herramientas de uso extenso en el área de reconocimiento de voz tales como HTK, CMU Sphinix, CSLU Toolkit, entre otras [5].

El reconocimiento de voz es un área de estudio que está teniendo cada vez más aplicación en lo investigativo y en lo comercial. El hecho de que existan algoritmos eficaces que reconozcan patrones auditivos y puedan realizar tareas con esta información hace que la experiencia de usuario sea mucho más placentera e intuitiva que con aquellos sistemas que únicamente tienen sistemas de entrada y salida de texto. Al tener tanta demanda y aplicabilidad en la actualidad, es importante que haya compendios de información en los que se dé una visión amplia desde lo teórico y lo práctico de las herramientas y algoritmos disponibles; con el objetivo de que los desarrolladores e investigadores puedan hacer uso de estos para su implementación o mejora. Desde la visión tomada por la presente investigación, estos compendios no existen o son insuficientes.

Con base en la información dada anteriormente, se plantea la siguiente pregunta de investigación: ¿Cómo ha sido el desarrollo de los sistemas de reconocimiento de voz a lo largo de la historia y con qué herramientas matemáticas y algorítmicas se han desarrollado?

1. Objetivos

1.1 Objetivo General

Caracterizar los sistemas de reconocimiento de voz más importantes desarrollados hasta la fecha, describiendo los componentes matemáticos y algorítmicos que permiten su funcionamiento.

1.2 Objetivos Específicos

- Describir las etapas y los elementos fundamentales en el desarrollo de los sistemas de reconocimiento de voz, mencionando también sus principales características.
- Comparar las herramientas matemáticas y algorítmicas más relevantes y utilizadas en el desarrollo de sistemas de reconocimiento de voz.
- Identificar las aplicaciones prácticas que han tenido los sistemas de reconocimiento de voz a lo largo de la historia.

2. Justificación

La cantidad de sistemas de reconocimiento de voz que se encuentran hoy en día en el mercado es tan grande y diversa como sus aplicaciones. La presente monografía cobra importancia bajo la premisa de realizar una síntesis académica de un conjunto de referencias que brinden claridad al lector respecto a cuáles son las técnicas que se han usado a lo largo de la historia y se siguen usando para desarrollar los SRAH.

Actualmente, con la variedad de técnicas resulta difícil decidir con certeza que técnica se debería utilizar para desarrollar un Sistema de Reconocimiento Automático del Habla y por qué. Con la ayuda de este compilado de técnicas, entre las cuales se encontrarán el uso de redes neuronales, lógica difusa y demás, se busca que el lector logre distinguir qué tan óptimas son unas técnicas frente a otras, así como una ilustración inicial de cuáles son los retos con los que es posible encontrarse a la hora de abordar una técnica sobre otra.

Finalmente, se busca que con el desarrollo del presente trabajo los profesionales y futuros profesionales tengan bases teóricas y experimentales en cuanto a cuáles son las técnicas más comunes y cuál o cuáles se debería utilizar al momento de realizar un SRHA. De igual forma, desde las diversas aplicaciones expuestas a lo largo de la monografía se espera inspirar en cuanto a la creación de nuevos desarrollos en campos como lo son la medicina, la domótica o asistentes de voz; siendo estas aplicaciones frecuentes y necesarias en la vida diaria relacionadas con los SRHA.

3. Metodología

En el presente trabajo se realizó un estado del arte que relaciona las técnicas matemáticas y algorítmicas más relevantes relacionadas con los sistemas de reconocimiento de voz. Por lo tanto, esta es una investigación descriptiva que, según lo sugerido en [6], trató de responder los siguientes interrogantes respecto al reconocimiento de voz a través de patrones de inteligencia artificial:

1. ¿Qué es?
2. ¿Cómo funciona?
3. ¿Qué partes o componentes tiene? ¿Cómo están relacionadas esas partes?

En concordancia con lo anterior, el presente trabajo corresponde a una monografía. En esta, se hizo un recuento histórico y teórico de las investigaciones más importantes hechas hasta la fecha relacionadas con los algoritmos de reconocimiento de voz. Además, el proceso metodológico que se llevó a cabo tuvo las siguientes partes:

1. Descripción: Se abordaron las técnicas más relevantes utilizadas para la extracción de características de la voz y su posterior reconocimiento, describiendo los elementos algorítmicos y matemáticos que posibilitan su funcionamiento.
2. Análisis: Se establecieron similitudes y/o diferencias entre las técnicas descritas, identificando las ventajas y desventajas de cada una.
3. Síntesis: Con base en la descripción y el análisis hechos anteriormente, se realizaron conclusiones acerca de la eficiencia y eficacia de las diferentes técnicas utilizadas para el desarrollo de sistemas automáticos de reconocimiento del habla, brindando también perspectivas respecto a las técnicas más usadas en la actualidad y a las que más proyección tienen en un futuro cercano.

Respecto a fuentes de información, para la presente monografía se utilizaron las publicaciones disponibles en los repositorios de las diferentes universidades y facultades de ingeniería alrededor del mundo, adoptando Google Académico como la principal herramienta de búsqueda y consulta.

4. Glosario de Siglas

A continuación se presenta el significado de las siglas que se utilizan en la presente monografía, adjuntando su traducción al español y su significado en su idioma original, en caso de que sean siglas que provienen de lenguas extranjeras.

AT&T: American Telephone and Telegraph (Teléfonos y Telégrafos Americanos, empresa estadounidense de comunicaciones).

RCA: Radio Corporation of America (Corporación de Radio de América).

MIT: Massachusetts Institute of Technology (Instituto de Tecnología de Massachusetts).

NEC: Nippon Electric Company (Compañía Eléctrica Nipona).

DTM: Dynamic Time Warping (Alineamiento Dinámico Temporal).

IBM: International Business Machines Corporation.

DARPA: Defense Advanced Research Projects Agency (Agencia de Proyectos de Investigación Avanzados de Defensa).

LPC: Linear Prediction Coding (Codificación Predictiva Lineal).

HMM: Hidden Markov Model (Modelo Oculto de Markov).

DNS: Dragon Naturally Speaking.

MAP: Mean average Precision (Precisión promedio de la media)

DTW: Dynamic Time Warping (Técnicas de programación dinámica)

MFCC: Mel Frequency Cepstral Coefficient (Deformación de tiempo dinámica)

BCRL: Brazo robótico para cirugía de laparoscópica.

5. Sistemas Artificiales de Reconocimiento de Voz a través de la Historia

Los primeros intentos realizados por el ser humano para desarrollar un sistema que le permitiera comunicarse con máquinas a través de su voz datan de la década de 1940, surgiendo este como necesidad de contar con una herramienta que permitiera que las personas con problemas auditivos pudieran percibir el habla como algo visible [2]. Estos primeros sistemas estaban basados en dispositivos mecánicos, pero estos fueron evolucionando poco a poco hacia los componentes electrónicos, hasta llegar a los complejos sistemas actuales enmarcados en el campo de la informática y las ciencias de la computación. En este apartado del presente trabajo se hará un recuento histórico del desarrollo de los sistemas automáticos de reconocimiento del habla.

Los primeros intentos para desarrollar un sistema artificial que lograra reconocer palabras dichas por un ser humano tuvieron lugar en la década de 1940, siendo impulsados por los laboratorios Bell de AT&T. Los investigadores de este laboratorio construyeron un dispositivo que se basaba en los principios de la fonética acústica, dependiendo su éxito de la capacidad del sistema para percibir información verbal compleja de alta precisión [3]. El desarrollo de este sistema se prolongó hasta la década de 1950, haciendo posible la identificación de dígitos mono-locutor teniendo como base la medición de las resonancias espectrales del tracto vocal para cada dígito que se percibía. Siguiendo la misma línea de investigación, RCA Labs logró reconocer 10 sílabas. Sin embargo, el mayor avance de la década de 1950 fue presentado por el University College de Londres y el MIT Lincoln Lab, quienes lograron desarrollar un sistema de reconocimiento limitado de vocales y consonantes [1]; representando esto una novedad para la época por el uso avanzado de elementos estadísticos y teniendo como objetivo mejorar el rendimiento en palabras que tuvieran dos o más fonemas.

Los años 60 del siglo XX trajeron consigo importantes avances en la investigación relacionada con los sistemas de reconocimiento automático del habla, siendo el primero de estos el sistema de hardware específico desarrollado por los laboratorios NEC en Japón [2]. Así mismo, es posible destacar 3 proyectos de suma importancia en la década de 1960 [3]:

- Los desarrollos de RCA Labs continuaron y se enfocaron en la búsqueda de soluciones para problemas relacionados con la falta de uniformidad existentes en las escalas de tiempo del habla, pues todos los seres humanos tienen velocidades distintas a la hora de pronunciar las palabras. Para corregir esto, diseñaron distintas técnicas de normalización en el dominio temporal, detectando así el inicio y el fin del discurso de manera fiable.
- T. K. Vintsyuk (Unión Soviética) propuso el uso de herramientas de programación dinámica con el fin de obtener el alineamiento temporal de parejas de realizaciones. Este fue el origen de la técnica de Dynamic Time Warping (DTM) que es ampliamente usada en la actualidad.
- D. R. Reddy (Universidad de Stanford) desarrolló una técnica de seguimiento dinámico de fonemas, logrando reconocer oraciones de amplio vocabulario en el campo del reconocimiento de habla continua.

En la década de 1970 creció el interés de diferentes empresas y grupos de investigación por el campo del reconocimiento automático del habla, así como ciertas críticas por la fiabilidad de los métodos desarrollados hasta el momento. Una de las empresas que demostró su interés fue IBM, desarrollando proyectos de reconocimiento de grandes vocabularios. Así mismo, el gobierno de los Estados Unidos de América puso en marcha varios proyectos de inversión para el desarrollo de sistemas automáticos de reconocimiento del habla, siendo estos conocidos como los proyectos DARPA [1]. Respecto a los campos de investigación, los años 70 fueron testigos de la exploración profunda de herramientas probabilísticas en el campo del reconocimiento de voz, siendo el mayor

avance el desarrollo de la codificación predictiva lineal (LPC por sus siglas en inglés), utilizándose este método en sistemas automáticos del reconocimiento del habla de amplio uso en la actualidad. Además, se implementaron avances importantes en el uso de programación dinámica y se implementaron sistemas de reconocimiento de palabras aisladas fundamentados en el encaje de patrones [3].

Para el inicio de la década de 1980 ya existía una buena base en la construcción de los sistemas de reconocimiento del habla. Había varios proyectos de desarrollo desde la década de los 70 que vieron la luz en los años 80 y representaron avances significativos en este campo. En proyectos anteriores solamente se reconocían vocablos aislados, pero a partir de estos avances se lograron reconocer palabras encadenadas de forma fluida. El principal avance investigativo dado en la década de 1980 fue el desarrollo de los Modelos Ocultos de Markov (HMM por sus siglas en inglés), siendo este un modelo probabilístico de alta fiabilidad que es ampliamente usado en la actualidad [3]. Por otro lado, en la década de los 80 se hicieron los primeros avances en la utilización de redes neuronales artificiales para el reconocimiento de voz y es posible destacar los siguientes proyectos e hitos de investigación [2]:

- Desarrollo de decodificadores fonéticos con base en el conocimiento y la experiencia de fonetistas en la interpretación de espectrogramas.
- Recopilación de grandes bases de datos de audio que sirvieron para estandarizar los resultados y comparar el desempeño de los diferentes proyectos de desarrollo.
- Importantes avances del programa estadounidense DARPA en la investigación y el desarrollo de sistemas automáticos de reconocimiento del habla.

En la década de 1990 continuaron los avances en el desarrollo de sistemas automáticos de reconocimiento del habla, centrándose la mayoría en la ampliación de los vocabularios y la

separación en los campos de aplicación. El reconocimiento automático de voz se pudo aplicar a la línea telefónica y a entornos con ruido ambiente excesivo, obteniendo resultados positivos [3]. Por otro lado, gracias a los avances en electrónica y al desarrollo de hardware potente a precio bajo, fue posible instalar sistemas de reconocimiento del habla en equipos de cómputo al alcance de la población común, por lo que su uso empezó a popularizarse. Se logró realizar la integración de este tipo de softwares con el sistema operativo de las máquinas y con herramientas de procesamiento de lenguaje natural, dándole así mayor sentido a las palabras reconocidas y una posible interpretación para estas [1].

En el siglo XXI se continuó ampliando el vocabulario de los sistemas automáticos de reconocimiento del habla, elevando así la efectividad de los mismos en su objetivo. Por otro lado, se optimizaron los algoritmos y se aprovechó el desarrollo de Hardware más potente para que este tipo de sistemas tuvieran una respuesta mucho más rápida que en el siglo XX [3]. Además, se popularizó el uso de redes neuronales y Deep Learning para el reconocimiento del habla, situación que permite que los sistemas lleven a cabo un proceso de aprendizaje que da la posibilidad de que mejoren a medida que son utilizados por el usuario, aprendiendo nuevas palabras y corrigiendo errores sobre la marcha [2].

6. Elementos a tener en cuenta para Diseñar un Sistema de Reconocimiento de Voz

En el presente capítulo se relacionan algunos elementos y conceptos teóricos sobre los que es necesario tener claridad para poder desarrollar un sistema de reconocimiento de voz efectivo. Estos conceptos son el de los tipos de aprendizaje, el decodificador acústico-fonético y el modelo del lenguaje.

6.1 Tipos de Aprendizaje

En los sistemas automáticos de reconocimiento de voz es necesario emular diversas formas de aprendizaje que el ser humano aplica, ya que el objetivo es que estos sistemas sean capaces de reconocer patrones auditivos, palabras y frases de la misma manera que lo hace cualquier persona. Por ese motivo es primordial identificar los tipos de aprendizaje que pueden ser aplicados a los sistemas de reconocimiento de voz, siendo este el motivo del presente apartado.

Camargo [4] afirma que los sistemas de reconocimiento de voz pueden utilizar aprendizaje inductivo, deductivo o una mezcla de estos dos, según las metodologías y objetivos que se plantean. En la siguiente figura es posible observar la diferencia entre estos tipos de aprendizaje:

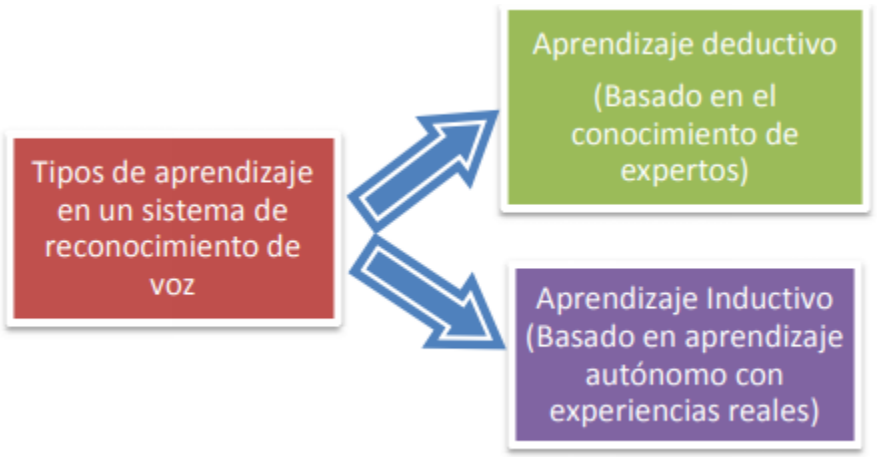

Figura 1 Tipos de aprendizaje en un sistema de reconocimiento de voz.

Fuente: [4]

Como se puede apreciar en la Figura 1, el aprendizaje deductivo está basado en bases de conocimientos, dando lugar a lo que en inteligencia artificial es conocido como sistemas expertos. Por el contrario, el aprendizaje inductivo da libertad al sistema para que aprenda con base en las experiencias que va adquiriendo. A continuación, se hablará de cada uno de estos aprendizajes de forma más detallada.

6.1.1 Aprendizaje Inductivo

El aprendizaje inductivo puede definirse como la capacidad de obtener conceptos nuevos a partir de ejemplos [5]. Cuando un sistema computacional aplica este tipo de aprendizaje, este es capaz de realizar procesos de generalización y especialización partiendo de un conjunto de ejemplos de entrada. Los algoritmos que aplican este tipo de aprendizaje son incrementales. Esto quiere decir que el procesamiento de los ejemplos se realiza uno a uno y paso a paso, permitiendo este hecho observar el efecto que tiene cada uno de los ejemplos de entrada en el resultado final.

Morales y González [6] mencionan que todo problema debe listar los siguientes cinco elementos para ser caracterizado de naturaleza inductiva:

- Clase de reglas: Se refiere a la clase de funciones a trabajar o al lenguaje que se toma en consideración. Por ejemplo, todas las expresiones regulares de un alfabeto específico, programas en Prolog, funciones recursivamente enumerables, lenguajes libres de contexto, entre otros.

- Espacio de hipótesis: Se trata del conjunto de todas las hipótesis que están enmarcadas dentro de un lenguaje de hipótesis definido. En este punto se elige de qué formas los métodos utilizan las reglas y de qué forma aprende de ellas. Cuando se determina el espacio de hipótesis es importante tener en cuenta lo que se quiere que el sistema realice, pues las hipótesis están direccionadas precisamente a que estas tareas se cumplan.

- Conjunto de ejemplos y presentación: Es necesario definir, según el tipo de problema y lo que se quiera lograr, qué ejemplos se le pasan al sistema y de qué forma se presentan. Por ejemplo, en algunos casos es bueno pasar solamente ejemplos de éxito, mientras que en otros es recomendable pasar ejemplos de éxito y fracaso, indicando al sistema qué ejemplo corresponde a qué caso.

- Clase de método de inferencia: En esta etapa es necesario definir el proceso computacional a través del cual se leen los ejemplos y se producen las hipótesis.

- Criterio de éxito: En este punto es necesario aclarar cuáles son las posibles salidas del sistema que podrían considerarse como exitosas o como fallidas. Es necesario además fijar este parámetro en valores numéricos, según la precisión deseada para el sistema que se está construyendo.

El aprendizaje inductivo puede ser supervisado o no supervisado [5]. En el primer caso se le indica al sistema a qué concepto pertenece cada ejemplo de entrada. En este tipo de sistemas el aprendizaje es realizado a través del contraste entre ejemplos, respondiendo a qué características distinguen a los ejemplos de un concepto de otros. En el caso del aprendizaje inductivo no supervisado, si bien siguen existiendo ejemplos de entrada, no existe una clasificación. En este caso el sistema debe encontrar la forma más adecuada de particionar los ejemplos.

Un buen ejemplo de los procesos computacionales inductivos es el aprendizaje de una función [6]. Los ejemplos serían pares (x, f(x)), siendo x la entrada y f(x) la salida. El proceso inductivo sería el siguiente: Dada una colección de ejemplos f, retornar una función h cuyas salidas sean aproximadas a las salidas de f. A esta función se le conoce como hipótesis o modelo.

6.1.2 Aprendizaje Deductivo

El aprendizaje deductivo consiste en la transmisión de conocimientos de un experto (o de una base de conocimientos) a una máquina [7]. Una buena aproximación al aprendizaje deductivo en máquinas puede darse a través del aprendizaje deductivo en seres humanos, concretamente al aprendizaje de una lengua como segundo idioma. Este proceso se asimila, por ejemplo, al aprendizaje que deben seguir los sistemas de reconocimiento de voz.

En un enfoque deductivo, un tutor experto explica todas las reglas gramaticales a sus estudiantes para que estos puedan estudiarlas de forma individual y de forma mecánica. Esto quiere decir que el punto de partida son las reglas y con base en estas el estudiante deduce los ejemplos [8]. Esto contrasta con el aprendizaje inductivo, en el que primero se dan los ejemplos y a partir de ahí es que el aprendiz deduce las reglas. El método deductivo de enseñanza es también conocido como método de enseñanza didáctica, ya que es un enfoque tradicional en el que el profesor da a

sus alumnos una explicación de las formas y requiere que estos presten atención consciente para las puedan entender, aprender y aplicar en su proceso de aprendizaje.

Una de las principales aplicaciones del método deductivo de aprendizaje a las máquinas se da en la lógica de predicados, siendo esta utilizada principalmente en el desarrollo de sistemas expertos. En este paradigma lógico se le pasan al sistema una serie de reglas (predicados en este caso) que este debe almacenar y sobre los cuales puede realizar actividades de reconocimiento. Estos predicados pueden adquirir, según el caso, un valor verdadero o un valor falso. García, et al. [9] mencionan las siguientes ventajas y desventajas de la lógica de predicados en la construcción de sistemas expertos, tomando como base los programas desarrollados en el lenguaje de programación PROLOG:

Manejo de Incertidumbre: En la lógica de predicados únicamente hay dos valores posibles (verdadero y falso). Esto constituye un problema en el manejo de la incertidumbre, ya que se trata de una lógica binaria. Hay muchos problemas de la vida cotidiana del ser humano que exigen un manejo de incertidumbre apropiado, pues muchos casos de aprendizaje cercanos al razonamiento humano (como el reconocimiento de voz) tienen una naturaleza difusa y no binaria.

Razonamiento Monotónico: La lógica de predicados y el paradigma de aprendizaje deductivo en general son un formalismo del razonamiento monótono, en el que se da una base de conocimientos cuya variación es compleja. Esto dificulta el modelado de problemas del mundo real en los que las condiciones y las reglas pueden cambiar a través del tiempo.

Programación Declarativa: Los sistemas expertos desarrollados con programación declarativa a través del paradigma deductivo siguen una serie de reglas y predicados para encontrar respuestas. Esto constituye una ventaja para el desarrollador del sistema, pues únicamente debe preocuparse por codificar las reglas del problema en cuestión. Sin embargo, esta misma

característica representa un problema para la eficiencia computacional, pues los algoritmos utilizados para clasificar información basada en predicados (como el algoritmo de Backtracking) suelen tener una complejidad computacional alta. Los tiempos de ejecución pueden llegar a ser bastante grandes cuando se trata de problemas complejos que requieren de muchos predicados.

6.2 Decodificador Acústico-Fonético

La primera etapa, y una de las más importantes, de un sistema de reconocimiento de voz es la captura de la voz y su transformación a señales digitales. La señal de voz normalmente es capturada a través de un micrófono y tiene datos relacionados con la palabra que se dice (son los datos de interés) y también ruido ambiental. Gracias a este último elemento es que se hace necesario que el procesador se adapte al entorno acústico en el que se encuentra, pues solo de esta forma es posible minimizar el ruido ambiente y fijar la atención en las palabras que se deben reconocer. Después de un proceso de análisis, el procesador de señal produce una serie de números o características que capturan los aspectos más relevantes de la expresión de forma compacta, dejando de lado aquellos datos que no sean de interés y que por el contrario puedan ralentizar el reconocimiento de voz y aumentar la complejidad computacional de los procedimientos. Una vez que se han extraído las características de la señal es que aparecen los decodificadores, siendo este un paso clave en el proceso de reconocimiento del habla [9]. Uno de los métodos más utilizados para realizar la decodificación es el enfoque acústico, dando origen así a los decodificadores acústico-fonéticos.

El modelo acústico del reconocimiento del habla trata de imitar el proceso natural de reconocimiento seguido por el ser humano, en el que la información utilizada por el oyente para identificar los fonemas se centra primero en la información acústica y después en la identidad de

los segmentos adyacentes [10]. En ese orden de ideas, los decodificadores acústico-fonéticos intentan realizar el reconocimiento automático de fonemas y palabras únicamente con los vectores de características correspondientes a la señal de voz, centrando su atención principalmente en las características acústicas de esta. La mayoría de los decodificadores acústico-fonéticos utilizan Modelos Ocultos de Markov, Redes Neuronales Artificiales o una combinación de estos dos elementos para ser entrenados y realizar la tarea de reconocimiento de voz.

6.3 Modelo del Lenguaje

En el contexto de los sistemas automáticos de reconocimiento de voz el modelo del lenguaje representa el conjunto de palabras que se pueden decir (vocabulario) y la forma en la que estas pueden decirse, es decir, la gramática [11]. Normalmente los modelos del lenguaje se definen a través de gramáticas de estados finitos, estableciendo todas las posibles reglas que pueden dar origen al compendio de palabras que el sistema tiene como objetivo reconocer [12]. El modelo del lenguaje también se puede definir como la descripción de un objeto matemático que reúne las características de la palabra hablada. Suele expresarse como un modelo estadístico de secuencias de palabras que tiene como objetivo conocer cuál fue la palabra articulada por el hablante [9]. El modelo del lenguaje se diferencia del modelo acústico en que este último se centra específicamente en las características sonoras de la señal de voz, mientras que el modelo del lenguaje trata de realizar el reconocimiento a través de reglas gramaticales que tienen cierta probabilidad asignada.

6.4. Clasificación de los Sistemas de Reconocimiento de Voz

No todos los sistemas de reconocimiento de voz son iguales entre sí. Algunos poseen algunas características o medios de funcionamiento específicos que requieren que el ambiente de

ejecución sea idóneo según los parámetros con que fueron desarrollados. A continuación se menciona cómo se clasifican los sistemas de reconocimiento de voz con base en criterios relacionados con su funcionamiento.

6.4.1 Sistemas Previamente Entrenados

Algunos sistemas de reconocimiento de voz requieren que se realice un entrenamiento previo por parte del usuario para que la precisión al momento de detectar las palabras sea óptima. Uno de los ejemplos más comunes a lo largo de la historia ha sido el sistema de reconocimiento de voz que integra Microsoft en sus sistemas operativos: la herramienta "Reconocimiento de Voz". En 2009, bajo el sistema operativo de Windows Vista, se observa que mediante una comparativa con el sistema de reconocimiento Dragon Naturally Speaking, sin realizar el entrenamiento, la herramienta de Windows presenta un porcentaje de error al realizar un ejercicio de dictado del 75.77% contra el 52.95% del DNS, y se demuestra en [2] que luego de un previo entrenamiento, la herramienta de Windows mejora sus números hasta un 25.59% frente al 30.10% obtenido en la segunda ocasión por el DNS. La ventana de entrenamiento del Sistema de Reconocimiento de Voz de Windows se puede observar en la Figura 2.

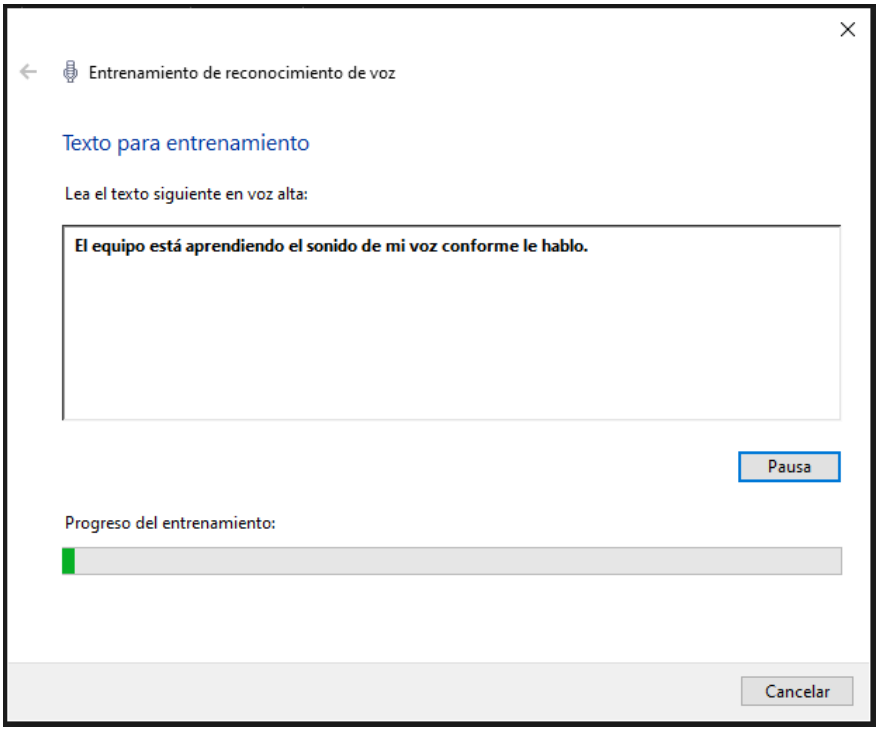

Figura 2 Ventana de entrenamiento de la herramienta de reconocimiento de voz de Windows.

Fuente: Elaboración propia.

Como se ha discutido a lo largo del presente trabajo, el entrenamiento es fundamental. En [1] se puede observar que se realizó un Dataset de 2000 registros del que se hizo una extracción de 1158 para entrenar un sistema que reconociera dígitos verbalizados, obteniendo resultados óptimos luego de utilizar una red neuronal compuesta por 130 nodos de entrada, 200 nodos ocultos y un nodo de salida.

En los dos casos presentados es de vital importancia recordar lo mencionado en el apartado 5.1.1: Para lograr que el sistema automático de reconocimiento del habla presente una tasa de error baja, se debe considerar realizar la grabación variando el tono, la intensidad, y la velocidad de la misma; para que el sistema pueda tener una gran variedad de parámetros que le sirvan como base.

6.4.2 Reconocimiento del Locutor

Dentro del campo de estudio de los sistemas de reconocimiento de voz existe un área que se especializa en reconocer a la persona que está realizando la grabación, también conocida como locutor. Si bien dentro de este campo es posible encontrar distintas ramas, las más estudiadas por sus aplicaciones en seguridad son la identificación y la verificación del locutor, siendo esta última más fácil de abordar.

Cuando se habla de verificación del locutor, se busca que esto se realice de forma tal que el reconocimiento no dependa de datos identificatorios por parte del hablante, como puede ser su nombre o su documento de identificación. Para llevar a cabo esta verificación actualmente se utilizan dos métodos: el Universal Background Model [5] y el Cohort Model [6]. El primer método se trabaja bajo la premisa de que, si el locutor de prueba se encuentra más cerca a una población promedio que el locutor objetivo, entonces es probable que el locutor de prueba no sea el objetivo. Por otro lado, el Cohort Model no utiliza la población, sino que realiza la tarea de reconocimiento enfocándose en el Cohorte. De este modo, si el locutor de prueba está más cerca que el locutor objetivo en comparación con el Cohorte, entonces la probabilidad de que el de prueba sea el objetivo es alta.

6.4.3 Continuidad del Sistema de Reconocimiento

Existe una diferencia entre los sistemas de reconocimiento de voz que permiten que el usuario hable de forma continua, y los que requieren que el usuario vaya haciendo pausas durante el uso del sistema de reconocimiento. Hasta el momento es poco lo que se ha dicho acerca del primer tipo de sistemas, estos requieren que la base de datos de entrenamiento del sistema se realice con grabaciones continuas en donde el número de palabras preferiblemente redondee más de 1000

palabras, lo cual hace que la complejidad del reconocimiento aumente considerablemente llegando al punto de realizar algoritmos híbridos para mejorar el rendimiento de estos sistemas.

6.4.3.1 A Continuous Speech Recognition System Embedding MLP into HMM.

Boulard y Morgan [13] a través de su artículo exponen cómo se desarrolló un sistema de reconocimiento de voz continuo mediante el método de embeber un perceptrón multicapa en un HMM, logrando así mejorar el rendimiento que ya se obtiene con técnicas como el Maximum Likelihood o Maximum a Posteriori.

Esto se logró al mezclar algoritmos como el perceptrón multicapas (MLP) y los modelos ocultos de markov (HMM), con la creación de una red neuronal cuyos valores de salida se utilizan como probabilidades MAP (Maximum a Posteriori) para el HMM. Es importante aclarar cómo esta combinación, aunque ayuda al rendimiento obtenido al analizar tramos, reduce el rendimiento cuando se trata de palabras. Adicionalmente en su desarrollo se recurre a la técnica del Dynamic Time Warping (DTW) explicada de una mejor forma durante el capítulo 6 en algunas aplicaciones realizadas.

Boulard y Morgan [13] se vieron en la necesidad de realizar modificaciones sobre el esquema básico. Uno de ellos, es el hecho de ajustar como criterio de parada sobre el MLP el hecho de que el rendimiento empiece a bajar sobre un segundo set de datos, y no cuando el nivel de error del entrenamiento baje.

Adicionalmente modificaron apartados como lo es la estimación de la probabilidad de las salidas del MLP, los costos de transición de palabras para el HMM subyacente, y el proceso cómo se segmentan los datos de entrenamiento.

Finalmente, como resultado Boulard y Morgan [13] evidencian el como la combinación de estas técnicas muestran mejoras sobre el uso de los HMM convencionales alcanzando porcentajes de hasta el 65.3% de acierto al adicionalmente usar técnicas como la segmentación Viterbi.

7. Herramientas Algorítmicas Utilizadas en los Sistemas Automáticos de Reconocimiento de Voz

En el presente capítulo se hará una reseña de las diferentes herramientas algorítmicas y matemáticas que suelen ser usadas en la actualidad por los diferentes Sistemas Automáticos de Reconocimiento del Habla, tales como Bancos de Filtros, Codificación Predictiva Lineal, Modelos Ocultos de Markov, Redes Neuronales Artificiales y Lógica Difusa.

7.1 Bancos de Filtros

El reconocimiento de patrones es solamente la etapa final de todo el proceso que se debe seguir para realizar un reconocimiento de voz efectivo [14]. Para poder reconocer patrones, sílabas y palabras de la voz humana, en primer lugar, es necesario captar el audio, hacerle un pre-procesamiento y extraer las características principales. Estas características son almacenadas en alguna estructura de datos y esta estructura se convierte en la entrada de los diferentes modelos que pueden reconocer patrones. Es en estas etapas previas al reconocimiento en las que los bancos de filtros cobran importancia vital, por lo que se analizarán algunos de sus elementos claves y su integración con los Sistemas de Reconocimiento de Voz en este apartado de la presente monografía.

7.1.1 Adquisición de Voz

En esta etapa se debe obtener la señal de voz a través de algún dispositivo que capture sonido. Es importante que en esta fase se definen algunos parámetros que pueden llegar a ser fundamentales en las etapas siguientes, tales como el tipo de canal (si es monofónico o

estereofónico), el formato de codificación de muestra, la frecuencia de muestreo, el formato del archivo, entre otros. Además, en esta etapa también se debe definir el número de palabras que conforman el corpus de voz [15].

En la etapa de adquisición de voz inicial (para la fase de entrenamiento del sistema) es fundamental elegir una buena muestra de fonemas y palabras. También es recomendable que la misma persona grabe las mismas palabras con diferentes tonos e intensidades, ya que estos son factores que pueden variar inconscientemente a la hora de hablar. Por ejemplo, en la investigación hecha por Martínez y Aguilar [14] se grabó la palabra "fijo" con cuatro diferentes variaciones de tono (normal, agudo, grave, nasal):

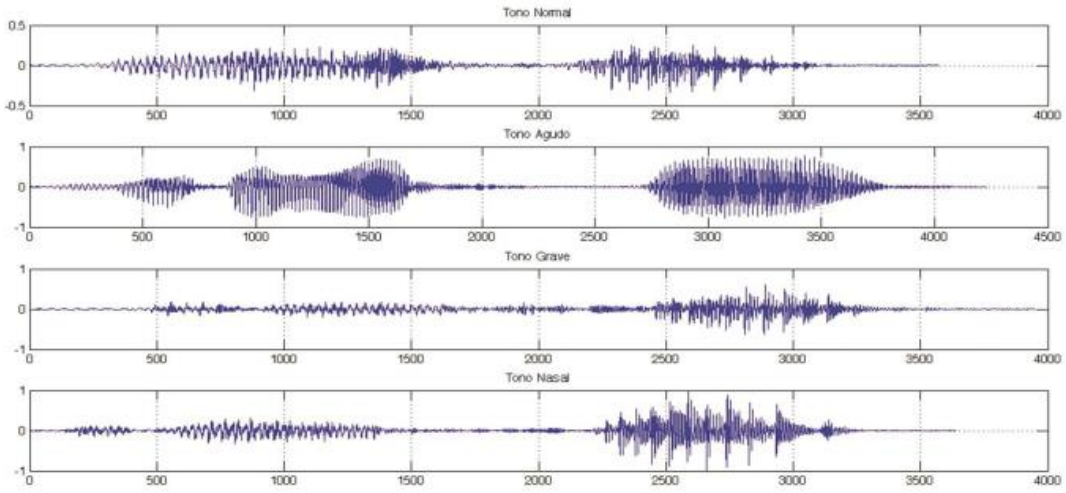

Figura 3 Comparación de la palabra "fijo" con cuatro variaciones de tono.

Fuente: [14].

Otra de las prácticas altamente recomendadas en la fase de adquisición de voz es que se graben palabras que tengan una gran variedad de fonemas. Por ejemplo, en una investigación hecha por Cruz y Acevedo [16] se decidió utilizar la palabra "zoológico", ya que esta contiene una buena cantidad de formantes de la voz y características espectrales. Además, se requiere que las

grabaciones sean hechas por un número significativo de personas y que, en lo posible, sean de diferentes regiones geográficas para que se validen los diferentes acentos que se pueden producir.

7.1.2 Pre-Procesamiento

El objetivo de la fase de pre-procesamiento es acondicionar la señal de audio de entrada para que esta pueda ser leída por el modelo de reconocimiento [16]. Para esto es necesario pasar la señal por una serie de procesos que, como lo explican Soto et al. [15], son los siguientes:

Filtro pasa bajas: Se eliminan las frecuencias más altas del audio. Normalmente, en estas frecuencias se encuentra gran parte del ruido ambiental.

Filtro pasa altas: Se eliminan las frecuencias más bajas del audio. Normalmente, en este rango de frecuencias se encuentran interferencias introducidas por el dispositivo de grabación o desperfectos en los cables y demás medios de transmisión.

Detección de la señal de voz: En esta fase se identifica cuáles son los fragmentos del audio en el que se está emitiendo la voz y se eliminan los silencios, identificando además el principio y el fin de cada palabra. Para realizar este proceso de forma automática es necesario calcular dos variables importantes: La energía promedio de la señal de voz y la energía de corto plazo, definiendo además un umbral de energía sobre el que se va a considerar que cierto fragmento corresponde a silencio, interferencia o voz humana. La energía promedio está dada por la siguiente ecuación:

$E_{promedio} = \frac{1}{N}\sum_{n=0}^{N-1} s(n)^2$, siendo s(n) la señal de la voz y N el total de muestras. Por su parte, la energía de corto plazo se define como la energía de una trama (intervalo determinado de tiempo) que se mide. Esta variable está dada por la siguiente ecuación:

$E_{trama} = \frac{1}{T}\sum_{n=0}^{T-1} \quad s(n)^2 t(m-n)$, en donde T representa el tamaño de la trama y el factor t(m-n) representa la trama a la que se le está calculando la energía.

Pre-énfasis: Se trata de un filtro digital parecido al pasa altas descrito anteriormente que tiene como objetivo hacer que el espectro de la voz tenga un rango parecido en todas las frecuencias. Este filtro está dado por la siguiente operación:

$S_{pp}[n] = S_{bp}[n] - \alpha S_{bp}[n-1]$, en donde $S_{pp}[n]$ representa la señal de salida actual del filtro, $S_{bp}[n]$ es la señal de entrada actual, α es una constante de suavizado que toma un valor entre 0.9 y 1, y $S_{bp}[n-1]$ es la señal de entrada previa. Como se puede apreciar en la operación descrita anteriormente, este es un proceso recursivo.

Segmentación y ventaneo: Es necesario hacer que los componentes de la voz que se va a analizar sean estacionarios. Para esto es necesario que se segmenta en tramas de entre 10 y 30 milisegundos, pudiendo considerar así la voz como una señal cuasi-estacionaria. Es necesario realizar este proceso porque las señales de voz realmente son estocásticas, variando su nivel de energía y contenido frecuencial en períodos largos de tiempo, dificultando esto su análisis. Cada una de las tramas generadas por este proceso debe ser superpuesta en una ventana adyacente con el objetivo de que las transmisiones entre tramas sean suaves y se eviten problemas por los cambios rápidos de la señal en los extremos de cada trama. A este último procedimiento se le conoce como ventaneo.

7.1.3 Extracción de Características con MFCC

Una vez que la señal de voz pasó por la etapa de preprocesamiento ya se encuentra lista para la extracción de sus características. En esta fase la voz es transformada en una serie de parámetros conocidos como coeficientes Soto et al. [15]. El principal desafío de esta etapa consiste en la elección de un modelo propició que logre estimar la envolvente espectral de la señal de la voz. Esta debe representarse por un número no muy grande de parámetros y además se debe utilizar un método que tenga una complejidad computacional tal que no se exija demasiado la máquina, ya que la segmentación hecha en el proceso anterior puede arrojar millones de unidades de datos sobre las que es necesario iterar.

Para lograr el objetivo propuesto anteriormente se ha vuelto popular el método de los Coeficientes Cepstrales en la Escala de Mel (MFCC) con bancos de filtros. Estos coeficientes pueden representar la amplitud del espectro de una forma compacta y sin tener que utilizar muchos parámetros [14]. Este método toma como base la percepción del sistema auditivo humano, haciendo especial énfasis en la variación de los anchos de banda de las frecuencias críticas al oído del ser humano Soto et al. [15]. En la Figura 4 se muestran las etapas que se siguen para elaborar un vector característico de MFCC:

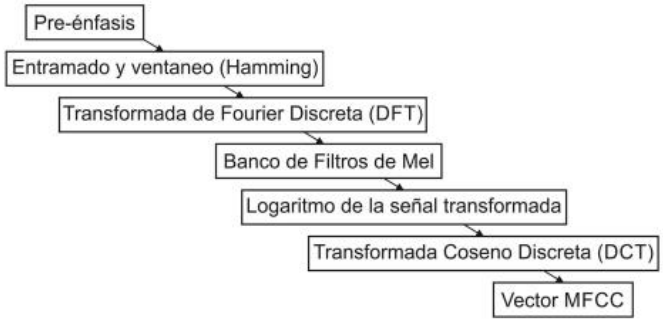

Figura 4 Proceso de obtención de los coeficientes MFCC.

Fuente: [14].

Las primeras dos etapas que se ven en la Figura 4 (Pre-énfasis y etapa de entramado y ventaneo) fueron explicadas anteriormente en el presente documento. Después de que la señal está segmentada por las etapas anteriores, esta debe pasar por la Transformada Discreta de Fourier (DFT) para obtener la magnitud y la densidad espectral de potencia de cada uno de los segmentos analizados Soto et al. [15]. El objetivo principal de esta etapa es identificar qué frecuencias contiene cada trama.

En la siguiente fase del proceso se deben agrupar las frecuencias resultantes de la Transformada Discreta de Fourier. Para cumplir esta tarea se utiliza el Banco de Filtros Mel. Este está compuesto por filtros triangulares que se distribuyen a lo largo de la escala de Mel. Esta escala tiene como objetivo determinar la frecuencia de un tono percibida por el oído humano, ya que se ha podido determinar a través de análisis psicoacústicos que este no percibe el tono de una manera lineal [18]. Los experimentos hechos han permitido hacer una aproximación a la forma en la que el oído percibe las frecuencias; aproximación expresada en la ecuación $F_{Mel} = 1127{,}01048\, log_e(1 + \frac{F_{Hz}}{700})$, en donde $F_{mel}(F_{Hz})$ corresponde a la frecuencia percibida por el oído humano. En concordancia con la ecuación anteriormente presentada, la escala de Mel puede interpretarse como una función logarítmica que comprime la frecuencia cuando está en valores altos. Para ilustrar esta relación se presenta la siguiente gráfica:

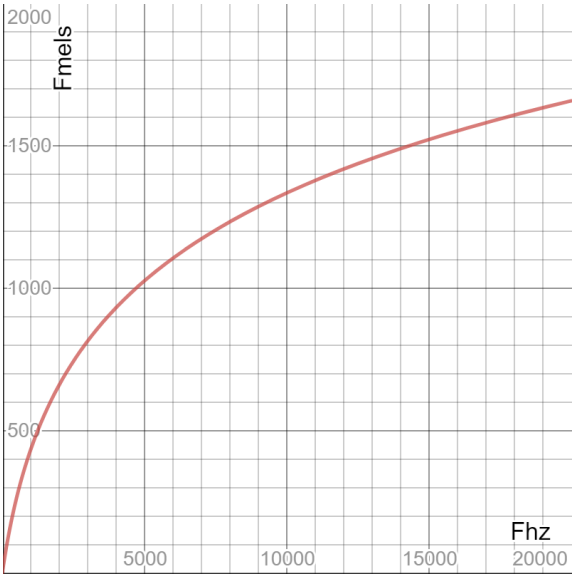

Figura 5 Escala de Mel en el rango de audición del ser humano.

Fuente: Elaboración propia

Como se puede ver en la Figura 5, la escala de Mel tiene un comportamiento aproximado a una función lineal por debajo de los 500 Hz, mientras que por encima de este umbral tiene un comportamiento logarítmico [14]. Teniendo en cuenta esta progresión se montan los filtros según los resultados de las frecuencias obtenidas con la Transformada de Fourier. Por lo tanto, los primeros filtros van a estar espaciados según un crecimiento lineal (más o menos equidistantes), mientras que a medida que se avanza por el espectro de frecuencias estos filtros se van distanciando más unos de otros, obedeciendo al crecimiento logarítmico de la escala de Mel. Se presenta un ejemplo gráfico de esto en la Figura 6, en la que se representa un banco de filtros triangulares, siendo este el tipo de filtro más utilizado para extraer características de la voz humana:

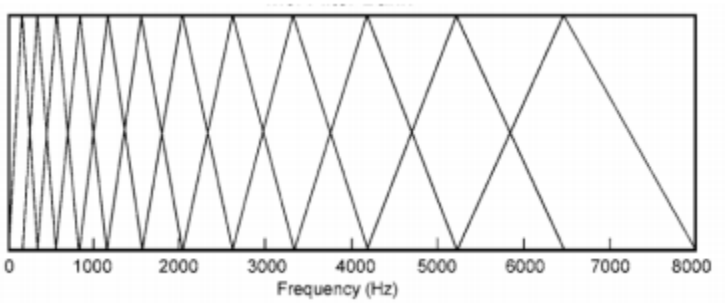

Figura 6 Representación de Banco de Filtros de Mel con 20 filtros entre 0 y 8000 Hz.

Fuente: [19].

Al tratarse de una función triangular, el cálculo de los Filtros de Mel debe expresarse matemáticamente a través de una función a trozos. La representación matemática es la siguiente:

$$B(m,k) = \begin{cases} 0 & si \quad k > f(m-1), \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & si \quad f(m-1) \leq k \leq f(m), \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & si \quad f(m) \leq k \leq f(m+1), \\ 0 & si \quad k > f(m+1), \end{cases}$$

Figura 7 Ecuaciones para calcular el Banco de Filtros de Mel.

Fuente: [15]

En las ecuaciones presentadas anteriormente, B(m,k) representa la matriz del banco de Filtros de Mel, m es el número del filtro que se está calculando, f es la frecuencia a la que se le calcula el filtro y k es el número de ventanas de análisis. Normalmente los primeros filtros que se obtienen son estrechos e indican cuánta energía hay cerca de los 0 Hz. Cuando la frecuencia aumenta, los filtros también se amplían y las variaciones terminan siendo menores. Después de obtener el banco de filtros, cada uno de estos se debe multiplicar por las ventanas de densidad de

potencia obtenidas en la Transformada de Fourier. con el objetivo de conocer la energía de los filtros y posteriormente sumar todos estos valores, resultando la siguiente expresión matemática:

$E(m,k) = \sum_{m=0}^{M} \quad B(m,k)P(k) \quad k = 1,2,\ldots,K$, siendo P(k) la densidad espectral de potencia de cada ventana y k el número de la ventana en cuestión. Después de esto es necesario calcular el logaritmo de las energías de los filtros con el objetivo de que las características obtenidas sean más parecidas a lo que el humano realmente escucha [15], pues como se explicó anteriormente, el oído humano no percibe el tono de forma lineal sino que tiene un comportamiento más parecido a una función logarítmica. Esta etapa del proceso simplemente consiste en calcular el logaritmo a cada uno de los elementos de la sumatoria en la que se calcula la energía de los filtros.

Después del cálculo del logaritmo de la energía falta solamente un paso para obtener el vector de características: Calcular la transformada del coseno discreto sobre el logaritmo de la energía de los bancos de filtros, obteniendo de esta forma el vector MFCC, que es el vector de características [18]. La transformada del coseno discreto (DCT) permite disminuir la complejidad computacional para la etapa del reconocimiento. Esta transformada está dada por la siguiente ecuación:

$MFCC(n) = \sum_{m=1}^{M} \quad E_{log}(m,k)\, cos(n(m - \frac{1}{2})\frac{\pi}{M}) \quad m = 1,2,\ldots,M,$ en donde n representa la posición en el vector de características, M representa el total de filtros del banco y k representa la ventana de análisis.

El vector MFCC obtenido con todo el proceso explicado anteriormente sirve como entrada para la etapa de reconocimiento, pudiéndose realizar con modelos probabilísticos (como los Modelos Ocultos de Markov) o de inteligencia artificial (como las redes neuronales y demás

sistemas expertos). Por lo tanto, la etapa de extracción de características es fundamental en el proceso de reconocimiento de voz.

7.2 Codificación Predictiva Lineal

En el ámbito de los Sistemas de Reconocimiento Automático del Habla, de la misma forma que los Bancos de Filtros con MFCC, el método de Codificación Predictiva Lineal (LPC por sus siglas en inglés) es utilizado para generar una estructura de datos con características de la voz humana, siendo este pasado como parámetro a las diferentes herramientas o algoritmos utilizados para reconocer los patrones, palabras o fonemas que se desean identificar.

El método de Codificación Predictiva Lineal está construido sobre la premisa de que cada muestra de la señal de voz puede representarse y predecirse a través de una combinación lineal de las muestras pasadas [16]. En términos matemáticos, esto indica que cada muestra de voz $s(n)$ puede aproximarse con la expresión $s(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p)$, en donde p representa el orden de predicción y cada valor de a corresponde a un coeficiente de predicción que es necesario calcular. El orden de predicción debe ser elegido según la frecuencia de muestreo y teniendo en cuenta además la longitud del tracto vocal de las personas que hacen las grabaciones de prueba [19]. Además, este número de coeficientes determina la resolución con la que el análisis de Codificación Predictiva Lineal puede representar la envolvente espectral de la señal de voz. La elección correcta de este parámetro es fundamental, porque un valor excesivamente reducido implica una resolución muy baja, mientras que un valor muy alto podría hacer que se capturen características que no son de interés, como por ejemplo el ruido ambiente o la interferencia por estática. En la siguiente figura se puede observar el comportamiento de los

coeficientes LPC según el parámetro p, tomando como entrada el espectro de Fourier que se ve en la parte inferior:

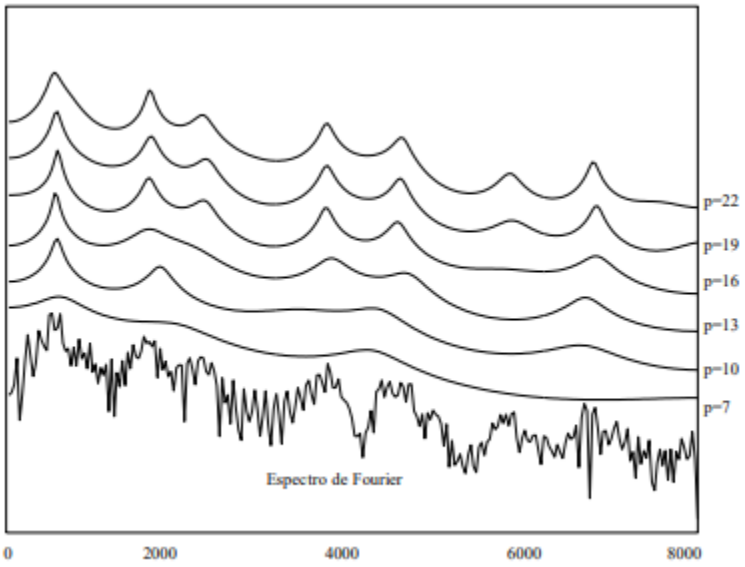

Figura 8 Variación del espectro LPC en función del número de coeficientes p.

Fuente: [19]

En el ejemplo presentado en la Figura 8, la frecuencia de muestreo fue de 16 kHz y es posible observar que con 16 coeficientes es suficiente para captar la información más relevante del espectro pasado como entrada. Utilizar un valor menor dejaría de lado ciertas características importantes, mientras que trabajar con un valor mayor hace que se capturen ciertas señales de ruido (conocidas como informantes falsos) y además incrementa el costo computacional del cálculo.

Las etapas principales de la Codificación Predictiva Lineal son las siguientes:

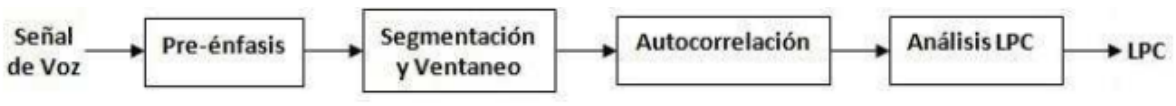

Figura 9 Diagrama de bloques para el cálculo de coeficientes LPC.

Fuente: [16]

La etapa de pre-énfasis y la de segmentación y ventaneo, como fueron explicadas en la sección de Bancos de Filtros de la presente monografía, tienen como objetivo preprocesar la señal de voz capturada para que la extracción de características pueda realizarse de forma efectiva. En la etapa de pre-énfasis se hace una selección de los elementos de la señal de voz que tienen una mayor importancia para el reconocimiento, descartando algunos datos para que el costo computacional del sistema disminuya. Por su parte, en la etapa de segmentación y ventaneo se divide la señal resultante del pre-énfasis en pequeñas tramas que constituyen la entrada del algoritmo LPC [16]. En la siguiente etapa se calculan los elementos de autocorrelación de cada trama enventanada, con el objetivo de analizar la periodicidad de las muestras obtenidas, logrando identificar cuáles son las tramas que se repiten o que pueden tener alguna similitud marcada. Finalmente, se calculan los coeficientes LPC a través del método recursivo de Levinson-Durbin [20]. Estos coeficientes son pasados como entrada a los Modelos Ocultos de Markov, Redes Neuronales o cualquier método que sea utilizado para identificar patrones y palabras. Estos métodos no toman como entrada la señal directa de voz, por lo que los algoritmos de LPC y MFCC tienen una importancia vital, pues logran extraer las características necesarias para que las diferentes herramientas de reconocimiento hagan su tarea de forma oportuna y eficiente.

7.3 Modelos Ocultos de Markov

Los modelos ocultos de Markov (HMM por sus siglas en inglés) son modelos probabilísticos que tienen la capacidad de representar procesos aleatorios paramétricos. Ya que el reconocimiento de patrones en la voz humana tiene determinadas propiedades que pueden tratarse e interpretarse a través de la estadística, los modelos estocásticos suelen ser muy utilizados en este

propósito, siendo precisamente los HMM el modelo estocástico más popular y de mayor éxito en ese sentido [21].

Los HMM tienen dos componentes fundamentales: Un proceso de Markov y un conjunto determinado de distribuciones de probabilidad que conforman la salida del modelo. La entrada de estos modelos es la voz humana, siendo esta codificada y almacenada en una secuencia de vectores con información espectral [22]. Esta secuencia de vectores permite observar de forma indirecta los estados del proceso de Markov, siendo esta la única forma posible en la que estos se pueden observar ya que se encuentran ocultos.

Según Villamil [21], los modelos ocultos de Markov utilizados en sistemas de reconocimiento de voz trabajan teniendo en cuenta dos hipótesis principales:

1. Es posible dividir la voz humana en segmentos y estados, permitiendo parametrizarla como un elemento estacionario. Esto significa que en los HMM la señal mantiene una estructura determinada de principio a fin en la ventana de análisis. Además, se asume que las transiciones entre estos segmentos son instantáneas.

2. La probabilidad de que se genere un vector de características depende únicamente del estado actual del modelo y no de símbolos anteriores.

Es importante tener en cuenta que estas dos hipótesis no son aceptadas en la comunidad científica como ciertas para las señales de voz. Sin embargo, los HMM trabajan asumiéndolas como ciertas y han logrado obtener resultados positivos considerables, al punto de ser considerados por muchos años como uno de los métodos principales para reconocer patrones y palabras en la voz humana.

7.3.1 Descripción Conceptual de los Modelos Ocultos de Markov

Un Modelo Oculto de Markov puede definirse como una máquina de estados finita y probabilística. Esto quiere decir que un HMM se trata de un conjunto de estados que están conectados con arcos, teniendo cada uno de estos un peso probabilístico. Por lo tanto, los Modelos Ocultos de Markov se representan desde la perspectiva matemática y computacional como un grafo dirigido. Para ilustrar este concepto se presenta la Figura 10:

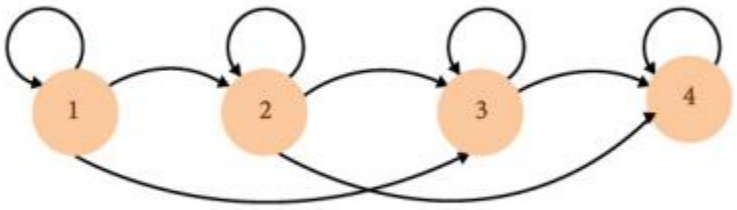

Figura 10 Representación gráfica de un Modelo Oculto de Markov

Fuente: [22]

El Modelo Oculto de Markov representado en la Figura 10 corresponde a un modelo de izquierda-derecha, también conocido como modelo de Bakis. Cómo es posible observar, todos los arcos de transición llevan desde determinado estado hacia sí mismo o hacia otro estado ubicado más a la derecha del modelo. De esta forma se representa una de las principales propiedades de los HMM: La secuencia oculta de estados incrementa su estado o permanece constante conforme transcurre el tiempo [22]. Las transiciones entre estos estados son representadas por los arcos del grafo, presentándose estas transiciones según la probabilidad asignada a cada uno de los arcos.

Es pertinente hacer una aclaración respecto al funcionamiento de los HMM: Los estados no son visibles para el observador externo. Lo que sí es visible es una salida que este estado genera,

teniendo cada uno de estos una función asociada para cada salida posible. En un instante discreto de tiempo se asume que el proceso está en algún estado y esto genera una observación que es dependiente de la función de probabilidad asociada al estado en el que se encuentra. En un instante siguiente, toda la cadena de Markov cambia de estado en coherencia con la matriz de probabilidades de transición entre los estados, siendo esta la representación computacional del grafo mostrado en la Figura 10. Cuando la cadena cambia de estado produce una nueva observación, siendo esto lo único que percibe el observador externo. Este observador solamente puede evidenciar la salida de las funciones probabilísticas que están asociadas a cada estado y no puede observar la secuencia completa de estados. Es por eso que este modelo recibe el nombre de "Modelo Oculto de Markov" [23]. Este tipo de modelos se utiliza en el habla porque es capaz de adaptarse a señales que tienen propiedades cambiantes a lo largo del tiempo, como la voz humana.

7.3.2 Algoritmos de Análisis Secuencial

Según Guevara [22], los Modelos Ocultos de Markov hacen uso de dos algoritmos de análisis secuencial para lograr reconocer patrones y palabras en la voz humana: El algoritmo de Viterbi y el algoritmo de Baum Welch. A continuación, se hará una explicación general de cada uno de estos algoritmos.

7.3.2.1 Algoritmo de Viterbi.

El algoritmo de Viterbi permite, a través de una observación, encontrar la secuencia de estados que tiene una probabilidad más alta en un Modelo Oculto de Markov [22], obteniendo así la secuencia óptima que mejor explica la secuencia de observaciones. Se trata de un algoritmo de programación dinámica y no iterativa, ya que si se tratara de un procedimiento iterativo, la

complejidad sería demasiado alta [23]. El resultado de este algoritmo es la mejor ruta entre todos los estados. Este resultado puede visualizarse en una matriz en la que la que cada fila corresponde a un estado del Modelo Oculto de Markov y cada columna corresponde a una unidad temporal, que para el caso del reconocimiento de voz también es conocida como marco del habla. La representación de los resultados de este algoritmo puede apreciarse en la Figura 11:

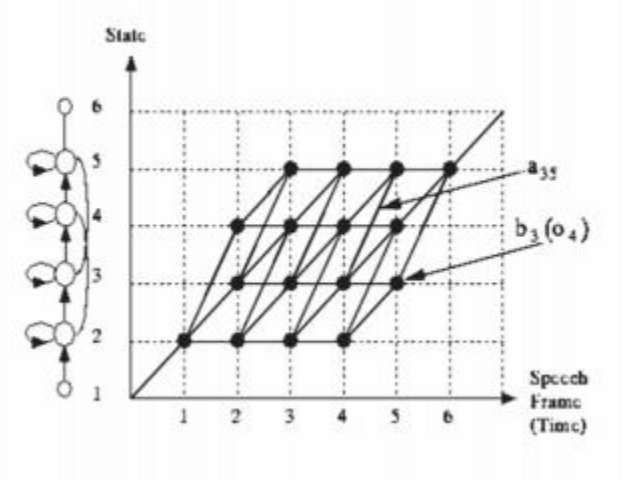

Figura 11 Representación gráfica del algoritmo de Viterbi.

Fuente: [22].

Como se mencionó anteriormente, el objetivo del algoritmo de Viterbi es encontrar la secuencia de estados más probable una vez que se tiene una secuencia de observaciones. Para encontrar esta tarea, el algoritmo debe realizar dos tareas importantes: Determinar el estado final más probable y generar una matriz en la que defina cuál es el estado anterior más probable para cada estado actual. Según lo explicado por [23], los pasos que el algoritmo sigue para realizar estas dos tareas y cumplir con el objetivo de encontrar la secuencia de estados más probable son los siguientes:

1. Determinar la probabilidad inicial de que en la primera observación se presente cada uno de los estados posibles. Esto se hace teniendo en cuenta la matriz de probabilidades para las

transiciones entre estados y la probabilidad de que se presente cada observación en cada uno de estos estados. Por ejemplo, para el modelo propuesto en la Figura 11, el único estado inicial posible es el estado 1. Esto simplifica el cálculo, porque simplemente se trata de calcular la probabilidad de que en ese estado se genere la primera observación. En caso de que haya más estados iniciales posibles, es necesario hacer este cálculo para cada uno de los estados y almacenarlo en alguna estructura de datos.

2. Para cada una de las transiciones posibles desde los posibles estados iniciales, se debe encontrar la probabilidad de que en cada siguiente estado se genere la observación correspondiente a esta iteración. Por ejemplo, para el modelo propuesto en la Figura 11 los estados posibles en una segunda observación son los estados 1, 2 y 3. En cada uno de estos estados se debe determinar la probabilidad de que se genere la segunda observación (en este caso). Es importante saber que para esto es necesario multiplicar esta probabilidad por la probabilidad asignada a la transición entre los posibles estados anteriores (estado 1 para el ejemplo) y los posibles estados actuales. De estas probabilidades es necesario elegir la ruta con mayor probabilidad. De esa forma se tiene, para cada estado posible en esa observación, la secuencia más probable desde el inicio. El estado anterior que tenga probabilidad mayor debe almacenarse en la matriz de estados anteriores. Por ejemplo, en caso de que en la segunda iteración se esté analizando el estado 2 (ejemplo de la Figura 11) es necesario analizar la probabilidad de que la ruta seguida haya empezado en el estado 1 (es el único estado inicial posible). En ese caso, el valor de la matriz de estados anteriores en la fila 2 corresponde y la columna 2 (segunda iteración) corresponde al estado 1.

3. Repetir el paso 2 para todas las observaciones. Por ejemplo, si son 5 observaciones, este proceso debe realizarse 5 veces. En resumen, en cada iteración se estaría encontrando la secuencia más probable hasta el estado correspondiente.

4. Una vez se llegue a la última observación se tendrá información suficiente para elegir cuál es el último estado más probable, según los caminos calculados en los pasos anteriores. Con toda esta información se debe determinar un último estado con probabilidad mayor.

5. Una vez que se determina cuál es el último estado más probable, se debe recorrer la matriz de estados anteriores para determinar cuál es la secuencia más probable que llega a este último estado. Esta secuencia sería la salida del algoritmo.

Para tener mayor claridad, a continuación se presentan los pasos enunciados anteriormente en forma de algoritmo formal:

Algorithm 1 Algoritmo de Viterbi

procedure VITERBI
1) Inicialización:
    $\delta_1(i) \leftarrow \pi_i b_i(O_1)$ ▷ i = 1,...,N
    $\psi_1(i) \leftarrow 0$
2) Recursión:
    $\delta_t(j) \leftarrow max_{i=1,...,N}\{\delta_{t-1}(i)a_{ij}\}b_j(O_t)$ ▷ t = 2,...,T y j=1,...,N
    $\psi_t(j) \leftarrow argmax_{i=1,...,N}\{\delta_{t-1}(i)a_{ij}\}$ ▷ t = 2,...,T y j=1,...,N
3) Terminación:
    $P^* = max_{i=1,...,N}\{\delta_T(i)\}$
    $q_T^* = argmax_{i=1,...,N}\{\delta_T(i)\}$
4) Recursión para obtener la secuencia de estados:
    $q_t^* = \psi_{t+1}(q_{t+1}^*)$ ▷ t = T-1, T-2,...,1

Figura 12 Algoritmo de Viterbi.

Fuente: [23].

A continuación se aclara el significado de cada una de las variables enunciadas en el algoritmo:

- N: Número de observaciones.
- T: Número de estados del modelo.
- a: Matriz de probabilidades de transición entre estados.

- b: Función de probabilidad para cada observación en el estado j (subíndice).
- $\boldsymbol{\delta}$: Estructura de datos en la que se guardan las probabilidades máximas de obtener la observación j en el estado actual t (subíndice).
- Ψ: Matriz de estados anteriores más probables.
- P*: Probabilidad de llegar al último estado más probable (estado con mayor probabilidad en la última observación).
- q*: Estructura de datos en la que se almacena la secuencia más probable para las observaciones presentadas. Es la salida del algoritmo.

7.3.2.2 Algoritmo de Baum Welch.

El segundo algoritmo de análisis secuencial utilizado por los Modelos Ocultos de Markov para el reconocimiento de voz es el algoritmo de Baum Welch. Este puede definirse como la sobreestimación de Modelo Oculto de Markov tomando como base otro Modelo Oculto de Markov [22]. Este algoritmo surge como respuesta a uno de los problemas principales de los HMM: Encontrar un modelo que maximice la probabilidad de una secuencia de observaciones. En ese orden de ideas, el objetivo es determinar cuál es el modelo que explica mejor la secuencia elegida como parámetro. Encontrar un modelo que cumpla con estas características no es considerado posible de forma analítica, por lo que es necesario utilizar algoritmos iterativos para cumplir esta tarea, tales como el Algoritmo de Baum Welch. Este algoritmo permite estimar los parámetros que hagan máxima la probabilidad de una secuencia de observables en un Modelo Oculto de Markov. La representación gráfica de este algoritmo puede verse en la Figura 13:

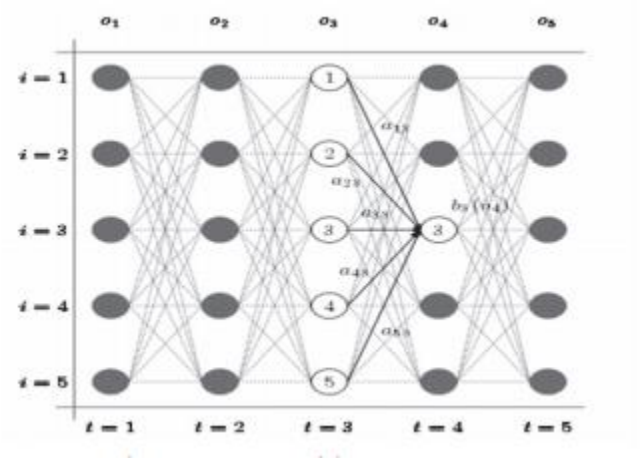

Figura 13 Representación gráfica del algoritmo de Baum Welch.

Fuente: [22]

En concordancia con lo anterior, el algoritmo de Baum Welch es usado para generar el Modelo Oculto de Markov propiamente dicho, con base en una serie de observaciones. Por ejemplo, para el caso del reconocimiento de voz se suelen generar modelos para cada palabra del vocabulario, siendo la digitalización de sus sonidos las observaciones que funcionan como parámetros del algoritmo [24]. Una vez que se tienen estructurados estos modelos, se podrá aplicar el algoritmo de Viterbi para definir las secuencias más probables y así tener una noción de los patrones más probables respecto a las palabras de entrada.

Según lo explicado por Ponce [25], el objetivo del algoritmo de Baum Welch es determinar el valor de λ; definiéndose este como λ = (π, A, B), siendo π la probabilidad del estado inicial, A la matriz de transiciones y B la matriz de estados ocultos del modelo. Por lo general, estos tres parámetros son desconocidos en un principio y deben ser estimados a través de una serie de observaciones proporcionadas.

El modelo generado por el algoritmo de Baum Welch debe definir la expresión $\xi t(i,j)$, pudiéndose interpretar esta como la probabilidad conjunta de que un proceso se encuentre en un estado *i* en un instante *t* y se genere una transición a un estado *j* en el instante *t + 1*. Como se

mencionó anteriormente, todo esto debe calcularse con base en una secuencia de observaciones O. Este algoritmo es del tipo forward-backward (adelanto-atraso), ya que sus cálculos no dependen solamente de la observación actual, sino que también analizan la observación anterior y la siguiente a la iteración en la que se encuentra.

El resultado del algoritmo de Baum Welch es un modelo similar al presentado en la Figura 13. Allí, según las observaciones representadas en el eje X, se establecen los estados que están representados en el eje Y. En cada transición se establecen las adyacencias entre estos estados y se define una probabilidad para estas transiciones. Después de realizar este proceso, se puede hacer una abstracción de la matriz de estados y la de adyacencias, armando así el Modelo Oculto de Markov deseado.

7.3.3 Modelos ocultos de Markov en los Sistemas de Reconocimiento de Voz

Los Modelos Ocultos de Markov tienen una amplia aplicación en los sistemas de reconocimiento de voz, principalmente porque el modelado del habla puede ajustarse de manera óptima a señales con propiedades que cambian repetidamente de estado mientras el tiempo transcurre [10]. Por eso es que este tipo de modelo es considerado como una herramienta casi imprescindible en la actualidad para el desarrollo de sistemas que reconocen patrones y palabras en la voz humana.

Una de las características de los problemas de reconocimiento de voz que hacen que los Modelos Ocultos de Markov sean óptimos para el caso es que, debido a la inercia propia de los órganos articulatorios, se puede suponer que las características de la señal de la voz no varían mucho en un intervalo de tiempo pequeño. En los sistemas de reconocimiento que utilizan HMM, se modela la evolución temporal de la secuencia de espectros que se le pasa como entrada al

Modelo Oculto y este contempla las diferentes fuentes de variabilidad de la señal. De esta forma es posible asociar los estados del Modelo Oculto de Markov a determinados tramos de la señal [23]. Los HMM permiten entonces que las funciones probabilísticas de los estados para la generación de observaciones modelan las características espectrales de cada tramo, mientras que las probabilidades de transición entre estados sirven para modelar la secuenciación y la duración de los sonidos. Así, mediante técnicas estocásticas, es posible realizar interpretaciones y predicciones basadas en las probabilidades de observación y transición que permitan reconocer palabras según las características espectrales y de duración en los sonidos emitidos por un ser humano. Todo esto se hace tomando como punto de partida una base de conocimiento que permite generar modelos ocultos para determinados fonemas o para palabras enteras.

7.4 Redes Neuronales Artificiales

Después de realizar el proceso de extracción de características de la voz humana, es necesario implementar modelos que permitan, con base en estas características obtenidas a través de algoritmos de MFCC o LPC, reconocer patrones, fonemas, sílabas o palabras en la voz humana. En este sentido, uno de los modelos computacionales más utilizados en la actualidad es el de las redes neuronales computacionales. En este apartado de la presente monografía se pretende dar algunas generalidades respecto a este modelo y explicar cómo puede aplicarse a los Sistemas Automáticos de Reconocimiento del Habla.

Con el interés de generar un procesamiento similar al del cerebro humano, funcionando este de manera no-lineal, compleja y paralela; se desarrolló de un modelo computacional en constante evolución que permite lograr dicho objetivo. Una red neuronal se caracteriza por

aprender de la experiencia, almacenándose esta de forma similar a como sucede en el cerebro. [26].

Desde el momento en el que Frank Rosenblant desarrolló e introdujo el modelo del perceptrón, las redes neuronales artificiales han tenido un gran avance, siendo hoy una herramienta poderosa de procesamiento en problemas de clasificación y reconocimiento de patrones, entre otros [27].

7.4.1 Definición

Una red neuronal artificial puede definirse como un sistema que posibilita definir una relación entre entradas y salidas que están inspiradas en el sistema nervioso humano [27]. Las salidas de estos sistemas se diferencian radicalmente de la computación tradicional, ya que no utilizan una algoritmia secuencial sino que la información se procesa en paralelo. Otra definición acertada es la que enuncia Kohonen [28], en la que se define a las redes neuronales artificiales como redes de múltiples elementos simples conectados de forma paralela, organizados de forma adaptativa y con determinada jerarquía. Estos elementos simples tratan de interactuar entre sí y con los objetos del mundo real, simulando de una forma simplificada y fiel el comportamiento del sistema nervioso humano. De hecho, a estos elementos simples se les conoce como neuronas.

7.4.2 Características

Las redes neuronales artificiales comparten una serie de características con el cerebro humano, entre las que es posible mencionar el aprendizaje, la generalización y la abstracción. Este modelo tiene la capacidad de auto ajustarse a lo largo de su proceso de aprendizaje y es capaz de generalizar sobre conjuntos de entradas cuyas variaciones sean mínimas. Además, las redes

neuronales artificiales tienen la capacidad de extraer características sobre un conjunto de entrada cuyos aspectos en común son nulos [29]. Según Izaurieta y Saavedra [26], las redes neuronales artificiales presentan las siguientes características:

Tienen una inclinación natural a adquirir el conocimiento por medio de la experiencia. Este conocimiento es almacenado, de forma similar al cerebro humano, a través del peso relativo de las conexiones entre las neuronas.

- Tienen una gran plasticidad y una adaptabilidad alta, siendo capaces de cambiar de forma dinámica según las modificaciones del medio y el contexto.
- Tienen un nivel alto de tolerancia a fallos. Podrían sufrir un daño considerable y, aun así, continuar teniendo un comportamiento aceptable. Esta característica es compartida con los sistemas biológicos.
- Tienen un comportamiento no-lineal predominante. Esto le permite procesar información que proviene de sistemas no-lineales; siendo este el comportamiento más frecuente en una gran cantidad de fenómenos naturales.

7.4.3 Estructura

Como se ha mencionado anteriormente, las redes neuronales artificiales buscan modelar de forma aproximada el comportamiento del sistema nervioso humano. Por eso, para entender algunos conceptos básicos de neurobiología, haciendo especial énfasis en el comportamiento y la estructura de la neurona como unidad mínima y fundamental. Una neurona tiene tres componentes principales: Las dendritas, el cuerpo o soma y el axón; conociéndose la conexión entre el axón de una célula y la dendrita de otra como sinapsis [27]. En la Figura 14 se puede apreciar gráficamente esta estructura:

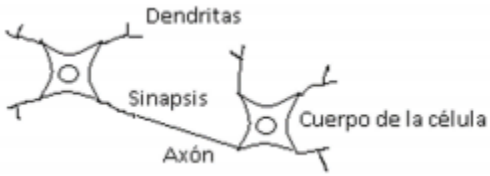

Figura 14 Componentes principales de una neurona.

Fuente: [27].

Para efectos del modelado computacional de las redes neuronales, Izaurieta y Saavedra [26] sugieren que se preste especial atención a dos comportamientos característicos de las neuronas biológicas:

- Los impulsos eléctricos de entrada y salida en la sinapsis no tienen la misma intensidad. La intensidad del pulso que sale del proceso de sinapsis depende de la cantidad de neurotransmisor que haya en la neurona. Estas magnitudes son modificadas constantemente en el proceso de aprendizaje, de tal forma que las señales resultantes de cierta capa de neuronas se ajusten a los conocimientos requeridos. Como se profundizará más adelante, este suceso se modela computacionalmente a través del peso que se le da a cada neurona.

- El núcleo de la neurona se suman todas las entradas provenientes de las dendritas. Si la suma de estas entradas sobrepasa determinado umbral, entonces esta señal será transmitida en el axón, negándose esta transmisión en caso de que el umbral no sea sobrepasado. Como se explicará más adelante, para modelar computacionalmente este comportamiento se utiliza una función de activación. Además, para el caso de las neuronas biológicas existe un intervalo de tiempo en el que no pueden volver a transmitir que oscila entre los 0,5 y los 2 ms. A este lapso se le conoce como periodo refractario.

Una sola neurona es diminuta en sí misma y se estima que, por ejemplo, son entre 5 y 6 veces más lentas que una compuerta lógica de silicio [26]. Sin embargo, el poder de procesamiento

que tiene el sistema nervioso humano está dado en que esta lentitud se compensa con la cantidad de neuronas, ya que se estima que en el cerebro hay aproximadamente 10 billones de conexiones sinápticas, haciendo que el sistema resultante sea muy eficiente. A este conjunto de conexiones es al que se le conoce como red neuronal [27]. Cuando se llevan estos sistemas al campo computacional no se pretende modelar de forma exacta el comportamiento fisiológico de cada neurona, sino que se quiere modelar correctamente el comportamiento de toda la red, pues esto es lo que realmente representa la forma en la que el ser humano aprende.

Cuando las redes neuronales son llevadas al campo de la computación, la neurona es representada a través de un elemento mínimo conocido como perceptrón o elemento procesador (PE por sus siglas en inglés). Este elemento tiene una cantidad determinada de entradas que son combinadas en su núcleo, normalmente a través de una suma básica en la que cada elemento es multiplicado por un peso específico proveniente de la neurona anterior (en algunos modelos esta multiplicación se hace directamente en el valor de salida de las neuronas). Esta suma de entradas es pasada como parámetro a una función de transferencia o de activación. El resultado de esta función es pasado directamente a la salida del perceptrón [29]. Normalmente, la salida de estos perceptrones es conectada como entrada a perceptrones de una capa siguiente, formando así la estructura de la red neuronal. El funcionamiento de cada elemento procesador puede verse con claridad en la siguiente gráfica:

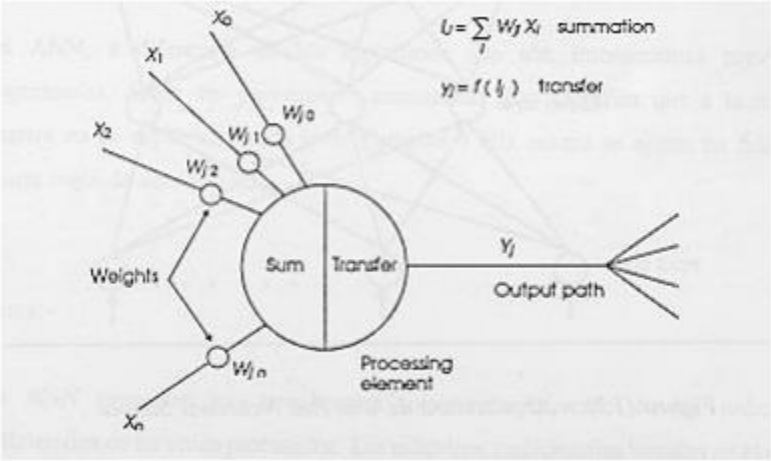

Figura 15 Diagrama de una neurona artificial.

Fuente: [29].

Las funciones de activación tienen el objetivo de modelar el concepto mencionado anteriormente en el que las neuronas "dejan pasar" la suma de pulsos de entrada si este supera determinado umbral [26]. Estas funciones suelen ser lineales, funciones a trozos según determinadas condiciones o funciones sigmoidales. Por ejemplo, a continuación se presentan las gráficas de dos de las funciones de activación más conocidas:

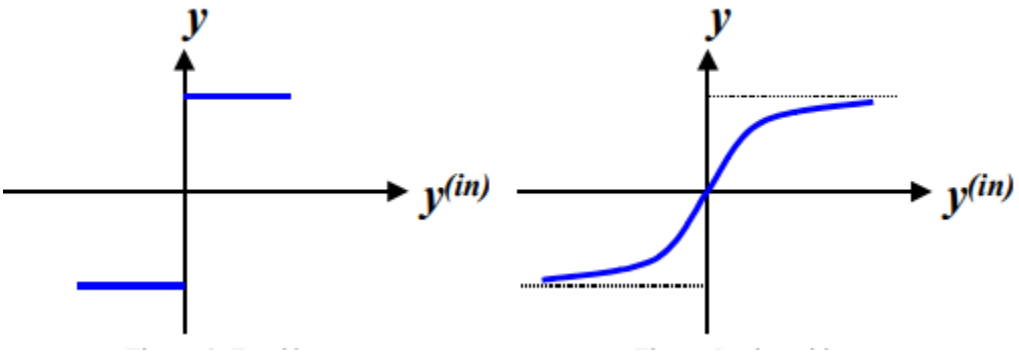

Figura 16 Función de activación de escalón (izquierda) y sigmoidal (derecha).

Fuente: [26].

Como se mencionó anteriormente, una red neuronal artificial es una serie de neuronas o perceptrones conectados entre sí. Para lograr una mayor eficiencia en el modelo computacional, estas neuronas suelen organizarse en capas de la siguiente forma:

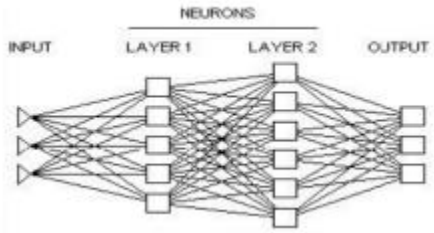

Figura 17 Capas de una red neuronal.

Fuente: [27]

En la Figura 17 presentada anteriormente es posible identificar 3 tipos de capas: Capa de entrada, capas ocultas y capa de salida. La capa de entrada es la que se encarga de recibir la información del exterior, las capas ocultas realizan el procesamiento necesario y la capa de salida presenta los resultados [26]. Si se hace la analogía con una neurona biológica, la capa de entrada tiene similitud con las dendritas, las capas ocultas tienen similitud con el soma o núcleo de la neurona y la capa de salida se asemeja al axón.

7.4.3 Aprendizaje

El aprendizaje o entrenamiento constituye la clave para que una red neuronal artificial funcione de forma correcta. Desde la óptica computacional, en esta etapa se ajustan los pesos sinápticos de cada memoria de tal forma que los resultados obtenidos sean los esperados. A través del aprendizaje es que las redes neuronales artificiales adquieren plasticidad y son capaces de adaptarse a los distintos estímulos del medio [26]. Acevedo, et. al. [27] clasifican y explican los

diferentes tipos de aprendizaje que soportan las redes neuronales de la siguiente manera, teniendo en cuenta que el aprendizaje supervisado es el más utilizado:

Aprendizaje supervisado: Se realiza un entrenamiento a la red neuronal que está estrictamente supervisado por el diseñador de la misma, determinando fácilmente si la respuesta que está arrojando el sistema es correcta según la entrada que se le pasa como parámetro. En caso de que los resultados obtenidos en la salida sean erróneos, se hace una modificación a los pesos de las neuronas para aproximar la salida a lo que se espera. El proceso se repite hasta que la salida del sistema sea lo más aproximada posible a un resultado correcto.

Aprendizaje por corrección de error: En el entrenamiento se le presenta a la red neuronal una serie de entradas con su correspondiente salida deseada. Este tipo de aprendizaje tiene como objetivo que exista una diferencia mínima entre la salida que se obtiene y la que se desea. Se hace una comparación estricta entre las salidas y según este resultado se ajustan los pesos en las conexiones de red.

Aprendizaje por refuerzo: En este tipo de aprendizaje supervisado no se le pasa a la red la salida esperada de forma específica, sino que según los resultados se le envía una señal de refuerzo en la que se indica si el resultado fue correcto o no, valiéndose para esto normalmente de una variable booleana. Así, mediante un mecanismo de probabilidades, se ajustan los pesos de tal modo que la red neuronal se acerque a los resultados deseados. El aprendizaje por refuerzo suele ser más lento que el aprendizaje por corrección de error, pues es probable que se reciban muchas señales de fracaso consecutivas y esto hace que los pesos tengan que ajustarse en un número elevado de ocasiones.

Aprendizaje estocástico: En este tipo de aprendizaje se realizan cambios de forma aleatoria en los pesos de las neuronas para cada iteración, comparando la salida obtenida con la

salida deseada. Cuando la diferencia entre la salida obtenida y la salida esperada es mínima, se concluye que la red ya aprendió lo suficiente. En caso de que haya diferencias significativas entre la salida obtenida y la deseada, se hacen cambios en los pesos según una distribución probabilística diseñada específicamente para que los resultados sean óptimos. Si después de que se realice este cambio la salida obtenida se acerca más a la deseada, el cambio es aceptado y guardado. En caso de que el resultado se aleje de lo correcto, entonces el cambio es descartado y se vuelve a la imagen anterior de pesos.

Aprendizaje no supervisado: En este tipo de aprendizaje no hay una supervisión y no existe una comparación entre salidas obtenidas y esperadas. En este tipo de aprendizaje se le da una gran autonomía a la red neuronal, pues esta descubre por sí misma características, categorías, correlaciones y regularidades en los datos de entrada. El objetivo del algoritmo de entrenamiento es modificar los pesos de la red neuronal de forma que los vectores de salida generados sean coherentes. La mayoría de los algoritmos de aprendizaje no supervisado están basados en el algoritmo de Hebb.

7.5 Lógica Difusa

7.5.1 Generalidades

Existen determinados fenómenos del entorno del ser humano que no es posible representar de forma correcta en el marco clásico de la teoría de probabilidades. Por ejemplo, en esta categoría se pueden incluir los casos en los que es necesario describir parámetros a través de expresiones lingüísticas. En estas situaciones puntuales es difícil utilizar técnicas matemáticas tradicionales para inferir con base en este tipo de expresiones [30]. Contrario a la lógica tradicional, las formas

de racionamiento del ser humano en determinados casos no son binarias, permitiéndole esto tomar decisiones complejas en ambientes hostiles.

Teniendo en cuenta lo anterior, el ingeniero y matemático iraní Lofty A. Zadeh formuló el modelo de lógica difusa en 1965. Se puede decir de manera formal que, si la lógica es la ciencia de los principios formales y normativos del razonamiento, la lógica difusa trata de estudiar los principios formales del razonamiento aproximado, teniendo la lógica clásica únicamente como límite [31]. En ese orden de ideas, la lógica difusa puede definirse como una lógica de múltiples variables en la que es posible representar de forma matemática la incertidumbre y la vaguedad, dando además herramientas formales que permiten tratarlas y realizar inferencias a partir de datos que serían imprecisos en el contexto de la lógica clásica. Se dice que cualquier problema puede resolverse dando un conjunto de entradas que finalmente producen un conjunto de salidas. En ese contexto, la lógica difusa permite lograr salidas más cercanas a la realidad que en modelos tradicionales, ya que atiende a criterios de significado y no de precisión [32].

En el marco de la lógica difusa se utiliza el concepto de conjuntos difusos como una generalización de la teoría de conjuntos convencional. Esta herramienta permite representar la borrosidad de muchos sucesos cotidianos que atañen a la vida del ser humano. Cuando se utiliza lógica aristotélica, un elemento puede pertenecer o no a un conjunto dado, siendo este un sistema totalmente binario. Por el contrario, en el paradigma difuso es posible que un elemento pertenezca a diferentes conjuntos en diferentes grados [33]. Estos conjuntos difusos permiten representar de forma matemática variables lingüísticas que en la vida cotidiana suelen representarse con palabras. Para ilustrar esta noción se presenta la siguiente figura:

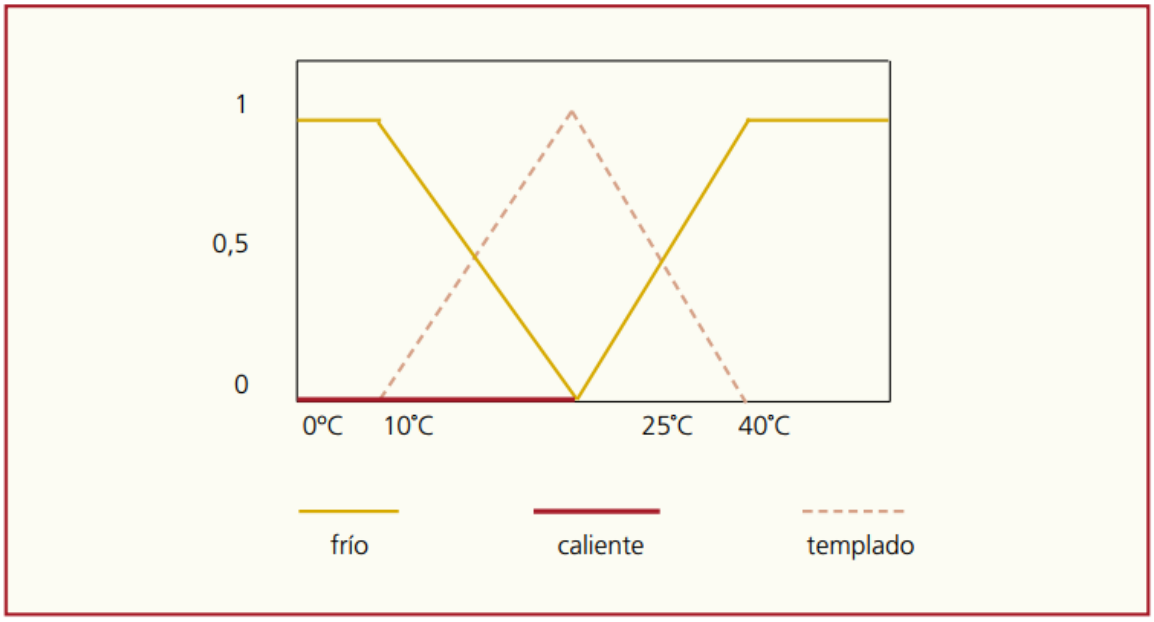

Figura 18 Representación gráfica de una clasificación de temperaturas usando conjuntos difusos.

Fuente: [33].

En la Figura 18 puede verse un sistema de clasificación de temperaturas con tres conjuntos difusos: frío, caliente y templado. En la vida cotidiana, según la subjetividad del observador, es posible clasificar determinado valor dentro de estos 3 conjuntos difusos en diferentes proporciones. El eje X de la gráfica representa la temperatura en grados Celsius y el eje Y representa la pertenencia de cada valor de temperatura a uno de los conjuntos difusos. Por ejemplo, para un contexto de 15°C se podría decir que la temperatura es fría y templada en diferentes medidas.

Los sistemas de control que trabajan con lógica difusa permiten tomar decisiones según determinadas reglas de pertenencia a conjuntos difusos. En un sistema de control real con lógica difusa los datos son captados del medio (normalmente con sensores y se someten a un proceso llamado fuzzificación, en el que se les asignan ciertas funciones de pertenencia similares a las mostradas en la Figura 18, aunque normalmente son más complejas. A los valores resultantes de

la fuzzificación se les aplican ciertas inferencias y el resultado difuso es asociado con un valor numérico, dando paso así al proceso de defuzzificación. Este proceso puede ser evidenciado de forma más específica en la siguiente figura:

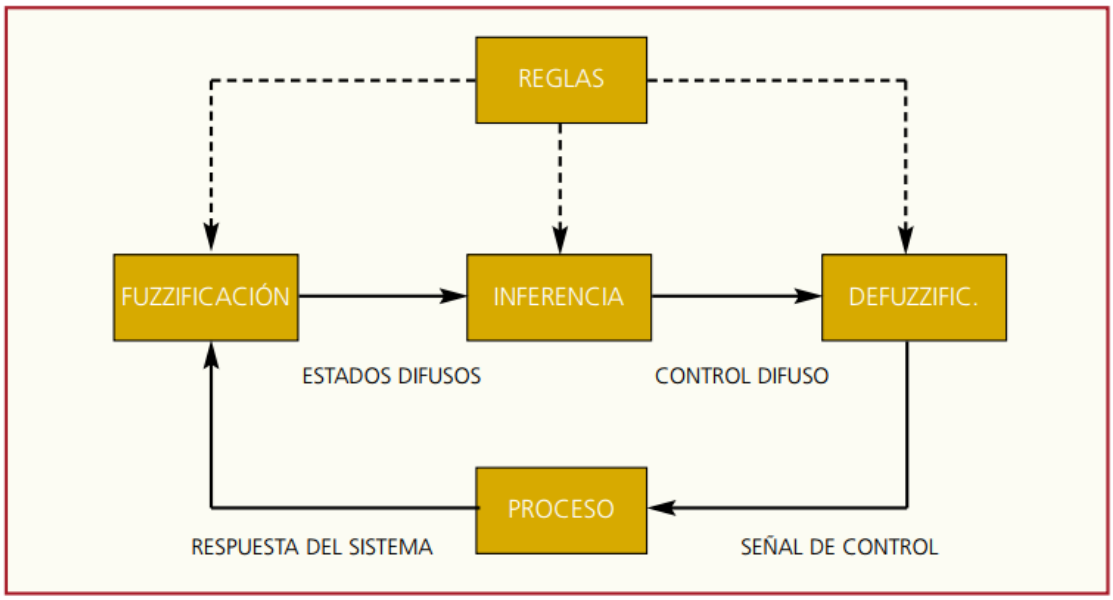

Figura 19 Diagrama de funcionamiento de un sistema de control difuso.

Fuente: [33]

Como se puede ver en la Figura 19, las reglas de pertenencia de los conjuntos difusos tienen una influencia directa sobre los tres procesos principales: Fuzzificación, inferencia y defuzzificación. Además, en la figura queda claro que las decisiones tomadas (respuesta del sistema) son consecuencia de algún proceso determinado que se le realiza a la señal de defuzzificación.

7.5.2 Lógica Difusa y Reconocimiento de Voz

El reconocimiento automático del habla es un problema cuyas soluciones se encuentran en constante investigación y desarrollo. A lo largo de la historia se ha trabajado con diferentes

enfoques que permiten crear sistemas automáticos de reconocimiento de voz, entre las que es posible mencionar los Modelos Ocultos de Markov, las Redes Neuronales Artificiales, la Codificación Predictiva Lineal, entre otros. Sin embargo, se dice que estos enfoques tienen un problema en común: La alta sensibilidad al ruido ambiente. Esta dificultad trae como consecuencia la necesidad de plantear algoritmos nuevos que permitan detectar y retirar el ruido, desarrollando además una compensación correcta entre la velocidad y la precisión del reconocimiento de voz [30]. Diversos estudios plantean que es posible solucionar este problema del aislamiento del ruido ambiente y mejorar la proporción entre velocidad y precisión del reconocimiento de voz a través de lógica difusa. En este apartado de la presente monografía se hará un recuento de las principales técnicas utilizadas para este fin.

La mayoría de sistemas de reconocimiento del habla basados en lógica difusa toman como base una representación visual del sonido, ya que los conceptos de lógica difusa han sido exitosamente utilizados en sistemas de reconocimiento de imágenes. Para esto se suele utilizar normalmente el espectrograma de la señal de entrada [30]. El espectrograma es una representación visual de las diferentes frecuencias auditivas que hacen parte de un intervalo de tiempo. El eje horizontal representa los instantes de tiempo que se miden y el eje vertical representa un rango de frecuencias. Para cada instante de tiempo se representan diferentes intensidades en el rango de frecuencias. El espectrograma se diferencia de la señal de representación acústica típica en que el primero aporta información referente a la frecuencia de sonido, mientras que la representación típica solamente permite apreciar la amplitud de la onda en un periodo de tiempo [31]. En ese orden de ideas, a pesar de que el espectrograma solamente tiene dos ejes, es posible decir que se trata de un gráfico tridimensional en el que se tienen en cuenta las siguientes dimensiones: Tiempo

(eje horizontal), frecuencia (eje vertical) e intensidad (color). Para dejar más claro este concepto se presenta la siguiente Figura:

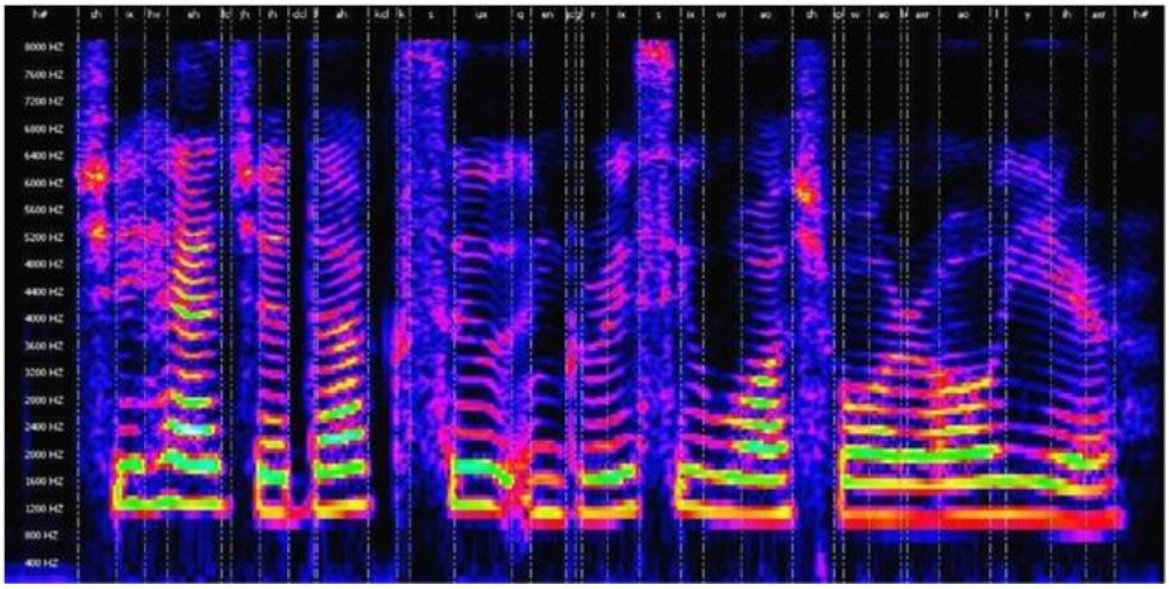

Figura 20 Espectrograma de ejemplo.

Fuente: [31].

En el espectrograma de la Figura 20 se puede observar que las frecuencias que más intensidad tienen en el audio capturado son aquellas que están en el rango entre 1200 y 2000 Hz. Se registran sonidos en otras frecuencias del espectro, pero estos son los que tienen una intensidad mayor.

Teniendo en cuenta que es posible representar una señal de audio de forma gráfica a través de un espectrograma, se hace posible aplicar las mismas técnicas de lógica difusa utilizadas en reconocimiento de imagen para reconocer patrones y palabras en la voz humana. Existen dos motivaciones principales para desarrollar modelos de lógica difusa en sistemas de reconocimiento del habla: En primer lugar, de forma parecida al cerebro humano, un sistema de reconocimiento de voz puede ignorar el ruido, en lugar de detectarlo y eliminarlo. De esta forma es posible utilizar la inferencia difusa y el reconocimiento de patrones para obtener resultados acertados en contextos

con altos niveles de ruido. En segundo lugar, se puede afirmar que cuando se trata de representar sucesos de la vida cotidiana, el exceso de cálculos precisos no representa necesariamente que se obtengan resultados precisos, ya que los problemas cognitivos (como el reconocimiento del habla) tienen una naturaleza difusa [30].

7.5.3 Sistemas Neurodifusos

Como se mencionó anteriormente, una de las estrategias más utilizadas para resolver problemas de reconocimiento del habla haciendo uso de la lógica difusa es utilizando los espectrogramas. De esta forma, el problema de reconocimiento de voz se convierte automáticamente en un problema de reconocimiento de imágenes. Teniendo en cuenta esto, es común que se combinen características de las redes neuronales y de los sistemas difusos, dando origen así a los sistemas neurodifusos. En este punto, suelen ser de amplio uso los modelos concurrentes, en los que las redes neuronales y los sistemas difusos trabajan juntos y se asisten entre ellos [34].

Uno de los algoritmos neurodifusos más utilizados es el ANFIS (Adaptive-Network-based Fuzzy Inference System). Este algoritmo permite construir un conjunto de normas fuzzy if-then, en las que según los resultados del proceso de fuzzificación se toman determinadas decisiones de inferencia. Estas funciones de membresía a los conjuntos difusos son las que generan los pares de entrada y salida que se pasan como parámetros de entrenamiento a la Red Neuronal Artificial.

8. Características y Aplicaciones de los Sistemas de Reconocimiento de Voz

Las diferentes aplicaciones de los sistemas de reconocimiento de voz difieren cuando se analiza si el sistema reconoce entradas continuas o entrecortadas de audio, siendo las segundas más fáciles de analizar puesto que se reconoce de una manera más eficiente y sencilla el límite entre cada palabra, haciendo muy importante que el locutor realice pausas lo suficientemente extensas como para que se identifiquen de esta forma.

Por otro lado, como ya se analizó en el capítulo 4 el reconocimiento de voz continuo necesita de algoritmos lo suficientemente complejos como para requerir la unión de dos técnicas como es el caso de los MLP y HMM.

A lo largo del capítulo se hará un recorrido donde se detallan las características y usos de los sistemas de reconocimiento de voz actuales.

8.1 Características de los Sistemas de Reconocimiento de Voz en la Actualidad

En la actualidad los sistemas de reconocimiento de voz comparten en muchas ocasiones una serie de características que han demostrado que facilitan y producen mejores resultados al momento de realizar el reconocimiento, como lo es el uso de algoritmos HMM con ayuda del algoritmo de Viterbi expuesto en el capítulo 5.

Como etapa inicial todos los sistemas requieren que se haga un pre-procesamiento de la señal recibida, siendo utilizadas diferentes técnicas como lo son el MFCC, la cual se basan en un análisis del oído humano el cual no puede percibir frecuencias sobre los 1Khz, para esto, se utilizan dos tipos de filtros uno linealmente espaciado para las frecuencias que están por debajo de los 1000 hz, y un segundo filtro logarítmico para las frecuencias altas que superan los 1000 hz. Todo esto

debido a la importancia que implica la primera etapa del reconocimiento de voz, teniendo en cuenta que entre mejor sea la parametrización que se le realiza a la señal, mejor serán los resultados del reconocimiento de voz [35].

Técnicas como el DTW adicionalmente permiten que dos series de tiempo con la misma información pero diferente duración puedan ser identificadas como similares, gracias a la deformación no lineal de alguna de las dos señales [35].

Posteriormente se realiza el procesamiento de dicha información con ayuda de los algoritmos o modelos previamente entrenados, como la ayuda de los diferentes algoritmos previamente estudiados explicados a través del capítulo 5 como lo son, los Bancos de Filtros, la Codificación Predictiva Lineal, los Modelos Ocultos de Markov, las Redes Neuronales Artificiales y Lógica Difusa.

Finalmente, los sistemas de reconocimiento de voz cuentan con interfaces desarrolladas para facilitar su uso a los usuarios del mismo, en el caso de los sistemas que se basan en HMM es común encontrar que la interfaz es desarrollada mediante HTK, el cual es un software patentado para desarrollar aplicaciones con HMM, como el mostrado en la Figura 21.

Figura 21 Interfaz de usuario en HTK. [21]

Otras interfaces son desarrolladas bajo el software de Matlab, como la mostrada en la figura 21, para un programa de reconocimiento de voz con aplicaciones domóticas.

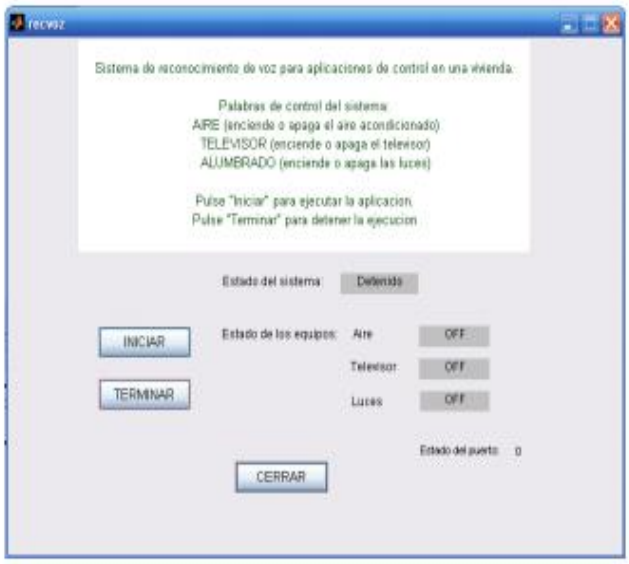

Figura 22 Interfaz de usuario en Matlab. [36]

8.2 Usos y Aplicaciones de los Sistemas de Reconocimiento de Voz

Son diversas las aplicaciones que han tenido los sistemas de reconocimiento de voz en la actualidad, pasando por áreas como la domótica, hasta áreas como la medicina, todas con el objetivo de ayudar a los usuarios a tener un control más indirecto con diversas actividades que realicen cotidianamente, a continuación se dará un paso por aplicaciones realizadas en las áreas previamente mencionadas.

8.2.1 Asistentes de Voz

Impulsados por el sueño de hablar y comunicarse con computadores, smartphones o cualquier dispositivo electrónico, nacen asistentes de voz como Siri (2011), Cortana (2013), Alexa

(2014) o Google Assistant (2016), siendo estos los asistentes más desarrollados y comercializados en la actualidad. [37]

Si bien los sistemas de reconocimiento de voz han ayudado a que cada una de estas herramientas tenga una mejora continua en su funcionamiento, es importante clarificar que el procesamiento del lenguaje natural juega un papel muy importante al momento de hablar de asistentes de voz. Es así como Google Assistant le permite al usuario formular distintas preguntas que pueden tener una misma respuesta como lo son por ejemplo ¿Donde yo parqué?, ¿Dónde deje mi carro? o ¿Recuerdas donde parquié? Bajo el asistente de Google Assistant, por ejemplo, se obtiene siempre la respuesta adecuada sin dar lugar a ambigüedades [37].

Entre las principales funcionalidades de los asistentes de voz en la actualidad es posible mencionar la escritura y lectura de mensajes de texto, realización de llamadas, escritura y lectura de correos electrónicos, programación de alarmas o eventos, narración de chistes o historias, programación de recordatorios y diversas aplicaciones en domótica gracias al internet de las cosas o IoT por sus siglas en inglés [37].

8.2.2 Aplicaciones en Domótica

Los diversos avances en sensores y desarrollo de sistemas domóticos -entendiendo un sistema domótico como un sistema inteligente que permite integrar diversas áreas del hogar-, han permitido que se incursione en procesos automatizados que emplean la voz humana como medio de activación [38].

Así pues, constantemente se busca la forma de que por medio de comandos de voz se puedan automatizar procesos como confirmaciones de acceso a la vivienda, la administración y realización de tareas cotidianas, o actos tan simples como activar y desactivar luces, electrodomésticos o sistemas inteligentes de riego; tareas que, al ser realizadas de forma automática o mediante órdenes emitidas por el usuario, permiten hacer realidad el concepto de vivienda inteligente. A continuación se listan algunas características y funcionalidades principales de este tipo de sistemas [41]

- Control de iluminación.
- Iluminación por detección de presencia.
- Automatización de persianas.
- Control y gestión de energía.
- Control de seguridad.
- Controles técnicos.
- Sistemas de mensajería.
- Realización de acciones preventivas.
- Climatización.
- Control de riego.
- Control de electrodomésticos.

8.2.2.1 Sistema de Reconocimiento de Voz para Aplicaciones de Control en una Vivienda.

Artículo presentando dentro del marco de la convención científica de ingeniería y arquitectura para el 7° Congreso Iberoamericano de ingeniería mecánica, Elizabeth Duarte y Bárbaro López [42], ambos pertenecientes a la Universidad de Pinar del Río Hermanos Saíz Montes de Oca exponen el diseño e implementación de un sistema de reconocimiento de voz con las capacidades de encender y apagar diversos electrodomésticos, apuntando al ideal de las casas inteligentes.

8.2.2.1.1 Procesamiento de Señal..

Con la mezcla de diferentes técnicas de procesamiento ya discutidas, Duarte y López [42] dividen el proceso en pre-procesamiento y extracción de coeficientes de reflexión, siendo el último el método por el cual se hace la caracterización de la señal previamente procesada con ayuda de técnicas ya expuestas a lo largo del documento, se realiza todo el proceso con el objetivo de obtener coeficientes que se aproximen a la señal de la voz, coeficientes que representan rasgos lineales del habla.

8.2.2.1.2 Alineamiento Temporal Dinámico.

Gracias a que el DTW emplea procedimientos matemáticos sencillos, permite que el procesamiento no sea costoso, procedimiento que se encarga de comparar dinámicamente las características extraídas de la señal con las características de la palabra de referencia, este algoritmo nos ayuda a evitar las variaciones de tiempo en entre dos palabras en las cuales varía el modo y/o la velocidad del habla [42]. La representación de este se puede observar en la Figura 23.

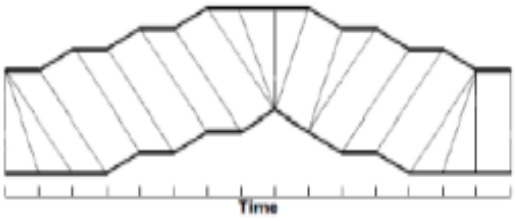

Figura 23 Alineamiento de dos señales con DTW

8.2.2.1.3 Implementación del Software.

El sistema se realizó con la intención de identificar las palabras, televisor, alumbrado y aire para así saber sobre qué electrodoméstico se desea actuar. La implementación final fue realizada mediante el software de matlab, el cual presentó etapas de entrenamiento y reconocimiento, ambos algoritmos expuestos en la Figura 24.

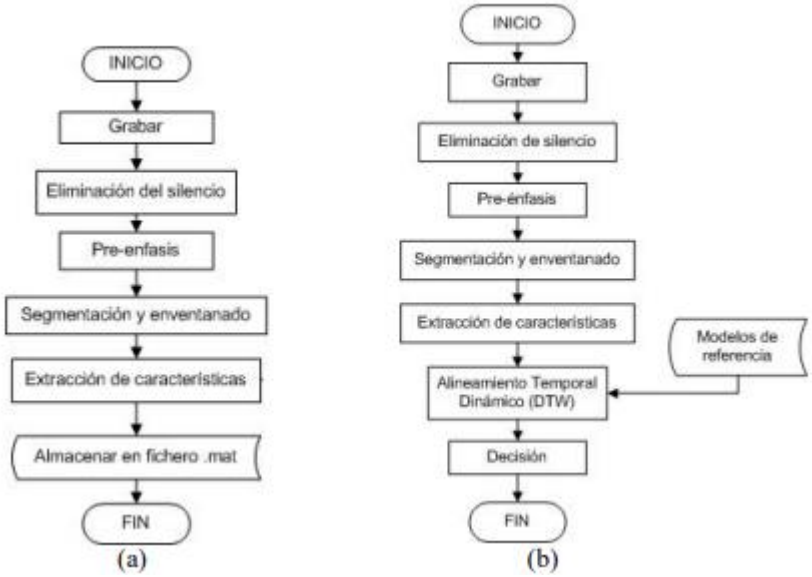

Figura 24 A) Algoritmo de entrenamiento B) Algoritmo de reconocimiento.

Fuente: [42]

Para la implementación se grabó bajo frecuencias de 11Khz y las grabaciones realizadas presentaron una duración de 3 segundos de voz. Posteriormente segmentaron la grabación en

tramas de 20ms, solapando las consecutivas 10ms. Luego de realizarla extracción de características, se obtuvieron 13 coeficientes con los cuales trabajaron.

8.2.2.1.4 Resultados.

Como se puede observar en la figura 25, se alcanzaron finalmente en condiciones reales porcentajes de acierto entre el 90% y 100%, bajo tiempos de ejecución menores a 1 segundo para el caso de la palabra Alumbrado.

Palabra	*Condiciones de laboratorio*		*Condiciones reales*	
	Aciertos	*%*	*Aciertos*	*%*
Aire	19	95	18	90
Televisor	20	100	19	95
Alumbrado	20	100	20	100
Promedio	19.7	98.3	19	95

Figura 25 Porcentaje de aciertos en condiciones de laboratorio y reales.

Fuente: [42]

Función	*Tiempo de ejecución*
Eliminación del silencio	0.047 s
Pre-énfasis	0.016 s
Segmentación y enventanado	0.109 s
Cálculo de los coeficientes	0.156 s
Cargar la base de datos	0.016 s
DTW	0.552 s
Total	0.896 s

Figura 26 Tiempos de procesamiento para la palabra Alumbrado.

Fuente: [42]

8.2.2.1.5 Conclusiones del Artículo.

En el artículo se concluye analizando la importancia del uso del DTW en su implementación debido a que aun con técnicas más recientes para el reconocimiento de voz, esta les permitió alcanzar porcentajes del 95% de precisión, aun realizando pruebas en ambientes contaminados con otras fuentes de ruido acústico.

8.2.3 Aplicaciones en Medicina

En el campo de la medicina, entre otras aplicaciones, se ha desarrollado una silla de ruedas inteligente controlada por voz y equipada con dos computadores con 64mb y 256mb de memoria RAM, un sensor láser de proximidad, una pantalla y un micrófono. Esta le permite al usuario desplazarse y hacer uso de su silla de ruedas gracias a 12 comandos de voz, los cuales son **Dusíla, Anda, Adelante, Detrás, Izquierda, Derecha, Lejos, Cerca, Medio, Para, Inicia y Termina,** [11].

Figura 27 Silla de Ruedas Inteligente

Fuente: [11].

El reconocedor de voz de la misma hace uso de dos componentes: uno que se encarga de reconocer las palabras y otro que se encarga de traducir las órdenes dichas por el usuario. El módulo de reconocimiento de voz está compuesto por dos estructuras: el módulo de lenguaje, en el que se describen las palabras a usar, y el modelo acústico, donde se describe o se entrena al reconocedor de voz a interpretar lo dicho por el usuario [11].

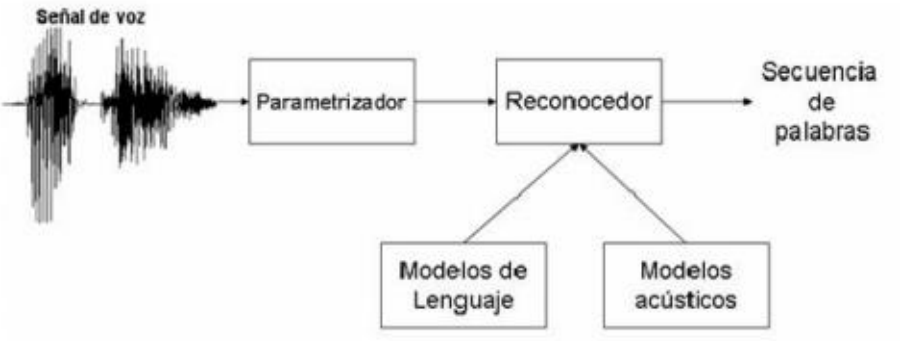

Figura 28 Diagrama del módulo de reconocimiento de voz

Fuente: [11].

Otra problemática que se ha atacado mediante el desarrollo de sistemas de reconocimiento de voz es la latente necesidad que tienen los adultos mayores en la ingesta de medicamentos, particularmente cuando se analiza la frecuencia de administración. Moguel y Azabal [39] se ocupan de esta problemática a través del desarrollo de una arquitectura de hardware y un sistema de software que permiten al paciente interactuar con el sistema general. Esto se logra con la adición de un módulo que permite a los profesionales de la salud realizar la administración y configuración remota del dispositivo mediante WiFi o Bluetooth, como lo ilustra la figura a continuación.

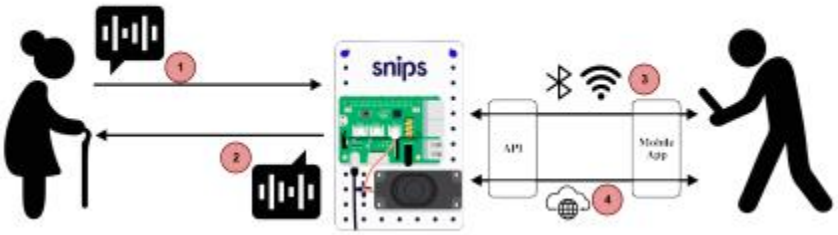

Figura 29 Asistente de voz para el recordatorio de tratamiento farmacológico

Fuente: [39]

8.2.3.1 Diseño de un Sistema de Reconocimiento de Voz para un Brazo Robótico para Cirugía Laparoscópica.

La tesis presentada por Ricardo Pastor Hernández en el 2018 [39] consta del diseño y desarrollo de un sistema de reconocimiento de voz para controlar un brazo robótico en cirugías laparoscópicas, la cual comprende la 3ra fase del procedimiento general que es el sistema de reconocimiento de voz.

8.2.3.1.1 Metodología.

La metodología de la tesis consta de los elementos habituales como lo son el planteamiento del problema, el cual nace como la necesidad de disminuir tiempos de recuperación, costos y estética al realizar la cirugía laparoscópica mediante de un sistema de reconocimiento de voz, evitando así el trabajo del ayudante del médico general, el cual se encarga de mover la cámara y demás enfermeros asistentes. Permitiendo así que el médico general, con ayuda del software pueda realizar todo el trabajo.

Adicionalmente se exponen los diversos objetivos, preguntas de investigación e hipótesis frente a la misma, finalizando con la definición de alcances y límites de la investigación.

8.2.3.1.2 Diseño del Algoritmo

Para el diseño del algoritmo Pastor [39], tuvo a consideración dos algoritmos de reconocimiento automático del habla, en donde los pasos del primero consisten en tomar la señal de audio, digitalizarla, hacerle un pre-énfasis, aplicarle unas ventanas de hamming, autocorrelacionado, calcular los coeficientes del LPC y finalmente tomar una decisión. Y el segundo consiste a grandes rasgos en realizar un pre-procesado, extraer las características de la señal como lo son coeficientes LPC y coeficientes polinomiales ortogonales, para finalmente con ayuda de la base de datos, y un DTW reconocer la palabra en cuestión.

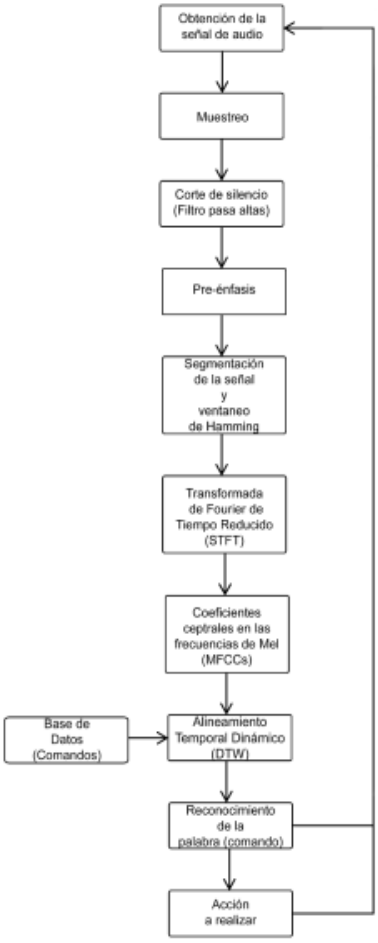

Figura 30 Algoritmo para el tratamiento de la señal [39]

En conjunto estos dos algoritmos sirvieron de bases e inspiración para que Pastor [39] creara el algoritmo expuesto en la Figura 30.

8.2.3.1.3 Implementación del Algoritmo de RAH

La implementación del algoritmo se llevó a cabo en una Raspberry pi 3, la cual contaba con un sistema operativo Ubuntu MATE de 64 bits 16.04, un micrófono, y una tarjeta de sonido manhattan de tipo USB, tal y como se muestra en la figura 31.

Figura 31 Sistema para obtener la señal de audio [39]

Finalmente, la programación y ajustes finales, fue realizada con ayuda del lenguaje de programación Python en su versión 3.

8.2.3.1.4 Pruebas de Funcionamiento

El funcionamiento experimental consistió en que el usuario en primer lugar se posiciona a una distancia aceptable del micrófono, posteriormente, presiona el botón, procede a pronunciar la

palabra, una vez la pronunciaba dejaba de presionar el botón, y se procedía a esperar la respuesta mediante una serie de luces led, que simulaban cada acción que debía realizar el sistema, la figura 32 muestra el proceso de funcionamiento.

Los resultados obtenidos luego de las pruebas en el ambiente controlado muestran que se obtiene un porcentaje del 84% en el peor de los casos, cuando se trata de la palabra Izquierda, pero en el mejor de los casos se obtiene un 93% como es el caso para la acción parar. La lista de comandos del sistema es Arriba, Abajo, Izquierda, Derecha, Avanzar, Atrás, Parar, con un tiempo promedio de determinación de si es o no la palabra de 2.19 segundos.

Figura 32 Proceso de funcionamiento del sistema [39].

8.2.3.1.5 Conclusiones de la Tesis

Finalmente, Pastor [40] concluye analizando como el sistema es funcional y soporta una variedad de señales de voz, pero adicionalmente aclara que se propone a futuro realizar una optimización en los tiempos de ejecución, adicionalmente menciona como es viable la manipulación de un BCRL que era uno de los objetivos principales, ya que, al realizar las simulaciones, el sistema respondía de forma positiva en un 89.2% de las pruebas realizadas.

9. Análisis comparativo de las técnicas y herramientas algorítmicas usadas para el desarrollo de sistemas artificiales de reconocimiento del habla

En este apartado se presenta un análisis en el que se pretende establecer un paralelo entre las diferentes herramientas algorítmicas y técnicas matemáticas analizadas en la presente monografía para el desarrollo de los sistemas artificiales de reconocimiento del habla. En la siguiente tabla comparativa se puede apreciar dicho análisis:

Técnica	Tipo	Características principales
Banco de filtros	Físico - Matemático	• Utiliza herramientas matemáticas y físicas (como la Transformada Discreta de Fourier) para digitalizar la señal de voz. • Su objetivo es obtener un vector que contenga las principales características de la señal de voz capturada en medios análogos. • Constituye una etapa previa al proceso de reconocimiento, pues proporciona la entrada a los diferentes modelos que reconocen los fonemas y las palabras. • No efectúa un reconocimiento completo, pero es una etapa previa indispensable para que los sistemas de reconocimiento del habla cumplan su objetivo correctamente. • Está presente en una gran cantidad de sistemas, sin importar la técnica que se utilice para la etapa del reconocimiento.
Codificación Predictiva Lineal	Matemático – Recursivo	• De la misma forma que en el banco de filtros, su objetivo es almacenar las principales características de la señal de voz en determinada estructura de datos que los diferentes algoritmos de reconocimiento puedan interpretar y modificar. • Utiliza técnicas matemáticas y algoritmos recursivos para extraer las características de la señal de voz. • Utiliza la técnica de ventaneo, en la que se toman determinados puntos discretos de la señal, desde los que se calcula el vector de características. • Una de sus principales dificultades es la elección correcta del parámetro *p,* que determina la

		extensión de las ventanas en las que se capturarán las características de la voz. Una elección incorrecta hará que se dejen características de lado (un valor de *p* muy bajo) o que se capture ruido innecesario (un valor de *p* muy alto). • Al tratarse de un modelo predictivo y recursivo, su coste computacional es alto si se le compara con los bancos de filtros.
Modelos Ocultos de Markov	Estocástico – Probabilístico	• Se trata de un modelo probabilístico en el que se determina la probabilidad de que el sistema pase de un estado a otro, representándose esto visual y computacionalmente a través de un grafo dirigido. • Toma segmentos de la voz humana y los convierte en estados del modelo, tratando de realizar predicciones que permitan determinar cuál será el siguiente fonema y de esa forma detectar unidades lingüísticas complejas. • Utiliza algoritmos de análisis secuencial, siendo los algoritmos de Viterbi y de Baum Welch los más utilizados. • Es una técnica de popularidad extendida desde hace aproximadamente 30 años, ya que su eficacia en la tarea del reconocimiento del habla es elevada. • Puede acarrear un costo computacional alto, pues es necesario generar modelos probabilísticos para cada fonema y esto hace que el uso de memoria y las iteraciones aumenten conforme el vocabulario crece.
Redes Neuronales Artificiales	Inteligencia Artificial	• Buscan simular la forma en la que el cerebro humano es capaz de reconocer la voz y las palabras del entorno que lo rodea. • Tienen la capacidad de ajustarse a sí mismas y mejorar sus resultados conforme transcurre el tiempo, pues puede aprender de experiencias previas y ampliar su vocabulario. • Dependiendo del tipo de aprendizaje y de red neuronal que se elija, se requiere una etapa de entrenamiento en la que el objetivo es alimentar la red con fonemas y/o palabras que le permitan realizar el reconocimiento de forma efectiva. • Es un modelo de amplia popularidad en los sistemas automáticos de reconocimiento del habla de la actualidad, pues el proceso de reconocimiento es complejo y difícil de expresar en términos deterministas o puramente

		algorítmicos. Al modelar el cerebro humano y la forma en la que este aprende, se pueden conseguir resultados efectivos y con costos computacionales reducidos.
Lógica Difusa	Inferencial - Matemático	• Se basa en el principio que dice que la mayoría de las interacciones del ser humano no pueden ser representadas de forma precisa a través de sistemas binarios o 100% deterministas, sino que casi siempre es necesario utilizar conjuntos o herramientas matemáticas que den lugar a la incertidumbre y a los estados intermedios. El reconocimiento del habla encaja en esta definición, pues no se trata de una señal binaria, sino de un fenómeno analógico. • Los modelos que utilizan lógica difusa para el reconocimiento del habla suelen utilizar el espectrograma de la voz, siendo ideal este gráfico tridimensional para ser la entrada de algoritmos que utilizan conjuntos difusos. • Se ha popularizado la combinación de los conjuntos difusos con las redes neuronales artificiales, dando lugar a los Sistemas Neurodifusos.

Fuente: Elaboración propia.

10. Conclusiones

En la presente monografía se ha podido recopilar una serie de características fundamentales relacionadas con los Sistemas Automáticos de Reconocimiento del Habla; pasando por su desarrollo histórico, las etapas a seguir para poder desarrollar un sistema de este tipo, algunos algoritmos que están presentes en el funcionamiento de estos sistemas, y algunas de las principales aplicaciones que estos tienen en la vida actual del ser humano.

Desde el punto de vista investigativo, con base en la información recopilada en el presente trabajo, es posible identificar dos corrientes o técnicas principales: Los métodos estadísticos y la inteligencia artificial. Los primeros han dominado el desarrollo de los sistemas automáticos de reconocimiento del habla en gran parte de su historia, proveyendo resultados fiables a medida que se les puede suministrar un vocabulario amplio. Sin embargo, todo indica que en la actualidad la apuesta mayor va por los sistemas basados en inteligencia artificial, haciendo uso principalmente de redes neuronales artificiales. Esto ocurre porque en los últimos años ha habido grandes avances en campos como el Machine Learning y el Deep Learning, permitiendo estos la realización de Sistemas de Reconocimiento de Voz autónomos que tienen la capacidad de mejorarse a sí mismos a medida que adquieren experiencia con uno o varios interlocutores. De esta forma se logran sistemas dinámicos que se adaptan a su entorno y cuyo vocabulario se incrementa con el paso del tiempo. Por lo tanto, si bien es cierto que los métodos estadísticos tienen una fiabilidad alta y han demostrado un buen funcionamiento, los métodos de aprendizaje automático tienden a dominar el campo de los Sistemas Automáticos de Reconocimiento del Habla.

Por otro lado, es fundamental mencionar que los Sistemas de Reconocimiento de Voz adquieren una importancia cada vez mayor en la vida cotidiana del ser humano. Este tipo de

sistemas se están convirtiendo en una pieza determinante en las interfaces hombre/máquina, por lo que la necesidad de mejorar las técnicas y desarrollar algoritmos eficientes y eficaces siempre estará latente. Se requiere que los Sistemas Automáticos de Reconocimiento del Habla cuenten con vocabularios extensos, pero así mismo que puedan adaptarse a cambios en el lenguaje que son cada vez más frecuentes y que requieren de una alta efectividad en el aprendizaje autónomo. Es indispensable minimizar los porcentajes de error y disminuir el tiempo en el que se genera el reconocimiento, así como mejorar las habilidades de procesamiento del lenguaje natural que permiten dar sentido a los fonemas que se detectan. Con el avance de diferentes herramientas como el Internet de las Cosas, la Domótica y otros sistemas inteligentes, la tarea del mejoramiento de los Sistemas de Reconocimiento de Voz es ineludible y debe estar dentro de las prioridades de los grupos de investigación alrededor del mundo.

Referencias Bibliográficas

[1] S. O. J. Suárez, «Reconocimiento de voz en español mediante sílabas,» *Polibits 34,* pp. 20-30, 2006.

[2] C. A. De Luna Ortega, J. C. Martínez Romo y M. Mora González, «Reconocimiento de Voz con Redes Neuronales, DTW y Modelos Ocultos de,» *Conciencia Tecnológica,* vol. I, nº 32, pp. 1-6, 2006.

[3] I. Villamil, Aplicaciones en Reconocimiento de Voz utilizando HTK, Bogotá: Pontificia Universidad Javeriana, 2005.

[4] J. Camargo y L. G. E. García, «Reconocimiento de voz humana aplicado a la domótica.,» *Ingenium,* vol. 13, nº 26, pp. 97-106, 2012.

[5] S. Ruiz, E. Miranda, M. Herlein, G. Etchart y C. Alvez, «Análisis Comparativo de Distintas Toolkits para el Reconocimiento Biométrico de Personas Mediante Voz,» de *XIX Workshop de Investigadores en Ciencias de la Computación*, Buenos Aires, 2017.

[6] F. Morales, «Tipos de investigación,» Bogotá, 2010.

[7] K. Barrios, J. López, S. Mendieta, R. Benavides y Y. Sáez, «Sistema de reconocimiento de voz: un enlace en la comunicación hombre-máquina,» *Revista de Iniciación Científica,* vol. 4, pp. 92-95, 2018.

[8] C. Camacho, Desarrollo de un sistema de reconocimiento de habla natural basado en redes neuronales profundas, Bachelor´s Thesis, 2016.

[9] J. Camargo, Sistema de reconocimiento de voz humana por hardware, BacherlorThesis, 2013.

[10] C. Cambronero y I. Moreno, Algoritmos de aprendizaje: knn & kmeans. Inteligencia en redes de comunicación, Madrid: Universidad Carlos III, 2006.

[11] E. Morales y J. González, «Aprendizaje computacional.,» *Aprendizaje Bayesiano http://ccc. inaoep. mx/emorales/Cursos/NvoAprend/node63. html*, 2007.

[12] M. Parreño, «Reconocimiento de voz con el robot Félix,» de *Doctoral dissertation*, 2017.

[13] S. Gabrielsson, Métodos para enseñar la gramática española como una lengua extranjera -un análisis cualitativo de los métodos deductivo, inductivo y aprender haciendo, Lunds University https://lup.lub.lu.se/student-papers/search, 2012.

[14] Z. Garcia, I. Bonet, P. Piñero y M. León, «Sistemas basados en conocimiento usandoProlog,» *Revista Cubana de Ciencias Informáticas*, vol. 1, nº 3, pp. 4-13, 2007.

[15] R. Hernández, Sistema de control activado por voz para uso en domótica, Xalapa Enríquez, México: Universidad Veracruzana, 2016.

[16] S. F. S. Fernández, Reconocimiento fonético en habla continua usando información de segmentos adyacentes, Portugal: Guimardes, 2004.

[17] J. Alcubierre, J. Minguez, L. Montesano, L. Montano, O. Saz y E. Lleida, «Silla de ruedas inteligente controlada por voz,» de *Primer congreso Internacional de Domótica, Robótica y Teleasistencia para todos*, 2005.

[18] C. Miranda, R. Camal, J. Cen, C. González, S. González, M. García y L. Narváez, Un juego de Gravedad con Reconocimiento de Voz para Niños con Problemas de Lenguaje, Yucatán: Universidad Autónoma de Yucatán, 2007.

[19] H. Bourlard y N. Morgan, A Continuous Speech Recognition System Embedding MLP into HMM, NIPS Proceedings, 1989.

[20] G. Martínez y G. Aguilar, «Reconocimiento de voz basado en MFCC, SBC y Espectrogramas,» *Ingenius,* vol. 10, pp. 12-20, 2013.

[21] M. Soto, J. De La Rosa y A. Moreno, Comparación de técnicas de parametrización espectral para reconocimiento de voz en idioma español, Zacatecas, México: Universidad Autónoma de Zacatecas "Francisco García Salinas", 2018.

[22] L. Cruz y M. Acevedo, «Reconocimiento de voz usando redes neuronales artificiales backpropagation y coeficientes Ipc. In 6to Congreso Internacional de Cómputo en Optimización y Software,» *CiCos,* nº 89-99, 2008.

[23] H. Torres y H. Rufiner, «Clasificación de fonemas mediante paquetes de onditas orientadas perceptualmente.,» *Anales del 1er Congreso Lationamericano de Ingeniería Biomédica, Mazatlán,* vol. 98, pp. 163-166, 1998.

[24] L. Toro, Análisis de estrés en la voz utilizando coeficientes cepstrales de Mel y máquina de vectores de soporte, Medellín: Universidad de San Buenaventura, 2018.

[25] M. Bezoui, A. Elmoutaouakkil y A. Beni-hssane, «Feature extraction of some Quranic recitation using mel-frequency cepstral coeficients (MFCC).,» de *5th international conference on multimedia computing and systems (ICMCS9*, pp. 127-131, 2016.

[26] P. Riaño, Introducción al Reconocimiento de la Voz, Cádiz: Universidad de Cádiz, 2004.

[27] C. De Luna, J. Martínez y M. Mora, «Reconocimiento de voz con redes neuronales, DTW y modelos ocultos de Markov,» *Conciencia Tecnológica*, nº 32, p. 0, 2006.

[28] C. Guevara, «Modelos ocultos de Markov para el desarrollo de un sistema de ayuda al habla para personas que sufren de disartria,» *Universidad de Lima (Ed.), Hacia la transformación digital. Actas del I Congreso Internacional de Ingeniería de Sistemas*, pp. 141-153, 2019.

[29] A. Pérez, L. Domínguez, P. Lotito, J. D'Amato y A. Rubiales, «Aplicación del algoritmo de Viterbi sobre modelos ocultos de Markov para la estimación de tráfico vehicular,» *Mecánica Computacional*, vol. XXXIV, pp. 2871-2888 (artículo completo), 2017.

[30] J. Rodríguez y J. Vidal, «Identificación ciega adaptativa basada en el algoritmo de Baum-Welch,» *URSI 1994: IX Simposium Nacional de la Unión Científica Internacional de Radio: Las Palmas de Gran Canaria*, pp. 513-517, 21-23 de septiembre 1994.

[31] G. Ponce, F. Álvarez y W. Medina, Diseño y análisis de un algoritmo predictivo de n-canales conjuntos disponibles en el rango 806 - 890 MHz basado en el método de Baum - Welch, Guayaquil: Espol, 2017.

[32] F. Izaurieta y C. Saavedra, Redes neuronales artificiales, Universidad de Concepción Chile, 2000.

[33] E. Acevedo y A. S. E. Serna, «Principios y características de las redes neuronales artificiales,» *Desarrollo e Innovación en Ingeniería,* p. 173, 2017.

[34] T. Kohonen, «An introduction to neural computing,» *Neural networks,* vol. 1, nº 1, pp. 3-16, 1988.

[35] X. Olabe, «Redes neuronales artificiales y sus aplicaciones,» *Publicaciones de la Escuela de Ingenieros,* 1998.

[36] M. Asís, Reconocimiento automático del habla basadas en lógica difusa y algoritmos genéticos, Huaraz, Perú: Universidad Nacional Santiago Antúnez de Mayolo, 2017.

[37] J. Cerrón, Usos de la lógica difusa e intervalos en el reconocimiento del habla, Navarro: Universidad Pública de Navarra, 2011.

[38] N. Hurtado y F. Cari, Agente Inteligente con reconocimiento de voz usando Lógica Difusa para mejorar el proceso de búsqueda de libros en la Biblioteca Central de la UNAMBA, Perú: Universidad Nacional Micaela Bastidas de Apurímac, 2019.

[39] Vidal, «La revolución de la lógica difusa,» *Antena de Telecomunicación,* nº https://www2.coitt.es/res/revistas/07b_Articulo_Revolucion.pdf, pp. 38-39, 2007.

[40] F. Paredes y W. Sarango, Desarrollo de un algoritmo neurodifuso para reconocimiento biométrico de voz en una tarjeta de arquitectura ARM, Quito: Universidad Politécnica Salesiana, 2018.

[41] L. Muda, M. Begam y I. Elamvazuthi, «Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques.,» *arXiv preprint arXiv:1003. 4083,* 2010.

[42] E. Pérez y B. Portilla, Sistema de reconocimiento de voz para aplicaciones de control en una vivienda, 2014.

[43] B. Matthew, «Hoy (2018) Alexa, Siri, Cortana, and More: An Introduction toVoice Assistants,» *Mayo Clinic Libraries, Mayo Clinic, Rochester, Minnesota, USA,* vol. 37, nº 1, pp. 81-88, 2018.

[44] L. Herrera, «Viviendas inteligentes (Domótica),» *Ingeniería e Investigación,* vol. 25, nº 2, pp. 47-53, 2005.

[45] E. Duarte Pérez y Á. López Portilla, «Sistema de reconocimiento de voz para aplicaciones de control en una vivienda,» de *7° Congreso Iberoamericano de Ingeniería Mecánica,* La Habana, 2014.

[46] E. Moguel, M. Zabala, D. Flores, J. Berrocal, J. García y J. Murillo, «Asistente de voz para el recordatorio de tratamiento farmacológico.,» *Jornadas de Ingeniería del Software y Bases de Datos (JISBD),* 2019.

[47] R. Pastor, Diseño de un sistema de reconocimiento de vozpara un brazo robótico para cirugía laparoscópica, Puebla, México: Benemérita Universidad Autónoma de Puebla, 2018.