Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Ciudad de México

División de Ingeniería y Arquitectura
Departamento de Mecatrónica
Cátedra de Tecnologías de la Información aplicada a la Biomédica

Síntesis de Voz Esofágica

Proyectos de Ingeniería II

Profesor:
Dr. Jorge Eduardo Brieva

Asesor:
M. en C. Alfredo Mantilla Caeiros

Autores:
Alejandro Hernández Araujo 995537
Jaime Mauricio Moreno 969396

14 de Noviembre de 2006

Índice general

1 Introducción

Para poder apreciar la importancia de la comunicación es importante entender los procesos básicos del lenguaje y la forma en que las personas usan dicho lenguaje para comunicarse. El objetivo principal del lenguaje es comunicar ideas y las palabras y frases que usamos no son tan importantes por sí mismas. El desarrollo de la actividad intelectual y la adquisición del lenguaje en los seres humanos comienzan en la infancia, así mismo la capacidad del lenguaje de codificar ideas de forma conveniente para procesarlas mentalmente. Por lo tanto se puede considerar a la voz como el medio de comunicación primordial entre las personas.

Este proyecto presenta el desarrollo de un sistema de síntesis de voz que permita mejorar las cualidades de percepción de la voz esofágica, que es una forma de producción de voz resultado de la extracción de la laringe debida a distintos padecimientos entre los que se encuentran el cáncer. El desarrollo de los algoritmos de síntesis de voz esofágica representa la etapa final del sistema que inicia con el análisis y el reconocimiento de voz.

Uno de los propósitos de este proyecto es generar un conjunto de algoritmos que sirvan como base para el diseño de dispositivos que permitan a las personas que hayan sido sometidas a laringectomía, tener una mejor calidad de voz y que esto a su vez repercuta principalmente en su calidad de vida.

La voz esofágica es una forma de comunicación usada comúnmente por pacientes que han sufrido una laringectomía total. Esta operación es resultado de cáncer en la laringe. En la laringectomía se extirpa la laringe separándola de la tráquea, la respiración posterior a esta operación se realiza a través de una apertura en el cuello que se conoce como *stoma*[1]. Esta apertura permite el flujo de aire hacia y desde la cavidad respiratoria. El esófago, el tubo muscular que conduce al estómago, actúa como el tracto vocal, entonces la voz se produce al introducir aire en el esófago, el aire se excita a su vez por la pulsación del área faringe-esofágica. Este aire se usa como fuente de excitación para la voz. El aire inhalado produce un sonido ruidoso controlado cuyas características de forma de onda se modifican en sus características por los articuladores y finalmente se radia a través de los labios. La voz esofágica se puede describir como ronca, tensa y de baja intensidad.

[1] Incisión parecida a la boca, realizada mediante cirugía en una superficie del cuerpo para crear una apertura hacia un órgano interno.

Las personas que ha sufrido de la laringectomía tienen varias opciones para recuperar la voz, ninguna de ellas totalmente satisfactoria. Uno de los métodos más fáciles de usar es la laringe artificial, el cual es un dispositivo manual que introduce una fuente vibratoria en el tracto vocal al hacer vibrar sus paredes externas, sin embargo no produce flujo de aire en el tracto lo que disminuye la inteligibilidad de las consonantes. La voz traqueo-esofágica utiliza una prótesis para desviar el aire de los pulmones hacia el esófago esto proporciona una vibración del esfínter superior del esófago, este método proporciona un flujo de aire para las consonantes y permite generar palabras de duración normal. No obstante este método requiere de una conexión entre el esófago y la traquea que se realiza mediante cirugía y no es apropiado para algunos pacientes.

En el método de la voz esofágica se requiere que se inhale o inyecte aire hacia el esófago esto limita el rango del período fundamental y la intensidad de la voz. Tanto la voz esofágica como la traqueo-esofágica se caracterizan por un promedio bajo y perturbaciones grandes entre ciclos del período fundamental de la voz, así como una intensidad promedio baja. A pesar de esto se ha encontrado que los patrones de los formantes[2] en la voz esofágica son muy similares a los patrones de los hablantes normales, excepto por algunas elevaciones en las frecuencias formantes que se atribuyen a la reducción del tracto vocal como resultado del procedimiento quirúrgico de la voz esofágica.

Para mejorar la forma de comunicación de los hablantes esofágicos, sería de mucha utilidad generar un sistema que convirtiera la voz esofágica en voz normal. Dicho sistema puede basarse en el análisis por LPC el cual se considera una herramienta efectiva para estimar los parámetros del sistema lineal de producción de voz. Para mejorar la calidad de la voz esofágica se puede reemplazar la fuente vocalizada usando un método por LPCs.

1.1. Problemática

Las personas que han sufrido Laringectomía Total tienen repercusiones no solo físicas como la pérdida de la voz, sino que esto conlleva también a tener dificultades psicosociales tales como la baja autoestima, posible pérdida de empleo y cambia significativamente su papel en la familia. Por razones como éstas se busca mejorar la calidad de la voz de las personas que han sido sometidas a una laringectomía.

Se han desarrollado diferentes dispositivos, como la laringe electrónica, y terapias de rehabilitación, como la voz esofágica, con el objetivo de permitir a la persona comunicarse con los demás.

[2] Formantes: frecuencias de resonancia del tracto vocal. Se explicará más a fondo en el capítulo 2.

Dentro de la voz esofágica, el cual es el objeto de nuestro estudio, se desea tener un dispositivo que mejore las características de la voz emitida. Para esto es necesario un estudio de análisis, reconocimiento y síntesis de la voz esofágica que se implementará en un DSP, el cual es un sistema electrónico capaz de procesar digitalmente señales.

Las partes de análisis y reconocimiento requieren un proceso especial basado en métodos de análisis y reconocimiento de señales de voz de un hablante normal, pero para el caso de voz esofágica se deben tener condiciones especiales que en conjunto con la etapa de síntesis se pueda tener un sistema capaz de producir voz de mejor calidad e inteligibilidad, las cuales suelen ser bajas debido a diferentes factores. Dichos factores pueden ser desde los problemas de coarticualción entre sonidos adyacentes dentro de una palabra hasta el desarrollo de una fuente de excitación que pueda proveer sonidos más naturales y no tan periódicos y monótonos.

1.2. Objetivos

Objetivo General:

- Diseño, implementación y validación de distintos algoritmos de síntesis de voz.

Objetivos Específicos:

- Implementación de algoritmos de síntesis de voz esofágica.
- Simulación en Matlab de algoritmos de síntesis de voz por formantes.
- Integración de un Sistema Total de Síntesis de Voz Esofágica: Análisis, Reconocimiento y Síntesis.

2 MARCO TEÓRICO

Se considera a la síntesis de voz como la generación automática de señales de voz, la cual ha tenido un desarrollo importante en años recientes. Los progresos recientes en la síntesis de voz han generado sintetizadores con gran inteligibilidad sin embargo la calidad y naturalidad del sonido continúan siendo el mayor problema.[2]

El desarrollo de los sistemas de síntesis y reconocimiento de voz puede facilitarse si se tiene un buen conocimiento de la forma en que los seres humanos generan la voz y la manera en que se puede modelar este proceso a través de circuitos eléctricos ó en una computadora. Por esto, si se cuenta con un modelo de generación de voz además de comprender mejor el proceso de producción de la voz dicho modelo puede servir como base para un sistema de síntesis.

Los órganos del cuerpo humano responsables de la generación de la voz son los pulmones, la laringe, la faringe, la nariz, y varias regiones de la boca. La fuerza muscular para expulsar el aire de los pulmones provee la fuente de energía, el flujo de aire se modula de distintas maneras para producir las componentes de potencia acústica en el rango de las frecuencias audibles y por último las propiedades del sonido resultante se modifican por el resto de los órganos vocales para producir la voz.

El proceso de resonancia acústica es de gran importancia para determinar las propiedades de los sonidos de voz. La estructura resonante principal, en el caso particular de las vocales, se conoce como el tracto vocal; este comienza en la laringe y se extiende hasta la faringe, y de la boca a los labios. La forma en la que se comportan en el tiempo las frecuencias resonantes y sus intensidades son de gran importancia para determinar la característica de la voz. Los modos de resonancia principal del tracto vocal se conocen como formantes, y por convención se enumeran a partir de la frecuencia mas baja a la que se presentan.

Desde el punto de vista técnico, se puede considerar al sistema de voz como un tubo acústico situado entre la glotis y la boca[5]. El tracto vocal, el cual se excita mediante la fuente glotal se puede aproximar como un tubo recto cerrado en la región de las cuerdas vocales con una impedancia $Z_g = \infty$, y abierto en la zona de la boca cuya impedancia es $Z_m = 0$. En este caso la función de transferencia del tracto vocal es:

$$V(\omega) = \frac{Z_m}{Z_g} = \frac{U_m}{U_g} = \frac{1}{cos(\frac{\omega l}{c})} \tag{2.1}$$

Donde l es la longitud del tubo, ω es la frecuencia en radianes y c es la velocidad del sonido. El denominador se hace cero a las frecuencias $F_i = \frac{\omega_i}{2\pi}$ en la cual i=1,2,3..., además:

$$\frac{\omega_i l}{c} = (2i-1)\frac{\pi}{2},\ y\ F_i = \frac{2i-1}{4l} \tag{2.2}$$

Si $l = 17cm$[1], $V(\omega)$ es infinito a las frecuencias Fi = 500, 1500, 2500,... Hz lo que significa que existen resonancias cada 1 kHz iniciando en 500 Hz. Si la longitud es diferente de 17 cm, las frecuencias F_i se escalaran por un factor de $\frac{17}{l}$ para que el tracto vocal se pueda aproximar mediante dos o tres secciones de tubos en donde las áreas de las secciones adyacentes son distintas y las resonancias se asocien con las cavidades individuales.

Para el caso particular de las vocales, estas se pueden aproximar a través de un modelo de dos tubos que se presenta en la Figura 2.1. En este caso la faringe esta representada por el tubo más angosto y está abierto hacia el tubo más ancho que representa la cavidad oral. Si se asume que la longitud de los dos tubos es igual a 8.5 cm., las frecuencias de resonancia ocurren al doble de la frecuencia para un solo tubo. Debido al acoplamiento acústico, las frecuencias resonantes no se acercan unas a otras a menos de 200 Hz.

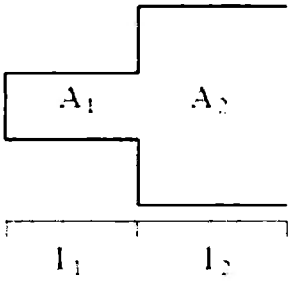

Figura 2.1: Modelo de dos tubos del tracto vocal.

Para las aplicaciones de síntesis de voz es necesario modelar la fuente del sonido y la estructura resonante del tracto vocal. Si se requiere modelar el proceso de producción de la voz de una forma muy aproximada es necesario modelar la mecánica de la vibración de las

[1]Longitud promedio del tracto vocal

cuerdas vocales detalladamente y tener modelos de la excitación que puedan alimentarse apropiadamente al tracto vocal.

La síntesis de voz se puede producir mediante diferentes métodos tales como la síntesis concatenativa, síntesisis por formantes y síntesis articulatoria. La diferencia entre estos métodos depende del tipo de aplicación y los objetivos esperados. Dentro de cada de uno de estos métodos presentan beneficios y deficiencias. Estos métodos se clasifican principalmente en tres grupos los cuales serán analizados y a partir de las metas especificadas para el proyecto de síntesis de voz esofágica se escogerá uno de ellos:

- **Síntesis concatenativa**, que utiliza muestras pregrabadas de diferente longitud a partir de la voz natural.
- **Síntesis Articulativa**, que intenta modelar el sistema de producción de voz humana de manera directa.
- **Síntesis por Formantes**, que modela las frecuencias resonantes de la señal de voz a través de la función de transferencia del tracto vocal basándose en un modelo de fuente-filtro.

Los métodos por Formantes y los Concatenativos son usados de manera usual en los sistemas de síntesis en la actualidad. El método articulatorio es aún muy complicado para implementaciones de alta calidad, pero puede ser un método con mucho potencial para el futuro.

2.1. Síntesis Concatenativa

La síntesis concatenativa se basa en la conexión de muestras naturales pregrabadas y este tipo de síntesis es probablemente la forma más fácil de producir sonidos de voz sintética inteligible y natural.

Las palabras son tal vez la unidad más básica para texto escrito y para algunos sistemas de mensaje con vocabulario muy limitado. La concatenación de palabras es relativamente fácil de realizar y los efectos de coarticulación entre palabras que se capturan en las unidades almacenadas son mínimos. Sin embargo existe una gran diferencia entre palabras habladas de manera aislada y en oraciones continuas lo que provoca que la continuidad de la voz suene poco natural. De esta manera, las sílabas como unidad de concatenación es considerablemente más pequeña que la cantidad de palabras, pero el tamaño de la base de datos de las unidades es todavía más grande para sistemas TTS (Texto a Voz).

A diferencia de las palabras, el efecto de coarticulación no se incluye en las unidades almacenadas, así que usar sílabas como unidades básicas no es muy razonable. Tampoco no hay forma de controlar los contornos prosódicos a lo largo de una oración.

Los sistemas de síntesis actuales se basan en el uso de fonemas, difonos, demisílabas y alguna combinación de estas. Los fonemas son probablemente las unidades más usadas en síntesis de voz debido a que son la presentación lingüística normal de la voz. El inventario de unidades básico es generalmente entre 40 y 50, que es mucho menor comparado con otras unidades. Al usar fonemas se tiene un máximo de flexibilidad con los sistemas basados en reglas. No obstante, algunos fonemas que no tienen una posición final de estado estacionario como las plosivas que son difíciles de sintetizar. Los fonemas se usan algunas veces como entrada de sintetizadores de voz para generar sintetizadores basados en difonos.

Los difonos (o pareja de fonos) se definen para extender el punto central de la parte de estado estable del fonema hacia el punto central del fonema siguiente, así que contienen las transiciones entre fonemas adyacentes. Esto significa que el punto de concatenación estará en la región de estado más estacionario de la señal, lo que reduce la distorsión de los puntos de concatenación. Otra ventaja con los difonos es que los efectos de coarticulación no necesitan formularse como reglas. En principio, el número de difonos es el cuadrado del número de fonemas (mas los alófonos), pero no se necesitan todas las combinaciones de fonemas.

Una vez descritas algunas de las unidades para la síntesis concatenativa, es necesario describir el inventario de unidad. La construcción del inventario de unidad consiste en tres fases principales. Primero la voz natural debe ser grabada para que se incluyan todas las unidades usadas (fonemas) dentro de todos los contextos posibles (alófonos). Después de esto las unidades deben ser etiquetadas o segmentadas a partir de los datos de voz hablada, y finalmente se deben escoger las unidades mas apropiadas.

2.2. Síntesis por Formantes

Las teorías acústicas han demostrado que las características de transmisión del tracto vocal se pueden aproximar por un conjunto de resonadores (y antiresonadores) cuyos anchos de banda y frecuencias de resonancia se pueden controlar de manera independiente. Por lo tanto el tracto vocal se puede representar como un filtro lineal invariante en pequeños intervalos de tiempo. Para sonidos vocalizados el modelo de fuente glotal se representa por un tren de pulsos cuasi periódicos con amplitud y período controlable. Para

sonidos no vocalizados el modelo de la fuente se representa por ruido blanco aleatorio. Las señales de voz se sintetizan excitando el filtro del tracto vocal con la fuente.

Las características personales de la voz están determinadas por las frecuencias formantes y por la fuente vocalizada. Las características de la fuente solo se pueden determinar de manera indirecta.

La teoría del modelo fuente-filtro para la producción de la voz es la base de la mayoría de los sintetizadores de hoy en día. Esta teoría establece que la fuente glotal y el filtro del tracto vocal se pueden separar de manera lineal y que no existe interacción entre ellos, esto implica que la característica de variación en el tiempo del tracto vocal no tiene efecto en la forma de onda de los pulsos de la fuente glotal.

El filtro se excita por la fuente que puede ser tanto una simulación de la vibración del tracto vocal para segmentos vocalizados o una señal de ruido que simule una opresión en algún punto del tracto vocal. La voz sintetizada mediante este método generalmente es inteligible pero a menudo suena poco natural. La Figura 2.2 muestra la representación del modelo Fuente -Filtro.

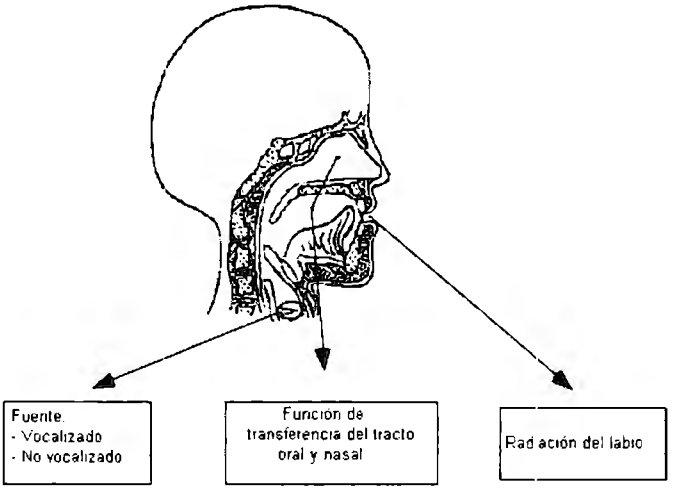

Figura 2.2: Concepto del modelo Fuente - Filtro.

2.2.1. Modelado de la Fuente

Los sonidos de la voz se dividen en aquellos producidos por una vibración periódica de las cuerdas vocales (sonidos vocalizados) y aquellos generados sin vibración de las cuerdas vocales, sino por ruido plosivo o de fricción (sonidos no vocalizados). Por lo tanto son necesarias dos fuentes de excitación:

- La fuente vocalizada, producida por una onda cuasi-periódica.
- La fuente no vocalizada, un generador de ruido.

El modelo de una fuente vocalizada esta compuesto por un generador de tren de impulsos que produce pulsos a una tasa de 1 por período fundamental. Esta señal alimenta un filtro lineal cuya respuesta en frecuencia $G(f)$ aproxima la forma de onda glotal. Al final se tiene un control de ganancia que permite el ajuste de la amplitud del segmento vocalizado. En la Figura 2.3 se presenta el diagrama a bloques del modelo de fuente vocalizada.

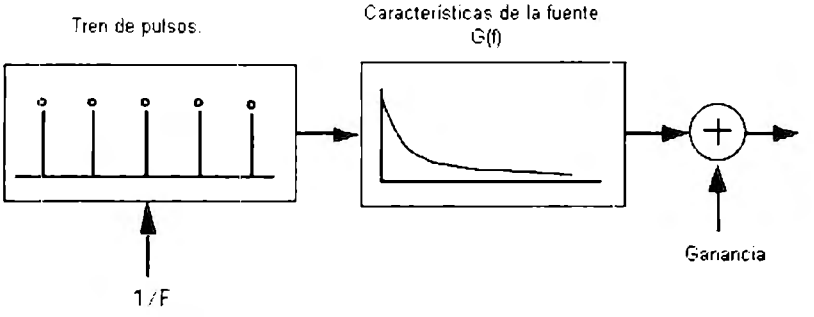

Figura 2.3: Diagrama a bloques del modelo básico de fuente vocalizada.

Existen diferentes modelos para representar la función de transferencia $G(f)$. De la Figura 2.3 se observa ue la función de transferencia tiene una característica de filtro pasa-bajas. El modelo más sencillo consiste de un filtro pasa-bajas cuya pendiente puede variar dependiendo del tipo de actividad de las cuerdas vocales. La pendiente espectral puede variar aproximadamente entre -12 dB/octave y -24 dB/octave. Este valor se relaciona directamente con la duración de la apertura glotal[5].

Los sonidos no vocalizados se generan cuando las cuerdas vocales están en un modo no vibrador y se mantienen abiertas. El modelo común para la fuente no vocalizada consiste de un generador de ruido blanco semi-aleatorio y un parámetro de ganancia.

El análisis de la voz para poder obtener una voz sintetizada de alta calidad requiere de modelos muy aproximados no tan solo para el tracto vocal sino también para la fuente de excitación. Se sabe muy bien que una representación precisa de la fuente de excitación es de gran importancia para obtener sonidos con alta naturalidad. Mientras que las características del tracto vocal se pueden modelar mediante un filtro solo-polo, aun se carece de modelos eficientes para la fuente de excitación. La razón principal de esto es que no se tiene suficiente conocimiento acerca del comportamiento fonador de la glotis, la cual se ha comprobado es bastante compleja. Como consecuencia de esto se

carece de métodos de análisis de voz que tomen en cuenta la fuente de excitación.

Los modelos glotales se pueden dividir en dos categorías principales: los modelos interactivos y los no interactivos. En los modelos interactivos, el flujo glotal se calcula a partir del área glotal y de funciones que incorporan al modelo las diferentes impedancias del sistema acústico. Estos modelos requieren de un conocimiento detallado de las características físicas de las diferentes partes que conforman la glotis, y dichas características no se poseen en la mayoría de los casos. Es por esto que la interacción entre la fuente y el tracto vocal no se tiene bien establecida. Esto provoca que en la actualidad los modelos no interactivos del flujo glotal sean los más usados para modelar la función de excitación[6].

Los modelos no interactivos se basan en parámetros para modelar directamente el flujo glotal ó la función derivativa del flujo, incluyendo de manera implícita algunos de los efectos que se producen en la interacción del flujo con el tracto vocal.

El comportamiento de la glotis se puede estudiar mediante métodos ópticos o electrográficos, que miden los movimientos de las cuerdas vocales, pero no entregan el flujo glotal directamente. Por lo tanto es deseable poder extraer las características esenciales de la fuente glotal directamente de la forma de onda de la voz. De esta manera el método del filtro inverso se ha convertido en uno de los métodos más importantes para el análisis de la fuente glotal de excitación. No obstante, el ajuste del filtro inverso para los modelos glotales se hace de forma subjetiva y cualitativa de acuerdo a la forma de onda glotal. Es por esto que los modelos matemáticos que son capaces de describir las características principales de la fuente de excitación son de gran importancia para poder obtener voz sintetizada de buena calidad.

2.2.2. Modelo LF de la Fuente

Un modelo más aproximado de la fuente es el modelo LF de Fant, Liljencrants y Lin, este modelo se basa en parámetros que modelan la respuesta temporal del tracto vocal cuando el aire fluye a través de la glotis.

Como la producción de voz se observa como un proceso de filtrado lineal que puede ser considerado como invariante en el tiempo para períodos cortos de tiempo, entonces se tiene que el flujo glotal actúa como la fuente que excita el tracto vocal y éste a su vez se ve modificado por la impedancia del labio. Como la relación presión contra volumen en los labios se puede aproximar por un diferenciador la forma de onda de la voz medida en los labios se puede expresar como la derivada del flujo glotal. Como el efecto de la radiación del labio de introduce se incluye en la fuente entonces la excitación del tracto vocal se convierte en el flujo glotal derivativo.

La relación entre el flujo glotal y su derivativo se presenta en la Figura 2.4. La rapidez en el cerrado de las cuerdas vocales resulta en una respuesta de tipo impulso negativo al cual se le llama el pulso glotal. El intervalo de tiempo en el que las cuerdas vocales permanecen cerradas y durante el cual no hay flujo de aire se conoce como la fase cerrada. El intervalo de tiempo en el cual las cuerdas vocales se encuentran total o parcialmente cerradas se conoce como la fase abierta. El intervalo de tiempo que va desde el valor más negativo del flujo derivativo al momento en que se cierran las cuerdas vocales es la fase de retorno.

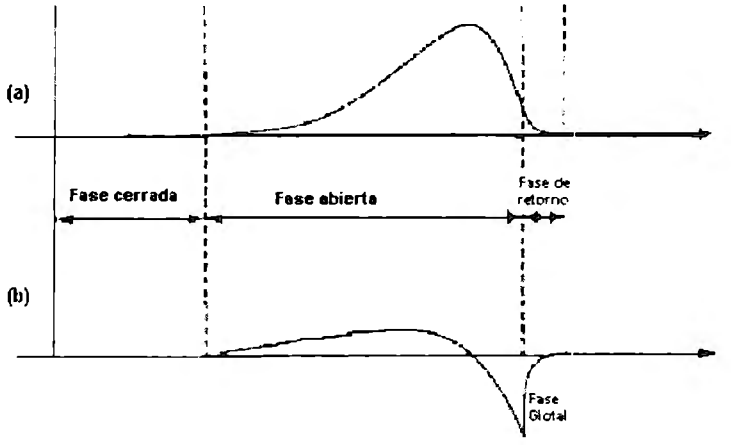

Figura 2.4: Relación entre el flujo glotal y su derivativo. (a) Flujo glotal. (b) Flujo glotal derivativo.

El modelo LF describe al flujo glotal derivativo en términos matemáticos como una señal sinusoidal que crece exponencialmente en la fase abierta y como una exponencial decreciente en la fase de retorno. El modelo utiliza cuatro parámetros para modelar el flujo glotal, la Figura 2.5 muestra los parámetros que definen el flujo glotal derivativo.

La primera parte se representa por tres de los cuatro parámetros del modelo:

$$g(t) = E_0 e^{\alpha t} \sin(\omega_g t) \tag{2.3}$$

$$\text{para } 0 \leq t \leq t_e$$

Esta parte del modelo representa el flujo desde la apertura de la glotis hasta el momento en que ocurre la excitación máxima, que es el instante en que ocurre la máxima discontinuidad del flujo y que coincide con el máximo pico negativo.

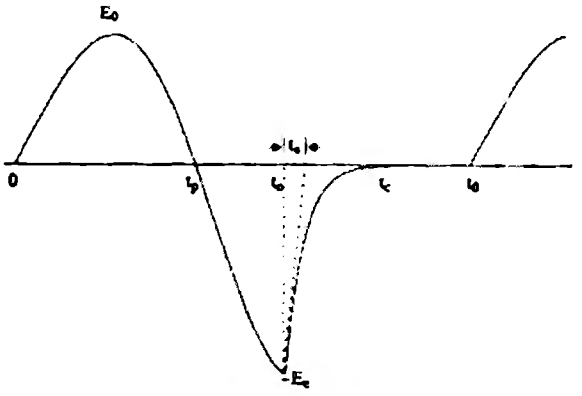

Figura 2.5: Parámetros del modelo LF.

Los tres parámetros que definen este segmento del flujo son:

1. E_0 que es un factor de escala definido por:

$$E_0 = \frac{E_\varepsilon}{e^{\alpha t} sin(\omega_g t_\epsilon)} \tag{2.4}$$

2. $\alpha = B\pi$ donde B es el ancho de banda del segmento exponencial creciente.

3. $\omega_g = 2\pi F_g$, donde $F_g = \frac{1}{t_p}$ y $t_p$es el tiempo de elevación (el tiempo desde la apertura glotal al flujo máximo).

La segunda parte del modelo es un segmento exponencial que permite un flujo residual después del máximo pico negativo, en el tiempo te que es cuando las cuerdas vocales se cierran. Este segmento se representa por:

$$g(t) = -\frac{E_\varepsilon}{\varepsilon t_\alpha}\left[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}\right] \quad para\ t_e \leq t \leq t_c \leq T_0 \tag{2.5}$$

Donde t_α es el cuarto parámetro. E_ε es la magnitud de la máxima amplitud negativa del flujo glotal derivativo y t_c es el instante en el que se cierra el tracto vocal. Además t_α es la constante de tiempo de la curva exponencial y se determina extrapolando la tangente del flujo derivativo en el tiempo t_e hasta que intersecta con el eje x. El parámetro ε se puede determinar de la ecuación haciendo $t = t_e$ y $g(t) = -E_\varepsilon$ y queda:

$$\varepsilon t_\alpha = 1 - e^{-\varepsilon(t_c-t_e)} \tag{2.6}$$

Para valores pequeños de t_α, ε se aproxima a $\frac{1}{t_\alpha}$. T_0 es el período fundamental.

Además de los cuatro parámetros del modelo LF se debe cumplir una condición extra que establece que el área total bajo la curva del flujo glotal derivativo debe ser cero:

$$\int_0^{T_0} g(t) = 0 \tag{2.7}$$

En este modelo la fase de retorno es muy importante ya que esta determina la cantidad de energía de alta frecuencia tanto de la fuente como de la voz. Entre más rápido se cierran las cuerdas vocales, la fase de retorno es mas pequeña lo que resulta en mas componentes de alta frecuencia.

2.2.3. Modelo de Rosenberg

Rosenberg propuso un modelo de fuente de excitación glotal derivativa que modela de manera muy sencilla el flujo glotal. Este modelo se compone de dos funciones trigonométricas con una sola discontinuidad en la fase de cerrado de la glotis[10].

Las ecuaciones que definen el modelo son:

$$g_R[n] = \begin{cases} \frac{1}{2}\left[1 - \cos\left(\frac{\Pi n}{N_1}\right)\right] & 0 \leq n \leq N_1 \\ \cos\left[\frac{\Pi(n-N_1)}{2N_2}\right] & N_1 \leq n \leq N_1 + N_2 \\ 0 & en\ otro\ caso \end{cases} \tag{2.8}$$

Donde los parámetros que controlan el pulso glotal son N_1y N_2. N2. Estos parámetros determinan la posición dentro del intervalo del período de cada pulso en que se cierra la glotis.

2.2.4. Modelo del Tracto Vocal

Existen dos estructuras básicas en general, en paralelo y en cascada; pero para mejorar el desempeño se puede usar una combinación de ambas.

Se requieren por lo menos tres formantes para producir voz inteligible y hasta 5 formantes para producir voz de alta calidad. Esto se logra debido a que cada formante se modela generalmente con un resonador de dos polos que permite especificar tanto la frecuencia formante (frecuencia de par de polos) y su ancho de banda.

Un sintetizador por formantes en cascada consiste de resonadores pasa-banda conectados en serie y la salida de cada resonador de formantes se aplica a la entrada del siguiente resonador. La estructura en cascada necesita tan solo de las frecuencias formantes como información de control. La Figura 2.6 muestra el diagrama del sintetizador con estructura en cascada[5, 7].

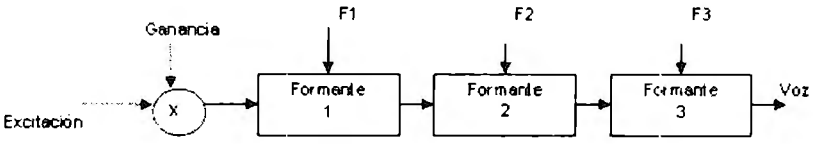

Figure 2.6: Estructura básica de un sintetizador por formantes en cascada

Otro sistema que se define para la síntesis por formantes es la estructura en paralelo que consiste de resonadores conectados en paralelo. Algunas veces se usan resonadores extras para los sonidos vocalizados nasales. La señal de excitación entra simultáneamente a todos los formantes y sus salidas se suman. Para los formantes adyacentes se deben de sumar en fase opuesta para evitar ceros no deseados o antiresonancias en la respuesta en frecuencia. La estructura en paralelo tiene como característica principal el que permite controlar individualmente el ancho de banda y la ganancia de cada formante, pero esto nos lleva a necesitar más información de control. La Figura 2.7 muestra la estructura en paralelo de un sintetizador.

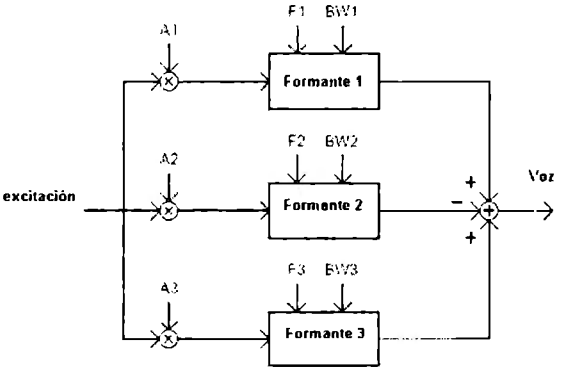

Figure 2.7: . Estructura básica de un sintetizador por formantes en paralelo.

Analizando las características de cada uno de las estructuras anteriores se puede notar que con tan solo una estructura básica es difícil lograr buenos resultados así que se han realizado esfuerzos para mejorar y combinar los modelos básicos. Con la combinación de los modelos de cascada y paralelo se obtiene el modelo PARCAS (Paralelo-Cascada) introducido por Laine (1982).

En el modelo PARCAS la función de transferencia uniforme del tracto vocal se modela con dos funciones de transferencia parciales, cada una incluyendo cada segundo formante

de la función de transferencia. De la Figura 2.8, los coeficientes k1, k2 y k3 son constantes y se escogen para balancear las amplitudes de los formantes en las vocales neutrales para mantener constantes las ganancias de las ramas paralelas para todos los sonidos.

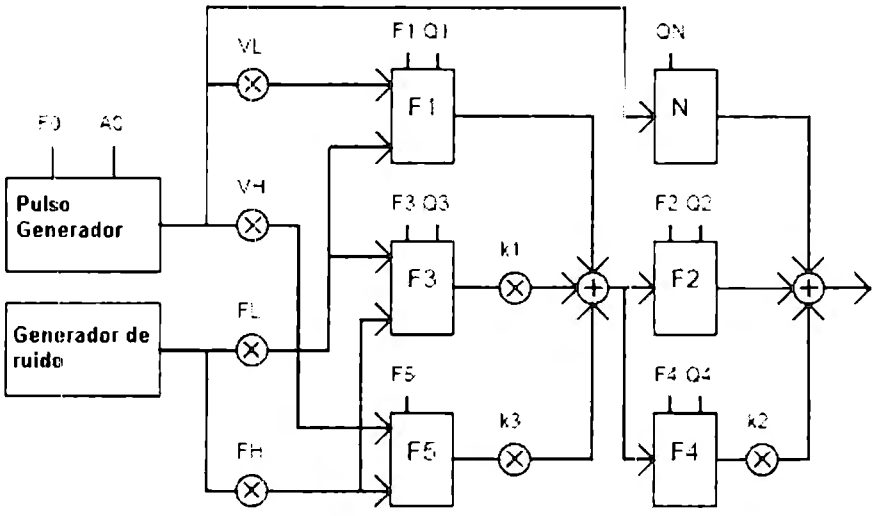

Figure 2.8: Estructura PARCAS (Paralelo-Cascada).

El modelo PARCAS usa un total de 16 parámetros de control:

- F0 y A0 - Frecuencia fundamental y amplitud de la componente vocalizada.
- Fn y Qn - Frecuencias formantes y valores de Q (frecuencia formante /Ancho de banda)
- VL y VH - Amplitud de la componente vocalizada, baja y alta.
- FL y FH - Amplitud de la componente no vocalizada, baja y alta.
- QN - Valor de Q del formante nasal a 250 Hz.

La señal de excitación usada en la síntesis por formantes consiste de una especie de fuente vocalizada o ruido blanco. Los filtros formantes representan tan solo las resonancias del tracto vocal, así que se necesita suministro adicional para los efectos del modelo de la forma de onda gutural y las características de radiación de la boca. Generalmente la forma de onda gutural se aproxima simplemente con un filtro a -12dB/octava y las características de radiación con un simple filtro a +6dB/octava.

2.3. Síntesis Articulatoria

La síntesis Articulatoria intenta modelar los órganos vocales humanos y es por esta razón de que sean potencialmente el método más satisfactorio para producir voz sintética de alta calidad.

La síntesis Articulatoria involucra modelos de los articuladores humanos y de las cuerdas vocales[3]. Generalmente se modelan los articuladores con un conjunto de funciones de área entre la glotis y la boca. El primer modelo articulatorio estaba basado en una tabla de funciones de área del tracto vocal desde la laringe a los labios para cada segmento fonético. Para síntesis basada en reglas los parámetros de control articulatorios podrían ser por ejemplo apertura de labio, proyección de labios, altura de la punta de la lengua, posición de la punta de la lengua, altura de la lengua, posición de la lengua y apertura velica. Los parámetros de excitación o fonadores podrían ser apertura glotal, tensión de las cuerdas, y presión de los pulmones.

El diagrama a bloque de la Figura 2.9 siguiente muestra los componentes principales de la síntesis articulatoria:

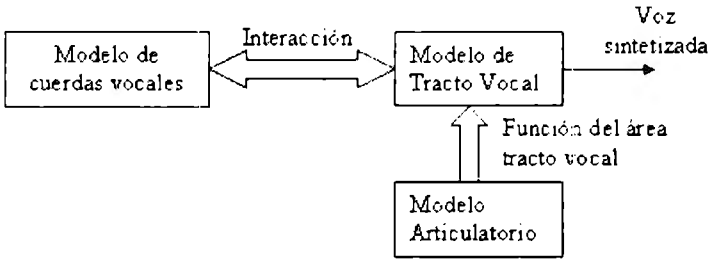

Figure 2.9: Diagrama a bloques básico de síntesis articulatoria.

El por qué se hace tan difícil modelar la síntesis articulatoria, se explica cuando se habla; los músculos del tracto vocal provocan que las articulaciones se muevan y cambien la forma del tracto vocal lo que causa los diferentes sonidos. Los datos para el modelo articulatorio se derivan generalmente de un análisis de rayos X de la voz natural. Sin embargo, estos datos son generalmente solo en 2-D cuando el tracto vocal real es naturalmente en 3-D. Dentro del modelo Articulatorio se representa de 6 a 10 parámetros de articulación que representa la posición de los articuladores de voz (lengua, labios, garganta, etc.). Dentro del modelo articulatorio se puede describir el "tubo acústico" el cual es manejado por un modelo de excitación, el cual simula la modulación del flujo de aire proveniente desde las cuerdas vocales y que también incluye una modelación detallada de

la vibración de las cuerdas vocales. Algunos otros parámetros son tomados de los efectos de la carga acústica del tracto vocal a la salida de las cuerdas vocales y también de la acústica del tracto vocal que son afectados por las cavidades subglóticas. Las características descritas anteriormente se representan como interacción entre las cuerdas vocales y el modelo del tracto vocal. Los mecanismos para modelar las turbulencias del flujo de aire, las aspiraciones y fricaciones son normalmente incluidas dentro del modelo de "tubo acústico".

Si los sistemas dinámicos (parámetros de constates de tiempo) son incorporados dentro del modelo articulatorio, el control de dicho sintetizador es proyectado por una simple alimentación de secuencias de parámetros objetivos de articulación para la vocalización que va ser sintetizada. La coarticulación entonces toma lugar naturalmente sin la necesidad de complejas reglas. Además, el sonido producido por el sistema corresponde a los movimientos de articulatorios que son físicamente posibles. Desde el instante en que los movimientos articulatorios son relativamente lentos, la velocidad de los datos para un sintetizador articulatorio puede ser teóricamente muy baja.

2.4. Síntesis por LPCs

El análisis realizado en un segmento de voz provee al sintetizador con dos parámetros de información necesarios para reconstruir la señal original. El primero es un conjunto de coeficientes que describen el filtro de síntesis, es decir, que forman el modelo del tracto vocal. El segundo parámetro es una señal necesaria para generar la señal de excitación que se alimentará al filtro del tracto vocal.

La predicción lineal es un método ampliamente usado para representar las características frecuenciales del tracto vocal. El método de la predicción lineal se conoce también como Codificación por Predicción Lineal, puesto que predice una muestra de voz en el dominio del tiempo basándose en una combinación ponderada lineal de muestras anteriores. Este método se puede considerar que elimina las redundancias en la correlación sobre pequeños segmentos adyacentes.

En el tracto vocal real, el área de su sección transversal varía de acuerdo a la posición a lo largo de tracto y con el tiempo. Estas variaciones producen los diferentes sonidos de la voz generados con la misma excitación. A medida que la excitación se propaga a lo largo del tracto vocal parte de su energía se refleja y parte de esa energía se propaga. Los coeficientes de reflexión representan que tanto de esa energía se refleja y que tanto se deja pasar. Esta reflexión de la energía es lo que genera el modelado espectral de la

excitación. Este modelado espectral se le considera el filtro del tracto vocal[4].

El filtro del tracto vocal se modela de acuerdo a la estructura de un filtro solo-polo. En esta estructura cada coeficiente predictor del filtro retrasa la señal por una unidad de tiempo y propaga solo una parte del valor de la muestra. Entonces el filtro del tracto vocal se representa por la ecuación 2.9 que muestra la función de transferencia.

$$H(Z) = \frac{1}{A(Z)} \tag{2.9}$$

donde A(z) es:

$$A(Z) = 1 - \sum_{k=1} a_k z^{-k} \tag{2.10}$$

$a_k, 1 < k < p$ son los coeficientes predictores

p el coeficiente del predictor

Si se realiza la transformación al dominio del tiempo de la ecuación 2.10 se puede ver que esta predice una muestra de voz mediante la suma ponderada de muestras anteriores, esto se puede ver en la ecuación 2.11:

$$s'(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{2.11}$$

Donde $s'(n)$ es el valor predicho mediante valores previos de la señal de voz de $s(n)$[1].

2.4.1. Estimación de los parámetros de LP

Para calcular los parámetros LP de un segmento de voz, se deben encontrar los coeficientes $a_k s$ para que la ecuación 2.11 nos de la mejor aproximación de las muestras de voz, es decir, para que $s'(n)$ se acerque má a $s(n)$ para todos los valores de n en el segmento.

La forma espectral de $s(n)$se asume estacionaria durante la duración de todo el segmento. El error entre el valor predicho y el valor real se calcula como:

$$e(n) = s(n) - s'(n) \tag{2.12}$$

Sustituyendo $s'(n)$:

$$e(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{2.13}$$

Los valores de a_k se pueden calcular minimizando el error cuadrático total en todo el segmento:

$$E = \sum_{n} e^2(n) \tag{2.14}$$

Al igualar a cero las derivadas parciales de E con respecto a a_k se obtiene un conjunto de ecuaciones que minimizan el error. Este conjunto de ecuaciones se puede resolver por varios métodos, aquí presentamos el método de la autocorrelación.

2.4.2. Uso de la autocorrelación para estimar parámetros

En este método se asume que el segmento de voz es cero fuera de las fronteras de dicho segmento. Las ecuaciones para las $a_k s$ se expresan de forma compacta en la matriz de la forma:

$$\begin{pmatrix} r(o) & r(1) & \cdots & r(p-1) \\ r(1) & r(2) & \cdots & r(p-2) \\ \vdots & \vdots & \vdots & \vdots \\ r(p-1) & r(p-2) & \cdots & r(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{pmatrix} \tag{2.15}$$

Donde $r(l)$ es la autocorrelación de la l en retraso y que se calcula como:

$$r(l) = \sum_{m=0}^{N-1-l} s(m)s(m+l) \tag{2.16}$$

y N es la longitud del segmento de voz $s(n)$.

Debido a la estructura de la matriz tipo Toeplitz (simétrica, conteniendo en sus diagonales el mismo elemento), se puede usar la recursion de Levinson-Durbin para resolver el sistema. Las ecuaciones quedan:

$$E^{(0)} = r(0) \tag{2.17}$$

$$k_i = \frac{r(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} r(i-j)}{E^{(i-1)}} \tag{2.18}$$

$$a_i^{(i)} = k_i \tag{2.19}$$

$$a_j^{(i)} = a_j^{(i-1)} - k_i a_{i-j}^{(i)} \tag{2.20}$$

$$E^{(i)} = (1 - k_i^2) E^{(i-1)} \tag{2.21}$$

Donde $1 \leq j \leq i-1$.

En todas las ecuaciones, i es el orden actual de la recursion, y las ecuaciones se resuelven de manera recursiva para todos los ordenes de i = 1, 2, ..., p.

El coeficiente de iésimo orden de la ecuación 2.19 para valores de $1 \leq i \leq p$ es el iésimo coeficiente de reflexión y si se cumple la condición de que:

$$|k_i| < 1 \quad 1 \leq i \leq p \tag{2.22}$$

Si se cumple la condición 2.22, las raíces del polinomio predictor caerán todas dentro del círculo unitario en el plano z, y el filtro solo-polo será estable.

2.4.3. Señal de excitación en la síntesis por LPC

Una práctica común en la síntesis por LPC para modelar la señal de excitación es usar un tren de impulsos, sin embargo esto produce un sonido vocalizado poco natural y genera una señal de baja calidad. Por otro lado no se han realizado muchas investigaciones para usar una señal mas real de excitación para la síntesis por LPC debido a que no existe una relación clara entre la señal residual y la señal glotal.

A pesar de que en la síntesis por formantes se tiene una representación bastante aproximada del flujo glotal a través del modelo del pulso glotal, la síntesis por LPC no utiliza explícitamente las frecuencias formantes para representar el tracto vocal, mas bien realiza una aproximación de la envolvente del espectro logarítmico de la voz. Esto da como resultado que la señal de excitación para la síntesis por LPC deba satisfacer otros requerimientos.

Debido a que el análisis mediante los predictores lineales realiza una aproximación de la envolvente es necesario que el pulso de excitación tenga un espectro que sea esencialmente plano. Para obtener una señal de excitación natural entonces es necesario conservar las características de fase de la forma de onda glotal. Ante todo esto para generar la señal de excitación es necesario iniciar con una estimación de la onda glotal.

Esta estimación se puede realizar de forma muy exacta mediante un proceso de filtrado inverso. Este proceso inicia con una función $q(n)$ que alimenta un modelo de producción lineal de voz. Este modelo no es mas que un estimado de la función de transferencia del tracto vocal que se puede obtener mediante el método de la covarianza y que se usa como filtro inverso para producir el estimado de la función de excitación efectiva $q'(n)$[10, 11].

Esta señal $q'(n)$ que se genera por el método mencionado no es más que la señal residual para el análisis por covarianza en el intervalo de apertura de la glotis, es decir, en el intervalo en el que fluye el aire a través de las cuerdas vocales para producir el

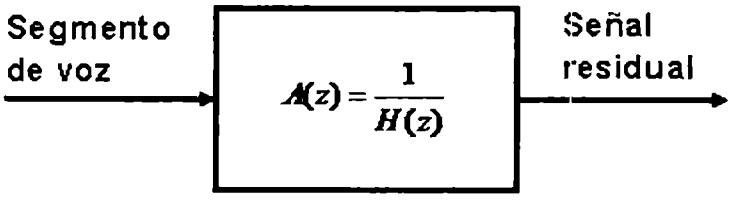

Figura 2.10: Diagrama de la obtención de la señal residual.

sonido vocalizado. Mediante este método se demuestra que si se obtienen las raíces del filtro de autocorrelación de décimo orden o mayor para un sonido vocalizado, este par de raíces complejas aproxima el comportamiento espectral de la señal $q'(n)$ que se obtiene mediante el análisis de la covarianza.

Para construir una señal de excitación general mediante predictores lineales primero se debe realizar el análisis por LPC de toda la señal para extraer la envolvente y de ahí obtener los polos para el filtro inverso. La señal se debe aplicar al filtro inverso para obtener el residual que se puede tomar como función de excitación.

El diagrama de la Figura 2.10 muestra el procedimiento general para la obtención de la señal residual.

La función $A(Z)$ representa la función de transferencia del filtro inverso dado por las raíces de los coeficientes predictores que conforman a $H(Z)$. Los coeficientes predictores se calculan por medio de las ecuaciones 2.10 en el dominio del plano Z, o se pueden obtener del plano s por medio de la ecuacion 2.11 y después transformar a Z.

2.5. Ventajas y Desventajas de los tipos de Síntesis

Dentro de los tipos de sintetizadores descritos en las secciones 2.1 a la 2.4, se plantea la necesidad de analizar cuales son las ventajas y desventajas de cada uno de los modelos.

Empezando por la síntesis concatenativa, decimos que una de las desventajas de los sintetizadores concatenativos es que se limitan por lo general a un solo hablante y a una voz y requieren más capacidad de memoria de almacenamiento que otros métodos.

Uno de los aspectos más importantes en la síntesis concatenativa es encontrar la longitud de unidad correcta. Es necesario escoger entre unidades más largas o más cortas. Con unidades mas largas se logra alta naturalidad, se tienen menos puntos de concatenación y buen control de la coarticulación, pero se incrementa la cantidad de unidades requeridas y de memoria. Con unidades más cortas, se necesita menos memoria, pero

la unión y etiquetamiento de las muestras se vuelven más difíciles y complejos. Existen varios problemas en la síntesis concatenativa comparada con otros métodos:

- Distorsión debida a discontinuidades en los puntos de concatenación, que pueden reducirse usando difonos o métodos para suavizar la señal.
- Los requerimientos de memoria son generalmente muy altos, especialmente cuando se usan unidades de concatenación grandes, como sílabas o palabras.
- La recolección y etiquetamiento de datos de cada muestra de voz consume tiempo. En teoría todos los alófonos posibles se deben incluir en el material, pero se debe tener un compromiso entre la calidad y el número de muestras.

Para la síntesis por Formantes se observa que una de las características más importantes del modelo es que puede proveer un número casi infinito de sonidos lo que la hace más flexible que los métodos por concatenación en el cual la capacidad de almacenamiento es una limitante.

Dentro del modelo por Formantes se describe al sintetizador en cascada y en paralelo. La ventaja principal de la estructura en cascada es que las amplitudes relativas de los formantes para las vocales no necesitan controles individuales. La estructura en cascada se adapta mejor para sonidos vocalizados no nasales y como necesita menos información de control que la estructura en paralelo es más sencilla de implementar y esta es una ventaja.

La estructura en paralelo se adapta mejor para sonidos vocalizados nasales, fricativas y consonantes cortas, pero a diferencia de la estructura en cascada, algunas vocales no se pueden modelar con un sintetizador en paralelo.

Para la síntesis Articulatoria se plantea la dificultad de que es uno de los métodos más difíciles de implementar, así también como la carga computacional que representa que es considerablemente alta. Así que la síntesis Articulatoria es muy difícil de optimizar debido a la insuficiencia de datos de los movimientos de las articulaciones durante el habla. Otra deficiencia en la síntesis articulativa es que los datos de los rayos X no caracterizan las masas ó los grados de libertad de las articulaciones. Así también, los movimientos de la lengua son tan complicados que es casi imposible modelarlos de manera precisa. Las ventajas de la síntesis articulativa son que los modelos del tracto vocal permiten un modelado mas preciso de los transitorios debidos a cambios abruptos de área mientras que la síntesis por formantes solo modela el comportamiento espectral.

2.5.1. Ventajas y Desventajas de la síntesis por LPC

Los mayores problemas que presenta la síntesis por LPC es la calidad de la voz sintetizada debido a inexactitud en el análisis por LPC y errores que se introducen por la detección del período fundamental y de los segmentos vocalizados. En muchos experimentos realizados se ha encontrado que la voz sintetizada por LPC es generalmente muy cercana a la voz real en un rango de frecuencia muy amplio.

Para poder comprender algunas de las razones por las cuales se compromete la calidad de la voz en la síntesis por LPC se debe analizar todo el proceso desde que se descompone la señal de voz en el modelo del filtro solo-polo y la señal de excitación residual. De manera ideal el modelo de LPC extrae toda la información espectral de la envolvente de la señal de voz, de tal forma que la señal residual es una señal con un espectro plano. El parámetro de la ganancia se puede incluir ya sea antes del filtro como una parte de la señal de excitación ó después del filtro como una etapa final del mismo.

La pérdida de naturalidad en la síntesis por LPC se puede determinar por dos factores importantes:

1. La adaptación de la envolvente y la cuantización de los coeficientes LPC.

2. El modelo de excitación que alimenta el filtro, ya se el modelo glotal ó el modelo residual.

2.5.2. Adaptación de la envolvente en síntesis por LPC

La técnica de análisis por LPC es bastante efectiva para extraer la envolvente espectral en segmentos pequeños de tiempo para una señal de voz, no obstante existen algunos problemas que se presentan al realizar este análisis.

Se considera al modelo de LPC como un modelo de filtro solo-polo, así que los ceros en el espectro no se representan correctamente. Se ha propuesto una gran cantidad de algoritmos que incluyen la representación de los ceros pero no se han logrado avances significativos en la calidad de la voz sintetizada como para justificar el agregarle esa complejidad a los algoritmos.

Se observa de manera común que para un sonido de vocal estable, el modelo de LPC puede variar en su aproximación dependiendo de la posición de la ventana.

La técnica de análisis por LPC para extraer los coeficientes del filtro se considera básicamente invariante en el tiempo. Para poder realizar el análisis sobre una señal de voz se necesita fragmentar la señal en segmentos de 10 a 25 mseg de duración. En muchos experimentos se ha encontrado que para señales de voz de alta frecuencia fundamental

se requieren de ventanas de más corta duración, mientras que para señales de alta frecuencia fundamental, se deben utilizar ventanas más amplias para lograr una voz con una respuesta más suavizada.

2.5.3. La señal de excitación para la síntesis por LPC

La mayor parte de los defectos en la síntesis por LPC se ha atribuido a la aproximación de la señal de excitación por modelos que se generan a partir de señales periódicas como son los modelos glotales de la fuente de excitación.

Los problemas básicos que presentan los modelos glotales se pueden categorizar entre los que se producen por las aproximaciones en los modelos básicos y los que se producen por la extracción incorrecta de los parámetros de la frecuencia fundamental y de los segmentos vocalizados.

Los modelos glotales omiten información importante en cuanto a la magnitud y la fase que está contenida en el espectro de la señal residual. La pérdida en la información de la magnitud perjudica de manera importante la calidad de la voz sintetizada. La extracción de la frecuencia fundamental es una fuente de error importante y que perjudica la calidad de la voz debido a que la señal con que se alimenta el tracto vocal puede variar dramáticamente de un hablante a otro.

La mayoría de los problemas de calidad de un sintetizador por LPC se debe a la pérdida de información de la señal residual que caracteriza la fuente de excitación. Se ha comprobado mediante la experimentación que para las regiones de baja frecuencia de voz, la señal residual genera una mejor calidad de voz. Estos resultados son congruentes con la forma en la que se percibe la voz por los humanos.

2.6. Análisis y Reconocimiento

2.6.1. Análisis espectral basado en Formantes

La presente sección de Análisis se ha realizado usando como base los fundamentos descritos para el "Análisis por síntesis". En esta sección se describe de manera parcial y solo con fines de tener un acercamiento a los métodos de análisis desarrollados para la etapa de Análisis de voz Esofágica. Para un estudio más a fondo de los modelos utilizados para el Análisis de Voz Esofágica se hace necesario una revisión de las referencias [1][2].

Para el análisis espectral basado en formantes es necesario realizar el cálculo de los Coeficientes de Predicción Lineal (o en ingles LPC). El cálculo de LPCs es un excelente

método usado para el análisis de señales de audio. Esto debido a que la naturaleza de las señales de audio que para segmentos de voz larga duración no son representativos de procesos estacionarios es necesario tomar segmentos de voz de alrededor de 10 milisegundos en el cual la obtención de LPCs nos proporciona las características necesarias para realizar el proceso de análisis de voz.

Los LPCs nos ayudan a encontrar las propiedades del tracto vocal que genera la señal de voz que se esta analizando. Como ventaja principal del cálculo de LPCs para análisis de voz se puede mencionar la precisión de las estimaciones obtenidas y la rapidez de cálculo con la que se realiza.

Para realizar el cálculo de LPCs se toma el supuesto que una muestra de voz puede ser aproximada mediante la ecuación 2.11 descrita en la Sección 4.4. De esta manera los coeficientes de predicción lineal permiten minimizar el error entre el valor de la función que se está analizando y la función calculada a través de las muestras anteriores.

Un factor importante para el cálculo de los LPCs es el orden del predictor lineal, pues este número determinará el número de formantes que se esperan encontrar en el análisis espectral de la voz. Con base en la teoría del muestreo de Nyquist el cual para una señal de voz muestreada a 8 KHz nos determina que nuestra señal de voz se encontrará en rango máximo de hasta 4Khz; por lo que si se espera que por cada 1000 Hz se obtenga un formante, tendremos hasta 4 formantes como mínimo. Debido a esto, se necesita como mínimo un orden de predictor lineal de 8.

Como se ha mencionado al inicio de la sección de Análisis los LPCs nos ayudaran a realizar el análisis por formantes, pues con ayuda de los LPCs es posible modelar el filtro del tracto vocal que hace posible la señal de voz. Los LPCs se usan para determinar las frecuencias de los formantes (picos del espectro) del filtro resonante de voz.

De los coeficientes de predicción lineal se obtienen las raíces que conforman a los formantes. Por medio de las raíces de los coeficientes predictores se construirá el filtro del tracto vocal de solo polos que describen la envolvente de las frecuencias resonantes. En el dominio espectral se observan que los polos son los que generan los picos de la función del tracto vocal. Se puede resolver la ecuación del filtro resonante para encontrar las raíces. Al resolver la ecuación en el plano s, para cada pico en el espectro de la señal se tienen un par de complejos conjugados, por lo que se puede determinar que por cada par complejo se tiene un formante o frecuencia resonante.

Una vez que se tienen las frecuencias resonante es posible determinar la amplitud de cada una de ellas y al realizar la comparación de la energía presente en las frecuencias resonantes que reproducen el sonido emitido se puede determinar si un sonido es vocalizado o no vocalizado.

La amplitud mínima necesaria para determinar si un sonido es vocalizado o no vocalizado es necesario que rebase el 20 % de la energía presente en la señal del segmento presente. Este valor de umbral se justifica debido a que para pacientes de voz esofágica la energía de la señal decrece conforme emiten las palabras.

Por lo tanto los pasos necesarios para realizar el análisis de la voz esofágica son:

1. La adquisición de datos. Un archivo de voz esofágica se digitaliza con una frecuencia de muestreo de 8KHz.

2. Filtrado. El filtrado se realiza con la finalida de eliminar las componentes de alta frecuencia que se encuentren presente en la señal de audio.

3. Segmentación y Ventaneo. Este paso se realiza debido a que es necesario trabajar con tramas pequeñas de muestras que nos permitan realizar una mejor caracterización de la señal. La caracterización en nuestro proceso de análisis será por formantes.

4. Caracterización de los segmentos. Para la caracterización se toman en consideración los parámetros espectrales y de energía de la señal. Los parámetros espectrales se adquieren mediante los LPCs y con ayuda de los LPCs se encuentran las frecuencias resonantes de la señal. Una vez que se tiene los formantes se obtienen los parámetros energéticos de la señal mediante la amplitud de la envolvente espectral de la señal.

5. Toma de decisión. La disyuntiva que se presenta en nuestro análisis es separar entre sonidos vocalizados y no vocalizados. Los vocalizados son los que sufrirán el proceso de reconocimiento y síntesis; mientras que los no vocalizados pasarán directamente sin sufrir ningun procesamiento. La decisión se realizará tomando en cuenta el criterio de la energía, la cual se menciono que tiene que rebasar el 20 % de la energía presente en el segmento.

2.6.2. Reconocimiento por Modelos Ocultos de Markov

La etapa de Reconocimiento se basa en los algoritmos de Modelos Ocultos de Markov debido a que como se sabe una señal de voz no es del todo constante a través de todos los segmentos por muy pequeños que estos segmentos sean. Por medio de otros tipos de reconocedores basados en modelos de percepción auditiva o los predictivos aprovechan que para ciertos sonidos repetitivos dentro de las palabras normalmente tienen patrones

acústicos similares que se pueden diferencia de una palabra a otra, pero esto no siempre ocurre y para un hablante de voz esofágica los parámetros que representan cada palabra puede variar y modificarse, esto como resultado ya sea por la coarticulación de palabras adyacentes o debido a que con el transcurso del discurso la presión del aire que expulsa es menor y la intensidad e inteligibilidad disminuye; por lo cual no se puede considerar que un particular comportamiento pueda representar de manera consistente cualquier sonido que se esta tratando de reconocer.

Por las razones mencionadas se hace necesario que para mejorar el funcionamiento de un sistema reconocedor es necesario tomar en cuenta las variaciones para que de esta manera se generen mejores condiciones que ayuden al reconocedor.

Una de las soluciones que considera las modificaciones para un sistema reconocedor se puede solventar con los Modelo Ocultos de Markov. Los modelos ocultos de Markov escogen el comportamiento que más se acerca al sonido que se intenta reconocer con base en comparar la mínima distancia acumulada a través de una óptima trayectoria. La trayectoria mencionada es la que los modelos ocultos de Markov siguen respecto la variabilidad del sonido y esto ayuda a escalar la distancia entre el sonido real emitido y una aproximación. La aproximación esta dada por el modelo generado para cada sonido en particular. Los modelos son generados por patrones que representan al sonido. Cada vez que un modelo es generado, el modelo producirá un vector de características que representa al sonido; sí el modelo es bastante bueno, la estadística de un largo número de vectores de observación será muy similar a la estadística medida por el oído humano.

3 SISTEMA PROPUESTO

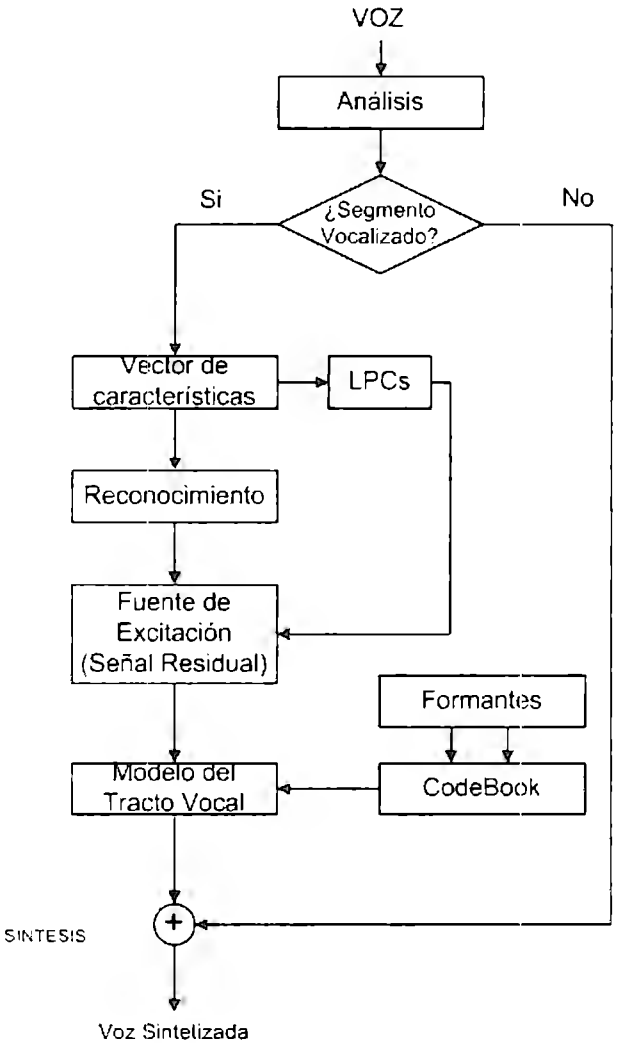

Figure 3.1: Diagrama a bloques del sistema propuesto.

El sistema propuesto presenta la concatenación de los algoritmos de las etapas de análisis y reconocimiento de voz esofágica con los algoritmos de la etapa síntesis de voz esofágica, desarrollados durante este proyecto, para generar un sistema completo que sea la base para un desarrollo posterior en DSPs con el objetivo de generar un dispositivo independiente que mejore la calidad de la voz de las personas que han sufrido la extracción de sus órganos fonadores.

3.1. Diagrama a bloques del sistema propuesto

El diagrama de la Figura 3.1 presenta de manera esquemática el proceso que sigue la señal de voz esofágica para lograr una mejora significativa en la inteligibilidad y naturalidad de la voz.

Para realizar la simulación del sistema la señal de voz es un archivo grabado de voz esofágica a la cual se le realizan los procesos de análisis, reconocimiento y síntesis. Estos archivos de audio son los que se alimentan a las diferentes etapas y al final la señal de voz sintetizada será también un archivo de audio pero modificado de acuerdo a los procesos mencionados en el diagrama a bloques. El sistema propuesto se encuentra en la etapa de simulación, posteriormente cuando los algoritmos de simulación hayan sido validados y depurados se podrá proceder a implementarlos en un DSP para hacer este sistema totalmente portátil.Para realizar la simulación del sistema la señal de voz es un archivo grabado de voz esofágica a la cual se le realizan los procesos de análisis, reconocimiento y síntesis. Estos archivos de audio son los que se alimentan a las diferentes etapas y al final la señal de voz sintetizada será también un archivo de audio pero modificado de acuerdo a los procesos mencionados en el diagrama a bloques. El sistema propuesto se encuentra en la etapa de simulación, posteriormente cuando los algoritmos de simulación hayan sido validados y depurados se podrá proceder a implementarlos en un DSP para hacer este sistema totalmente portátil.

3.1.1. Análisis

En la primera parte del sistema, el archivo de audio entra a la etapa de análisis, esta etapa separa los segmentos vocalizados de los segmentos no vocalizados mediante el análisis de energía de los formantes. Para realizar este procedimiento este algoritmo en primer lugar divide el archivo de voz tomando segmentos de 1000 muestras, es decir, la longitud de cada segmento es de 125 mseg. Cada segmento de 125 mseg se divide

en segmentos mas pequeños de 12.5 mseg para poder realizar el análisis por LPCs. La Figura 3.2 muestra un diagrama de este proceso.

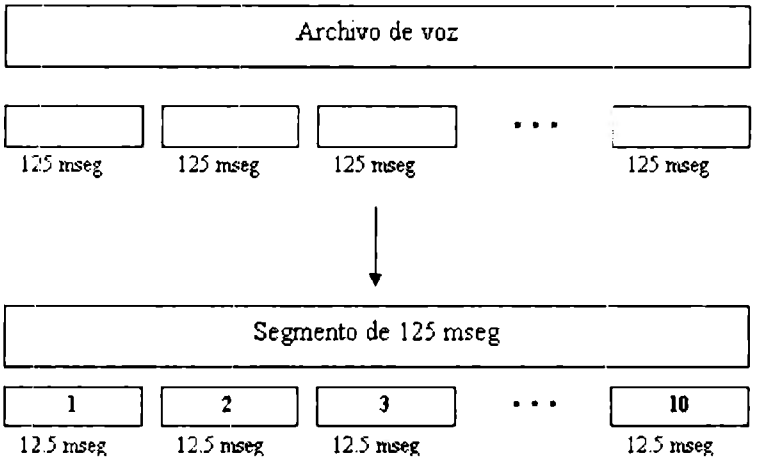

Figure 3.2: Proceso de segmentación.

Entonces se tienen 10 segmentos de 12.5 mseg a los cuales se les va a extraer los coeficientes LPCs. Se extraen 12 coeficientes para cada segmento, posteriormente a partir de dichos coeficientes se calculan los formantes, en total se obtienen 6 formantes, esto de acuerdo a la teoría que establece que se tiene 1 formante por cada 2 coeficientes LPC. De los 6 formantes calculados solo se toman los 3 primeros ya que estos son los que tienen la mayor amplitud y entregan la información suficiente para realizar el análisis. La Figura 3.3 muestra de forma esquemática el proceso de extracción de los formantes mencionado.

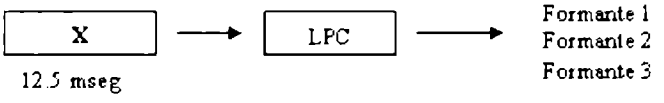

Figure 3.3: Proceso de extracción de formantes

Una vez que se tiene los 3 formantes para todos los 10 segmentos de 12.5 mseg se seleccionan los formantes que tengan la mayor amplitud haciendo un análisis con la transformada rápida de Fourier pero ahora para todo el segmento de 125 mseg, esto es

para determinar si se encontró un sonido vocalizado. Para determinar esto, se compara la amplitud de los formantes con respecto a un umbral, si la amplitud de los formantes es mayor que el umbral entonces ese segmento es un segmento vocalizado. La Figura 3.4muestra un diagrama de la elección del valor máximo para la amplitud de los formantes en el segmento de 125 mseg.

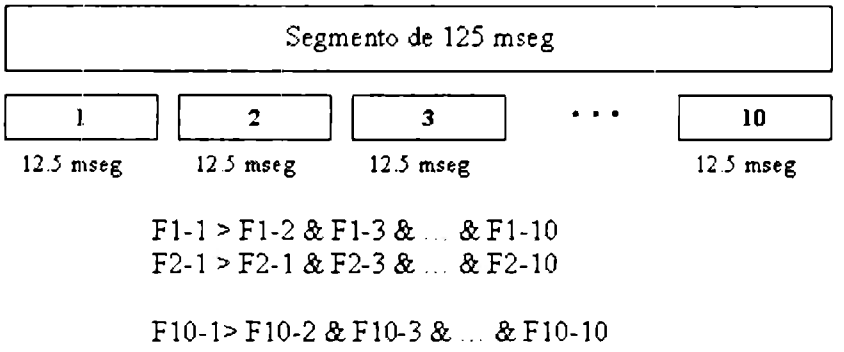

Figure 3.4: Localización del valor máximo del segmento

Este análisis se repite para toda la señal de voz y de esa manera se extraen los sonidos vocalizados de toda la señal de voz. Para separar los sonidos vocalizados de los no vocalizados, se le aplica una mascara al archivo de voz en el que se le asigna un valor de 1 a los sonidos vocalizados encontrados y un valor de 0 a los sonidos no vocalizados. En la Figura 3.5 se muestra la aplicación de la máscara lógica para los segmentos vocalizados. Los sonidos vocalizados se guardan en un vector que va a pasar a la etapa de reconocimiento. Los sonidos no vocalizados se guardan en otro vector que se pasa directamente a la etapa final de síntesis en donde se le van a insertar los segmentos vocalizados sintetizados.

3.1.2. Reconocimiento

El vector de sonidos vocalizados se entrega como parámetro de entrada a la etapa de reconocimiento y este se considera como un vector de características. En esta etapa se vuelve a seccionar el vector de sonidos vocalizados en segmentos de 1000 muestras. Cada uno de estos segmentos se subdivide a su vez cada 200 muestras. Estos segmentos de 200 muestras son analizados para reconocer el tipo de segmento vocalizado presente. La Figura 3.6 muestra el proceso de segmentación del vector de sonidos vocalizados.

Como el reconocimiento de los sonidos vocalizados se realiza mediante Modelos Ocultos de Markov, primero se debe entrenar el algoritmo, una vez que esta entrenado se procede

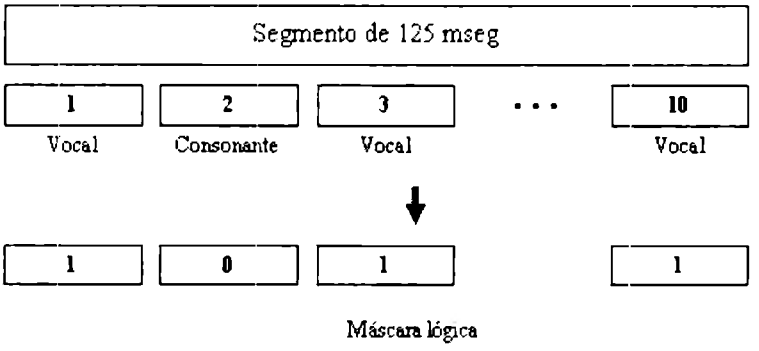

Figure 3.5: Mascara lógica para obtener el vector de segmentos vocalizados

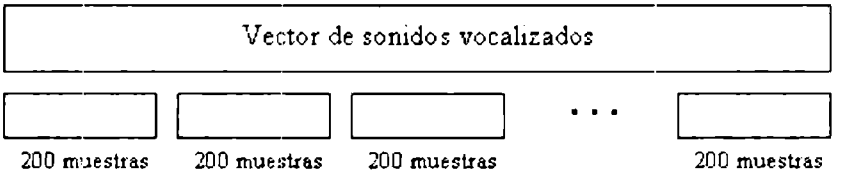

Figure 3.6: Segmentación del vector de sonidos vocalizados

a realizar el reconocimiento con los segmentos del vector de características.

En cada segmento de 200 muestras se calculan los coeficientes LPC, estos coeficientes se cuantizan linealmente, es decir se normalizan a un valor máximo y mínimo para que se tenga una muestra de coeficientes que este dentro de un rango y se puedan comparar de forma adecuada con los vectores de observación en el siguiente proceso de reconocimiento. Los coeficientes LPC cuantizados se guardan en un codebook de 64 elementos, esto genera un vector de observación que se va a comparar con el modelo de Markov. Este proceso descrito se muestra en la Figura 3.7.

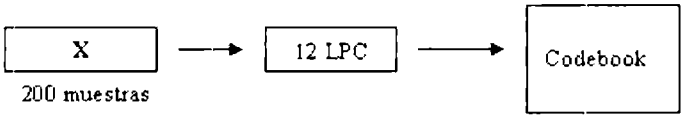

Figure 3.7: Generación de los vectores de observación

Este vector se compara con las ventanas del modelo de Markov. Estas ventanas contienen los parámetros óptimos después del entrenamiento y a los cuales se les ha aplicado el mismo procedimiento de normalización mencionado antes. La comparación arroja la probabilidad de que cada segmento pertenezca a cierto modelo, se asigna así a ese segmento el valor de la vocal que tuvo una probabilidad mayor. En caso de que no se reconozca ninguna de las 5 vocales no se asigna ningún valor. En la Figura 3.8 se observa el chequeo de la probabilidad.

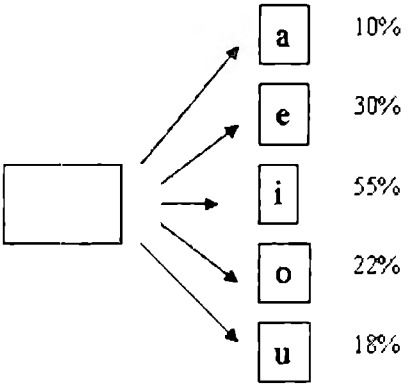

Figure 3.8: Probabilidad de que cada segmento pertenezca a una vocal

Una vez que se tiene el segmento vocalizado reconocido se incrementa un contador de igualdad de la vocal reconocida. Este proceso se repite para cada uno de los segmentos de 200 muestras. Posteriormente para saber cual fue la vocal reconocida para el segmento de 1000 muestras se analizan los contadores de cada vocal y el mayor va a determinar el tipo de vocal que se reconoció para todo el segmento de 1000 muestras.

Así al final de la etapa de reconocimiento se tiene una vocal reconocida por cada segmento de 1000 muestras del vector de sonidos vocalizados. Estas vocales se guardan a su vez en un vector de reconocimiento que se entrega a la etapa de síntesis.

3.1.3. Síntesis

Después de que se reconoció el tipo de segmentos vocalizados que están presentes en la señal de voz, se inicia la etapa de síntesis de dichos segmentos vocalizados. La etapa de síntesis se divide en dos bloques importantes, y son el modelo de la fuente de excitación y el modelo del tracto vocal. En la figura 3.9 se muestra un diagrama simplificado de la síntesis de voz.

Figure 3.9: Diagrama a bloques del proceso de síntesis de voz

Para modelar la fuente de excitación se probaron varios métodos que ya se mencionaron en el marco teórico. Los métodos que modelan la fuente de excitación como una señal periódica cuya frecuencia depende del período fundamental de la voz a sintetizar como el modelo LF y el modelo de Rosenberg nos entregan un segmento de voz sintetizada cuya característica es un sonido monótono y metalizado, con buena inteligibilidad pero su naturalidad se ve comprometida debido a que no considera información en el espectro de frecuencias. En la Figura 3.10 se muestra la forma de las pulso.

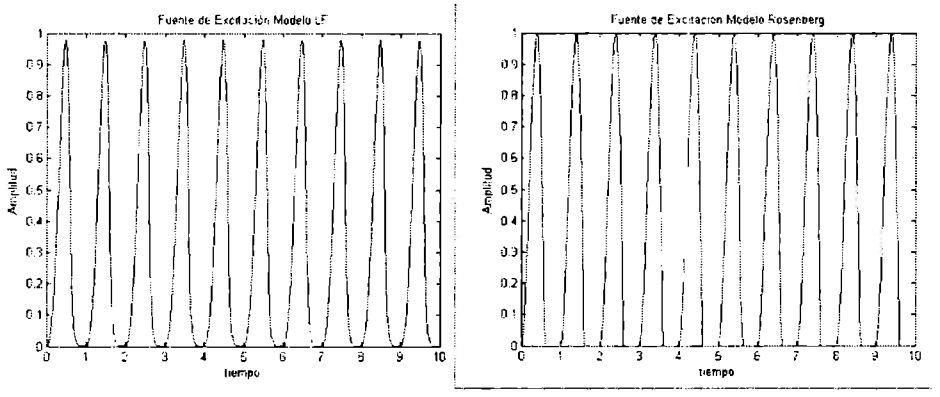

Figure 3.10: Formas de onda de los modelos LF y Rosenberg de fuente de excitación

Por otra parte se analizó la señal residual, la cual mejora notablemente la naturalidad de la voz sintetizada ya que una de las características de esta señal es que tiene un espectro plano, es decir, tiene un número importante de componentes frecuenciales que al momento de alimentarse al filtro del tracto vocal generan una señal de voz mucho mas parecida a la del hablante original. Debido a esto se decidió usar la señal residual para modelar la fuente de excitación. En la figura 3.11 se ve la señal residual que se extrajo del segmento vocalizado 'o'.

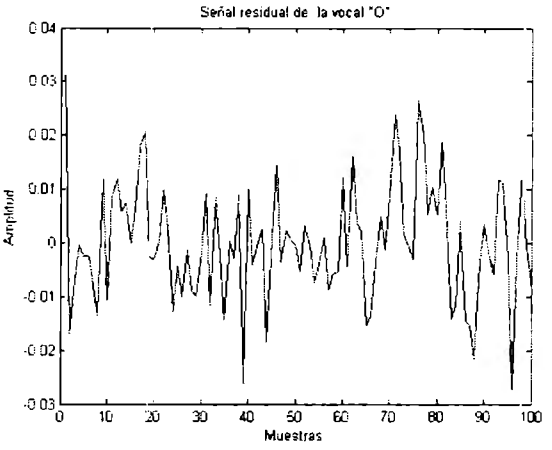

Figure 3.11: Señal residual

El vector de sonidos vocalizados que viene de la etapa de síntesis se secciona en segmentos de 1000 muestras para obtener la señal residual de cada uno de los segmentos. La señal residual se obtiene filtrando el segmento vocalizado con un filtro inverso, los coeficientes del filtro son los coeficientes LPC del segmento vocalizado que se obtuvieron en la etapa de análisis. Esta señal residual es la señal que se va a alimentar al filtro del tracto vocal. Este procedimiento se muestra en la figura 3.12.

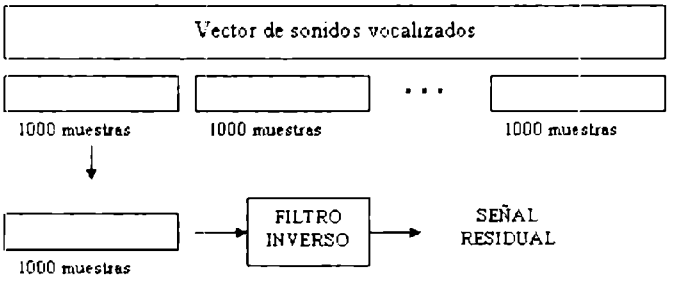

Figure 3.12: Extracción de la señal residual en los segmentos vocalizados

En el bloque del filtro del tracto vocal se usó el modelo de formantes, la ventaja de usar este modelo es que se reduce el número de variables del filtro, esto debido a que solo se usan 3 formantes por cada filtro de cada vocal. En esencia lo que se tiene es un codebook con las frecuencias formantes de cada vocal. Las frecuencias formantes almacenadas en

el codebook se extrajeron haciendo el análisis sobre segmentos vocalizados de hablantes normales, esto para mejorar las características del filtro en cuanto a ganancia y para tener una mejor característica en frecuencia.

El filtro del tracto vocal entonces se modela como un filtro solo-polo donde los coeficientes del filtro son las frecuencias formantes de los segmentos vocalizados. La información acerca del segmento vocalizado que se va a sintetizar se toma del vector de vocales que entrega la etapa de reconocimiento. Con esta información el procedimiento para la síntesis se realiza de la siguiente manera, se toma el primer segmento de 1000 muestras en donde se encontró un segmento vocalizado, se calcula la señal residual, luego se toma la vocal correspondiente del vector de reconocimiento y se extraen los formantes del codebook para modelar el filtro del tracto vocal, finalmente se filtra la señal residual y el segmento vocalizado sintetizado se guarda en un vector de voz sintetizada. Ver figura 3.13.

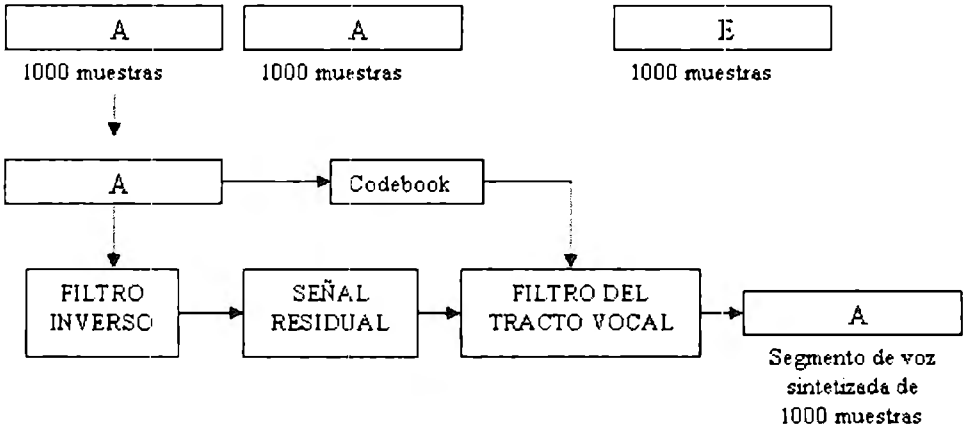

Figure 3.13: Diagrama a bloques del proceso de síntesis

Finalmente para reconstruir la señal de voz se suman los segmentos vocalizados sintetizados con los segmentos no vocalizados, como el tamaño de los segmentos vocalizados sintetizados se asegura que es del mismo tamaño de los segmentos vocalizados que se separaron en la etapa de análisis, al momento de insertar la voz sintetizada simplemente se suman los dos vectores y se tiene la señal de voz reconstruida.

4 RESULTADOS

El sistema completo de Síntesis de Voz Esofágica se desarrolla dentro del código "VozEsofagica.m" de Matlab. El código sigue el proceso descrito en el diagrama de bloques presentado en la sección de Sistema Propuesto. Primero se lleva a cabo el análisis del vector de audio, segundo el sistema de reconocimiento no entrara en funcionamiento si el análisis no se ha llevado a cabo y por último se lleva a cabo el procedimiento de síntesis de voz esofágica. Es necesario recalcar que los procesos de reconocimiento y síntesis se llevan a cabo sólo sobre los segmentos de sonidos vocalizados; los no vocalizados pasan directamente al sumador que tiene como parámetros de entrada los sonidos no vocalizados y los segmentos vocalizados que han sido objeto del proceso de reconocimiento y síntesis.

A continuación se describe y se muestran los resultados obtenidos para cada etapa del sistema completo.

4.1. Etapa de Análisis

La etapa de Análisis es la parte inicial de todo el proceso de Síntesis de Voz Esofágica y el desempeño óptimo de todo el sistema en parte depende de una caracterización de sonidos vocalizados y no vocalizados lo más exacta posible, pues un correcta o incorrecta decisión en cuanto a que tipo de sonido que se ha analizado afectará las etapas siguientes del sistema.

Las funciones que realizan el proceso de análisis se encuentran dentro de la función "analisis.m" de Matlab (ver Anexo).

Los factores que afectan la salida de la etapa de Análisis son los siguientes:

1. El tamaño del segmento a analizar.
2. El número de formantes
3. El umbral de selección.

El tamaño del segmento a analizar actúa de manera directa en la cantidad de información que se debe analizar. Se debe recordar que se toman segmentos de tiempo muy

pequeños con el objetivo de que pueda hacer una mejor caracterización de la señal. Si los segmentos son muy largos no se podrá obtener de manera exacta los parámetros que representan la señal.

En cuanto al número de formantes, se encuentra determinado directamente del orden de LPC que se desea obtener, pero dentro del código de "analisis.m" se toman en cuenta solo los tres primeros formantes, esto debido a que la teoría de los formantes nos especifica que la mayor concentración de la energía de los sonidos se encuentran en las frecuencias menores, además nuestro método para determinar si un sonido es vocalizado o no vocalizado se encuentra directamente ligado a una combinación de formantes y la energía máxima de los formantes de la señal. Esto se puede ver directamente en la Figura 4.1, en la cual se puede observar que la envolvente de la señal de audio que se representa en el tiempo en la Figura 4.2, tiene amplitudes mayores en las frecuencias resonantes .

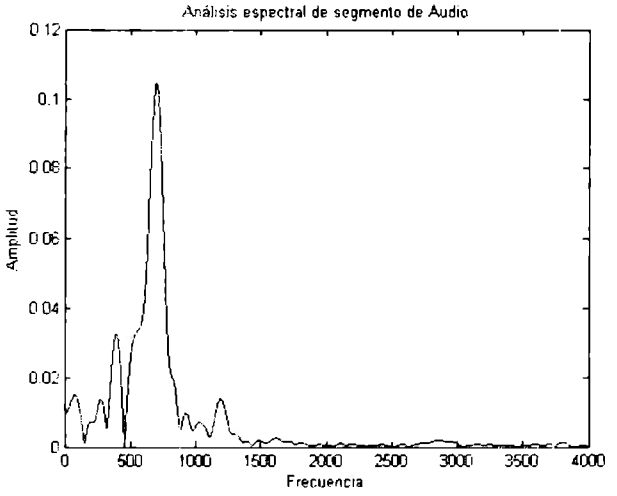

Figure 4.1: Formantes del segmento de voz correspondiente a una "A".

De esta manera se puede demostrar que con al menos 3 formantes se puede representar y construir una señal de audio de muy buena calidad, pues una gran cantidad de información se encuentra dentro de estos primero formantes y si se desea reconstruir una señal esto es posible tomando solo estos 3 primeros formantes. Si se tomaran menos formantes todavía se podrá reconstruir la señal pero no de manera muy exacta; y si se tomaran más formantes es posible obtener una señal más exacta pero el tiempo de procesamiento aumentará considerablemente, pues la cantidad de segmentos que se tienen que analizar es grande dependiendo del tamaño del vector de audio.

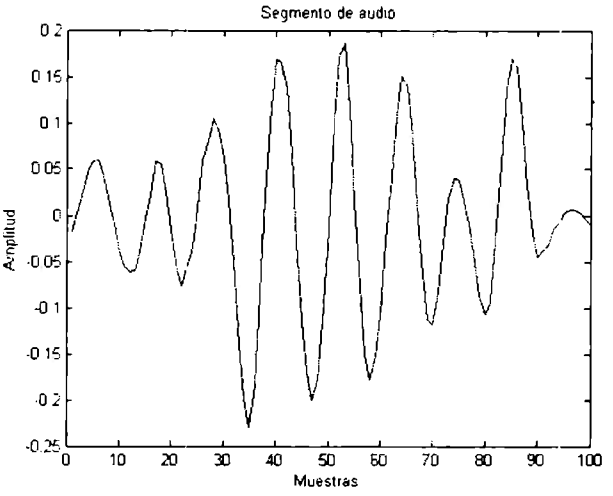

Figure 4.2: Señal de audio del segmento correspondiente a la "A".

Un factor que es el que más fuerte determina el tipo de sonido que se ha analizado es el umbral de selección. Para un hablante normal se puede tomar como un sonido vocalizado si tiene arriba de un 20% de la energía de la señal; pero para un hablante de voz esofágica dicho umbral debe variar, pues los sonidos emitidos por el paciente no suelen entrar dentro del criterio del 20% de la energía, por lo que se considera necesario aumentar dicho umbral. El aumentar el umbral tiene como objetivos primordiales, el no permitir que sonidos no vocalizados muy ruidosos se consideren como vocalizados, y el que se pueda discernir correctamente entre sonidos que no se tienen considerados dentro de la etapa de reconocimiento (como algunas consonantes) y las vocales. El como afecta el umbral la salida de la etapa del bloque de Análisis se muestra en la Figura 4.3, Figura 4.4 y en la Figura 4.5 las cuales corresponden a la palabra BOCA de un paciente de voz esofágica.

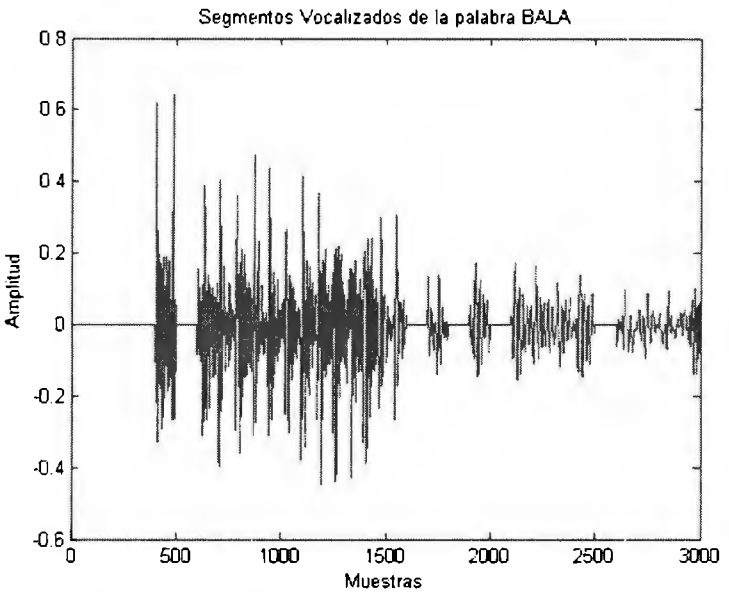

Figura 4.3: Segmentos Vocalizados para un umbral del 10 %.

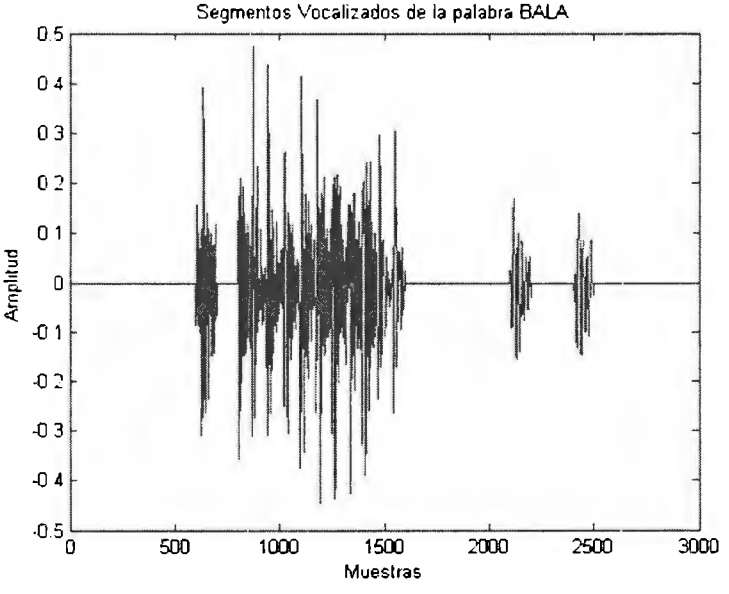

Figura 4.4: Segmentos vocalizados cuando el umbral es del 29 %.

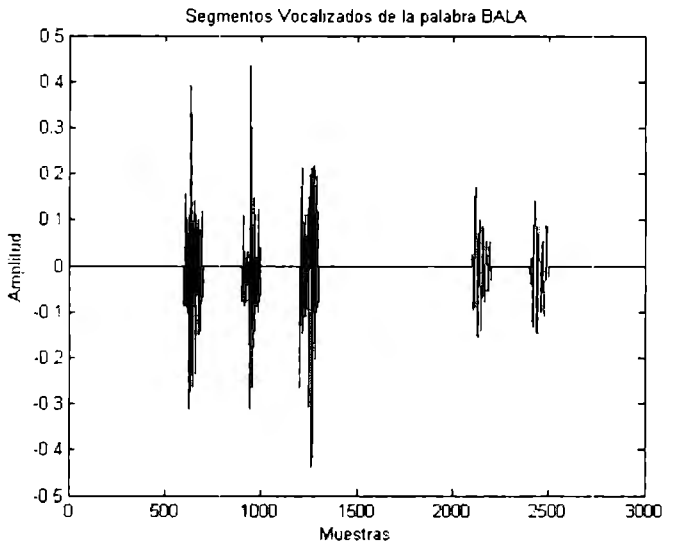

Figura 4.5: Segmentos vocalizados con un umbral del 45 %.

Como se puede observar en las figuras anteriores, si el umbral es menor al 20 % se puede permitir que sonidos no vocalizados e inclusive ruido pasen a la etapa de reconocimiento lo cual puede arrojar resultados erróneos y la etapa de síntesis se verá bastante afectada pues esta fuertemente ligada a los resultados que le provee el bloque de reconocimiento.

Cuando se aumenta el umbral es posible obtener una mejor caracterización del vector de audio en sonidos vocalizados y no vocalizados, pero nuestro sistema de reconocimiento debe aceptar solamente sonidos vocalizados correspondientes a las vocales y existen algunas consonantes que también corresponden a los sonidos vocalizados que cruzan el umbral y se envían al bloque de reconocimiento, lo cual provoca que el reconocedor entre en conflicto.

Cuanto mayor sea el umbral menor será la probabilidad de que se transmita ruido y sonidos vocalizados de algunas consonantes, pero también se da el caso de que parte de la señal que corresponde a los sonidos vocalizados de las vocales no pase al bloque de reconocimiento y se puede ver afectada la salida.

Con base en las conclusiones anteriores y valorando de manera subjetiva el umbral se ha fijado en un 29% y con un tamaño de 100 muestras para cada segmento que se va a analizar. Las 100 muestras representan 12.5 milisegundos del archivo de audio. El número de formantes se ha fijado en 3 formantes por segmento debido a que se puede obtener una buena caracterización espectral con solo 3 formantes.

4.2. Etapa de Reconocimiento

La etapa de Reconocimiento se basa en los Modelos Ocultos de Markov. La principal característica de los modelos de Markov que queremos aprovechar es que el sistema reconocedor aproxima la muestra de observación a un determinado modelo, pero solo si previamente a hecho una comparación con los otros modelos. Debido a que los Modelos Ocultos de Markov utiliza fundamentos de probabilidad es susceptible de errores, por lo cual se tendrá que aplicar ciertas consideraciones para que la probabilidad de acierto sea mayor.

El procedimiento para desarrollar la etapa de reconocimiento se puede ver en el código de Matlab "procedure.m" que se encuentra dentro del código "VozEsofagica.m", en el cuál se muestra la secuencia preescrita para un sistema reconocedor basado en Modelos Ocultos de Markov.

En la primera parte del código se debe comprobar si el sistema ya se encuentra entrenado. El entrenamiento se refiere a calcular los modelos que servirán como base de comparación, es decir la referencia con respecto a los vectores de observación (en nuestro caso los sonidos vocalizados enviados por el bloque de análisis). Se ha mencionado que entre mejor sea el modelo para cada vocal en particular mejor será el sistema reconocedor. El que un modelo sea bueno dependen de la cantidad de vectores de audio con el que fue generado cada modelo. Si el número de vectores de entrenamiento es más grande y con más variaciones referentes al sonido del cual se desea generar el modelo, el modelo será mucho mejor.

Después de comprobar que se han generado los modelos para cada vocal se procede a recibir los vectores de observación que provienen del bloque de análisis. A continuación se describe gráficamente lo que ocurre dentro del bloque de reconocimiento.

El diagrama de la Figura 4.6 nos muestra el proceso general del sistema reconocedor.

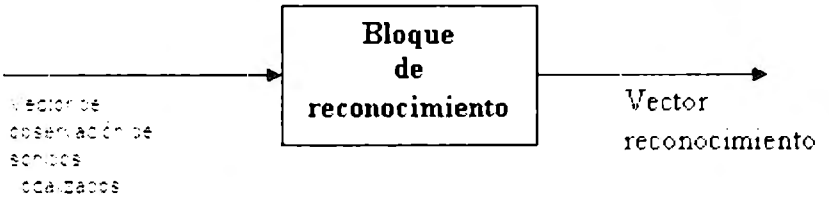

Figure 4.6: Diagrama general de la etapa de Reconocimiento.

Lo que el sistema reconocedor realiza de manera en particular es:

1. 1. Toma segmentos de 1000 muestras del vector de sonidos vocalizados.
2. Se toma el vector de audio de los sonidos vocalizados.Ver Figura 4.7.

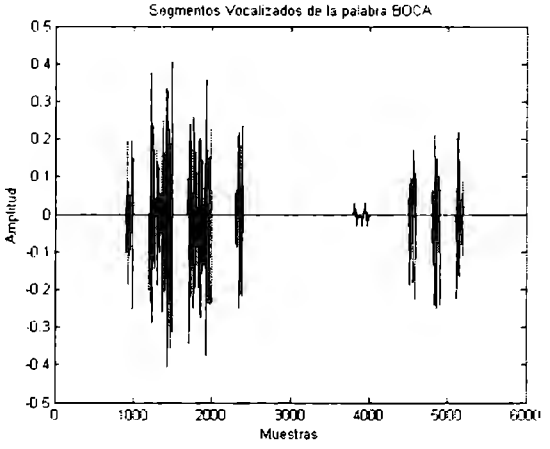

Figure 4.7: Vector de sonidos vocalizados entregado por la etapa de análisis.

3. De los segmentos vocalizados se segmento en muestras de 1000 para ser ingresados en el reconocedor. Ver Figura 4.8.

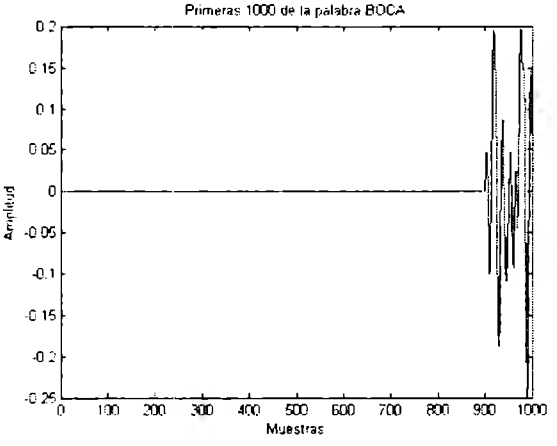

Figure 4.8: Primeras 1000 muestras tomadas del vector de sonidos vocalizados.

4. El reconocedor vuelve a segmentar para tomar los vectores de observación. Cada nuevo segmento corresponde a 200 muestras y dichas muestras son las que toma como vectores de observación.

5. Cada vector de observación se compara con cada uno de los modelos que representan a las vocales. Por medio del algoritmo de 'avance.m' se calcula la probabilidad de que dicho vector de observación corresponde al modelo con el cual se esta comparando.

6. Después de calcular la probabilidad para cada modelo de las vocales se toma la probabilidad mayor la cual se considera que es la que mejor representa al modelo de observación dado.

7. En el punto 5 y 6 se ha descrito el procedimiento para cada segmento de 200 muestras, por lo cual se tendrá que realizar el mismo procedimiento un total de 5 veces para un segmento de 1000 muestras.

8. Después de calcular el modelo más aproximada por cada 200 muestras se tiene que realizar una nueva decisión, la cual consiste en tomar como modelo representativo del segmento de 1000 muestras como aquel modelo que mayor probabilidad tiene dentro del segmento de 1000 muestras.

9. Al final del paso 8 se genera un vector en el cual se coloca el modelo que más se acerca a los vectores de observación dados por cada 1000 muestras.

Una vez que se ha construido el vector que contiene los modelos (la vocal) que mejor representa al vector de observación por cada 1000 muestras se envía al bloque de síntesis para que realice la síntesis de cada uno de los sonidos vocalizados reconocidos.

Los modelos para cada vocal en particular fueron generados a través de la caracterización por formantes pues es el que mejores resultados de reconocimiento nos arroja.

Algunos de los ejemplos que se tiene para el bloque de reconocimiento son los siguientes:

En la tabla 4.1 se muestran las vocales reconocidas para cada archivo de audio. Cada vocal presente representa un segmento de observación de 1000 muestras. En algunas de las palabras de la tabla 4.1 se puede ver que algunos segmentos se han reconocido como vocales que no pertenecen a la palabra que ha ingresado al reconocedor. Esto sucede debido a que como se ha mencionado en la etapa de análisis algunos sonidos vocalizados que sobrepasan el umbral del 20% de la energía no son necesariamente vocales, por lo cual dichos consonantes vocalizadas si sobrepasan el umbral y entran al reconocedor. En

Palabra de Voz Esofágica	Vector de reconocimiento
BECERRO	_E_AEOOO
BOCA	OOAAAA
ACA	I_A_IA
ABRAZAR	UOAAUAAUA
ABRA	OAAAIAI
BALA	AAA

Table 4.1: Vectores de reconocimiento de voz esofágica.

la teoría de los modelos ocultos de Markov se ha mencionado que el modelo calcula la probabilidad de que el vector de observación se acerque lo más posible a un determinado modelo, por esta razón el sistema reconocedor cuando se encuentra enfrente de un sonido vocalizado lo aproxima a un determinado modelo y lo interpreta como una vocal.

Como se ha mencionado en la sección de la etapa de Análisis, debido a que el umbral no discrimina correctamente entre sonidos vocalizados de algunas consonantes y las vocales, algunas consonantes sobrepasan el umbral fijado en la etapa de análisis. Aún cuando ésto sucede, se ha observado que las amplitudes, de partes de consonantes o ruido que sobrepasan el umbral, no son significativas con respecto a las amplitudes de las vocales. A pesar de que dichos sonidos no deseados se reconocieron como vocales al transmitir el vector de reconcomiento al bloque de síntesis, el bloque de síntesis realizará un procedimiento tal que minimizará los efectos de estas vocales que de manera incorrecta fueron reconocidas.

4.3. Etapa de Síntesis

La última etapa del sistema es el de Síntesis el cual se encargará de revisar el vector de reconocimiento para poder realizar la síntesis de cada segmento vocalizado que haya sido enviado por la etapa de análisis.

El procedimiento para la Síntesis de los segmentos vocalizados se puede ver en el código "conexion.m".

La etapa de Síntesis recibe parámetros de los bloque de análisis y reconocimiento. Los paramétros para el correcto funcionamiento del bloque de síntesis son:

- Vector de segmentos vocalizados.
- Vector de segmentos no vocalizados.

- Máscara de segmentos vocalizados.
- Vector de reconocimiento.

Enumerando los pasos para el sistema reconocedor se puede ver de que manera se hace uso de cada uno de los paramétros mencionados.

1. Se modificó el código de "análisis.m" para que pudiera construir un vector de máscara para todo el archivo de audio. El vector indica por medio de 0s y 1s los sonidos no vocalizados o vocalizados por cada 100 muestras de todo el archivo de audio. Un 0 representa un segmento no vocalizado y a este segmento no se le realizará ningún procesamiento. Un 1 representa un segmento vocalizado y este segmento es el que se utiliza para la generación de la fuente de excitación.

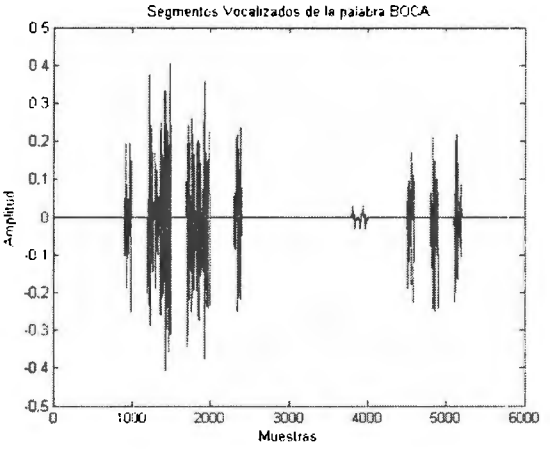

Figure 4.9:

2. Con ayuda de la máscara se toman sólo los segmentos vocalizados. Del vector de segmentos vocalizados se toman las muestras para poder generar la fuente de excitación. La fuente de excitación, el cual es una fuente residual, se realiza por medio del código "arcoef.m" el cual recibe como parámetro el segmento vocalizado. La Figura 4.10 nos muestra el proceso que sufre el sonido vocalizado se hace pasar por el filtro de LPCs generados para el segmento en particular. La señal de salida, la fuente de excitación residual, contiene las características en frecuencia, intensidad

de la señal y además de la inflexión de la voz necesaria para que el sonido sea más natural. La señal residual generada se muestra en la Figura 4.11.

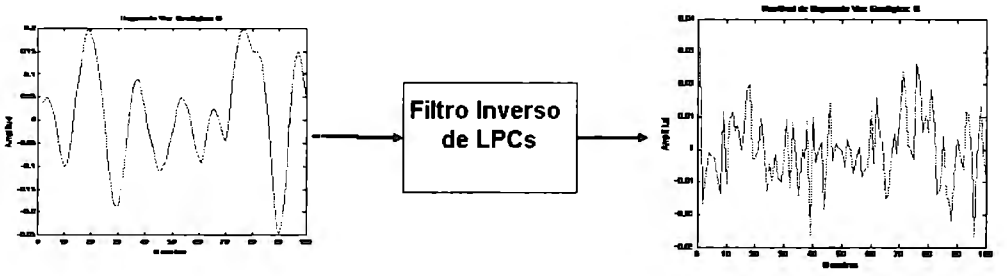

Figure 4.10: Proceso para generar la fuente de excitación.

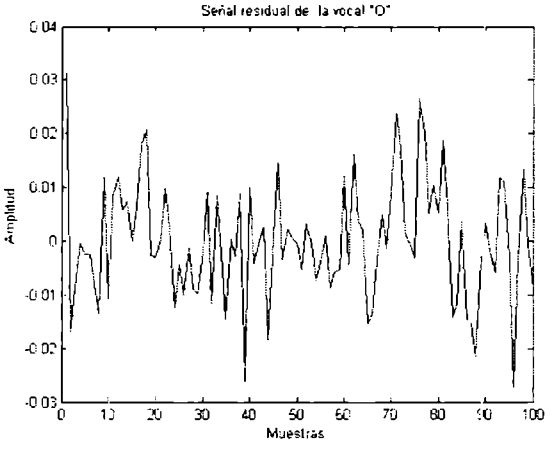

Figure 4.11: Señal residual de la vocal 'O'.

3. La señal residual generada para la fuente de excitación se hace pasar por un filtro pasa bajas que tiene una frecuencia de corte de 1 Khz con el objeto de suavizar los picos de la señal y tener una fuente de excitación que nos ayude a mejorar los efectos de coarticulación entre los segmentos adyacentes, pues con una señal residual que no ha sido filtrada contiene altas frecuencias que afecta la coarticualción. La señal residual después del filtro se puede ver en la Figura 4.12.

4. Con ayuda de la máscara se puede determinar en que segmento de 1000 muestras se encuentra el sonido vocalizado. De esta manera al conocer la posición de la vocal dentro del vector de reconocimiento se procede a extraer del codebook los

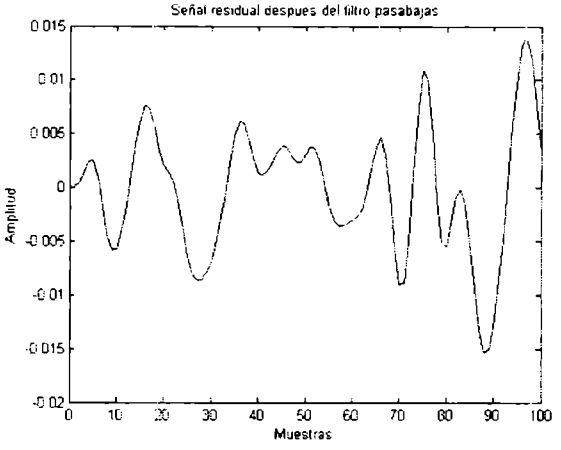

Figure 4.12: Señal residual a la salida del filtro pasabajas.

formantes requeridos para sintetizar la vocal que nos da el vector de reconocimiento. Un ejemplo de modelo de tracto vocal se puede ver en la Figura 4.13 . El tracto vocal se genera a partir de las frecuencias resonantes presentes en el codebook, el cual son diferentes para cada vocal. Los frecuencias resonantes dentro del codebook pertenecen a los formantes de las vocales de un hablante normal.

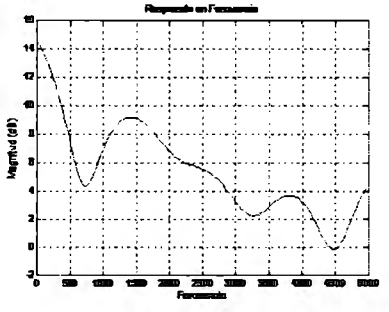

Figure 4.13: Modelo de Tracto Vocal por formantes.

5. Una vez que se tiene la señal de excitación y el filtro que represental el Tracto Vocal, se hace pasar la señal residual por el filtro de Tracto Vocal para generar una señal sintetizada que tendrá mejores características en intensidad y energía que la señal de voz esofágica. Debido al filtro de formantes la salida de la señal tiene una ganancia grande comparada con la segmento vocalizado que había a la entrada

lo cual al conjuntar la palabra completa producirá variaciones de amplitud muy grandes que provoca que la palabra sintetizada suene ruidosa. Por esta razón es necesario reajustar la amplitud de la señal sintetizada. Una vez que la amplitud ha sido ajustada se procede a reintegrar la señal sintetizada en la posición original de la que fue extraída. En la Figura 4.14 se puede ver el segmento vocalizado antes de pasar por el proceso de síntesis y en la Figura 4.15 se puede ver lo que resulta del proceso de síntesis.

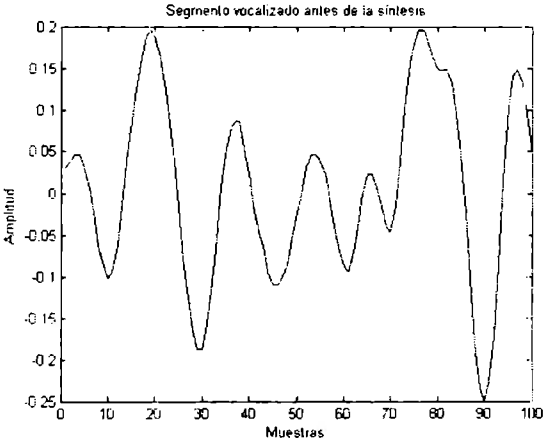

Figure 4.14: Segmento vocalizado antes del proceso de Síntesis.

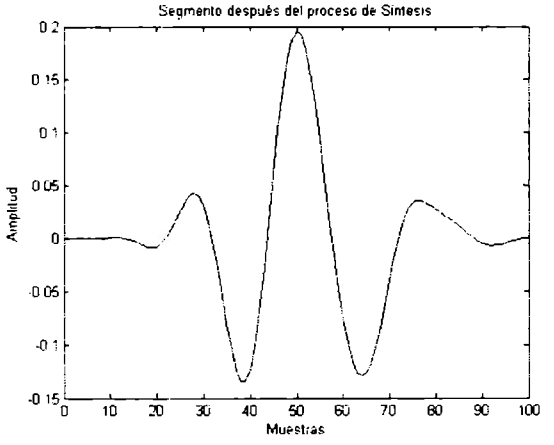

Figure 4.15: Segmento vocalizado después de la Síntesis.

6. Una vez que ya se han sintetizado los segmentos vocalizados provenientes de la etapa de análisis se procede a realizar la suma entre los segmentos vocalizados sintetizados y los segmentos no vocalizados para reconstruir la palabra original pero con las modificaciones realizadas para los segmentos vocalizados. La Figura 4.16 nos muestra la señal de salida que se tiene a la salida del bloque de síntesis. Pero la diferencia entre esta señal sintetizada y la señal de voz esofágica entrante no es apreciable en el dominio del tiempo. Por esta razón se procede a realizarse un análisis espectral para determinar las diferencias. En la Figura 4.17 se observa la señal de salida en el dominio espectral.

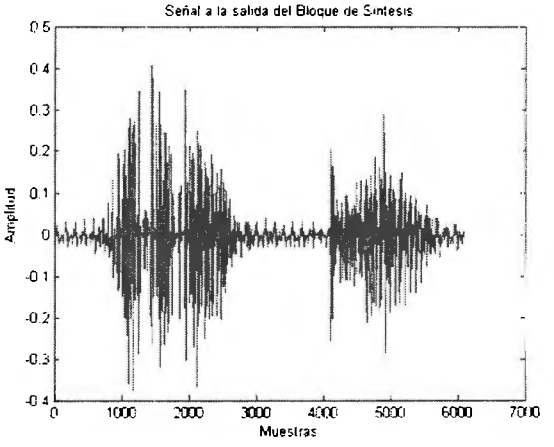

Figure 4.16: Señal a la salida del bloque de Síntesis.

7. Como se puede ver la Figura 4.17, existen ciertas componentes en las bajas frecuencias que al escuchar el audio de la señal sintetizada se escucha como ruido. Este ruido es provocado por los segmentos no vocalizados a los cuales no se les realizó ningún procesamiento, por esta razón se procede a filtrar nuestra señal de salida por medio de un pasabandas para limitar el ancho de banda de nuestra señal y tratar de limpiar nuestra señal. En la Figura 4.18 se observa el resultado del filtro pasabandas.

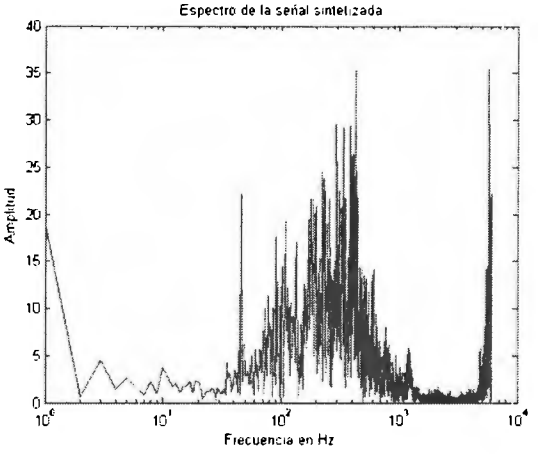

Figure 4.17: Espectro de la señal a la salida del bloque de Síntesis.

4.4. Evaluación de la calidad

Realizar una evaluación de la calidad de la voz de un sintetizador ó codificador de voz es una tarea difícil puesto que no existen fórmulas o cálculos matemáticos que puedan proporcionar una evaluación de la calidad de la voz sintetizada. El problema recae en que la calidad de la voz está relacionada inherentemente a la percepción de la misma. La calidad percibida y la comprensión de la señal de voz sintetizada dependen de un gran número de condiciones entre las que se incluyen el contenido de voz, la individualidad de los hablantes, el ruido de fondo y la persona que percibe la voz.

Debido a que la gran mayoría de los sistemas sintetizadores de voz hacen uso de sus cualidades perceptibles, de manera general las mediciones de calidad subjetivas basadas en pruebas de audición resultan más relevantes que las medidas objetivas tal como la SNR.

Las pruebas de calidad intentan calificar que tan bueno es un sonido de voz sintetizado relacionado a la presencia o ausencia de factores que degraden la señal sintetizada. Estas pruebas de calidad son totalmente subjetivas, y los resultados pueden variar significativamente dependiendo del tipo de persona que emite la voz y de la que la escucha.

Para evaluar la calidad del sistema de síntesis de voz de este proyecto se realizó la prueba de MOS[1] , esta prueba avalúa de manera general la opinión de un grupo de personas en donde se califica una palabra sintetizada basándose en una escala fija. Dicha escala se encuentra en un rango de 1 a 5. Cada nivel describe la calidad de la voz de

[1]Mean Opinión Score: Calificación de opinión promedio.

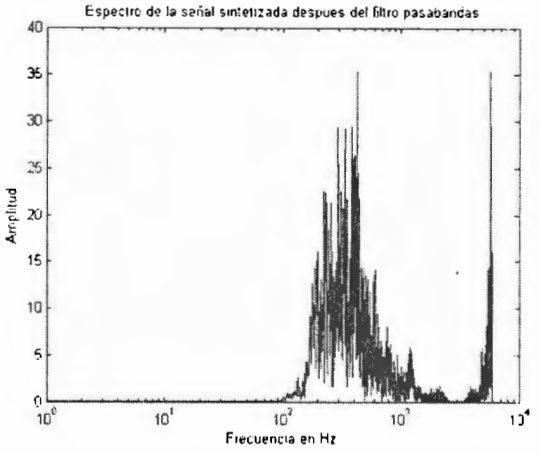

Figure 4.18: Señal después del filtro pasabandas.

acuerdo a la siguiente relación:

- 5 = Excelente
- 4 = Buena
- 3 = Aceptable
- 2 = Pobre
- 1 = Inaceptable

El proceso para realizar la evaluación se describe a continuación. Se tomó una muestra de 25 personas, 10 hombres y 15 mujeres con rangos de edades entre los 18 y 25 años de edad. Se formaron 5 grupos de 5 personas a las que se les hizo escuchar un total de 30 palabras presentándoles en primer lugar la voz esofágica y posteriormente la voz sintetizada, después de escuchar cada palabra sintetizada se les pidió que evaluaran la calidad de la voz con respecto a la voz esofágica de acuerdo al rango de calificaciones mencionado antes. En la tabla siguiente presentamos las calificaciones promedio de cada grupo de encuestados y el promedio de cada palabra. Finalmente se generó el promedio general para calificar el desempeño general del sistema.

En las siguientes gráficas se muestra como varió la percepción de cada palabra para cada grupo de encuestados. Estas gráficas demuestran las variaciones que se pueden tener debido a la forma en la que un distinto grupo de personas perciben la voz.

Evaluación MOS						
	Grupo de Encuestados					
Palabra	1	2	3	4	5	Calificacion Promedio
abajo	4.6	4.2	4.4	4.2	4	4.28
abolir	3.4	4	4.4	3.6	3.2	3.72
abono	4.2	3.8	4	4.4	4.4	4.16
abra	5	5	4.4	4.8	4.8	4.8
abrazar	3.2	3.2	3	3.6	3	3.2
aca	4.2	4	4.4	4.2	4	4.16
acero	4.4	4.4	4.2	4	4.2	4.24
acervo	3.4	3.2	3.6	3.2	3.2	3.32
acolito	5	5	4.6	5	4.2	4.76
adios	4.4	4.2	4.4	4.8	4.6	4.48
adobe	4	4.2	4.4	4	4.2	4.16
adobo	5	4.8	4.6	5	5	4.88
alistar	2.6	3	3.4	2.8	3	2.96
bacalao	4.8	5	4.8	5	4.6	4.84
bache	3.2	3	3.4	3	3.4	3.2
bacteria	4	4.4	4	4.2	4	4.12
bala	5	4.6	4.8	5	4.6	4.8
balada	4.8	4.6	5	5	4.8	4.84
bebedero	4	4.2	4	4.4	4	4.12
bebido	3.4	3.6	3	3.8	3.2	3.4
becerro	4	3.6	3.8	4.2	4	3.92
billar	2.6	3	2.8	3	2.6	2.8
boca	4.8	5	5	4.8	5	4.92
bruja	5	4.8	5	4.6	5	4.88
bulbo	4.4	4.2	3.8	4	4.2	4.12
burdo	3.4	3.2	3	3.4	3	3.2
cama	5	5	4.6	4	4.8	4.68
cana	4.6	4.8	5	4.6	5	4.8
foco	3	3.2	3	3	3.2	3.08
foca	4.2	4.8	5	4.6	5	4.72
	Promedio general					4.3

Table 4.2: Calificaciones del grupo de encuestados.

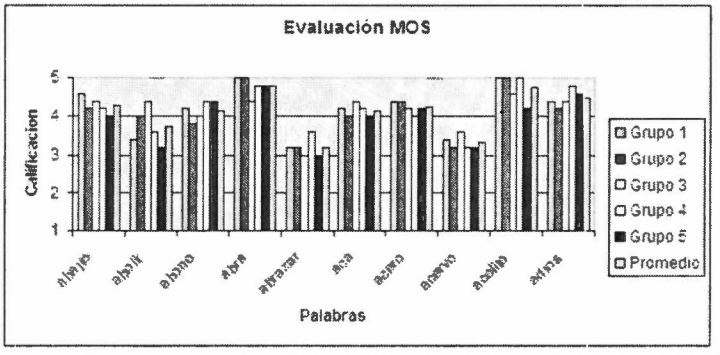

Figure 4.19: Promedio de calificación para cada grupo de personas.

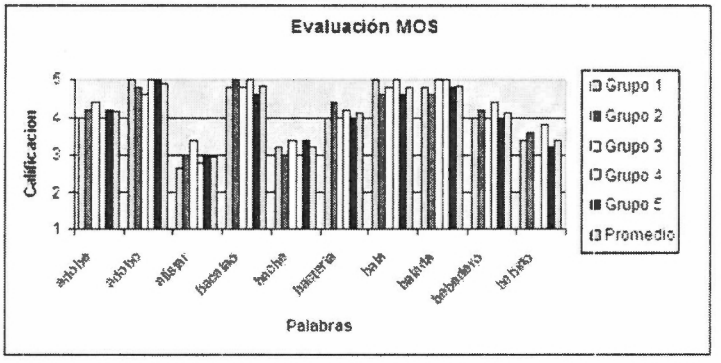

Figure 4.20: Comparación de calificaciones para los 5 grupos encuestados.

En la gráfica de la figura 4.19se observa en las primeras 5 barras las calificaciones promedio para cada palabra de cada uno de los cinco grupos de personas y en la barra final se tiene la calificación promedio de los cinco grupos.

Es evidente que los resultados de la evaluación MOS sufren de una variación significativa dependiendo del grupo de personas. Esto debido a que la opinión en cuanto a la calidad de la voz varía de persona a persona, esto se observa muy claro en la Gráfica de la figura 4.20 y también en la Gráfica de la figura 4.21 que se muestra a continuación y que relaciona las calificaciones de las últimas 10 palabras evaluadas por nuestro grupo de encuestados.

A pesar de las variaciones inherentes a este tipo de evaluación, la calificación promedio general de nuestro sistema se encuentra por encima del rango 4, lo cual para muchos sintetizadores y codificadores de voz se considera una calificación aceptable. Mostramos

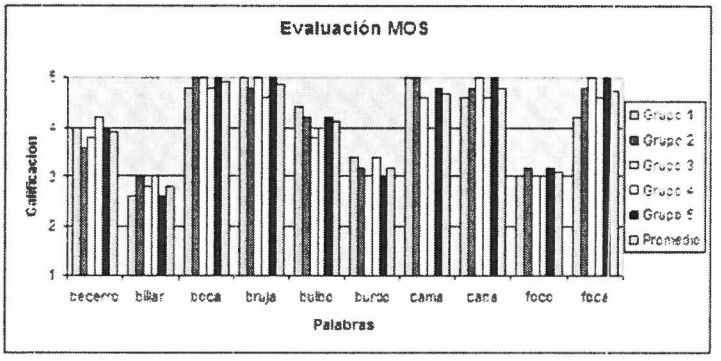

Figure 4.21: Calificaciones promedio del grupo de encuestados.

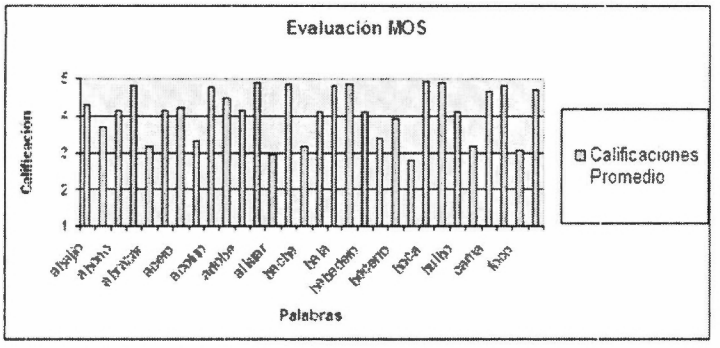

Figure 4.22: Calificaciones promedio de las palabras usadas.

también las calificaciones promedio de la mayoría de las palabras que se sometieron a esta evaluación, se puede ver en la Gráfica de la figura 4.22 que las calificaciones de la mayoría de las palabras se encuentran en el rango de 3 a 5, es decir que se consideraron como por encima de una calidad aceptable pero sin llegar a excelentes.

En términos de porcentaje, en la tabla 4.3 presentamos las calificaciones en donde el valor máximo es el 100 %, esto con el objetivo de tener una evaluación de la calidad y de la confiabilidad de los resultados del sistema de manera general, el resultado se aprecia en la tabla y se puede ver que tenemos una calidad del sistema del 83.6 %. Esta calificación la consideramos aceptable puesto que algunas de las señales de voz que se evaluaron no ofrecieron los resultados esperados debido a que el archivo de voz no era lo suficientemente confiable por cuestiones relacionadas a los métodos de grabación.

Evaluación MOS	
Palabra	Porcentaje
abajo	87.20%
abolir	74.40%
abono	83.20%
abra	96.00%
abrazar	64.00%
aca	86.40%
acero	84.80%
acervo	66.40%
acolito	95.20%
adios	91.20%
adobe	89.60%
adobo	97.60%
alistar	59.20%
bacalao	96.80%
bache	66.40%
bacteria	82.40%
bala	97.60%
balada	96.80%
bebedero	86.40%
bebido	68.00%
becerro	80.00%
billar	56.00%
boca	98.40%
bruja	99.20%
bulbo	87.20%
burdo	72.00%
cama	93.60%
cana	96.00%
foco	61.60%
foca	94.40%
Promedio general	83.60%

Table 4.3: Calificaciones en porcentajes de las palabras sintetizadas.

5 CONCLUSIONES Y TRABAJO FUTURO

En la primera etapa del proyecto se completó la investigación bibliográfica referente a la síntesis de voz. Se consideraron tres tipos de síntesis: Concatenativa, por Formantes y Articulatoria que de acuerdo a la teoría representan los modelos más utilizados tanto en el desarrollo de sistemas como en la investigación en los años recientes. Bajo el mismo esquema de la síntesis por formantes se desarrolló un algoritmo basado en la síntesis por LPC?s la cual ofrece una mejor caracterización de la envolvente de la señal de voz.

En la etapa de síntesis de voz esofágica uno de los factores que determinan de manera predominante la calidad de la voz sintetizada está determinado por los modelos de la fuente de excitación, ya que de estos modelos depende la inteligibilidad y la naturalidad de la voz sintetizada. Mientras que para los modelos de forma de onda periódica se tiene una buena inteligibilidad, la naturalidad del sonido de voz no ofrece buenos resultados. Por otro lado al usar la señal residual se garantiza que la voz sintetizada sea mas parecida a la del hablante ya que esta señal retiene la naturaleza periódica de la señal de voz original así como su período fundamental, sin embargo para el caso de la voz esofágica la señal residual requiere de un procesamiento posterior para mejorar sus características en frecuencia e intensidad.

En cuanto al sistema total consideramos que es necesario proveer una señal de voz mas limpia para alimentar al sistema ya que el ruido introduce errores tanto en el análisis como en el reconocimiento. En la parte del análisis se buscó mejorar la decisión para distinguir los sonidos vocalizados de los no vocalizados variando el umbral, sin embargo si se eleva considerablemente el umbral algunos segmentos vocalizados se omiten y al contrario si se baja el umbral se pueden tomar segmentos que no son vocalizados como si lo fueran y de esta manera introducir un error a las siguientes etapas. De acuerdo a lo estudiado en esta etapa se considera que se puede mejorar el algoritmo de análisis introduciendo otro método de análisis de voz que complemente el método implementado por formantes. En cuanto al reconocimiento los puntos mas importantes que se discutieron fue el tratar de mejorar los vectores de observación que entrenan el modelo de Markov, así mismo se

podría contemplar usar otro método de reconocimiento como las redes neuronales para evaluar su desempeño y determinar que método entrega mejores resultados.

Como trabajo futuro queda el desarrollo de los algoritmos de análisis, reconocimiento y síntesis en otro lenguaje de programación esto para garantizar que pueda ser implementado en un DSP y así generar un sistema que pueda usarse de manera independiente y en tiempo real.

Bibliografía

[1] Mata Daniel, Angeles Carlos, Alvarado Jorge, Cabrera Laura. "Análisis y Reconocimiento de Voz Esofágica". 2005

[2] L. Rabiner y B. Juang. "Fundamentals of Speech Recognition". Prentice May. New Jersey, EU. 1993

[3] R. Goldberg & L.Riek. (2000). A Practical Handbook of Speech Coders. New York: CRC Press.

[4] J. N. Holmes (1988). Speech Synthesis and Recognition. Londres: Chapman & Hall.

[5] K. Matsui, N. Hara. Enhancement of Esophageal Speech using Formant Synthesis. Proc. IEEE ICASSP pp. 81-84 (1999).

[6] R. G. Tull, J. C. Rutledge. Linear Predictive Synthesis of Vowels for Pitch enhancement of Female Geriatric Esophageal Speech. Proc. Conference of the IEEE pp. 1359-1360 (1993).

[7] J. Makhoul, et al. A Mixed Source Model for Speech compression and Synthesis. IEEE ICASSP, pp 163-166, (1978).

[8] A.K. Krishnamurthy. Glottal Source Models for Speech Coding and Synthesis. Proc of the 32nd Midwest Symposium, pp 93-96, (1989).

[9] D. Malah. Efficient Spectral Matching of the LPC Residual Signal. IEEE International Conference ICASSP. pp 1288-1291, (1981).

[10] W. T. Hartwell, D. T. Prezas. A Pulse Driving Function Generator for LPC Synthesis of Voiced Segments of Speech. IEEE Transactions on Signal Processing, pp 1113-1116, (1981).

[11] D. Wong, J. Markel. An Excitation Function For Lpc Synthesis Which Retains The Human Glottal Phase Characteristics. Conference IEEE ICASSP, pp 171-174. (1978).

Anexos: Algoritmos del Sistema de Síntesis

VozEsofagica.m

```
% SÍNTESIS DE VOZ ESOFAGICA

clc

% La variable 'tr' es la bandera que indica si el sistema de
% reconocimiento ya ha sido entrenado. La primera vez que se
inicializa el
% sistema es necesario entrean los modelos Ocultos de Markov.
% Bandera  'tr' en 0 indica sistema sin entrenamiento. Bandera 'tr' en
1
% indica sistema entrenado.

archivo='sounds\adobo';
tr=1;
% El código de analisis recibe como variable el nombre del archivo de
audio
% que se va a analizar. La variable archivo es un string que indica la
ruta
% de acceso del archivo.
    clear X, clear EX, clear EX2, clear letras, clear VOZ;
    [X,EX,EX2,mask]=analisis(archivo);

% Después del análisis se realiza el reconocimiento
    letras=procedure2(EX,tr,X)
% Se realiza la síntesis.
    [V VOZ]=conexion(EX,EX2,X,letras,mask);
```

Analisis.m

```
function [X,EX,EX2,mask]=analisis(archivo)
clc
[X,FS]=WAVREAD(archivo);    %convierte el archivo *.wav a matriz
SEG=1000;             %tamaño del segmento
RF1=1;
TX=length(X)-1;      %tamaño del archivo
DX=[0:TX]/FS;        %dominio del archivo
CLPC=4+FS/1000;      %coeficiente del LPC
TM=100;              %tamaño de la muestra
VENTANA=hamming(TM);
LIM=.29;             %umbral de seleccion
mask=zeros(1,TX/SEG);
[c,d]=butter(5,0.25);

while   RF1<=(TX/SEG)
        Y=X((1+(RF1-1)*SEG):(SEG*RF1)); %segmentacion

        REF1=1;

        while   REF1<=SEG/TM
                SMPL=Y((1+(REF1-1)*TM):(TM*REF1));  %muestreo
                YLPC=lpc(SMPL,CLPC);        %espectro LPC
                ROOTS=roots(YLPC);          %raices del espectro LPC
                ROOTS=ROOTS(imag(ROOTS)>0.011);
                FORM=sort(atan2(imag(ROOTS),real(ROOTS))*FS/(2*pi));
                F1(REF1)=round(FORM(1));    %Formantes 1,2 y 3
                F2(REF1)=round(FORM(2));
                F3(REF1)=round(FORM(3));
                SMPL=SMPL.*VENTANA;          %Ventaneo
```

```
                FFTY=abs(fft(SMPL,FS))*2/length(SMPL);  %Transformada
de Fourier
                aF1(REF1)=FFTY(F1(REF1));   %Amplitud de los formantes
                aF2(REF1)=FFTY(F2(REF1));   %1,2 Y 3
                aF3(REF1)=FFTY(F3(REF1));
                REF1=REF1+1;
        end

        MaF1=max(aF1);      %seleccion de las amplitudes mayores del
segmento
        MaF2=max(aF2);
        MaF3=max(aF3);
        aF1=aF1./MaF1;      %Normalizacion de los valores máximos
        aF2=aF2./MaF2;
        aF3=aF3./MaF3;

        laF1=aF1>LIM;       %Mascara individual por umbral
        laF2=aF2>LIM;
        laF3=aF3>LIM;

        AUX=length(aF1);
        AUX2=laF1&laF2&laF3;    %Mascara para la vocalizacion con AND
        n=1+((RF1-1)*length(AUX2));
        mask(1,n:RF1*length(AUX2))=AUX2;
        REF1=1;
        REF2=1;
        AUX3=~AUX2;

        while   REF2 <=AUX
            while   REF1 <=TM
                OUT(REF1+(REF2-1)*TM)=Y(REF1+(REF2-1)*TM)*AUX2(REF2);
%Aplicacion
                OUT2(REF1+(REF2-1)*TM)=Y(REF1+(REF2-1)*TM)*AUX3(REF2);
%Aplicacion
                %OUT2=filter(c,d,OUT2);
                REF1=REF1+1;
            end

            REF1=1;
            REF2=REF2+1;
        end

        EX((1+(RF1-1)*SEG):(SEG*RF1))=OUT;
        EX2((1+(RF1-1)*SEG):(SEG*RF1))=OUT2;
        RF1=RF1+1;

end

if length(EX)~=length(X)
    EX2(length(EX):length(X))=X(length(EX):length(X));
end


TEX=length(EX)-1;       %tamaño del archivo a la salida
TEX2=length(EX2)-1;
DEX=[0:TEX]/FS;         %dominio de la salida
DEX2=[0:TEX2]/FS;         %dominio de la salida
```

Conexión.m

```
function     [V VOZ]=conexion(EX,EX2,X,letras,mask)
clc
ref=1;
size=100;
SEG=1000;
indexado=1/(SEG/size);

while ref<=(length(mask))
   if(mask(ref)== 1)
       ix=ceil(ref*indexado);
       voc=letras(ix);

       if (voc == ' ');
           FILE2((ref-1)*size+1:(ref)*size)=EX((ref-
1)*size+1:(ref)*size);
       else
           vesof=EX((ref-1)*size+1:(ref)*size);
           ar=arcoef(vesof);
           fuent=filter(ar,1,vesof);
           [c,d]=butter(5,0.25);
           fuent=filter(c,d,fuent);
           cf=formantes(voc);
           voz=filter(1,cf,fuent);
           voz=voz/(max(voz)/max(vesof));
           wd=hann(length(fuent));
           voz=(voz(:).*wd);
           FILE2((ref-1)*size+1:(ref)*size)=voz;
       end
   end
   ref=ref+1;
end
if(length(X) ~= length(FILE2))
    FILE2(length(X))=0;
end

VOZ=EX2+FILE2;
[b,a]=butter(5,[0.0625 0.475]);
V=filter(b,a,VOZ);
dif=X-V';
dif=dif*2;
[b,a]=butter(5,.12,'high');
V=filter(b,a,dif);
```

Arcoef.m

```
function ar=arcoef(s)
p=12;
t=length(s);
ar=zeros(1,p+1);
ar(1,1)=1;
r=(0:p);

dd=s;
ww=hamming(t);
y=zeros(1,t+p);
c=(1:t)';
wd=dd(:).*ww;
y(1:t)=wd;
z=zeros(t,p+1);
z(:)=y(repmat(c,1,p+1)+repmat(r,t,1));
rr=wd'*z;
rm=toeplitz(rr(1:p));
rk=rank(rm);
    if rk
        if rk<p
            rm=rm(1:rk,1:rk);
        end
        ar(1,2:rk+1)=-rr(2:rk+1)/rm;
    end
```

Formantes.m

```
function coeficientes=formantes(cual)
if cual=='A'
    F1=[788 1327 2332 2600];
    B1=[100 120 350 350];
elseif cual=='E'
        F1=[188 431 1826 2337];
        B1=[150 180 200 200];
elseif cual=='I'
       F1=[216 258 2244 2653];
       B1=[300 200 100 100];
elseif cual=='O'
        F1=[320 457 868 2511];
        B1=[100 120 200 200];
elseif cual=='U'
        F1=[321 718 1789];
        B1=[100 150 200];
end

teta=(F1(1,:))*pi/5000;  % frecuencia
base=exp(-(B1(1,:))*pi/10000);  % Ancho de banda
raices=base.*exp(j*teta);
raices=[raices conj(raices)];
coeficientes=real( poly(raices) );
```