



Escuela
Politécnica
Superior

Caracterización del locutor mediante el pulso glotal.



Grado en Ingeniería en Sonido e Imagen
en Telecomunicación.

Trabajo Fin de Grado

Autor:

Pablo Esquerdo Roig

Tutor/es:

Sergio Bleda Perez

Junio 2017



Universitat d'Alacant
Universidad de Alicante

JUSTIFICACIÓN Y OBJETIVOS

El propósito de este trabajo es el estudio de un algoritmo de reconocimiento mediante el pulso glotal como discriminador del locutor. Generalmente el reconocimiento de voz se suele dividir en dos vertientes, la primera de ellas es el reconocedor del habla, es decir, del mensaje transmitido y la segunda de ellas consiste en el reconocimiento del locutor. En nuestro caso, nos centraremos en el segundo caso que puede ser de gran utilidad en el futuro en el ámbito de seguridad, como podría ser por ejemplo, en un banco.

En primer lugar, tendremos que realizar un sistema que sea capaz de modelar el aparato fonador que como explicaremos más tarde, se excitará la glotis e irá por el tracto vocal hasta desencadenar en un filtro de radiación para así finalmente formarse la voz.

Por lo tanto, el objetivo del proyecto será el estudio del pulso glotal como factor característico del locutor. La principal ventaja de éste estudio, es que no podrá ser falsificado mediante la variación de las frecuencias fundamentales de los formantes, es decir, ni aun imitando nuestra voz casi a la perfección se podría inhabilitar el algoritmo creado, debido a que la información que estamos analizando está sumergida en nuestra glotis o vibraciones emitidas.

Dicho de otra manera, suponiendo que cada locutor tenga un pulso glotal único y característico respecto a los demás, nuestro objetivo a modo resumen será el aislamiento del pulso de cada persona para la posterior caracterización, desembocando en un banco de pruebas entre distintos pulsos glotales para la obtención de la estadística de veracidad del algoritmo creado, o lo que es lo mismo, el porcentaje de acierto y de error de nuestro sistema.

AGRADECIMIENTOS.

En primer lugar me gustaría agradecer ante todo a mi tutor Sergio Bleda, tanto por su paciencia como por su habilidad de desglosar un gran problema en pequeñas partes, debido a que sin él nunca habría conseguido terminar y justificar el documento.

Además, también quiero agradecer a la Universidad de Alicante y a todo el profesorado que me ha impartido clases por brindarme la oportunidad de aprender sobre lo que realmente me gusta y expandir mis horizontes de cara al futuro.

Por otra parte me gustaría agradecer a mi familia por todo lo que han hecho, nunca me ha faltado de nada y siempre han sido el último empujón necesario para conseguir mis metas. Además, quiero mencionar a mi abuela Isabel por el especial entusiasmo que tiene en que su nieto acabe la carrera, ¡ya queda poco yaya!.

Finalmente pero no menos importante, me gustaría agradecer a los diferentes amigos que han querido colaborar para el desarrollo de mi trabajo, como a los amigos que no han podido, ¡gracias! vosotros sois la familia que se elige.

Además, una mención especial a mi amiga Andrea por haberme ayudado durante todo mi transcurso escolar desde Bachiller y personal por haber sido un pilar fundamental en mi desarrollo personal y profesional.

Gracias de corazón.

ÍNDICE

JUSTIFICACIÓN Y OBJETIVOS.....	2
AGRADECIMIENTOS.....	3
ÍNDICE.....	4
2. FUNDAMENTOS TEÓRICOS	6
2.1 FISILOGIA DEL SISTEMA FONADOR.....	6
2.1.1 CAVIDAD INFRAGLÓTICA.....	7
2.1.2 CAVIDAD GLÓTICA.....	7
2.1.3 CAVIDADES SUPRAGLÓTICAS	9
2.2 MODELADO DEL SITEMA FONADOR.....	10
2.2.1 MECANISMO DEL HABLA.....	10
2.2.2 MODELO LINEAL DE PRODUCCIÓN DE VOZ	12
2.3 DIGITALIZACIÓN DE SEÑALES DE AUDIO.....	14
2.4 ENVENTANADO DE LA SEÑAL	20
2.5 PREDICCIÓN LINEAL DE LA SEÑAL DE VOZ (LPC).....	23
2.6 FILTRO ADAPTATIVO LEAST MEAN SQUARE (LMS).....	28
2.7 ESPECTROGRAMA.....	32
3. DESARROLLO.....	33
3.1 PLANTEAMIENTO GENÉRICO	33
3.2 MÓDULO DE ANÁLISIS	34
3.2.1 MÓDULO DE TRATAMIENTO PREVIO DE LA SEÑAL.	35
3.2.2 MÓDULO DE ENVENTANADO.	36
3.2.3 MODULO DE FILTRO DEL TRACTO VOCAL Y CONSTRUCCIÓN DE RESIDUO.....	38
3.2.4 MÓDULO DE EXTRACCIÓN DE LOS PULSOS GLOTALES.....	44
3.2.5 MODULO DE ELIMINACIÓN DEL RUIDO DE BAJA FRECUENCIA	49
3.2.6 MODULO DE OBTENCIÓN DEL PULSO GLOTAL MEDIO.	51
3.2.7 MÓDULO DE EXTRACCIÓN DE INFORMACIÓN A PARTIR DEL ESPECTOGRAMA.....	52

3.2.8 EXTRACCIÓN DE CARACTERÍSTICAS A PARTIR DEL MÓDULO DEL ESPECTRO	55
3.3 MÓDULO DE OBTENCIÓN.	58
3.3.1 MÓDULO DE GRABACIÓN	59
3.3.2 MÓDULO DE OBTENCIÓN DEL VECTOR CARACTERÍSTICO "A"	61
3.3.3 MÓDULO DE OBTENCIÓN DEL VECTOR CARACTERÍSTICO "B"	63
3.3.4 MÓDULO DE OBTENCIÓN DEL VECTOR CARACTERÍSTICO "C y D" ..	66
3.3.5 ALMACENAMIENTO DE LOS VECTORES CARACTERÍSTICOS "A, B, C y D"	68
3.3.6 CÁLCULO DE LA DISTANCIA ENTRE LOS DISTINTOS VECTORES	69
3.3.7 MATRIZ RESULTANTE CON TODAS LAS POSIBILIDADES EXISTENTES.	71
3.3.8 SUBDIVISIÓN DE LA MATRIZ RESULTANTE EN DOS MATRICES DISTINTAS.	73
3.4 MÓDULO DE INTERPRETACIÓN.	73
3.4.1 MÓDULO DE INTERPRETACIÓN PRIMERA VERSIÓN.....	74
3.4.2 MÓDULO DE INTERPRETACIÓN MEDIANTE LA OBTENCIÓN DEL UMBRAL.....	77
CONCLUSIONES.....	81
BIBLIOGRAFÍA	84

2. FUNDAMENTOS TEÓRICOS

En los siguientes apartados explicaremos todos los conceptos teóricos que nos harán falta para la realización de nuestro proyecto. Gracias a éste apartado pretendemos aclarar las posibles dudas que pueden surgir sobre la ejecución e implementación de nuestros algoritmos. Nuestra idea es ofrecer la información necesaria para el entendimiento del proyecto, en caso de querer profundizar sugerimos leer la bibliografía que posteriormente ofreceré.

2.1 FISIOLOGIA DEL SISTEMA FONADOR

El aparato fonador está dividido en tres partes fundamentales (Figura 2.1): las cavidades infragloticas (órganos respiratorios), la cavidad glótica (órgano fonador) y las cavidades supraglóticas (órganos de la articulación) [1] [2]

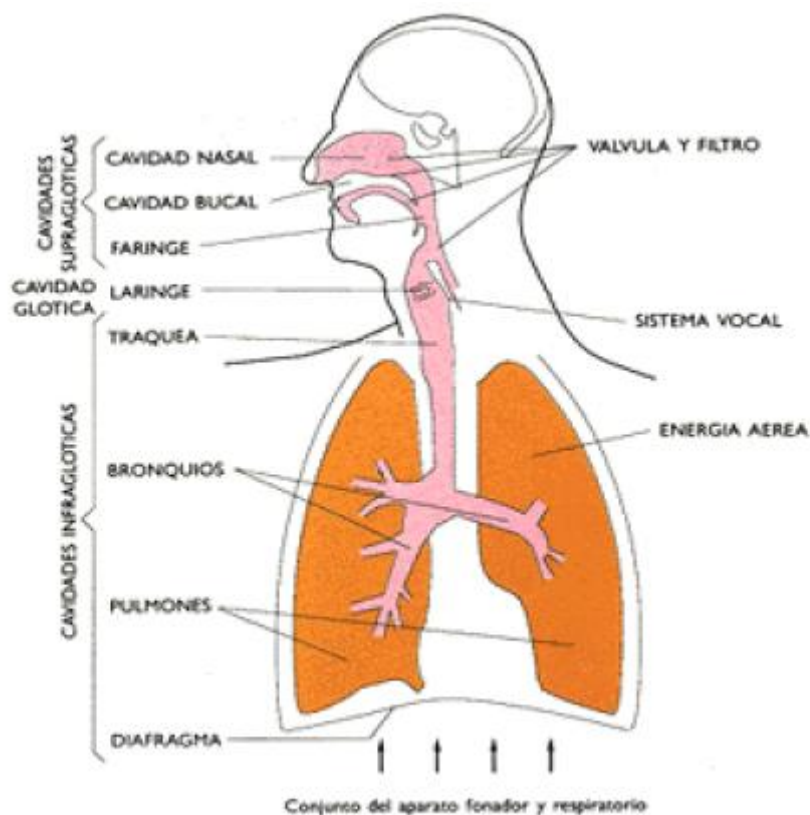


Figura 2.1: Aparato fonador [3]

2.1.1 CAVIDAD INFRAGLÓTICA

La cavidad infraglotica está situada por debajo de la glotis, a su vez está compuesta por pulmones, bronquios diafragma y tráquea.

El diafragma es un tabique muscular situado entre la cavidad abdominal y torácica. Cuando es contraído produce un elevamiento de las costillas y aumenta el volumen de la cavidad torácica por lo que el aire entra en los pulmones. Éste efecto es denominado inspiración.

De la misma manera, cuando el diafragma se relaja provocará una expulsión del aire que ha contenido en los pulmones con la fuerza y ritmo necesario para la formación de la fonación, éste efecto es denominado como la espiración.

2.1.2 CAVIDAD GLÓTICA

En cuanto a la laringe, está especialmente adaptada para actuar en función a un vibrador. Está comprendida por:

-El aparato fibroso (Figura: 2.2), soporte de las cuerdas vocales y esqueleto: Cartílagos (epiglótico, aritenoides, Santorini, Morgagni, cricoide, tiroide y sesamoideos), ligamentos y articulaciones.

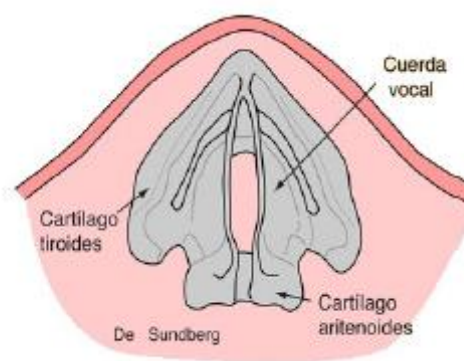


Figura 2.2: Aparato fibroso [4]

-El aparato tensor de las cuerdas vocales. Los elementos que vibran son las cuerdas vocales que son pliegues a lo largo de las caras laterales de la laringe que son movidos y estirados por diversos músculos específicos de la propia estructura que compone la laríngea. Cada cuerda vocal se estira entre el cartílago aritenoides y el cartílago tiroideos por los diferentes músculos inversamente, la contracción de los músculos cricoaritenoides laterales tiran de los cartílagos aritenoides hacia delante lo que genera la respiración normal. El músculo Tiro-aritenoides cuyo musculo interno conocido como músculo vocal, constituye en su mayoría el grosor de la cuerda vocal, cuya tensión y consistencia incrementa a la vez que estrecha la glotis.

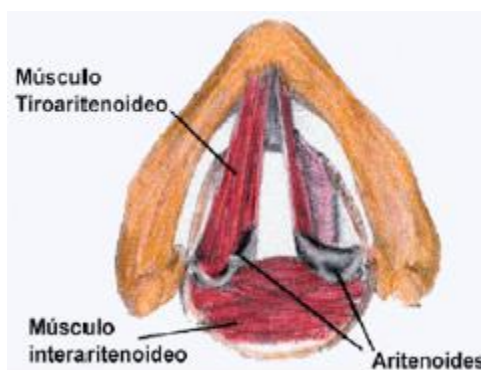


Figura 2.3: Musculatura correspondiente del aparato tensor [5]

Las cuerdas vocales podrán adoptar diferentes formas en función de las cual provocará un cambio en el tono, por lo tanto, unas cuerdas vocales planas o delgadas producen sonidos agudos y unas cuerdas vocales gruesas provocan sonidos graves.

-El aparato motor de las cuerdas vocales (Figura 2.4). Está constituido por los músculos que se encargan de aproximar o separar las cuerdas vocales. Así, los músculos crico-aritenoides provocarán una separación de las cuerdas abriendo ampliamente la glotis durante la inspiración. Los músculos cricoideos laterales ejercerán la función antagonista provocando la aproximación de las cuerdas y preparando la acción del músculo vocal. El músculo aritenoides transversal tirará de los aritenoides produciendo una aproximación de las cuerdas vocales de forma que vibren con el aire espirado. [6]

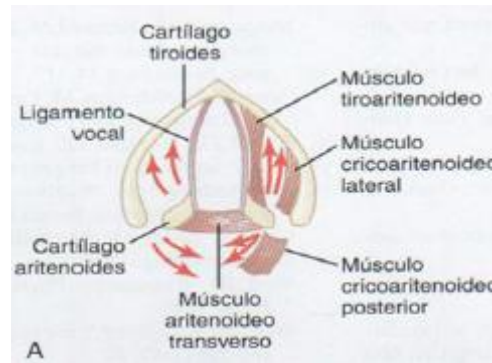


Figura 2.4: Aparato motor [7]

2.1.3 CAVIDADES SUPRAGLÓTICAS

En esta cavidad se encuentran los órganos encargados de la articulación y la resonancia. Los tres órganos principales de la articulación son la lengua, los labios y el paladar blando, por otra parte, los responsables de la resonancia son la nariz, la boca, los senos paranasales, la faringe e incluso el tórax. Es el último tramo que recorrerá el aire donde se generará la gran variedad de sonidos gracias a la funcionalidad de la movilidad de la cavidad bucal.[5]

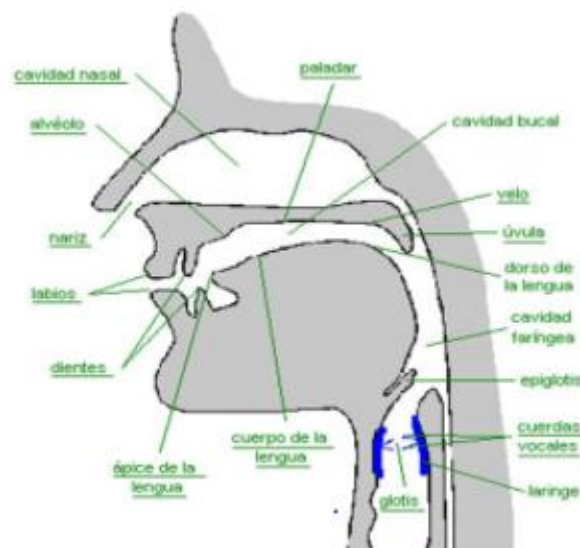


Figura 2.5 cavidades supraglóticas [8]

2.2 MODELADO DEL SITEMA FONADOR

Con el objetivo de la aplicación de técnicas de procesamiento digital de señales a los problemas del reconocimiento del locutor, es esencial comprender los fundamentos del proceso de producción de voz como los fundamentos del procesado digital de señales. Este apartado resumirá la teoría acústica necesaria para la producción de voz proporcionando por tanto la base que será necesaria para la aproximación clásica del modelado de la señal de voz como salida de un sistema lineal variante en el tiempo siendo excitado por una secuencia de pulsos cuasi periódicos o ruido blanco, dependiendo así también por el tipo de voz producida pudiendo ser sorda o sonora respectivamente.

Dicha aproximación será aplicada para la obtención en tiempo discreto para la señal de voz.

2.2.1 MECANISMO DEL HABLA

El habla es el resultado final de la aplicación de un acto voluntario en el que se ha intervenido los órganos pertenecientes al sistema digestivo y respiratorio. El control del proceso lo lleva a cabo el sistema nervioso central utilizando la realimentación de la información transmitida a través del oído. El aparato fonador encargado de la producción de la voz se puede dividir en tres bloques:

- Generador de energía. La energía necesaria para la producción de voz partirá de los músculos abdominales y torácicos, que al aumentar la presión en los pulmones producirá una corriente de aire como ya he detallado anteriormente.
- Sistema vibrante. Constituido por la glotis o cuerdas vocales situadas en la laringe, separando el tracto vocal de la tráquea. Las cuerdas vocales en los sonidos sordos producirán una vibración cuasi periódica. De dicha vibración obtendremos una señal para su posterior modulación y modificación por el tracto vocal lo que dará lugar a diferentes armónicos que generarán la tonalidad y las diversas características psicoacústicas a la señal de voz.

- Sistema resonante. Estará constituido por el tracto vocal que se caracteriza por ser un tubo no uniforme de 17 centímetros de longitud, cuyos límites son los labios u orificios de la nariz por un lado, y por el otro lado las cuerdas vocales. Por lo tanto, en cuanto a el tracto vocal, estará constituido por la faringe (conexión del esófago con la boca) y la cavidad bucal. Posee una sección variable en función de la posición de los órganos articulatorios (mandíbula, labios y lengua) pudiendo variar entre 0 y 20 centímetros cuadrados. La cavidad nasal a su vez comenzará en el velo del paladar hasta terminar en la abertura de la nariz, cuando el velo está abierto, la cavidad se acoplará acústicamente con el tracto vocal dando como resultado la producción de los sonidos nasales. Por otra parte, en los sonidos no nasales, el velo impedirá el paso del aire hacia la nariz. Los órganos articulatorios y la cavidad nasal permitirán concentrar la energía en determinadas frecuencias o formantes lo que dará como resultado, resonadores conmutables.

En la figura 2.6 se mostrará una representación del mecanismo fisiológico completo de la producción de la voz. La función primaria será la inhalación, que es posible gracias a la expansión de la cavidad torácica mediante la cual descenderá la presión en los pulmones entrando aire a través de las fosas nasales. La energía necesaria para expulsar el aire residirá en los músculos torácicos y abdominales. Cuando la cavidad torácica se contraiga aumentando la presión en los pulmones, el aire saldrá pasando antes por los bronquios y la tráquea, actuando por tanto como excitación del tracto vocal.

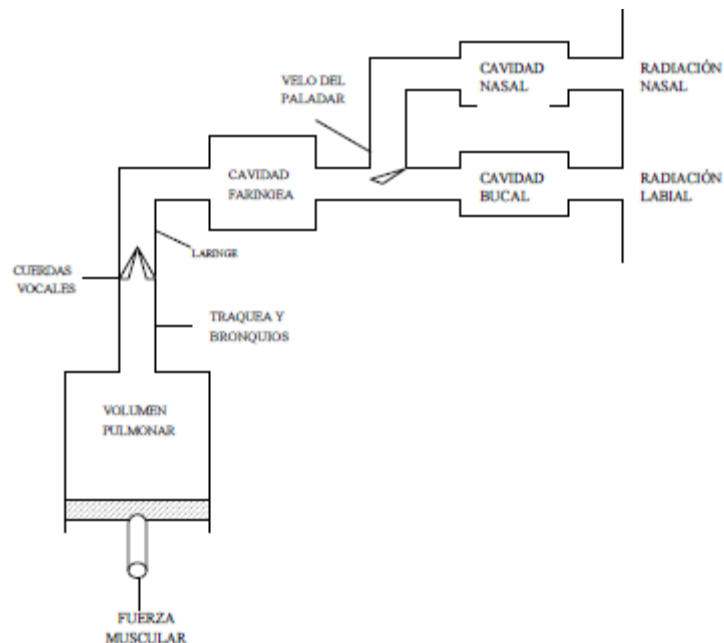


Figura 2.6. Representación esquemática de la producción de voz. [9]

2.2.2 MODELO LINEAL DE PRODUCCIÓN DE VOZ

Para la aceptación de la validez de un modelo de producción lineal, deberá ser equivalente al real en sus distintos terminales, pero su estructura interna no tendrá por qué reproducir los mecanismos físicos internos. Además, deberá estar controlado por diversos parámetros que estarán relacionados con la producción de voz. El modelo lineal incluye un sistema excitado por una señal capaz de evolucionar en el tiempo desde pulsos cuasi periódicos para voz sonora como a ruido aleatorio para voz sorda. En la figura 2.7 se representará el modelo lineal clásico, también denominado modelo fuente - filtro siendo la fuente los pulmones y la glotis y el filtro las cavidades supraglóticas.

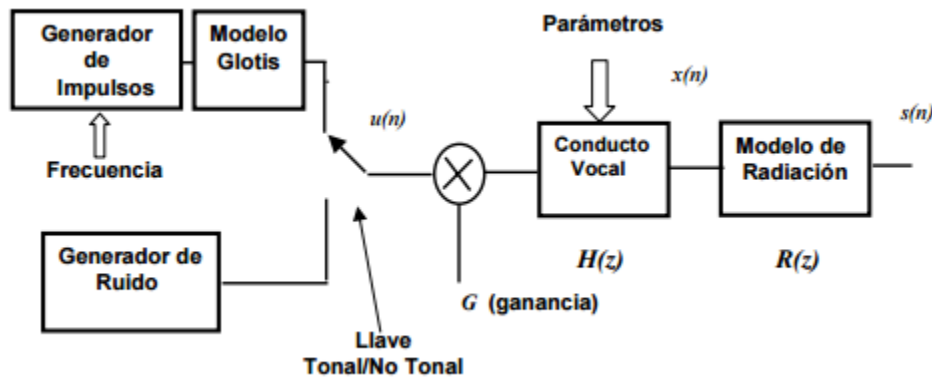


Figura 2.7 Modelo lineal de producción de voz. [10]

En él se podrán distinguir los siguientes elementos:

1) El tracto Vocal $V(z)$. Generalmente emplea una función todo polos con la siguiente fórmula:

$$H(z) = \frac{G}{1 - \sum_{i=1}^N a_i z^{-i}} = \frac{G}{\prod_{i=1}^N (1 - p_i z^{-1})} \quad \text{Fórmula(2.1)}$$

2) La radiación de los labios $R(z)$. La técnica utilizada para la introducción del efecto de radiación en el modelo digital es la inclusión de un filtro de paso alto, o lo que es lo mismo, un diferenciador de primer orden de la forma

$$R(z) = 1 - z^{-1} \quad \text{Fórmula (2.2)}$$

Dicho diferenciador tendrá un efecto de subida de 6 dB/octava para las frecuencias fundamentales.

3) La excitación. Para la voz sonora, el generador producirá un tren de impulsos que excitará por lo tanto un sistema lineal $G(z)$ con la forma de la glotis deseada. Para la voz sorda, el modelo será más sencillo debido a que todo requerirá una fuente de ruido aleatoria. Para ambos casos, existirán controladores de ganancia que controlarán la intensidad de la señal de excitación.

Finalmente, es necesario combinar los modelos previamente explicados para obtener la función de transferencia global $H(z)$, quedando así la siguiente fórmula:

$$H(z) = V(z)R(z)G(z) \quad \text{Fórmula (2.3)}$$

Para involucrar un modelo lineal para un sistema de reconocimiento deberíamos restringir ciertas características. La primera de ellas, para calcular los parámetros del modelo es necesario suponer una casi estacionariedad, es decir, las características de la señal no pueden cambiar durante el intervalo en el que son estudiadas. Esta suposición es válida en algunos tramos de la señal, pero en otros como por ejemplo las transiciones de los fonemas no. La diferencia, es que si queremos ser estrictos la voz es un proceso no estacionario por definición. La segunda suposición, será asumir un filtro todo polos para el tracto vocal implicando por lo tanto la falta de ceros para la modelación de ciertos fonemas, como pueden ser los nasales. La última, la simplificación de división en sonidos sordos o sonoros no es siempre válida, debido a que puede actuar las fricativas sonoras.

2.3 DIGITALIZACIÓN DE SEÑALES DE AUDIO

A continuación detallaremos en profundidad los conceptos relacionados con

- Muestreo (Sampling)
- Cuantización (Quantization)
- Codificación (Codification)

Las vibraciones sonoras se pueden representar como señales electrónicas con el material necesario como puede ser un micrófono que convierte las vibraciones en una señal de voltaje dependiente del tiempo. El resultado de éste fenómeno de conversión es denominado señal analógica. Este tipo de señales son continuas debido a que consisten en un continuo de valores.

Las señales analógicas pueden ser manipuladas, amplificadas y grabadas mediante técnicas analógicas. Dicha señal es adecuada para algunas aplicaciones, el problema es que cuando una grabación analógica es copiada, añadimos una cantidad de ruido importante, además de que cuando amplificamos una señal, también ampliaremos el ruido que ésta contiene.[11]

Sin embargo, los ordenadores son máquinas digitales, es decir, sus operaciones son basadas en las matemáticas discretas que es concretamente lo opuesto a continuo. Las entidades son contadas por lo que debemos trabajar con números finitos y exactos.

La mayor dificultad de trabajar en ordenador para la síntesis de sonido es que se trabaja en el dominio discreto debido a que el conocimiento científico que se tiene es esencialmente analógico. Además, los ordenadores únicamente trabajan con números binarios o bits que son combinaciones de 0 y 1.[11]

Por lo tanto, para trabajar con pistas de audio en el ordenador provenientes de un micrófono, las señales analógicas deberán ser convertidas a formato digital, es decir, se deberá de transformar la pista de audio en un conjunto de números binarios. Además, para reproducir la señal debe ser convertida a analógico, por lo tanto, el ordenador necesitará dos tipos de conversores de datos. El primero será de analógico a digital (AD) y el siguiente de digital a analógico (AC).

En la siguiente figura se representará un esquema de los pasos de digitalización de una señal analógica:

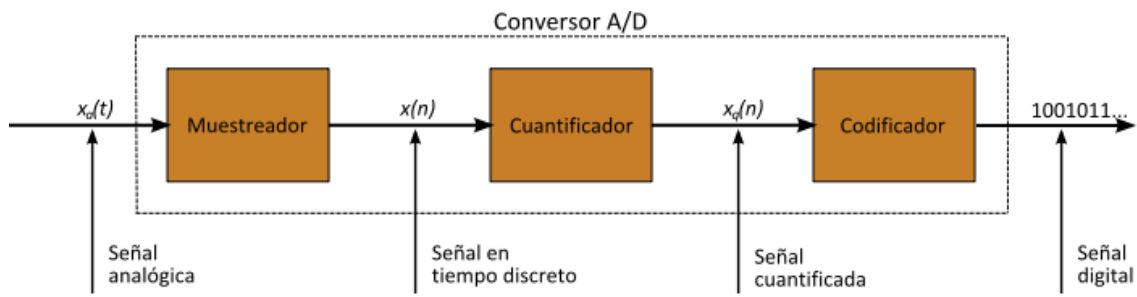


Figura 2.8 Esquema de funcionamiento de un conversor Analógico / Digital [12]

Muestreo

El bloque de muestreo funciona mediante la medición de la amplitud de la señal continua con intervalos de misma duración. Cada valor medido es denominado muestra (o sample) de la señal. Matemáticamente se puede expresar de la siguiente manera: para muestrear una señal se toma valores de una señal continua $x(t)$ a determinados instantes de tiempo t_n :

$$x(t = tn) = x(n), \quad tn = n \cdot Tm \quad \text{Fórmula 2.4}$$

Siendo Tm el periodo.

La distancia temporal entre dos muestras consecutivas es denominada periodo de muestreo que es medido en segundos. la inversa de dicho periodo es:

$$f_m = \frac{1}{T_m} \quad \text{Fórmula 2.5}$$

Denominado frecuencia de muestreo, que se mide en ciclos por segundos o Hz.

Por lo tanto, en el proceso de muestreo la señal sufre una transformación de señal continua a un conjunto de muestras, o dicho de otra manera, un conjunto de puntos discretos en el tiempo.

Además, es importante muestrear la señal lo suficientemente rápido para poder capturar toda la información de ésta. El teorema de Nyquist o teorema de muestreo, dice que para muestrear la señal adecuadamente es necesario tener al menos dos muestras por cada ciclo de la senoide, por lo tanto, para una correcta representación del sonido, la frecuencia de muestreo f_m tendrá que ser mayor como mínimo al doble de la frecuencia más alta contenida en la señal.[11]

$$f_m \geq 2 f_{maxima} \quad \text{Fórmula 2.6}$$

La frecuencia de Nyquist es denominada a la frecuencia más alta que se podrá capturar con una determinada frecuencia de muestreo f_m .

$$f_{Nyquist} = \frac{f_m}{2} \quad \text{Fórmula 2.7}$$

Aliasing

Una onda compleja puede estar compuesta de sinusoides con altas frecuencias, las cuales oscilará tan rápidamente que no podrán ser representadas correctamente las muestras de la señal debido a que hay demasiado espaciado entre sí. A este fenómeno se le denomina Aliasing y ocurrirá cuando la señal muestreada tiene componentes de frecuencia mayores a la mitad de la frecuencia de Nyquist o frecuencia de muestreo. Debido a que a estas frecuencias no se cumple el teorema de Nyquist, ocurrirá dicho fenómeno. [11]

Estas componentes de frecuencia corromperán la señal original introduciendo componentes denominados alias. Las frecuencias que aparecen se podrán calcular como

$$f_r = f_m - f_x \quad \text{Fórmula 2.8}$$

Donde f_m es la frecuencia de muestreo y f_x la frecuencia de la señal.

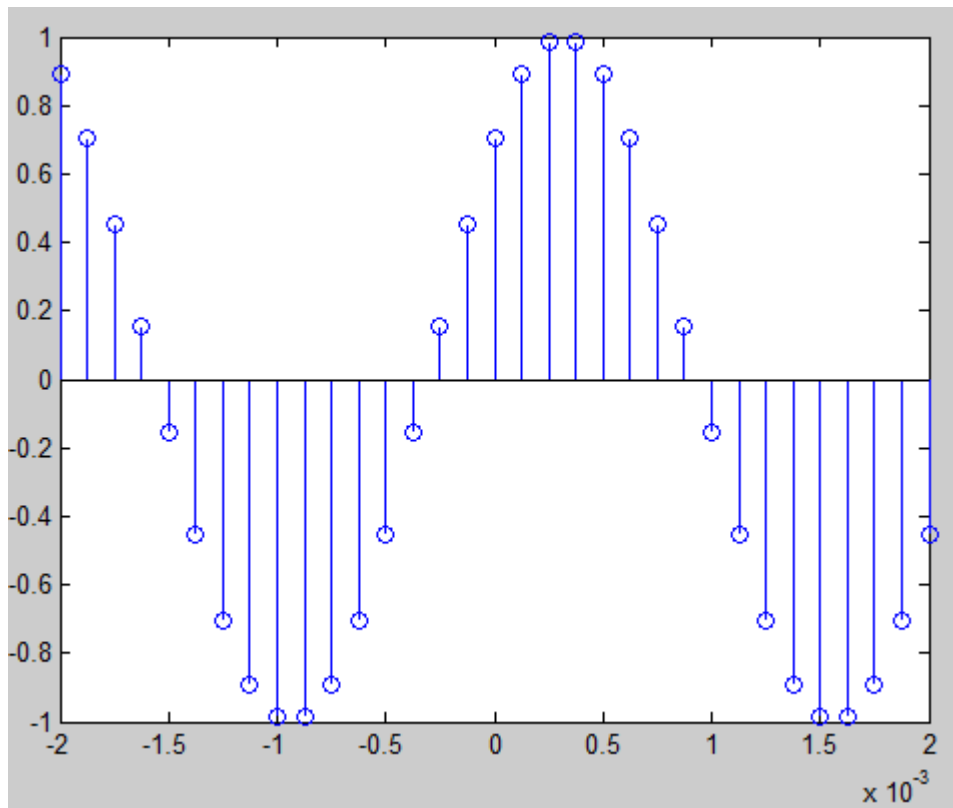


Figura 2.9 Señal muestreada

Hay dos posibles soluciones para solucionar el problema del aliasing:

- Aumentar la frecuencia de muestreo hasta que sea igual o mayor al doble de la frecuencia máxima de la señal.
- Realizar un filtro denominado filtro antialiasing que es un filtro de paso bajo que se situará por encima de la frecuencia de Nyquist.

Cuantización

Una vez la señal es muestreada obtenemos un conjunto de muestras o valores continuos con la amplitud de la señal. La cuantización será realizada al limitar los posibles valores de la amplitud en una señal definiendo una serie de valores.

El número de posibles valores viene determinado por la resolución del convertidor Analógico/Digital o Digital/Analógico. La resolución de éstos depende de la palabra que se utiliza para la representación de las muestras de la señal. La resolución se mide en número de bits, y un convertidor de n bits utilizará 2^n valores de la señal. Es decir, si un sistema 8 bits, tendrá por lo tanto $2^8=256$ posibles valores diferentes. Como es de suponer, a mayor valor de resolución del convertidor, mayor precisión tendrá la representación de la señal.

Ruido de cuantización

El ruido de cuantización aparece durante el proceso de cuantización debido a que se sustituye la amplitud de la muestra por la amplitud más cercana al conjunto de valores permitidos. Es definido como la diferencia entre la señal antes de cuantizar y la señal muestreada cuantizada cumpliendo la siguiente ecuación:

$$r(n) = x_s(n) - x_Q(n) \qquad \text{Fórmula 2.9}$$

Donde $x_s(n)$ es la señal discreta y $x_Q(n)$ la señal discreta cuantizada. Dicho ruido representa la pérdida de calidad del sonido una vez ha sido cuantizado.

Además, se define la relación señal a ruido de cuantización (SNR_Q) como la relación entre la potencia P_s de la señal y la potencia P_N del error $\varepsilon[n]$, medido en decibelios.

$$P_s = \frac{1}{N} \sum_{n=-\infty}^{\infty} x_s^2 [n]$$

$$P_N = \frac{1}{N} \sum_{n=-\infty}^{\infty} \varepsilon^2 [n]$$

$$SNR_Q(dB) = 10 \cdot \log \frac{P_s}{P_N} = 10 \cdot \log \frac{\frac{1}{N} \sum_{n=-\infty}^{\infty} x_s^2 [n]}{\frac{1}{N} \sum_{n=-\infty}^{\infty} \varepsilon^2 [n]} \quad \text{Fórmula 2.10}$$

Codificación

El proceso de codificación por lo tanto consiste en asignar un conjunto de bits a cada valor posible de las muestras. Hay diversas posibilidades para realizar el proceso de codificación, además, cada codificación incluirá parámetros referentes al proceso de digitalización como pueden ser los siguientes:

- Frecuencia de muestreo
- Número de canales
- Bit rate que es la velocidad de transferencia (medido en bits por segundo)
- Resolución número de bits.
- Pérdida: Algunos codificadores realizan una compresión de la pista de audio dando como resultado cierta pérdida de información.

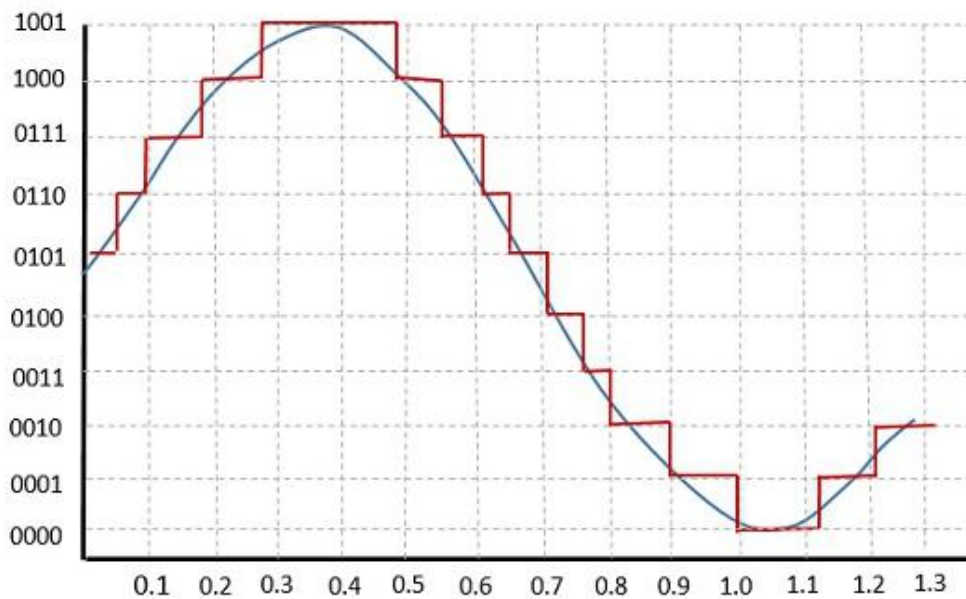


Figura 2.10 Ejemplo de señal cuantificada [13]

2.4 ENVENTANADO DE LA SEÑAL

Debido a que la señal de voz tiene una serie de particularidades como la pseudo periodicidad y la diferenciación entre los dos estilos de sonidos, es necesario estudiar estos comportamientos debido a que son los que caracterizarán la señal de voz. En primer lugar, es necesaria la realización de una multiplicación por un pulso cuadrado lo que supone una convolución en el dominio frecuencial. El problema surge a que estaríamos imponiendo un aliasing espectral no deseado, por lo que para evitar lo mencionado anteriormente es necesario el enventanamiento de la señal.

El propósito principal de la multiplicación de la onda por la función de ventana tiene dos efectos. En primer lugar, atenúa progresivamente la amplitud tanto en el principio como en el final de los extremos del intervalo de extracción para evitar un cambio abrupto en los puntos finales. En segundo lugar, producirá una convolución para la transformada de Fourier de la función de ventana y el espectro de la voz o la media móvil ponderada en el dominio espectral. Es por lo tanto deseable que la función de la ventana satisfaga las dos características para reducir la distorsión espectral causada por la ventana. Una es una resolución en alta frecuencia, principalmente un agudo, la otra es una pequeña fuga espectral de los otros elementos espectrales producidos por la convolución, en otras palabras, una gran atenuación lateral. [14]

Dado que estos dos requisitos son contrarios entre ellos y debido a que es imposible satisfacer los dos, se ha escogido diversas ventanas de compromiso. Una de esas ventanas por compromiso es la ventana de Hamming $W_H(n)$ definida como:

$$W_H(n) = 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right) \quad \text{Fórmula 2.11}$$

Esta ventana es normalmente usada para análisis de voz. La ventana de Hamming tiene la ventaja de que su resolución en el ámbito frecuencial es relativamente grande y su fuga espectral es pequeña desde que la atenuación lateral es mayor a 43 dB.

Por otra parte, la ventana rectangular, $W_R(n) = 1$ ($0 \leq n \leq N-1$), que corresponde a la extracción simple de $N - 1$ muestras de la señal de audio, tiene la mayor resolución frecuencial mientras que la atenuación lateral es solo de 13 dB. La ventana rectangular no

es adecuada para el análisis de una señal de audio debido a que tiene un amplio rango dinámico de componentes espectrales.

Otra ventana denominada la ventana de Hanning,

$$WN(n) = 0.5 - 0.5 \cos\left(\frac{2n\pi}{N-1}\right) \quad \text{Fórmula 2.12}$$

es también utilizada. Aunque la ventana de esta ventana es que las bandas laterales de orden superior son menores que las de la ventana de Hamming, la atenuación de la primera capa es únicamente 30 dB.

Las formas de las ventanas anteriormente explicadas para un espectro de 10 periodos de ondas sinusoidales de 1-KHz extraídas mediante el uso de estas ventanas se podrá observar en la figura 2.11.

La relación entre el periodo de muestreo $T[s]$, el número de muestras para el análisis N y la resolución de la frecuencia nominal del espectro calculado $\Delta f[Hz]$ se expresa como:

$$\Delta f = \frac{1}{TN} \quad \text{Fórmula 2.13}$$

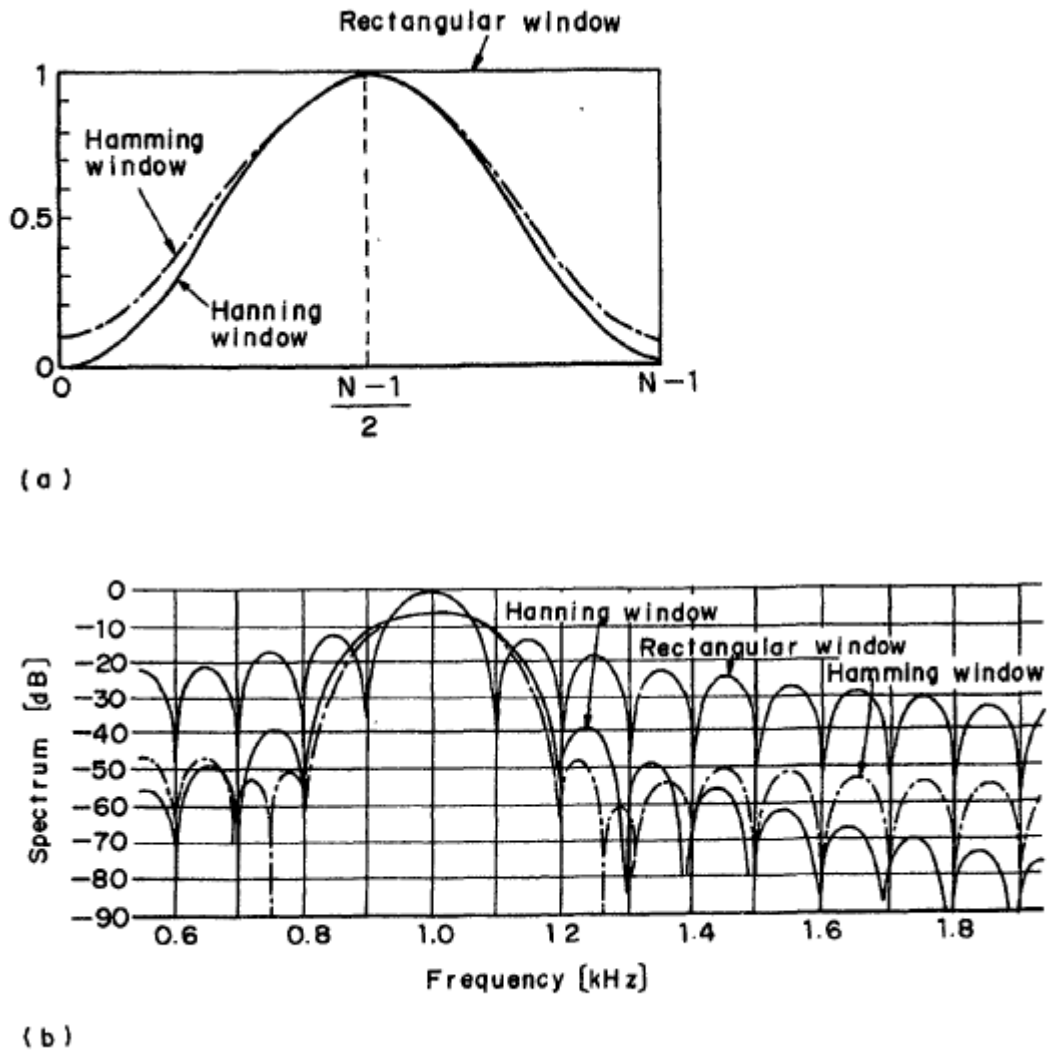


Figura 2.11 Formas de las ventanas (a) y el espectro de los 10 periodos de las señales sinusoidales extraídas de 1-KHz de las diversas ventanas usadas (b). [14]

A partir de esto, está claro que la resolución aumentara en proporción de la longitud del intervalo de análisis de la pista de audio. Por ejemplo, cuando $T = 0.125$ ms (8 KHz de muestreo) y $N=256$ (32- ms de duración),

$$\Delta f = \frac{10^3}{0.125 \cdot 256} = 31 \text{ [Hz]}$$

Fórmula 2.14

Cuando el análisis de la longitud de la ventana aumenta, la resolución espectral aumenta mientras que la resolución temporal decrece. Por otra parte, cuando el análisis de longitudinal de la ventana se acorta, la resolución temporal aumenta mientras que la resolución espectral disminuye. Estas relaciones son fáciles de entender por el hecho de que al multiplicar la forma de una pista de audio por la función ventana corresponde a la media móvil del espectro en el dominio espectral.

Además, cuando la forma de onda se multiplica por la ventana Hamming o Hanning, el intervalo de análisis efectivo se vuelve aproximadamente 40% más corto desde que la forma de onda cercana a ambos extremos, dando por consiguiente una disminución del 40% en la forma espectral.

Por lo tanto, la multiplicación de la onda de voz por una ventana reduce la fluctuación espectral debido a la variación de la excitación de la posición del pitch dentro del intervalo de análisis. Esto es eficaz en la producción de espectros estables durante el análisis de la voz, ya que la multiplicación de la ventana reduce la efectividad de la longitud del intervalo de análisis, el intervalo de análisis debería desplazarse de forma superpuesta a lo largo de la onda de voz para facilitar el rastreo de los espectros que varían en el tiempo. [14]

El intervalo de análisis de corto tiempo multiplicado por la función ventana y extraído de la forma de onda es denominado marco.

2.5 PREDICCIÓN LINEAL DE LA SEÑAL DE VOZ (LPC)

El tracto vocal es modelado como un sistema todo polos formado por una cascada de un pequeño número de resonadores de dos polos. Cada resonancia se define como un formante con su frecuencia central y su ancho de banda correspondientes. No se considera el efecto de la cavidad nasal en la producción de los sonidos nasales. [15]

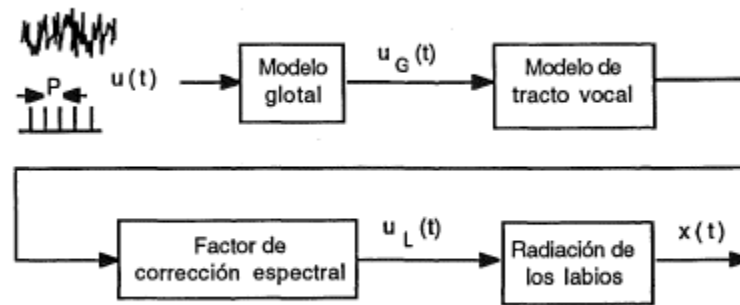


Figura 2.12 Modelo lineal de producción del habla.

La señal de velocidad volumétrica en los labios $u_L(t)$ es transformada en una señal de presión acústica $x(t)$ a una determinada distancia de los labios mediante el modelo de radiación de los labios.

Si suponemos la invarianza con el tiempo, el modelo puede ser descrito en notación de la transformada Z para su posterior implementación mediante la siguiente ecuación:

$$X(z) = U(z)G(z)V(z)L(z) \quad \text{Fórmula 2.15}$$

Siendo $X(z)$ y $U(z)$ la transformada Z de las secuencias discretas $x(n)$ y $u(n)$, que se obtienen como resultado de muestrear $x(t)$ y $u(t)$ a un período de muestreo T , y $G(z)$, $V(z)$ y $L(z)$ son las funciones de transferencia de los sistemas discretos que modelan los efectos de la glotis, el tracto vocal y los labios, respectivamente. Hay que hacer notar que en la representación discreta puede eliminarse el factor de corrección espectral que figuraba en el modelo original. [15]

Se puede simplificar este modelo mediante la combinación de la glotis, los labios y el tracto vocal desencadenando todo en una función de transferencia $H(z)$ que cumple la siguiente ecuación:

$$X(z) = U(z)H(z) \quad \text{Fórmula 2.16}$$

Pero en la práctica, las aplicaciones modelan el filtro $H(z)$ como un filtro todo-polos que cumple la siguiente ecuación:

$$H(z) = \frac{G}{1 + \sum_{i=1}^p a_i z^{-i}} \quad \text{Fórmula 2.17}$$

Desde que el término de la predicción lineal fue acuñado por Wiener, ésta técnica ha sido aplicada en un gran abanico de aplicaciones bajo distintas circunstancias. Fue utilizado por primera vez por Saito e Itakura y Atal y Schroeder produciendo un gran impacto en todos los aspectos de tratamiento y síntesis del sonido.

La técnica de la predicción lineal que corresponde con las siglas LPC (Linear Predictive Coding) consistirá en estimar el valor de la señal actual $x(n)$ como una combinación lineal de las muestras anteriores, quedando el valor estimado x_e de la siguiente manera:

$$x_e = - \sum_{i=1}^p a_i x(n-i) \quad \text{Fórmula 2.18}$$

Siendo p el orden del predictor lineal y a_i los coeficientes de predicción obtenidos. El problema principal alberga en la obtención de los coeficientes a_i de forma que la aproximación x_e sea lo suficientemente buena siguiendo algún criterio.

El error del valor real $x(n)$ y del valor estimado $x_e(n)$ se denomina error de predicción y sigue la siguiente fórmula:

$$e(n) = x(n) - x_e(n) = x(n) + \sum_{i=1}^p a_i x(n-i) \quad \text{Fórmula 2.19}$$

Utilizando esta expresión se puede considerar el error de predicción como respuesta a $x(n)$ de un sistema, siendo denominado filtro del error de predicción, a continuación definiremos su función de transferencia.

$$A(z) = 1 + \sum_{i=1}^p a_i z^{-i} \quad \text{Fórmula 2.20}$$

Además, a partir de B también podemos definir la siguiente fórmula

$$x(n) = -\sum_{i=1}^p a_i x(n-i) + e(n) \quad \text{Fórmula 2.21}$$

Por lo tanto, para simplificar el modelo de predicción lineal utilizaremos el siguiente esquema:

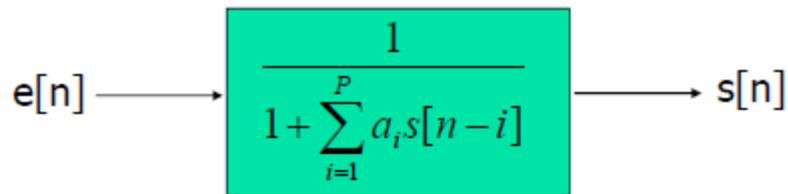


Figura 2.13 Simplificación de predictor lineal [16]

Además, si combinamos las expresiones (1.16) y (1.17) obtenemos la siguiente ecuación:

$$X(z) = \frac{G U(z)}{1 + \sum_{i=1}^p a_i z^{-i}} \quad \text{Fórmula 2.22}$$

Si tomamos la transformada Z a ambos lados de la ecuación anterior, podemos escribir la fórmula (2.22) como:

$$x(n) = -\sum_{i=1}^p a_i x(n-i) + G u(n) \quad \text{Fórmula 2.23}$$

Por lo que si comparamos las expresiones (2.23) y (2.21) el filtro de predicción A(z) será un filtro inverso del filtro H(z) de la expresión (2.17), quedando así:

$$H(z) = \frac{G}{A(z)} \quad \text{Fórmula 2.24}$$

Por lo que como he mencionado anteriormente, el problema básico de la predicción lineal es la obtención del conjunto de coeficientes a_k de la señal de tal forma que se obtenga una estimación buena de las diversas propiedades espectrales de la señal de voz mediante el uso de la ecuación (2.24).

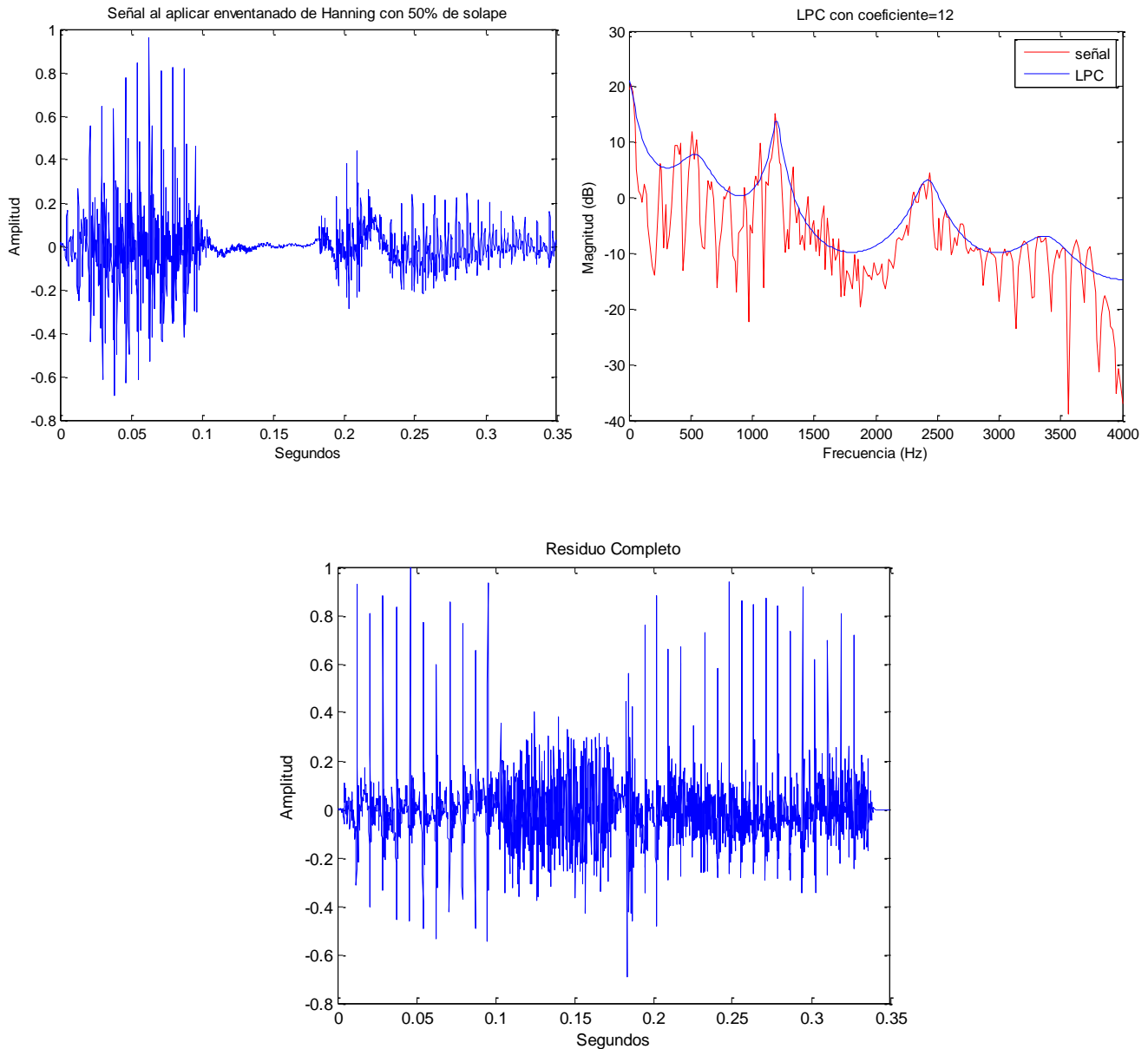


Figura 2.14 a) Señal al aplicársele el enventanado adecuado, b) Obtención de los coeficientes del filtro LPC, c) Reconstrucción del residuo mediante el filtrado inverso LPC.

2.6 FILTRO ADAPTATIVO LEAST MEAN SQUARE (LMS)

Los filtros adaptativos son caracterizados por modelar la relación entre señales de forma iterativa en tiempo real, a diferencia de los filtros IIR o FIR, los filtros adaptativos pueden cambiar su forma de comportamiento variando sus coeficientes con un algoritmo interno adaptativo, es decir, los coeficientes no se conocen al principio, se calculan con la implementación del filtro y por consiguiente son ajustados en cada iteración.

Por otra parte, los filtros adaptativos son variantes en el tiempo y no son lineales, otorgando mayor complejidad que otros filtros. El filtro adaptativo, se caracteriza por la presencia de dos señales a la entrada: $x(n)$ y $e(n)$ siendo $e(n)$ el error de la señal de salida del filtro $y(n)$. Además, internamente tendremos los coeficientes del filtro denominados $w(n)$ que multiplicarán la entrada $x(n)$ para dar como resultado la salida $y(n)$.

El filtro adaptativo que nosotros vamos a considerar tendrá la siguiente forma:

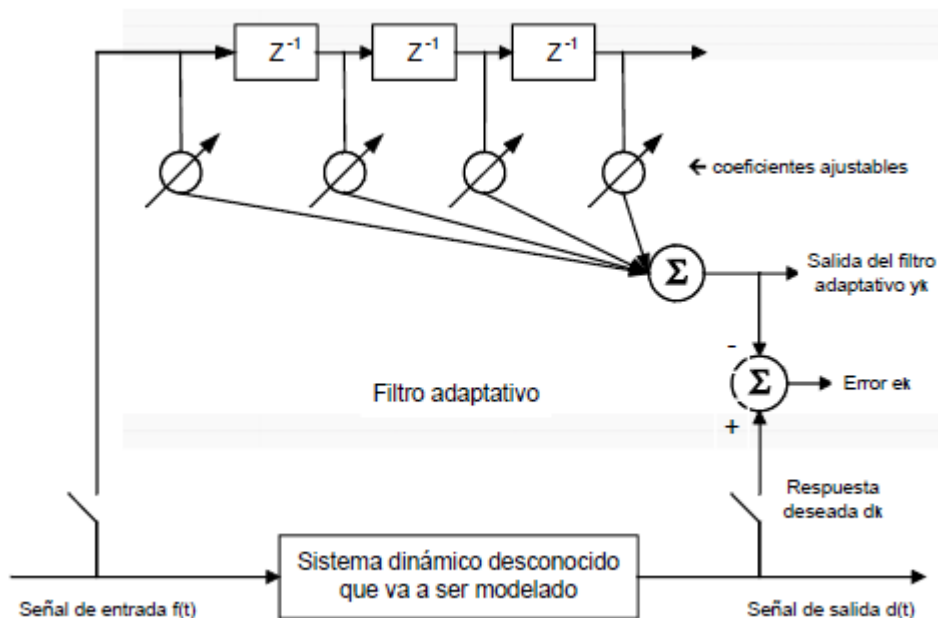


Figura 2.15: Esquema de funcionamiento de un filtro discreto adaptativo. [17]

Además de las señales comentadas será necesario la obtención de una señal más, la señal deseada para poder generar por lo tanto la señal de error.

El filtro adaptativo de la figura 2.15 es un filtro digital, sus respectivas entradas son la señal de entrada y salida del sistema desconocido. Los coeficientes por lo tanto se ajustan automáticamente mediante un algoritmo que minimizará el error cuadrático medio. Cuando los coeficientes convergen y el error se hace pequeño, la respuesta impulsiva del filtro se asemejará a la del sistema desconocido.

Se define la potencia media de la señal de error como $\xi = E\{(e[n])^2\}$. Ésta es la función del valor de los coeficientes del filtro transversal y su representación se denomina superficie de error. En la figura 2.16 se muestra un ejemplo de superficie de error ξ en función del valor de los coeficientes w_0 y w_1 de un filtro adaptativo sencillo. [18]

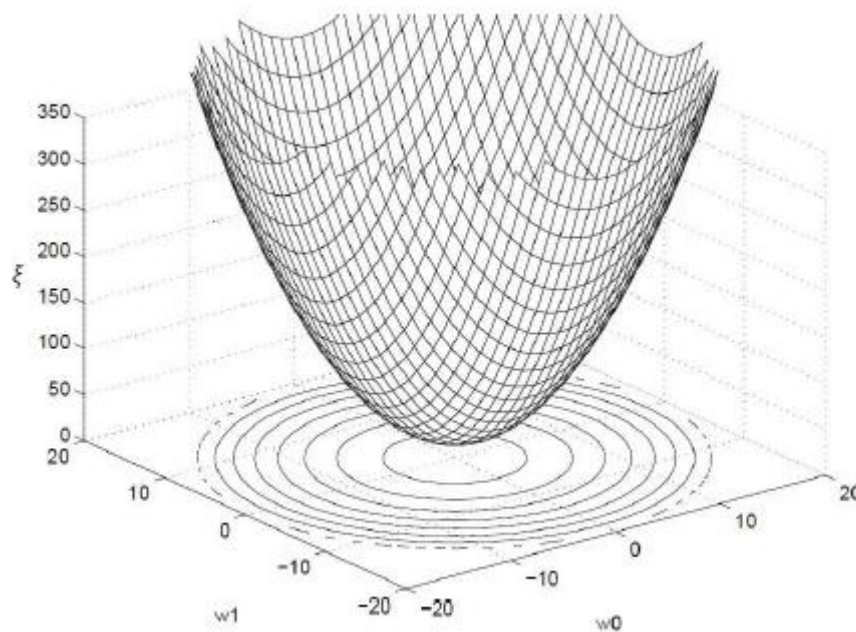


Figura 2.16: La función ξ , dependiente del valor de los coeficientes W , es cuadrática y presenta un único mínimo global.[18]

El análisis del filtro adaptativo se puede desarrollar considerando primero el combinador lineal adaptativo de la Figura 2.17 y un subsistema de la Figura 2.15.

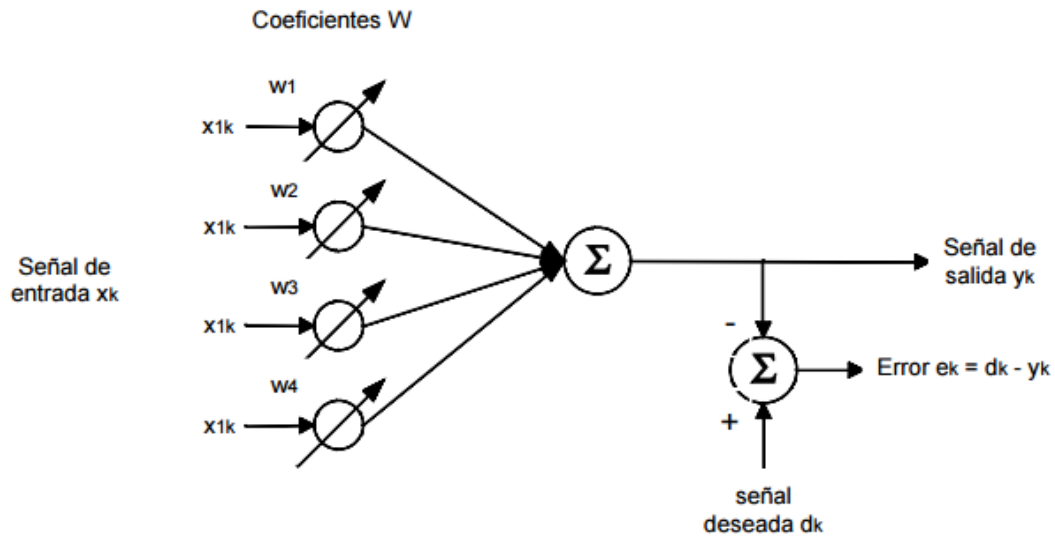


Figura 2.17 Combinador adaptativo lineal [17]

Por lo que tenemos el siguiente conjunto de señales:

- En primer lugar la señal de entrada, que es un conjunto de n señales que seguirá la siguiente fórmula:

$$X_k = [x_{1k}, x_{2k}, x_{3k} \dots x_{nk}]^T \quad \text{Fórmula 2.25}$$

- El conjunto de coeficientes designados con el siguiente vector

$$W^T = [w_1, w_2 \dots w_n] \quad \text{Fórmula 2.26}$$

- La señal de salida k -ésima

$$y_k = \sum_{l=1}^n w_l x_{lk} = W^T X_k \quad \text{Fórmula 2.27}$$

- Suponiendo la señal deseada como d_k , el error en el instante k -ésimo es:

$$e_k = d_k - y_k = d_k - W^T X_k \quad \text{Fórmula 2.28}$$

- El error cuadrático medio es el valor esperado de e_k^2

$$MSE = E[d_k^2] - 2P^T W + W^T R V \quad \text{Fórmula 2.29}$$

Siendo P la correlación cruzada entre la señal de entrada x y la señal deseada d , y R la autocorrelación de la señal de entrada $E[x_k x_k^T]$. Se puede observar que el MSE es una función cuadrática de los coeficientes, con lo que tendrá la forma de una parábola. El proceso adaptativo estará continuamente ajustando los coeficientes, buscando la parte más baja de la parábola. Para obtener los coeficientes óptimos habrá que derivar el MSE e igualarlo a cero. [17]

Además, el algoritmo Least Mean Square (LMS) utiliza el método steepest descent. Éste algoritmo realiza los cambios proporcionalmente al gradiente como se puede observar en la siguiente ecuación:

$$W_{k+1} = W_k + \mu (-\nabla_k) \quad \text{Fórmula 2.30}$$

Donde ∇ es la estimación del gradiente.

$$\nabla_k = \nabla + N_k \quad \text{Fórmula 2.31}$$

y N_k el ruido del gradiente.

Finalmente, el gradiente se define como la derivada del cuadrado del error respecto a cada uno de los coeficientes. Si se hace esa derivada, se obtiene como conclusión que el gradiente tiene el valor: [17]

$$\nabla_k = -2e_k X_k \quad \text{Fórmula 2.32}$$

Obteniendo finalmente la ecuación que rige el algoritmo LMS

$$W_{k+1} = W_k + 2 \mu e_k X_k \quad \text{Fórmula 2.33}$$

2.7 ESPECTROGRAMA

El espectrograma es una representación de las distintas variaciones espectrales en el eje vertical y de la intensidad del sonido a lo largo del tiempo en el eje horizontal.

Para la obtención del espectrograma es necesario la aplicación de una transformada de Fourier a la señal. En relación de la ventana utilizada para el análisis de Fourier se obtendrá diferentes niveles de detalle o resolución del espectrograma, por lo que si la ventana es muy grande el espectrograma será muy detallado, y por otra parte si es muy pequeño el efecto será inverso imposibilitando la distinción de los distintos armónicos en el espectrograma. No todo van a ser beneficios, debido a que si la ventana es muy grande el coste computacional será aumentado en gran medida como era de suponer.

Por lo tanto, la funcionalidad del espectrograma es generalmente para el análisis de la intensidad, el ritmo, la sonoridad, la duración y la estructura de los distintos formantes denominados timbre.

3. DESARROLLO

En el siguiente apartado explicaremos el desarrollo seguido para la implementación del algoritmo de discriminación del locutor mediante el pulso glotal. En primer lugar, el software utilizado ha sido "MATLAB" junto con la ayuda de "Audacity". Mediante Audacity hemos seleccionado la parte interesada del audio a analizar, y mediante MATLAB y los fundamentos teóricos adquiridos en el grado realizaremos el software necesario. Dentro de MATLAB, utilizaremos diversas funciones ya implementadas como puede ser LPC, como otras funciones que implementaremos.

Podemos dividir nuestro proyecto en tres grandes módulos, el primero será la obtención del trozo de audio a analizar junto con sus características el cual llamaremos módulo de análisis, el segundo será el módulo de obtención el cual se encargará de obtener las características glotales de cada audio a analizar. Por último pero no menos importante, el tercer módulo será el de interpretación de datos que nos servirá para ver la efectividad del algoritmo utilizado como su credibilidad.

3.1 PLANTEAMIENTO GENÉRICO

El desarrollo de nuestro proyecto se podrá resumir por lo tanto en tres partes.

- 1) En primer lugar, grabaremos al locutor una frase sencilla para luego escoger una palabra de esa frase que será la misma para todos los locutores, analizaremos dicha palabra para más tarde con diversas técnicas aplicadas a ésta (como inventanar la señal para luego aplicarle el LPC y así obtener el residuo) obtendremos la parte que contiene información de dicha señal, es decir, la obtención del pulso glotal.
- 2) Una vez obtenida la señal que del pulso glotal, trabajaremos con ella tanto temporal como frecuencialmente para así obtener diversos parámetros que puedan caracterizar dicha señal, para así en el siguiente módulo poder compararlos entre ellos.

3) Finalmente, una vez obtenidas las características glotales a analizar, las compararemos con las diversas grabaciones para así finalmente poder obtener un porcentaje de error a nuestro algoritmo.

3.2 MÓDULO DE ANÁLISIS

A continuación, realizaré una explicación de la implementación del módulo de obtención. El módulo de obtención, será el encargado de obtener los pulsos glotales a partir de una pista de audio grabada con un micrófono, es decir, éste módulo será el más importante debido a que con una captación buena de los datos, obtendremos unos resultados más fáciles de analizar y por lo tanto unos resultados más óptimos. El procedimiento dentro de este módulo, en líneas generales será el siguiente:

En primer lugar, grabaremos al locutor con una frecuencia de muestreo de 8000 Hz generalmente utilizada para teléfonos, la frase "había cuatro vacas" seleccionando únicamente "cuatro" de toda la frase, lo hacemos así para que la glotis actúe de forma normal dentro de una frase, es decir, sin estar forzada para así obtener un resultado lo más parecido a la normalidad.

A continuación, enventanaremos la señal mediante el tipo de ventana Hanning con un 50% de solape para más tarde aplicarle el LPC (Linear prediction filter coefficients).

Finalmente, una vez aplicado el LPC a la señal, realizaremos el filtrado inverso para así obtener la señal residuo, que se caracteriza por contener toda la información de la glotis, que es la que queremos analizar y clasificar.

En los siguientes subapartados, explicaremos todos los pasos anteriormente mencionados con más detalle, detallando su propósito y la función que ejercerá en un futuro además de justificando las decisiones tomadas para la realización de su implementación.

3.2.1 MÓDULO DE TRATAMIENTO PREVIO DE LA SEÑAL.

En este módulo se realizará una adecuación de la señal para su posterior tratamiento. A éste módulo le llegará la señal de voz digitalizada mediante un micrófono. En primer lugar, la señal grabada será monocal, es decir, solo tendrá un canal de información para trabajar únicamente con lo necesario por lo que realizaremos un programa que grabe 4 segundos que convertirá la pista en monocal, así nos ahorraremos envolventes o factores que puedan alterar la información de dicha señal. A continuación, estableceremos una frecuencia de muestreo de 8000 Hz, o lo que es lo mismo, 8000 muestras por segundo. Dicha frecuencia es utilizada generalmente para telefonía, además tenemos que recordar que el rango de frecuencia de la señal de voz está comprendida de los 0 a los 4000 Hz.

Cada locutor dirá la frase "Había cuatro vacas" de la cual únicamente seleccionaremos la palabra cuatro para ser posteriormente analizada, a continuación veremos cómo mediante el programa Audacity seleccionamos la palabra "cuatro" para ser luego analizada con más detalle en MATLAB. Una vez recogida la señal correspondiente a "cuatro" la normalizaremos para que a la hora de trabajar con ella sea más funcional.

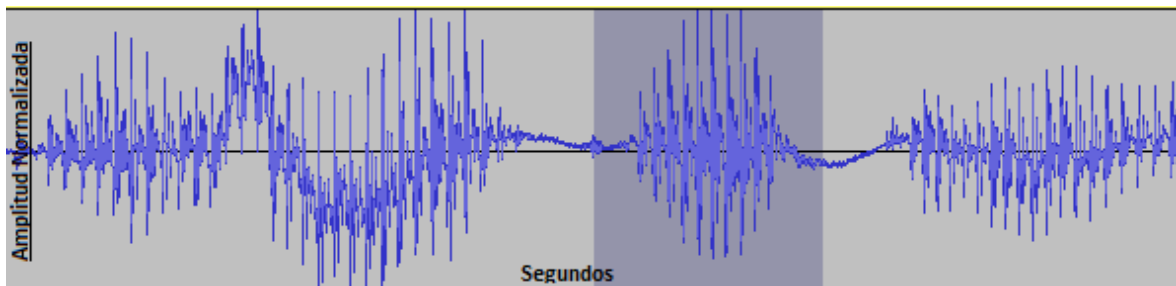


Figura 3.1 Señal perteneciente a Había cuatro vacas, apareciendo la parte del cuatro seleccionada.

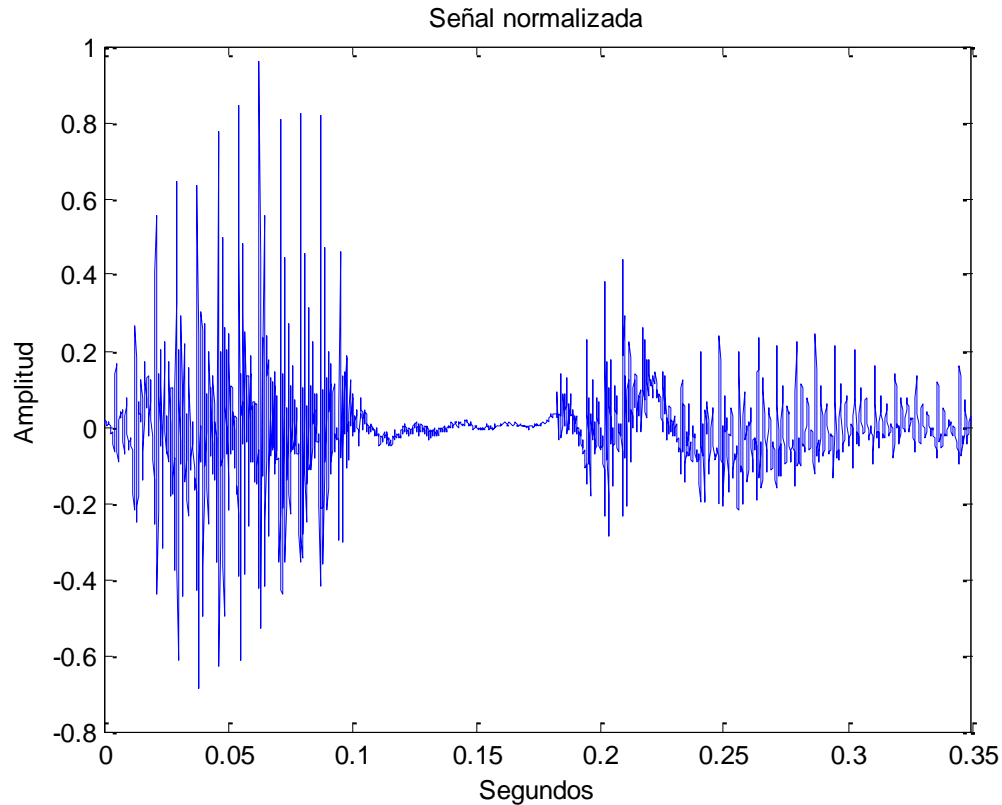


Figura 3.2 Palabra 4 Normalizada representando Amplitud frente a Tiempo

3.2.2 MÓDULO DE ENVENTANADO.

En el siguiente módulo realizaremos el enventanado de la señal, es decir, segmentaremos la pista de audio en diversas ventanas temporales iguales o inferiores a 20 milisegundos más pequeñas para luego ser procesadas y así no introducir un solape espectral. Hay diversos tipos de ventana como la Kaiser, la Hamming, la Hanning, pero para éste proyecto la más óptima será la ventana Hanning con 20 milisegundos de ventana temporal. Dicha ventana es la más óptima debido a que no introducirá una componente continua, como hace la Hamming. A la hora de implementar el algoritmo, utilizamos un solape del 50% dando como resultado una señal que apenas ha sido modificada, pudiendo reconstruirla sin apenas modificaciones.

Como ya hemos comentado anteriormente, debido a que la frecuencia de muestreo es de 8000 Hz, es decir, 8000 muestras por segundo, si queremos conseguir una ventana temporal de 20 milisegundos deberemos de coger de 160 en 160 muestras. Además, debido a que tendrá un solape del 50% el enventanado de la señal se tendrá que producir cada 80 muestras, un ejemplo para entender lo anteriormente explicado sería el siguiente: Si la primera ventana abarca de la muestra 1 a la 160, la siguiente debería de abarcar de la 80 a la 240 y así sucesivamente, dando como resultado una multiplicación de toda la señal componente por componente por la ventana elegida.

A continuación en las siguientes gráficas podremos observar una sección de 160 muestras a la cual ya se ha aplicado el enventanado, y la señal total tras ser enventanada, que debería de ser muy parecida a la original.

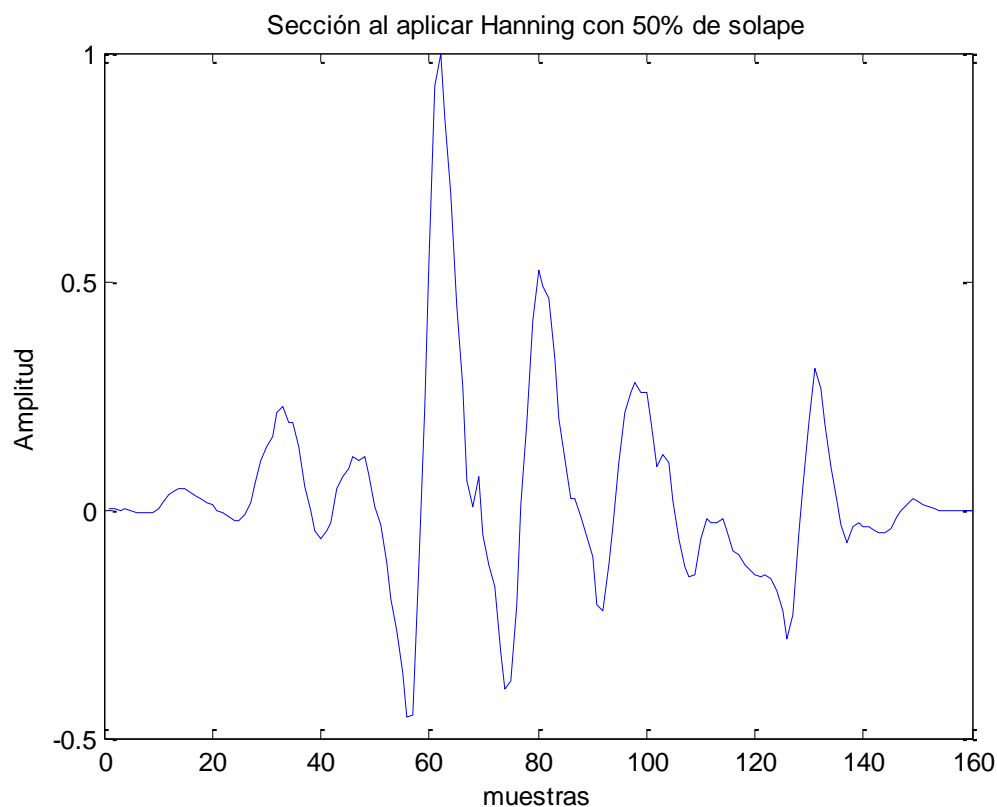


Figura 3.3 160 muestras al ser aplicadas mediante enventanado de Hanning.

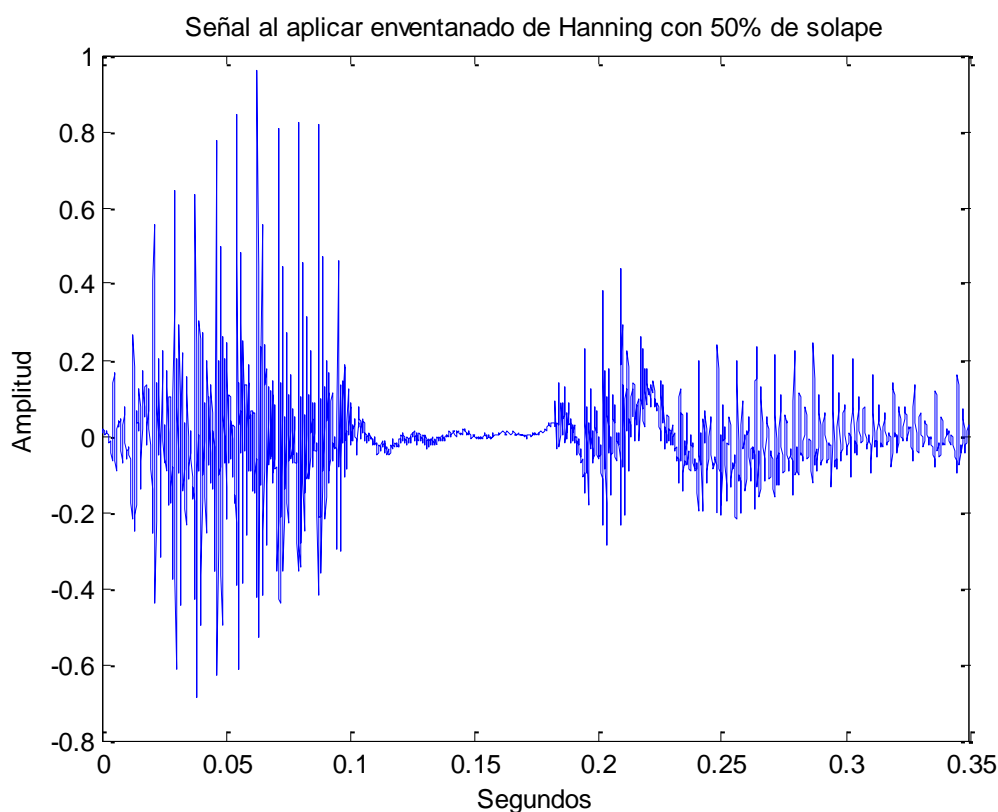


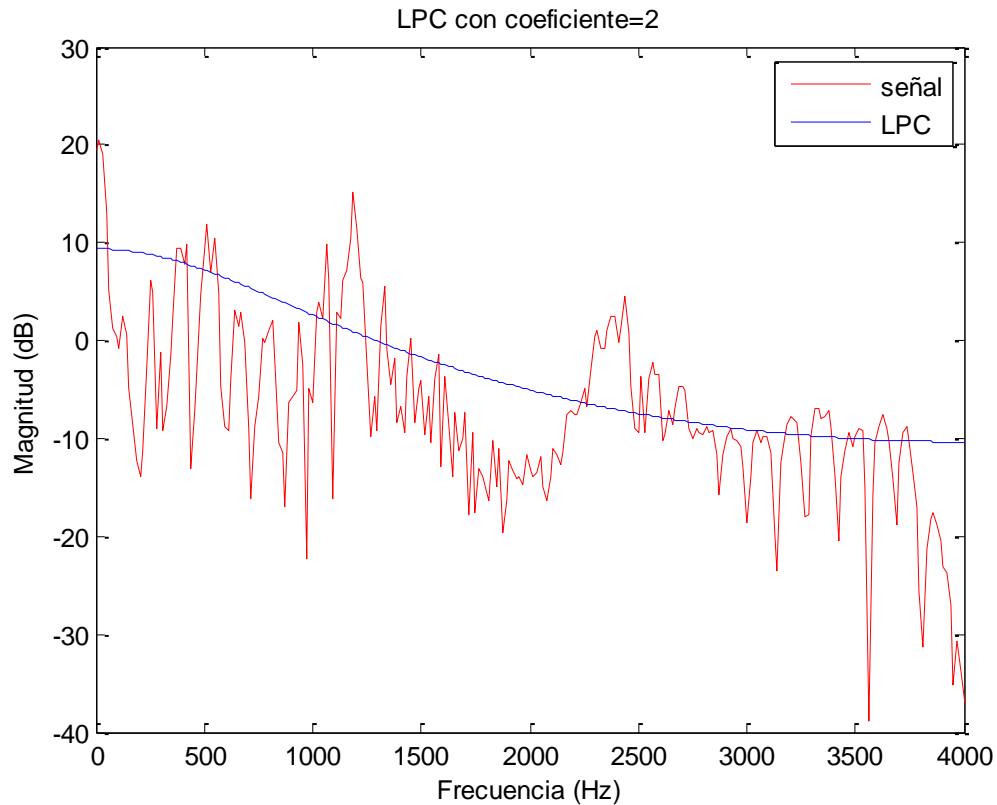
Figura 3.4 Señal al ser reconstruida después de aplicar el enventanado.

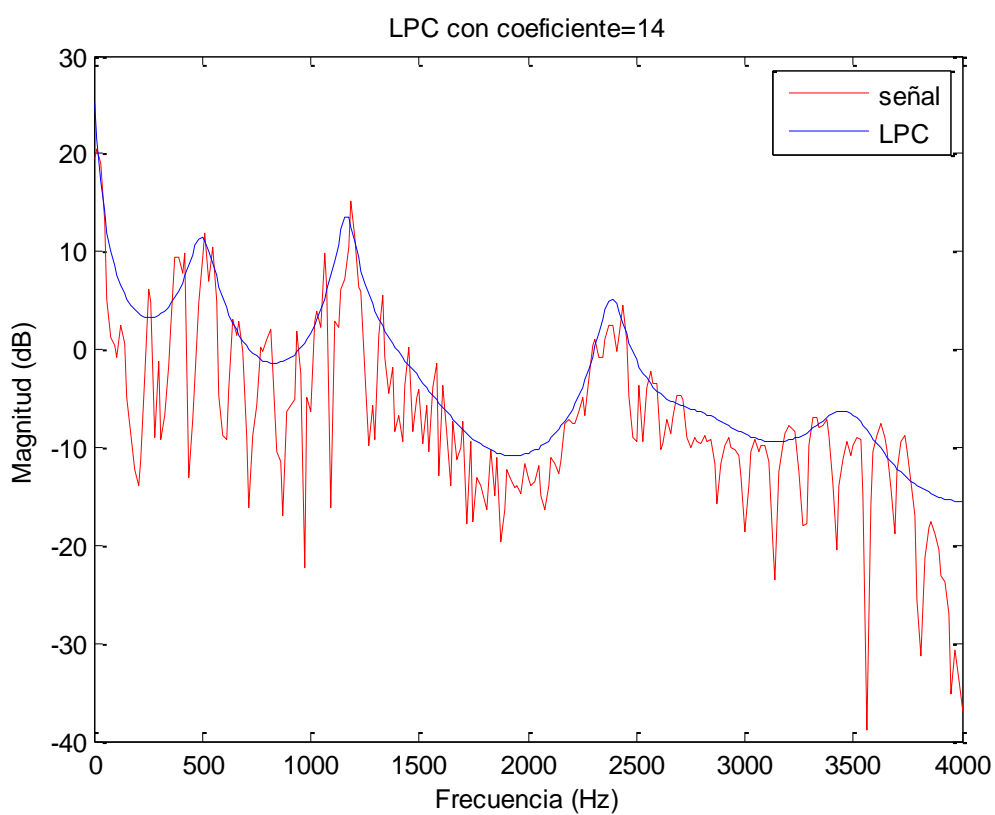
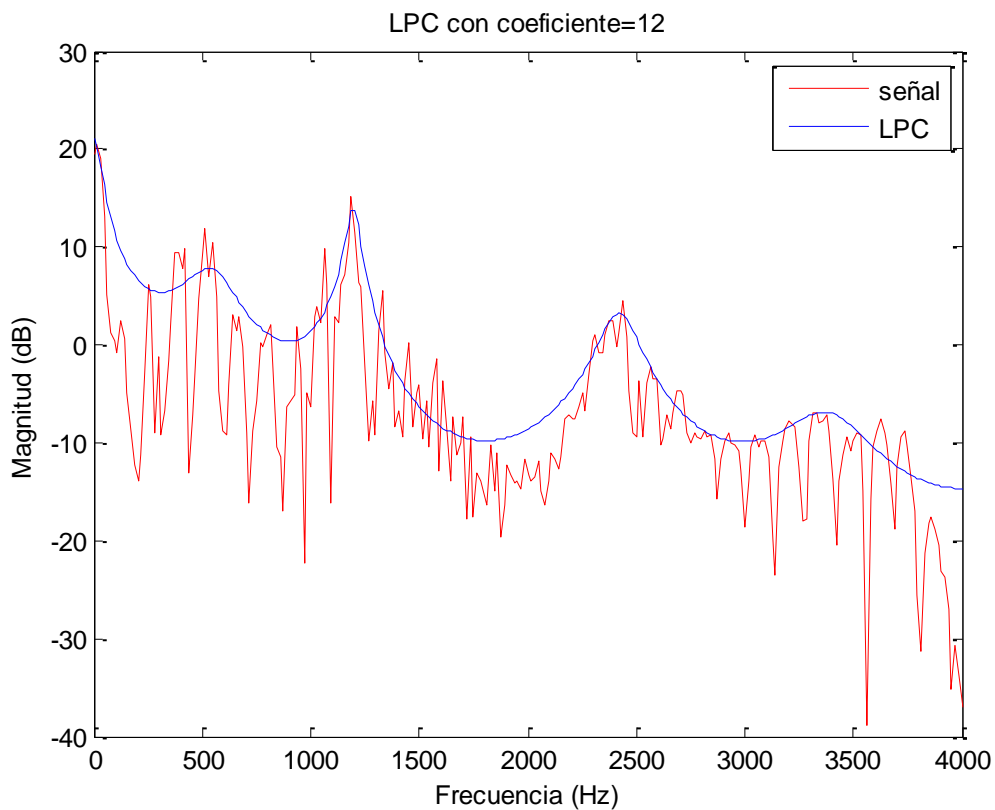
3.2.3 MODULO DE FILTRO DEL TRACTO VOCAL Y CONSTRUCCIÓN DE RESIDUO.

En el siguiente módulo, calcularemos el filtro del tracto vocal de cada segmento de voz a analizar. Una vez obtenido dicho filtro, aplicaremos el filtrado inverso del tracto vocal lo que dará como resultado la eliminación de los efectos formantes sobre cada segmento de audio. El resultado de todo lo anteriormente mencionado es el residuo, donde se encontrará la información que queremos analizar, es decir, los pulsos glotales.

Para calcular los coeficientes del filtro utilizaremos el algoritmo de Durbin que está implementado en MATLAB con la función "LPC" el cual si le pasamos un segmento de audio y el grado del predictor, nos devolverá los coeficientes necesarios para la implementación del filtro junto con la potencia del error de predicción.

En cuanto al grado del predictor lineal, a mayor grado mayor será la exactitud, es decir, menor será el error obtenido. A continuación mediante la gráfica que introduciré, podremos comprobar el espectro de una señal de un segmento de voz frente a la forma de ondas de los diversos filtros de tractos vocales que hemos extraído con el algoritmo de Durbin, variando entre cada una de ellas el grado del predictor.





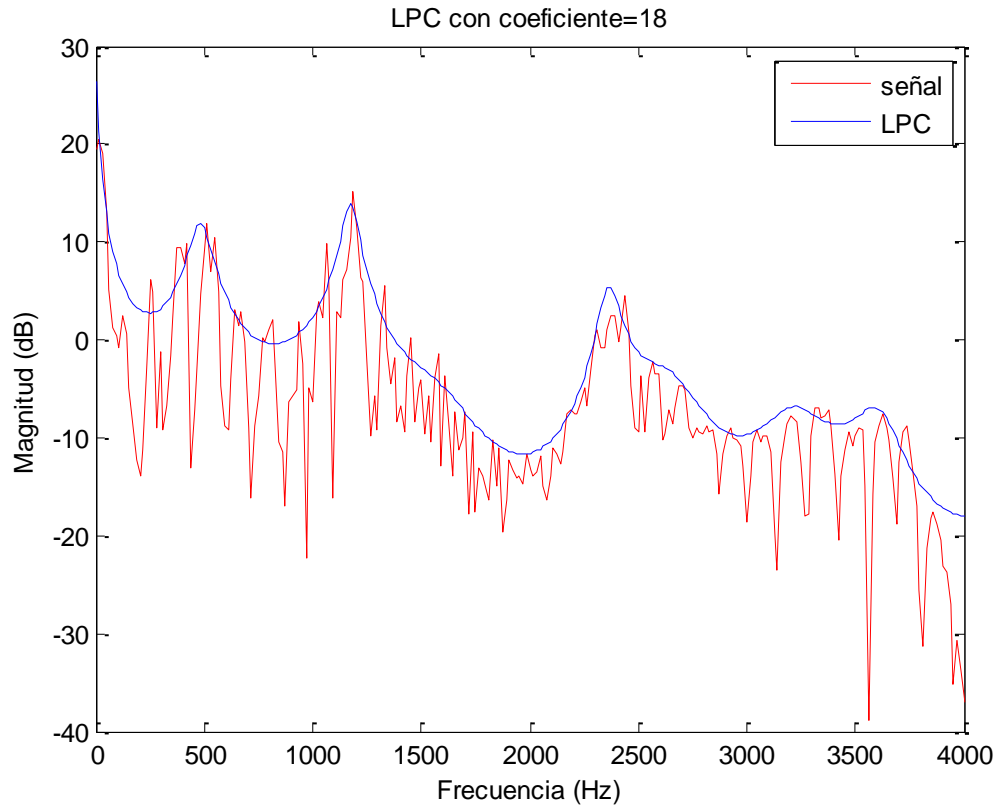


Figura 3.5 (a) Representación de la forma de onda del filtro de tracto vocal con grado 2, (b) grado 12, (c) grado 14, (d) grado 18.

Como se puede apreciar, a mayor grado del filtro, mayor exactitud. Sin embargo, debido a que estamos calculando todo al momento, es decir, de manera iterativa, a mayor grado mayor coste computacional por lo que tenemos que llegar a una solución de compromiso. El grado óptimo para analizar la voz está entre 10 y 14 debido a que menor de 10 nos ofrecerá una curva que distará demasiado de la real, y cuando sea mayor de 14 no obtendremos unos resultados muy notorios pero el coste computacional se incrementará en gran medida, por lo que nuestra solución de compromiso será 12 debido a que con grado 10 en transiciones muy bruscas no se ajusta lo demasiado, y en grado 14 no se nota mucha diferencia con nuestra solución propuesta. Dicho de otra manera, hemos considerado que un grado doce se ajusta equitativamente tanto en forma de onda como en coste computacional.

Una vez obtenido los diversos coeficientes del filtro junto con su error, tendremos que definir el filtro que aplica nuestro tracto vocal hasta emitir el sonido que será el siguiente:

$$y = \frac{b_o}{a_k} \quad b_o = \sqrt{e} \quad \text{Fórmula 3.1}$$

y = Señal de audio emitida

e = potencia del error de predicción.

a_k = Coeficientes para la obtención del filtro del tracto vocal

Debido a que nuestra intención es eliminar dicho filtro, es decir, realizar el filtro inverso para obtener la señal residual, aplicaremos a toda la señal de audio el siguiente filtro que será programado mediante un algoritmo.

$$\text{Filtro inverso} = (y)^{-1} = \frac{a_k}{b_o} \quad \text{Fórmula 3.2}$$

A continuación, podremos observar la diferencia de una parte de la señal al aplicar el filtrado inverso, es decir, al obtener el residuo de ésta.

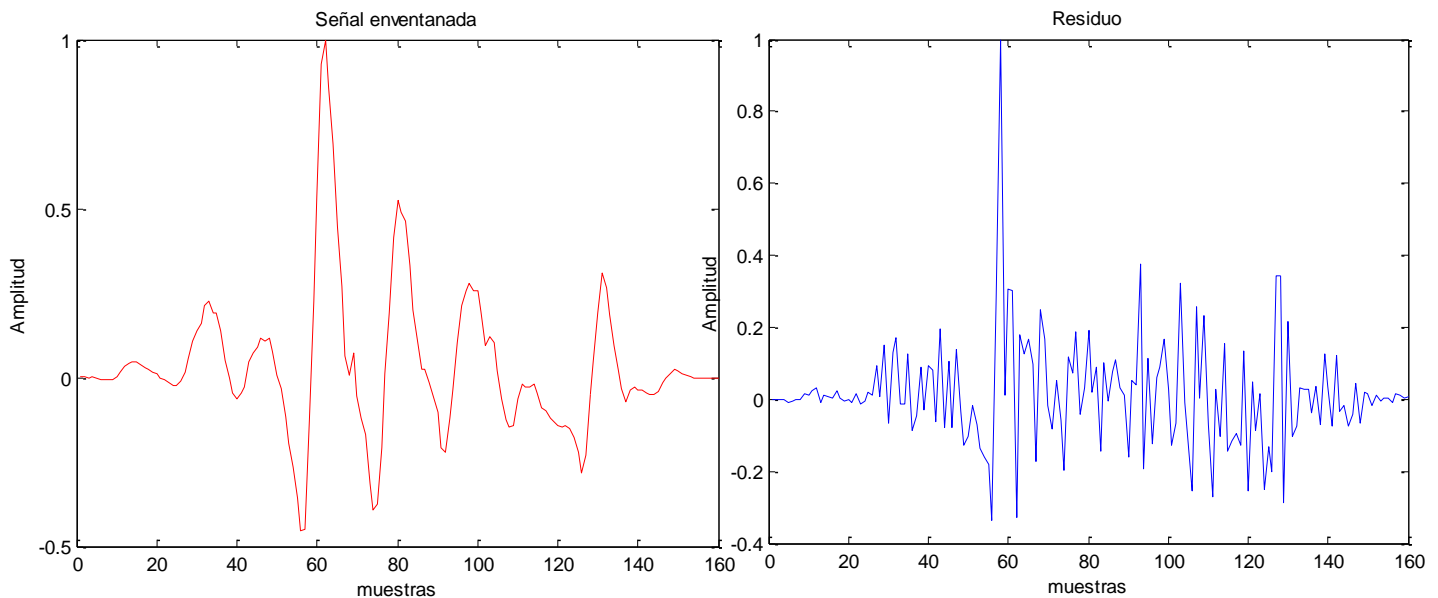


Figura 3.6 (a) Un trozo de 160 muestras de la señal, (b) Residuo de dicho trozo.

Como comentaremos posteriormente, la información útil será localizada en los picos de los pulsos glotales y las muestras continuas a éstos, debido a que el resto es señal sin información.

Tras haber conseguido el residuo de un enventanado, deberemos deshacer la segmentación realizada para el módulo de enventanación, es decir, el solape de la reconstrucción deberá de ser el mismo que el mencionado previamente, o sea del 50%. Si todo éste proceso es realizado correctamente obtendremos una señal de residuo o señal de excitación de nuestro sistema fonador buena. Una vez obtenida la señal residuo podremos diferenciar los distintos pulsos glotales. A continuación mostraré una gráfica con la representación de la señal residuo.

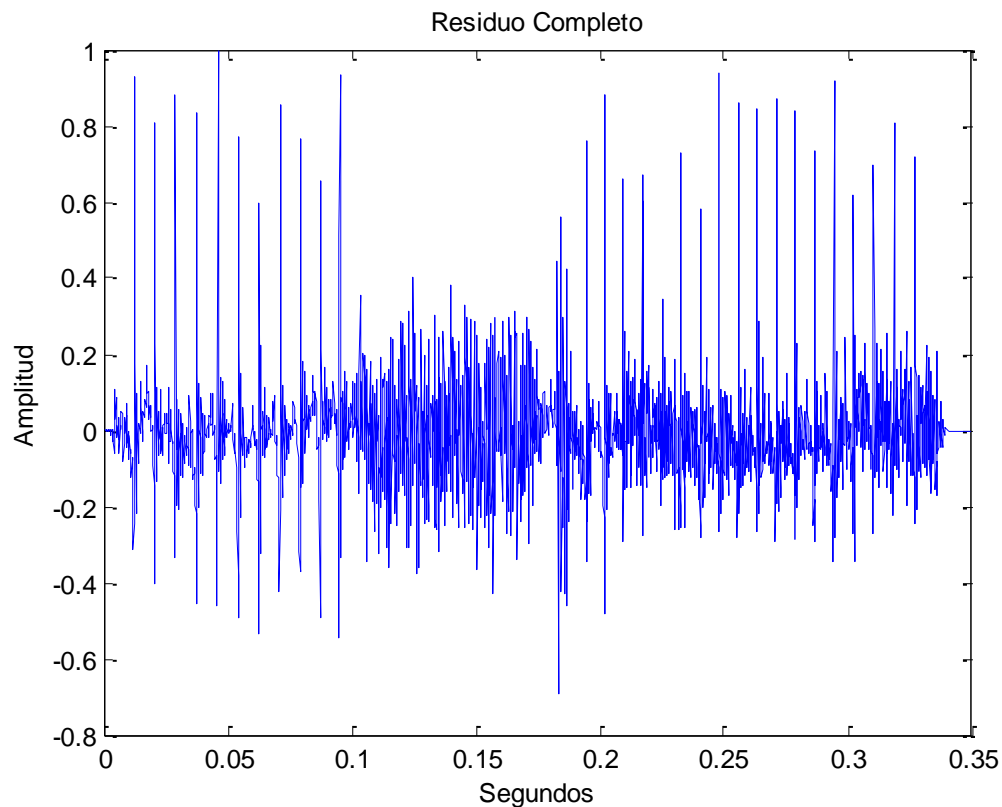


Figura 3.7 Señal Residuo reconstruida.

Además como se ha explicado en el marco teórico, se puede apreciar que los pulsos glotales ofrecen información redundante donde se encuentran las vocales, en nuestro caso la primera parte pertenece a "ua", luego va dividida por información sin utilidad para finalmente finalizar en "o" de la palabra "cuatro".

3.2.4 MÓDULO DE EXTRACCIÓN DE LOS PULSOS GLOTALES.

En el siguiente módulo realizaremos la extracción de los pulsos glotales a partir de la señal del residuo. Los pulsos glotales se caracterizan por un pico rodeado de ruido, esto es debido a que la señal de residuo se compone con el flujo de aire de las cavidades infragloticas, los pulmones que se consideran ruido blanco y las vibraciones de los impulsos vibratorios proporcionados por la glotis.

Como hemos podido observar en la gráfica de residuo, el pulso se genera pseudoperiódicamente debido a que siempre existirá cierto grado de modificación debido a su generación temporal. Hay que recordar que una misma persona tiene un rango de valores de Pitch que ocasionará que nunca tengamos la voz exactamente igual, aun así esto no nos influirá debido a que nuestra intención no es encontrar un patrón temporal o frecuencial, nuestra intención será en la forma de onda en sí, es decir, nuestro algoritmo se centrará en la extracción de los pulsos glotales.

Para la detección y extracción de los pulsos glotales deberemos fijarnos en la forma de onda del residuo. En primer lugar realizaremos una normalización de la señal para que así el máximo esté siempre encontrado en uno, una vez hemos normalizado podemos comprobar que los pulsos glotales aparecen como máximos muy diferenciados respecto al ruido introducido por los pulmones, dicho de otra manera, las zonas que tendremos que seleccionar son las que suponen picos con una amplitud mayor respecto a sus muestras mas cercanas, escogiendo así los pulsos glotales con mayor amplitud. Nuestra hipótesis se fundamentará por lo tanto en que a mayor amplitud, mayor nitidez dando como resultado unos resultados más óptimos.

Debido a que la realización de todo el proyecto se fundamentará en la elección del pulso glotal adecuado, se ha realizado la derivada de la señal que contiene el residuo para así ofrecer unos picos más pronunciados lo cual nos ayudará a nosotros para elegir el umbral establecido para la elección de los pulsos óptimos, a continuación colocaremos dos gráficas para poder apreciar la diferencia entre la señal y la derivada de ésta en cuanto a amplitud nos referimos.

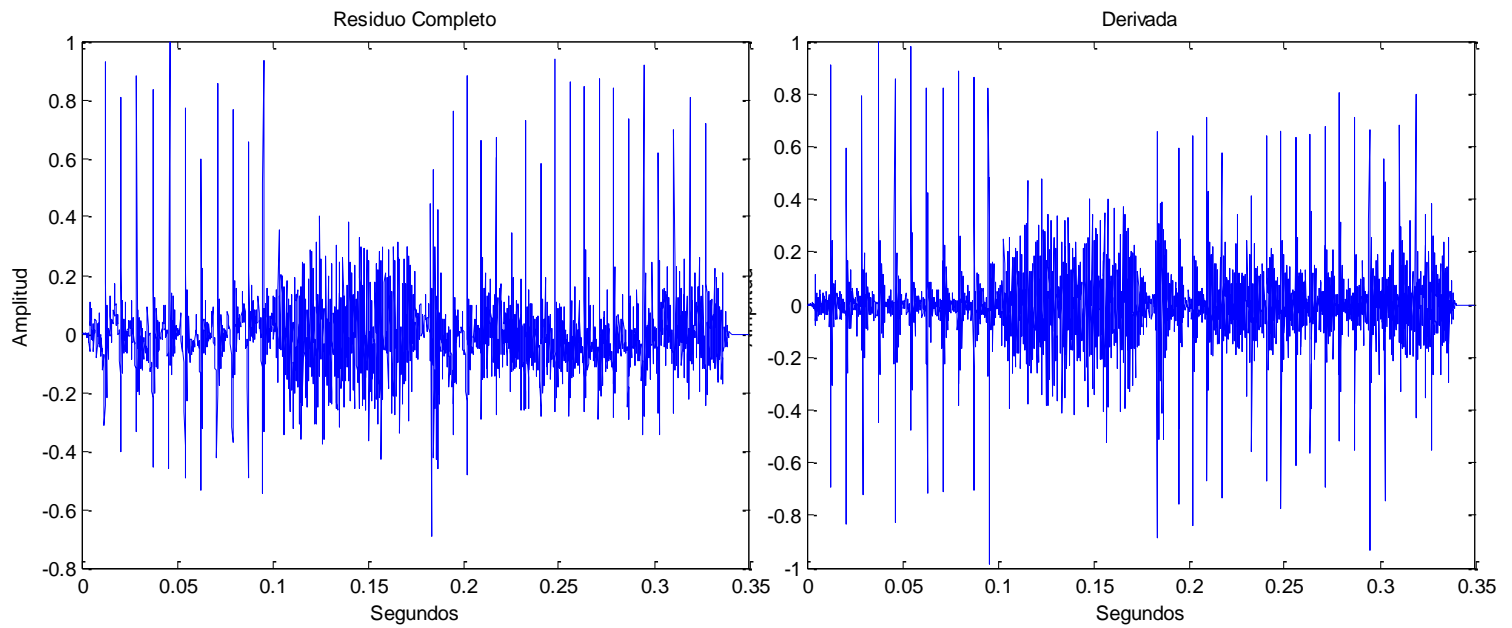


Figura 3.8 (a) Señal del Residuo, (b) Señal de la primera derivada del Residuo.

Como podemos comprobar, los picos de amplitud han sido pronunciados respecto al ruido, esto nos ofrecerá un beneficio abismal debido a que a continuación lo que haremos será escoger los pulsos glotales óptimos mediante la decisión de un umbral en la señal derivada para más tarde coger la misma posición de dicho pulso pero de la señal del residuo, debido a que nosotros tenemos que trabajar con la señal del residuo no con su derivada, ésta última únicamente nos ofrecerá una ayuda para nosotros.

Una vez escogidos los picos óptimos que tendremos que extraer de dicha señal, podremos afirmar que éstos serán los más óptimos para su reconocimiento de la forma ofreciéndonos a lo largo del programa unos resultados óptimos. Por último, únicamente nos quedaría decidir el número de muestras que consideraremos como pulso glotal. La primera vez que decidimos el número de muestras que cogeríamos fue un total de 20 muestras para atrás y 29 para adelante, dándonos un total de 50 muestras.

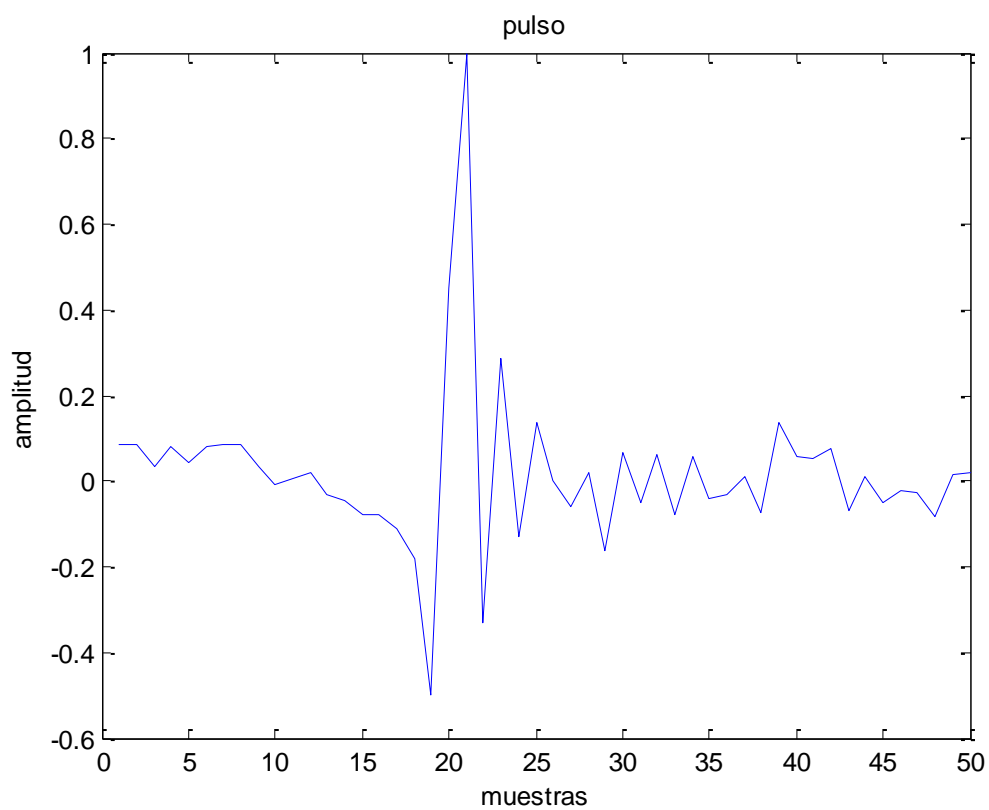
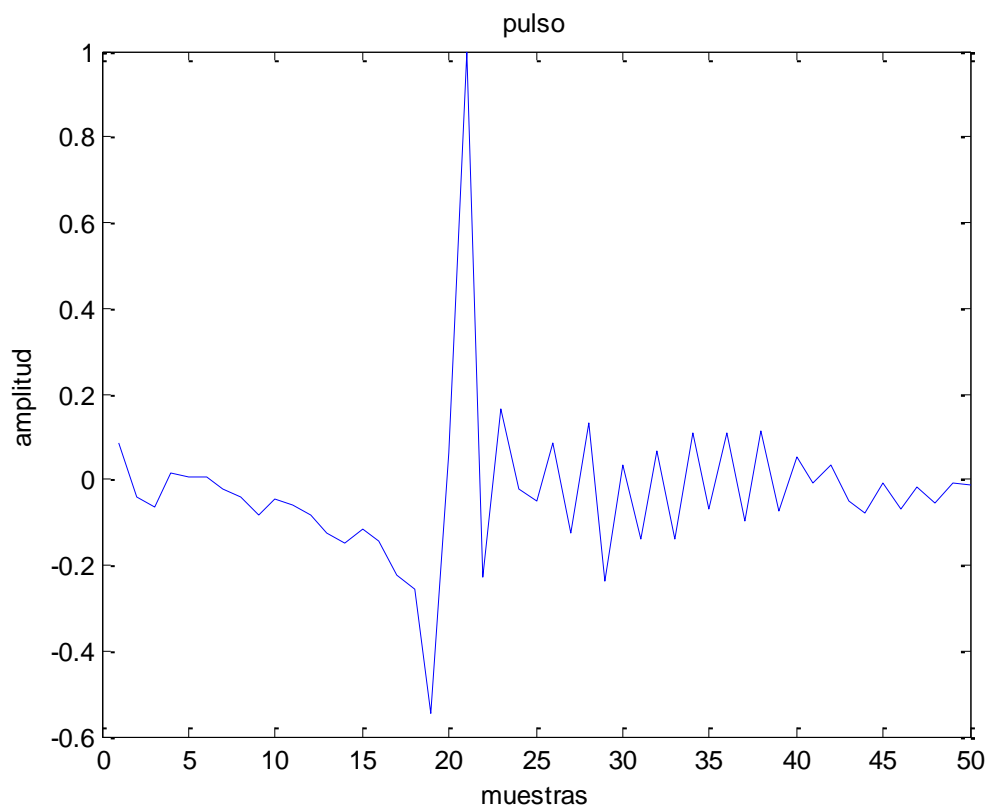


Figura 3.9 Dos pulsos de la señal de residuo del mismo locutor con 50 muestras.

Más tarde, nos dimos cuenta que la información detrás del pico era demasiado amplia debido a que había la posibilidad que estuviéramos cogiendo información de otro pulso por lo que decidimos reducirla de 20 muestras a 5, dándonos un total de 35 muestras y aquí observamos la diferencia.

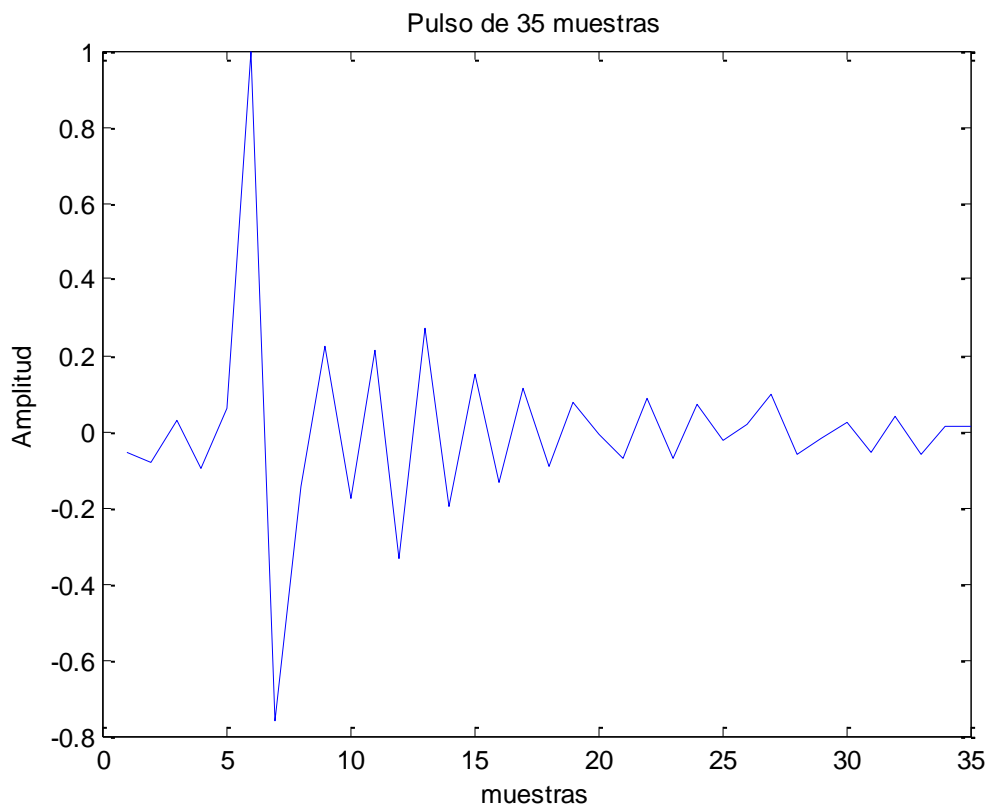


Figura 3.10 Pulso glotal de 35 muestras.

Una vez conseguido los diversos pulsos glotales con 35 muestras, procedimos a elevar la frecuencia de muestreo para obtener mayor resolución temporal. Con una frecuencia de muestreo de 8kHz la señal obtenida es muy abrupta, y necesitábamos más resolución, por lo que la elevamos a 32 kHz, quedando la señal resultante más suavizada.

Para variar la frecuencia de muestreo se realiza una interpolación y/o un diezmado, ambos algoritmos están implementados en la función `resample` de matlab. En este caso bastaba con una interpolación.

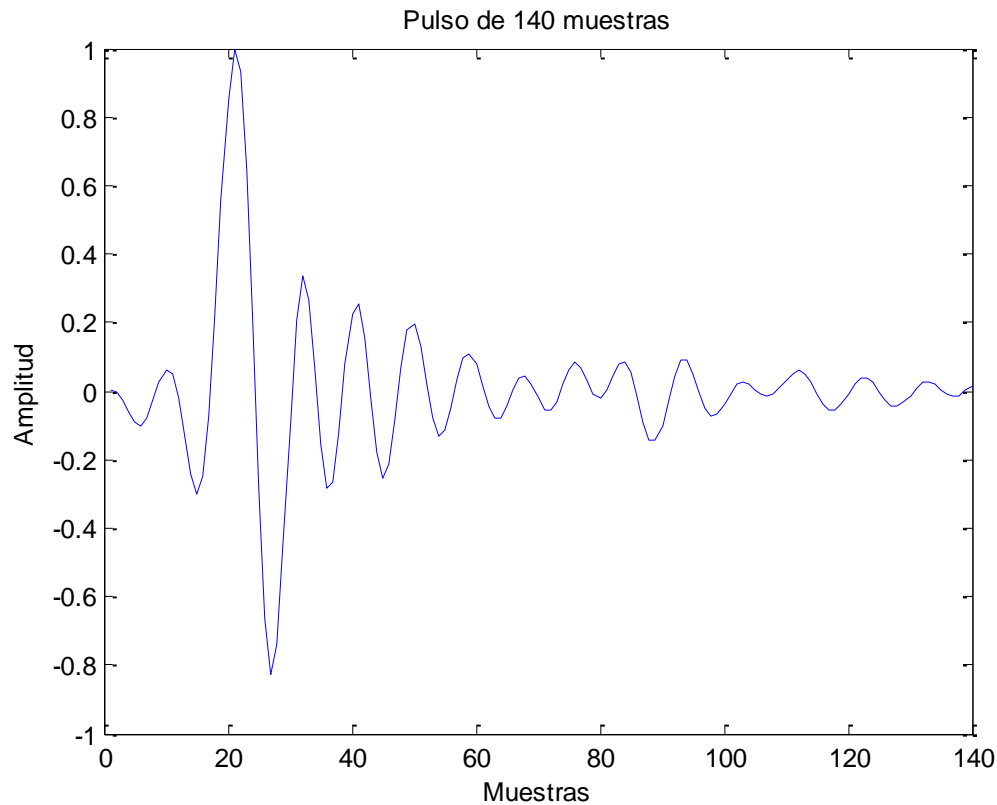
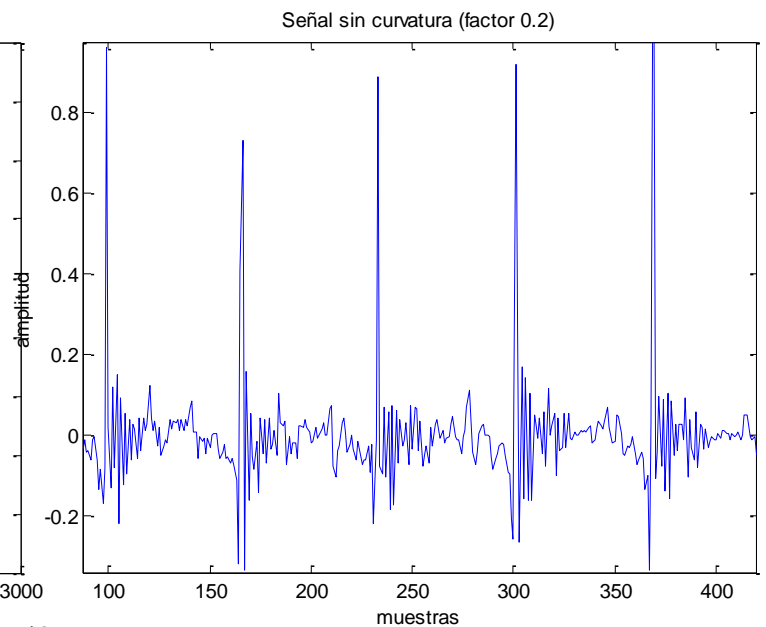
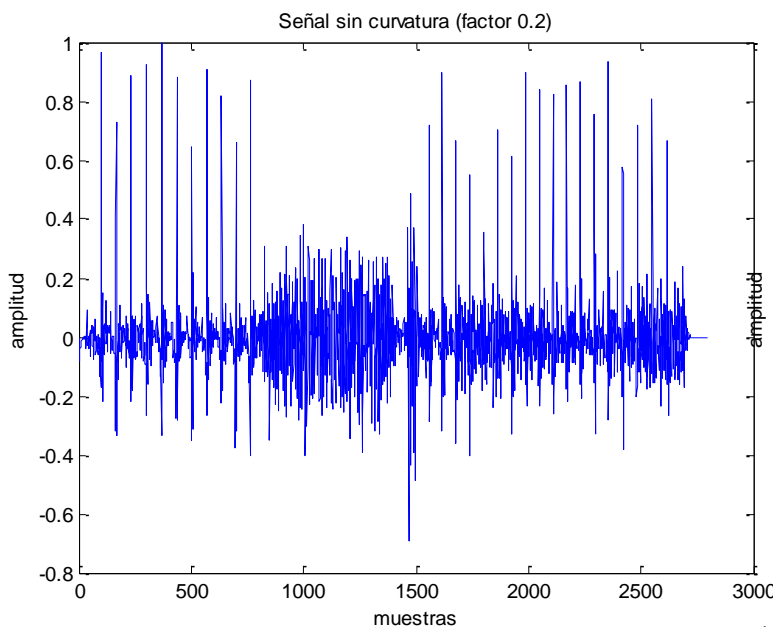
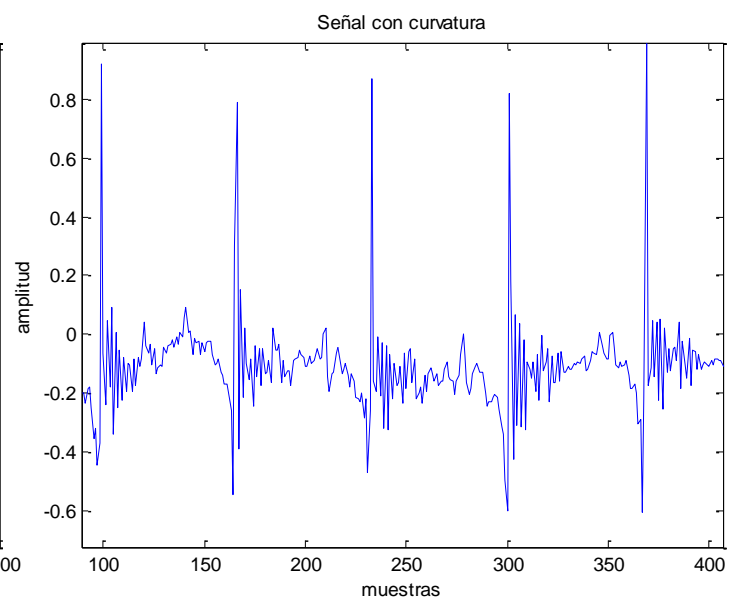
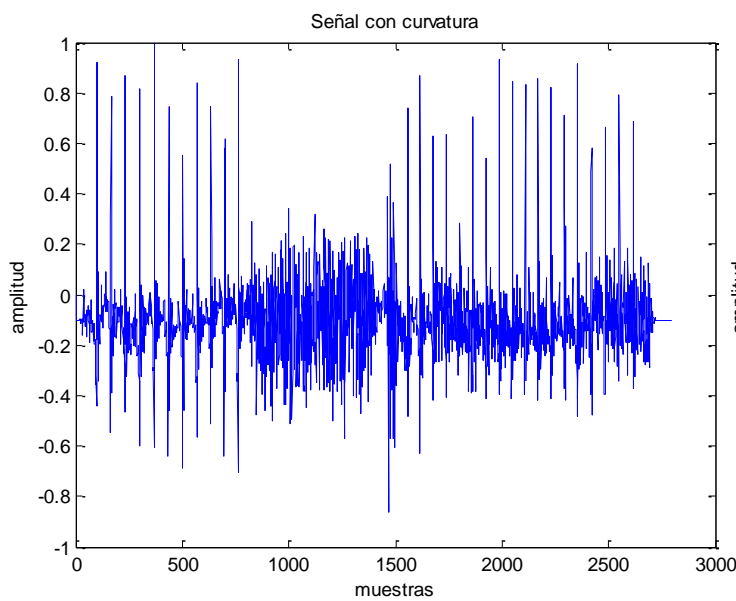


Figura 3.11 Pulso glotal de 140 muestras (tras haberle aplicado el resample).

Una vez hemos obtenido los diversos pulsos glotales, los almacenaremos en una matriz para en su posterioridad realizar la media de todos ellos, ahora mismo si realizásemos la media no sería correcta debido a que la señal de residuo contiene una curvatura que tendremos que eliminar mediante un filtrado adaptativo o "lms" que explicaremos en el siguiente módulo.

3.2.5 MODULO DE ELIMINACIÓN DEL RUIDO DE BAJA FRECUENCIA

A continuación, procedemos a eliminar las variaciones de baja frecuencia que tiene nuestra señal con la finalidad de tener todos los pulsos en torno a 0. Para ello, empleamos el algoritmo LMS variando su factor hasta obtener el resultado esperado.



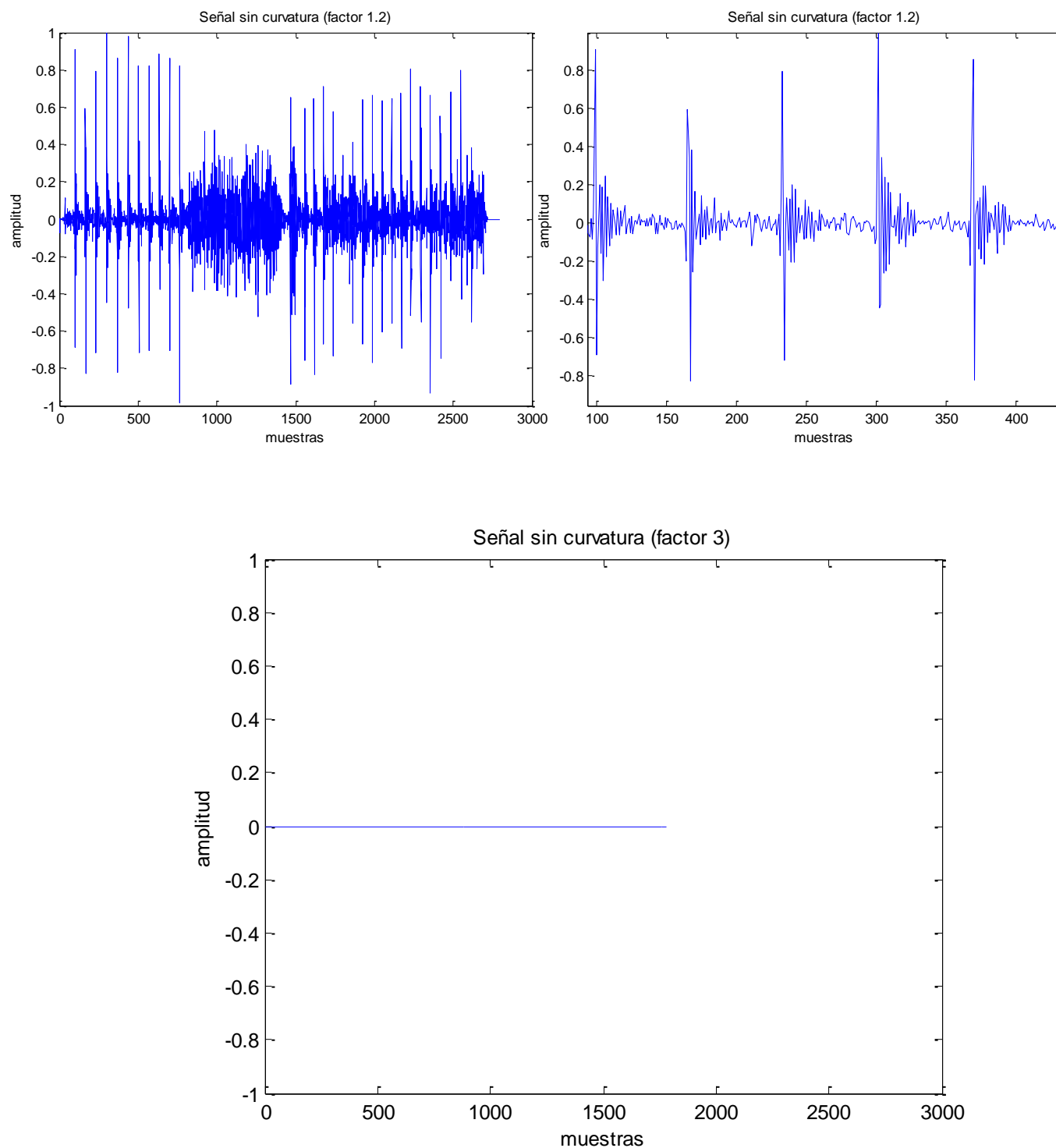


Figura 3.12 (a) Ejemplos de la señal original y ampliación de sus pulsos al aplicar un factor 0, (b) un factor 0.2, (c) un factor 1.2, (d) un factor 3.

Como podemos observar, conforme vamos aumentando el valor del factor X obtenemos una señal más recta, con 0.2 aún se puede apreciar un poco de curvatura, con 3 destruimos la señal por completo, por lo que la solución de compromiso ha sido escoger un valor igual a 1.2.

Este paso es muy importante debido a que como he mencionado anteriormente, a la hora de hacer la media de los distintos pulsos el resultado es muy distinto habiendo una curvatura debido a que no están en el mismo punto de partida, dándonos como resultado una media errónea.

3.2.6 MODULO DE OBTENCIÓN DEL PULSO GLOTA MEDIO.

Una vez hemos aplicado la función "resample" para aumentar el número de muestras y hemos aplicado el filtro adaptativo LMS (Least Mean Square) realizaremos la media de todos los pulsos obtenidos anteriormente. Es necesario hacer la media debido a que así conseguimos el pulso en el cual cada punto está equidistante del conjunto de puntos que tiene cada pulso, de esta manera obtenemos el pulso más característico y con menos variación de cada locutor.

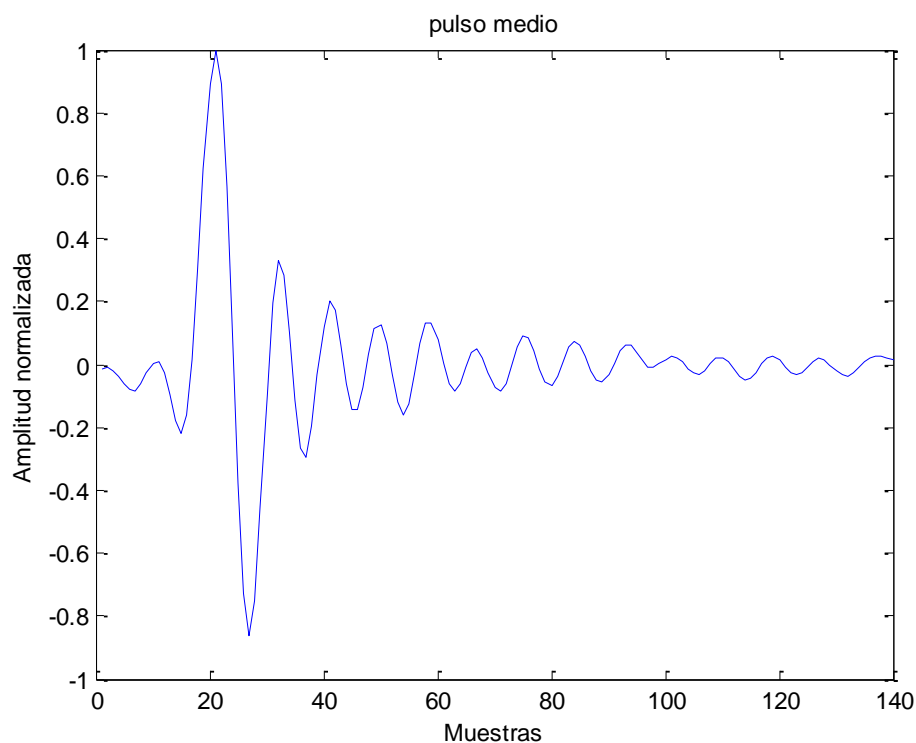


Figura 3.13 Representación del pulso medio de un locutor.

3.2.7 MÓDULO DE EXTRACCIÓN DE INFORMACIÓN A PARTIR DEL ESPECTROGRAMA

Como hemos mencionado anteriormente, el espectrograma realiza el cálculo del contenido espectral de las muestras según va variando en el tiempo. A continuación podremos observar que hemos realizado diversos espectrogramas sobre distintas señales y distintos locutores.

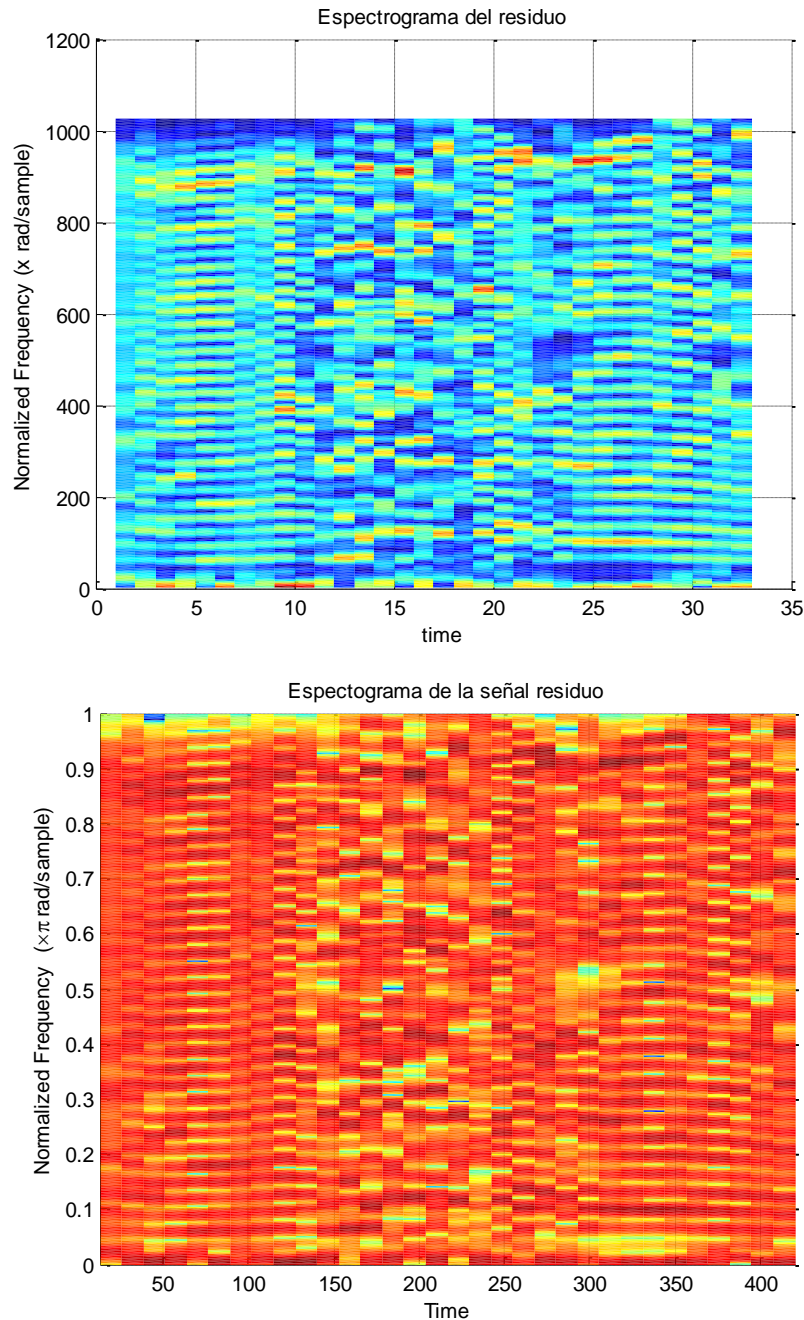
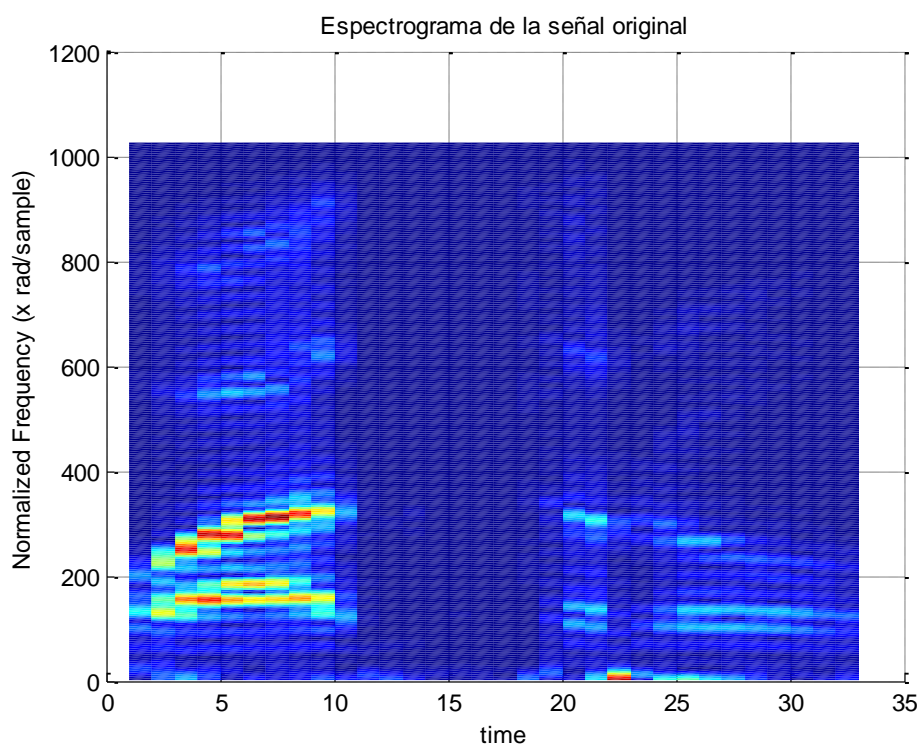
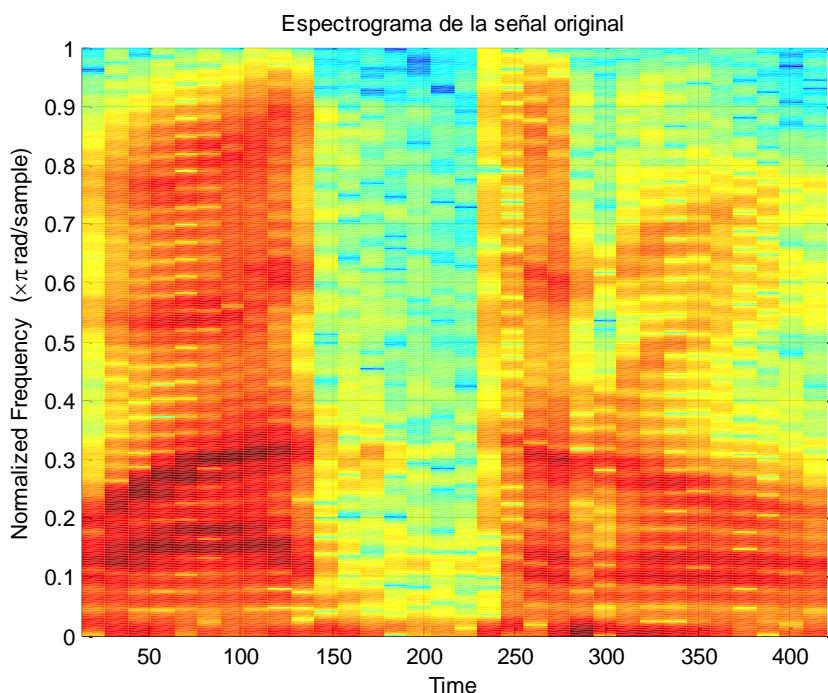


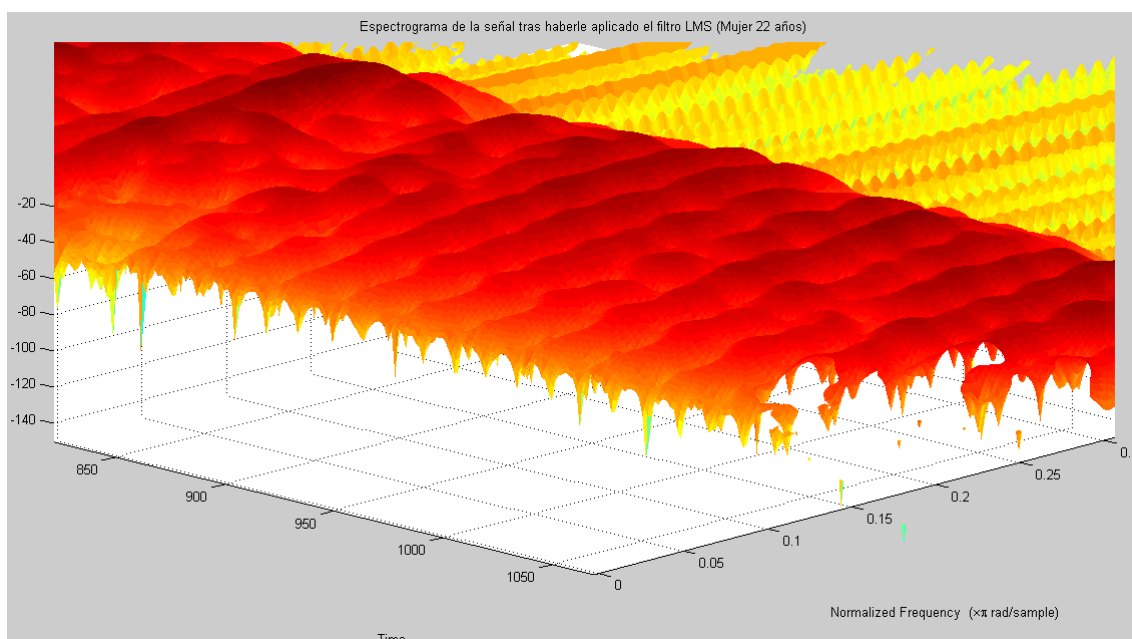
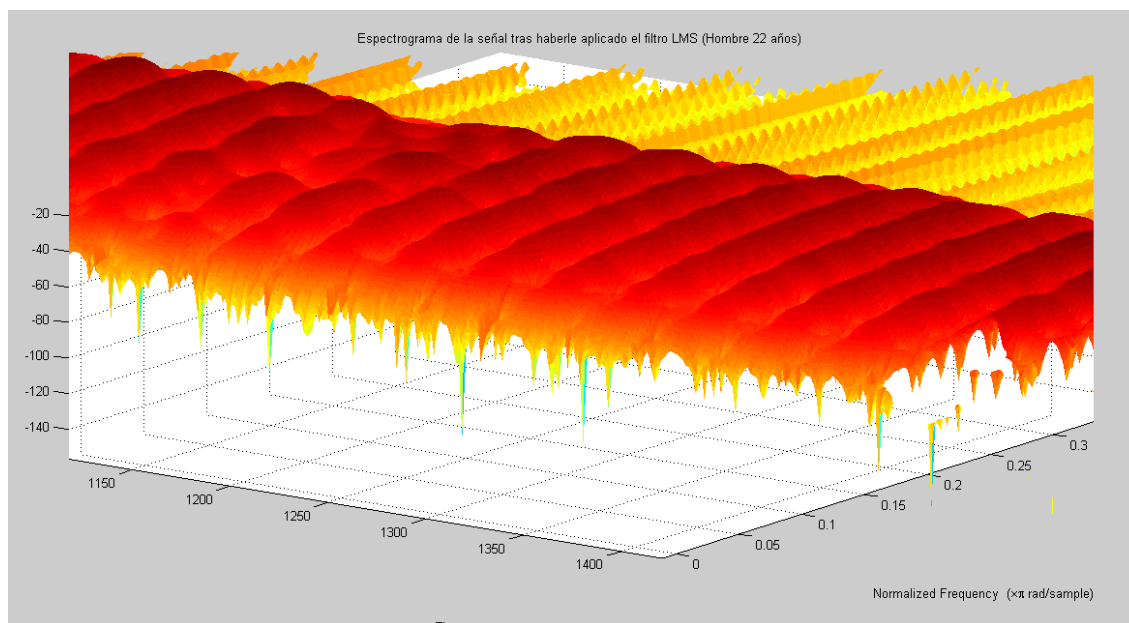
Figura 3.14 Espectrograma de la señal residuo obtenida en 2D representado de dos maneras.

Como podemos comprobar, todo el contenido de la señal tiene un gran índice de energía, es decir, todo el contenido contiene información relevante.



3.15 Espectrograma de la señal de entrada de audio en 2D representado de dos maneras.

Como podemos comprobar, la diferencia entre la representación de la señal original frente a la representación de la señal residuo, en la señal original únicamente se puede encontrar información útil donde están colocadas las vocales, esto ocurre como ya he explicado anteriormente debido a la glotis.



3.16 Comparación del espectrograma en 3D de un hombre de 22 años y una mujer de 22 años.

Finalmente, hemos realizado el espectrograma a dos locutores diferentes además con distinto sexo, de ésta manera comprobamos que los pulsos son diferentes entre una persona y otra como era de esperar pero a simple vista mediante estas imágenes no sabríamos caracterizar al locutor, por lo que únicamente las utilizaremos de manera visual y en ningún momento trabajaremos con ellas.

3.2.8 EXTRACCIÓN DE CARACTERÍSTICAS A PARTIR DEL MÓDULO DEL ESPECTRO

En primer lugar hay que decir que cuando trabajamos en modo frecuencial hemos quitado el resample debido a que únicamente nos podría afectar negativamente, por lo tanto en este módulo trabajaremos con pulsos de 35 muestras.

Lo único que hemos hecho en este sector es utilizar la función establecida de MATLAB llamada "fft" la cual se encarga de realizar la fast fourier transform la cual permite calcular la transformada de Fourier Discreta (DFT). La DFT sigue la siguiente fórmula:

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j \frac{2\pi kn}{N}} \quad k=0, 1, 2, \dots, N-1 \quad \text{Fórmula 3.3}$$

Sin embargo, el algoritmo tiene algunas limitaciones en la señal y en su espectro, la más importante es que la señal de la cual se tomaran las muestras y que va a ser transformada debe consistir de un número de muestras igual a una potencia de dos como por ejemplo 512, 1024, 2048, 4096. En nuestro caso, escogeremos 2048.

Una vez hemos explicado brevemente el funcionamiento de la Transformada discreta de Fourier (DFT) cogeremos en nuestro programa los diversos pulsos obtenidos previamente y les aplicaremos la "fft" consiguiendo así trabajar en el modo espectral. Cabe destacar que a la hora de representar las gráficas en este modo no se representarán de manera lineal, las gráficas para su visualización se utilizará una escala logarítmica siendo el eje y por lo tanto decibelios y el eje x el número de muestras que hemos escogido que en nuestro caso serán 2048.

Cuando obtenemos la Transformada de Fourier de los diversos pulsos lo que realizamos como hemos hecho anteriormente en el modo temporal será la media de todos los pulsos. A continuación veremos una gráfica con diversos pulsos pasados a modo espectral y en negro y con un grosor mayor la media de todos ellos.

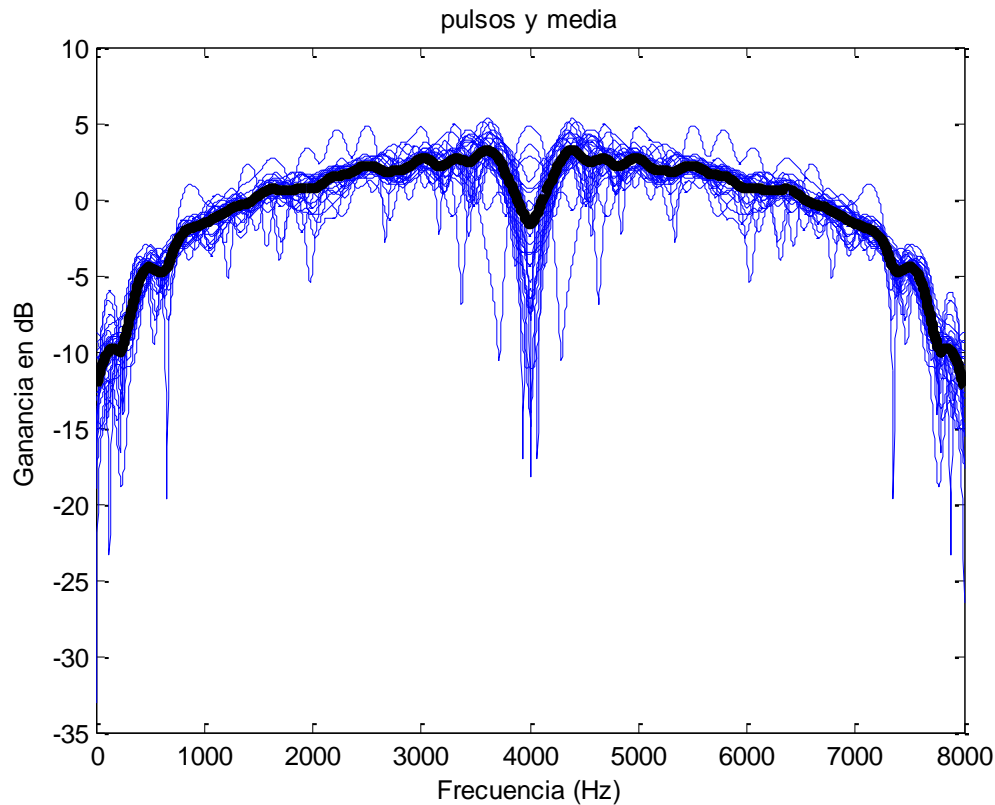


Figura 3.17 Representación gráfica de los diversos pulsos y su media.

Como podemos apreciar en la gráfica, la señal resultante es simétricamente par, es decir si la partimos por la mitad es igual a un lado que al otro por lo que a partir de ahora escogeremos la mitad de muestras a la hora de representar el resultado dándonos el siguiente resultado:

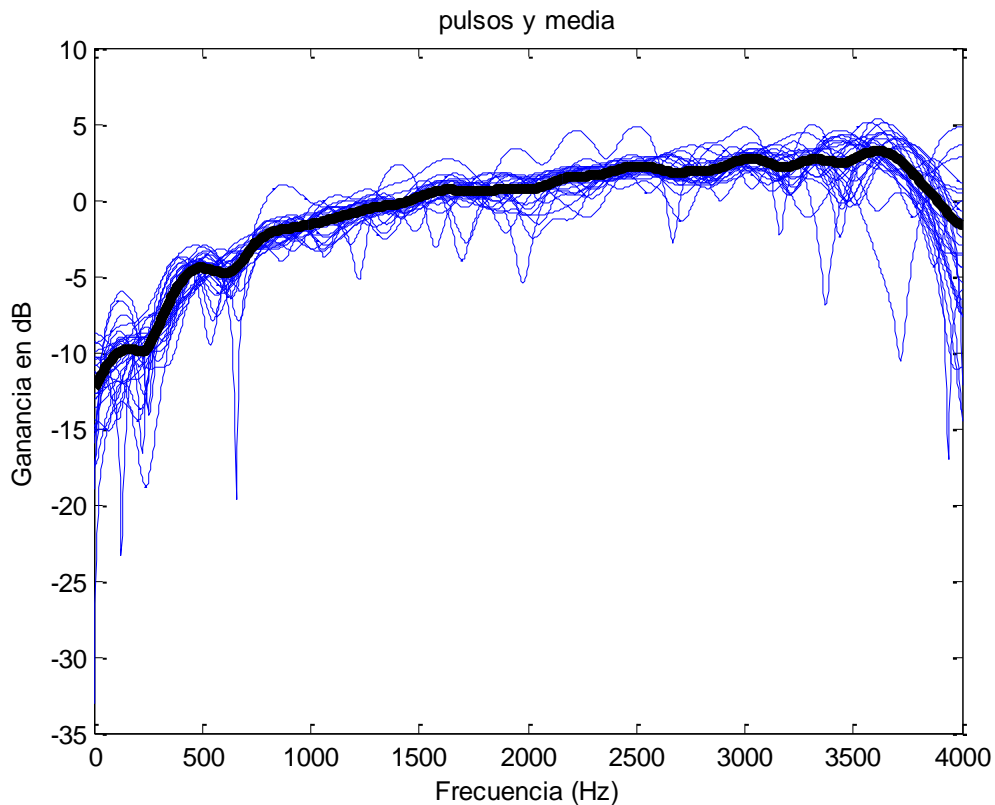


Figura 3.18 Representación grafica hasta la mitad de la señal original.

Una vez hemos conseguido los diversos pulsos y su media, realizamos la primera derivada sobre cada uno de ellos. La primera derivada nos sirve debido a que si se compara luego esa señal con otra derivada de otro locutor aparecerían picos abruptos, y en el caso ideal de compararla con otra señal del mismo locutor aparecería una línea recta, pero esto es muy difícil por las condiciones en las que se ha grabado, pero idealmente sí que funcionaría.

Cuando obtenemos la derivada de cada pulso calculado previamente, realizamos como en los demás subsectores la media de todos los pulsos para obtener el más aproximado al real que sería el siguiente:

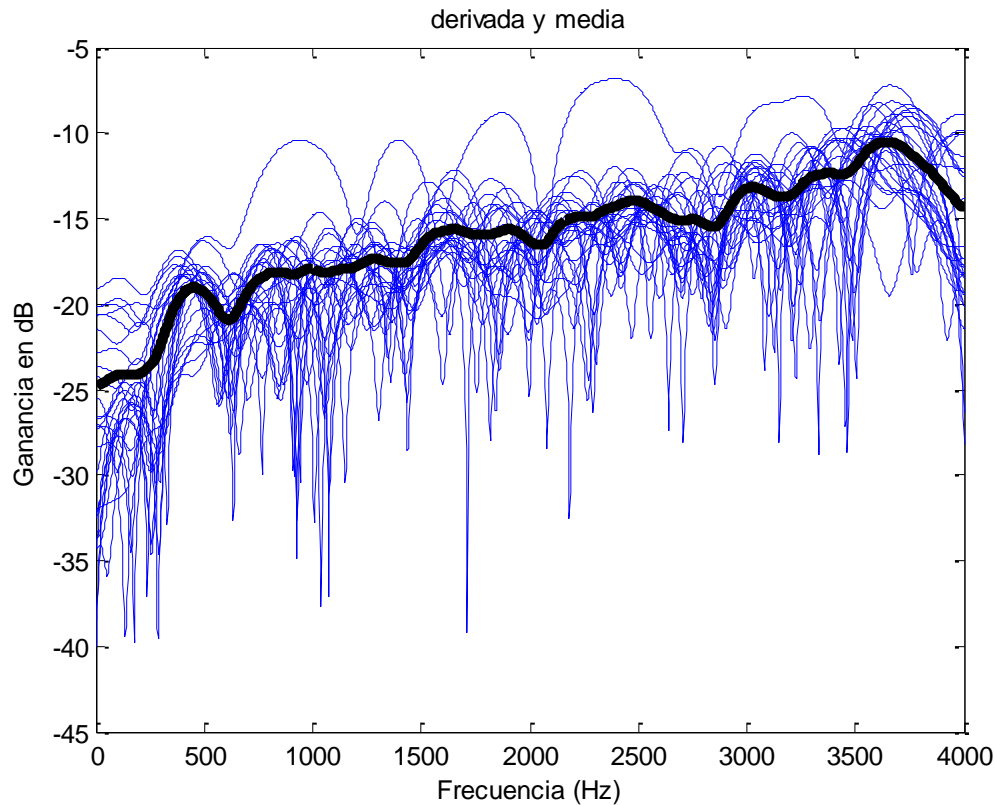


Figura 3.19 Representación de la derivada de la transformada discreta de Fourier de cada pulso con la media de todos ellos en color negro y con un grosor mayor.

3.3 MÓDULO DE OBTENCIÓN.

A continuación, realizaré una breve explicación de la implementación del módulo de análisis. El módulo de obtención tendrá como finalidad conseguir una biblioteca de datos los cuales como veremos serán la obtención de cuatro vectores característicos de cada audio de cada locutor. Habrá un total de diez locutores los cuales serán cinco de sexo masculino y cinco de sexo femenino. El rango de edades será 4 hombres entre 20 y 25 años, otro hombre de 50 a 55 años y por la otra parte cuatro mujeres entre 20 y 25 años y una mujer entre 50 y 55 años. De cada locutor se tendrá un total de cuatro audios seleccionados previamente, por lo que habrá un total de cuarenta audios a analizar.

También hay que decir que cada locutor pertenecerá a un número comprendido entre el 0 y el 9, por lo que primero pondremos a todos los hombres cada uno de ellos con un número comprendido entre 0 y 4 y luego a todas las mujeres con un número comprendido entre 5 y 9, lo que nos facilitará saber en cada momento de quien es el audio y su sexo.

En resumidas cuentas, un total de diez locutores con cuatro audios de cada uno harán un total de cuarenta audios a analizar, siendo A0, A00, A000, A0000 los cuatro audios del primer locutor y A1, A11, A111, A1111 los cuatro del siguiente y así sucesivamente.

Además, de cada audio obtendremos cuatro vectores característicos de su locutor, lo que hará un total de 160 vectores con información de los diversos locutores.

3.3.1 MÓDULO DE GRABACIÓN

Como ya he resumido anteriormente, necesitamos cuatro grabaciones de cada persona. Debido a que las condiciones de grabación no son las más óptimas se realizará un total de veinte grabaciones por persona para la posterior visualización de la forma de onda y la selección del audio con más información, hay que destacar que en algunos casos ha hecho falta realizar hasta un total de cuarenta grabaciones a algún locutor debido a su manera poco óptima de realizar la frase.

Hay ciertas condiciones que deberá cumplir cada locutor a la hora de realizar la frase, no se podrá entonar, es decir, se deberá de hablar linealmente sin mostrar ninguna entonación y además se deberá alto y claro pero sin llegar a saturar. A continuación mostrare dos pistas de audio poco óptimas.

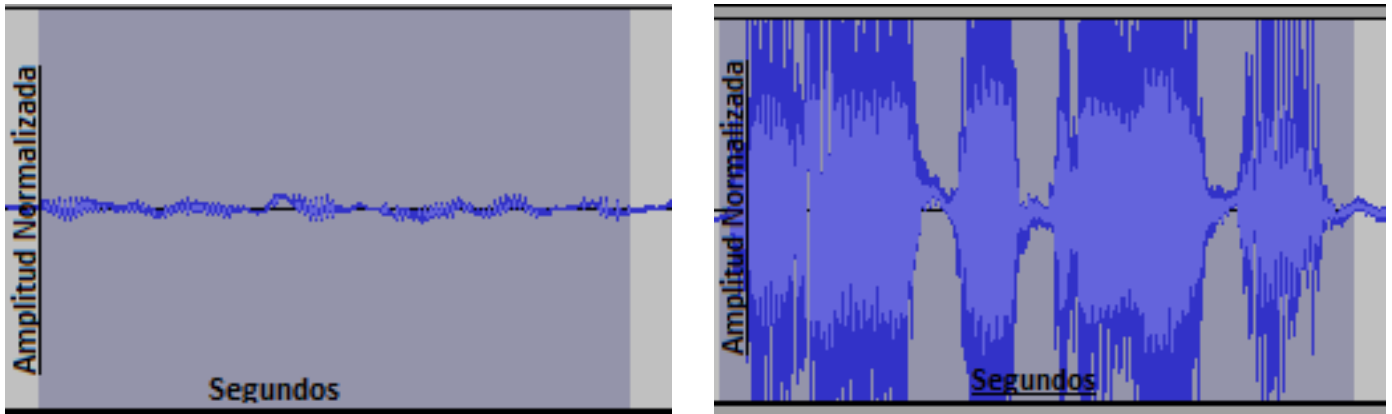


Figura 3.20 (a) Audio con poca información, (b) audio con saturación.

Como podemos imaginar, el primer audio no contendrá apenas información por lo que obtener los pulsos glotales de ahí será poco efectivo. Sin embargo, en el segundo audio contienen demasiada información, tanta que llegan a saturar por lo que las diversas muestras de audio que contendrá se verán cortadas tanto por arriba como por abajo de manera cuadrada, lo que nos dará como resultado un pulso medio glotal que se asemejaría poco a la realidad, por lo tanto, debemos escoger un audio en el cual sea un término medio, ni tanto ni tan poco como podremos observar en la siguiente figura:

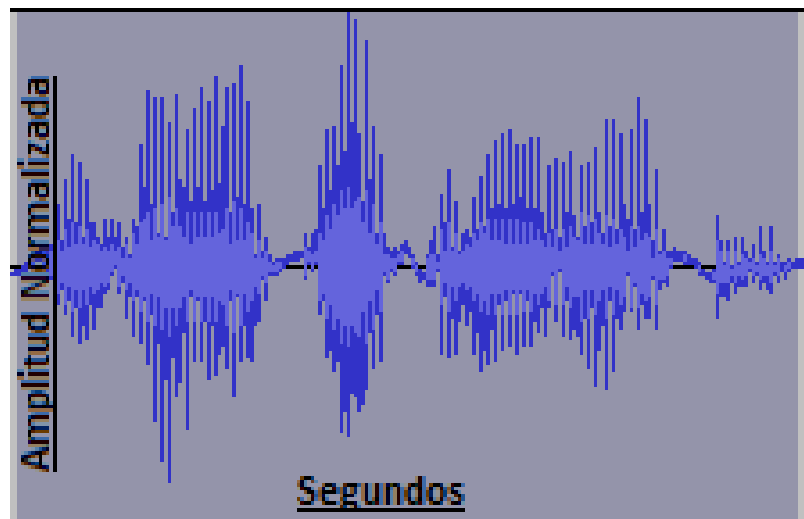


Figura 3.21 Sección de audio correcta.

3.3.2 MÓDULO DE OBTENCIÓN DEL VECTOR CARACTERÍSTICO "A"

El total de señales sobre las que podemos trabajar son tres, la primera de ellas será el pulso medio del cual obtendremos dos vectores, el A y el B, y luego tendremos la media de los pulsos tras haberles aplicado la transformada discreta de Fourier (DFT) que corresponderá con el vector C y por último la derivada de los pulsos obtenidos para el vector C, es decir la media de la derivada de los pulsos en el ámbito frecuencial que pertenecerá al vector D.

El vector A corresponderá a la posición de los máximos y mínimos de toda la señal promedio, es decir la distancia en el eje X entre los distintos puntos rojos que veremos en la siguiente imagen:

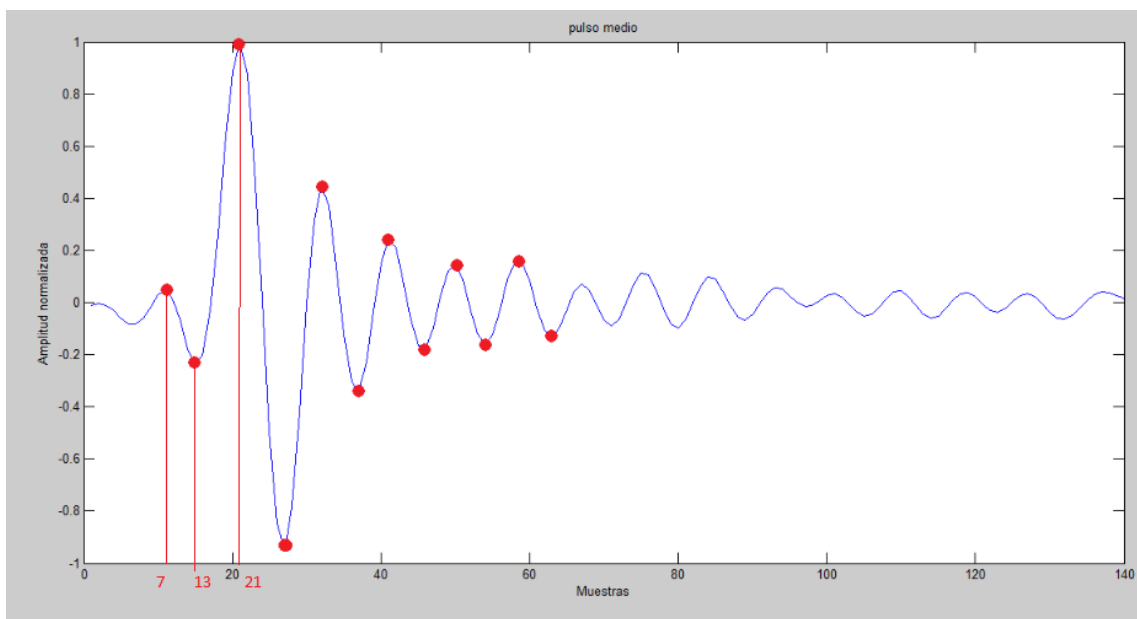
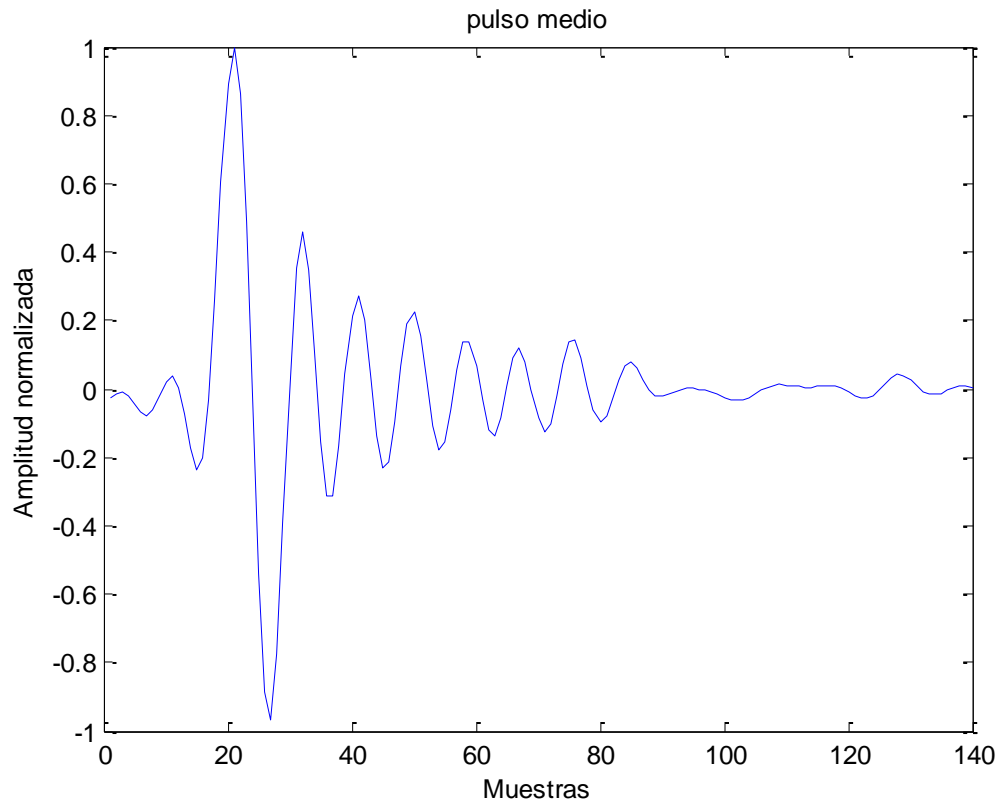


Figura 3.22 Pulso medio y referencia en el eje X.

Es decir, este vector contendrá la posición de la muestra donde se encuentran los distintos máximos y mínimos, por lo que hemos tenido que crear un algoritmo que localice esa posición de la muestra y la almacene en un vector.

El principal problema, es que como podremos comprobar a primera vista, la diferencia en cuanto a la posición no será muy diferente entre distintos locutores, por lo que llegados a éste punto podremos imaginar que dicho vector no será un buen factor para caracterizar al locutor, cosa que más tarde confirmaremos, pero a simple vista ya nos hacemos a la idea de que éste factor no será decisivo.



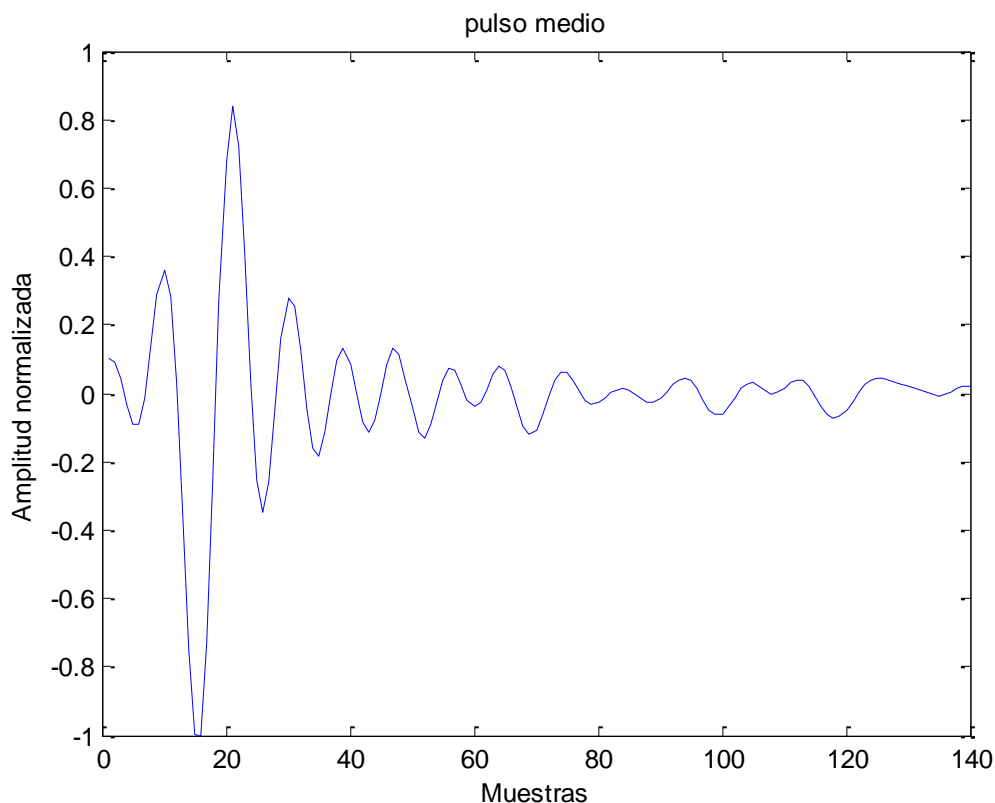


Figura 3.23 Comparativa entre pulsos medios de dos locutores distintos.

A simple vista en cuanto a la posición de los máximos y los mínimos se aprecia que son más o menos similares, pero se puede apreciar una variación en las amplitudes, es decir, en la comparación respecto al eje y entre las dos, lo que hará que lleguemos al siguiente paso, la obtención del vector característico "B".

3.3.3 MÓDULO DE OBTENCIÓN DEL VECTOR CARACTERÍSTICO "B"

Una vez hemos obtenido el vector característico "A", tendremos un vector con las posiciones de las muestras que contienen un máximo o un mínimo, por lo que mediante otro algoritmo lo que haremos será obtener el valor que tiene el eje Y en ese punto.

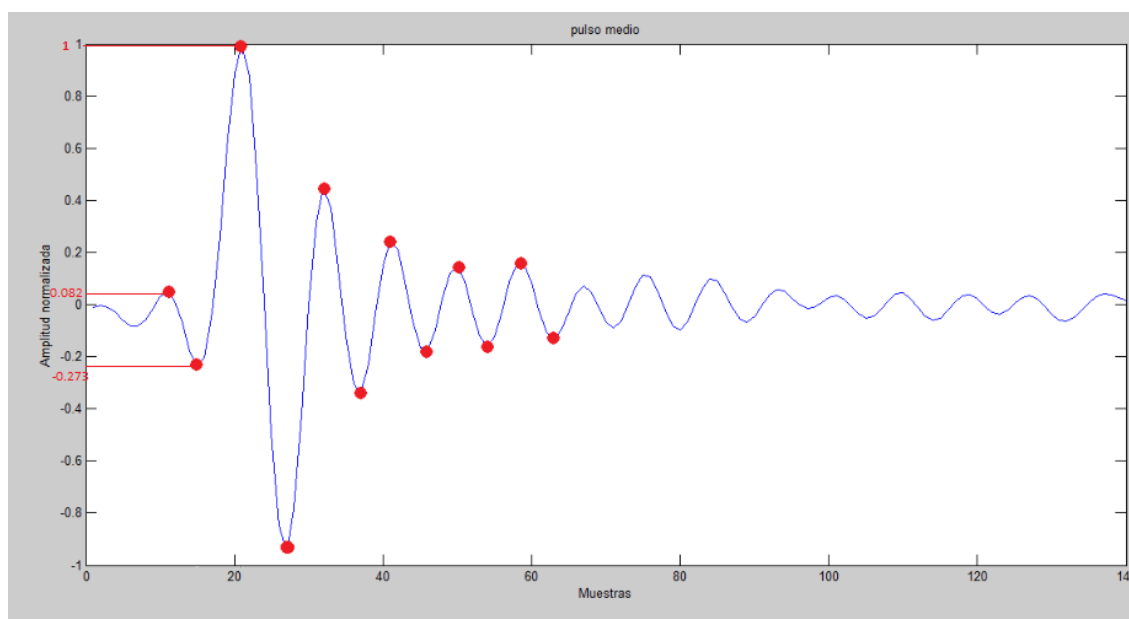


Figura 3.24 Pulso medio y referencia en el eje Y.

En este vector a simple vista será más clasificatorio que el vector "A" debido a que tras ver todos los pulsos medios de cada locutor, se puede apreciar que las amplitudes generalmente nunca van a ser iguales a la de ningún otro, ni aun siendo la misma persona. Además, como podemos ver, aquí afectan los números decimales por lo que podemos afirmar que el grado de exactitud de este vector será muchísimo más característico que el anterior. Esta suposición la confirmaremos más adelante.

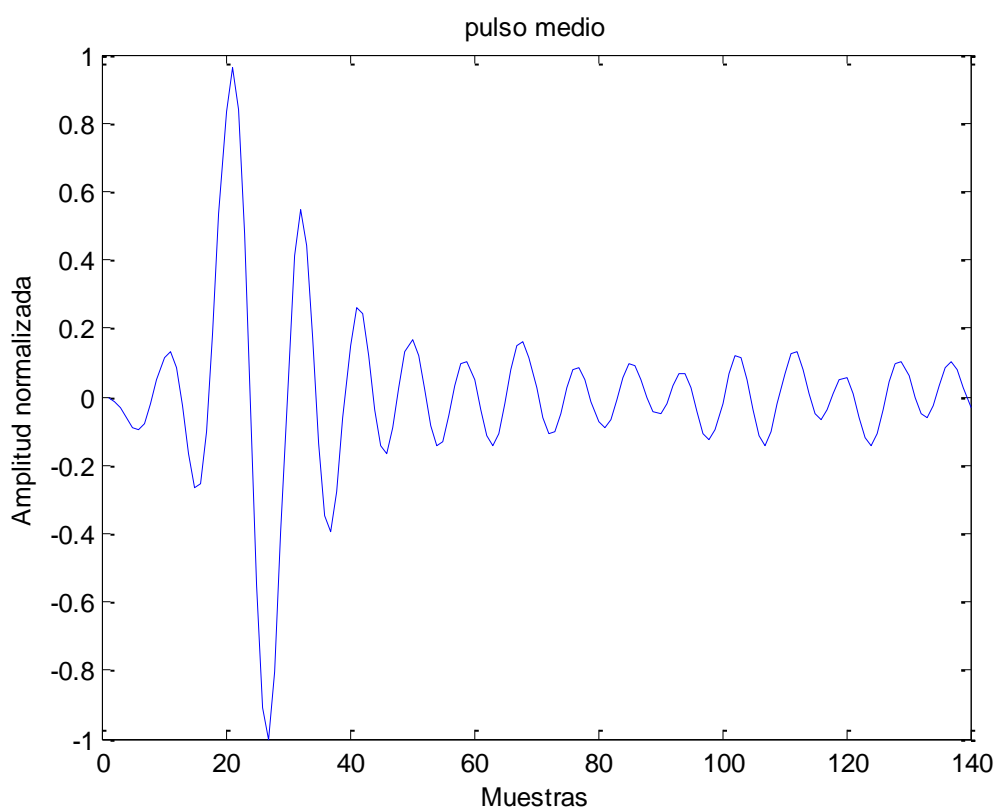
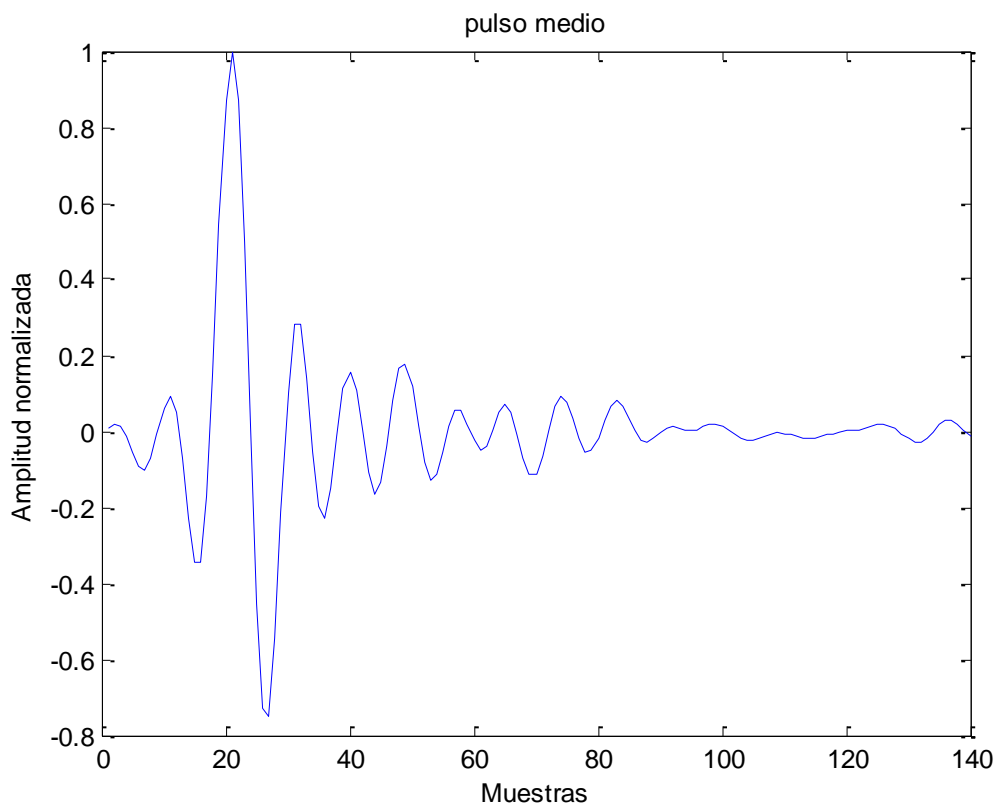


Figura 3.25 Diferencia entre dos pulsos medios de distintos locutores.

Resumiendo, el vector A contendrá la posición de la muestra que contiene un máximo o un mínimo, y el vector B tendrá el valor de la Amplitud que contendrá esa posición.

3.3.4 MÓDULO DE OBTENCIÓN DEL VECTOR CARACTERÍSTICO "C y D"

El vector C corresponderá a la media de los pulsos tras haberles aplicado la "fft" que ya hemos calculado previamente, es decir, contendrá los valores de la curva de color negro y con mayor grosor de la siguiente imagen:

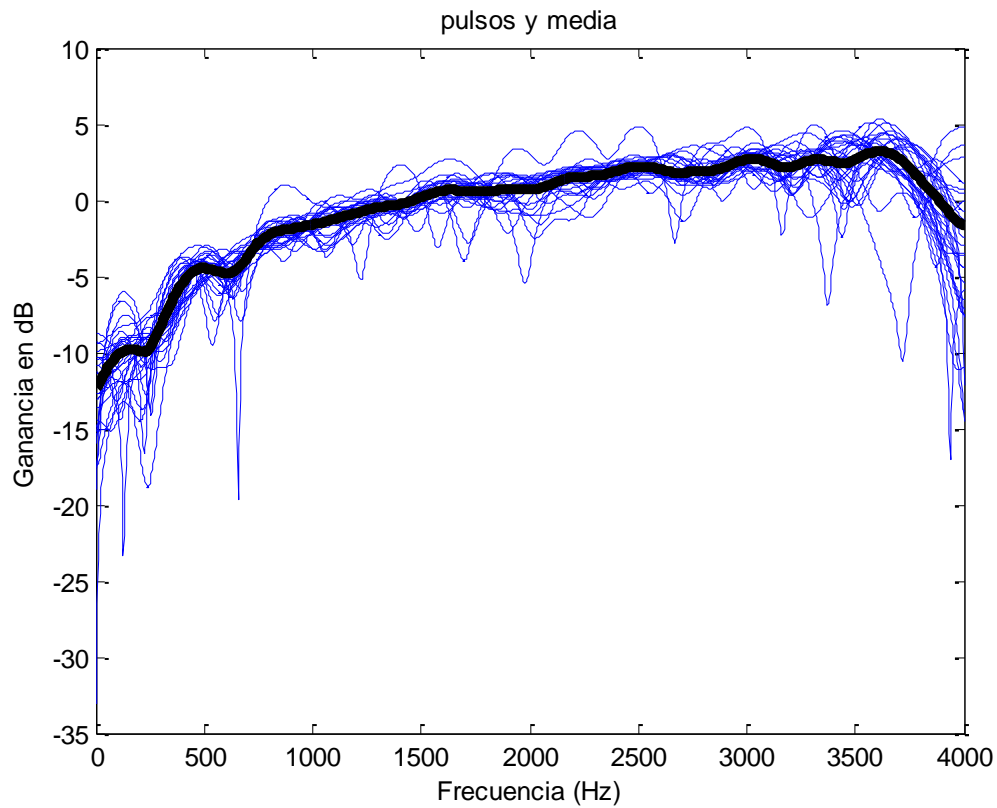


Figura 3.26 Media del módulo del espectro de los pulsos.

El vector D corresponderá a la derivada de la media de los pulsos tras haberles aplicado la "fft", es decir, a la derivada de los pulsos en ámbito frecuencial y corresponderá con la curva de color negro y con mayor grosor de la siguiente imagen:

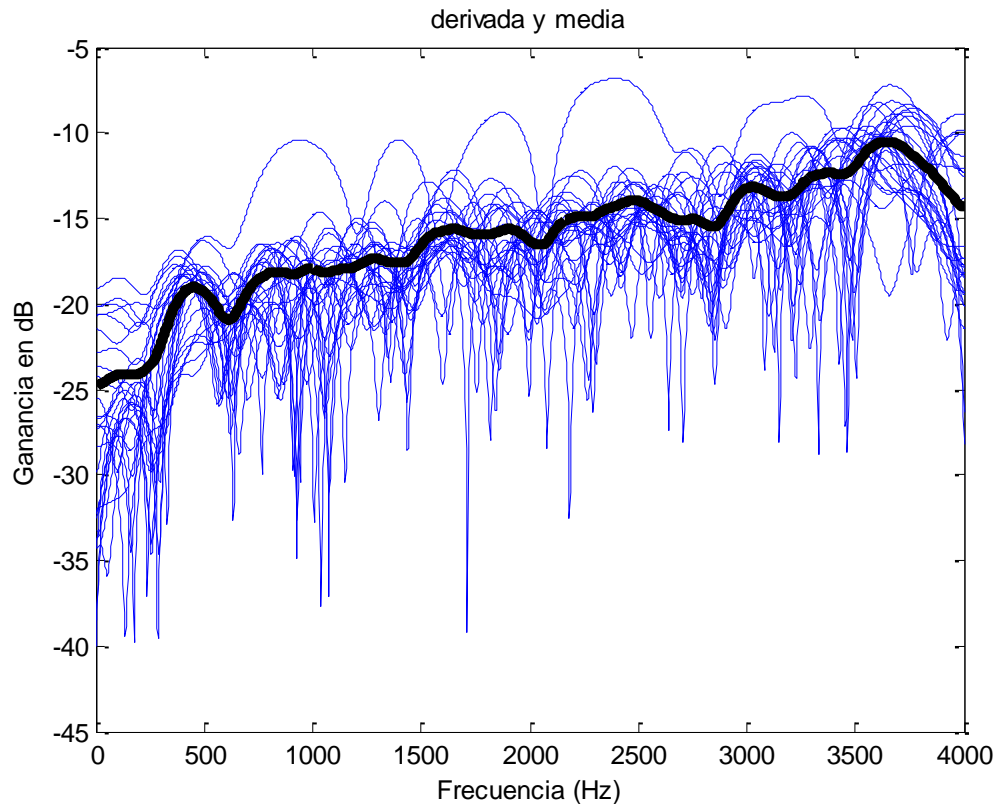


Figura 3.27 1ª Derivada del módulo del espectro.

La primera derivada del módulo del espectro sirve como medida para poder apreciar la velocidad del cambio del espectro.

3.3.5 ALMACENAMIENTO DE LOS VECTORES CARACTERÍSTICOS "A, B, C y D"

Como he mencionado previamente, cada pista de audio generará cuatro vectores característicos, debido a que cada locutor tendrá cuatro audios podemos afirmar que cada uno tendrá 16 vectores característicos. Además, como cada persona se le asignará un número para reconocerla, como por ejemplo el 0, meteremos cada uno de esos vectores en una matriz como podemos comprobar a continuación:


 PABLO.mat	A0	<1x17 double>
	A00	<1x17 double>
	A000	<1x17 double>
	A0000	<1x17 double>
	B0	<1x17 double>
	B00	<1x17 double>
	B000	<1x17 double>
	B0000	<1x17 double>
	C0	<1x2048 double>
	C00	<1x2048 double>
	C000	<1x2048 double>
	C0000	<1x2048 double>
	D0	<1x2047 double>
	D00	<1x2047 double>
	D000	<1x2047 double>
	D0000	<1x2047 double>

Figura 3.28 Locutor 0 con todos sus vectores característicos de los distintos audios.

Realizaremos la misma acción entre los distintos locutores, dándonos un total de diez matrices. Esas diez matrices luego las metemos en una misma, obteniendo una matriz de 160 vectores y además, también hacemos matrices con cada letra, es decir, una matriz que contenga todas las "A", otra con todas las "B", otra con todas las "C", y por último con todas las "D" obteniendo cuatro matrices con 40 vectores. Esto lo hacemos ya que en un futuro nos hará falta para realizar cálculos entre las distintas posibilidades.

3.3.6 CÁLCULO DE LA DISTANCIA ENTRE LOS DISTINTOS VECTORES

A continuación tendremos que definir qué tipo de cálculo realizamos entre los distintos vectores para poder conseguir algún parámetro caracterizador. Hemos optado por la ecuación de la distancia Euclídea, que como bien sabemos es la raíz cuadrada de la diferencia de un punto menos otro elevado al cuadrado.

$$Valor = \sum (X_1 - X_0)^2$$

$$Distancia = \sqrt{Valor}$$

Fórmula 3.4

En nuestro caso, X1 será el primer valor de nuestra señal 1 y el X0 será el primer valor de nuestra señal 2, esto se repetirá sucesivamente hasta que la señal finalice debido a que está dentro de un bucle para calcular toda la señal, de ahí que haya un sumatorio al principio de la ecuación. Una vez hemos obtenido el valor, lo único que tendremos que hacer será la raíz cuadrada de éste, consiguiendo como solución un número que generalmente suele ser menor que 1.

Cuando vemos la ecuación, podemos obtener una cosa en claro y es que suponiendo que tengamos un vector A0 cuyo valor fuese 2 y un vector A1 cuyo valor fuese 3, cumplirá la propiedad conmutativa, es decir, el orden no alterará el resultado debido a que dicho resultado va elevado al cuadrado. Un ejemplo claro es que $(2-3)^2 = (-1)^2 = 1$ y por otra parte $(3-2)^2 = 1^2 = 1$. Esto nos será muy útil debido a que en un futuro cuando tengamos que calcular todos los vectores unos entre otros, cuando haya dos resultados iguales uno se podrá obviar debido a que es técnicamente imposible que dos valores sean iguales mediante la resta de dos vectores distintos sucesivamente.

Una vez definida la relación matemática que vamos a utilizar, la aplicaremos sobre los distintos vectores "A B C D" para así obtener cuatro valores. A continuación se mostrará una gráfica además de la diferencia entre los espectros de pulsos y entre la derivada de los espectros de los pulsos.

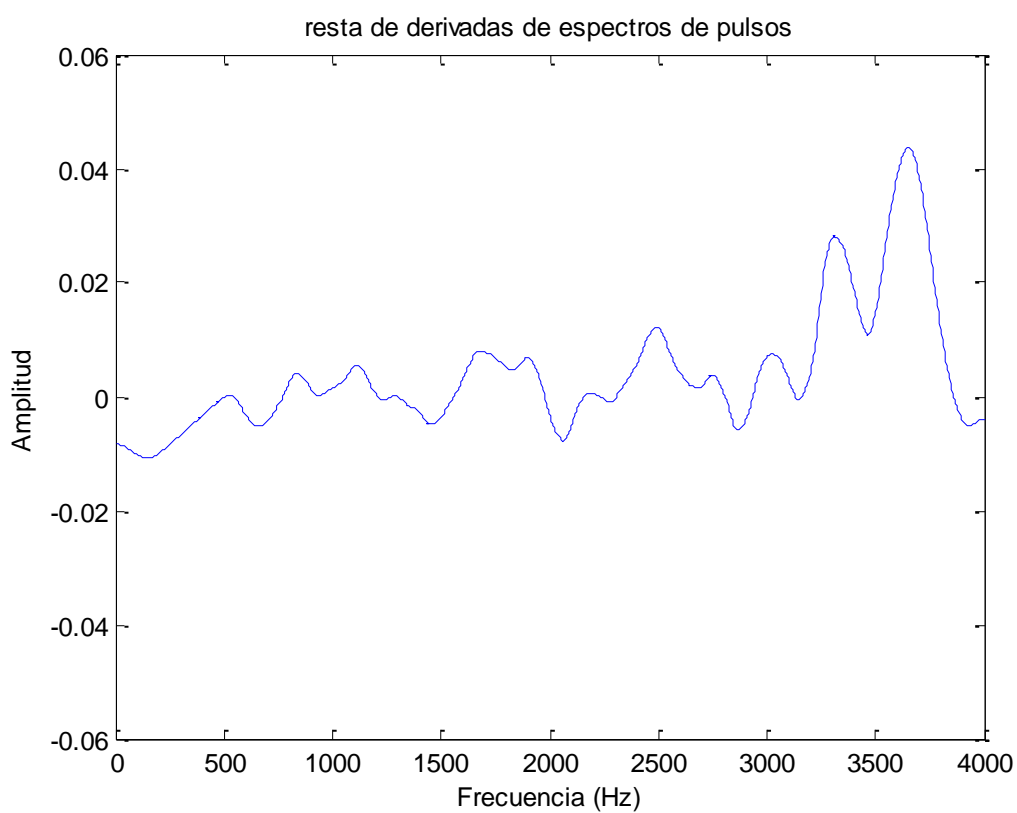


Figura 3.29 Representación de la diferencia entre dos espectro distintos y dos derivadas de espectros distintas.

A simple vista parece que hay una curvatura muy grande, pero si nos damos cuenta con más detalle y vemos el eje Y, podemos apreciar que en el primer caso el valor únicamente está entre 0.5 y -0.5 en el primer caso, y en el segundo caso aun es más pequeño debido a que el valor está más o menos entre -0.02 y 0.02.

	Caso 1	Caso 2	Caso 3
Valor obtenido mediante las distancias	4,2426	7,0711	6,5574
Valor obtenido mediante las alturas	0,1352	0,9133	0,5281
Valor obtenido mediante la resta espectral	10,8538	9,4483	15,9242
Valor obtenido mediante la resta de la 1ª derivada espectral	0,4566	0,4514	0,5247

Como podemos comprobar, conforme vamos cambiando los valores característicos, el valor de cada uno de las distancias calculadas sobre "A B C D" irá variando, por lo que en el siguiente apartado hablaré de todas las posibilidades y como almacenarlas.

3.3.7 MATRIZ RESULTANTE CON TODAS LAS POSIBILIDADES EXISTENTES.

Como hemos comentado anteriormente, dividiremos los vectores característicos "A B C D " cada uno en una matriz que contendrá todos los equivalentes a ese mismo vector, es decir, haremos una matriz con todos los vectores "A" y así sucesivamente. Una vez hemos realizado la matriz, crearemos un algoritmo el cual se encargué de realizar el cálculo matemático anteriormente explicado con todos los demás existentes de la matriz.

Aplicando la lógica sabremos que habrá un total de 1600 posibilidades, obtenido de multiplicar cuarenta audios por los restantes, es decir, cuarenta por cuarenta. Sin embargo, de esas 1600 posibilidades no todas serán utilizables debido a que al restar un vector consigo mismo obtendremos un cero, esto ocurrirá un total de cuarenta veces por lo que ya podemos afirmar que cuarenta audios seguro que no son útiles.

Debido a que como he comentado anteriormente, al elevar el cuadrado el resultado será lo mismo $A-B$ que $B-A$, por lo que si queremos un porcentaje real tendremos que crear un algoritmo el cual se encargue de buscar cada valor con los restantes, y si coincide igualar uno de ellos a cero porque si no estaríamos realizando la estadística con el mismo valor. Además, como he mencionado anteriormente, es técnicamente imposible que un valor se repita si ha sido con otro vector distinto a los aplicados para obtener dicho valor.

Por lo tanto, mediante un cálculo, nos damos cuenta de que en el bucle cada vez habrá un valor que deberemos igualar a cero, debido a que la primera vez que se compare la primera fila con las 39 restantes todos los valores serán válidos, pero como podemos imaginar, cuando pasemos a la segunda fila y se resté con todos los demás también se restará con la primera por lo que ya tendremos un valor repetido. Así continuará aumentando de uno en uno los valores repetidos hasta llegar a la última fila que se obtendrá 39 resultados iguales más la resta de ella misma que será cero como hemos comentado anteriormente.

Entonces lo que tendremos que hacer en primer lugar, será el cálculo de posibilidades válidas que lo calcularemos de la siguiente manera:

En primer lugar realizaremos un sumatorio de los valores comprendidos entre 1 a 39 dándonos de resultado 780. A continuación a ese 780 le sumamos los 40 ceros obtenidos al restar cada uno de ellos consigo mismo, por lo que obtenemos un total de 820. Debido a que 820 serán el número de resultados que no los podremos utilizar para la estadística, los igualaremos a cero para no tenerlos en cuenta. Por lo tanto, tendremos un total de $1600 - 820$ posibilidades, o dicho de otra manera, tendremos un total de 780.

El cálculo del vector "A" no lo hemos calculado debido a que como hemos mencionado anteriormente, no pertenecerá a un factor discriminante. Como he comprobado mediante un contador de distintas posibilidades, los diversos contadores de los distintos vectores obtienen el mismo número de posibilidades que habíamos calculado teóricamente, por lo que podemos afirmar que el algoritmo realizado trabaja correctamente.

3.3.8 SUBDIVISIÓN DE LA MATRIZ RESULTANTE EN DOS MATRICES DISTINTAS.

La última parte de este subsistema consistirá en dividir la matriz resultante con 780 posibilidades anteriormente mencionada en dos matrices. La primera contendrá únicamente los cálculos realizados entre los 4 vectores del mismo locutor, es decir, la primera contendrá únicamente los resultados de A0-A00, A0-A000, A0-A0000, A00-A000, A00-A0000, A000-A0000 y así con el resto de locutores. La segunda, contendrá por lo tanto el resto de cálculos entre diversos interlocutores.

Como sabemos que de cada locutor obtenemos 6 posibilidades al restar sus vectores entre sí, sabemos que de esos 6 por 10 serán un total de 60 cálculos serán con el mismo locutor, por lo que por una parte tendremos una matriz de 720 posibilidades que pertenecerá a la falsa aceptación y otra matriz de 60 posibilidades que pertenecerá al falso rechazo, como explicaré en el siguiente apartado. Por lo tanto, únicamente nos quedará crear un algoritmo que se encargue de encontrar los cálculos del mismo locutor con sus diversos audios y separarlos, sabiendo entonces que la estadística de la falsa aceptación contemplará 720 posibilidades y la del falso rechazo tan solo 60 posibilidades.

3.4 MÓDULO DE INTERPRETACIÓN.

A continuación, realizaré una breve explicación de la implementación del módulo de interpretación. El módulo de análisis tuvo como finalidad en primer lugar obtener el porcentaje de acierto de nuestro algoritmo mediante la colocación de umbrales, pero más tarde se transformó en el cálculo de probabilidad de que se realice una falsa aceptación, o un falso rechazo, como explicaré a continuación.

La Tasa de Falso Rechazo (False Rejection Rate, FRR) y la Tasa de Falsa Aceptación (False Acceptance Rate, FAR). La tasa de falso rechazo es la probabilidad que el sistema de autenticación rechace a el mismo usuario porque no sea capaz de identificarlo correctamente, por lo tanto la tasa de falsa aceptación es la probabilidad de que el sistema autentique correctamente a un usuario ilegítimo.

La Tasa de Falso Rechazo y la Tasa de Falsa Aceptación son inversamente proporcionales debido a que cada uno empieza donde desemboca el otro, es decir, marcarán valores opuestos, por lo tanto si queremos una seguridad muy restrictiva como puede ser para un banco, necesitaremos una tasa de falsa aceptación muy baja por lo que por consiguiente obtendremos a su vez un aumento de la tasa de falso rechazo.

Para evitar un desequilibrio entre las tasas lo que crearemos será un umbral, el cual igualará las dos tasas para que sea lo más igual posible, dicho de otra manera, haremos que las gráficas corten en un punto que será denominado Tasa de Error Igual (EER) que determinará la capacidad de identificación del sistema.

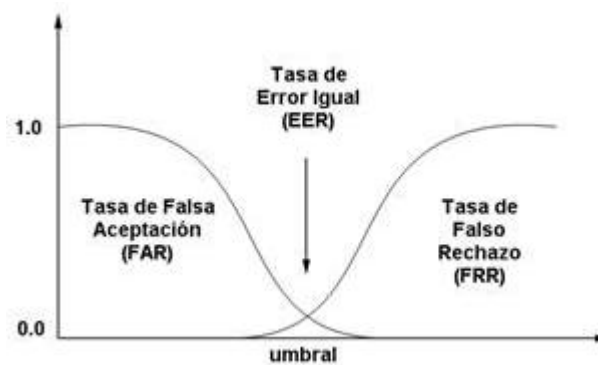


Figura 3.30 Representación gráfica de las distintas Tasas [19]

3.4.1 MÓDULO DE INTERPRETACIÓN PRIMERA VERSIÓN.

En primer lugar, hicimos un algoritmo el cual se encargase de comprobar los distintos valores tanto de la matriz que únicamente tiene los de los cálculos con el mismo locutor tanto como la que contiene los valores de los cálculos entre distintos locutores. Después lo que hicimos fue colocar un umbral tras ver personalmente todos los valores, el cual en teoría si el valor del sumatorio calculado previamente superaba dicho umbral se suponía que el cálculo había sido entre dos locutores distintos, y si el valor era menor a dicho umbral se suponía que se había calculado con el mismo locutor.

Por lo tanto, debido a que tenemos una matriz únicamente con los valores obtenidos tras restar audios del mismo locutor, en éste caso únicamente tendremos que ver si el valor es menor al establecido, en caso de que sea menor lo sumaremos al porcentaje de acierto. Como es de suponer, si el valor en dicha matriz supera el umbral, el programa se estará equivocando, por lo tanto no se podrá sumir al porcentaje de acierto. En la matriz de cálculos entre diversos locutores ocurre exactamente lo contrario, si el valor obtenido tras los cálculos supera el umbral acertaría que son distintos locutores y por lo tanto el programa habría acertado, en cambio, en caso de no superarlo el programa se estaría equivocando.

Como es lógico, el programa no puede acertar el 100% de las veces, en cambio debido a que una matriz únicamente cuenta con 60 audios (la que contiene los cálculos entre mismos locutores), y otra 720 (la que contiene entre distintos), suponiendo que en el primer caso acertase más de 30 veces y en el segundo más de 360 veces ya superaríamos el 50% de acierto.

Sin embargo, el único problema de éste método es que el umbral lo ponemos a ojo, mediante ensayo y error por lo que no podremos encontrar la solución de compromiso adecuada, lo que hará que tengamos que mejorar el algoritmo para encontrar dicha solución.

A continuación mostrare diversos porcentajes de acierto con diversos umbrales establecidos para comprobar la efectividad del programa.

```
Umbral de las alturas 0.3600
El porcentaje de acierto para el mismo locutor mediante la altura 88.3333 %
El porcentaje de acierto para distinto locutor mediante la altura es 84.8611 %
Umbral de la resta frecuencial 0.4650
El porcentaje de acierto para el mismo locutor mediante la derivada es 73.3333 %
El porcentaje de acierto para distinto locutor mediante la derivada es 74.7222 %
Umbral de la resta de las derivadas 10.0000
El porcentaje de acierto para el mismo locutor mediante la derivada es 75.0000 %
El porcentaje de acierto para distinto locutor mediante la derivada es 68.6111 %
```

```
Umbral de las alturas 0.2500
El porcentaje de acierto para el mismo locutor mediante la altura 53.3333 %
El porcentaje de acierto para distinto locutor mediante la altura es 95.0000 %
Umbral de la resta frecuencial 0.5800
El porcentaje de acierto para el mismo locutor mediante la derivada es 80.0000 %
El porcentaje de acierto para distinto locutor mediante la derivada es 49.4444 %
Umbral de la resta de las derivadas 12.3400
El porcentaje de acierto para el mismo locutor mediante la derivada es 88.3333 %
El porcentaje de acierto para distinto locutor mediante la derivada es 42.7778 %
```

Figura 3.31 Porcentajes de acierto con distintos umbrales.

El umbral establecido para los resultados de los vectores "C", es decir, para los de la diferencia espectral, dista de los otros dos debido a que ahí se han obtenido unos valores mucho más grandes.

Además, como podremos imaginar, si ponemos un umbral muy alto el programa siempre pensará que está hablando el mismo locutor debido a que nunca llegará a superar dicho umbral, por lo tanto, si ponemos un umbral muy alto el programa supondrá que nunca es el mismo locutor debido a que siempre superará dicho valor establecido. A continuación colocaré otras imágenes para demostrar nuestra hipótesis.

```
Umbral de las alturas 10.0000
El porcentaje de acierto para el mismo locutor mediante la altura 100.0000 %
El porcentaje de acierto para distinto locutor mediante la altura es 0.0000 %
Umbral de la resta frecuencial 10.0000
El porcentaje de acierto para el mismo locutor mediante la derivada es 100.0000 %
El porcentaje de acierto para distinto locutor mediante la derivada es 0.0000 %
Umbral de la resta de las derivadas 100.0000
El porcentaje de acierto para el mismo locutor mediante la derivada es 100.0000 %
El porcentaje de acierto para distinto locutor mediante la derivada es 0.0000 %

Umbral de las alturas 0.0000
El porcentaje de acierto para el mismo locutor mediante la altura 0.0000 %
El porcentaje de acierto para distinto locutor mediante la altura es 100.0000 %
Umbral de la resta frecuencial 0.0000
El porcentaje de acierto para el mismo locutor mediante la derivada es 0.0000 %
El porcentaje de acierto para distinto locutor mediante la derivada es 100.0000 %
Umbral de la resta de las derivadas 0.0000
El porcentaje de acierto para el mismo locutor mediante la derivada es 0.0000 %
El porcentaje de acierto para distinto locutor mediante la derivada es 100.0000 %
```

Figura 3.32 Porcentajes para demostrar la hipótesis previamente explicada.

3.4.2 MÓDULO DE INTERPRETACIÓN MEDIANTE LA OBTENCIÓN DEL UMBRAL.

Finalmente, decidimos mejorar el algoritmo anterior, debido a que se basaba en el establecimiento de umbrales al azar, por lo que lo retocamos para obtener la tasa de falsa aceptación y falso rechazo explicado previamente. Además, en éste caso cuanto menor sea el porcentaje mayor será la efectividad, debido a que aquí estamos analizando la capacidad de errar de programa, es decir, la probabilidad de que el resultado no sea el correcto, por lo tanto, cuanto menor sea mayor será el porcentaje de acierto.

A continuación mostraré las gráficas obtenidas tras haber cambiado el algoritmo que nos ayudarán a decidir el umbral.

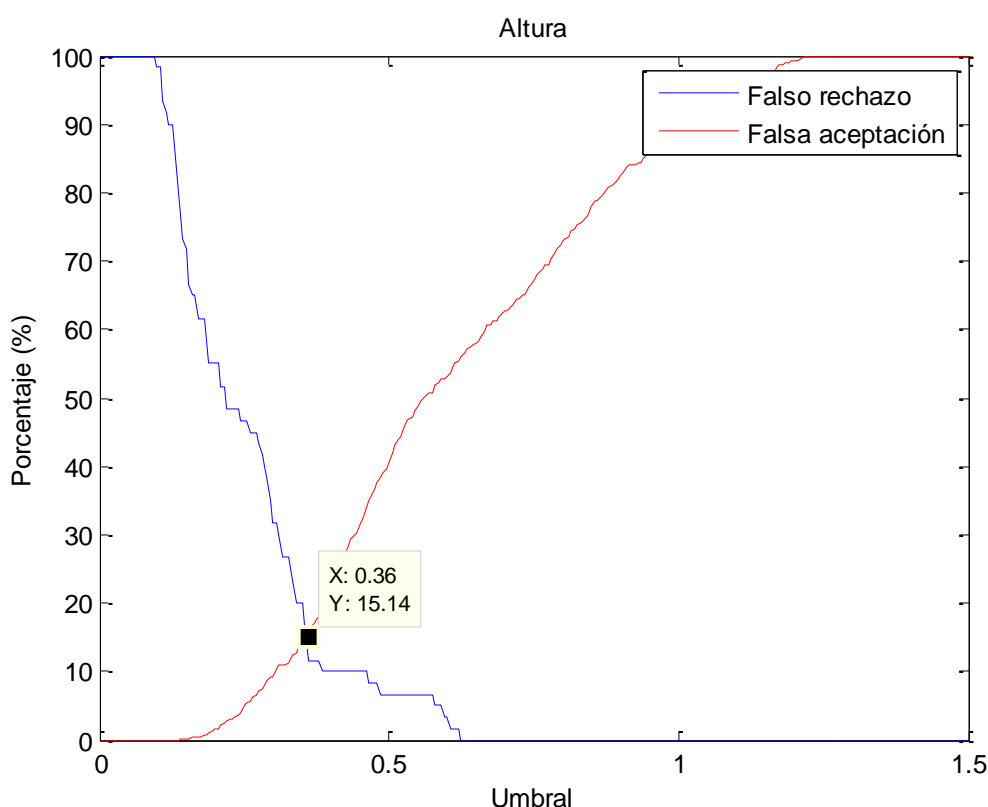


Figura 3.33 Representación de umbrales obtenidos mediante los cálculos con el vector "B" (el vector que contenía la altura de los máximos y mínimos)

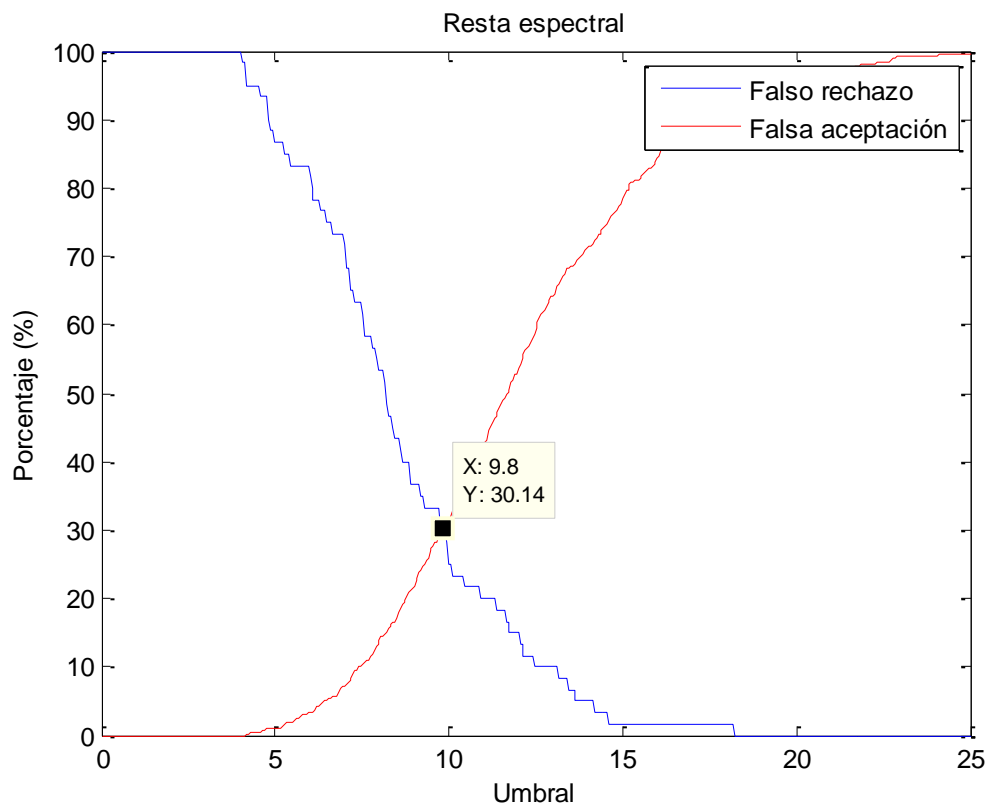


Figura 3.34 Representación de umbrales obtenidos mediante los cálculos con el vector "C" (el vector que contenía la resta en frecuencial)

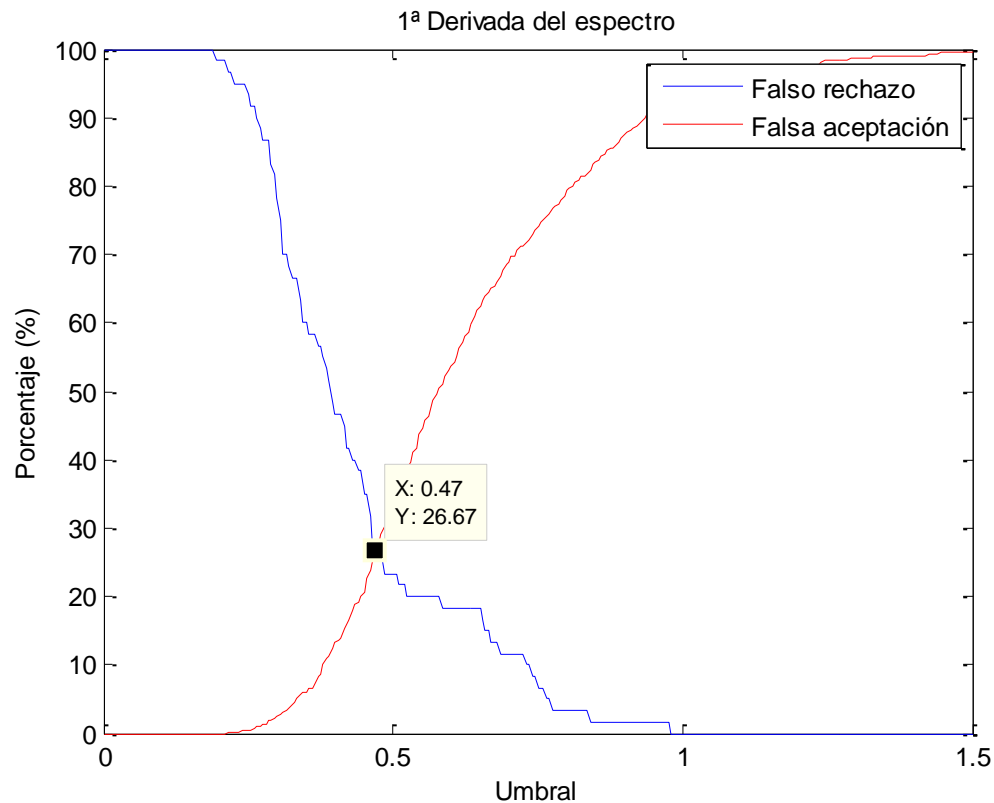


Figura 3.35 Representación de umbrales obtenidos mediante los cálculos con el vector "D" (el vector que contenía la derivada de el pulso en frecuencial)

Estos han sido los resultados obtenidos, como podemos observar en el primer caso, el cálculo del vector "B" el umbral debería de ser 0.36 y obtendría una tasa de error del 15.14%. En el segundo caso, el cálculo del vector "C" el umbral debería de ser 9.8 obteniendo una tasa de error del 30.14%, y por último, el cálculo del vector "D" el umbral debería de ser 0.47 obteniendo una tasa de error del 26.67.

A continuación mostraré una imagen para comprobar si nuestro algoritmo funciona correctamente y calcula la probabilidad de error obtenida con dicho umbral:

```
Umbral "B"  0.3600
Falsa aceptación y Falso rechazo mediante la altura es  15.1389 %
Umbral "C"  9.8000
Falsa aceptación y Falso rechazo mediante la resta frecuencial es  30.1389 %
Umbral "D"  0.4700
Falsa aceptación y Falso rechazo mediante la derivada es  26.6667 %
```

Figura 3.36 Comprobación de los resultados obtenidos.

Además, como podemos comprobar, la hipótesis de que los valores eran opuestos se cumple debido a que como podemos comprobar gráficamente, si colocamos un umbral muy pequeño el porcentaje de "Falso Rechazo" será el 100% de las veces, dicho de otra manera, nunca reconocerá que el locutor es el mismo y por lo tanto el porcentaje de "Falsa aceptación" será 0% debido a que nunca aceptará a nadie. Ocurriendo lo mismo pero inversamente cuando se coloque un umbral muy grande en el cual el algoritmo aceptará siempre a cualquier locutor y nunca rechazará.

Por lo que para finalizar, podemos afirmar que la característica más óptima para caracterizar el pulso glotal de cada locutor, será el valor de la altura donde se sitúan los máximos y mínimos.

CONCLUSIONES

En este apartado recopilaremos tanto los objetivos previstos al empezar el proyecto como el desarrollo y resolución de éstos con un final bastante satisfactorio.

El objetivo principal de éste proyecto fue la búsqueda de parámetros que nos sirviesen para caracterizar a un locutor a partir del pulso glotal, es decir, la extracción del pulso glotal y su debida caracterización posterior para analizar la fiabilidad que tendrá cada uno de las características extraídas. Además, conseguimos simular mediante una pista de audio la obtención de un sistema fonador, con su señal residuo incluida.

Analizaremos por tanto detalladamente cada una de las características o vectores característicos que hemos decidido analizar, de menor a mayor grado de fiabilidad.

A modo de resumen, se ha obtenido a lo largo del proyecto cuatro vectores característicos "A" , "B", "C", "D". Cada uno de ellos contiene información obtenida, teniendo por lo tanto el vector "A" la posición donde se encuentran mínimos y máximos en el pulso glotal medio, el vector "B" el valor numérico de la amplitud que tienen los máximos o mínimos encontrados en el vector "A", el vector "C" contiene la resta en el ámbito espectral y finalmente el vector "D" contendrá los valores al aplicar la primera derivada a los valores recogidos en el vector "C" en el ámbito espectral.

En primer lugar, en cuanto al vector característico "A" que es el vector correspondiente a la localización de los máximos y los mínimos fue fácil de descartar porque como pudimos comprobar a lo largo del proyecto, los pulsos glotales eran muy similares en cuanto a la posición de los máximos y los mínimos, era suposible debido a que tras la visualización de diversos pulsos de diversas materias como pueden ser de la frecuencia cardiaca, todos se asemejaban en cuanto a la forma por lo que resulto ser este vector inutilizable.

Seguidamente, hablando del vector característico "C" que es el vector correspondiente a la transformación espectral de los diversos pulsos así como de su media, se ha obtenido un porcentaje de error del 30.14%, o lo que es lo mismo un porcentaje de acierto del 69.86% . Es un resultado bastante bueno, debido a que en el remoto caso de que los demás vectores no hubiesen obtenido mejor resultado, acertar un 70% de las veces sería un resultado bastante optimista.

A continuación, hablando del vector característico "D" que es el vector correspondiente a la derivada de la transformación espectral previamente mencionada, sinceramente desde que decidí analizar dicha característica imaginé que sería el factor más discriminante del locutor, pero no fue así, debido a que tiene un porcentaje de error del 26.67% o lo que es lo mismo, un porcentaje de acierto del 73.33%. Tras haber comprobado los resultados del vector "C" y "D" nos damos cuenta de que también es necesario siempre trabajar en el ámbito espectral además del temporal, debido a que si no fuese así habríamos perdido dos características glotales con unos muy buenos resultados.

Finalmente, hablamos del vector característico "A" que es el vector correspondiente al valor de la amplitud tanto en los máximos como en los mínimos, éste vector ha sido el más óptimo con tan solo un porcentaje de error del 15.14%, o lo que es lo mismo, un porcentaje de acierto del 84.86%. Además, conforme íbamos realizando el proyecto se podía apreciar que las alturas no solían coincidir entre distintos locutores, por lo tanto al hacer la resta matemática se obtendría unos valores muy distantes a comparación de la resta de valores entre el mismo locutor.

En cuanto a conclusiones extraídas del proyecto podemos obtener las siguientes:

En primer lugar, la calidad del pulso glotal extraído se verá principalmente condicionada por las condiciones de grabación para dicha pista, es decir, si realizamos la grabación en una cámara anecoica o en una habitación en la cual no influya el ruido, con material de trabajo óptimo se obtendrá un pulso glotal más real. En nuestro caso no ha sido así, se ha realizado en unas condiciones normales por lo que si hubiesen sido unas condiciones óptimas el resultado hubiese sido más exacto.

Debido a que el medio influye en gran medida a la obtención del pulso glotal y por lo tanto a su posterior procesado y caracterización, considero que la implantación de un sistema de seguridad mediante la voz únicamente debería de servir como complemento, es decir, que primero se implantase otro sistema de seguridad completamente diferente a éste y para la fase de confirmación/ comprobación se implantase éste, debido a que en la caracterización por el pulso glotal se ve altamente influenciada por el medio de grabación como ya he mencionado.

Para finalizar, recalcar que a la hora de la realización del algoritmo siempre se puede mejorar para obtener unos porcentajes de error menores, debido a que podemos utilizar otras características como la energía para la comprobación de ciertas características u otras muchas, pero por suerte se ha realizado el proyecto siempre por el camino planteado desde el principio, no ha habido que tomar caminos distintos y finalmente hemos conseguido un resultado bastante óptimo para las condiciones y con los materiales que se ha realizado.

BIBLIOGRAFÍA

- [1] Arthur C. Guiton, "Tratado de fisiología médica", Ed. Interamericana, Quinta edición, 1976.
- [2] H Rouviere "Anatomía humana, descriptiva, topográfica y funcional", Ed. Baily-Bailiere, 1980.
- [3] Fonética articuladora, [última consulta: 18 jun. 2017] Disponible en <http://paginaspersonales.deusto.es/airibar/Fonetica/Apuntes/02.html>
- [4] Vocal sound production, [última consulta: 18 jun. 2017] Disponible en <http://hyperphysics.phy-astr.gsu.edu/hbase/Music/voice.html>
- [5] Miguel Ros San Juan, "Preprocesado para la mejora de la segmentación del área glotal, [última consulta: 18 jun. 2017] Disponible en: http://oa.upm.es/43903/1/TFG_MIGUEL_ROS_SAN_JUAN.pdf
- [6] FinK B.R, "The human larynx: a funtional study", New York, Ed. Raven press, 1974
- [7] Musculatura, [última consulta: 18 jun. 2017] Disponible en: <http://mayullantalla.blogspot.com.es/2011/01/musculatura.html>
- [8] Aparato fonador, [última consulta: 18 jun. 2017] Disponible en: <http://slideplayer.es/slide/5552202/>
- [9] José Luis Navarro Mesa, "Procesador Acústico: El bloque de extracción de características", [última consulta: 18 jun. 2017], Disponible en: <https://www2.ulpgc.es/hege/almacen/download/25/25296/apuntesextraccioncaracterisitcas.pdf>
- [10] Juan Carlos Gómez, "Modelos de producción de voz", [última consulta: 18 jun. 2017] Disponible en: http://www.fceia.unr.edu.ar/prodivoz/Modelo_Produccion_Voz_bw.pdf

- [11] Emilia Gómez Gutiérrez, "Digitalización del Sonido", [última consulta: 18 jun, 2017] Disponible en: <http://www.dtic.upf.edu/~egomez/teaching/sintesi/SPS1/Tema2-Digitalizacion.pdf>
- [12] Cuantificación digital, [última consulta: 18 jun. 2017], Disponible en: [://es.wikipedia.org/wiki/Cuantificaci%C3%B3n_digital](http://es.wikipedia.org/wiki/Cuantificaci%C3%B3n_digital)
- [14] Cuantificación digital Comunicación, [última consulta: 18 jun. 2017], Disponible en: http://www.w3ii.com/es/digital_communication/digital_communication_quantization.html
- [14] Sadaoki Furui, "Digital Speech, Processing, Synthesis, and Recognition", Ed. Board, Segunda edición, 2001
- [15] Francisco Javier Hernando Pericas, "Técnicas de procesado y representación de la señal de voz para el reconocimiento del habla en ambientes ruidosos Capítulo 3", [última consulta: 18 jun. 2017], Disponible en: http://www.tdx.cat/bitstream/handle/10803/6911/04_hernandoPericas_capitol_3.pdf?sequence=4
- [16] Extraído de los apuntes técnicas de reconocimiento y síntesis del habla de la Universidad de Alicante.
- [17] José Julio Hernández Fernández, "Control inverso adaptativo Capítulo 2", [última consulta: 18 jun. 2017], Disponible en: <http://bibing.us.es/proyectos/abreproy/11284/fichero/Volumen+1%252FCap%C3%ADtulo+2.pdf>
- [18] Extraído de la práctica 5 de Tratamiento Digital de la Señal (TDS) impartida en la Universidad de Alicante
- [19] Sergio D. Werner, "Aplicación de Nuevas Tecnologías al Sistema Electoral-Biometría y Voto electrónico", [última consulta: 18 jun. 2017], Disponible en: <http://www.monografias.com/trabajos82/biometria-y-voto-electronico/biometria-y-voto-electronico2.shtml>