CAPITULO I: Reconocimiento de voz

1.1 Introducción

El reconocimiento de voz es una rama de la inteligencia artificial cuyo principal objetivo es poder establecer una comunicación entre un humano y una computadora mediante la voz. El principal objetivo que se plantea en un sistema de reconocimiento de voz es el de procesar adecuadamente un mensaje oral bajo ciertos obstáculos como son: la fonética, acústica, léxica, ambigüedades, incertidumbres entre muchos otros.

Para el desarrollo de este trabajo, nos enfocaremos al reconocimiento de un hablante, donde el principal objetivo es identificar a la persona en cuestión y no el mensaje que se emite. Decidí trabajar con los coeficientes Wavelet como elementos de medición. Una wavelet es una pequeña onda cuya energía se encuentra concentrada en el tiempo y sirve como herramienta para el análisis de fenómenos transitorios, no estacionarios y variantes en el tiempo. Así pues, "Un Sistema de Reconocimiento Automático del Habla (SRAH) es aquel sistema automático que es capaz de gestionar la señal de voz emitida por un individuo. Dicha señal ha sido pasada por un proceso de digitalización para obtener elementos de medición (muestras), las cuales permiten denotar su comportamiento e implementar procesos de tratamiento de la señal, enfocados al reconocimiento". (Oropeza, J. L., 2006)

A partir de estas definiciones, el reconocimiento de voz contiene dos áreas de procesamiento importantes: el entrenamiento y el reconocimiento. En cuanto al entrenamiento se refiere, es una etapa crítica y sumamente importante dentro del reconocimiento de voz. Se puede afirmar que gran parte del éxito de un sistema de reconocimiento de voz, recae en la etapa de entrenamiento. Por otro lado, la etapa de reconocimiento la asociaremos con la etapa de clasificación, donde se determina mediante un clasificador, a que grupo corresponden de manera más adecuada las características de la voz previamente analizadas.

1.1.1 Antecedentes

El ser humano es un ser que fabrica y desarrolla dispositivos y sistemas para su bienestar y comodidad. Domótica es una palabra que nace a partir de palabras en latín *domus* – casa y *tica* – automática. Ésta es entonces la rama de la robótica que nace a partir de la necesidad de la comodidad que demanda el ser humano a partir de la tecnología que desarrolla. La domótica es en sí, la automatización del hogar para la comodidad y seguridad del usuario.

Si bien podemos encontrar indicios que desde hace muchos años en el siglo XV comenzaron con el desarrollo de sistemas de reconocimiento de voz a través de mecanismos como lo fue el trabajo publicado en 1668 por B. J. Wilkins enfocado a las posiciones del tracto vocal para producir sonidos. Quizá como tal, éste no fue un sistema de reconocimiento de voz, sin embargo, fue un gran preámbulo y sobre todo la base en la que muchos otros se inspiraron.

Posteriormente alrededor del año 1770 el fisiólogo alemán C. G. Kratzenstein desarrolló un mecanismo de viento para producir los sonidos vocales a partir del aire. Los instrumentos que construyó consistían en formas alargadas en forma de tubos con ciertas peculiaridades que al momento de soplar producían los sonidos A, E, I , O, U. Muy cercano a esa fecha el ingeniero y arquitecto húngaro W. R. Kempelen desarrolló su “máquina parlante” (Figura 1). Esta máquina podía producir sonidos que eran comunes a todas las lenguas europeas. Kempelen aseguraba que cualquier persona podía lograr en tres semanas realizar síntesis de voz, realmente sorprendente, en las lenguas: francesa, italiana y latina, sin embargo, idiomas como el alemán eran más complejos dado a la prevalencia de los sonidos consonánticos. Sin embargo fue el físico inglés Wheatstone que hasta principios de 1800 la desarrolló.

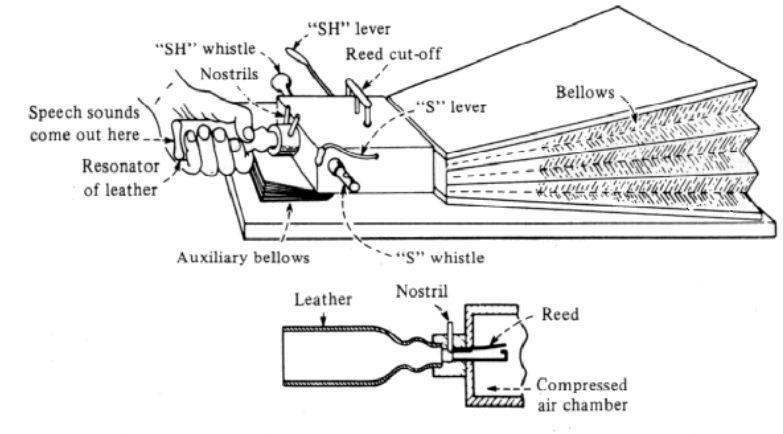


Figura 1: Máquina parlante de Wheatstone 1800

En el año 1888 el gramófono de Berliner dio pie a los primeros registros de voz y fue hasta 1922 cuando C. Paget descubre que existen dos componentes frecuenciales fundamentales. Realizó una tabla con estas componentes al estudiar su propia voz. Estos hoy en día se conocen como las componentes de Paget. Sin embargo, no son las únicas componentes en frecuencia que conocemos hoy en día.

Fue en el año de 1939 en la feria mundial de Nueva York cuando se presentó el VODER por sus siglas en inglés (VOice DEmostratoR) el cual como lo explica Verona Fernández (1997) e podía producir discurso continuo inteligible con facilidad. El Voder tenía dos fuentes de sonido: un oscilador que generaba un zumbido periódico, análogo al interruptor de Stewart para los sonidos sonoros, y un ruido aleatorio para los sonidos sordos. Publicado en la misma fecha el VOCODER se presentó por parte de SIEMENS en Múnich (1939) (Figura 2). Este instrumento parte de un análisis del habla real, por ello, más que una creación a partir de cero es una reconstrucción de algo ya dado. El funcionamiento del Vocoder empieza a ser algo complejo ya que empieza con una etapa de codificación mezclando una señal análoga de audio a partir de una voz humana para posteriormente procesarla y multiplexarla a una salida decodificadora mezclándola con una señal

digital. Anteriormente su principal aplicación fue la de seguridad en radiocoumincaciones. En la actualidad el vocoder lo podemos encontrar incluso en instrumentos musicales como un sintetizador.

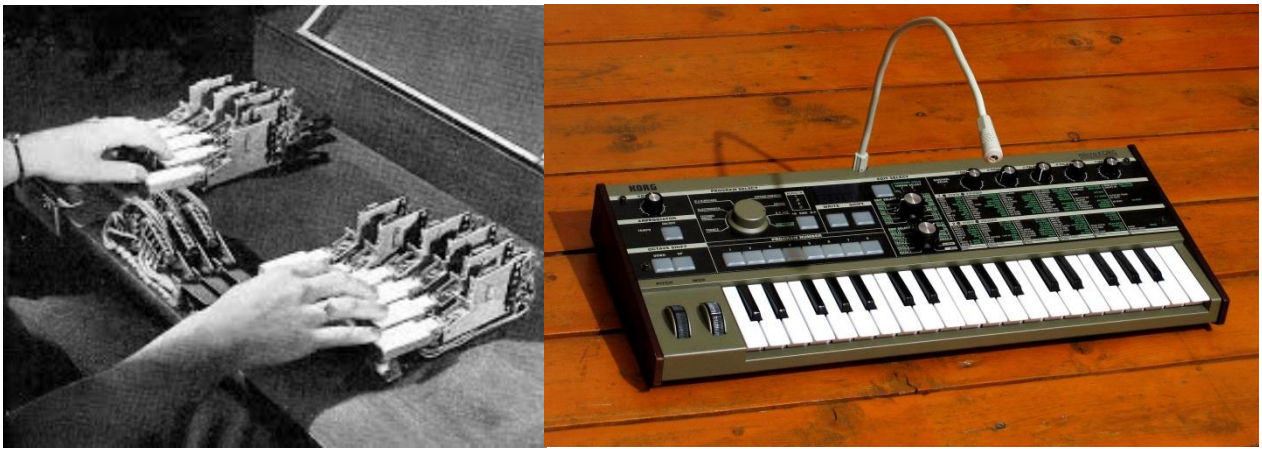


Figura 2: Comparación de aplicaciones del Vocoder en distintas fechas (1939 y 2009)

En 1969 los Laboratorios Haskinsde Nueva York llevaron a cabo la síntesis del lenguaje de una manera completa a partir del espectrograma en el llamado reproductor de patrones (Pattern Playback) de Hanskins. En la figura 3 se muestra el diagrama de funcionamiento del Pattern Playback.

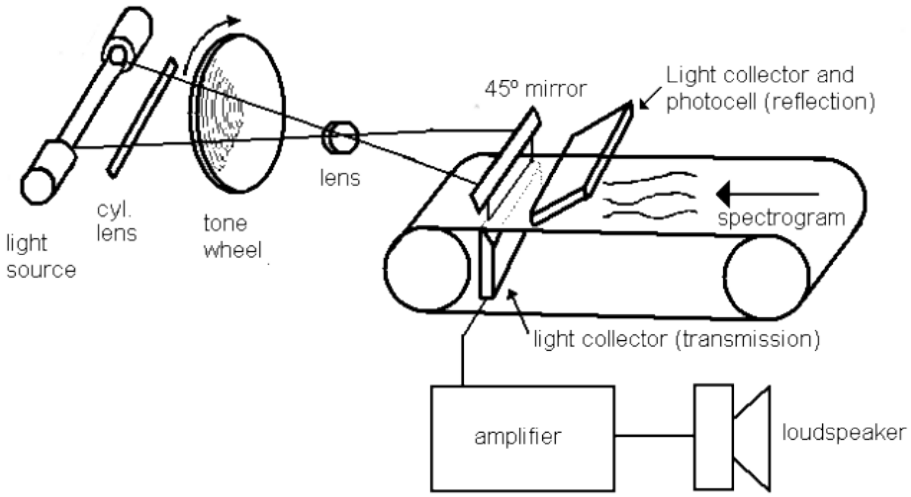


Figura 3: Diagrama del funcionamiento del Pattern Playback

Dejando de lado estos inventos que dieron base al estudio de la voz, ahora nos enfocamos a lo que podemos llamar el reconocimiento de voz pues data de fechas mucho más recientes como lo fue 1940 cuando se crea el espectrógrafo, el primer dispositivo para el reconocimiento de voz, un dispositivo capaz de visualizar la señal acústica sobre papel. Fue en los laboratorios de AT&T y Bell, así podían visualizar físicamente el espectro producido de una persona y otra.

Los Laboratorios RCA en 1956 según Rabiner [1993]. H.F. Olson y H. Bellar desarrollaron la máquina de escribir fonética. La señal leída del micrófono pasaba a través de una etapa de amplificación y una etapa de compresión antes de ser aplicada a un arreglo de 8 filtros. El propósito de la compresión era ajustar la señal a un valor medio, para que fuera lo más parecida entre los locutores.

Más adelante en 1960 se desarrolló un sistema de reconocimiento de voz caracterizado por utilizar un sistema de pautas entre palabras. Es hasta 1970 cuando realmente se desarrolla la tecnología de reconocimiento de voz que no requería que el usuario haga pausas entre palabras. Esta tecnología se volvió practica en los años 80 y sigue siendo desarrollada y afinada hasta hoy en día.

En 1970 G. Fant realizó un modelo de reconocimiento de voz el cual se dividía en 5 etapas. Éstas eran:

- Extracción de parámetros.
- Detección de segmentos.
- Transcripción fonética.
- Identificación de palabras.
- Interpretación semántica.

1.1.2 La transformada de Fourier

Jean-Baptiste Joseph Fourier física-matemático francés conocido por sus trabajos sobre la descomposición de funciones periódicas en series trigonométricas convergentes conocidas hoy en día como Series de Fourier.

En 1807, Fourier demostró que una función podía ser desarrollada en términos de series trigonométricas, y que se podían obtener, por integración, fórmulas para los coeficientes de Fourier. La Transformada de Fourier es ampliamente utilizada en el procesamiento y análisis de señales, con resultados satisfactorios en los casos en las que las señales son suficientemente regulares y periódicas.

1.1.2.1 Limitaciones del Análisis de Fourier

Por otro lado al analizar señales cuyo espectro varía con el tiempo, es decir, señales no estacionarias, el resultado no es satisfactorio. La Transformada de Fourier detecta la presencia de una determinada frecuencia pero no brinda información acerca de la evolución en el tiempo de las características espectrales de la señal (Figura 4).

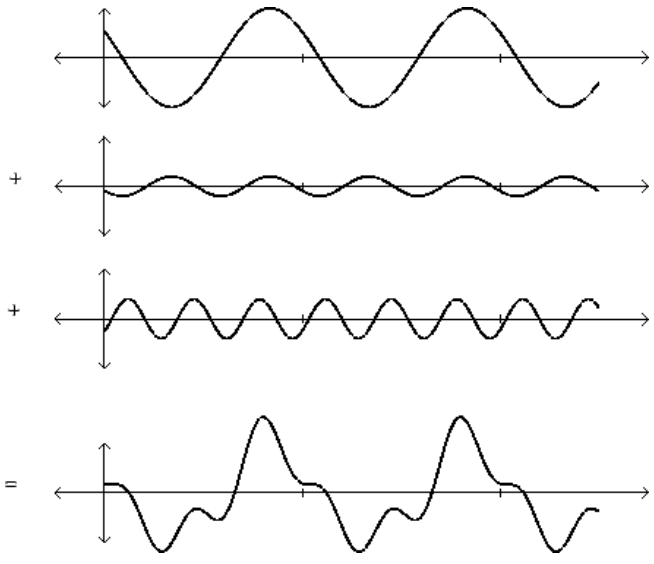


Figura 4: Cuatro ejemplos de señales periódicas estacionarias

Algunas características de la señal, tales como el inicio y el fin de una señal finita y el instante de aparición de una singularidad en una señal transitoria (Figura 5), no pueden ser analizados adecuadamente por el análisis de Fourier. Para los casos de señales no estacionarias y transitorias se utiliza generalmente la Transformada de Fourier con Ventana.

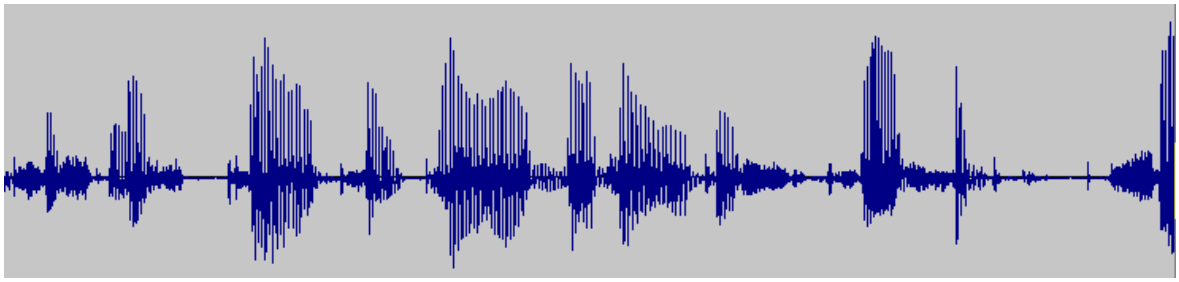


Figura 5: Señal transitoria, discontinuidad en el tiempo, no periódica (voz)

Así que la forma de analizar una señal de voz no estacionaria es realizar un análisis espectral dependiente del tiempo. De manera que una señal es partida secuencialmente para posteriormente analizarla con Fourier a cada una de estas partes. Gabor en 1940 es el primero en introducir el concepto de trasformada de Fourier de tiempo corto, conocida también como la transformada de Fourier de Ventana deslizante.

Al analizar la señal con una transformada de Fourier de Ventana deslizante es más fácil localizar esas singularidades que aparecen a lo largo de la señal. Sin embargo, la única información que nos proporcionará éste análisis será el periodo de tiempo donde aparece dicha singularidad. Esto se debe a que al correr un análisis de Fourier de Ventana deslizante una vez que se elige el tamaño de ventana a utilizar, los análisis se efectuarán con la misma resolución de tiempo y frecuencia. Aunado a esto, también será difícil identificarlos en los casos donde las singularidades aparezcan muy cercanas entre éstas.

1.1.3 La transformada Wavelet

La Transformada Wavelet, también conocida como la transformada de ondícula nace a partir del trabajo de Alfred Haar a principios de 1900´s. A partir del año de 1980 existieron varios colaboradores que a partir del trabajo realizado por Haar, dejaron grandes contribuciones dentro del análisis de señales por Transformada Wavelet. Tales son los casos en 1983 de Jan Olov-Strömberg para presentar su trabajo de óndulas discretas, en 1988 la francesa Ingrid Daubechies con su propuesta de óndulas ortogonales de soporte compacto, En 1991 Delrat y Newland con sus propuestas de óndulas de tiempo-frecuencia y armónicas respectivamente entre muchos otros. Algunas de las aplicaciones recientes del estudio de audio con Wavelets corresponden a Steve Hanov en 2008 con su software *"Wavelet Sound Explorer".* Programa que permite la visualización de archivos de audio el dominio de frecuencia-tiempo. Más reciente son las aplicaciones de procesamiento de señales biomédicas con Wavelets en el área médica, y cómo estos métodos en frecuencia- tiempo están mejorando la calidad de los diagnósticos médicos.

Específicamente, una función wavelet es una pequeña onda cuya energía se encuentra concentrada en el tiempo y sirve como herramienta para el análisis de fenómenos transitorios, no estacionarios y variantes en el tiempo. La transformada discreta se le ha dado un uso principalmente de codificación de señales, mientras que la transformada continua al análisis de señales. Por eso en este trabajo nos enfocaremos en la transformada Wavelet discreta ya que nos permite a partir del análisis trabajar con los detalles omitidos. Alguna de las aplicaciones que se desarrollan en la actualidad a partir de la transformada Wavelet son el análisis de Voz y la compresión y procesamiento digital de imágenes, siendo esta última realizada por la transformada Wavelet 2D. También es una herramienta útil en el análisis de señales sísmicas, electrocardigráficas, audio y reconocimiento de patrones.

Podemos dividir las Wavelets en dos tipos, las wavelets ortonormales y las wavelets discretas. Dentro del primer grupo encontramos a las wavelets ortonormales, las cuales no presentan información redundante y son una representación de la señal en forma unívoca. Mientras que en el segundo grupo encontramos a las transformadas Wavelets continuas con factores de escala y traslación discreteos, que se les conoce como Wavelets discretas.

Como anteriormente se mencionó, uno de los problemas por los cuales la transformada de Fourier no es de gran utilidad al analizar señales discontinuas no estacionarias es el hecho de que Fourier es incapaz de determinar las características de señales con anomalías en el tiempo. Una herramienta que es muy útil para poder estudiar estos comportamientos en señales no estacionarias es la Transformada Wavelet. Debido a que la Transformada Wavelet puede trabajar bajo señales transitorias y a muy altas frecuencias la hace una herramienta sumamente útil para poder resolver el problema que se presenta. Además una ventaja que nos ofrece a diferencia de la Transformada de Fourier de Ventana deslizante es que el tamaño de la ventana en la Transformada Wavelet es adaptada a la frecuencias que se le presentan.

De tal manera podemos concluir que El análisis WT es superior a los distintos tipos de análisis de Fourier ya que proporciona una localización tiempo-frecuencia adaptiva. A un nivel de escala grande de la wavelet se obtiene buena resolución en frecuencia mientras que a una escala baja se tiene una buena resolución en tiempo Así pues podemos enlistar las ventajas de utilizar la Transformada Wavelet en vez del análisis de Fourier:

- Habilidad para trabajar con señales transitorias.
- Adaptabilidad de la Ventana dependiendo de la frecuencia.
- Al realizar la transformada inversa de Wavelet no existe pérdida de información.
- Mejores resultados para análisis de señales de alta frecuencia.

1.1.3.1 Tipos de Wavelets

A las Wavelets se les asocia al término de onda pequeña, haciendo referencia a la función de "encuadre" (de ventana), cuya longitud es finita. La parte de Wave se refiere al carácter oscilatorio de la función. Cuando se hace referencia al término 'madre', se está indicando el hecho de que las funciones usadas, con diferente zona de acción, derivan de una función principal, es decir, la Wavelet madre es un prototipo a partir del cual se generan el resto de funciones. Como ejemplo se muestran algunas de las Wavelets madre más utilizadas.

- Wavelet Haar: (Figura 6)

- Wavelet Coiflet: (Figura 7)

- Wavelet Mexican Hat: (Figura 8)

- Wavelet Symlet (Figura 9)

- Wavelet Deubechies: (Figura 10)

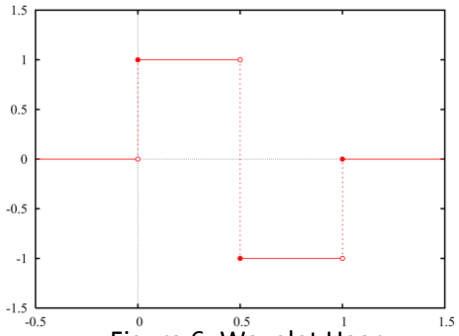


Figura 6: Wavelet Haar

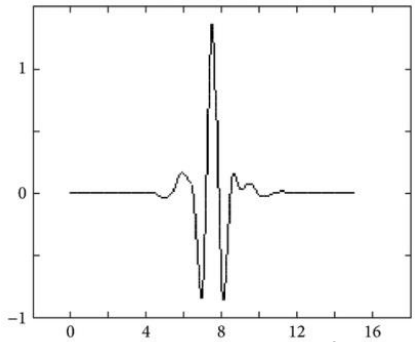


Figura 7: Wavelet Coiflet

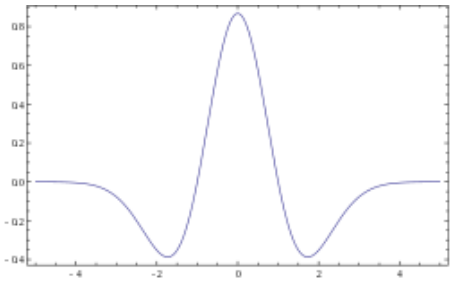


Figura 8: Wavelet Mexican Hat

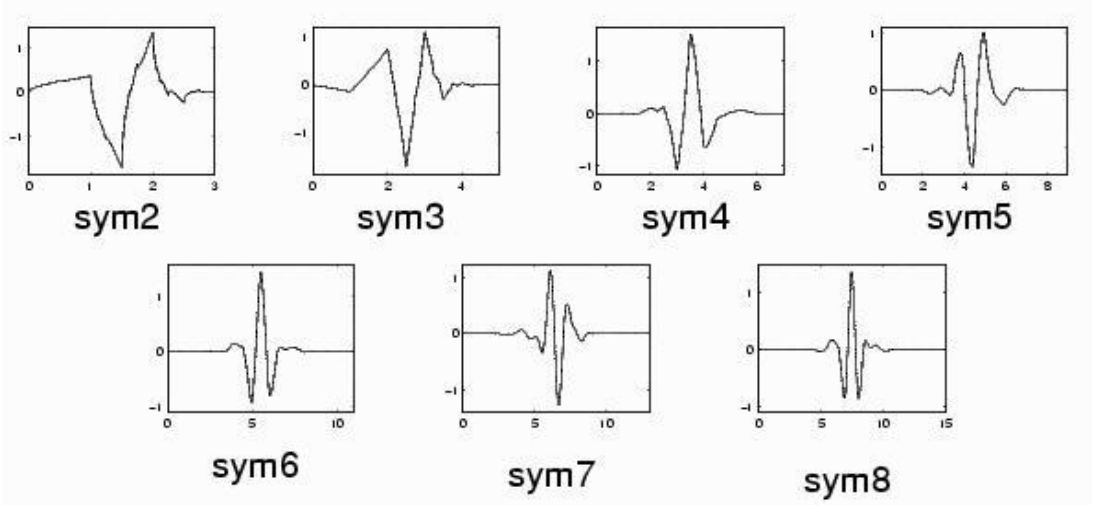


Figura 9: Las ocho familias de ondas de la Wavelet Symlet

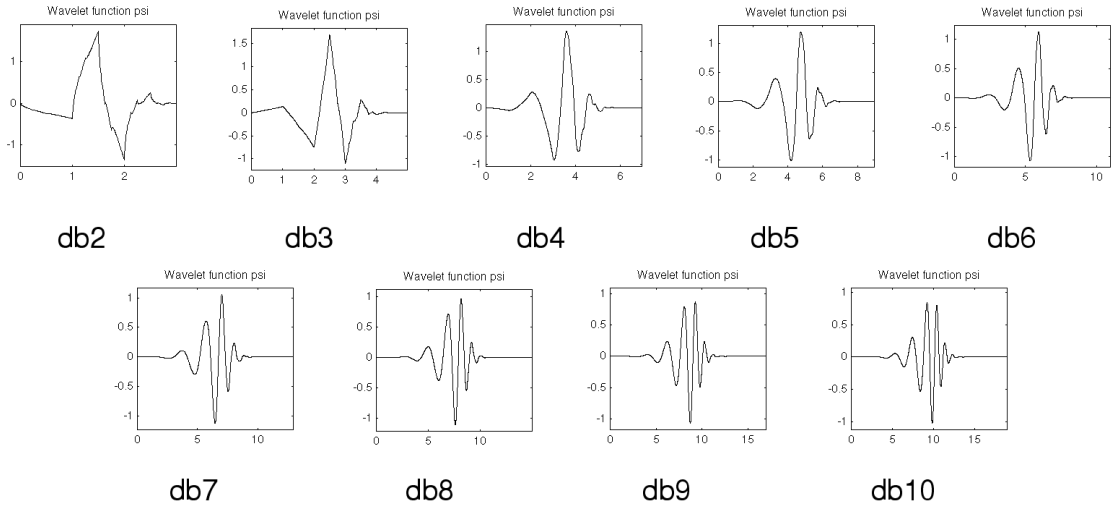


Figura 10: Las diez formas de onda de la Wavelet Daubechies

1.1.3.2 Los coeficientes Wavelet

La ecuación que define la Transformada Wavelet Continua (CoWT) es la siguiente:

$$CoWT(b,a) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{(t-b)}{a}\right)dt; a,b \in R; a \neq 0$$

En donde la variable x(t) es la señal a analizar, ψ(t) la forma de Wavelet (o Wavelet Madre), a = el parámetro de dilatación y b = parámetro de traslación. El análisis wavelet entrega una serie de coeficientes que indican que tan similar es la señal analizada a la señal de una función madre. Dado que la CoWT es un proceso reversible, es posible también reconstruir la señal a partir de los coeficientes Wavelet. González González (2010) sugiere que para poder definir más acertadamente los coeficientes es mejor relacionar el concepto como un filtro.

Al realizar la transformada Wavelet al momento de desfragmentar la señal, se realizar la etapa de "detallado" por filtrado en dos etapas, uno donde se realiza con un filtro desvanecedor (pasa-bajas) y otro como un filtro de detalles (pasa-altas). A este concepto de análisis de una señal mediante bancos de filtros es conocido como descomposición por árbol de Mallat. Son nombrados como los Coeficientes de Detalle [CD = $d_j(n)$] a los datos extraídos por el filtro pasa-altas y los Coeficientes de Aproximación [CA = $a_j(n)$] a los datos recuperados por el filtro pasa bajas. Así los coeficientes de detalle y aproximación respectivamente son:

[**CD = d_j(n)**] y [**CA = a_j(n)**] donde **n** indica el nivel (octava) de descomposición del árbol.

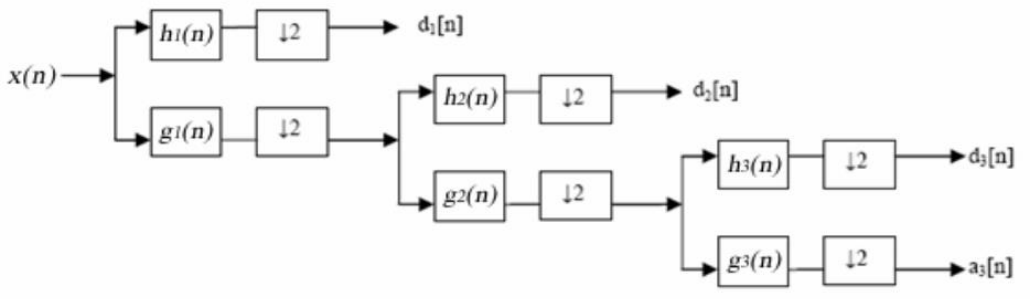


Figura 11: Árbol de descomposición Wavelet de 3 Niveles

Podemos determinar entonces que la calidad del análisis o la resolución de la "ventana" depende del nivel máximo de fragmentación que se realice.

1.2 Planteamiento del Problema

La razón por la que se desarrolla este trabajo de tesis, es el hecho de que actualmente la seguridad en las casas, y en especial en México, requieren de un gran apoyo debido a la gran ola de inseguridad que sacude actualmente al país. Además, a pesar de que en el mercado existe una cantidad considerable de sistemas de seguridad para las casas, éstos por lo general son bastante costosos. Verona Fernández (1997) nos menciona que "Aunque se pueden encontrar en el mercado sintetizadores de muy altas prestaciones, los sistemas de reconocimiento son aún muy restrictivos: palabras aisladas (speaker recognition), vocabulario muy limitado, dependencia del locutor, etc." Este trabajo se enfocará en desarrollar un algoritmo para identificación del hablante o *speaker identification.* Según Furui (2008), el reconocimiento de voz puede ser clasificado en dos tipos. En verificación del hablante e identificación del hablante. La principal diferencia entre verificación e identificación de un hablante es el número de alternativas de decisión. En identificación, el número de alternativas de decisión es igual al tamaño de la población, mientras que en verificación, sólo existen dos opciones: Aceptado o rechazado, sin importar el tamaño de población.

Muchos de los sistemas de seguridad que podemos encontrar hoy en día en el mercado no involucran sistemas de identificación de hablante, razón por la cual es otro aspecto a favor para trabajar y desarrollar esta tecnología. Al trabajar en Python para el desarrollo del software nos brinda cierto apoyo el hecho de contar con librerías y herramientas para el análisis de audio, sin embargo, a pesar de que existían algunos métodos de análisis de archivos de audio no existe una documentación para el desarrollo de un sistema de seguridad como el que en este trabajo de tesis planteo.

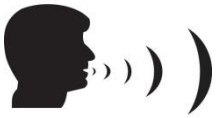
1.3 Marco Teórico.

Actualmente se han desarrollado muchas aplicaciones para el reconocimiento de voz, entre ellas destacan lo que es el control de voz. El *"State of the Art"* en reconocimiento de voz, deja muchas expectativas en cuanto a un futuro muy prometedor de las nuevas tecnologías que se busca desarrollar en un futuro próximo. Desde teléfonos inteligentes que integran estas nuevas características de control por voz hasta consolas de videojuego que incluyen un módulo para reconocimiento de comandos por voz, automóviles con sistemas de audio, navegación (GPS) y telefonía por control de voz entre muchas otras.

El panorama que me atrajo a realizar este proyecto es que existe un potencial muy grande para desarrollar esta tecnología. Últimamente surgen nuevas propuestas y nuevos prototipos y todo indica a que este tipo de tecnologías tendrán un futuro prometedor y todo indica que el área de domótica tendrá un crecimiento considerable próximamente.

1.4 Metodología

La plan de trabajo que se propuso para este proyecto consta de la utilización de ciertos conceptos matemáticas que han sido probados para estudiar y analizar los problemas que en mi proyecto se presentan. Empezando a partir de herramientas para poder estudiar y analizar señales de voz como lo es la Transformada Wavelet y la aplicación de ciertos módulos computacionales para realizar funciones que nos ayudan a crear un clasificador. Las teorías de modelado de mezclas para poder construir modelos necesarios para identificar características que comparten datos dentro de un grupo y la aplicación de dependencias entre características para formar modelos de árbol que nos ayudarían a generar una clasificación más precisa. En el diagrama siguiente podemos observar los pasos que se seguirán para lograr nuestro objetivo.



Hablante

python

Convierte archivo .wav a **arreglo numérico** (digitalizar)

Obtención de Coeficientes Wavelet por medio de **PyWavelets**

Mezcla de distribución Normal a partir de coeficientes:

nCA = mixture.NormalDistribution(cA.**mean**(),cA.**std**())
nCD = mixture.NormalDistribution(cD.**mean**(),cD.**std**())
m = mixture.MixtureModel(2,[0.5,0.5], [nCA,nCD])

medias
varianzas

Árboles de dependencia entre características:
(Se generan las dependencias)

tree = {}
tree[0] = -1
tree[1] = 0
n1 = **mixture.ConditionalGaussDistribution**(2,[**mean**()],[0,5],[**std**()],tree)

*La figura 30 muestra a detalle los árboles de dependencia

Mezcla Dist. Condicional n1, n2, n3,….

Arreglo a partir de los coeficientes CA y CD

Expectation Maximization
(Máximo de verosimilitud para datos incompletos).

data = m.sampleDataSet(550)
data.modelInitialization(data)
m.EM(data,40,0.01)

Componentes (mu, sigma, w, parents)

Tras proceso de dependencia por árboles y maximización de datos incompletos.

componentes

data

Graficación de resultados con soporte de la librería **matplotlib**

Clustering

clust = m.classify(**data**)