UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

Diseño de un sintetizador de voz por difonemas

TESIS

Que para obtener el título de

Ingeniero en Computación

P R E S E N T A

Fernando del Río Ávila

DIRECTOR DE TESIS

Dr. José Abel Herrera Camacho

Ciudad Universitaria, Cd. Mx., 2003

INDICE

INTRODUCCION

Un sistema de síntesis de voz es aquel sistema que genera por medios mecánicos, eléctricos o electrónicos una salida de audio que simule la voz humana. Actualmente estos sistemas son muy usados. Algunos ejemplos de sus usos pueden ser los sistemas lectores de correo electrónico por medio de un sistema telefónico, lectores de libros para ciegos, sistemas de información electrónicos por medio de teléfono, en películas para generar efectos especiales de sonido o simplemente como un método adicional de interacción con una computadora personal. Además, varios de estos sistemas hacen también uso de la contraparte de los sintetizadores de voz: los sistemas de reconocimiento de voz.

Los sistemas de síntesis de voz han existido desde el siglo XVIII. Aunque los primeros sistemas eran completamente mecánicos, de difícil operación y solo podían generar algunos fonemas, actualmente los sistemas de síntesis de voz son electrónicos, sea por medio de una computadora o en circuitos integrados de síntesis. Debido al avance de las computadoras, en cuanto a capacidad de almacenamiento y velocidad, la inteligibilidad (comprensión) de estos sistemas es bastante grande, así como su facilidad de uso.

Los sistemas de síntesis de voz aun no tienen una naturalidad completa (semejanza con una voz humana real), es esta un área de investigación amplia en la actualidad.

En el desarrollo de esta tesis se pretende el diseño y construcción de un sistema de síntesis de voz básico que pueda generar una salida de voz para la entrada de un texto cualquiera. Este sistema será de tipo concatenativo (unión de segmentos previamente almacenados y clasificados) utilizando como unidades básicas los difonemas (segmentos que contienen las mitades final e inicial de dos fonemas consecutivos).

Este programa recibirá una entrada de texto y la almacenara en un archivo tipo ‘wav’ para posteriormente ser reproducida.

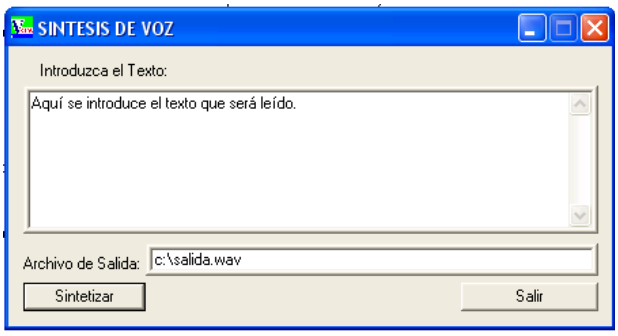

Figura 0.1: Pantalla principal del sistema.

En la figura se observa la pantalla principal del sistema que va a ser diseñado. El sistema consta de una ventana donde puede ser escrito o pegado el texto a ser leído y un renglón donde se introduce el nombre de archivo de la salida. En el caso de que uno o ambos sean omitidos se llenaran automáticamente con valores por omisión.

1.EL SONIDO

1.1. Características Físicas del Sonido

El sonido se produce cuando un cuerpo vibratorio transmite esta vibración hacia un medio elástico. El sonido se propaga de forma longitudinal, es decir, las vibraciones mecánicas que constituyen la onda siguen la dirección de propagación. De esta forma el sonido se comporta de forma similar a un resorte donde la onda se propaga como un conjunto de capas que empuja o jala a la capa siguiente.

Así la onda sonora consiste de compresiones (zonas de alta presión) y ramificaciones (zonas de baja presión) con relación a la presión de equilibrio circundante. Debido a que el sonido debe ser transmitido a través de un medio este no puede generarse en el vacío a diferencia de la luz u otras radiaciones electromagnéticas.

El patrón generado de zonas de alta y baja presión se denomina onda sonora y esta es transmitida a diferentes velocidades de acuerdo al medio (por ejemplo: 340 m/s en el aire y 1,500 m/s en el agua).

Las principales características del sonido son su amplitud y su frecuencia. Para explicar estos términos se necesita observar una onda de tono puro. Esta es una onda sonora donde la variación de presión en el tiempo corresponde a la función matemática seno.

En esta onda pura la amplitud es la variación máxima de presión con respecto a la presión atmosférica normal y la frecuencia es el número de veces en una unidad de tiempo que se repite el ciclo.

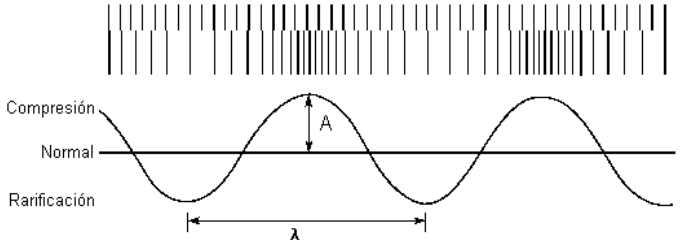

Fig. 1.1: Comparación del aire en equilibrio y el aire en presencia de una onda sonora y su representación transversal indicando las zonas de compresión y ramificación del aire. A corresponde a la amplitud de la onda y λ corresponde a la longitud de la onda.

1.1.1. La Amplitud

La magnitud de la variación de la presión en un medio debido a la onda sonora se conoce como amplitud de la onda, esta variación se mide normalmente en pascales o newtons por metro cuadrado. El desplazamiento de una onda de sonido de un tono puro se puede describir con la ecuación general de ondas:

$$y(x,y) = A\sin\left(2\pi\left(ft - \frac{x}{\lambda}\right)\right)$$

Esta ecuación indica la magnitud de la perturbación a una distancia (x) de la fuente sonora en el tiempo (t). Esta variación es de forma senoidal y depende de la magnitud máxima de la variación o amplitud (A), la frecuencia (f) de la señal y su longitud de onda (λ).

Además podemos obtener la Intensidad acústica que es el promedio de la energía de transmisión por unidad de área perpendicular a la dirección de la propagación:

$$I = \frac{A^2}{2S\rho}$$

ρ es la densidad del aire en equilibrio (kg/m^3), S es la velocidad del sonido en el medio de transmisión y A es su amplitud. La intensidad se obtiene en wats por metro cuadrado.

El rango de amplitudes que puede percibir el oído humano es muy grande. Por ejemplo el umbral de dolor (el ruido mas fuerte que puede ser percibido sin dañar el oído) es un millón de veces mas grande

que el umbral de ruido (el ruido mas pequeño perceptible por el oído humano) . Sin embargo, debido al rango tan grande, este es percibido de forma no linear, teniendo una mejor recepción a bajas intensidades y percibiendo las variaciones de forma menos eficiente en intensidades mas grandes.

Debido a esto la amplitud no se mide directamente sino en una unidad llamada decibeles, la cual nos da un comportamiento similar. Estos son calculados de la siguiente manera:

$$dB = 20\log\left(\frac{I}{I_0}\right)$$

donde: dB es la medida en decibeles, I es la intensidad del estimulo y I_o es una intensidad predeterminada. La I_o que se utiliza normalmente es de 10^{-12} Watt/m^2, que es aproximadamente el umbral de ruido para una frecuencia de 1000 Hz., por lo que a esta intensidad le corresponden los 0dB.

Con esta escala un aumento de un orden de magnitud corresponde a un aumento de 10dB, por lo que el umbral de dolor (1 Watt/m^2) tiene un valor de 120dB.

Decibel (dB)	Intensidad (Watt/m^2)	Ejemplo
0	10^{-12}	Umbral de ruido
30	10^{-9}	Biblioteca
60	10^{-6}	Conversación Normal
90	10^{-3}	Interior de un camión
120	1	Turbina de Jet (Umbral de dolor)

Tabla 1.1: Algunos valores de decibel con su valor en amplitud y un ejemplo del valor.

1.1.2. La Frecuencia

La distancia que recorre una onda pura antes de regresar a la posición inicial para reiniciar el ciclo es conocido como longitud de onda. De acuerdo a la velocidad de la onda sonora esta longitud de onda tarda cierto tiempo en recorrer un ciclo (periodo de la onda). Si obtenemos el inverso de este periodo obtenemos el numero de ciclos que lleva acabo la onda en una unidad de tiempo (frecuencia). Esta

frecuencia se mide normalmente en Hz (1/seg) De esta forma entre mayor sea la frecuencia de una señal menor serán su periodo y su longitud de onda.

1.1.3. La velocidad del sonido

Medio	**Velocidad (m/s^2)**
GASES	
Aire (0^oC)	331.29
Oxigeno (0^oC)	316
Vapor (134^oC)	494
LIQUIDOS	
Agua (0^oC)	1402.3
Agua (50^oC)	1542.5
Agua Salada (20^oC)	1522.6
SÓLIDOS	
Granito (20^oC)	6000
Aluminio (20^oC)	5100
Plomo (20^oC)	1230

Tabla 1.2: Velocidad del sonido en diferentes medios(13)

La velocidad de la onda sonora depende del medio en el que se transmite.

Para ondas longitudinales como el sonido en un medio gaseoso esta se obtiene por:

$$S = \sqrt{\frac{B}{\rho}}$$

Donde ρ es la densidad y B es el modulo entre la presión y el cambio de volumen por unidad de volumen del medio. A partir de aquí se puede obtener la velocidad utilizando la presión o temperatura presión:

$$S = \sqrt{\frac{p\gamma}{\rho}}$$

Donde S es la velocidad del sonido, p es la presión de equilibrio, ρ es la densidad de equilibrio en kilogramos por metro cúbico y γ es el modulo del calor especifico a presión constante y el calor especifico a volumen constante.

Temperatura:

$$S = \sqrt{\frac{R\theta\gamma}{M}}$$

Donde θ es la temperatura en grados Kelvin y M es el peso molecular del gas. En un gas especifico R,γ y M son constantes por lo que resulta:

$$S = Cte_{gas}\sqrt{\theta}$$

1.1.4. Ondas Esféricas

Hasta el momento se han considerado las ondas sonoras como un fenómeno lineal, pero en realidad las ondas sonoras se expanden en un conjunto de frentes de ondas esféricos. El mecanismo de propagación conocido como Principio de Huygens dice que cada punto de la onda es una fuente de ondas esféricas. Este conjunto de ondas se propaga hacia el frente generando una onda esférica coherente. La propagación hacia atrás interfiere con la propagación ya existente eliminándose, por lo que solo se mantiene el desplazamiento hacia el frente. Este fenómeno se conoce como principio de superposición de ondas. Además este mismo principio permite que una onda mantenga sus características independientemente de la existencia de otras ondas.

1.1.5. Atenuación de las ondas sonoras

1.1.5.1. Atenuación por propagación

Cuando se considera que las ondas se transmiten de forma lineal, estas deberían poder transmitirse una distancia indeterminada sin atenuación, sin embargo, como las ondas se propagan en forma esférica, la energía se propaga hacia una esfera de radio cada ves mayor, por lo que la intensidad debe ir disminuyendo al aumentar el área. La intensidad es inversamente proporcional al cuadrado del radio. SI

consideramos esta atenuación en decibeles, la intensidad disminuye 6 decibeles por cada aumento de distancia en un factor de dos. Además, si la onda se propaga cerca de un medio absorbente, este causara una atenuación de 12 decibeles en su cercanía.

1.1.5.2. Atenuación por Absorción.

Además de la atenuación debida a la propagación, una pequeña parte de la onda sonora es absorbida por el aire por diferentes medios. Uno de estos es la producción de calor debido al choque de las moléculas de aire y la viscosidad del medio. Esta atenuación es dependiente de la frecuencia, siendo mayor para frecuencias mas altas y es menor en materiales mas densos como el agua o algunos sólidos como se observa en la tabla. Además esta absorción no es lineal, cada material absorbe mejor ciertas frecuencias que otras.

El coeficiente de absorción se calcula obteniendo el promedio de absorción del medio a 250,500,1000 y 2000 Hz.

Material	Coeficiente de Absorción
Oxigeno	165.0
Aire	137.0
Agua (0^oC)	0.569
Agua (20^oC)	0.253
Agua (80^oC)	0.079

Tabla 1.3: Coeficientes de absorción Sonora para varios materiales

1.1.5.3. Impedancia acústica

Diferentes materiales tienen una resistencia diferente a la propagación de las ondas sonoras y depende de la densidad del material y la velocidad del sonido y se mide en $Pa*m^2$:

$$Z = V\rho$$

Cuando una onda sonora pasa de un medio a otro con diferentes impedancias una parte de esta onda es reflejada por el medio y el resto es transmitido hacia el segundo medio.

1.1.6. Otras propiedades del sonido

1.1.6.1. Difracción, Refracción y Reflexión

Debido al principio de Huygens el sonido puede rodear obstáculos o expandirse después de pasar a través de un pequeño agujero. A esta propiedad se le conoce como difracción. Entre menor sea la frecuencia de una señal mejor será su capacidad para difractarse, por lo que señales de alta frecuencia pueden no rodear completamente el objeto y crear una sombra sonora tras el objeto. Además algunas ondas son reflejadas por el objeto creando interferencia y como resultado zonas con diferentes intensidades.

Además cuando una onda sonora varia su velocidad por pasar de un material a otro o por variaciones de temperatura en el mismo medio, la onda puede refractarse o desviarse, causando así que se propague a una distancia menor o mayor de acuerdo a las condiciones que desvíen la onda hacia el suelo aumentando su distancia de propagación o hacia el aire disminuyéndola,

Cuando las ondas sonoras llegan a un medio diferente, estas son reflejadas en cierta medida dependiendo del material en dirección perpendicular al Angulo con el que llegan a la frontera entre los medios.

1.1.6.2. Interferencia

Cuando dos ondas de la misma frecuencia se cruzan pueden sumarse si llegan en fase y aumentar su intensidad o restarse si llegan fuera de fase pudiendo llegar incluso a eliminarse ambas.

1.1.6.3. Efecto Doppler

Cuando el receptor o la fuente están en movimiento con respecto al otro la frecuencia se percibe de forma diferente, aumentando si esta se esta acercando y disminuyendo si esta se aleja.
Si es un movimiento en línea recta la frecuencia que se escucha es:

$$f_o = f_s \frac{S + v_o}{S - v_o}$$

Donde fs es la frecuencia original, S la velocidad del sonido y Vo la velocidad relativa del receptor con el transmisor, siendo esta negativa si se están alejando.

1.1.6.4. El análisis espectral

Una sonido cualquiera en un momento determinado esta compuesta por una o mas componentes senoidales. Este conjunto de componentes se denomina espectro de la señal. La señal de mas baja frecuencia es la fundamental, siendo las demás sus armónicas, donde la frecuencia de cada armónica es un múltiplo de la frecuencia fundamental.

Cuando una señal es periódica se puede obtener un análisis de su espectro por medio de su transformada de Fourier, sin embargo si se aplica la transformada de directamente se obtiene también en el espectro el ruido de la cuantización, aunque este se puede minimizar en señales periódicas tomando diferentes secciones y promediando el espectro de cada una.

Además, como no se esta tomando la señal original sino solo una sección de esta el espectro resultante aparece deformado.

Si es una señal no periódica, es decir que cambia su frecuencia con el transcurso del tiempo se utiliza un espectrograma que consiste en una grafica de tres dimensiones donde se hace un análisis de Fourier

de la señal en bloques de un tamaño determinado graficando como varia la intensidad de cada frecuencia en el transcurso del tiempo.

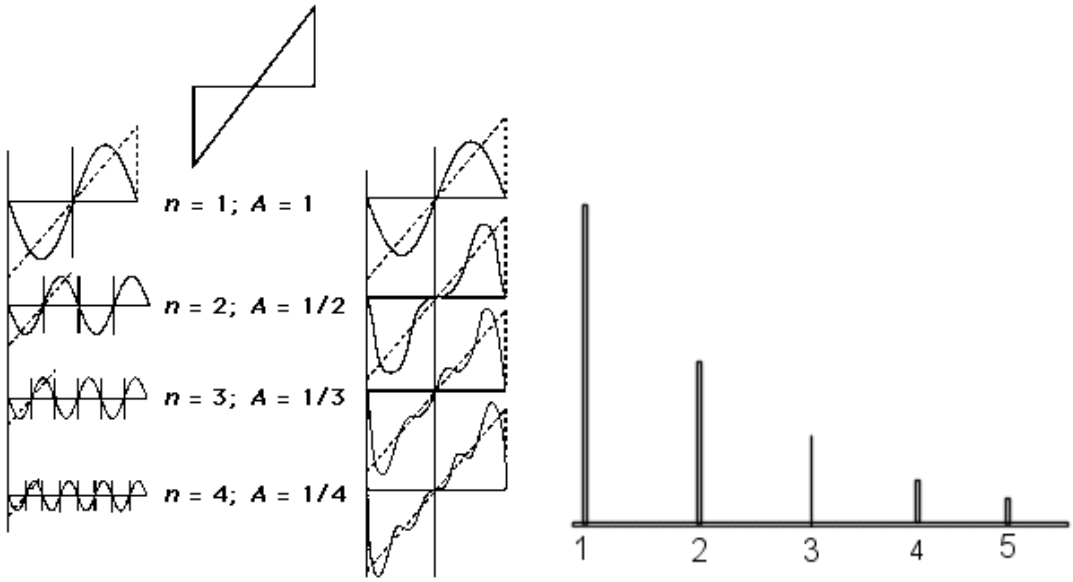

Figura 1.2: Aquí se observa una onda triangular de frecuencia 1 y su espectro. Nótese como la frecuencia de las armónicas son múltiplos de la frecuencia fundamental.

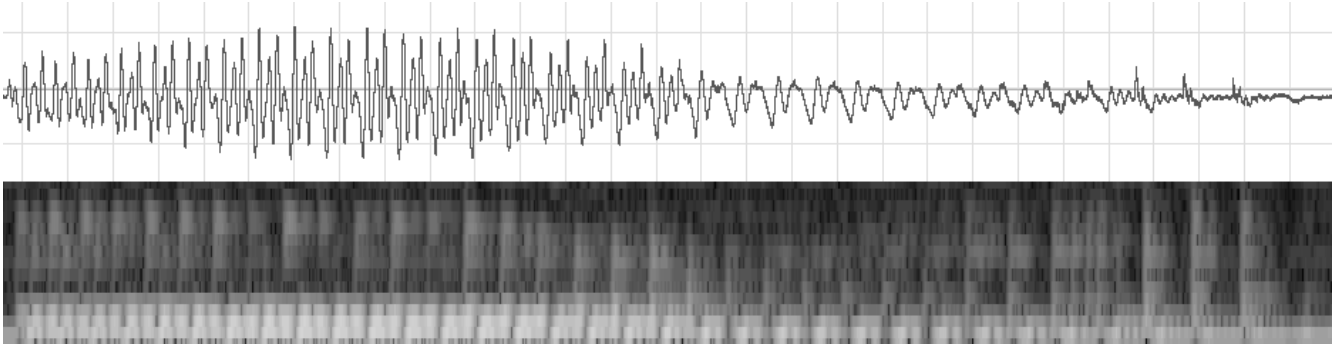

Figura 1.3: Onda y espectrograma de la palabra "hola". El eje horizontal corresponde al tiempo, el vertical a la frecuencia y el color a la intensidad (mas claro mayor intensidad) Nótese como varia el espectro con el paso del tiempo.

1.2. Almacenaje y reproducción digital del sonido.

1.2.1. Conversión Analógica-Digital

Aunque el sonido se presenta de forma continua, en un sistema digital como una computadora no puede ser almacenado de esta forma ,por lo que se necesita transformarlo de alguna manera para poder ser introducido a esta. Para hacer esto se utilizan los ADC (convertidores analógicos-digitales) para almacenar la señal y los DAC (Convertidores digitales-analógicos) para regresarla (aunque de forma aproximada) a una onda discreta.

En un ADC el primer paso consiste en introducir la señal en un "Sample and Hold" cuya función es tomar el valor de la señal cada cierto tiempo y mantener ese valor constante entre cada lectura ignorando los cambios que ocurran en ese intervalo, con lo que convierte la variable independiente (tiempo) de continua a discreta. Este valor se mantiene durante el periodo para que el modulo de cuantización tenga el valor a la entrada el tiempo suficiente para convertirlo a un valor digital.

1.2.2. Cuantización

Cuando se desea convertir una señal analógica a una señal digital se necesita tomar un numero infinito de valores y convertirlo a un numero finito de valores que pueda ser representado con un numero finito de símbolos.
Estos símbolos son en la mayoría de los casos números binarios, por lo que el numero máximo de símbolos que se pueden utilizar para representar diferentes amplitudes es 2^n, donde n es el numero de bits que se usan para representar cada muestra.

1.1.1.1. Cuantización Uniforme

La forma mas fácil de cuantizar una señal es utilizando una Cuantización uniforme donde los valores asignados a cada símbolo tienen un espaciado uniforme. Existen dos tipos de cuantizadores uniformes, el mid-riser y el mid-tread.

La separación entre los valores se obtiene por:

$$\Delta = \frac{2X_{\max}}{2^n}$$

donde X_{max} es el valor máximo que puede tomar la señal, Δ es el espaciado entre cada valor siendo n el numero de símbolos a utilizar.

En el mid-tread se utiliza el cero como un valor, por lo que hay un valor menos del lado positivo que del lado negativo, por lo que el rango es de +(Max-Δ) a –(Max)
En el Mid-Riser no se utiliza el cero como valor, por lo que hay el mismo numero de valores positivos y negativos. En este los valores centrales son +/- (Δ/2), por lo que el rango es de +(Max-Δ/2) a –(Max-Δ/2).

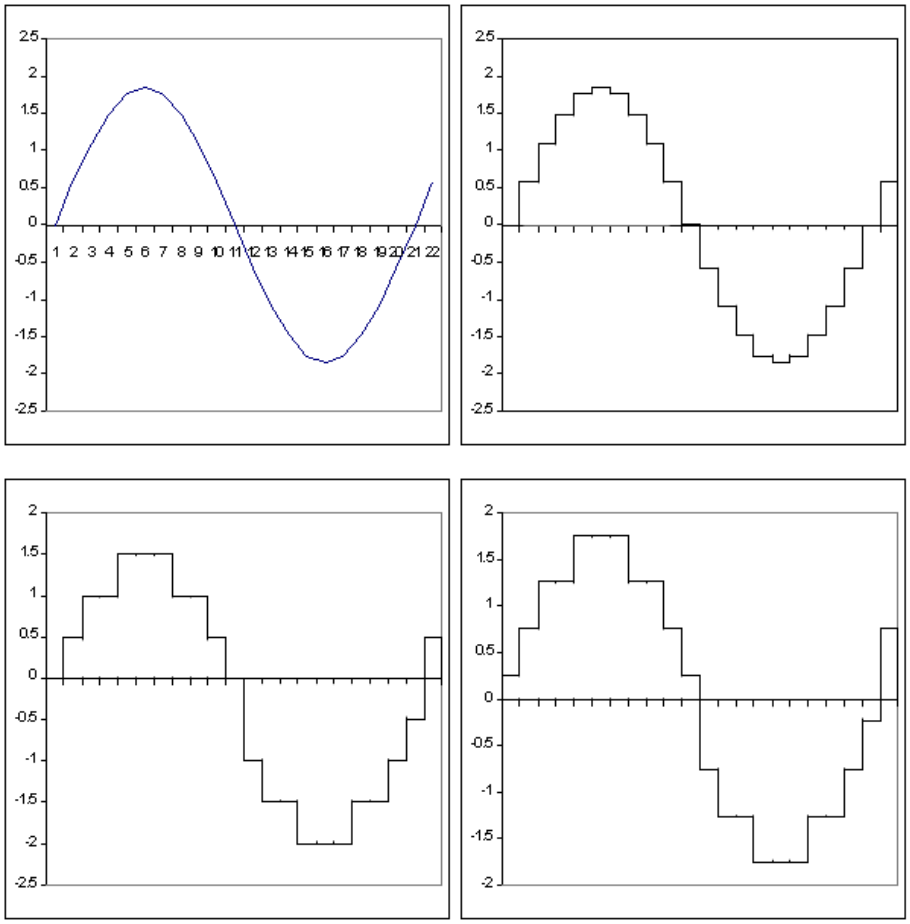

Fig. 1.4 En esta figura se observa la señal original, la señal al pasar por el Sample and Hold y la señal cuantizada a 8 niveles (3 bits) con mid-tread y mid-riser.

En este cuantizado el error máximo es de Δ/2 positivo o negativo. En condiciones normales tiene una media de cero y una desviación de Δ/sqr(12). A partir de aquí podemos calcular la relación del ruido agregado con la señal original (SNR):

$$SNR = \frac{(3)2^{2n}}{\left(\frac{X_{\max}}{\sigma_x}\right)^2}$$

$$\sigma_x = \sqrt{\Sigma x^2(n)}$$

y si expresamos esta ecuación en dB:

$$SNR(dB) = 10\log\left(\frac{\sigma_x^2}{\sigma_e^2}\right) = 6B + 4.77 - 20\log\left(\frac{X_{\max}}{\sigma_x}\right)$$

Sin embargo, en señales de voz donde hay segmentos con amplitudes mas bajas esto no se cumple, porque al no llegar la señal al máximo es equivalente a que se estuviesen utilizando menos bits para cuantizar la señal.

1.2.2.1. Cuantización Ley μ

Para obtener un mejor cuantizado los niveles no son uniformes, sino variados logaritmicaménte. O utilizamos un cuantizado uniforme pero cuantizamos el logaritmo de la señal de entrada.

y(n)=log|x(n)|

con esto logramos que la relación señal-ruido sea independiente de la magnitud de la señal:

$$SNR = \frac{1}{\sigma_\varepsilon^2}$$

Sin embargo, este tipo de cuantizacion no se puede utilizar de forma practica debido a la forma en que se distribuyen los valores y que requerirían un numero infinito de valores de cuantizacion. Por esto se utiliza una variación llamada ley-μ:

$$y(n) = \frac{\log\left[1 + \mu \frac{|x(n)|}{X_{\max}}\right]}{\log[1+\mu]} signo[x(n)]$$

En este tipo de cuantizacione el SNR es:

$$SNR = 6B + 4.77 - 20\log[\ln(1+\mu)] - 10\log\left[1 + \left(\frac{X\max}{\mu\sigma_x}\right)^2 + \sqrt{2}\left(\frac{X_{\max}}{\mu\sigma_x}\right)\right]$$

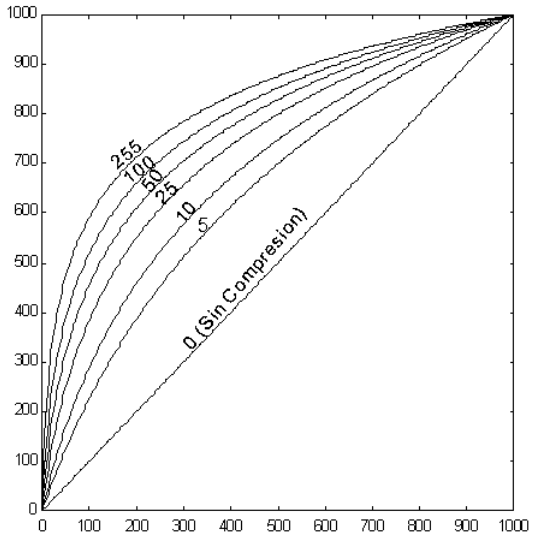

Fig 1.5 Relacion Entrada/salida para diferentes valores de μ

1.2.2.2. Cuantización Diferencial

Los tipos anteriores de cuantizacion dan un buen resultado en señales con amplitudes mas o menos estables, sin embargo en señales como las voz donde hay grandes variaciones presentan problemas. Una solución a este problema seria el hacer que la Δ del cuantizador varíe de acuerdo con las características de la señal, variando la ganancia de la señal de entrada para mantener una varianza constante.

Al observar la señal se observa que la variación entre una muestra y la siguiente es pequeña, por que en ves de cuantizar la señal cuantizamos $X(n) - \tilde{X}(n)$, donde $\tilde{X}(n)$ es el valor predicho de X(n) de acuerdo a las muestras anteriores, es decir, alimentamos el cuantizador con el error de predicción que tiene una varianza mucho menor a la señal.

El valor de la predicción se obtiene con:

$$\tilde{x}(n) = \sum_{k=1}^{P} a_k x_q(n-k)$$

$$x_q(n) = d(n) + \tilde{x}(n)$$

donde a_k es un valor de peso para cada muestra, P es grado del predictor. Y d(n) es la diferencia entre el valor predicho y el valor normal.

Uno de los cuantizadores diferenciales mas sencillos es el LDM (Modulación delta lineal) que utiliza solo un bit con la siguiente tabla:

Signo de d(n)	c(n)	D(n)
+	0	+Δ
-	1	-Δ

Tabla 1.4: Cuantizador LDM

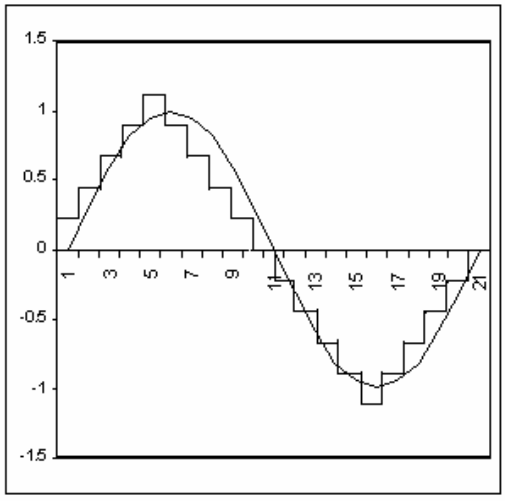

Fig. 1.6 Modulación LDM de una señal senoidal.

1.2.3. Teorema del muestreo y teorema de Nyquist

Para que exista un muestreo correcto de la señal la versión digital de la misma debe ser suficiente para reconstruir la señal analógica original. Sin embargo en algunos casos esto no sucede, sino que al tratar de reconstruir la señal se obtiene una señal diferente.
Para que una señal analógica pueda ser reconstruida se requiere que las muestras se tomen a una frecuencia de al menos el doble de la frecuencia máxima de la señal. A esto se le conoce como el teorema de Nyquist.

Cuando existen frecuencias por encima de la frecuencia de Nyquist estas son destruidas y se obtiene una señal diferente de una frecuencia menor que se encuentra dentro del rango de señales almacenables.

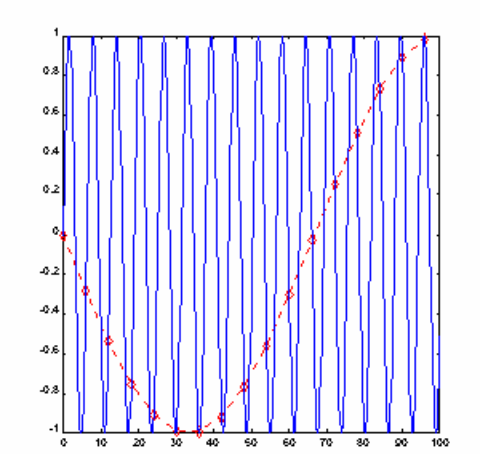

Fig. 1.7 Una señal es muestreada a .95 de su frecuencia original. Nótese que se obtiene una señal de una frecuencia mucho menor.

Este efecto también se produce a la inversa. Si observamos el espectro de una señal digital observamos que el espectro de frecuencias menores a ½ de la frecuencia de Nyquist se encuentra repetido en las partes mas altas del espectro. Nótese que las copias iguales están intercaladas por copias inversas. Cada copia inicia en $nf_s/2$, donde n=1,2,3... y f_s es la frecuencia de muestreo.

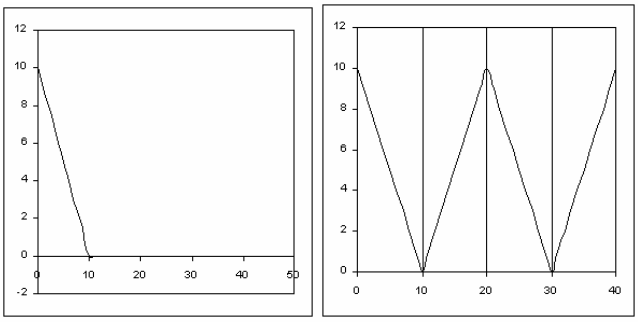

Fig. 1.8 Espectro de una señal continua y el espectro de su equivalente discreto.

Este fenómeno es el que produce que cuando existe una señal por encima de ½ de la frecuencia de nyquist aparezca una nueva señal de frecuencia menor al encimarse la parte inferior de la copia sobre la señal original. A esto se le conoce como Aliasing, debido a que la frecuencia toma un alias o una nueva identidad.

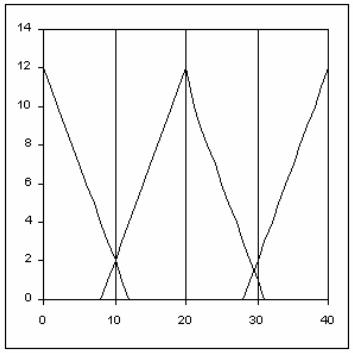

Fig. 1.9 Señal con aliasing debido a una frecuencia de muestreo muy baja.

1.2.4. Conversión digital-analógica

Una vez que la señal ha sido almacenada en forma digital es necesario regresarla a una señal analógica para su reproducción.

Como el tren de impulsos es igual a la señal original mas un numero infinito de copias la forma mas sencilla de reconstruirlo seria alimentar el tren de pulsos a un filtro paso bajas para reobtener la señal original, sin embargo, debido a las características de los circuitos no se puede transmitir un tren de

pulsos, sino que se envía el valor de cada muestra durante un periodo n de tiempo, además de el tiempo que tarda en llegar al valor deseado.

Esto genera que cada pulso este en realidad multiplicado por una señal rectangular, por lo que el espectro de la señal a la salida del filtro es en realidad la función original multiplicada por una función denominada sinc.

$$\sin c(x) = \frac{\sin(\pi x)}{\pi x}$$

Esta distorsión se puede compensar utilizando un amplificador variable, que amplifique cada frecuencia de forma diferente, multiplicando el espectro por 1/sinc.

2.ANATOMÍA Y FISIOLOGÍA DE LA VOZ HUMANA

2.1. El sistema generador de voz

La producción del sonido desde el punto de vista físico ocurren principalmente en tres secciones: el tracto pulmonar, la laringe y el tracto vocal. Además interviene también el sistema nervioso en el control de las funciones a realizar.

La producción física del sonido consiste básicamente en la generación de un chorro de aire (tracto pulmonar) que después es filtrada y enviada a través de un sistema vibratorio (en la laringe) y por ultimo modulada (tracto vocal) para generar los sonidos correspondientes.

2.1.1. El tracto pulmonar

Los pulmones están compuestos por una masa esponjosa con una capacidad aproximada de 4-5 litros. Estos se encuentran dentro de una cámara de aire conocida como pleura que se encuentra contenida por las costillas y sobre el diafragma.

El diafragma es un músculo con forma de domo que al comprimirse se extiende hacia fuera lo que hace que aumente el volumen de la pleura por lo que los pulmones se llenan de aire, obteniéndose el efecto contrario cuando este se relaja.

Cuando se produce una salida de aire voluntaria se utiliza fuerza adicional proporcionada por los músculos abdominales con lo que se aumenta la presión con la que el aire sale de los pulmones. Para producir sonidos se utilizan presiones de entre 4 y 20 cmH_2O.

Una vez que el aire sale de los pulmones este es conducido por medio de los bronquios hacia la traquea que es un conjunto de cartílagos con forma anular unidos por un tejido de donde llega a la laringe. Aunque este es un tubo rígido puede curvarse de acuerdo a los movimientos de la cabeza.

Estos movimientos de aire son los responsables principales de la creación del sonido y cuya intensidad esta relacionada de acuerdo con la presión de aire que es generada.

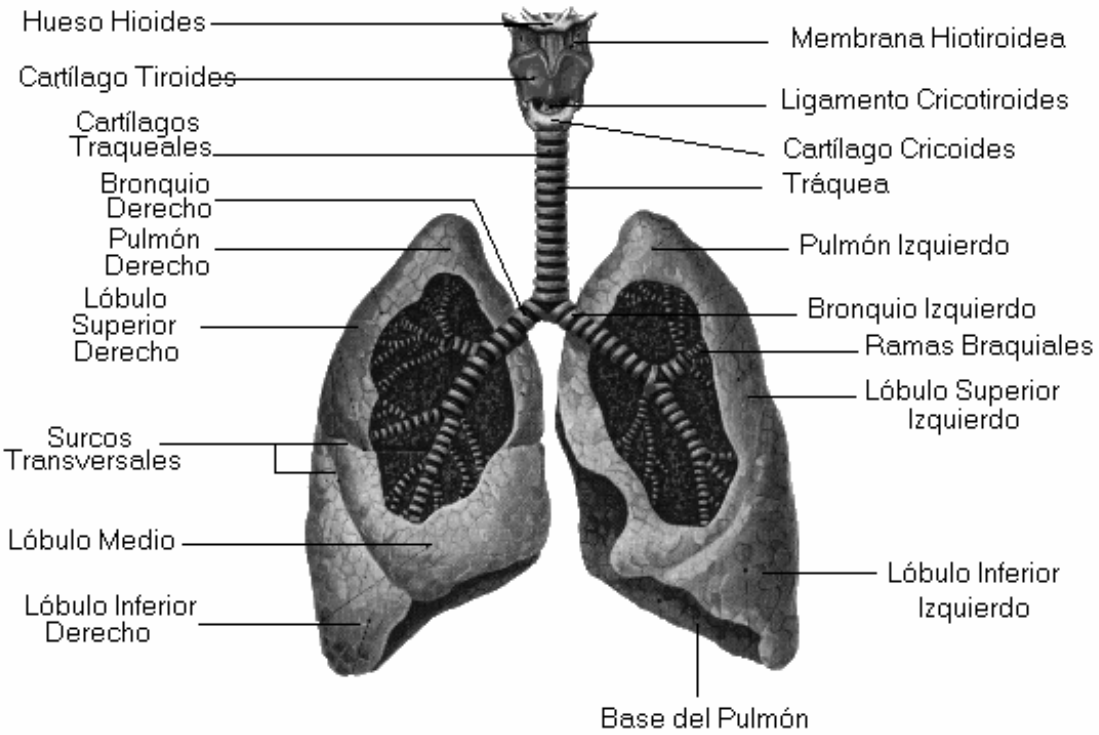

Fig. 2.1- El tracto pulmonar

2.1.2. La Laringe

Esta está formada por tres cartílagos (cricoides, tiroides, arteniode), un conjunto de músculos y las cuerdas vocales. Los cartílagos cricoides y tiroides son los que contienen y controlan las cuerdas vocales cuyas funciones son cerrar la traquea para impedir la entrada de objetos en los pulmones (por

ejemplo al deglutir) y permitir que se forme presión en el tórax y el abdomen y son los responsables de la generación de sonidos.

El cartílago cricoides es un anillo en la parte superior de la traquea y de altura mayor en su parte trasera. La tiroides esta colocada al frente de este y es el que resiste el empuje de las cuerdas vocales. El artenoides soporta la parte posterior de las cuerdas vocales y esta conectado en la parte alta del cricoides.

Estos cartílagos se encuentran controlados por un conjunto de músculos que permiten abrir y cerrar la abertura que se encuentra entre las cuerdas vocales.

Las cuerdas vocales están formadas por un tejido sólido con dobleces entre el frente y la parte posterior de la laringe. Cuando estas se encuentran separadas el conducto se encuentra abierto y permite el paso normal del aire (respiración). Si esta está cerrada se impide el paso hacia los pulmones para poder deglutir sin que lleguen objetos a los pulmones.

Para generar sonidos estas cuerdas se abren y cierran parcial o totalmente en ciertas secuencias regulares se generan vibraciones en la corriente de aire con lo que se generan sonidos.

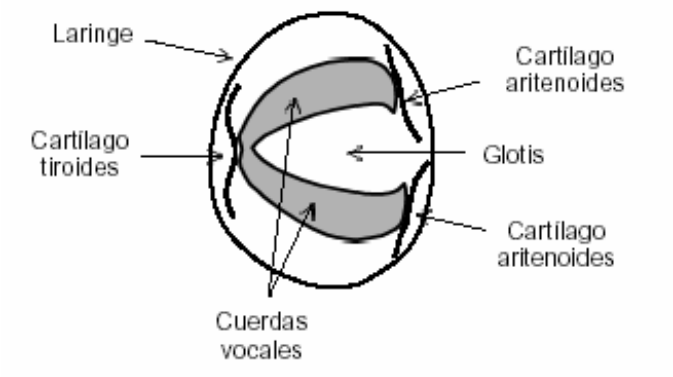

Fig. 2.2- Las cuerdas vocales

2.1.3. El tracto vocal

Por encima de las cuerdas vocales se encuentra el tracto vocal que consta de faringe (laringeal, oral y nasal) y las cavidades nasales y orales.

La parte superior (techo de la boca) esta dividida en dos secciones. Al frente se encuentra el hueso palatal atrás del cual se encuentra tejido conectivo con un músculo denominado velo el cual se puede elevar para sellar el tracto nasal. Atrás de el velo se encuentra la úvula que es un apéndice carnoso.
Frente al paladar se encuentra la arista alveolar que es donde los dientes se encuentra insertos. Además se encuentra la epiglotis que es un cartílago en forma de plato sobre las cuerdas vocales pero su función es independiente de la producción de voz.

La resonancia de este tracto modifica las señales acústicas y esta se puede modificar de acuerdo a la posición de su estructura moviendo la lengua y demás componentes móviles.

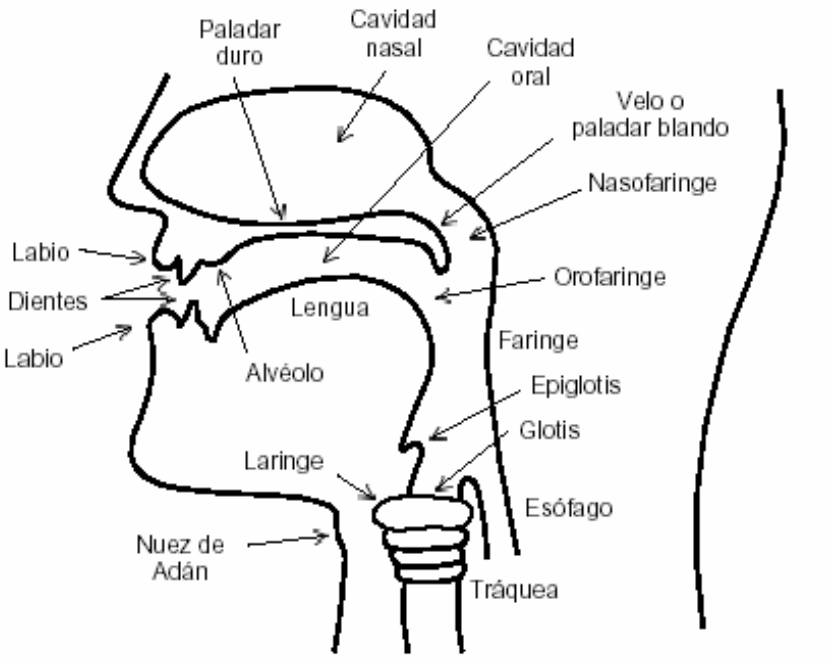

Fig. 2.3- EL tracto vocal

2.1.4. Tipos de excitación

De acuerdo a la forma en que se modifica el chorro de aire por los diferentes elementos de la laringe y el tracto vocal se pueden producir diferentes tipos de sonidos. Los principales son:

- Fonación: Este se produce por oscilaciones en las cuerdas vocales con lo que al ser forzada una corriente de aire estas vibran de acuerdo a la tensión de las cuerdas y debido al efecto de Bernoulli. La apertura y cierre de las cuerdas produce pulsos cuasi-periódicos de aire (pulsos glotales) con formas de onda similares a la triangular con ciclos de trabajo de entre .3 y .7. Además se produce un efecto de paso bajas por lo que contienen una fundamental fuerte y armónicas débiles. Los sonidos producidos por medio de fonación se denominan sonoros, mientras que los demás se consideran como sordos.

- Susurro: Estos son generados en la laringe. Las cuerdas vocales están solo semi-cerradas por lo que el aire genera turbulencias ocasionando ruido de banda ancha. Estos son de amplitud menor que los producidos por las fonaciones pero tienen mayor energía en las frecuencias alta.

- Fricación: Esta puede ocurrir con fonación o sin ella y esta formado al igual que el susurro con ruido de banda ancha pero en este caso la turbulencia se genera o es modificada en el tracto vocal en lugar de las cuerdas vocales. Es de mayor amplitud que el susurro debido a que el filtrado en el tracto vocal es menos intenso y existen menos perdidas

- Compresión: El tracto vocal se encuentra casi cerrado por lo que al exhalar el aire no sale sino que aumenta en presión hasta que se produce un transitorio por lo que se produce un silencio seguido por una ráfaga de ruido abrupto (plosiva) o con una caída gradual y turbulenta similar a un fricativo (africativa)

- Vibración: es cuasiperiódica y ocurre en el tracto vocal y pueden utilizar o no sonidos producidos por fonación y producen una modulación rápida y repetitiva.

2.2. El sistema receptor de voz

Para percibir el sonido se necesitan tres pasos básicos: el impulso sonoro debe llegar a los receptores, estos receptores deben transducir (transformación de un tipo de energía a otro diferente) las variaciones de presión en el aire en señales eléctricas y estas deben ser enviadas al cerebro para su procesamiento. Este proceso se lleva a cabo en el oído, el cual consta de tres partes: externa, media e interna.

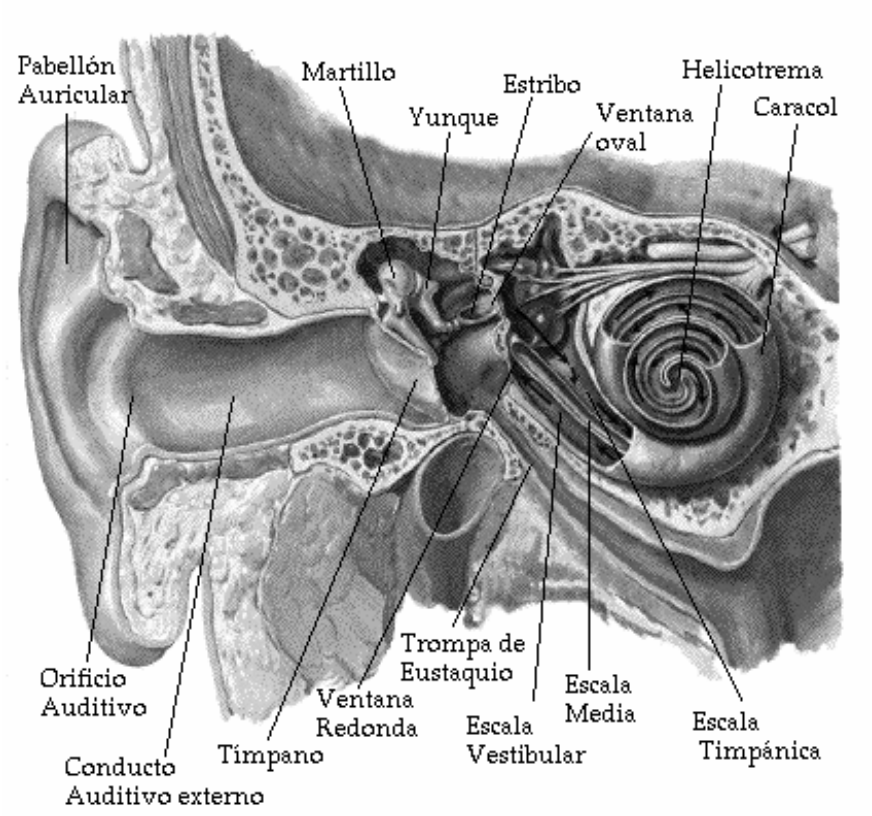

Fig. 2.4- El aparato auditivo

2.2.1. El oído externo

La parte mas externa del oído es la pinna (oreja) que es la estructura que se encuentra afuera de los oídos a los lados de la cabeza y que posiblemente ayuda a la localización de los sonidos.

En esta los sonidos son reflejados en las curvas de la misma de forma distinta según la frecuencia lo que genera un conjunto de ecos que parecen ser interpretados en la ayuda de la localización de los sonidos principalmente en el plano medio (plano perpendicular a la línea entre los tímpanos) . Al final de esta se encuentra la concha que es una cavidad de resonancia.

De aquí el sonido pasa al canal auditivo que es solo un tubo de aproximadamente 3 cm. de profundidad por 1 de diámetro. Su función es proteger los demás componentes del oído y mantenerlos a una temperatura mas o menos constantes. Este canal esta cubierto de una cera que ahuyenta a algunos insectos.

Además este canal junto con la concha ayuda a amplificar algunas frecuencias por medio de la resonancia. Al entrar las señales cercanas a la frecuencia de resonancia son reflejas en la parte interna y así ayudan a amplificar las señales de la misma frecuencia que entran. Hay una ganancia de aproximadamente 15-20 dB en los 2.5 kHz y de unos 10-17dB en los 5.5 kHz.

El ultimo punto de este canal es una membrana llamada tímpano que es una estructura cónica que se encuentra al final del canal o meatus.

2.2.2. El oído medio

Cuando las ondas sonoras llegan al tímpano hacen que este vibre, transmitiendo esta vibración a los demás componentes del oído medio del otro lado de la membrana.

EL oído medio es una cavidad llena de aire, dentro de la cual se encuentran el martillo (maellus), yunque (incus) y estribo (stapes) , que son los tres huesecillos (osciculos) mas pequeños del cuerpo humano. Estos están unidos entre si de forma articulada.

Además se encuentra aquí la trompa de Eustaquio el cual es un canal que comunica al oído medio con el exterior por medio de las vías respiratorias para igualar la presión a ambos lados del tímpano.

El tímpano esta unido al martillo que a su vez esta unido al yunque, estribo y ventana oval. De esta forma la vibración del tímpano es enviada a la ventana oval que es la entrada al oído interno.

La función básica de esta cadena de huesos es el acoplamiento de impedancias entre el aire y el fluido del oído interno. Esto es debido a que si las vibraciones se transmitieran directamente al liquido del oído interno debido a la diferencia de densidades se perdería casi en su totalidad (solo se transmitiría entre del 3 al 8% de la energía sonora).

Esto se logra por varios medios:

El área del tímpano es de aproximadamente 0.6 cm^2 mientras que el área de la ventana es de 0.032cm^2. La diferencia de áreas es de aproximadamente 17:1 con lo que se obtiene una amplificación de las fuerzas que llegan al tímpano.

La forma cónica del tímpano disminuye el movimiento de los huesecillos disminuyendo la velocidad y aumentando la energía.

El brazo del yunque es menor que el brazo del martillo, lo que genera un efecto de palanca que multiplica la vibración por un factor de aproximadamente 1.3

Todos estos efectos aumentan considerablemente la vibración que llega al oído interno (los cálculos mas conservadores indican un factor de 20, aunque algunos autores calculan factores de hasta 100). Este acoplamiento es mas efectivo en las cercanías de 1kHz.

Además aquí se localizan los músculos tensores del oído medio. El tensor tympany que esta unido al martillo en las cercanías del tímpano y que es activado por el nervio trigenial (quinto par) y el músculo stapedius que esta conectado al estribo y es activado por el nervio facial (séptimo par).

Estos músculos se contraen a intensidades muy grandes (aprox. 90 dB) con lo que se atenúa la vibración de los huesecillos principalmente en las frecuencias abajo de 1kHz. Esto es conocido como reflejo acústico y que sirven así para limitar los daños al oído por sonidos muy intensos. Sin embargo este reflejo tarda de 40 a 160 ms en activarse.

Además en frecuencias mayores a 2kHz las características de los mismos limitan la transmisión por lo que se genera un efecto de pasobandas.

2.2.3. El oído interno

El oído interno consta de el aparato vestibular (que es el encargado de mantener el equilibrio), la ventana oval, la ventana redonda y la coclea.

La coclea es una estructura rígida en forma espiral (2 ¾ vueltas) de aproximadamente 35 mm. de longitud por 2 mm. de diámetro, el cual se va adelgazando conforme se acerca a la punta. Esta está dividida en de forma longitudinal en 3 partes por medio de dos membranas, la membrana basilar y la membrana de Reissner. De esta forma la coclea queda dividida en tres compartimentos denominados escalas: vestibular, media y timpánica.

La escala superior (vestibular) e inferior (timpánica) además se encuentran unidas por medio de una abertura en el extremo de la coclea llamada helicotrema. Ambas escalas además se encuentran llenas de un liquido llamado perilinfa, mientras que la escala media contiene endolinfa.

La señal acústica llega a la coclea por medio del estribo que se encuentra en contacto con la ventana oval de donde se transmite a la perilinfa. Debido a que este es un liquido incompresible por lo que en la ventana redonda se encuentra otra membrana flexible para compensar los movimientos del liquido.

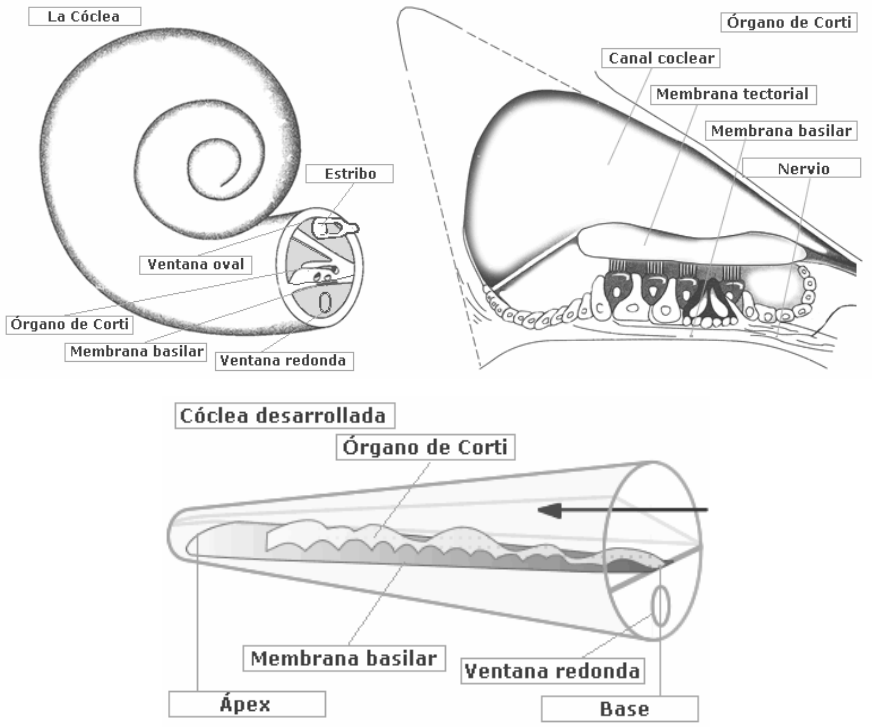

Fig. 2.5- La coclea

2.3. La voz en la transmisión de información

El conjunto de sonidos generados por medio del sistema generador de voz pueden utilizarse para la transmisión de mensajes por medio del lenguaje hablado. Normalmente en el estudio de este se divide en diferentes niveles de acuerdo al punto de vista desde el cual se esta estudiando.

2.3.1. Nivel acústico:

Este nivel solo se preocupa por la representación física del sonido como variaciones de presión en el aire y sus componentes físicos como frecuencia fundamental, amplitud, etc.

Desde este punto de vista la voz es un conjunto de sonidos y silencios alternados. Los sonidos pueden ser periódicos (sonoros) o estar formados principalmente por ruido de naturaleza estocástica (sordos)

2.3.2. Nivel Fonético:

El sonido es producido por medio de el aire producido en los pulmones y la modulación que le producen diferentes elementos (cuerdas vocales, lengua, etc.). Según la forma en que interactúen estos elementos se producen diferentes sonidos que se pueden dividir en grupos de acuerdo a la forma en que se articulan:

- Vocales: En estas el tracto vocal se encuentra abierto y libre de obstáculos, por lo que la única función de la boca es una variación en el timbre. En la lengua española existen 5 vocales principales (a,e,i,o,u), aunque según la región o su posición en la palabra se pueden producir diferencias en su pronunciación (estos son conocidos como alófonos). Las vocales a su vez se pueden subdividir en otras categorías según:
    - El modo de articulación
        - Abertura Máxima (a)
        - Abertura Media (e,o)
        - Abertura Mínima (i,u)
    - El punto de articulación:
        - En la parte anterior de la boca o palatales (i,e)
        - En la parte del centro o centrales (a)
        - En la parte posterior o velares (o,u)

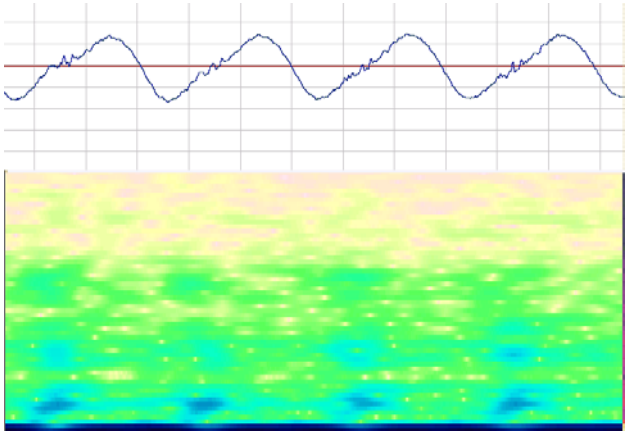

Fig.2.6 Onda y espectro del fonema 'i'. Nótese como la onda es cíclica.

- Diptongos: Es un monosílabo que empieza en la posición de una vocal y cambia hacia la posición de otra vocal diferente. La vocal de mayor abertura es el núcleo silábico y la de menor abertura es la silaba marginal. Existen dos tipos de diptongos:
    - Crecientes: el núcleo silábico precede al margen silábico, el margen se conoce como semiconsonante. Existen dos semiconsonantes conocidas como [j] (hac**ia**,rad**io**) y [w] (ag**ua**,antig**uo**).
    - Decrecientes: el núcleo antecede al margen. En estos el margen silábico se conoce como semivocal que son de mayor duración y con una articulación mayor que en las semiconsonantes. Las semivocales son [i] y [u]

- Consonantes: Las consonantes se clasifican de acuerdo a su manera de articulación las cuales ya se mencionaron:
    - Plosivas: El tracto vocal cerrado en el punto de articulación, el pasaje también cerrado. Se produce una exhalación cortante con característica de respuesta transitoria. En el español mexicano existen 6: /p/,/t/,/b/,/d/,/g/,/k/.

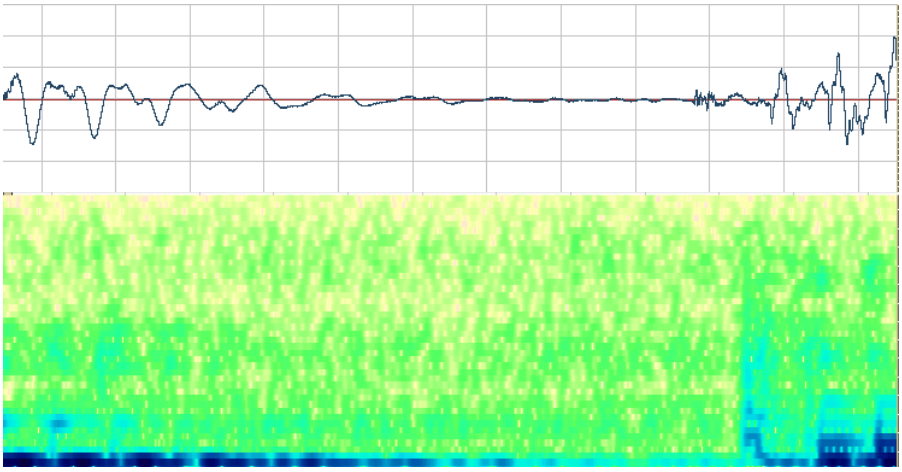

Fig.2.7 : Fonema 't' dentro de una palabra. Antes de un fonema plosivo existe una pequeña pausa en lo que se obtiene la presión de aire suficiente para efectuar la exhalación.

    - Fricativas: El tracto vocal esta abierto parcialmente con el velo cerrado. Se genera ruido en el punto de articulación. Existen 3: /s/,/f/,/j/,/x/. también existe /θ/ que corresponde a la letra 'z' pero este fonema es prácticamente inexistente en México.

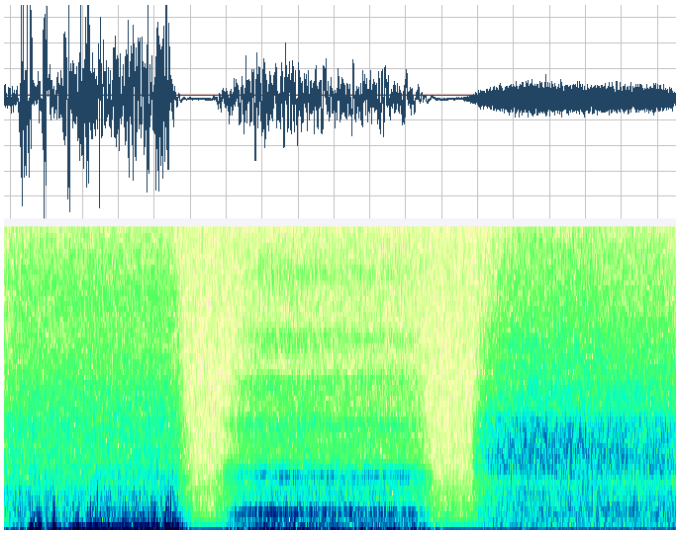

fig. 2.8: fonemas 'f','j' y 's'. Los fonemas fricativos están compuestos principalmente por ruido

- Aficativas: Existe un cierre inicial del tracto vocal seguido de una expiración gradual que produce turbulencia, en español solo existe uno: /c/ que corresponde al símbolo 'ch'.

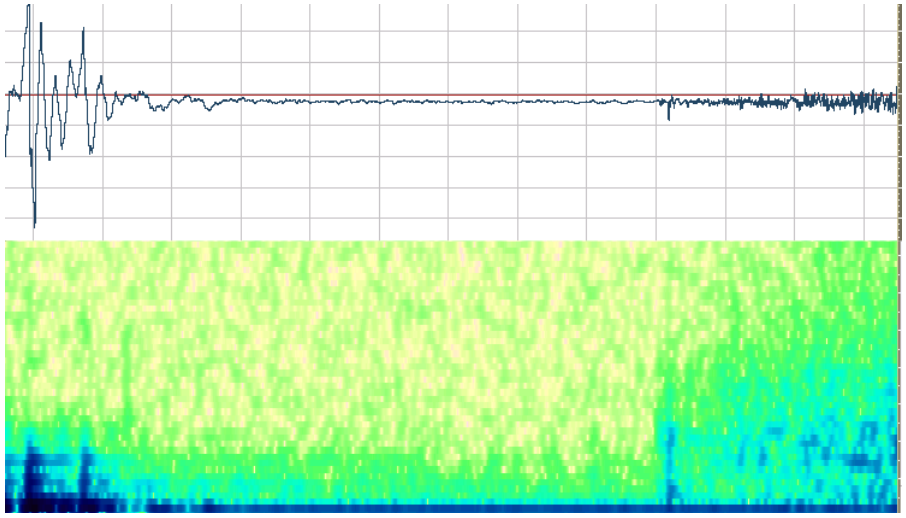

Fig. 2.9. Fonema 'ch'. Nótese la pausa que existe al inicio del fonema al igual que en las plosivas.

- Semivocales: El tracto vocal esta parcialmente abierto en el punto de articulación sin turbulencia. Este se divide en dos tipos:
  - Vibrantes: /r/, /ř/ que corresponde a 'rr'
  - Laterales: /l/,/ʎ/ que corresponde a 'll'

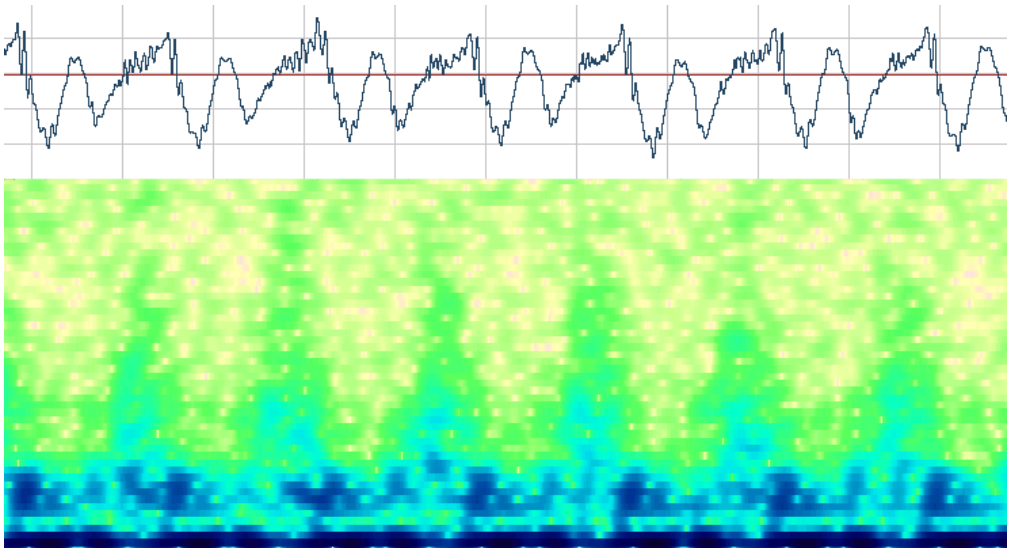

FIg 2.10. Fonema 'll'. Nótese que las semivocales son periódicas al igual que las vocales.

- Nasales: El tracto vocal esta cerrado y el velo abierto: /m/,/n/,/ñ/

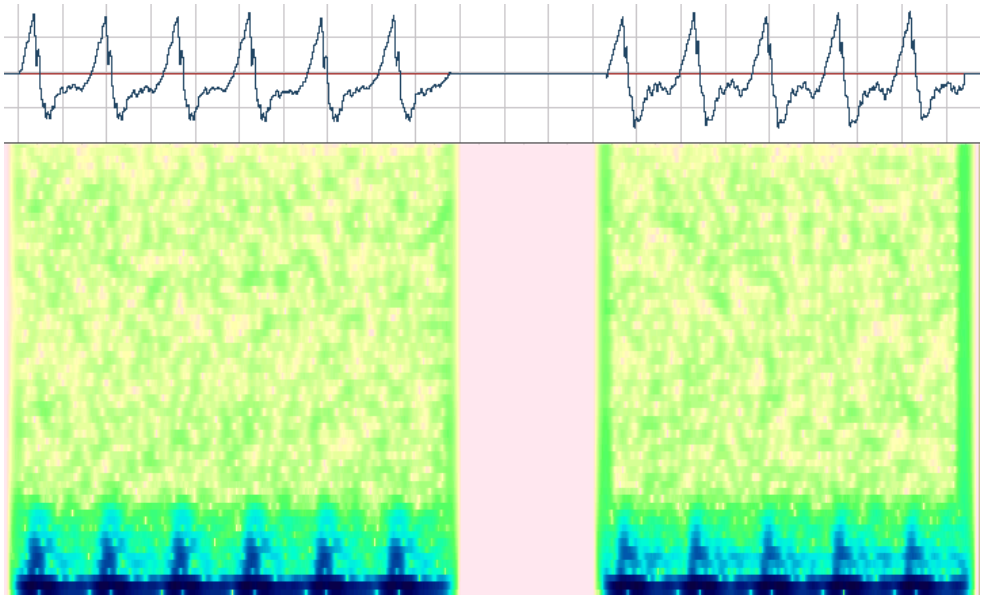

Fig 2.11. Fonemas 'm' y 'n'

También pueden clasificarse por el punto en donde se articulan. Sin embargo este punto es aproximado ya que debido a los fonemas adyacentes puede no alcanzarse perfectamente el punto de articulación y sufrir variaciones o ser variable con el tiempo:

- Bilabiales: Se pronuncian con los labios : /p/ y /m/

- Labiodentales: La punta de la lengua hace contacto con la parte posterior del diente incisivo superior: /f/,/b/
- Interdentales: La lengua se sitúa entre los dientes: /d/,/t/,/θ/
- Alveolares: La punta de la lengua se acerca o toca la punta alveolar en el techo de la boca: /s/,/l/,/r/,/ř/,/n/
- Palatares: La lengua se apoya en el paladar: /c/,/ʎ/,/ñ/,/j/
- Velares: La lengua toca el velo del paladar: /g/,/k/,/x/

Por la acción de las cuerdas vocales:

- Sonoras: Se producen vibraciones en las cuerdas vocales: /b/, /d/, /g/, /j/, /f/, /θ/, /l/, /ʎ/, /r/, /ř/
- Sordas: En estas consonantes no existe vibración en las cuerdas vocales: /k/, /p/, /t/, /s/, /x/, /c/, /m/, /n/, /ñ/

	BILABIALES		LABIO-DENTALES	INTER-DENTALES	DENTALES		ALVEO-LARES		PALATALES		VELARES	
	Sonoros	Sordos	Sordos	Sordos	Sonoros	Sordos	Sonoros	Sordos	Sonoros	Sordos	Sonoros	Sordos
Oclusivos	/b/	/p/			/d/	/t/					/g/	/k/
Fricativos			/f/	/Ø/				/s/	/y/			/x/
Africados										/č/		
Laterales							/l/		/ľ/			
Vibrantes							/r/ /r̄/					
Nasales	/m/						/n/		/ň/			

Fig. 2.12. : Fonemas consonantes según su clasificacion. (10)

2.3.3. Nivel Fonológico

El nivel fonético se estudian los fonos que son los diferentes los diferentes sonidos que conforman un lenguaje desde el punto de vista físico. En el nivel fonológico estos mismos sonidos se estudian desde el punto de vista lingüístico, es decir, desde el punto de vista de su significado. Estas unidades lógicas se llaman fonemas y son las unidades mínimas de un lenguaje. Un fonema se define como la unidad mínima que al cambia el significado de una palabra al ser sustituida por otra.

También se debe notar que la relación entre fonos y fonemas no es uno a uno, debido a que un fonema se puede pronunciar de diferentes maneras (alófonos) debido a modificaciones introducidas por los fonemas circundantes (debido a las limitaciones dinámicas del tracto vocal) o por variaciones regionales.

Debido a las limitaciones físicas del tracto vocal existen 3 efectos básicos:

- Coarticulación: se debe a que en el habla fluida se necesita cambiar constantemente de un sonido al siguiente por lo que unos sonidos se ven afectados por los otros.
- Asimilación: este problema al igual que el siguiente se agravan al aumentar la velocidad con que se habla. En este un fonema es eliminado al no poder alcanzar el punto central del mismo
- Undershoot: Aquí aunque el fonema no desaparece no alcanza su punto central

fonema /a/	se representa siempre con la grafía "a"
fonema /b/	se representa con tres grafías diferentes: "b", "v", "w" ejemplos: *bobina*, *vivir*, *wolframio*
fonema /ch/	se representa siempre con el dígrafo "ch"
fonema /d/	se representa siempre con la grafía "d"
fonema /e/	se representa siempre con la grafía "e"

fonema /f/	se representa siempre con la grafía "f"
fonema /g/	se representa a veces con la grafía "g" ejemplo: *gárgola* y otras veces con el dígrafo "gu" ejemplo: *guerra*
fonema /i/	se representa a veces con la grafía "i" ejemplo: *ilícito* y otras veces con la grafía "y" ejemplos: *y*, *rey*, *buey*.
fonema /x/	se representa a veces con las grafías "j" o "g". otras veces con la grafía "x", aunque en estos casos es licito usar también "j", ejemplo: *mexicano* o *mejicano*, *Texas* o *Tejas*.
fonema /k/	se representa a veces con las grafía "c", “qu” o “k”
fonema /l/	se representa casi siempre con "l", ejemplo: *libélula*. Aunque el fonema /ll/ al final de palabra ha sido sustituido por /l/ en casi todos los ámbitos de la lengua. Ejemplo de "ll" leída como /l/ a final de palabra: *Sabadell*.
fonema /m/	se representa siempre con la grafía "m" ejemplo: *mamífero*.
fonema /n/	se representa con la grafía "n".
fonema /ñ/	se representa siempre con la grafía "ñ",.
fonema /o/	se representa siempre con la grafía "o".
fonema /p/	se representa siempre con la grafía "p"
fonema /r/	se representa siempre con la grafía "r"
fonema /rr/	se representa a veces con la grafía "rr" ejemplo: *arroba* y otras veces con la grafía "r" ejemplo: *rosa*.
fonema /s/	el fonema /s/ se representa con la grafía "s",
fonema /ɵ/	se representa con la grafía "z” y "c": *césar*, *cero*. Sin embargo en Mexico este fonema no se utiliza y es reemplazado siempre por /s/

fonema /t/	se representa siempre con la grafía "t"
fonema /u/	se representa casi siempre con las grafías "u", y "ü". En algunas palabras tomadas del inglés, se representa con "w"
fonema /y/	Se representa con las grafías "y" y "ll".

Tabla 1 :Fonemas del español y las grafias con que se representan.(10)

grafía "a"	representa siempre el fonema /a/
grafía "b"	representa siempre el fonema /b/,
grafía "c"	representa el fonema /k/ en "ca", "co", "cu y /s/ en "ce", "ci".
dígrafo "ch"	representa el fonema /ch/
grafía "d"	representa siempre el fonema /d/
grafía "e"	representa siempre el fonema /e/
grafía "f"	representa siempre el fonema /f/
grafía "g"	representa a veces el fonema /g/: "ga", "go", "gu", "gü" representa a veces el fonema /x/: "ge", "gi"
dígrafo "gu"	representa el fonema /g/ en "gue", "gui"
grafía "h"	Si no forma parte del dígrafo "ch", la h es estrictamente muda y por tanto no representa ningún fonema.
grafía "i"	representa siempre el fonema /i/

grafía "j"	representa siempre el fonema /x/
grafía "k"	representa siempre el fonema /k/
grafía "l"	cuando no forma parte del dígrafo "ll", representa siempre el fonema /l/
dígrafo "ll"	el dígrafo "ll" representa el fonema /y/ (salvo a fin de palabra)
grafía "m"	siempre representa el fonema /m/
grafía "n"	representa siempre el fonema /n/
grafía "ñ"	representa siempre el fonema /ñ/
grafía "o"	representa siempre el fonema /o/
grafía "p"	siempre representa el fonema /p/
grafía "q"	se usa aislada en contados casos para representar el fonema /k/: *Qatar*, *Iraq*.
dígrafo "qu"	representa el fonema /k/ en "qui", "que".
grafía "r"	representa a veces el fonema /rr/: "rosa" y a veces /r/: *mormón*
dígrafo "rr"	representa siempre el fonema /rr/
grafía "s"	representa siempre el fonema /s/
grafía "t"	representa siempre el fonema /t/
grafía "u"	representa siempre el fonema /u/
grafía "ü"	representa el fonema /u/ en "güe", "güi", como *güisqui*, *agüero*.

grafía "v"	representa siempre el fonema /b/
grafía "w"	representa a veces el fonema /b/: *wolframio* y a veces el fonema /u/ o el grupo /gu/ en palabras tomadas del inglés.
grafía "x"	representa a veces el grupo /ks/: "examen", *taxonómico* representa a veces el fonema /s/: *excepción* representa a veces el fonema /x/: *México*, *texano*. En estos casos se puede sustituir por "j".
grafía "y"	representa a veces el fonema /y/: *cayo* y a veces el fonema /i/: *y*, *rey*
grafía "z"	según la norma, representa el fonema /ɵ/ . Sin embargo en Mexico se sustituye por /s/

Tabla 2. : Grafias del español y su correspondencia fonémica. (10)

3.SINTETIZADORES DE VOZ

3.1. **Sintetizadores de Voz**

Un sintetizador de voz es un aparato o software que genera una salida sonora que simula o imita a la voz humana. Aunque desde el siglo XVIII existen aparatos mecánicos que trataban de simular la voz humana mediante medios mecánicos, solo en la ultima mitad del siglo XX se volvió posible el hacer sistemas que generaran automáticamente una salida de voz a partir de texto escrito (sistemas de texto a voz).

Estos sistemas constan principalmente de dos bloques. El primer bloque toma una entrada de texto y a partir de esta genera una trascripción fonética y opcionalmente información sobre duración, entonación, etc. Esta trascripción fonética se alimenta al segundo bloque que es el que genera la salida final de voz.

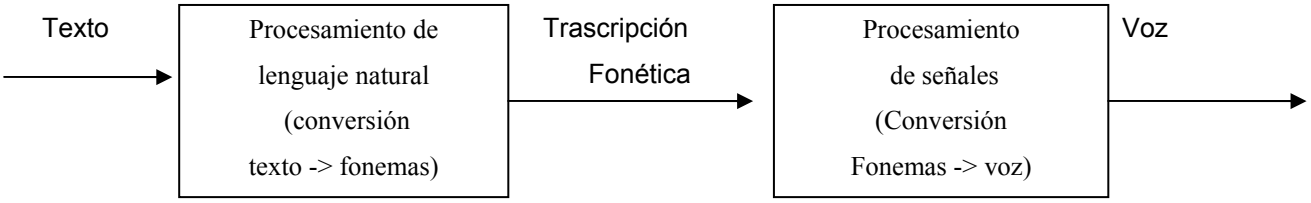

Fig.3.1 : Estructura básica de un sistema de síntesis de voz.

3.2. **Historia de los sintetizadores de Voz**

3.2.1. Inicio de los sintetizadores de voz: Sintetizadores mecánicos

Desde la segunda mitad del siglo XVIII se han construido diversos sistemas, primero mecánicos, luego electrónicos y actualmente digitales que sean capaces de generar una salida de voz de forma artificial.

El primer intento registrado fue realizado en 1773 donde Ch. G. Kratzenstein, un profesor de fisiología de Copenhague logro producir sonidos vocálicos a partir de tubos de resonancia conectados a tubos de órgano.

Simultáneamente a este invento Wolfgang von Kempelen (Hungaria 1734 – Viena 1804) trabajaba en la construcción de una maquina que pudiera generar sonidos que simularan a la voz humana. En 1791 publico sus descubrimientos y la forma en que se podía construir su maquina de forma que otras personas pudieran continuar su investigación. Esta máquina funcionaba por medio de un fuelle que generaba una salida de aire. Este después era enviada a través de varias tuberías y aberturas que podían ser modificadas manualmente para producir palabras o frases cortas.

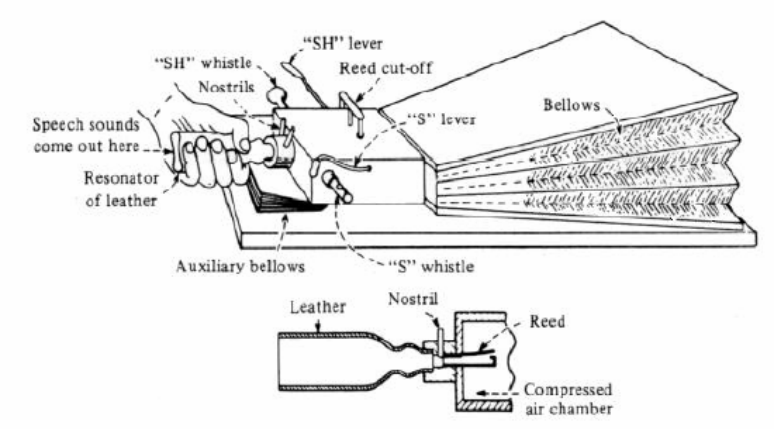

Fig3.2.: Aparato de Von Kempelen (6)

Durante el siglo XIX se construyeron otros aparatos similares que mejoraron levemente el modelo de Kempelen al introducir elementos adicionales que simularan la lengua o los cambios en la forma de la boca que se realizan al hablar.

3.2.2. Sintetizadotes de voz eléctricos

En 1922 Stewart construyo el primer sistema que intentaba simular la voz mecánica por medios eléctricos. Este dispositivo constaba de dos circuitos resonantes activados por un buzzer. Esto permitía aproximarse a vocales estáticas al ajustar dos de las frecuencias fundamentales (resonantes) de cada una de las vocales.

En la década de los 30's se diseño el VOCODER en los laboratorios Bell. Este aparato estaba diseñado para analizar la voz humana y obtener de esta parámetros acústicos. Después a partir de estos parámetros se podía reconstruir una salida similar a la onda de voz original.

A partir de este sistema se construyo una segunda versión que fue mostrada en la feria mundial de Nueva Cork (1939). Este sistema llamado VODER y desarrollado por Homer Dudley generaba una salida de ruido o una salida de audio senoidal de acuerdo a un selector. La frecuencia de esta salida podía ser controlada por un pedal. Esta salida era después filtrada por 10 filtros pasobanda cuya amplitud se modificaba por medio de los dedos. Aunque la inteligibilidad de este sistema es mínima fue el primer sistema en mostrar la posibilidad de generar voz por medios eléctricos. Este sistema es la base de la síntesis por formantes. El sistema de filtros fijos que utiliza este sistema es demasiado limitado para generar las diferentes salidas requeridas, por lo que en los sistemas mas modernos este sistema no es utilizado.

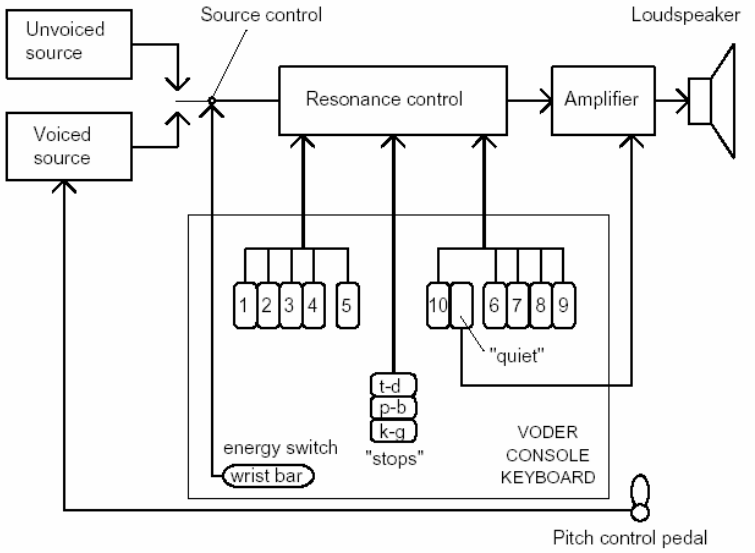

Fig.3.3 : Diagrama de el sistema VODER (6)

En 1950 se presento el reproductor de patrones desarrollado en los laboratorios Haskins. Este utilizaba espectrogramas los cuales se iluminaban y eran enviados a un conjunto de celdas fotovoltaicas, cada una de las cuales controlaba la intensidad de una onda fundamental de diferentes frecuencias en saltos

de 120Hz que podían reconstruir aproximadamente la señal del espectrograma. Franklin Cooper, Alvin Liberman, Pierre Delattre y otros asistentes experimentaron con espectrogramas reales y con adaptaciones dibujadas a mano para experimentar con la importancia de diferentes factores. Este sistema genero una inteligibilidad mucho mas alta (de mas del 90% con espectrogramas reales y de poco mas del 80% con espectrogramas estilizados)

Estos dos sistemas funcionaban copiando los patrones espectrales de la voz. Poco después de estos sistemas se le dio un nuevo enfoque a la teoría de síntesis de voz. Este nuevo enfoque fue la generación de una teoría acústica de la forma en que se produce la voz y no solo en los resultados del proceso. Esta teoría es la base de la síntesis por formantes que se vera mas adelante. Según esta teoría la voz se puede considerar como la salida de un filtro lineal excitado por una o mas fuentes, principalmente las cuerdas vocales y por ruido turbulento debido a diferencia de presiones a través de un estrangulamiento. El filtro en este caso es una simulación de los efectos del conducto acústico (faringe, cavidad oral y labios). Este tracto vocal es simulado por una función de transferencia con pares de polos complejos conjugados que producen picos en el espectro de salida (llamados formantes). Además de los polos se requiere la introducción de ceros (antiresonadores) para modelar la absorción de las ramas laterales en algunas articulaciones como nasales y fricativas.

En 1953 se crearon los primeros sintetizador de formantes (Paramteric Artificial Talker (PAT) construido por Walter Lawrence y el orador Verbis Electris(OVE I) que construyo Gunnar Fant). Estos sistemas fueron enfrentados en una conversación en 1956 en el MIT.

El PAT tenia tres resonadores en paralelo. Se tenia una señal de entrada de ruido o periódica y de ahí a través de patrones dibujados sobre un vidrio que se deslizaba se controlaban las 3 frecuencias del formante, amplitud del ruido, amplitud de fraseo y fundamental.

El OVE utilizaba filtros en cascada en lugar de en paralelo. Los dos mas bajos eran controlados por movimientos en 2 dimensiones de un brazo mecánico, mientras que la amplitud y la fundamental eran controlados por potenciómetros. Sin embargo, este sistema solo podía generar vocales.

Es interesante notar que aunque ambos sistemas parten de la misma teoría y usan los mismos principios utilizaron diferentes métodos para llevarla a la practica y hasta la fecha aun existe controversia sobre cual de los dos métodos es mejor o si la mejor opción es utilizar una mezcla de ambos, teoría propuesta en 1972 por Klatt.

Estos sistemas sufrieron varias modificaciones. A PAT se le introdujeron controles individuales para los formantes y un circuito independiente para fricativas, llegando a ser un sistema en cascada. A OVE I se le agrego otra rama estática para simular murmullos nasales y una cascada de dos formantes y un antiformante para simular mejor la función de transferencia del tracto vocal y la excitación de los sonidos fricativos y se transformo en OVE II. Estos sistemas mejorados fueron enfrentados una vez mas en 1962 en una conferencia en Estocolmo.

Además con mayores o menores modificaciones y mejoras (modular la amplitud del ruido en fricativas sonoras o agregado nuevos parámetros)estos dos sistemas continúan siendo la basa de los sistemas modernos de síntesis por formantes.
Una de las mas grandes modificaciones a esta clase de sistemas fue la introducción de sistemas híbridos (cascada y paralelos). En el sistema propuesto por Klatt cada sistema se utilizaba para modelar diferentes tipos de sonidos (en paralelo para sonidos sonoros y en cascada para sonidos sordos). Este sistema propuesto por Klatt fue además presentado como un listado en Fortran en 1980 lo que permitió su uso mas extendido.

Otro punto importante en la historia de los sintetizadores por formantes ocurrió en una conferencia en Boston en 1972 cuando John Holmes presento una salida de voz generada que era prácticamente indistinguible de una voz natural. Desafortunadamente esta señal fue generada de forma manual y basada en un proceso de prueba y error de varios meses de duración. Aunque este experimento demostró varios factores importantes en la generación de oraciones y otros factores que podían ser despreciados su método no ha podido ser llevado a métodos automáticos.

El siguiente avance importante en este tipo de sintetizadores es cambiar el tipo de señal de entrada de una señal monótona (triangular, tren de pulsos) en una señal que mas se asemeje a la señal que entra al tracto vocal.

El primer avance de este tipo se dio en 1975 (Rothenberg) donde se utilizo un sistema de tres parámetros de acuerdo a la apertura de la glotis, la respiración, etc. Se han creado métodos mas avanzados que simulen mas parámetros, pero hasta la fecha el resultado aun no es completamente natural, debido posiblemente a la falta de conocimiento del modelo real.

Aparte de este tipo de sintetizadores, otra línea paralela de investigación es la generación de líneas de transmisión que simulen un tubo similar al tracto vocal. Sin embargo, debido a restricciones en el conocimiento del tracto vocal y en cantidad de cálculos, se ha avanzado poco en esta arrea, aunque existen algunos modelos de sintetizadores de este tipo.

Una vez que se obtuvieron sistemas que pudieran simular la voz humana, una aplicación muy importante que solo se hizo posible con el advenimiento de computadoras y circuitos integrados es la generación de una señal de voz a partir de una entrada fonemica o de texto.

El primer programa de este tipo se desarrollo en 1961 (Kelly y Gerstman) con un sintetizador en cascada de tres formantes, cuyos parámetros posteriormente se modificaban a mano.

En 1964 apareció otro sistema (Holmes) que funcionaba a partir de síntesis por formantes en paralelo y un conjunto de tablas que permitía generar resultados mas complejos como coarticulación. En 1966 (Mattingly) modifico el programa para dar transiciones mas realistas, pero con poca mejora perceptiva y el uso de alófonos.

El primer uso practico que se le intento dar a este sistema, fue un adaptación como parte de una maquina de lectura para ciegos, pero nunca se concretizo por falta de recursos.

A finales de los 60’s y principio de los 70’s se continuo la investigación de los sistemas de síntesis por regla ajustando diferentes parámetros para hacerlos mas similares a la voz natural, principalmente por Klatt, dando como resultado el sistema de síntesis del M.I.T. MITalk (1976), el cual fue vendido y

cambio de manos varias veces durante los siguientes años. Después de este sistema se desarrollo el Klattalk que continuo siendo mejorado hasta finales de los 80's.

Otro importante desarrollo fue el Votrax SC-01 (1976) que fue el primer sistema de síntesis por formante en estar incorporado dentro de un circuito integrado que fue integrado dentro de sistemas de síntesis de bajo costo y el TMS-5520 de Texas Instruments que es la base del Echo , un circuito de síntesis por concatenación de segmentos pregrabados.

Otra línea de investigación es tomar segmentos de voz pregrabados como bloques para construir una frase cualquiera. Debido a las características de la voz no se pueden usar palabras o silabas (son demasiadas) ni fonemas (aunque son pocos no toman en consideración los efectos de coarticulación ni la transición entre fonemas). Debido a esto en 1958 (Peterson) propuso una unidad denominada di fonema que corresponde al segmento entre el centro de un fonema al centro del siguiente, debido a que así si se toma en cuenta la transición y los efectos de la coarticulación son pocos en el centro de un fonema. En teoría se requieren solo el cuadrado del numero de fonemas de una lengua de difonemas, pero hay combinaciones que no pueden existir con lo que se reduce el numero, pero se pueden agregar algunos difonemas para grabar diferencias entre silabas acentuadas y no, alófonos, etc. Peterson estimo que se requieren unos 8000 difonemas en ingles, aunque el numero normal en un sistema es mas cercano a 1000.

En 1961 (Sivertsen) propuso mezclar difonemas y unidades mas largas llamadas diadas, que contienen la mitad final de un fonema, un fonema completo y la mitad inicial del siguiente (VCV) para conservar algunos fenómenos que pueden no estar previstos en un di fonema.

Aunque tienen algunos problemas como discontinuidades en algunas uniones, estos sistemas son muy utilizados debido a su relativa sencillez y alta inteligibilidad. El primer sistema de este tipo fue mostrado en 1967 en el MIT, pero este sistema se cancelo por falta de recursos.

En 1976 (Olive y Spickenagle) intentaron extraer las características de los fonemas para crear un sistema que generara un catalogo de difonemas de forma automatizada.

Este sistema de síntesis es el mas utilizado actualmente y la investigación actual es sobre métodos para mejorar la naturalidad de la voz, disminuyendo o eliminando discontinuidades y generando métodos de generación de contornos para mejorar a la entonación, así como en análisis sintácticos y semánticas que permitan mejorar los contornos de acuerdo al contenido de una oración.

3.3. Sintetizadores de voz en la actualidad

En la actualidad la mayoría de los sistemas de síntesis están basados en la unión de segmentos pregrabados, esto porque aunque los otros tipos de sistemas en pueden generar diferentes tipos de voces y en teoría pueden generar audio de gran calidad y generar muchos tipos de variantes son de muy alta complejidad por lo que aun no se ha podido generar sonido de alta calidad, en cambio los sistemas de concatenación tienen menos variabilidad, como por ejemplo solo pueden tener un tipo de voz y para tener un tipo de vos diferente se requiere grabar toda una nueva base de datos y tienen menor rango de flexibilidad, además de requerir un mucho mayor espacio de almacenamiento, pero al ser de una mayor sencillez pueden generar audio de mejor calidad.

Otro de los avances actuales es la mejora en los modelos prosódicos (entonación, ritmo) y la utilización de métodos empíricos (estocásticos) en vez de métodos lingüísticos. Esto permite que reglas que no están bien definidas de puedan inferir estadísticamente de l estudio de grabaciones en vez de tratar de generarlas a partir de reglas.

Estos métodos generan el ritmo y la variación de entonación y frecuencia fundamental en el tiempo a partir de análisis sintáctico y gramático e información estadística por medios varios (sistemas lineales, redes neuronales, sumas de productos, árboles sintácticos, modelos de Markov)

El problema mas grande de estos sistemas estadísticos en que requieren grandes cantidades de información y algunos modelos poco frecuentes pueden ser totalmente ignorados en el proceso o tipos de oraciones que varían de forma muy drástica en el proceso pueden quedar totalmente fuera de rango.

Para aliviar este problema se están generando también sistemas que sirvan para analizar grandes cantidades de audio pre-grabado de forma automática, a diferencia de tener que analizarlos a mano como se hace normalmente. Sin embargo, hay tipos de diálogos, principalmente coloquiales o regionales que no tienen parámetros fácilmente reconocibles o que tienen demasiadas variantes posibles.

Otro avance importante es la relación entre el contorno de la frecuencia fundamental y el ritmo, ya que son interdependientes.

Dados estos avances la calidad de la voz sintética ha mejorado considerablemente, pero aun quedan fuertes problemas por resolver, principalmente en la generación de mejores modelos prosódicos y una mayor variabilidad de parámetros y tipos de voz, posiblemente generando nuevos sistemas de síntesis por formantes, ya que las restricciones de velocidad de los sistemas son mucho menores que antes con la llegada de equipos de mucho mayor velocidad.

3.4. Tipos de sintetizadores de voz

Actualmente la mayoría de los sistemas de síntesis de voz son creados por métodos electrónicos (computadoras, circuitos integrados). Los métodos utilizados para estos sistemas pueden dividirse principalmente en 3 grupos:

1- Articulatorios: Tratan de modelar directamente el sistema generador de voz directamente.
2- Por Formantes: Modelan la función de transferencia o de polos de frecuencias del tracto vocal.
3- Por Concatenación: Utilizan segmentos pregrabados que son unidos (concatenados)

Los dos tipos mas usados son por formantes y por concatenación. Aunque el primero fue mas usado inicialmente debido a limitaciones en la capacidad de almacenamiento, actualmente el segundo es mas usado debido a que son posibles mayores capacidades de almacenamiento.

3.4.1. Síntesis Articulatoria

Este tipo de síntesis trata de modelar los órganos vocales lo mas perfectamente posible, por lo que teóricamente es el sistema que podría generar síntesis de mas alta calidad, pero a la vez es el sistema mas complicado y de mas alta carga computacional.

Este tipo de síntesis involucra normalmente modelos de las cuerdas humanas, de la lengua (posición, altura, etc.), apertura del velo, presión de los pulmones, apertura glotal, etc.

El modelo de articulación se genera normalmente a partir de radiografías pero estas no proporcionan información suficiente para conocer todos los parámetros necesarios. La ventaja de este es que puede utilizar efectos que difícilmente se podrían llevar a cabo en otros sistemas.

Debido a las complejidades de análisis y la carga computacional requerida este tipo de síntesis ha recibido poca atención por lo que ha tenido muy poco desarrollo.

3.4.2. Síntesis por Formantes

Este es uno de los métodos de síntesis mas usados. Se basa en un modelo de entrada-filtro-salida del cual existen básicamente 2 tipos (filtros en cascada, filtros en paralelo) y combinaciones de ambos, lo cual produce un mejor resultado. Además este sistema proporciona mayor flexibilidad que la síntesis por concatenación y una menor dificultad que la síntesis articulatoria.

Normalmente se usan al menos 3 formantes para producir la señal de voz, aunque a veces se usan hasta 5 para mejorar la calidad. Cada formante se modela por medio de resonador basado en un filtro centrado en la frecuencia del formante y modelado con un par de polos, con lo que se puede indicar el ancho de banda del filtro.

La síntesis por formantes utiliza cierto conjunto de reglas que determinan los parámetros necesarios para cada sonido. Algunos de estos parámetros pueden ser: Frecuencia fundamental (F0), Grado de excitación (VO), Frecuencias y amplitudes formantes (F1,F2,F3,A1,A2,A3), etc.

Los resonadores de cascada funcionan mejor con los sonidos no nasales pero tienen problemas con las fricativas y plosivas. Los resonadores tienen problemas con las vocales pero funcionan bien para nasales, fricativas y plosivas. Debido a esto el modelo mixto ideado por Klatt (1980) que consiste en un sistema mixto con 6 formantes, filtros extras, la adición de un ruido de alta frecuencia y con un sistema de excitación compleja es el mas utilizado en los sistemas comerciales actuales como el MITalk, DECtalk y Prose-2000.

Otro sistema mixto es el PARCAS introducido por Laine en 1982 que es modelado por pares de ecuaciones de transferencia parciales y un conjunto de constantes para mantener las amplitudes balanceadas en diferentes salidas.

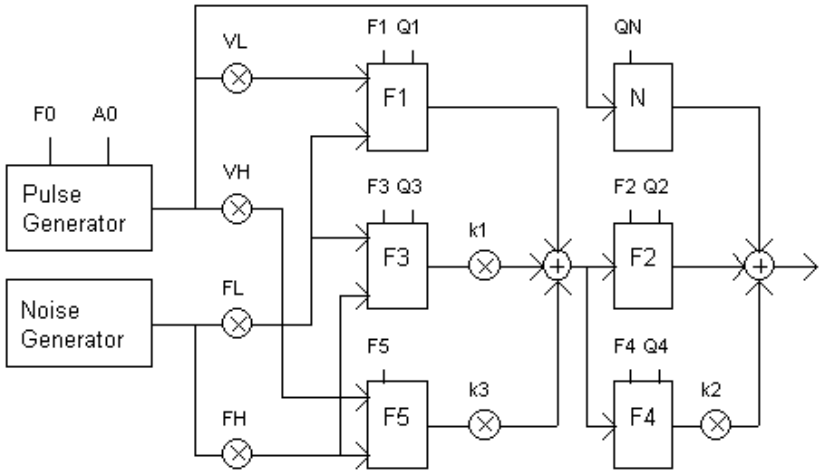

Fig 3.4: Diagrama de PARCAS (Laine 1982). Aquí se ve un sistema mixto con filtros en paralelo y cascada. también se observa que se usa una entrada periódica o una entrada ruido dependiendo del fonema a sintetizar. (10)

Una vez que se tiene el conjunto de resonadores a utilizar, estos se alimentan con un tren de pulsos con una frecuencia igual a la de la fundamental (Fo) para sonidos con fundamental (p. ej. Vocales) o con una fuente de ruido para sonidos sin fundamental (p. ej. /s/).

En algunos sistemas modernos se introducen señales diferentes a un tren de pulsos para mejorar la salida.

3.4.3. Síntesis por concatenación.

Este es el tipo de sintetizador que se va a desarrollar. En este tipo de sistema se basa en conectar segmentos de voz previamente grabados y almacenados. Es el sistema que presenta menor complejidad aunque tiene la limitación de que solo se pude usar un tipo de voz.

Uno de los aspectos mas importantes de este tipo de síntesis es el de seleccionar un tipo y tamaño de segmento de acuerdo al tipo de sistema que se quiere desarrollar. En unidades largas se requieren menores puntos de concatenación por lo que se obtiene un mejor control de coarticulación pero se requiere un numero muy grande de unidades. En unidades pequeñas se tienen mas puntos de concatenación pero se reduce el numero de unidades requeridas.

La unidades mas grandes son frases y palabras. Estas funcionan bien para sistemas que tienen un vocabulario limitado, por ejemplo un sistema que de la hora del día. Pero para sistemas de entrada libre hay demasiadas unidades para ser prácticos, además de que al ser libre la entrada el usuario tiene la libertad de ,por ejemplo, insertar palabras inexistentes o frases mal construidas.

Otra unidad que se podría considerar son las silabas. Su numero es considerablemente menor al de palabras o frases pero aun tiende a ser muy grande (10,000+). Además en un sistema basado en silabas no se pueden almacenar los efectos de coarticulación entre silabas ni la prosodia. El uso de silabas como unidad es solo factible en lenguajes silábicos, como por ejemplo el japonés donde solo existen menos de 100 silabas.

Debido a estos problemas no existe al momento ningún sistema de conversión texto-habla que utiliza estas unidades.

La siguiente unidad que se puede considerar son los fonemas. Su numero es bastante reducido (normalmente entre 25 y 50 según el idioma). Los fonemas presentan el problema de la falta de información de coarticulación por lo que son poco usados, aunque en muchos sistemas se utilizan los fonemas como unidades lógicas que son transformadas a la unidad correspondiente después de su análisis en el proceso de concatenación.

Otra unidad son las Demisílabas que representan la parte inicial y final de las silabas. Su numero es grande pero aceptable (aprox. 1000). Estas cubren un buen numero de problemas como la coarticulación, algunos alófonos y requieren menos puntos de concatenación que los fonemas. Su numero es grande pero aun es aceptable, desafortunadamente su numero no puede ser determinado fácilmente y hay algunas combinaciones que no pueden ser generadas con demisílabas, por lo que normalmente se usan solo en sistemas mixtos.

La siguiente unidad a considerar son los difonemas. Un difonema se define como el segmento que inicia del punto central del estado estable de un fonema al punto central del estado estable del siguiente fonema. Esto ayuda a disminuir la distorsión al estar el punto de concatenación en una zona de relativa estabilidad. Esto además permite evitar los problemas de coarticulación al estar esta presente explícitamente en los segmentos. El numero existente de difonemas es el cuadrado de los fonemas existentes. De estos existen algunos que pueden ser eliminados al no presentarse en una lengua. Su numero esta cerca de los 1000 en la mayoría de los lenguajes. El numero es grande pero aun aceptable y debido a las ventajas que presenta es un tipo de unidades muy utilizado. Estas son las unidades que serán utilizadas en el sistema a desarrollar.

Un sistema difonemico que requiere mención es el sistema MBROLA al ser uno de los pocos sistemas de síntesis de voz de distribución libre y sin costo, además de ser de los pocos sistemas de síntesis que ofrecen síntesis de la lengua española (con voces española y mexicana).

Existen otros tipos de unidades como los trifonemas (contienen un fonema entero entre dos medios fonemas) son poco usados, aunque actualmente se encuentra en investigación métodos para optimizar los sistemas de síntesis usando unidades de diferentes longitudes, donde estas unidades podrían formar una parte importante.

Otro tipo de unidad que esta en investigación actualmente es el uso de microfonemas que son segmentos de fonemas en diferentes contextos. Estos segmentos (llamados prototipos) son después concatenados y en las uniones se hace una interpolación para disminuir las discontinuidades.

Otro problema básico en estos sistemas es la distorsión que se presenta en los puntos de unión. Esto se ha minimizado un poco con el uso de difonemas donde la unión se lleva a cabo en un punto mas favorable. Otro proceso para eliminar esta distorsión es el uso de filtros o de interpolación para suavizar la unión, este es otro punto de investigación actual para obtener un buen suavizado sin alterar demasiado la señal original.

Un tipo de interpolación es el PSOLA (Suma con Traslape Sincronizado al Tono). En esta se toman los diferentes segmentos y se suman con un cierto traslape, ajustando por medio de ventanas centradas a una distancia igual a la frecuencia fundamental estimada previamente con lo que además de mejorar los puntos de concatenación se puede modificar el tono al modificar la frecuencia e interpolar puntos intermedios o la duración al repetir ventanas. El problema básico de este método es en aquellos fonemas que no tienen frecuencia fundamental en los cuales se genera un pequeño ruido tonal debido al análisis ventaneado.

Este sistema de interpolación es utilizado en los dos sistemas de síntesis de voz gratuitos mas importantes, el Festival (desarrollado por Alan Black y Paul Taylor en la Universidad de Edinburgo) y el MBROLA (desarrollado en la Faculté Polytechnique de Mons, Bélgica). Ambos sistemas se encuentran actualmente en desarrollo, pero solo se encuentra disponible al publico en forma de código fuente el sistema Festival. Ambos sistemas proporcionan bases de datos para generar voz en lengua española en variantes Ibérica y Mexicana.

4.CONVERSIÓN TEXTO – FONEMAS

La primera parte de un sistema de síntesis de voz es obtener el texto y generar la secuencia de fonemas que se requieren para generar la salida de voz así como información acerca de el lugar en que cada una de las palabras se encuentra acentuada.

Este proceso en el idioma español es sencillo en comparación con otros idiomas como por ejemplo el ingles o el francés, ya que a diferencia de estos en el español existe una relación mucho mas directa entre los símbolos utilizados y el fonema que representan por lo que se requieren muchas menos reglas para obtener los fonemas a ser reproducidos a partir del texto.

Las reglas básicas utilizadas para obtener los fonemas a partir de los grafos son:

- La letra ‘c’ toma el sonido [s] si esta seguida de una vocal débil (e,i) o el sonido [k] si esta seguida de una vocal fuerte (a,o,u). Además existe el fonema [c] que se representa por los símbolos ‘ch’:
- La letra ‘s’ tiene el sonido [s] excepto si esta seguida por una ‘h’
- La letra ‘l’ tiene el sonido [l] excepto si forma ´ll´ en cuyo caso forma el sonido [λ]
- La letra ‘r’ tiene sonido [r] cuando esta en medio de una palabra, después de una vocal y la letra siguiente es diferente de ‘r’. En caso contrario tiene el sonido [ř]
- ‘qu’ tiene sonido de [k]. Si después de la ‘q’ no existe una ‘u’ se mantiene el sonido [k] para poder leer palabras mal escritas ya que este caso no existe en el español.
- La letra ‘y’ tiene sonido de [λ] a menos que sea la ultima letra de una palabra o este seguida por una consonante en cuyo caso tiene el sonido [i].
- La letra ‘g’ tiene sonido de [x] si esta seguida de ‘e’ o ‘i’ y [g] en otro caso. Si la letra siguiente es ‘u’ seguida de una vocal débil (‘e’ o ‘i’) esta no se pronuncia a menos que tenga un símbolo de diéresis (ü) en cuyo caso si se pronuncia.
- La letra ‘x’ en español tiene pronunciación ‘ks’

Para llevar a cabo este proceso se va tomando palabra por palabra (en este punto se considera palabra cualquier secuencia de letras que se encuentra entre dos símbolos diferentes a una letra). Además si se encuentran símbolos o números estos deben ser eliminados, reemplazados por pausas o reemplazados por palabras según se requiera.

```
/* separar por palabras */
ctemp="";
tipo=0;
for (i=0;i<=j-1;i++) {
l1=cadena[i];
if(i<(j-1)) l2=cadena[i+1];
if(i<(j-1)) l3=cadena[i+2];
if (( (l1>='a' & l1<='z') | (l1>='• & l1<='•) | l1=='• | l1=='•) &
(tipo==1 | tipo==0))
{ctemp=ctemp+l1;tipo=1;}
else if ( (l1>='1' & l1<='9') & (tipo==0 | tipo==2)  )
{ctemp=ctemp+l1;tipo=2;}
else if ((l1=='.' | l1==',') & ( l1>='1' & l1<='9') & tipo==2)
{ctemp=ctemp+l1;}
else
{
if (tipo==1) salida=salida+palabra(ctemp);
if (tipo==2) salida=salida+numero(ctemp);
ctemp="";
tipo=0;
if (l1==',') salida=salida+"-0-";
else if (l1==';') salida=salida+"-1-";
else if (l1=='.') salida=salida+"-2-";
if((l1>='a' & l1<='z') | (l1>='• & l1<='•) | l1=='• | (l1>='1' &
l1<='9') | l1=='•  )
{i=i-1; }
}
```

Listado 4.1 – Este programa analiza la cadena de entrada y la separa en palabras.

En la siguiente tabla se encuentra el modulo que toma una palabra y la reemplaza por su representación fonética de acuerdo a las reglas mencionadas.

```
b=palabra.GetLength();
b=b-1;
for (a=0;a<=b;a++) {
l1=palabra[a];
if(a<b)   l2=palabra[a+1];
else l2=' ';
if(a<b-1) l3=palabra[a+2];
else l3=' ';
if(a>0)   p1=palabra[a-1];
else p1=' ';
switch (l1){
case 'c':
    l1='k';
      if (l2=='h') {l1='C';a=a+1;}
      if (vdebil(l2)) {l1='s';}
    break;
case 's':
      if (l2=='h') { l1='S';a=a+1;}
break;
case 'l':
      if (l2=='l') {l1='L';a=a+1;}
break;
case 'r':
      if (l2=='r') {l1='R';a=a+1;}
if (!vocal(p1)) {l1='R';}
break;
case 'q':
      l1='k';
      if(l2=='u') {a=a+1;}
      break;
case 'v':
      l1='b';
      break;
case 'z':
      l1='s';
```

```
      break;
case 'y':
      if(!vocal(l2)) {l1='i';}
      break;
case 'g':
      if (l2=='e' | l2=='i') {l1='j';}
      if (l2=='u' & (vdebil(l3))) {a=a+1;}
      break;
case 'ü':
      l1='u';
      break;
case ' ':
      l1='-';
 break;}
salida = salida + l1; }
```

Listado 4.2 – Esta función recibe una palabra y la reemplaza por su representación fonética. La palabra se recibe en la variable "palabra" y el resultado se almacena en la variable "salida"

```
int vocal(TCHAR a) {
      if (a=='a' | a=='á'  | a=='o' | a=='ó' | a=='u' | a=='ú' | a=='e' |
a=='é' | a=='i' | a=='í')
    return(1);
      else
      return(0);}

int vdebil(TCHAR a) {
      if ( a=='e' | a=='é' | a=='i' | a=='í')
    return(1);
      else
      return(0);}
```

Listado 4.3 – Funciones adicionales que son utilizadas en el listado anterior. Vocal regresa 1 si la letra es una vocal, 0 en caso contrario. Vdebil regresa 1 si la letra es una vocal débil (e,i) o 0 en caso contrario.

Nótese que estas reglas solo se aplican a palabras originarias de la lengua española. Debido a esto la pronunciación de algunas palabras, al ser prestamos de otros idiomas y no seguir las reglas de la lengua española, no puede ser obtenida a partir de ellas. Para estas palabras se requieren métodos adicionales para obtener su pronunciación correcta.

Una vez que se ha obtenido la secuencia de fonemas a reproducir necesitamos encontrar la silaba que lleva el énfasis dentro de la palabra llamado acento tónico, el cual puede es representado por un acento escrito (´) . El acento será escrito en los siguientes casos de acuerdo a la silaba que lleva el acento tónico:

Palabras Agudas: La silaba que lleva el énfasis es la ultima, llevan acento escrito solo si terminan en n, s o vocal.

Palabras Graves: La silaba que lleva el énfasis es la penúltima, llevan acento escrito solo si no terminan en n, s o vocal.

Palabras Esdrújulas: La silaba que lleva el énfasis es la antepenúltima y siempre llevan acento escrito.

Palabras Sobre-esdrújulas: Llevan el énfasis antes de la penúltima silaba y siempre llevan acento escrito.

Si la palabra lleva acento escrito ya no se requiere hacer nada para encontrar el acento tónico. En caso contrario se debe obtener la ultima o las dos ultimas silabas de la palabra para encontrar el punto donde se colocara el acento tónico.

Para separar las palabras en silabas existen 10 reglas básicas:

REGLA 1.- En las sílabas, por lo menos, siempre tiene que haber una vocal. Sin vocal no hay sílaba.

REGLA 2.- Existen conjuntos de consonantes que deben ser mantenidos juntos y pertenecen siempre a la misma silaba: br, bl, cr, cl, dr, fr, fl, gr, gl, kr, ll, pr, pl, tr, rr, ch.

REGLA 3.- Cuando una consonante se encuentra entre dos vocales, se une a la segunda vocal.
Ejemplo: une -> u-ne

REGLA 4.- Cuando hay dos consonantes entre dos vocales, cada vocal se une a una consonante excepto si son consonantes consideradas inseparables (ver regla 2)

Ejemplo: componer -> com-po-ner

Aprender -> a-pren-der

REGLA 5.- Si son tres las consonantes colocadas entre dos vocales, las dos primeras consonantes se asociarán con la primera vocal y la tercer consonante con la segunda vocal excepto si la segunda y tercera consonantes están dentro del grupo de inseparables.

Ejemplo: transporte -> trans-por-te

Cumple -> cum-ple

REGLA 6.- Las palabras que contienen una h precedida o seguida de otra consonante, se dividen separando ambas letras.

Ejemplo. Anhelo -> an-he-lo

REGLA 7.- El diptongo es la unión inseparable de dos vocales. Se pueden presentar tres tipos de diptongos posibles:

1) Una vocal abierta + una vocal cerrada
2) Una vocal cerrada + una vocal abierta
3) Una vocal cerrada + una vocal cerrada

Son diptongos sólo las siguientes parejas de vocales: ai, au, ei, eu, io, ou, ia, ua, ie, ue, oi, uo, ui, iu, ay, ey, oy.

Ejemplo: jaula -> jau-la

La unión de dos vocales abiertas o semiabiertas no forman diptongo, es decir, deben separarse en la segmentación silábica. Pueden quedar solas o unidas a una consonante. Ejemplo: aéreo -> a-é-re-o

REGLA 8.- La h entre dos vocales, no destruye un diptongo.

Ejemplo: ahuyentar -> ahu-yen-tar

REGLA 9.- La acentuación sobre la vocal cerrada de un diptongo provoca su destrucción.

Ejemplo: María -> Ma-rí-a

REGLA 10.- La unión de tres vocales forma un triptongo. La única disposición posible para la formación de triptongos es la que indica el esquema:
Vocal cerrada + vocal abierta o semiabierta + vocal cerrada
Sólo las siguientes combinaciones de vocales, forman un triptongo: iai, iei, uai, uei, uau, iau, uay, uey.

De acuerdo a estas reglas existen 4 tipos de silabas:

1) V -> vocal (1º 2)
2) VC -> vocal (1 o 2) + consonante (1)
3) CV -> consonante (1 o 2) + vocal (1,2 o 3)
4) CVC -> consonante (1 o 2) + vocal (1,2 o 3) + consonante (1 o 2)

En el siguiente listado se encuentra la rutina que encuentra la letra acentuada

```
* aqui hay que encontrar la vocal acentuda */
/* paso 1 -> si ya esta acentuada no hacer nada*/
if (!palabra_acentuada(salida)) {
/* paso 2 -> penultima silaba?? */
            found=0;
      if (nsv(salida)) {
            for (a=0;a<=b;a++) {
                  l2=0;
                  l3=0;
                  p1=0;
                  p2=0;
                  l1=salida[a];
                  if(a<b) l2=salida[a+1];
                  if(a<(b-1)) l3=salida[a+2];
                  if(a>0) p1=salida[a-1];
                  if(a>1) p2=salida[a-2];
                  if (vocal(l1) & !found) {
                        remember=a;
```

```
                        look=0;
                        if((a<=(b-2)) & (l2=='u' | l2=='i')) look=a+2;
                        else if(a<b) look=a+1;
                        l1=0;
                        while(look<=b      &      (!vocal(salida[look])))
look=look+1;
                        /*if (look>=0 & look<=b) */
                        l1=salida[look];
                        if(look>=(b-1) & vocal(l1)) found=1;
                  }
            }
      }
/* paso 3 -> ultima silaba */
    if (!found) {
            if (b==0 & vocal(salida[0])) {remember=0;found=1;}
if(b>0){
      if (vocal(salida[b-1])) {remember=b-1;found=1;}
      else if(b>1 & vocal(salida[b-2]) & (!vocal(salida[b-1]) | salida[b-
1]=='i' | salida[b-1]=='u') & !vocal(salida[b])) {remember=b-2;found=1;}
}

      }

if (found==1) salida=poner_acento(salida,remember);
}
```

Listado 4.4- Esta función analiza la palabra ya convertida y agrega un símbolo de acento si no lleva acento escrito.

```
int acento(TCHAR a) {
            if (a=='á' | a=='é' | a=='í' | a=='ó' | a=='ú')
    return(1);
      else
      return(0);}

int palabra_acentuada(CString palabra) {
int a,b;
```

```
TCHAR l1;
b=palabra.GetLength();
      b=b-1;
for (a=0;a<=b;a++) {
l1=palabra[a];
if (acento(l1)) return(1);
}
return(0);
}

int nsv(CString palabra) {
int b;
TCHAR l1;
b=palabra.GetLength();
b=b-1;
l1=palabra[b];
if(l1=='n' | l1=='s' | vocal(l1) ) return(1);
else return(0);
}

CString poner_acento(CString palabra,int lugar) {
LPTSTR p = palabra.GetBuffer( 50 );
if (palabra[lugar]=='a') p[lugar]='á';
else if (palabra[lugar]=='e') p[lugar]='é';
else if (palabra[lugar]=='i') p[lugar]='í';
else if (palabra[lugar]=='o') p[lugar]='ó';
else if (palabra[lugar]=='u') p[lugar]='ú';
palabra==p;
      return(p);
}
```

Listado 4.5- Incluye las funciones adicionales que son mandadas llamar por el bloque anterior.

Una vez hecho esto ya tenemos la palabra lista para ser enviada al modulo de generación de voz que la transformara a una salida de audio.

5.CONVERSIÓN FONEMAS - VOZ

Una vez que se tiene ya la salida del conversor texto-fonemas necesitamos convertir estos fonemas a una salida de audio.

Para un sistema concatenativo se requiere como primer paso el generar la base de datos de segmentos. Para este sistema en particular se utilizaran difonemas que corresponden a la sección desde la mitad de un fonema hasta la mitad del siguiente. Como cada difonema consta de 2 medios fonemas el numero total debe ser el numero de combinaciones existentes de todos los fonemas en grupos de 2. Al haber 24 fonemas en el español mexicano se requieren 24^2 (576) difonemas diferentes. Aunque algunas combinaciones no se pueden dar en la lengua española por lo que su numeo se reduce. Estas combinaciones fueron omitidas en su mayoría por la dificultad de su pronunciacion, por lo que no se pudieron generar difonemas utilizables para algunas, y en el caso de llegar a presentarse se presentan como dos sonidos independientes separados por una pequeña pausa.

Además, para simplificar el procesamiento cada fonema vocalico se considero para fines prácticos como 2: una versión llana y una acentuada. En teoría deberian utilizarse el numero de alófonos y no solo el numero de fonemas, pero esto es poco utilizado debido a la cantidad de espacio de almacenamiento requerido. En los sistemas comerciales de síntesis en español se consideran normalmente solo 2 alófonos (para /i/ y /u/) .

Una vez que se decidió el tipo de segmentos a utilizar, se requiere hacer una grabación que contenga todos los segmentos necesarios, ya sea leyendo un texto lo suficientemente extenso o con un conjunto de palabras sin sentido pero creadas de tal manera que contengan todos los segmentos necesarios,. Para generar este sistema se utilizaron palabras sin sentido ya que esto disminuye la cantidad de texto que se requiere grabar.

Una vez que se tiene la grabación se procedió a cortar y clasificar los segmentos. Para los sonidos donde la zona de estabilidad es un ciclo (vocales, semivocales y nasales) se tomo un punto donde

existiera un cruce con el origen (para disminuir las discontinuidades) como inicio del ciclo. Para las consonantes fricativas que consisten principalmente en ruido se tomo un cruce con cero cercano al centro. Para las consonantes plosivas y africativas se utilizo la zona de silencio que existe en el momento en que se esta generando la presión necesaria para la explosión de sonido. Cada uno de los difonemas una vez cortado se almacena en un archivo con el nombre de los dos medios fonemas que contiene.

Por ultimo se agrega un modulo al programa que tome los fonemas que se obtuvieron de analizar el texto de entrada y los separe en grupos de dos para unir los difonemas requeridos. Estos segmentos son almacenados en un archivo nuevo en formato WAV que puede ser reproducido con cualquier reproductor de audio. Finalmente este archivo de salida es reproducido.

Para generar este archivo WAV primero se copia un header donde se indica el tipo de archivo y sus propiedades. después se va almacenando la señal de sonido a partir de los diferentes difonemas almacenados. Al ir almacenando los difonemas se requiere almacenar la longitud del archivo.

Una vez que se ha generado la salida completamente se regresa al inicio del archivo para añadir al header la longitud final del archivo de sonido.

Este header ha sido previamente generado (excepto el valor de longitud, para el cual solo se ha dejado el espacio lleno con ceros)

Segmento RIFF (12 bytes)

Byte	
0 - 3	"RIFF" en caracteres ASCII
4 - 7	Longitud total del archivo en binario
8 - 11	"WAVE" en caracteres ASCII

Segmento de Formato (24 bytes)

Byte	
0 - 3	"fmt_" en caracteres ASCII
4 - 7	Longitud del segmento de Formato (0x10)
8 - 9	(0x1)
10 - 11	Numero de canales(1=Mono, 2=Stereo)
12 - 15	Velocidad de muestreo en Hz (binario)
16 - 19	Bytes por segundo
20 - 21	Bytes por muestra : 1=8 bit Mono, 2=8 bit Stereo or 16 bit Mono, 4=16 bit Stereo
22 - 23	Bits por muestra

Segmento de Datos

Byte	
0 - 3	"data" en caracteres ASCII
4 - 7	Longitud del segmento de datos
8 -	Datos

Tablas 5.1-3: Secciones de la cabecera de un archivo wav

Aquí podemos ver el Header ya generado, dividido en secciones. El formato de la señal de salida es Mono a 22.05 Khz. Se indica con XX las secciones que se modificara de acuerdo a la longitud de la salida:

```
SEGMENTO RIFF:
52 49 46 46 XX XX XX XX 57 41 56 45
R  I  F  F  |- LONG. -| W  A  V  E

SEGMENTO DE FORMATO (8 bits, mono, 22.050 KHz)
66 6D 74 20 10 00 00 00 01 00 01 00 22 56 00 00 22 56 00 00 01 00 08 00
f  m  t  _  |- LONG. -| |-1-| MONO  |- FREC. -| |- B/seg -| |B/m| |b/m|

SEGMENTO DE DATOS:
64 61 74 61 XX XX XX XX
d  a  t  a  |- Long. -|
```

El primer segmento es igual para cualquier archivo de sonido. En el segundo segmento se indica que el archivo de salida generado es Mono a 22.05 Khz. A 8 bits. En este caso se usa 1 byte por muestra (8

bits) y el numero de bytes por segundo es la frecuencia multiplicada por el numero de bytes por muestra.

La longitud final del segmento de datos es el numero de bytes por muestra multiplicado por el numero de muestras que se generaron.

además de estos segmentos existen otros donde se puede almacenar información adicional que depende del programa con el que se genero el wav, pero en este caso no se utiliza ninguno.

```
SetWindowText("->"+salida);
salida=salida+"-";
cad1 = (CEdit*) GetDlgItem(IDC_EDIT2);
cad1->GetWindowText(archson);
if (archson=="") {archson="default";}
archson=archson+".wav";

if (!fsale.Open("data\\header.wav",CFile::modeRead ,&e1)) {}
if (!fentra.Open(archson,CFile::modeCreate | CFile::modeWrite ,&e1)) {}
leido=fsale.Read(pbuf,44);
fsale.Close();
fentra.Write(pbuf,44);
j=salida.GetLength();
j=j-2;
cadena="";
tamano=0;

for (i=1;i<=j;i++) {
l1=salida[i];
l2=salida[i+1];
ctemp=l1;
ctemp2=l2;

switch (l1){
case 'C':
ctemp="ch";
```

```
break;
case 'L':
ctemp="y";
break;
case 'R':
ctemp="rr";
break;
case 'S':
ctemp="sh";
break;}
switch (l2){
case 'C':
ctemp2="ch";
break;
case 'L':
ctemp2="y";
break;
case 'R':
ctemp2="rr";
break;
case 'S':
ctemp2="sh";
break;}

cadena= "data\\"+ctemp+ctemp2+".pcm";
if (fsale.Open(cadena,CFile::modeRead ,&e1)) {
leido=fsale.Read(pbuf,30000);
if (!(vocal(l1) & !vocal(l2) & vocal(l3) ))
fentra.Write(pbuf,leido);
tamano=tamano+leido;
fsale.Close();}

else {
cadena="data\\"+ctemp+"-.pcm";
if (fsale.Open(cadena,CFile::modeRead ,&e1)) {
leido=fsale.Read(pbuf,30000);
fentra.Write(pbuf,leido);
```

```
tamano=tamano+leido;
fsale.Close();}
cadena="data\\-"+ctemp2+".pcm";
if (fsale.Open(cadena,CFile::modeRead ,&e1)) {
leido=fsale.Read(pbuf,30000);
fentra.Write(pbuf,leido);
tamano=tamano+leido;
fsale.Close();}  }  }

fentra.Seek(40,CFile::begin);
fentra.Write(&tamano,4);
tamano=tamano+40;
fentra.Seek(4,CFile::begin);
fentra.Write(&tamano,4);
fentra.Close();
cad1 = (CEdit*) GetDlgItem(IDC_EDIT2);
cad1->GetWindowText(cadena);
if (cadena=="") {cadena="default";}
cadena=cadena+".wav";
PlaySound(archson, NULL, SND_ASYNC | SND_FILENAME);
```

Listado 5.1 – Este listado toma la cadena de salida y apartir de el genera un archivo tipo wav que contiene la salida de voz generada para posteriormente reproducirla.

CONCLUSIONES

En esta tesis se llevo a cabo el diseño de un sistema de síntesis de voz por concatenación de difonemas. A la entrada de un texto cualquiera en lengua española puede generar una salida de voz.

Para este sistema se utilizaron difonemas debido a da buenos resultados y su cantidad es aceptable. El numero teórico de difonemas es de el cuadrado de fonemas del español (23^2=529), pero debido a que existen muchas combinaciones que no se pueden dar (por ejemplo rrr o nñ) el numero es menor. además de esto se contaron las vocales acentuadas como fonemas independientes para generar de forma mas sencilla las silabas acentuadas.

El numero total de difonemas que se utilizo fue de 400 elementos. Con este numero de elementos se logro generar una salida de voz comprensible, aunque es probable que si se utilizaran algunos alófonos (versiones alternativas de los fonemas) o se mezclaran unidades de tamaño mayor (por ejemplo trifonemas) para combinaciones donde el cambio es muy rápido y existe mayor contaminación entre fonemas (por ejemplo en los triptongos) se podría mejorar la salida del sintetizador.

Este sistema se compone de dos bloques básicos. El primer bloque toma una entrada de texto y la convierta a una cadena de fonemas que indican la forma en que se pronuncia el texto de acuerdo con las reglas del español, además de encontrar la silaba tónica en aquellas palabras que no tienen un acento escrito.

Una vez que se tiene esta cadena se procede a tomar los segmentos y concatenarlos dentro de un archivo de sonido tipo WAV, al que se le agrega una cabecera para que posteriormente pueda ser reproducido por cualquier software reproductor de sonido para después reproducir este archivo. Este archivo se queda almacenado en la computadora para que se le pueda dar un uso posterior.

Un problema del sistema es la falta de entonación. Actualmente no existe un sistema en español que pueda generar una entonación de forma automática y prácticamente no existe ninguno que pueda generar una entonación adecuada para cualquier tipo de frase. Otra mejora que se le podría hacer al sistema es un sistema que pudiera generar ciertas entonaciones aunque fuera de forma manual, como sucede en los sistemas comerciales.

Otra mejoría importante que se le podría hacer al sistema es en la conversión a fonemas. Este bloque puede generar correctamente la lectura de una palabra en español, sin embargo se podría agregar una base de datos de conversiones para aquellas palabras tomadas de lenguas extranjeras que se usan comúnmente en el español, ya que estas no se pueden inferir sus lecturas por medio de reglas. además un diccionario de conversiones ayudaría también a la conversión de abreviaturas que podrían incluirse dentro de la base de datos.

BIBLIOGRAFÍA

Libros

(1) E. Keller et al **Improvements in Speech Synthesis**
Willey and Sons, 2002, Inglaterra

(2) G. Bailly / C. Benoit **Talking Machines: Theories, Models and Designs**
Ed. North Holland, 1992, Holanda

(3) Dutoit Thierry **An introduction to text-to-speech synthesis**
Kluwer Academic,1997, Holanda

(4) Holmes, J.N. **Speech synthesis and Recognition**
Von Nostrand Reinbold,1988,Inglaterra

(5) Goldstein, Bruce **Sensation and Perception**
5th Edition, Brooks Cole Publishing Co., 1999, U.S.A.

Paginas Electronicas

(6) Lemmetty, Sami **Review of Speech Synthesis Technology**
http://www.acoustics.hut.fi/~slemmett/dippa/

(7) **History of speech sinthesys 1770-1970**
http://www.ling.su.se/staff/hartmut/kemplne.htm

(8) G. Stork, David **HAL's Legacy: 2001's Computer as Dream and Reality**
http://mitpress.mit.edu/e-books/Hal/

(9) Lleida Solano, Eduardo **Conversion Texto-Voz**
http://www.gtc.cps.unizar.es/~eduardo/investigacion/voz/ctv.html

(10) Cano,Rafael **Apuntes de Gramatica Española**
http://members.ferrara.linux.it/elfenor/espanol/gramatica.htm

(11) Castejon Lapeyra, F. Et al **Un conversor de texto-voz para español**
http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol52/artic8/8.html

(12) Figueroa Mora, Karina **Tesis de Licenciatura**

http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol52/artic8/8.html

(13) Hyperphysics Web Site
http://230nsc1.phy-astr.gsu.edu/hbase/hframe.html