CAPITULO IV

SISTEMAS DE RECONOCIMIENTO DE VOZ

HUMMINGBIRD - 1 HIR

4.1. EL HABLA

Cada vez se encuentra más adeptos que están trabajando en el reconocimiento automático del habla por ordenador. Se está acumulando una amplia experiencia en el ámbito de la tecnología informática de la lengua.

La historia de la industria informática es un vertiginoso viaje hacia lo sorprendente. Desde su origen, las modernas tecnologías de la información viven en un permanente cambio y despliegan una capacidad innovadora de alcance e intensidad nunca antes vistas.

El compromiso que esta industria mantiene con el futuro ha impulsado asombrosos avances en los que las ideas y proyectos que en su origen rozaban la ciencia-ficción se están convirtiendo en realidad. Hoy, uno de esos grandes mitos, la posibilidad de que un ordenador sea capaz de recoger y procesar el lenguaje hablado es ya una exitosa realidad en el mercado y a un precio asequible.

Desde hace 25 años el mundo se está incursionando en el reconocimiento automático del habla por el ordenador. Se ocupan del tratamiento del lenguaje natural a través del ordenador y constituye una de las líneas de investigación más estratégicas de la actual industria informática. El objetivo es conseguir que el usuario pueda conversar con el ordenador de la manera más aproximada posible a como lo hace con otras personas, y en el camino surgen aplicaciones tan útiles como herramientas lingüísticas de ayuda a la escritura, la traducción o la enseñanza de idiomas.

Esta tecnología representa no sólo un salto cualitativo en el modo en que se utiliza el ordenador, y por tanto en que se extienden sus posibilidades, sino que resultará vital para el valor de un idioma dentro de la creciente globalización. El trabajo de desarrollado en este campo sirve para potenciar el reconocimiento de la voz en casi todas las lenguas y las tecnologías a ella asociadas.

En los últimos años, la mejora sustancial en la capacidad de proceso del hardware y en el desarrollo de los algoritmos asociado al conjunto de instrucciones que indica al sistema cómo interpretar lo que "escucha" han permitido avanzar hasta conseguir las primeras aplicaciones de reconocimiento del habla por ordenador.
El sistema de reconocimiento de voz está basado en complejos métodos probabilísticos y modelos lingüísticos. La conversión de la palabra hablada en texto se realiza a través de

sofisticados algoritmos que aíslan, identifican e interpretan los componentes fonéticos individuales del habla humana. Actualmente este proceso resulta altamente eficaz y permite alcanzar una tasa media de aciertos del 96 por ciento. El ordenador es capaz, por ejemplo, de elegir correctamente entre palabras homófonas, como "a" y "ha", y diferenciar la palabra "coma" del signo de puntuación ",".

Las aplicaciones de estos sistemas son múltiples. Además de permitir a cualquier usuario sustanciales mejoras de comodidad y ahorro de tiempo en la habitual tarea de introducir textos en el ordenador, el sistema de reconocimiento de voz resulta de extraordinaria utilidad. Así, el radiólogo puede dictar sus conclusiones mientras examina con total libertad una radiografía y el periodista escribir un reportaje al tiempo que consulta otros documentos.

El reconocimiento automático del habla es una tecnología, que día a día, está siendo introducida como la interface idónea para la comunicación entre hombre y máquina debido a la naturalidad de la comunicación y la robustez que comienzan a presentar los sistemas actuales de reconocimiento automático del habla. Cuando ponemos a trabajar un sistema de reconocimiento automático del habla en aplicaciones reales con usuarios no cooperantes aparecen una gran cantidad de problemas entre los que cabe destacar la pronunciación de palabras de fuera del vocabulario del sistema, la aparición de sonidos extraños como pueden ser los producidos para expresar una duda ("eh", "uh", etc), la falta de gramaticalidad que se produce en muchas ocasiones al construir frases de forma espontanea y el ruido existente en el ambiente donde trabaja el sistema de reconocimiento automático del habla como: el ruido de impresoras, ordenadores y aire acondicionado en oficinas, el ruido de coche en aplicaciones de telefonía móvil, etc. Estos problemas hacen que la transcripción completa de la frase pronunciada sea una tarea dificultosa, lo que provoca que la tasa de reconocimiento del sistema se reduzca dramáticamente cuando un sistema que trabaja bien en condiciones de laboratorio pasa a ser utilizado en condiciones reales.

Esta problemática esta siendo tratada como el Reconocimiento del Habla Conversacional o Espontanea. En una escala más reducida, pero muy interesante de cara a aplicaciones reales, las técnicas de localización de palabras o más conocidas, intentan detectar la presencia de un conjunto más o menos reducido de palabras clave en un contexto de habla conversacional o espontanea.
En muchas ocasiones, y dentro de la comunicación oral entre dos personas, no somos capaces de entender perfectamente todas las palabras que pronuncia nuestro interlocutor

pero comprendemos la semántica del mensaje al entender las palabras con más significado del mensaje que nos transmite nuestro interlocutor. Este fenómeno ocurre muy frecuentemente cuando escuchamos una conversación en un idioma que no dominamos a la perfección. Bajo esta idea, y mediante una adecuada selección del conjunto de palabras clave como aquellas con mayor contenido semántico en la aplicación donde se va a utilizar el sistema de reconocimiento automático del habla. Los sistemas reconocimiento automático del habla basados en las técnicas de localización de palabras son los candidatos idóneos para trabajar en aplicaciones reales donde el vocabulario es más o menos reducido y controlable, como por ejemplo en servicios de telecomunicaciones tales como los sistemas de audiotex, telefonía móvil libre de manos o automatización de servicios de operadora.

Dentro del desarrollo de sistemas de reconocimiento automático del habla. Actualmente, se desarrolla un sistema de reconocimiento de habla continua en aplicaciones con semántica restringida y en el desarrollo de un sistema para detectar comandos definidos por el usuario para aplicaciones en condiciones reales

4.2. CODIFICACION DE VOZ Y AUDIO

Un primer acercamiento hacia la comprensión de un codificador de voz, será la definición de criterios que permitan determinar la calidad de la señal de voz recibida en el extremo receptor. A este criterio se le denominará Criterio de fidelidad.

Cualquier evaluación de una señal implica una medida de fidelidad. Para la mayoría de los sistemas de comunicación, esta medida es difícil de especificar, puesto que esta envuelve la percepción humana. La calidad de voz es evaluada generalmente a través del criterio según un oyente entiende qué es lo que se dijo o quién lo dijo (de aquí en adelante, se entenderá el término calidad de voz como la calidad de señal de voz en el extremo receptor). Mediciones objetivas que reflejen con acuciosidad dichos factores son difíciles de establecer.

A pesar de este incompleto estado del conocimiento, existen variados sistemas que cuantifican la calidad de voz. Estos derivan de pruebas realizadas a través de reconocimiento de palabras y sonidos, con distintos tipos de oyentes (humanos). Usando estos datos, se han establecido guías para el diseño de codificadores de voz. A ello se agregan las mediciones de densidad espectral de muestras de corta duración, relación

señal ruido, que analizadas correctamente, significan un paso hacia una definición objetiva de dicha cuantificación de la percepción.

Una amplia gama de codificadores de voz son denominados codificadores de forma de onda. Como su nombre lo indica, dichos codificadores reproducen la forma de onda de la señal. En un principio, fueron diseñados para ser independientes del tipo de señal, dado que pueden codificar con calidad una variedad de señales, por ejemplo música, tonos y datos dentro de la banda de voz. Además, tienden a conservar la mayoría de las características de la voz en un ambiente ruidoso. Para mantener dichas ventajas con un mínimo de complejidad, los codificadores de forma de onda típicos apuntan a economizar su tasa de transmisión de bits.

Los codificadores de forma de onda pueden ser optimizados y hechos para señales específicas, logrando una gran eficiencia de código. Un desarrollo típico es realizado utilizando observaciones estadísticas sobre un conjunto de señales, haciendo que el codificador de forma de onda permita un mínimo de codificación de código para un tipo de señal (por ejemplo, la voz). La construcción de dicho código es basado en un estudio estadístico de la forma de onda de la voz, distinto de la parametrización de la información obtenida de algún modelo físico de la señal. Las propiedades utilizadas en la creación de un código para un codificador de forma de onda corresponden a la explotación de la redundancia de las características de la señal de voz, ya sea en el dominio del tiempo o en el dominio de la frecuencia.

Es así como en el dominio del tiempo se utilizan las siguientes redundancias:

- Distribución no uniforme de la amplitud.
- Correlación entre muestra y muestra.
- Correlación ciclo a ciclo (periodicidad).
- Correlación entre intervalos de igual duración (pitch interval).
- Factores de inactividad de la voz (silencios).

4.3. COMPRESION DE LA VOZ

La integración de los sistemas de comunicación, junto con el constante crecimiento de la digitalización de las redes que utilizan dichos sistemas, hacen necesario una administración eficiente de los recursos disponibles. Uno de estos recursos, corresponde a

la capacidad de memoria de almacenamiento de datos. Más específicamente, se busca que dicha capacidad sea maximizada, a través de la reducción de los paquetes de información almacenados. Es así como surge la compresión de datos como una de las posibles soluciones quizás la más lógica al problema planteado en la administración eficiente de la capacidad de memoria.

La búsqueda de métodos de compresión de datos está ligada al tipo de datos almacenados, entendiéndose por tipo al origen de dichos datos. Dentro de estos tipos de datos, tenemos los datos procedentes del proceso de digitalización de la voz humana. A su vez, dichos datos pueden ser nuevamente clasificados según el método de digitalización de voz efectuado.

Dentro de las aplicaciones de la compresión de audio, se distinguen 4 áreas: difusión, almacenamiento, multimedia y telecomunicaciones. Ejemplos de estos son: almacenamiento en disco (CD audio, video, etc.), televisión por cable y satelital, Internet (audio y video), aplicaciones ISDN, etc.

En la actualidad, la mayoría de las aplicaciones de software que hacen uso de la compresión de audio para distintos fines, siguen el estándar creado por MPEG (Moving Picture Experts Group). Este grupo trabaja para crear estándares para la codificación de audio y vídeo definiendo para la parte de audio 3 estándares para comprimir conocidos como capas, cada uno define su propio formato de trama y el tipo de codificador que necesita, diferenciándose también en su complejidad y en la tasa de compresión lograda.

4.4. ALGORITMOS DE COMPRESION DE VOZ

Los algoritmos de codificación explotan las distintas características de la voz humana. Un sitial importante ocupa entre dichas características las propiedades de la voz humana en el dominio de la frecuencia. A continuación se presentará el desarrollo de un algoritmo uno de los más populares y antiguos codificadores de voz, el codificador de voz de canal. Se utiliza una moderada descripción de la voz, sabiendo de antemano cómo la señal fue generada en la fuente. La idea es que alguna compresión física durante la generación de la señal pueda ser cuantificada, utilizándola como descripción eficiente de la señal. Esto implica que la señal debe ser puesta en un molde específico que permita parametrizar la señal correctamente (en este caso, la señal de voz). Esta técnica utiliza las variaciones de la generación de señal como una fuente de código. Por ello, se le denomina a esta técnica

Codificadores de fuente de voz o más comúnmente como VoCoders, de los términos Voice (voz) y Encoder (codificador).

El objetivo es la preservación de la amplitud espectral de muestras de corta duración de señales de voz en una audición de voz pronunciable.

4.4.1. VOCODERS

Esta técnica está diseñada específicamente para señales de voz, por lo tanto no es aplicable su uso en las redes de telefonía pública, en las cuales otros tipos de señales como la señal de un módem, son transmitidas. Es más, los vocoders típicos producen sonido de voz "artificial" o "poco natural".

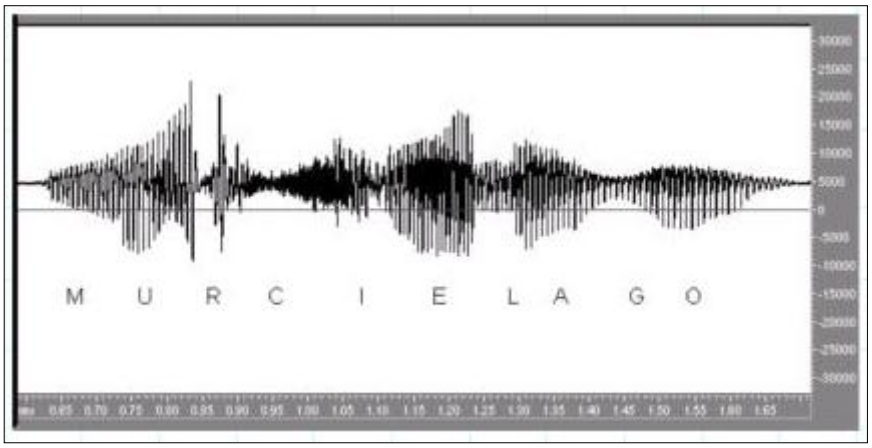


Figura 4.1 *Espectro de los fonemas*

El objetivo principal de un vocoder es codificar solo las características perceptivas importantes de la voz, con la menor cantidad de bits que el común de los codificadores. Debido a esto, los vocoders son utilizados en aplicaciones de limitado ancho de banda, donde otras técnicas no pueden aplicarse.

Algunas de las principales aplicaciones de los vocoders son:

- Grabación de mensajes almacenados (Ej. "Número equivocado").
- Encriptación de voz en transmisión por línea telefónica.
- Salida de audio de un computador o máquina.
- Sintetizadores musicales y experimentación electrónica del sonido.

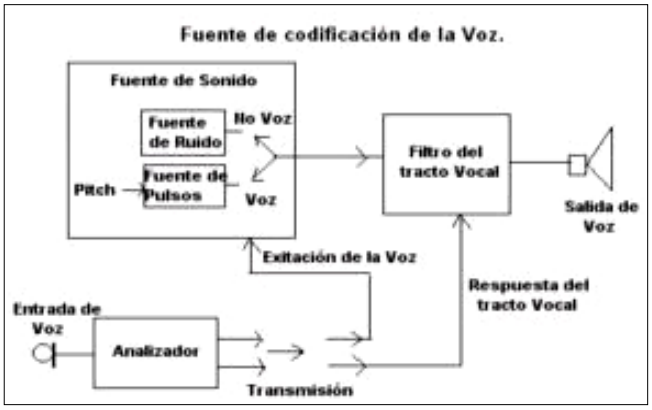


Figura 4.2 *Modelo del sistema generador de Voz*

Se puede decir que el vocoder depende de una rígida parametrización de la señal de voz, que concuerda con el modelo lineal estacionario de la generación de voz. En el modelo tradicional, mostrado en la figura, la fuente de sonido es independiente del sistema resonante que modula el sonido. Para comprender mejor el modelado de la generación de voz, se analizará algunas características de dicha fuente.

A continuación, se analizará las características del sistema, con el cual se ha modelado la generación de señales de voz, y que ha sido la guía para el diseño de diversos tipos de vocoders.

4.4.2. CARACTERIZACION DEL SISTEMA

De acuerdo al concepto de fuente lineal, la salida de sonido del tracto vocal, corresponde a la convulsión en el tiempo de las formas de onda de excitación y de la respuesta a impulso del sistema vocal. Consecuentemente, la resonancia acústica del tracto vocal modula o envuelve el espectro de la fuente. Diferentes sonidos de voz corresponden únicamente a distintas envolventes espectrales. Los vocoders han sido realizados sobre la base de una descripción paramétrica del tracto vocal. Esta descripción toma una variedad de formas, por ejemplo, amplitudes espectrales de muestras de pequeña duración de señales de voz evaluadas en frecuencias específicas, como se verá en el codificador de voz de canal.

Coeficientes de predicción lineal, que describe el comportamiento de la envolvente espectral, valores de las frecuencias que presentan una mayor resonancia en la densidad espectral de la muestra analizada, funciones de autocorrelación de muestras de voz de corta duración.

4.4.3. VOCODERS EN EL DOMINIO DE LA FRECUENCIA

La voz consiste en una sucesión de "fonemas" sonidos articulados por el tracto vocal. Cada sonido de voz se caracteriza por su potencia espectral. La envolvente espectral de cada sonido de voz es determinada por el mecanismo humano de generación de voz. A su vez, el sistema de audición humana permite el reconocimiento de dichos fonemas, los cuales forman la voz. Este hecho de analizar el espectro de la voz, es utilizado en el codificador de canal de voz o más conocido por Codificador de voz de canal.

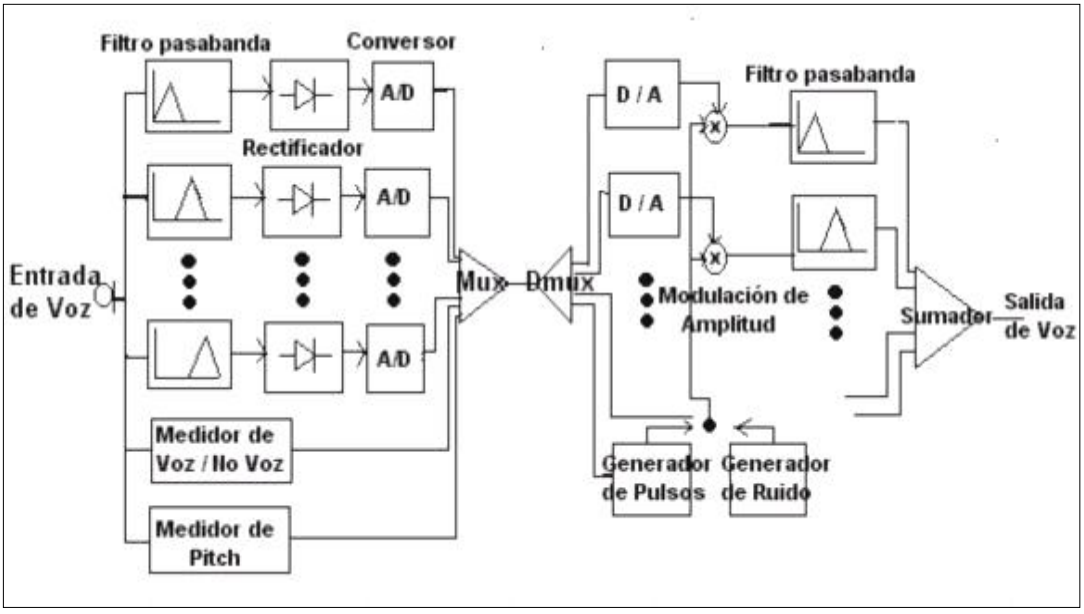


Figura 4.3 *Diagrama en bloques de un Codificador de voz de canal*

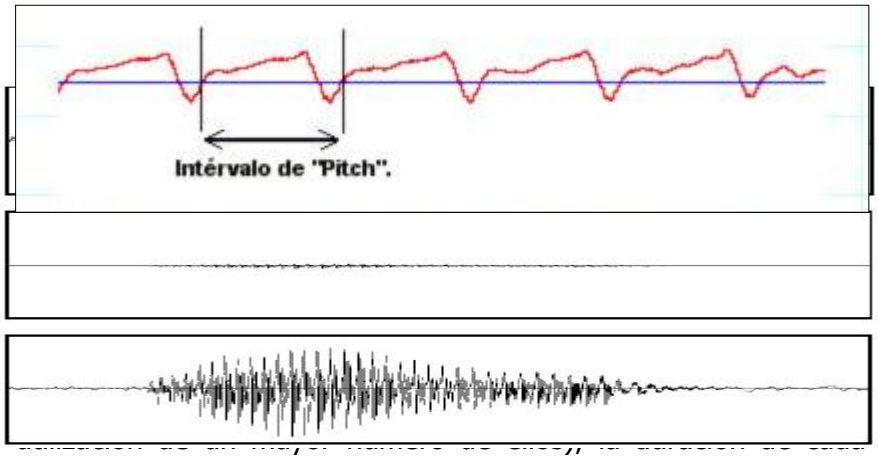
La mayor parte del proceso que envuelve a un codificador de voz de canal corresponde a determinar el espectro de muestras de corta duración en función del tiempo. Como se puede ver en la figura, un banco de filtros pasabanda es utilizado para separar la energía de la voz en sub-bandas, la cual es rectificada en forma completa, para así ser filtrada nuevamente por un filtro pasabajos y obtener de esta manera su nivel de potencia relativa dentro de dicha subbanda. Estos niveles de potencia individuales son codificados, multiplexados y transmitidos hacia el receptor.

Además de medir el espectro de la señal, los vocoders determinan la naturaleza del sonido, si corresponde a un sonido de voz o no, junto con medir la frecuencia de los pitch de señales de voz. La medición de la excitación es utilizada para sintetizar la señal de voz en el extremo receptor, es decir, en el decodificador, a través de una adecuada selección de la fuente de señal, según el modelo de generación de voz, mencionado anteriormente. La excitación de la voz es simulada con un generador de pulsos usando una frecuencia igual a la medida en el pitch de la muestra de voz. Los sonidos que no corresponden a un sonido de voz, son simulados con un generador de ruido. Debido a la naturaleza sintetizada de la excitación.

Como se puede apreciar en la figura, el decodificador implementa el modelo de generación de voz con un banco de filtros pasabanda, cuyos niveles de potencia de entrada son determinados por la respectiva subbanda en el codificador. Superponiendo las bandas individuales que han sido recreadas en el espectro de frecuencias, se obtiene la señal original.

Muchas variaciones de un codificador de voz de canal se han desarrollado, utilizando la naturaleza de la excitación y el promedio de los niveles de potencia. La mayor dificultad que enfrenta el desarrollo de un codificador de voz de canal corresponde a la determinación del pitch de la señal de voz. Inclusive, algunos sonidos no se clasifican dentro de los sonidos de voz ni fuera de ellos. Esto exige una mayor minuciosidad en el análisis de las características de la señal de excitación. Sin esta minuciosidad se tienen resultados pobres, dependientes del tipo de persona que habla y de los sonidos particulares que ha producido. Algunos de los más avanzados codificadores de voz de canal desarrollados han producido señales de voz bastante inteligibles, con un sonido sintético, a velocidades de 2400 [bps].

Un codificador de voz de canal maneja una serie de valores que hacen de su codificación un proceso eficiente en mayor o menor grado, según sea el caso. No se debe perder de vista que la señal obtenida en el extremo receptor tendrá solo las características más notorias de la voz codificada, por lo que es escuchar un sonido sintético o artificial no debe ser motivo de sorpresa.



Finalmente no receptor de un *codi* ue tenga el codificador o de banda implican la muestra, la correcta detección de los *pitch* de la señal de voz y su respectiva generación (excitación) en el decodificador.

Una mayor eficiencia en la descripción de la señal de voz se puede obtener especificando solo las frecuencias de los niveles altos de la densidad espectral de la muestra de voz analizada y las respectivas amplitudes de dichas frecuencia.

El proceso de codificación de voz permite transmitir y almacenar la señal de voz en forma digital eficientemente y sin pérdida de calidad. Desde el punto de vista de la transmisión de la señal de voz, la codificación de voz permite optimizar la utilización del canal de comunicación, transmitiendo el máximo de información, como transmitir varias comunicaciones por un solo canal, con la mínima pérdida de calidad optimizando la relación entre velocidad de transmisión (bits/segundo) e inteligibilidad del mensaje. Desde el punto de vista de almacenar señal de voz en formato digital, la codificación de voz permite minimizar el número de bits necesarios para el almacenamiento manteniendo un nivel de calidad adecuado.

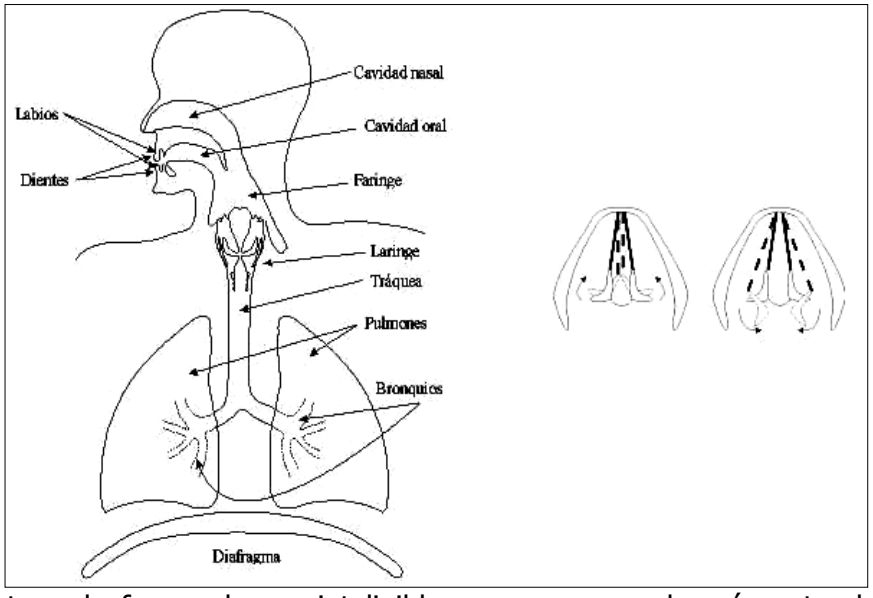
Como valor añadido al proceso, la codificación digital de voz permite incorporar algoritmos de cifrado para establecer comunicaciones privadas seguras o realizar grabaciones indescifrables para terceras personas. Las investigaciones van al desarrollo de codificadores de baja velocidad para aplicaciones de telefonía y en codificadores de audio de banda ancha (7 kHz a 20 kHz) para aplicaciones de teleconferencia y multimedia.

Figura 4.5 *Espectro de la VOZ humana*

Figura 4.6 *Sistemas que Generan y Modula la VOZ*

4.5. SINTESIS DE VOZ

El proceso de síntesis de voz dota a las máquinas de la capacidad de producir mensajes orales no grabados previamente como es el caso de los sistemas de respuesta oral. Tomando como entrada cualquier texto, los sistemas de síntesis de voz realizan el



proceso de lectura de forma clara e inteligible y con una voz lo más natural humana posible. La síntesis de voz conforma la interfaz oral de comunicación entre una máquina y el usuario de la misma.

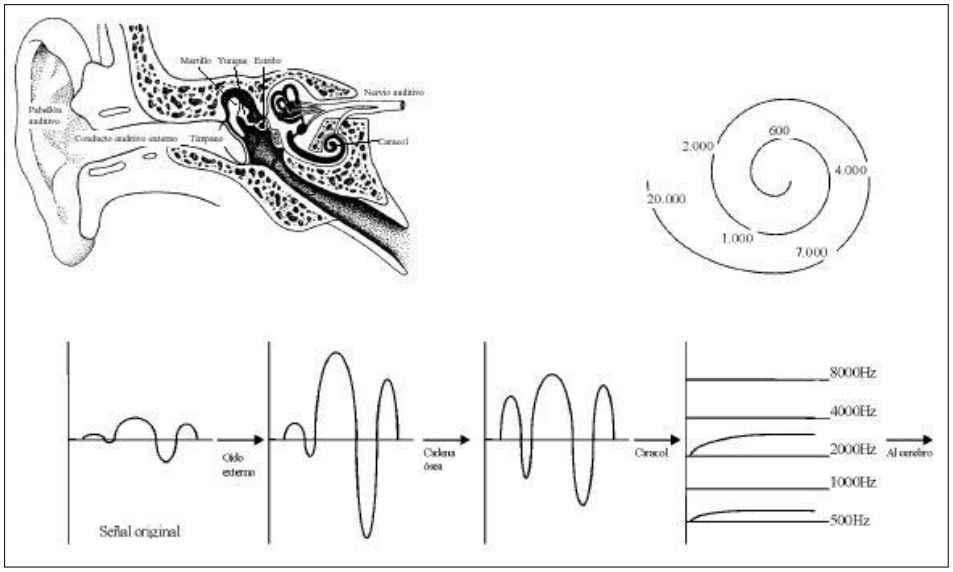
4.6. ANALISIS DE VOZ

El análisis de la señal de voz es el primer paso necesario en cualquier sistema basado en tecnologías del habla. Dejando de un lado las técnicas clásicas de análisis de la señal de voz, la investigación básica está encaminada al estudio de nuevas representaciones tiempo-frecuencia y su aplicación al análisis de la voz, la utilización estadísticas de orden superior y su aplicación en algoritmos de reducción y cancelación de ruido, la utilización de modelos auditivos para la representación de la señal de voz en sistemas de reconocimiento del habla, así como el desarrollo de algoritmos de detección de voz/silencio, pitch y sonoridad.

Figura 4.7 *Adquisición del Sonido*

El sistema auditivo está compuesto por complejos sistemas encargados de tareas como filtrar ruido, marcar rangos o niveles de audición, entre otros. Luego mandan las señales al cerebro para que sea éste quién tome las decisiones sobre las acciones a seguir.

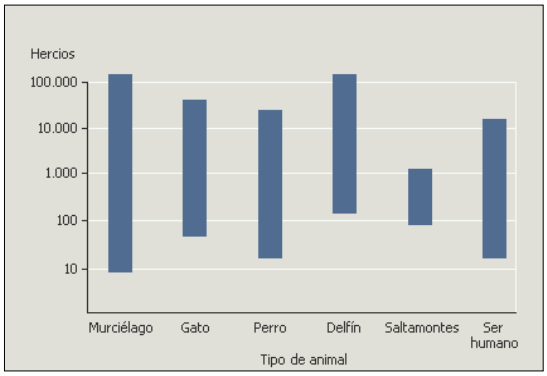
La red auditiva está distribuida a lo largo de toda la superficie craneal, y la información obtenida es recogida por el oído interno, a continuación se verá la forma en que el oído hace su trabajo, así como los niveles de audición que un ser humano es capaz de percibir.



A continuación se presenta un gráfico en el que se muestra los rangos de frecuencia a la que los animales escuchan.

Figura 4.8 *Rangos de Frecuencias audiibles*

4.7. SOFTWARE ESPECIALIZADO EN EL RECONOCIMIENTO DE VOZ



Se ha encontrado algunas aplicaciones especializadas en el tratamiento de la voz, a continuación se hace un análisis de las características principales de dos de las más reconocidas y localmente comerciales ViaVoice de IBM y Dragon Naturally Speaking de Dragon Systems. Sin ningún orden en especial.

4.7.1. VIAVOICE DE IBM

Permite crear documentos sin necesidad de teclear.
Se puede dar formato al texto a través de sencillas órdenes habladas.
Dispone de un analizador de la voz que aumenta la precisión del dictado.
Permite escuchar los documentos sin necesidad de leerlos.
Busca palabras nuevas en documentos existentes y las añade al vocabulario personal.
Durante el dictado, y de acuerdo con el contexto, las palabras que suenan de forma igual o parecida son escritas correctamente.
Cuenta con un Vocabulario activo de más de 100.000 palabras.
Dispone de un diccionario de respaldo con 475.000 palabras más.
Navega por Internet hablando con Internet Explorer.
La función de "Voice Mouse" permite utilizar el ratón sin usar las manos.
Incluye dictado directo en la mayoría de aplicaciones de Windows.
Utiliza mandatos naturales para decirle al sistema las acciones que ha de realizar.
Permite realizar funciones de Comando y Control del sistema mediante la voz.

Permite corregir los textos directamente con la voz.

4.7.2. DRAGON NATURALLY SPEAKING

- Dicte en prácticamente cualquier aplicación.
- Aprenda con un nuevo Tutorial interactivo.
- Inicie el programa en menos tiempo.
- Dicte con mayor precisión de reconocimiento.
- Corrija los errores mientras dicta.
- Acceda a los comandos sin memorizarlos .
- Añada palabras de documentos rápida y fácilmente.
- Administre su correo electrónico con la voz.
- Utilice nuevos métodos para explorar la Web.
- Cree abreviaturas de dictado.
- Dicte ahora y corrija después.
- Trabaje en Lotus Notes.
- Organice y guarde sus comandos de voz.

4.8. USO DE LAS RN PARA RECONOCIMIENTO DE VOZ

Las redes neuronales pueden usarse para varias tareas relacionadas con el tratamiento de señales. Como caso particular del tratamiento de señales está el tratamiento de voz. Dentro de este campo se pueden aplicar a los distintos problemas o tareas que existen: síntesis, codificación, compresión y reconocimiento.

La inteligencia artificial pretende acercar el comportamiento de las máquinas al comportamiento humano. Esto pretende liberar al hombre de tediosas tareas que hasta ahora sólo él podía realizar. Para que la supuesta máquina pueda conocer las necesidades del hombre es necesario que tenga una forma de comunicarse con él. Entre las formas de comunicación creo que la voz juega un papel primordial y es muy importante que se realice en ambos sentidos. Si bien se puede pensar que las máquinas podrían aprender por sí solas a entender la voz, quizá sea más conveniente ayudarles mediante la creación de reconocedores y sintetizadores de voz.

Las RN de manera eficiente para el reconocimiento de voz. Para ello debemos elegir las entradas de nuestra red, las salidas y la estructura necesaria para que produzca las salidas deseadas para las entradas dadas. Después habrá que elegir un algoritmo de

entrenamiento entre los posibles y unos parámetros para después realizar los entrenamientos.

4.8.1. PARAMETROS DE ENTRADA A LA RED

Los parámetros de entrada a la red para el reconocimiento de voces es conveniente que cumplan ciertas condiciones:

Que sean **simples**. En principio no interesan vectores de entrada de dimensiones muy grandes porque eso implica usar muchas neuronas. Esto no sólo significa que vamos a usar muchos recursos en el producto final sino que también aumentaremos mucho más el número de pesos y por tanto lo normal es que aumente notablemente el tiempo de entrenamiento.

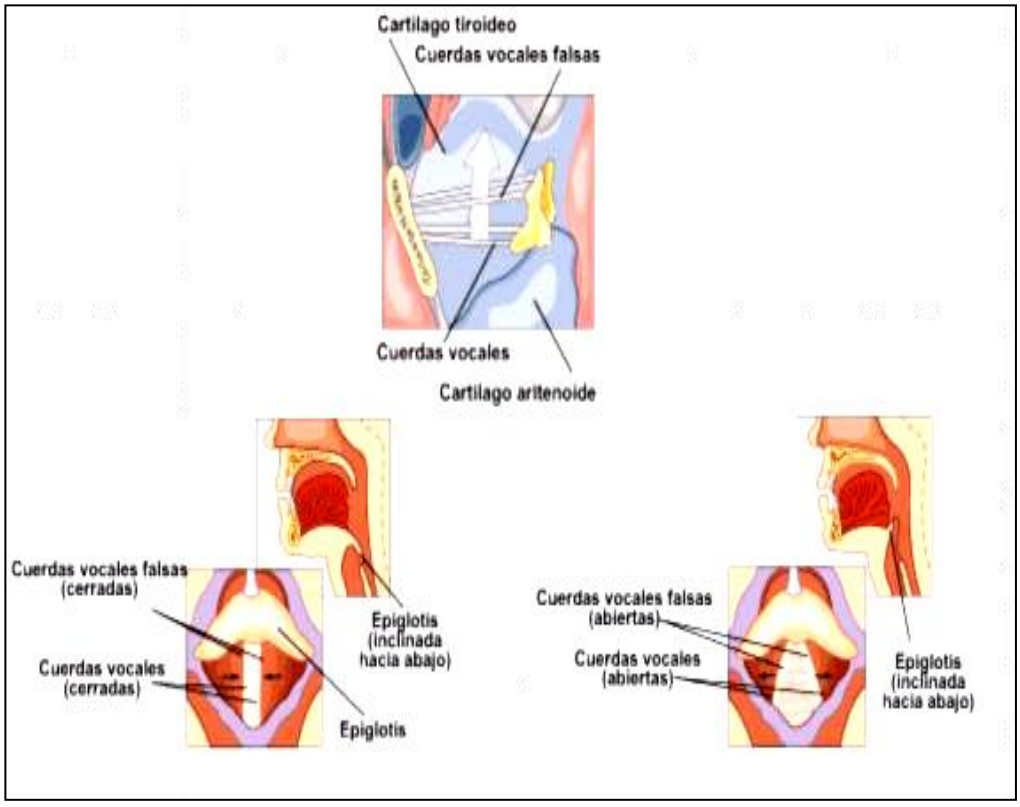
Que tengan la **mayor relación posible** con el problema que tratamos. En el caso del reconocimiento de voz conviene tener en cuenta las características del oído si es posible. Por ejemplo no tiene mucho sentido usar como entradas las componentes de bajas frecuencias de la voz si luego el oído no es capaz de distinguirlo. Para el tratamiento de voces se usa un rango de frecuencias entre 100 Hz. y 4000 Hz. y para las vocales las componentes espectrales más importantes se encuentran entre 100 ó 200 Hz y 2500 Hz). Por tanto hay muchos valores que no influyen nada en el reconocimiento de la voz tal como lo hace el ser humano. Aunque quizá hay aspectos que el oído no capta que puedan ser interesantes para el reconocimiento automático del habla.

Si es posible interesa que tengan **valores concordantes**, es decir, voces que suenen de manera similar a las palabras que queremos reconocer y sobre todo que tengan valores suficientemente distintos para salidas deseadas distintas. Por tanto no sería lógico por ejemplo introducir las muestras de voz directamente a la entrada de la red porque muestras de voz de aspecto bastante diferente debido a ruidos o distorsiones suenan de manera similar. Sin embargo, como una red neuronal puede aproximar prácticamente cualquier función, si retardamos la entrada un número de veces igual a la longitud de la ventana que usamos para obtener parámetros podríamos entrenar una subred para calcular parámetros y luego combinar esta subred con otra para obtener los mismos resultados... Que se calculen de forma fácil y sobre todo rápida a partir de la señal de voz original. Esto interesa sobre todo para una implementación final en tiempo real. Si las entradas además de ser pocas son las adecuadas habremos simplificado muchísimo tanto la red como el proceso de aprendizaje.

Las **altas componentes espectrales** se corresponden con variaciones rápidas en el espectro y por tanto corresponden al **rizado del espectro** el cual se relaciona estrechamente con la frecuencia fundamental y el carácter periódico de la **excitación** aplicada al tracto vocal. Las **bajas componentes espectrales** se corresponden con variaciones lentas de las componentes espectrales y por tanto contienen información de la **envolvente del espectro**, la cual se relaciona con la respuesta en frecuencia del **filtro que modela el tracto vocal.**

Figura 4.9 *Cuerdas Vocales*

Como lo que interesa para distinguir voces normalmente no son las características de la excitación, sino las características del tracto vocal o el filtro que le modela por tanto usaremos las bajas componentes espectrales para reconocer voces y no locutores. Hay que saber que la mayor información del locutor está en las cuerdas vocales.



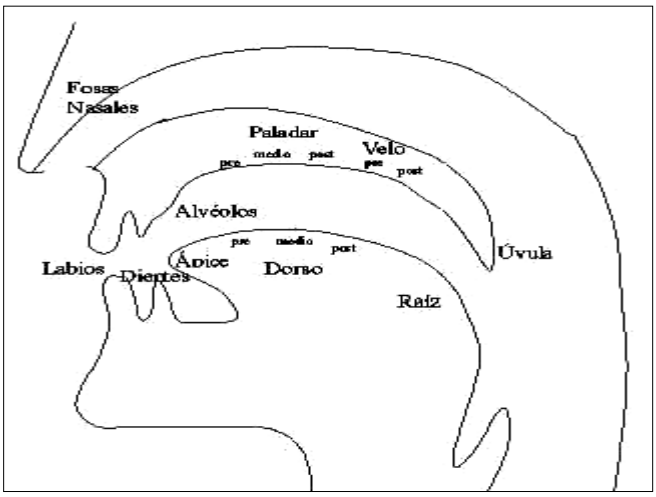
Según estas ideas y el desarrollo hecho al definir los espectros se comprende cómo se pueden usar los espectros para separar la información correspondiente a la excitación y la que corresponde al filtro resonador (tracto vocal).

Si bien es cierto que dos voces que suenan parecido tienen parámetros parecidos lo que no está muy claro es que dos voces que suenan diferentes tengan que tener espectros diferentes. Puede ser que dos vocales diferentes tengan los mismos espectros o muy parecidos.

Figura 4.10 *Zonas Donde se Produce la Modulación de los Sonidos*

4.8.2. VECTOR DE CARACTERISTICAS

Se basa en la idea de usar como entrada a la red una recopilación de otros parámetros para que aprenda a reconocer en la base a ellos. Deben ser bien escogidos para el caso



que corresponda.

Los formantes tienen relación con la envolvente y con el tracto vocal como filtro. Por tanto pueden ser tan buenos o mejores que los espectros para distinguir entre vocales o distinguir vocales de algunas consonantes. Además al incluir un parámetro de periodo fundamental que se anula para los tramos sordos tiene mayor capacidad de distinguir consonantes. Sin embargo, esto puede mejorarse filtrando la señal para simular el oído.

4.8.3. PARAMETROS DE SALIDA

Al hablar de reconocimiento de voz el problema casi coincide con el de clasificar las entradas, es decir, asignar a cada entrada la clase a la que pertenece. Por tanto parece lógico que la salida de nuestra red neuronal tenga un valor para cada clase.

Cuando sólo haya dos clases la salida puede ser una neurona que saque un cero para los valores de una clase y un uno para los de otra clase.

Cuando haya más de dos clases podemos asignar una neurona a cada clase que por ejemplo saque valor uno si las entradas pertenecen a esa clase y cero en caso contrario. Sin embargo, también existen otras posibilidades interesantes: Asignar un código binario a cada clase o asignar un vector a cada clase de forma que se asigne a las clases similares vectores similares.

4.9. NOTAS BIBLIOGRAFICAS

El sistema de reconocimiento de voz está basado en complejos métodos probabilísticos y modelos lingüísticos. La conversión de la palabra hablada en texto se realiza a través de sofisticados algoritmos que aíslan, identifican e interpretan los componentes fonéticos individuales del habla humana.

Las aplicaciones de estos sistemas son múltiples. Además de permitir a cualquier usuario sustanciales mejoras de comodidad y ahorro de tiempo en la habitual tarea de introducir textos en el ordenador, el sistema de reconocimiento de voz resulta de extraordinaria utilidad. Así, el radiólogo puede dictar sus conclusiones mientras examina con total libertad una radiografía y el periodista escribir un reportaje al tiempo que consulta otros documentos.

Un primer acercamiento hacia la comprensión de un codificador de voz, será la definición de criterios que permitan determinar la calidad de la señal de voz recibida en el extremo receptor. A este criterio se le denominará Criterio de fidelidad.

Las propiedades utilizadas en la creación de un código para un codificador de forma de onda corresponden a la explotación de la redundancia de las características de la señal de voz, ya sea en el dominio del tiempo o en el dominio de la frecuencia.

Es así como en el dominio del tiempo se utilizan las siguientes redundancias:

- Distribución no uniforme de la amplitud.
- Correlación entre muestra y muestra.
- Correlación ciclo a ciclo (periodicidad).

- Correlación entre intervalos de igual duración (pitch interval).
- Factores de inactividad de la voz (silencios).

Dentro de las aplicaciones de la compresión de audio, se distinguen 4 áreas: difusión, almacenamiento, multimedia y telecomunicaciones. Un algoritmo uno de los más populares y antiguos codificadores de voz, el codificador de voz de canal.

- INTERNET

  - www. IBM.com
  - www. Dragon_Speaking.org