IDENTIFICACION DEL HABLANTE EMPLEANDO CEPSTRO Y CURVA MELODICA

CARLOS ARTURO GARCIA GOMEZ

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRIA EN INGENIERIA ELECTRÓNICA
SANTAFÉ DE BOGOTÁ, D.C.
2015

IDENTIFICACION DEL HABLANTE EMPLEANDO CEPSTRO Y CURVA MELODICA

CARLOS ARTURO GARCIA GOMEZ

Trabajo de Investigación de Maestría Ingeniería Electrónica

Director
PEDRO R. VIZCAYA GUARÍN Ph.D. Profesor Titular

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERÍA
MAESTRIA EN INGENIERIA ELECTRÓNICA
SANTAFÉ DE BOGOTÁ, D.C.
2015

Tabla de Contenido

Listado de Figuras

Listado de Tablas:

1. INTRODUCCIÓN

El procesamiento del habla es un campo diverso con muchas aplicaciones. La figura 1 muestra algunas de estas áreas y cómo el reconocimiento del hablante se divide en dos grandes ramas como son la verificación e identificación.

Cuando hablamos de la Verificación automática del habla (ASV- Automatic Speaker Verification), nos referimos al uso de un sistema para verificar la identidad declarada de una persona por su voz, es decir, la tarea consiste en decidir si un hablante es quien dice ser. En identificación automática del hablante (ASI- Automatic Speaker Identification), el cual es nuestro objeto de estudio, no hay una notificación de identidad a priori, y el sistema decide a partir de un grupo de personas, cual persona es la que se encuentra hablando [1].

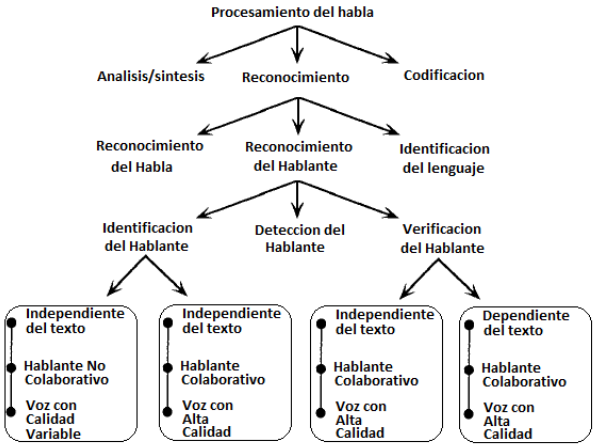


Figura 1. Procesamiento del Habla. [1]

Es importante notar que en la identificación automática del hablante (ASI), existirán casos en los que la persona que pretende ser identificada, no pertenezca al grupo de usuarios, o de igual forma no quiera ser identificada, de ahí que la herramienta desarrollada para la identificación automática del hablante, deberá considerar todos estos casos.

En los sistemas actuales desarrollados para la identificación del hablante [2], se encuentra que un tipo de parámetros óptimo para un reconocimiento de voz son los coeficientes cepstrales.
Los coeficientes cepstrales son un conjunto de parámetros apropiado para representar las características espectrales de la señal de voz, aunque solo contienen información instantánea de la señal, por tanto es necesaria la incorporación de características espectrales dinámicas, las cuales fueron sugeridas por Furui [3], incorporando los coeficientes que representan las primeras derivadas de los coeficientes cepstrales; sin embargo, estos sistemas desarrollados no arrojan el 100% de certeza en su identificación, además, debido a la tasa de error de estos

sistemas (5% aproximadamente, dependiendo de los factores que afectan la señal de voz como ruido, codificación de la voz, o errores en la toma de las muestras), es obligación valerse de otros calificadores para la correcta identificación de la persona, es decir, no está demostrado aún que la voz sea un parámetro único en la identificación del hablante.

En literatura, se define la curva melódica como el resultado acústico de los cambios en la frecuencia de vibración de las cuerdas vocales, es decir, la variación de la frecuencia del fundamental (f_0) en el tiempo.
Para el desarrollo de este trabajo, partiremos de la premisa o hipótesis que todas las personas no tienen la misma curva melódica, es decir, tanto la variación como la frecuencia fundamental f_0 es diferente en cada hablante, por tanto nos valdremos de la extracción de esta característica para adicionarla al sistema de identificación del hablante y de esta manera mejorar la tasa de error.

2. MARCO TEÓRICO.

El sistema propuesto se enfoca inicialmente en la adquisición, tratamiento, entrenamiento y posterior validación de los parámetros de audio extraídos con ayuda de los coeficientes cepstrales y mezclas de gaussianas.
Para la etapa de adquisición y tratamiento es necesario conocer los aspectos requeridos en la toma de información, ya que esto afecta de manera significativa nuestro sistema implementado.
Para la etapa de Entrenamiento, se exploran las técnicas de extracción de coeficientes cepstrales, así como del modelo de mezclas de Gaussianas.
La validación de nuestro sistema se realizara con ayuda de las relaciones de verosimilitud o LRs, que serán las encargadas de comprobar o rechazar la hipótesis planteada.

2.1 Producción, procesamiento y Tratamiento del habla.

Cuando hablamos del procesamiento de la voz, nos referimos a un amplio conjunto de métodos y técnicas que permiten, clasificar y procesar la señal de audio para diversos tipos de aplicaciones. Existen por ejemplo, aplicaciones enfocadas a la síntesis de la voz, es decir, a la generación artificial o computacional del habla, también encontramos aplicaciones enfocadas al análisis de voz, que busca entre otras cosas, analizar disfunciones vocales presentes en algunos hablantes, otra aplicación que se puede encontrar es la del reconocimiento del habla, que tiene como objetivo identificar a un hablante dentro de un grupo de personas [1].

Cuando nos referimos al sonido audible por el oído humano, lo definimos como una sensación percibida en el órgano del oído producida por la vibración que se propaga en un medio elástico en forma de ondas. El sonido audible para los seres humanos está formado por oscilaciones originadas por la presión del aire que el oído convierte en ondas mecánicas y finalmente, en impulsos nerviosos para que el cerebro pueda percibirlas y procesarlas. El sistema de producción del habla no forma parte estricta del sistema sensorial humano, pero su importancia es indudable. Para determinar las operaciones de un sistema automático de reconocimiento de voz y hablante, es fundamental conocer y determinar los mecanismos que han producido un mensaje hablado, para, a continuación, procesarlos, compararlos y analizarlos automáticamente. [4].

Los sonidos del habla implican una corriente de aire vibrante a la cual le sucede algo mientras avanza. Una cosa que puede sucederle es que el paso de la corriente de aire sea obstruido (bloqueado), en cierto grado, y en alguna parte o partes del mecanismo vocal. Esto altera también la dimensión y tamaño de las cavidades de resonancia. El modo de articulación de un sonido describe el grado de obstrucción a la corriente de aire y el tipo de cierre que produce esa obstrucción [5].

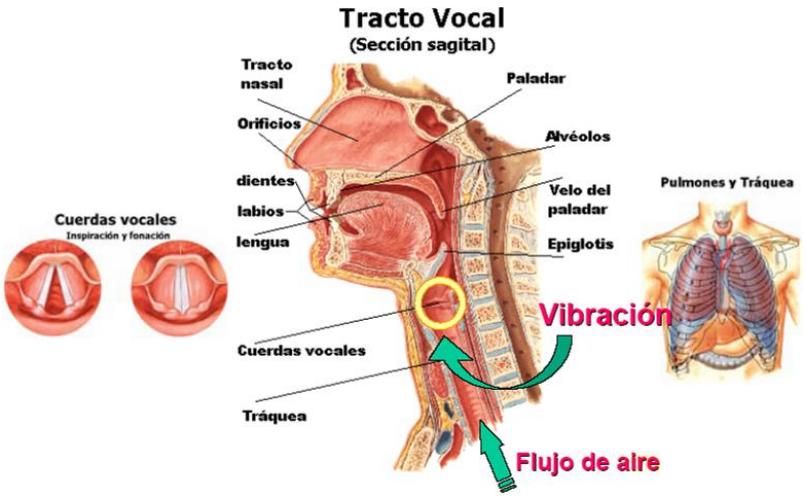


Figura 2. Sección del tracto vocal. Tomado de: http://vox-technologies.com/blog/aparato-fonador

Como la onda acústica pasa a través del tracto vocal, su contenido de frecuencia (espectro) se ve alterada por las resonancias del tracto vocal. Las resonancias del tracto vocal se denominan formantes. Por lo tanto, la forma del tracto vocal puede estimarse a partir de la forma espectral (por ejemplo, la ubicación formante y la inclinación espectral) de la señal de voz. Sistemas de verificación de voz suelen utilizar características derivadas sólo del tracto vocal. Como se ve en la Figura 2, el mecanismo vocal humano está impulsado por una fuente de excitación, que también contiene información dependiente del orador. La excitación es generada por el flujo de aire desde los pulmones, llevado por la tráquea a través de los pliegues vocales (o los cartílagos aritenoides) [1].

2.2 Características del sonido

La excitación puede ser caracterizado como la fonación (sonidos sonoros), susurrante, fricación (sonidos fricativos), compresión (Sonidos oclusivos), o una combinación de estos [5]:

2.2.1 Sonidos Sonoros: Si la corriente de aire no es bloqueada en gran medida, sino que es simplemente formada o dirigida por cierta parte de la lengua y quizás por los labios, se produce una vibración de las cuerdas vocales, A esta frecuencia de vibración de las cuerdas se le llama frecuencia fundamental (f0). La frecuencia fundamental depende de la presión ejercida al pasar el aire por las cuerdas y de la tensión de estas. En un hombre la frecuencia fundamental se encuentra en el rango 50-250 Hz, mientras en la mujer el rango es más amplio, encontrándose entre 100 y 500 Hz. Los ejemplos de estos sonidos incluyen a las vocales.

2.2.2 Sonidos Fricativos: Si la corriente de aire no es bloqueada totalmente, se produce una turbulencia audible. La corriente de aire para un sonido caracterizado por una turbulencia audible, se dice que está parcialmente obstruida, y el sonido se llama

fricativo. (El término fricativo se asemeja a la palabra fricción, y podemos pensar en una fricativa como sonido con fricción audible.) Los ejemplos de fricativas incluyen **s**, **f**, y **z**. Todas las fricativas son además consonantes.

2.2.3 Sonidos Oclusivos: Si para determinado sonido la corriente de aire se bloquea totalmente, se dice que hay bloqueo total. Tales sonidos, que son consonantes, se los conoce como oclusivos (ocasionalmente como plosivos), por ejemplo, **p**, **t**, y **k**.

En general, la clasificación de los sonidos se basa en su característica de sonoridad. En los sonidos sonoros, se observa un patrón regular tanto en su estructura temporal como en su frecuencia (ver figura 3), patrón del que carecen los sonidos sordos (ver figura 4).

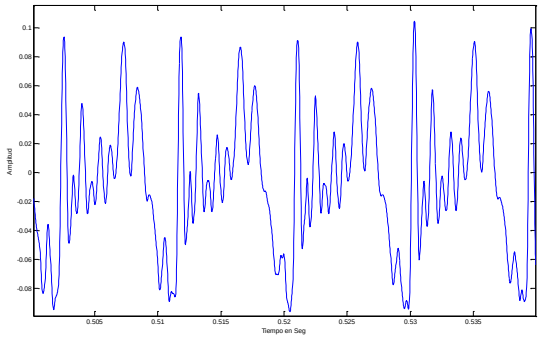


Figura 3. Forma de onda de un sonido sonoro (vocal a).

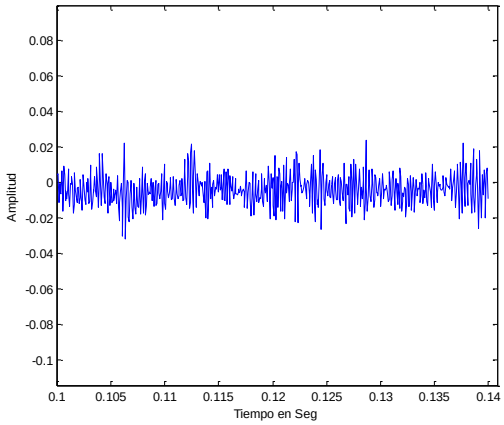


Figura 4. Forma de onda de un sonido sordo (consonante s).

2.3 Percepción del habla

La capacidad de comprender el lenguaje oral se deriva del funcionamiento de un conjunto muy complejo de procesos perceptivos, cognitivos y lingüísticos que permiten al oyente recuperar el significado de un enunciado cuando lo oye. La figura 5, muestra la estructura del sistema periférico auditivo.

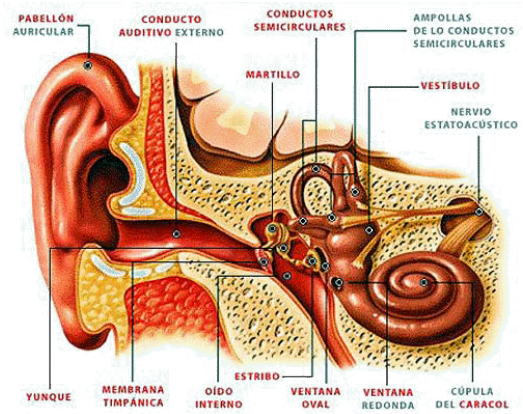


Figura 5. Estructura del sistema periférico auditivo. Tomado de: Anatomía y fisiología del oído, http://rabfis15.uco.es/lvct

Cuando el sonido llega al oído, las ondas sonoras son recogidas por el pabellón auricular (o aurícula). El pabellón auricular, por su forma helicoidal, funciona como una especie de "embudo" que ayuda a dirigir el sonido hacia el interior del oído. Sin la existencia del pabellón auricular, los frentes de onda llegarían de forma perpendicular, el proceso de audición resultaría ineficaz y gran parte del sonido se perdería.
Una vez que ha sido recogido el sonido, las vibraciones provocadas por la variación de presión del aire cruzan el conducto auditivo externo y llegan a la membrana del tímpano. El conducto auditivo actúa como una etapa de potencia natural que amplifica automáticamente los sonidos más bajos que proceden del exterior.

En la membrana timpatica, se produce la transducción, es decir, la transformación la energía acústica en energía mecánica. En este sentido, la membrana actúa como un transductor mecánico acústico.

La presión de las ondas sonoras hace que la membrana timpatica vibre, y a su vez, transmiten el movimiento del tímpano al oído interno, llegando hasta la ventana oval. Esta presión ejercida sobre la ventana oval, penetra en el interior de la cúpula de caracol, la cual se comunica directamente con el nervio auditivo o estatoacustico, conduciendo una representación del sonido al cerebro. La cúpula de caracol es un tubo en forma de espiral (de 3.5 cm aproximadamente).

La cúpula de caracol (también llamada cóclea), puede ser aproximada como un banco de filtros. Los filtros correspondientes al extremo más próximo a la ventana oval y al tímpano responden

a las altas frecuencias, ya que la membrana es rígida y ligera. Por el contrario, en el extremo más distante, la membrana es pesada y suave, por lo que los filtros correspondientes responden a las bajas frecuencias. Por ello los investigadores emprenden trabajos psicoacusticos experimentales para obtener las escalas de frecuencias que modelen la respuesta natural del sistema de percepción humano (ver figura 6).

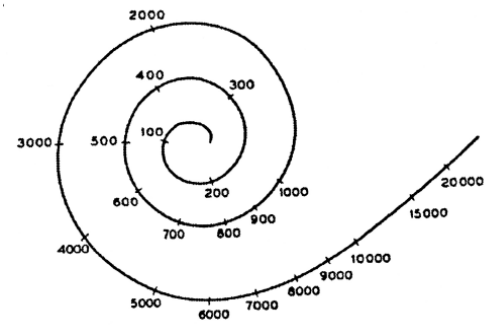


Figura 6. Distribución (experimental), de la frecuencia de resonancia del oído humano alrededor de la cúpula de caracol. [16]

Una característica fundamental del sistema auditivo humano es su capacidad de resolución de frecuencia e intensidad. En este aspecto, es fundamental el concepto de banda crítica. Una forma de entender el funcionamiento del sistema auditivo es suponer que contienen una serie o banco de filtros pasa-banda solapados conocidos como filtros auditivos [6].
Estos filtros se producen a lo largo de la membrana timpatica y tienen como función permitir discriminar entre distintos sonidos. Este banco de filtros no sigue una configuración lineal, y el ancho de banda y morfología de cada filtro depende de su frecuencia central.

El ancho de banda de cada filtro auditivo se denomina banda crítica [6]. Las bandas críticas, son rangos de frecuencia dentro de los cuales un sonido bloquea o enmascara la percepción de otro sonido. Es importante destacar que el concepto de banda crítica es una construcción teórica y no algo físicamente comprobado.

2.3.1 La escala BARK

La escala Bark (en honor al físico alemán Georg Heinrich Barkhausen) es la unidad de frecuencia perceptual; específicamente, un bark mide la tasa de banda crítica, es decir, una banda crítica tiene un ancho de un bark. La escala bark (Tabla 1) relaciona la frecuencia absoluta (en Hz) con las frecuencias medidas perceptualmente (el caso de las bandas críticas). Usando el bark, un sonido en el dominio de la frecuencia puede ser convertido a sonido en el dominio psicoacústico. De esta manera, un tono puro (representado por una componente en el dominio de la frecuencia) puede ser representado como una curva de enmascaramiento psicoacústico. Eberhard Zwicker modeló el oído con 24 bandas críticas arbitrarias para frecuencias por debajo de 15 KHz, con una banda adicional que ocupa la región entre 15 y 20 KHz. El bark (ancho de una banda crítica) puede calcularse con las siguientes fórmulas [7]:

$$1bark(Hz) \cong \frac{f}{100} \qquad \text{para } f < 500\,Hz$$

(2.1)
$$1bark(Hz) \cong 9 + 4\log\left(\frac{f}{1000}\right) \qquad \text{para } f > 500\,Hz$$

donde f = *frecuencia*

Banda crítica (Bark)	Frec. central (Hertz)	Ancho de banda (Hertz)	Frec. min. (Hertz)	Frec. máx. (Hertz)
1	50	-	-	100
2	150	100	100	200
3	250	100	200	300
4	350	100	300	400
5	450	110	400	510
6	570	120	510	630
7	700	140	630	770
8	840	150	770	920
9	1000	160	920	1080
10	1170	190	1080	1270
11	1370	210	1270	1480
12	1600	240	1480	1720
13	1850	280	1720	2000
14	2150	320	2000	2320
15	2500	380	2320	2700
16	2900	450	2700	3150
17	3400	550	3150	3700
18	4000	700	3700	4400
19	4800	900	4400	5300
20	5800	1100	5300	6400
21	7000	1300	6400	7700
22	8500	1800	7700	9500
23	10500	2500	9500	12000
24	13500	3500	12000	15500
25	18775	6550	15500	22050

Tabla 1. Escala de Bark para estimación de las bandas criticas del sistema auditivo. [16]

La figura 7, muestra un banco de filtros típico teniendo en cuenta las bandas críticas en barks.

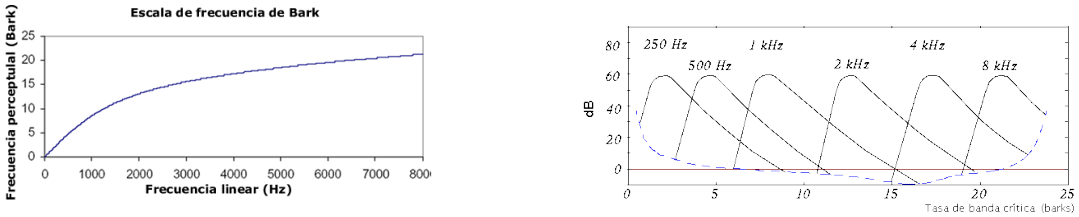


Figura 7. Umbral de enmascaramiento para un banco de filtros en escala bark. [16]

2.3.2 La escala MEL

Al igual que sucede con la frecuencia perceptual, resulta útil contar con una escala perceptual de frecuencias que pueda representar de manera más fidedigna nuestra percepción de la frecuencia de un sonido. Esta escala se conoce como escala Mel (ver figura 8), y fue propuesta por Stevens, Volkman y Newmann en 1937.

El nombre Mel deriva de melodía, como una forma de explicitar que se trata de una escala basada en comparaciones entre frecuencias. La escala Mel se construye equiparando un tono de 1000 Hz a 40 dBs, por encima del umbral de audición del oyente, con un tono de 1000 Mels.

Sobre los 500 Hz, los intervalos de frecuencia espaciados exponencialmente son percibidos como si estuvieran espaciados linealmente. En consecuencia, sobre este punto, cuatro octavas en la escala lineal de frecuencias medida en Hz se comprimen a alrededor de dos octavas en la escala Mel. La escala Mel ha sido ampliamente utilizada en modernos sistemas de reconocimiento de habla y puede ser aproximada en función de la frecuencia lineal como:

(2.2) $$\hat{f} = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

Donde f representa la frecuencia en escala lineal y $\hat{f}$ la frecuencia en escala Mel.

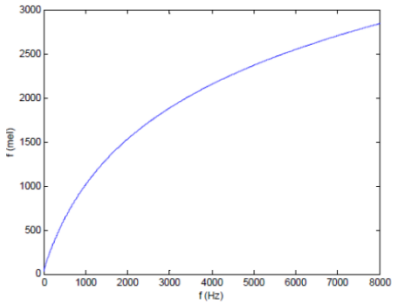


Figura 8. Representación de la escala Mel. [4]

2.4 La Curva Melódica

Desde el punto de vista articulatorio, la curva melódica se produce gracias a las variaciones en la frecuencia de abertura y cierre de los pliegues vocales que se producen en la etapa de la producción del habla correspondiente a la fonación. La representación de las curvas melódicas es uno de los principales problemas que se plantean en el estudio de la entonación. La primera dificultad para el análisis prosódico del habla aparece ya en el propio proceso de detección de la frecuencia fundamental (f_0). La figura 9, muestra un ejemplo del cálculo de la curva melódica en un audio de 2 segundos.

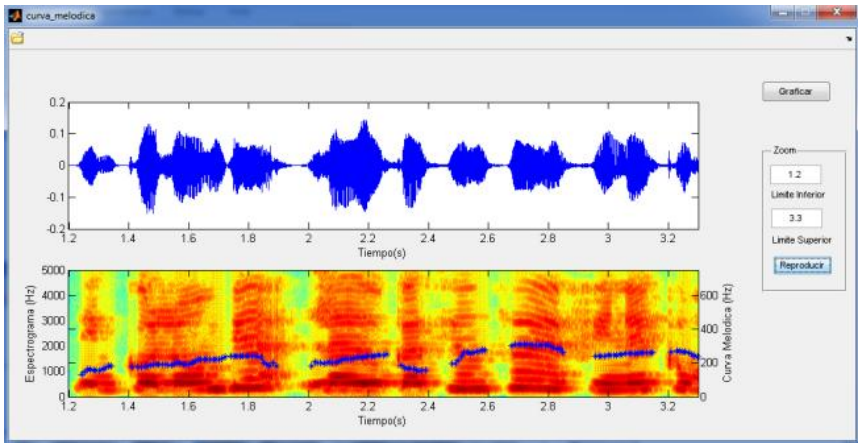


Figura 9. Calculo de la curva melódica en un segmento de audio de 2 seg.

Para obtener una representación acústica de la evolución temporal de la F_0 a lo largo de un enunciado se emplean normalmente algoritmos de detección de la F_0 que actúan directamente sobre la señal acústica para detectar la periodicidad de la misma y la longitud del periodo.

La frecuencia fundamental a nivel de percepción de la voz, cuyo valor medio habitual está entre 70 y 400 Hz, es más bajo para hombres (sonidos graves) que para mujeres (sonidos agudos). Se trata de un parámetro característico de cada locutor y está vinculado con el tono o frecuencia fundamental de vibración de las cuerdas vocales. Para sonidos sonoros, los cuales tienen cierta periodicidad ya que el sonido fluye desde las cuerdas vocales sin excesivo impedimento, la curva melódica puede tomar valores dentro del rango especificado, sin embargo, para sonidos sordos (sonidos oclusivos), resulta imposible de calcular debido a su naturaleza ruidosa.

2.4.1 Técnicas para el cálculo de la frecuencia fundamental

2.4.1.1 Función de Autocorrelación.

Un proceso estocástico se define como estacionario si todas sus características de aleatoriedad (distribución y aleatoriedad esencialmente) presentan invariancia temporal. Adicionalmente, se asume la ergodicidad del proceso, cuando los valores promedio de tiempo y ensamble son idénticos en el intervalo de análisis T_a . Luego, puede definirse la función de autocorrelacion (FAC) $R_x(\tau)$, para un proceso que posea las particularidades anteriormente descritas, a partir de [8]:

$$R_x(\tau) = \lim_{Ta\to\infty} \frac{1}{Ta} \int_{Ta} x(t)\, x^*(t+\tau)\, dt = \langle x(t), x(t+\tau) \rangle \tag{2.3}$$

Por cuanto la FAC tiene un papel básico en la descripción de los procesos ergodicos, es importante tener en cuenta las siguientes propiedades [9]:

- ✓ ***Paridad.***

$$R_x(\tau) = R_x(-\tau) = R_x(\tau, t+\tau) = E\left[x(t), x(t+\tau)\right] \forall\, \tau \in T_a \tag{2.4}$$

- ✓ ***Máximo Valor***

$$\left|R_x(\tau)\right| \le R_x(0) \tag{2.5}$$

Asumiendo $E\left[x(t)\right] = 0$,

$$R_x(0) = E\left[x(t), x(t)\right] = E^2\left[x(t)\right] = \sigma_x^2 \tag{2.6}$$

- ✓ ***Aditividad.*** Sean dos proceso ergódicos independientes $x_1(t)$ y $x_2(t)$, se asume que:

(2.7) $$R_{xT}(\tau) = R_{x1}(\tau) + R_{x2}(\tau)$$

Donde $R_{xT}(\tau)$ corresponde a la función de autocorrelacion resultante de la interacción entre dichos procesos.

- ✓ ***Convergencia.*** Si el proceso aleatorio $x(t)$ es no periódico, esto es, $x(t) \neq x(t+T_0), T_0 \in T_a$, además de considerar que es centralizado, esto es, $E[x(t)] = 0$, entonces se cumple que:

(2.8) $$\lim_{|\tau| \to \infty} R_x(\tau) = 0$$

- ✓ ***Periodicidad***. Si el proceso aleatorio x(t)es periódico para un valor T_0, esto es, $x(t) = x(t+T_0), T_0 \in T_a$, entonces la respectiva FAC, también será periódica:

(2.9) $$R_x(\tau) = R_x(\tau + kT_0), \forall \tau = T_0, \ \ k \in Z$$

Por tanto la existencia de valores iguales al máximo global, $R_x(kT_0) = R_x(0), k \in Z$ muestra que el proceso es periódico con frecuencia fundamental $F_0 = 1/T_0$. Sin embargo, aunque exista un solo máximo global en $\tau = 0$, pueden a veces presentarse otros máximos locales, cuyo mayor valor $R_x(\tau_{\max})$ sea suficientemente grande, como para asumir que el proceso tiene una parte periódica en $T_0 = \tau_{\max}$ [8].

2.4.1.2 Función Cepstral

Las técnicas de análisis espectral que operan en el dominio de la potencia espectral logarítmica tienen la limitación de que, debido a que los espectros de los filtros en bandas adyacentes están bastante correlacionados, originan coeficientes espectrales también bastante correlacionados. Es deseable eliminar esa correlación manteniendo solo la información que sea útil para el reconocimiento. Para ello, se utiliza un filtro de decorrelacion homomorfica o cepstrum que, mediante la transformada inversa de Fourier del logaritmo del espectro de potencias, lleva los coeficientes cepstrales al dominio de la cuefrencia convirtiéndolos en coeficientes cepstrales.

El cepstro (o cepstrum llamado así por invertir las cuatro primeras letras de la palabra spectrum), puede ser visto como una información del ritmo de cambio de las diferentes bandas de un espectro se calcula determinando el logaritmo natural de la magnitud de la transformada de Fourier de la señal x, y de obtener la transformada inversa de Fourier de la secuencia resultante:

(2.10) $$\hat{x} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log[X(e^{jw})] e^{jw} dw$$

La figura 10, muestra la transformación obtenida al aplicar el cepstro a una señal de voz, el pico más notorio, si se excluyera la energía de la señal (punto cero), corresponde a la frecuencia fundamental.

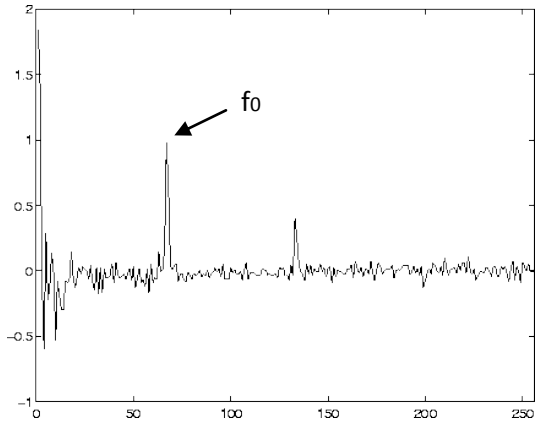


Figura 10. Calculo de la frecuencia fundamental a partir del análisis del cepstro

La variable independiente de un gráfico de cepstro se llama "cuefrencia". La cuefrencia es una medida del tiempo, pero no en el sentido del dominio temporal. Por ejemplo, si la velocidad de muestreo de una señal de audio es de 44.100 Hz y hay un gran pico en el cepstrum en la "cuefrencia" de 100 muestras, este pico indica la presencia de un pitch (frecuencia fundamental percibida) a 44100 / 100 = 441 Hz. Este pico que aparece en el cepstrum indica entonces el período en el que están los armónicos del espectro. Nótese que una sola onda sinusoidal no se debería usar como test para el cepstrum ya que esta no contiene armónicos. En su lugar, debería usar una señal de test que sí contenga armónicos, como la suma de dos senos en la cual el segundo tenga una frecuencia múltiplo de la primera (armónico).

2.5 Síntesis para la identificación del habla.

2.5.1 Preprocesado de la Señal

La señal de voz producida por un hablante, se considera una señal cuasi-estacionaria, es decir, que sus características permanecen constantes durante unos instantes, cambiante en instantes cortos de tiempo. Esto implica, que es posible analizar la señal de voz en pequeños instantes o tramas, garantizando que sus propiedades no cambien. Además, Para procesar una señal por un computador, la señal debe ser representada en forma digital para que pueda ser utilizado por un ordenador digital. La figura 11, muestra las etapas que sigue la señal de voz, antes de ser clasificada para extraer los parámetros fundamentales que nos servirán como base en el entrenamiento del sistema.

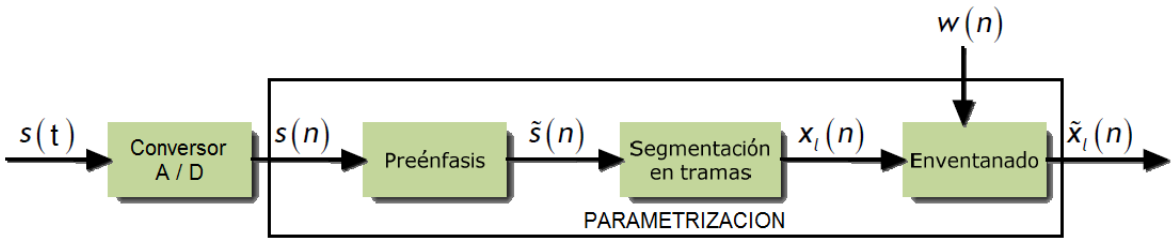


Figura 11. Preprocesado de la señal de voz, dividido en cuatro bloques: conversión, preénfasis, segmentación y enventanado.

2.5.2 Adquisición de la señal de voz.

Inicialmente, la señal acústica debe transformarse en una señal digital adecuada para el procesamiento de voz. Un micrófono o teléfono se utiliza para convertir la onda acústica en una señal analógica. Esta señal analógica se acondiciona con un filtro antialiasing (y posiblemente un filtrado adicional para compensar cualquier degradación del canal). El filtro antialiasing limita el ancho de banda de la señal a aproximadamente la velocidad de Nyquist (la mitad de la tasa de muestreo) antes del muestreo. La señal analógica es muestreada luego para formar una señal digital realizada por un convertidor analógico-a-digital (A / D) [1].

2.5.3 Parametrización

En el desarrollo de aplicaciones de voz será necesaria la reducción de la cantidad de información disponible, así como la extracción de dicha información en dominios donde ésta sea suficientemente robusta e independiente. En este sentido se propone la extracción de parámetros a partir de muestras de la señal, con el mencionado objetivo doble de reducir la cantidad de información a procesar y de expresar dicha información en dominios más adecuados.

A esta extracción es a la que denominaremos parametrización. La señal de voz muestra características pseudo-estacionarias solo a corto plazo, en órdenes de decenas de milisegundos. Por consiguiente, si se desea aplicar técnicas de análisis y tratamiento de voz, debemos limitar el segmento a procesar en este orden de magnitud. Esto da origen al denominado análisis localizado de la señal, que obligará al uso de tramas de voz de la duración reseñada. El mecanismo que nos permite, dada una señal de voz, realizar un análisis localizado mediante el uso de tramas consecutivas se denomina enventanado de la señal. Para la realización de este enventanado se aplicará (multiplicación) sobre la señal de voz completa una función limitada en el tiempo (ventana) lo que produce una nueva señal de voz cuyo valor fuera del intervalo definido por la ventana es nulo. Podemos expresar esto como.

(2.11) $$x(m) = s(n) * w(m-n)$$

Siendo s(n) la señal original (de larga duración), $w(n)$ la ventana temporal aplicada y $x(n)$ la trama de señal enventanada, que valdrá cero fuera del intervalo $n \in [n-N+1, m]$, siendo N

la duración en muestras de la ventana aplicada. De esta forma, la necesaria aplicación de técnicas de enventanado, que nos permitirán el análisis de tramos estacionarios, conlleva el efecto multiplicativo (ponderación) en el tiempo de la trama actual por los coeficientes de la ventana; y de forma equivalente, la convolución del espectro deseado de señal con la transformada de Fourier de la ventana correspondiente.

La ventana aplicada, es la denominada tipo Hamming, cuya estructura temporal está definida de la siguiente forma:

$$w(n) = \begin{cases} 0.54 - 0.46 * \cos\left(\frac{2\pi k}{N-1}\right) & : \text{si } k = 0,1,\ldots N-1 \\ 0 & : \text{en caso contrario} \end{cases} \tag{2.12}$$

Donde *N* es el tamaño de muestras en la ventana escogida.

Por otro lado, se debe considerar también el efecto de ponderación temporal, puesto que con las ventanas tipo coseno, las muestras de los extremos de la ventana quedan minimizadas frente a las muestras de la zona central de la ventana. Para compensar este efecto, se suelen tomar ventanas temporales solapadas, en las que las muestras extremas de una ventana sean las centrales de las ventanas consecutivas. El suavizado espectral es un efecto menos determinante que el lobulado espectral, razón por la que predomina la elección de ventanas con lóbulos secundarios bajos. La figura 12, muestra el efecto antes y después del enventanado, utilizando la ventana de Hamming.

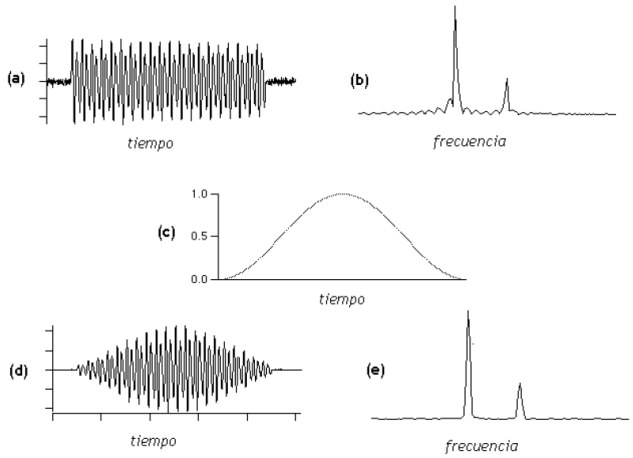


Figura 12. Utilidad de la ventana de Hamming. (a) Trama de voz segmentada. (b) Su espectro distorsionado debido a la segmentación. (c) La ventana de Hamming. (d) La trama enventanada. (e) Su espectro sin distorsión. [8]

2.6 Análisis Localizado en el Dominio de la Frecuencia

Una vez establecidas las ventanas de análisis de la voz, tenemos que en el análisis del espectro de una señal de voz aparecen dos componentes convolucionadas. Una es la proveniente de la frecuencia fundamental y sus armónicos y la otra de los formantes del tracto vocal.

El problema es que para el reconocimiento automático, es necesario el procesamiento digital de los datos y al ir ambas componentes convolucionadas será imposible tratarlas por separado teniendo que contar por ello con una componente (la frecuencia fundamental y sus armónicos) que no nos aporta ningún tipo de información útil.

Por ello habrá que utilizar algún método de análisis que, conservando las propiedades de la frecuencia sea capaz de separar adecuadamente cada una de las componentes sin que éstas pierdan sus propiedades concretas. Una transformación homomórfica es una transformación (^) que convierte una convolución en una suma:

(2.13) $$s(n) = e(n) * h(n) \rightarrow \hat{s}(n) = \hat{e}(n) + \hat{h}(n)$$

Ese método es el denominado análisis en el dominio cepstral que describimos a continuación.

2.7 Análisis en el Dominio Cepstral

El cepstrum (/kepstrum/), o coeficiente cepstral, $c(\tau)$, se define como la transformada inversa de Fourier del logaritmo del módulo espectral $|X(w)|$ (ver figura 13).

(2.14) $$c(\tau) = IDFT\left[\log|X(w)|\right]$$

La variable independiente en el dominio cepstral se denomina (siguiendo la misma lógica) "cuefrencia". Dado que el cepstrum representa la transformada inversa del dominio frecuencial, la "cuefrencia" es una variable en un dominio pseudo-temporal. La característica esencial del cepstrum es que permite separar elegantemente las dos contribuciones del mecanismo de producción: estructura fina y envolvente espectral.

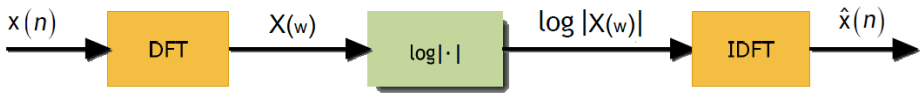


Figura 13. Pasos para el cálculo del cepstrum de la voz. [4]

Si denominamos $x(n)$ a la señal de voz, derivada de la convolución de la señal de excitación, g(n), con la respuesta impulsiva del tracto vocal h(n), y siendo $X(w), G(w), y\, H(w)$ sus DFTs respectivas, tendremos que:

(2.15) $$X(w) = G(w)H(w)$$

Si tomamos logaritmos sobre el módulo de esta expresión, tendremos:

(2.16) $$\log|X(w)| = \log|G(w)| + \log|H(w)|$$

Calculando ahora la transformada inversa, IDFT, resultará:

(2.17) $$c(\tau) = IDFT\left[\log\left|X(w)\right|\right] = IDFT\left[\log\left|G(w)\right|\right] + IDFT\left[\log\left|H(w)\right|\right]$$

Como se observa de la expresión anterior, en el dominio cepstral, las componentes de estructura fina y de envolvente espectral aparecen ahora como sumandos, en lugar de convolucionarse en el dominio temporal original. Además, en el dominio cepstral se verifica que las componentes debidas a la estructura armónica aparecen como picos equiespaciados a altas cuefrencias, justamente separados por el valor de τ que se corresponde con el periodo fundamental del tramo analizado. Puede demostrarse que el cepstrum de la voz tiene una forma como la que se muestra en la figura 14.

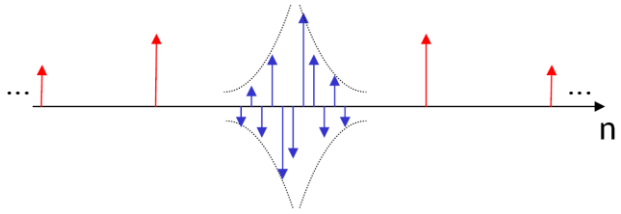


Figura 14. Representación del cepstrum de la señal de voz. Las muestras cercanas al origen corresponden al cepstrum del filtro, $\hat{\mathbf{h}}(\mathbf{n})$. Las muestras exteriores no nos interesan pues corresponden al cepstrum de la excitación [10].

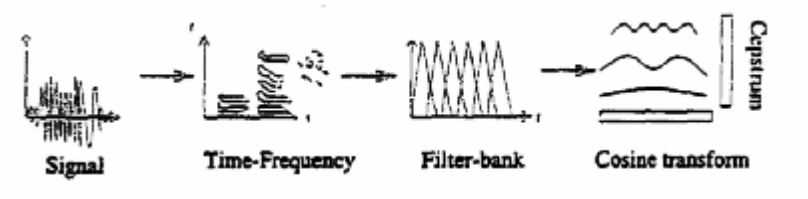
Según podemos observar en la anterior grafica, la respuesta al tracto vocal aparece en bajas cuefrencias, como señal impulsiva que abarca los primeros coeficientes cepstrales. Mediante un proceso de "liftering" o de filtrado en el dominio cepstral, podremos seleccionar y separar las componentes deseadas. Con un filtrado paso bajo tendremos la estructura de envolvente espectral. Si nos quedamos, por el contrario con los coeficientes altos, tendremos una estimación precisa del "pitch" de la trama de la señal bajo estudio. Si seleccionamos los primeros coeficientes cepstrales, que representan la estructura de formantes, y calculamos nuevamente la transformada de Furier sobre ellos, obtendremos ahora una buena estimación de la envolvente espectral. De esta forma, este proceso nos permitirá el cálculo de las formantes de la trama bajo análisis [2].

2.8 Cálculo de los Coeficientes Mel-Frequency Cepstrum, MFCC

Los coeficientes cepstrales derivados del análisis sobre la escala mel, o coeficientes mel-frequency cepstrum (MFCC) responden al uso de la escala mel, como escala en frecuencia ajustada al mecanismo de percepción auditiva, para realizar el cálculo de los parámetros cepstrales. Este cálculo se lleva a cabo a través de una representación definida como los coeficientes de la DCT del logaritmo de la energía de la señal de voz en cada banda perceptual:

(2.18) $$C_{MFCC}(m) = DTC\left\{\log\left(\sum_{k=0}^{N-1}\left|H_m^{\frac{1}{2}}(k)\,DFT\left\{s(n)\right\}\right|^2\right)\right\}, m = 0,\ldots,M_M$$

La figura 15, muestra un diagrama de bloques indica las operaciones a realizar sobre la señal de voz $s(n)$ para obtener los $C_{MFCC}(m)$.



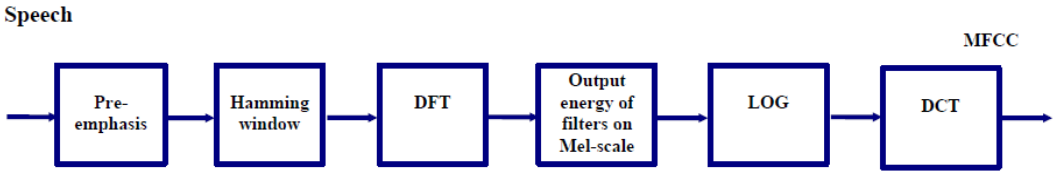


Figura 15. Diagrama de flujo para el cálculo de los coeficientes MFCC. [17]

El filtro perceptual tendrá la forma mostrada en la figura 16:

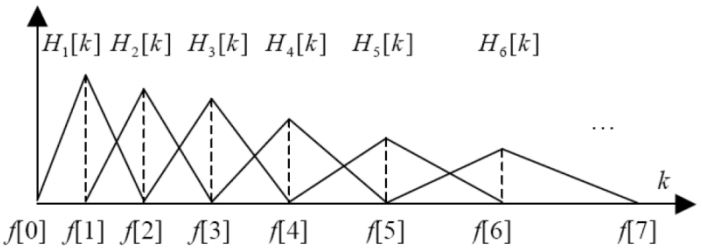


Figura 16. Filtro perceptual del oído humano. Está elegido de manera que los anchos de banda son uniformes en unidades de frecuencia Mel [11].

2.8.1 Parámetros Derivados: Parámetros Diferenciales

La información paramétrica se puede mejorar si se amplía el análisis para incorporar información diferencial (tanto de primer como de segundo orden). Para ello, se debe tener en cuenta que la derivada parcial respecto al tiempo del módulo espectral logarítmico tiene la siguiente expresión aproximada.

(2.19) $$\frac{\partial c_m(t)}{\partial t} = \Delta c_m(t) \approx \mu \sum_{k=-L}^{L} k * c_m(t+k)$$

Donde μ es una constante de normalización y $(2L+1)$ es el número de tramas sobre el que se extiende el sumatorio, con valores de $L=2$ para nuestro caso.

En el proceso de parametrización de las señales de voz, se calculan los coeficientes Δ y Δ Δ (diferencial de primer y segundo orden) de los MFCC. De esta manera obtenemos 19 coeficientes diferenciales de primer orden Δ y otros 19 coeficientes diferenciales de segundo orden ΔΔ, adicionales a los 19 MFCC que obteníamos para un total de 57 coeficientes cepstrales.

Por tanto, el vector de observación utilizado en reconocimiento del habla es una combinación de todas estas características:

(2.20) $$\vec{X}_k = \begin{bmatrix} \overrightarrow{c_k} \\ \Delta\overrightarrow{c_k} \\ \Delta\Delta\overrightarrow{c_k} \end{bmatrix}$$

2.9 Modelos de Mezclas Gaussianas (GMM)

Una vez que hemos obtenido los vectores de parámetros de cada una de las tramas que componen el fichero audio el resultado será que tenemos un flujo de vectores de dimensión 57, y tantos más vectores, cuanto más largo sea el archivo de audio.

En 1995, Reynolds introdujo el uso de modelos de mezclas de Gaussianas (GMM) para el reconocimiento del habla [12]. Desde entonces, se ha convertido en el método de referencia para esta tarea y es la base de los enfoques más exitosos que han surgido en los últimos años. Un GMM es una función de densidad de probabilidad de que, para un vector de características x, se define como:

(2.21) $$P(x \mid \lambda) = \sum_{k=1}^{K} w_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Donde K es el número de componentes de la mezcla, y los pesos w_k, cumplen la condición $\sum_{k=1}^{K} w_k = 1$, y la función:

(2.22) $$\mathcal{N}(x \mid \mu_k, \Sigma_k) = \frac{1}{|2\pi\Sigma_k|} \exp(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)$$

Representa la distribución gaussiana con media μ_k y matriz de covarianza Σ_k del vector D-dimensional, por tanto el modelo quedara denotado para cada hablante s como:

(2.23) $$\lambda_s = \left\{ p_k^s, \mu_k^s, \Sigma_k^s \right\}, \quad k = 1, \ldots, M$$

Reynolds también demuestra que mientras que el modelo general se puede aplicar para matriz de covarianza completa (full covariance), el uso de una matriz de covarianza diagonal representa una buena aproximación, ya que existe evidencia empírica de que el uso de las dos matrices de covarianza obtiene resultados similares, además, es posible demostrar que incrementando el orden de mezclas en la covarianza completa, este valor tiende a asemejarse a la matriz de covarianza diagonal [12] (Teorema de Karhunen-Loève)

2.10 Entrenamiento del GMM

El proceso de entrenamiento del GMM a partir de un conjunto de vectores X del locutor a modelar se realiza a través a través del cálculo de Máxima verosimilitud o ML.

El entrenamiento de Máxima verosimilitud o ML es el método más común de entrenamiento cuando no se posee ningún dato o modelo anterior. Se realiza a través del algoritmo Esperanza-Maximización (EM-Expectation Maximization) que se resume en:

1. Construimos un modelo inicial λ.

2. A partir de este modelo inicial λ , estimamos un nuevo modelo λ' que cumpla $P(x|\ \lambda') \geq P(x|\ \lambda)$

3. mientras que $P(x|\lambda') - P(x|\lambda) > \text{umbral de convergencia}$, volvemos al paso número 2.

De este modo, el nuevo modelo se convierte siempre en el modelo inicial y el proceso se repite hasta que se alcanza algún umbral de convergencia. Con este método Reynolds comenta el uso de 10 iteraciones suficientes para alcanzar la convergencia, ya que de lo contrario, existirá un excesivo procesamiento por parte del computador.

En [13] Reynolds usa un novedoso método de entrenamiento denominado UBM, (Universal Background Model), o modelo universal, El objetivo de este método, es realizar un modelo representativo de la generalidad del habla del ser humano, es decir, el efecto de este proceso será una "extracción" de las características que separan ese audio de las típicas que tiene el ser humano al hablar. Precisamente son esas diferencias las que harán modificar el modelo original, ya que las características que sean comunes a todos los hablantes serán absorbidas por nuestro UBM de origen (ver figura 17).

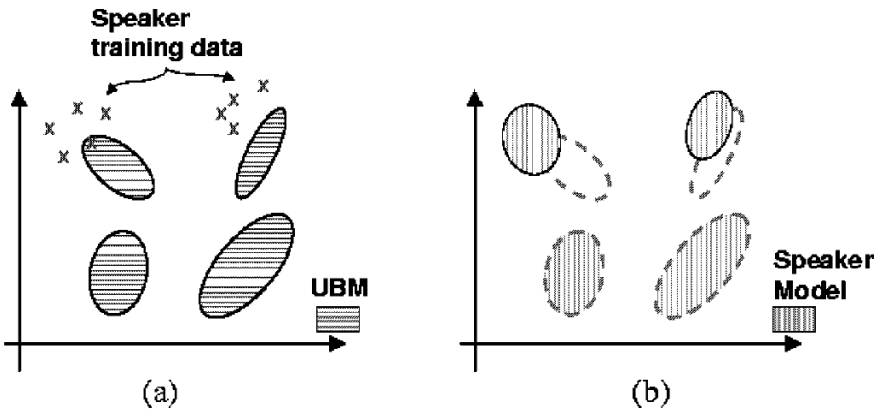


Figura 17. Ejemplo de entrenamiento por UBM, (a) vectores de entrenamiento, las x representan vectores de entrenamiento de un sujeto, mapeados en el modelo universal. (b) Los parámetros de mezcla adaptados se obtienen utilizando las estadísticas de los nuevos datos y los parámetros de mezcla UBM. [13]

2.11 Identificación del Habla Basada en GMM

El objetivo de una identificación de locutores es el de clasificar una señal de voz, cuyo origen no conocemos, como perteneciente a uno de entre un conjunto de N posibles locutores. Dentro

de estos sistemas. La figura 18, muestra un diagrama de bloques para un sistema de identificación del habla.

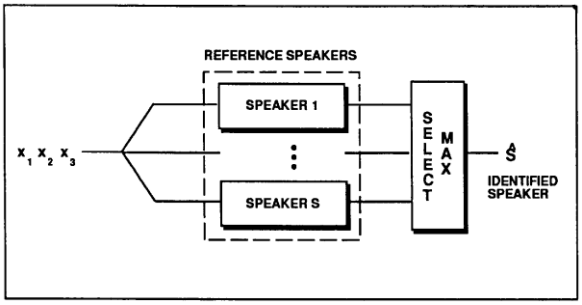


Figura 18. Diagrama de bloques de un sistema de identificación del habla. [13]

Veamos ahora cómo aplicar el modelado de locutores mediante GMMs para la identificación automática de los mismos mediante la voz. Si tenemos un grupo de S locutores $\{1,\ 2,\ldots,\ S\}$ representados por los modelos de mezclas gaussianas (GMM) $\{\lambda_1,\ \lambda_2,\ldots,\ \lambda_S\}$, el objetivo será encontrar el modelo de locutor que tenga la máxima probabilidad a posteriori de haber generado una secuencia de observaciones dada $X = \{x_1,\ldots, x_T\}$. El error mínimo para esta decisión basado en la regla de bayes nos da la siguiente ecuación:

$$\hat{s} = \arg \max_{1\le s\le S} P_r(\lambda_s \mid X) = \arg \max_{1\le s\le S} \frac{P_r(X \mid \lambda_s)}{P(X)} P_r(\lambda_s) \tag{2.24}$$

Sin embargo, dado que $p(X)$ es igual para todos los locutores, y suponiendo locutores equiprobables**,** entonces La regla de decisión se reduce a:

$$\hat{s} = \arg \max_{1\le s\le S} P_r(X \mid \lambda_s) \tag{2.25}$$

Usando logaritmos y la independencia entre locutores, el sistema de identificación del habla solo se tiene que calcular:

$$\hat{s} = \arg \max_{1\le s\le S} \sum_{t=1}^{T} \log p(x_t \mid \lambda_s) \tag{2.26}$$

2.12 Verificación del Habla Basada en GMM

El sistema de verificación es más complejo que el sistema de identificación, ya que el sistema tendrá dos entradas. La figura 19, muestra un diagrama de bloques para un sistema de verificación del habla.

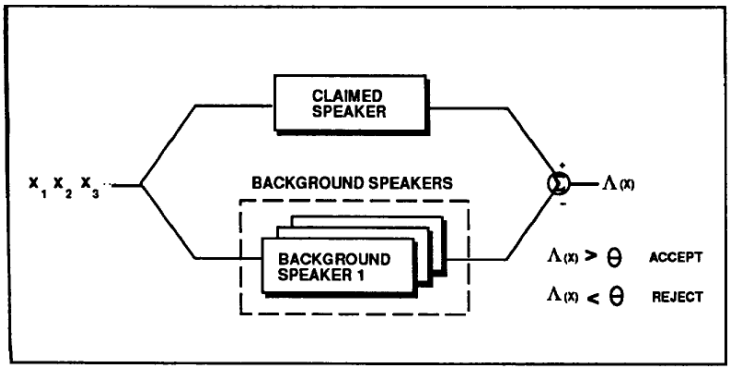


Figura 19. Diagrama de bloques de un sistema de verificación del habla [13]

Una de ellas es la señal de voz a verificar y la otra es una solicitud de identidad, que puede ser realizada de diversas formas. De este modo, las dos únicas salidas o decisiones del sistema son la aceptación H_0, o el rechazo H_1, de la hipótesis de que ambas locuciones pertenezcan a la misma persona. Por tanto, se plantean las hipótesis:

H_0: X pertenece al modelo del sujeto
H_1: X no pertenece al modelo del sujeto

La decisión de aceptar o rechazar la locución de entrada como correspondiente al locutor solicitado dependerá de si el valor del parecido o probabilidad obtenido supera o no un determinado umbral de decisión.

Las Relaciones de Verosimilitud o LRs (Likelihood Ratios) son una aproximación a la tarea de verificación introduciendo para ello las teorías bayesianas de apoyo a la decisión. Para una secuencia de observaciones dada $X = \{x_1, \ldots, x_T\}$. Y la solicitud de identidad de un sujeto con modelo GMM dado por λ_s, el LR dado esta dado por:

$$\frac{\Pr(\text{X pertenece al modelo del sujeto})}{\Pr(\text{X no pertenece al modelo del sujeto})} = \frac{P_r(\lambda_s \mid X)}{P_r(\lambda_{\bar{s}} \mid X)} \tag{2.27}$$

Aplicando la regla de bayes y usando las propiedades del logaritmo obtenemos:

$$LR = \log p(X \mid \lambda_s) - p(X \mid \lambda_{\bar{s}}) \tag{2.28}$$

3. ESPECIFICACIONES:

La herramienta escogida para la implementación del presente proyecto ha sido MATLAB. Siendo un lenguaje de programación de alto nivel basado en matrices y vectores que incorpora la posibilidad de una visualización grafica de los resultado. Además, puede incorporar una gran variedad de librerías denominados toolbox que extienden la cantidad de funciones contenidas en el programa principal. Uno de los toolbox que ha resultado muy útil para la implementación de este proyecto ha sido el de "Voicebox".

Con el fin de analizar los resultados obtenidos mediante MATLAB en el ámbito del reconocimiento, y dado que uno de los objetivos planteados en el presente proyecto, es el de mejorar la tasa de error del 5%, El trabajo se dividirá en dos grandes partes.

La primera, será la de imitar un sistema de identificación del habla, para esto trabajaremos con las características que usa el software de Agnitio, llamado Batvox [2], estas características serán reproducidas con ayuda del software de análisis matemático Matlab R2012b, para de esta manera, acercarnos lo mejor posible a los resultados que el software Batvox obtiene. La segunda etapa consta de implementar el mismo sistema de identificación del habla, pero en este paso incluiremos la característica de la curva melódica al sistema, para determinar las mejoras en el sistema.

3.1 Características de Batvox vs Matlab.

En los sistemas de identificación del habla, es muy importante encontrar una representación paramétrica de la señal que contenga información sobre la variación temporal de las propiedades de la señal. El objetivo de dicha representación paramétrica, es la extracción de las características acústicas más relevantes para el proceso de identificación del hablante.

El sistema desarrollado en Matlab consta de las siguientes características:

- **19 coeficientes cepstrales derivados del análisis sobre la escala mel.** Incluyendo la primera y segunda derivada para un total de 57 coeficientes.
- **40 Audios de entrenamiento.** Los audios de entrenamiento son aquellos audios con los que se generará un modelo del sospechoso que se utilizará en un cálculo biométrico.
- **40 Audios de test.** Los audios de test son aquellos audios de los que se desconoce la identidad del hablante y que se quiere enfrentar al modelo del sospechoso en un cálculo biométrico.

La tabla, muestra una comparación de las características usadas en el trabajo con Matlab, frente a las características de comparación presentes en el software Batvox de Agnitio.

ESPECIFICACIONES TÉCNICAS DEL SISTEMA	
Batvox Tecnología de reconociendo automático de locutor COREVOX 2.0 exclusiva de AGNITIO.	Matlab R2012b
CARACTERÍSTICAS DE LOS AUDIOS	
Batvox utiliza audios en formato WAV (PCM) con las siguientes características: 1. Mono 2. Frecuencia de muestreo: 8000 Hz 3. Resolución: 16 bits	Para Matlab también utilizamos audios en formato WAV con las siguientes características: 1. Mono 2. Frecuencia de muestreo: 16000 Hz 3. Resolución: 16 bits
CARACTERÍSTICAS DE LA EXTRACIÓN DE PARÁMETROS	
Las características de la extracción son: ✓ Tipo de ventana: Hamming ✓ Tamaño de ventana: 20 ms ✓ Solapamiento de 10 ms ✓ Número de filtros mel: 20 entre 0 y 4000 Hz ✓ Número de coeficientes por ventana: 38 (19 MFCC más sus deltas correspondientes)	Las características de la extracción son: ✓ Tipo de ventana: Hamming ✓ Tamaño de ventana: 20 ms ✓ Solapamiento de 10 ms ✓ Número de filtros mel: 20 entre 0 y 4000 Hz ✓ Número de coeficientes por ventana: 57 (19 MFCC más sus deltas y doble deltas correspondientes)
CARACTERÍSTICAS DEL MODELO UNIVERSAL (UBM)	
Batvox posee dos modelos universales, uno de mujer y otro de hombre. Las características de cada uno de ellos es la siguiente: ✓ Es un modelo GMM realizado con una mezcla de 1024 gaussianas ✓ Entrenado con el método ML ✓ Contiene más de mil horas de voz ✓ Cada uno incluye audios telefónicos, de móvil y de distintos tipos de micrófono	En Matlab utilizamos un Modelo Universal, creado a partir de 40 audios femeninos de duración 1 min. Para un total de 40 minutos de entrenamiento. ✓ Cuenta con un modelo GMM realizado con una mezcla de 1024 gaussianas ✓ Entrenado con el método ML ✓ Contiene 46 minutos de voz ✓ Solo se incluye audio grabado en formato wav.
CARACTERÍSTICAS DE LA NORMALIZACIÓN	
Los procesos realizados son CMN y filtros Rasta seguidos de un módulo de Feature Warping. Compensación de canal NAP para los modelos entrenados y Channel Factors para los test a identificar a partir de una matriz de compensación elaborada por Agnitio. Se normaliza con las poblaciones de referencia con la técnica T-Norm según se ha descrito en apartados anteriores y finalmente se normalizarán las puntuaciones con la técnica D-Norm para que sean independientes de la longitud de entrenamiento.	Con Matlab utilizamos la normalización CMN con Feature Warping, también se finaliza con la normalización T-Norm. No se usa ni compensación de canal NAP, ni Channel Factors por tratarse de audios grabados en estudio.

Tabla 2. Características de la Base de datos de Audio

Para la base de datos de los audios, se descargaron desde la página de la Universidad Nacional Autónoma de México [14] Con las siguientes características:

Se descargan 40 audios femeninos de duración aproximada de 8:00 min c/u, los cuales se dividen de la siguiente manera:

- ✓ 40 audios utilizados para entrenamiento (duración de 3:00 min).
- ✓ 40 audios de test, estos audios se recortan de tal manera que no hagan parte del mismo audio de entrenamiento (duración 15 seg).
- ✓ Adicional a lo anterior, se descargan 40 audios de test, los cuales no pertenecen a la misma grabación, este audio validara el sistema de una mejor manera (duración 15 seg).

Todos los audios fueron descargados con formato WAV, con frecuencia de muestreo de 16000 y 256 kbps.

3.2 Preprocesamiento de Audio.

Con ayuda del programa Matlab R2012b, se implementarán cada uno de los grandes bloques que se muestran en la figura 20.

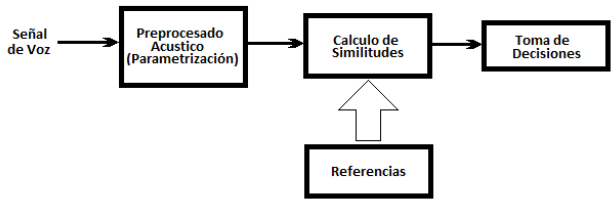


Figura 20. Proceso de identificación del hablante.

Para el sistema de identificación del habla implementado en Matlab R2012b, se trabajará con el toolbox denominado "VOICEBOX", desarrollado por Mike Brookes, Department of Electrical & Electronic Engineering, ya que estas rutinas ya han sido comprobadas, y permiten un ahorro de tiempo considerable para su uso.

3.3 Extracción de características para el entrenamiento del Modelo

La siguiente figura 21, se muestra el diagrama de bloques utilizado para realizar el cálculo de los coeficientes cepstrales.

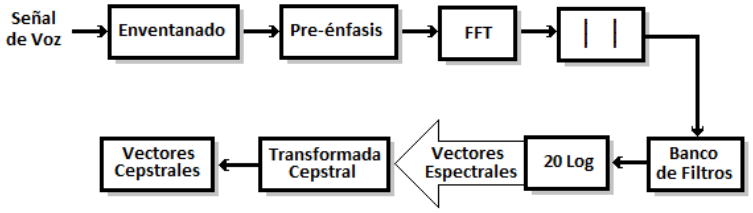


Figura 21. Proceso de extracción de los vectores cepstrales.

Debe tenerse en cuenta que al momento de calcular los coeficientes cepstrales (19 por cada 20 ms con solapamiento de 10 ms), también deberá incluirse al vector 38 coeficientes adicionales (correspondientes a la primera y segunda derivada).

3.4 Extracción de características de la curva melódica.

Para la extracción de la curva melódica, se partió del uso de PRAAT, la cual es una herramienta gratuita para el análisis fonético del habla desarrollada por Paul Boersma y David Weenink en el Instituto de Ciencias Fonéticas de la Universidad de Ámsterdam. Esta herramienta sirvió como base comparativa para la elaboración de un algoritmo en matlab que permita extraer y comparar la curva melódica de cada hablante.

La representación de las curvas melódicas es uno de los principales problemas que se plantean en el estudio de la entonación. La primera dificultad para el análisis prosódico del habla aparece ya en el propio proceso de detección de la frecuencia fundamental (F_0). Por otra parte, no se dispone aún de un sistema generalizado de representación de los fenómenos y unidades entonativas, por tal razón, el presente trabajo utilizará la característica del cepstro para la obtención de la frecuencia fundamental f_0.

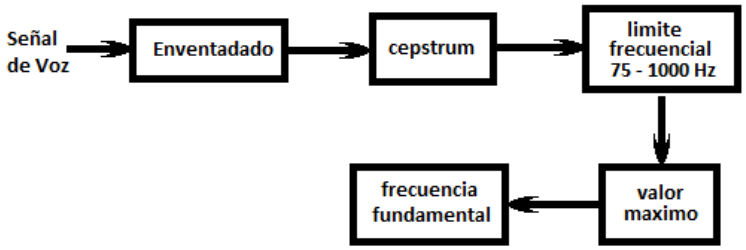


Figura 22. Proceso de obtención de la curva melódica.

Como se puede observa en la grafica 22, estos valores de frecuencia fundamental f_0 son los que permiten la obtención de la curva melódica, la cual se incluirán al vector cepstral junto con su primera y segunda derivada.

3.5 Análisis de resultados (Curvas DET).

Para el análisis de los resultados obtenidos se emplea el análisis y comparación de las curvas DET, Estas son las curvas más ampliamente utilizadas para evaluar el rendimiento de un sistema de reconocimiento de locutor. La figura 23 muestra un ejemplo de curva DET obtenido mediante la herramienta Matlab.

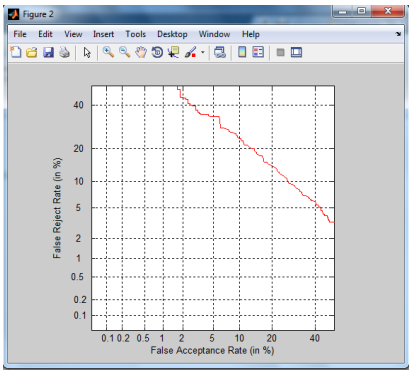


Figura 23. Ejemplo de curva de reconocimiento del habla

Cuando un sistema de reconocimiento de locutor exporta sus resultados, lo hace en forma de puntuaciones. La bondad de esas puntuaciones vendrá dada por la capacidad que tengan éstas en separar a las identificaciones correctas de las que no lo son.

Obviamente, las puntuaciones no tendrán implícitas en sí mismas una decisión de identidad o no identidad. Por esta razón, para realizar la verificación de si realmente un audio y un modelo pertenecen a la misma persona hará falta determinar un umbral a partir del cual las puntuaciones serán consideradas como identidad y que por debajo de ellas indicarán una no identidad.

Si, por ejemplo, queremos usarlo en una aplicación de alta seguridad en que lo que prima es que no haya ninguna probabilidad de identificación falsa. Es decir, que ninguna persona sea identificada erróneamente por el sistema (esta probabilidad es la que se denomina de Falsa Aceptación o FA) pondremos un umbral muy alto, aunque esto implique que en ocasiones sea no identificado algún individuo que debería haberlo sido.

Si por el contrario lo que prima es que nunca demos un resultado de no identidad a una comparación en la que ambos audios son de la misma persona (a la que denominamos probabilidad de Falso Rechazo o FR), tendremos que poner un umbral muy bajo aunque eso implique que el sistema realice falsas aceptaciones de vez en cuando.

De todo esto podemos deducir que la probabilidad de Falsa Aceptación y la de Falso Rechazo están íntimamente relacionadas y que cuando una aumenta la otra disminuye y viceversa. Además, también habrá un punto en el que las dos probabilidades se igualen que es el denominado punto de EER o Equal Error Rate. Para calcular el punto de EER de una curva DET no habrá más que ver cuál es el punto de la curva que corta la bisectriz de los dos ejes.

La otra gran propiedad que tienen las curvas DET es que si los resultados son lo suficientemente numerosos y la distribución de puntuaciones es en forma de una gaussiana normal, la curva resultante es recta y paralela a las otras lo que facilita enormemente la comparación de los sistemas. De esta manera, cuanto mejor sea un sistema más cerca estará su curva DET del origen de coordenadas.

4. DESARROLLOS:

4.1 Pre-procesamiento de Audio.

Luego de extraer y descargar los audios que permitirán realizar el entrenamiento y posterior comparación de los resultados, es importante notar que antes deben atravesar una serie de bloques que irán acondicionando la señal de audio a las características solicitadas.

- ✓ Estudio de cada audio, para comprobar que ningún otro hablante participa en la conversación.

Esta etapa fue realizada de manera empírica, es decir, se tomaron 46 audios, para posteriormente eliminar audios que no cumplen las características exigidas por el sistema, en total se escogieron 40 audios, eliminando cualquier componente de audio extraño (debido a la naturaleza de la voz, en algunos audios encontrábamos música de fondo, otro sujeto incluido en la conversación, o incluso audios en malas condiciones de escucha), de esta manera se garantiza que la base de datos es homogénea y no va a presentar alteraciones de resultados.

- ✓ Eliminación de silencios tras la descarga de los archivos de audio.

Se tomaron los 46 audios para la posterior eliminación de los silencios, ya que como se sabe, los silencios no aportan ningún tipo de información al sistema, y solo extienden el archivo de audio. Por esta razón se trabaja con 3:00 min de audio total para entrenamiento sin silencios.

- ✓ Conversión de los audios al formato especificado.

El formato de los archivos extraídos por la pagina de la Universidad Nacional Autónoma de México [14], se encuentra en formato mp3, para poder asemejarnos a las características de audio de Batvox, los audios atravesaron una etapa de re-muestreo y filtrado de silencio, para esta etapa utilizamos el software de edición de audio gratuito llamado “Audacity”, la figura 24, muestra el entorno visual del programa Audacity.

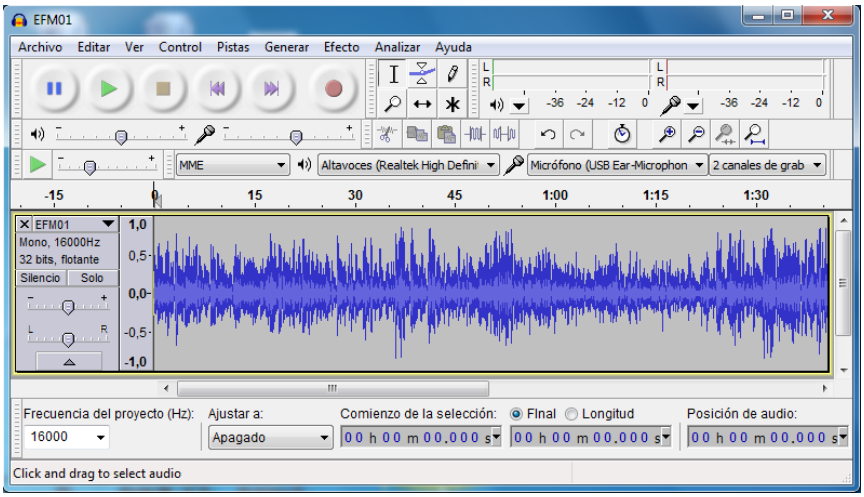


Figura 24. Programa Audacity para el preprocesamiento de la señal de audio

4.2 Base de datos de los audios en Matlab.

En esta etapa los audios fueron clasificados en 3 carpetas, para luego ser incluidas en el software matlab, los audios de entrenamiento como los archivos de test fueron almacenados en el entorno de trabajo de Matlab, con ayuda de una rutina sencilla incluida en el programa principal "IDENTHABLA.m".

Cada uno de los ficheros de audio se identifica a través de la duración de la señal, es decir el audio llamado "EFM09.wav", corresponde a un audio femenino de entrenamiento (del sujeto 9), ya que al observar las propiedades del archivo, se puede extraer la información del sujeto como la duración del audio (3:00 min). La figura 25, muestra las propiedades de cada audio, en este se puede identificar la identidad del sujeto, así como la duración del audio.

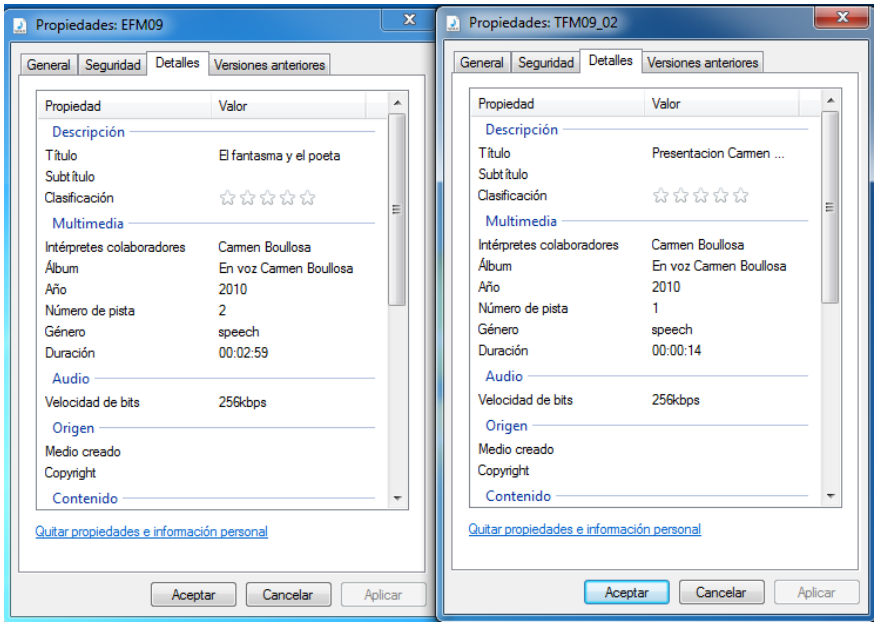


Figura 25. Propiedades del audio de entrenamiento y test para el sujeto 9.

4.3 Extracción de características para el entrenamiento del Modelo

A continuación se describen cada una de las etapas utilizadas en el proceso de extracción de características del modelo.

4.3.1 Calculo de los coeficientes cepstrales

Para este cálculo se trabajo con las rutinas propias del toolbox "voicebox". Las características de la extracción de los coeficientes son:

- ✓ Tipo de ventana: Hamming
- ✓ Tamaño de ventana: 20 ms
- ✓ Solapamiento de 10 ms
- ✓ Número de filtros mel: 20 entre 0 y 4000 Hz (no se incluye en el análisis el coeficiente cero, por tratarse de la energía de la señal)
- ✓ Número de coeficientes por ventana: 57 (19 MFCC más sus deltas y doble deltas correspondientes)

La Figura 26, muestra la extracción de 19 coeficientes cepstrales para una señal de voz de 10 segundos.

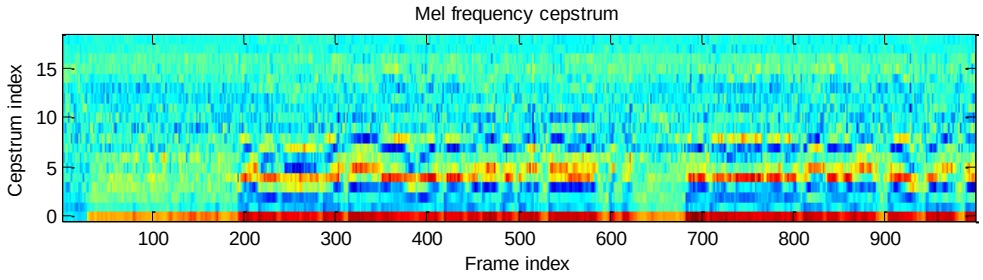


Figura 26. Obtención de los Coeficientes cepstrales mediante Matlab.

La rutina empleada en Matlab para el cálculo de los coeficientes cepstrales se denomina "melcepst.m", a continuación se muestra el uso del comando con los parámetros requeridos en la rutina.

C=melcepst(speech,fs,w,nc,p,n,inc,fl,fh);
Entradas Requeridas:
% speech Señal de voz a analizar
% fs frecuencia de muestreo (para nuestro caso 16000)
% w configura los deltas y doble deltas para el caso w=dD
% nc numero de coeficientes cepstrales sin incluir el coeficiente 0 (para nuestro caso nc=19)
% p numero de filtros a utilizar en el sistema (en nuestro caso usamos 20 filtros)
% n longitu de de la ventana de analisis (20 ms, es decir, n = 320)
% inc solapamiento entre muestras (10 ms, es decir inc=n/2)
% fl posición del filtro más bajo como una fracción de fs (fl=0)
% fh posición del filtro más alto como una fracción de fs (fh=0,5)
Salidas Obtenidas:
% C Coeficientes cepstrales de la señal de audio (57 por columna).

Adicional a la extracción de parámetros, se crea una rutina en Matlab, que permitirá realizar una normalización de media cepstral CMN (Cepstral Mean Normalization) con Feature Warping, para minimizar los efectos de las características de canal.

La figura 27, muestra un ejemplo de cómo se afectan los coeficientes antes y después de la normalización.

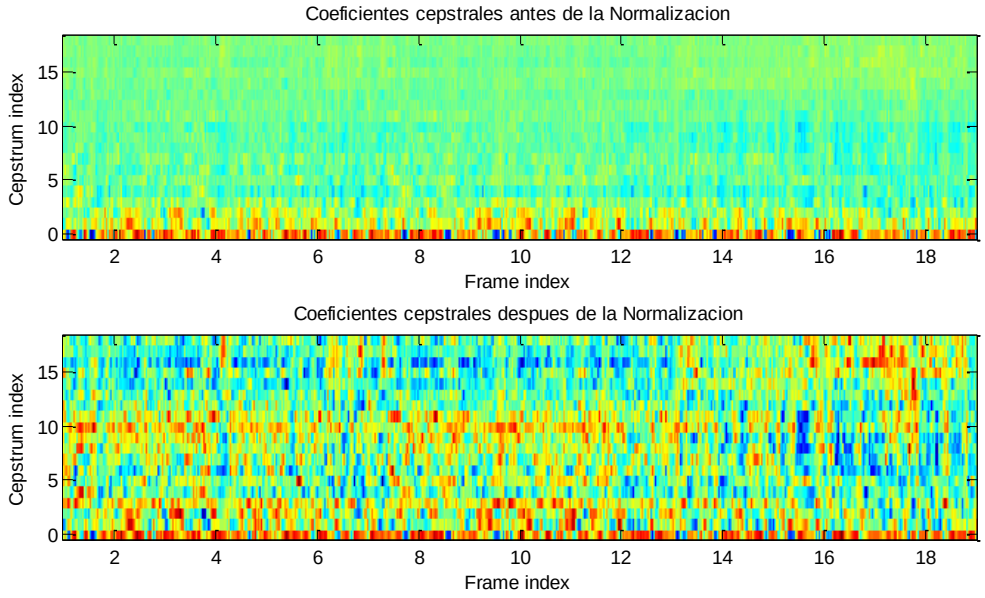


Figura 27. Comparación de los coeficientes Mel antes y después de realizada la normalización de Media Cepstral CMN y Feature Warping

4.5.2 Entrenamiento de los coeficientes con el modelo GMM

Una vez extraídos los coeficientes, se procede a entrenar cada uno de los 40 modelos. Para este entrenamiento se utiliza la rutina "gaussmix.m" propia del toolbox VOICEBOX de Matlab, que se incluye en la rutina realizada en el presente proyecto "IDENTHABLA.m", y que permitirán extraer, y almacenar los valores de media, matriz de covarianza y pesos relativos de 1024 gaussianas, para cada uno de los 40 audios, es decir, $\lambda_s = \left\{ p_k^s, \mu_k^s, \Sigma_k^s \right\}, \quad k = 1, \ldots, M$ para M=1024 y S=40. A continuación se muestra el uso de la rutina "gaussmix.m" en el presente trabajo.

- **Para la rutina de entrenamiento del modelo universal UBM**

[m0,v0,w0]=gaussmix(x,[],[],1024)

Entradas Requeridas:

% X(n,p) vector de coeficientes cepstrales (n= muestras p=57 coef.)
% 1024 Corresponde al valor de mezclas requeridas 1024

Salidas Obtenidas:

% m0(k,p) Vector de medias (k=1024 p=57)
% v0(k,p) Matriz de covarianza (k=1024 p=57)

% w0(k,1) pesos de las mezclas (k=1024)

- **Para el entrenamiento de cada modelo**.

[m,v,w]=gaussmix(x,[],[],m0,v0,w0)

Entradas Requeridas:

% X(n,p) vector de coeficientes cepstrales (n= muestras p=57 coef.)
% m0(k,p) Vector de medias del modelo universal UBM (k=1024 p=57)
% v0(k,p) Matriz de covarianza del modelo universal UBM (k=1024 p=57)
% w0(k,1) pesos de las mezclas del modelo universal UBM (k=1024)

Salidas Obtenidas:

% m(k,p) Vector de medias (k=1024 p=57)
% v(k,p) Matriz de covarianza (k=1024 p=57)
% w(k,1) pesos de las mezclas (k=1024)

El numero de audios S=40, fue escogido gracias al análisis que previamente realizó la empresa AGNITIO, la cual establece una escogencia entre 30 y 50 modelos para identificación, y para que los resultados sean acordes al análisis.

En Matlab realizamos un Modelo Universal, creado a partir de 40 audios femeninos en formato wav de duración 1 min. Para un total de 40 minutos de entrenamiento. Cuenta con un modelo GMM de 3 iteraciones realizado con una mezcla de 1024 gaussianas. El comando realiza el entrenamiento del modelo con el método ML. Luego de esto se ejecuta una sola iteración del modelo a calcular, pero partiendo de los parámetros de media, varianza y pesos obtenidos con el modelo universal.

Luego de tener los 40 modelos, procedemos a calcular la probabilidad de cada audio de test para proceder al cálculo de similitudes. Para este cálculo trabajamos con la rutina "gaussmixp.m", este comando retornará un vector con la probabilidad calculada para cada vector de coeficientes cepstrales que se ingresen. Adicional a esto, la rutina "gaussmixp.m", permite graficar en 1-D y 2-D los valores de probabilidad de salida. La figura 28, muestra la probabilidad de salida para un modelo de 1024 mezclas. Recordemos que gráficamente es imposible representar el plano de 1024-D, por tanto, se escogen solo 2 componentes para representarlo en el plano 2-D.

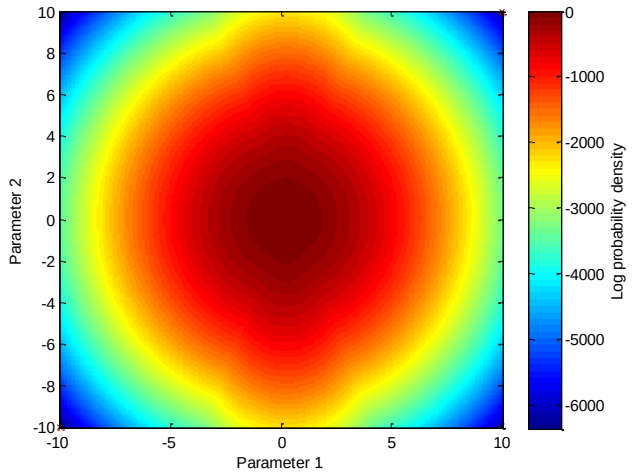


Figura 28. Representación de la probabilidad de salida en el plano 2-D de las 1024 mezclas de gausianas.

4.5.3 Verificación de los resultados obtenidos a través de las curvas DET.

Para el análisis de los resultados obtenidos, se emplea una rutina elaborada en matlab que permite calcular el error obtenido tras la comparación de comparación de las puntuaciones de falsa aceptación, que para este caso son las puntuaciones que cumplen la característica $p(X \mid \lambda_s)$ junto con las puntuaciones de falso rechazo, que cumplen $p(X \mid \lambda_{\bar{S}})$, estas puntuaciones darán una medida de semejanza entre las puntuaciones. La figura 29, muestra las puntuaciones de falsa aceptación (azul), junto a la de falso rechazo (rojo).

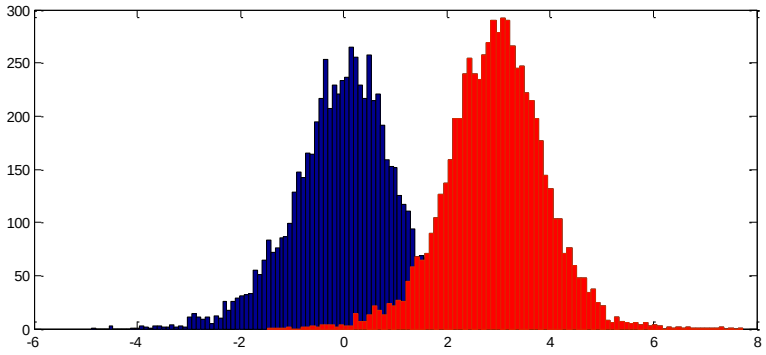


Figura 29. Puntuaciones de falsa aceptación (azul), junto a la de falso rechazo (rojo) para 40 audios.

Comparando estas puntuaciones a través de una rutina en matlab "compute_DET" se obtiene la tasa de error. La figura 30, muestra la Evaluación del sistema a través del cálculo de la taza de error o EER (Equal Error Rate) para un audio de test frente a 39 audios impostores. Es posible concluir que mientras el audio de test no pertenezca al sospechoso, los histogramas se verán más juntos, ya que las probabilidades arrojadas serán muy similares para el sospechoso como para los impostores, además, esto ocasionaría que grafica de error se encontrara más cercana a valores altos de probabilidad.

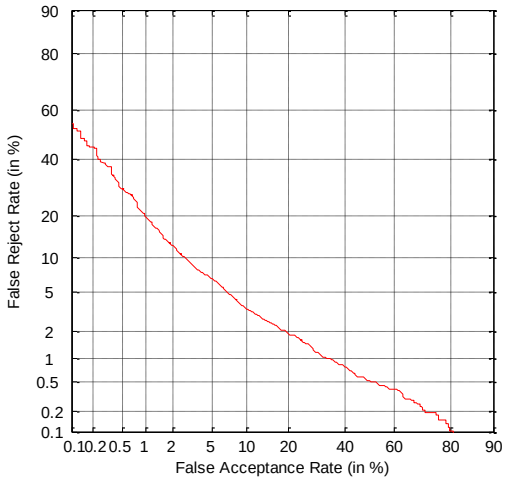


Figura 30. Evaluación de la taza de error para un audio de test frente a 39 audios EER=5.6%

4.5.4 Inclusión de Curva melódica al sistema de identificación del habla.

Para la extracción de la curva melódica se utilizo una rutina propia que opera con el concepto del cepstro, esta rutina denominada "curvmel.m", extrae la curva melódica de los sonidos sordos, para así incluir solamente sonido sonoro al sistema. La figura 31, muestra una comparación de la interfaz grafica desarrollada por el software praat y la creada en Matlab, Praat utiliza la función de autocorrelacion para el cálculo de la curva melódica, mientras que nuestra rutina utiliza el cepstro.

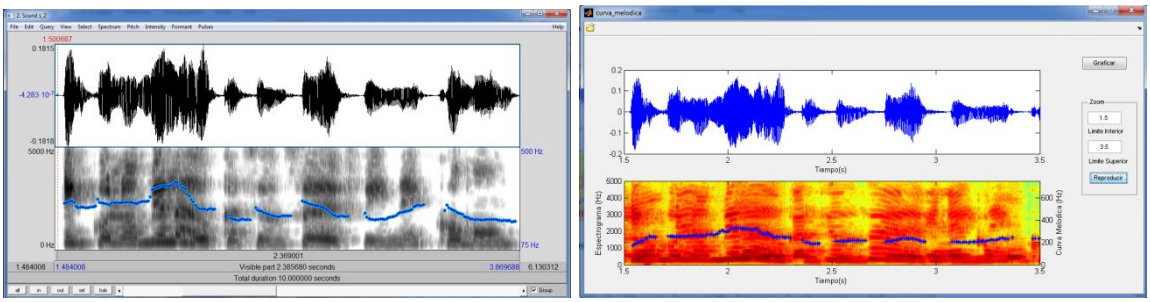


Figura 31. Comparación de la curva melódica obtenida mediante Praat vs Matlab.

Como se puede evidenciar en la figura 32, las curvas melódicas contienen variaciones y rupturas, relacionadas con las características intrínsecas de los elementos segmentales que componen los enunciados, como son las interrupciones en la curva debidas a la presencia de segmentos sordos, o las pequeñas variaciones en el curso de la curva melódica debidas a la aparición de sonidos fricativos u oclusivos. Estas variaciones, llamadas micromelódicas, no aportan información lingüística relacionada con la interpretación de la curva melódica, aunque pueden constituir un indicio adicional para el reconocimiento del segmento en cuestión [15].

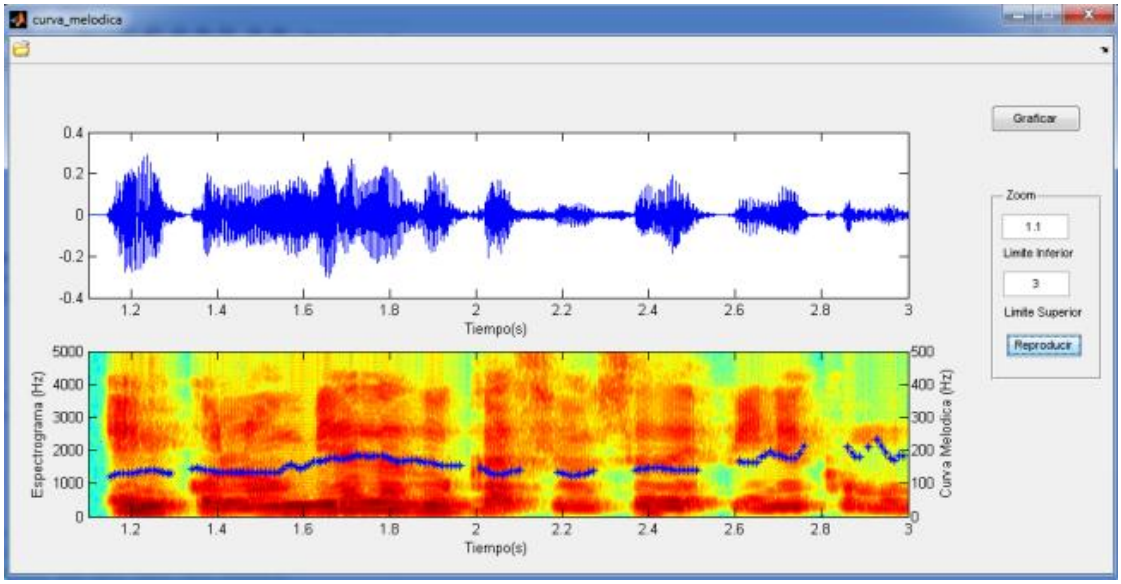


Figura 32. Calculo de la curva melódica para un segmento de habla de 3 segundos aprox.

Interrupciones debidas a la aparición de segmentos sordos a lo largo del enunciado

Los segmentos sordos, al carecer de frecuencia fundamental F_0, provocan interrupciones en la curva melódica, que pueden causar problemas en el proceso de interpretación. Estas interrupciones no son perceptibles para el oyente, que posiblemente realiza un proceso de reconstrucción del contorno, y no son pertinentes en el estudio de la entonación desde un punto de vista puramente lingüístico.

Variaciones abruptas de F_0 debidas a la naturaleza de los elementos segmentales

Las variaciones de F_0 debidas a la propia naturaleza de los elementos segmentales afectan, al igual que las relacionadas con la sonoridad, a la forma de la curva melódica. En una curva melódica pueden encontrarse fundamentalmente tres tipos de variaciones debidas a la naturaleza de los elementos segmentales [15]:

- El llamado 'fundamental intrínseco' de las vocales, que depende del grado de abertura de las mismas, de forma que cuanto más cerrada es la vocal, más alto es el valor de F_0
- Los pequeños descensos en la frecuencia fundamental ocasionados por las características de ciertas clases de consonantes, como las aproximantes y las vibrantes.
- Los ascensos y descensos que se observan antes y después de un segmento sordo, debidos a los efectos de la coarticulación.

Para el programa realizado en Matlab, se eliminaron las variaciones abruptas, debido a una etapa de estilizado incluida en la rutina de matlab. Lo que pretenderemos realizar en la presente investigación es determinar cuanto afecta el incluir los valores de sonidos sordos al sistema de identificación del habla, y cuanta mejora obtendríamos si eliminamos estos sonidos clasificados como sordos.

5. ANÁLISIS DE RESULTADOS:

Para la evaluación y mejora en la tasa de error en los sistemas de identificación del hablante, uno de los primeros retos a los cuales nos enfrentamos fue el de realizar un sistema de identificación del hablante que arrojara una tasa de error del 5%, para esto se diseño un sistema de identificación con ayuda de Matlab, que contenga gran parte de las características que relaciona el software profesional de reconocimiento automático de locutor llamado BATVOX de la empresa AGNITIO [2].

La figura 33 y figura 34, muestra el valor de error obtenido para la comparación para un audio de Test, con 39 modelos de población de referencia. Vale la pena notar que todos los audios escogidos para entrenamiento y prueba son femeninos, ya que según AGNITIO, la mezcla entre audios de diferente sexo, puede ocasionar desviaciones atípicas de las medidas.

Se determinó además que el error medio se mantiene en 5%, siempre y cuando las características de los modelos de la población de referencia sean iguales, es decir, es posible incluir o disminuir el número de modelos de comparación, siempre y cuando todos tengan las mismas características de señal.

Para los audios de test, pese a que pertenecen a conversaciones diferentes, deben tener las mismas características, es decir, debe ser grabadas en las mismas condiciones que el audio de entrenamiento, de lo contrario podría efectuarse un efecto del canal que no permita distinguir correctamente el audio, ocasionando que el error aumente.

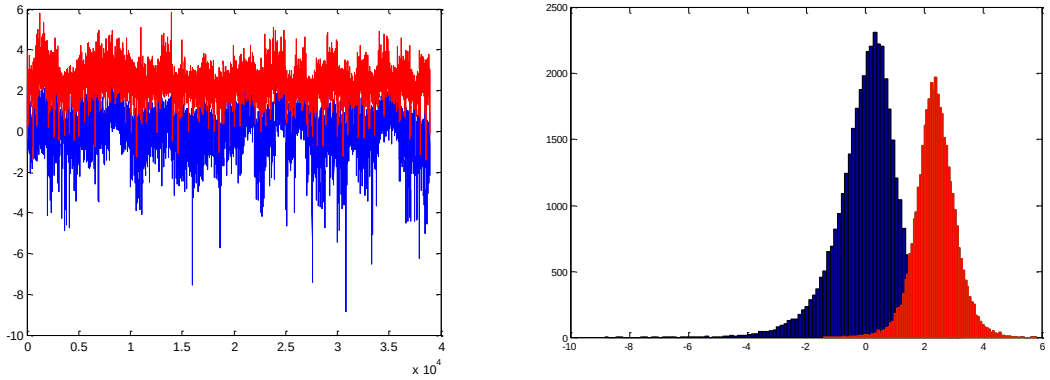


Figura 33. Puntuaciones (score) obtenidas mediante Matlab, para un audio de test, puntuación de falso rechazo (rojo) vs puntuación de falsa aceptación (azul).

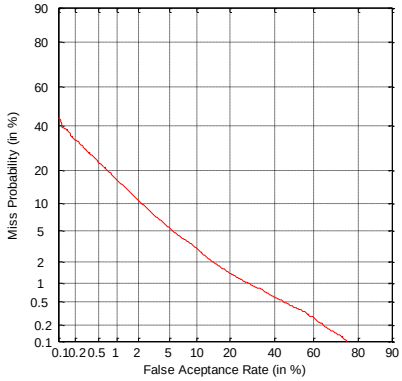


Figura 34. Rendimiento del sistema de identificación del habla para un audio de prueba y 39 sospechosos, EER= 4.99%

Luego de realizar el sistema para el cálculo del error, se realiza un estudio de la duración del audio de entrenamiento en Matlab, y se determino empíricamente que estos audios deben tener una duración de 1:00 minuto como mínimo para que el sistema arrojara un error de 5% como se indica en la figura 35. Además, más allá de este valor, observamos que el error medio no tiene una mejoría significativa.

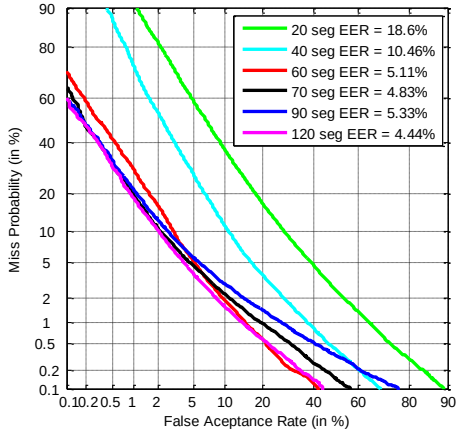


Figura 35. EER de audio de test en función del tiempo de entrenamiento

Numero de mezclas de gaussianas: Cuál número de mezclas de gaussianas utilizamos, si bien es cierto, mientras mas gausianas utilicemos, más fina será la resolución del cálculo de distribución frecuencial, pero, ¿cuántas mezclas estaremos dispuestos a utilizar?, recordemos que aumentar el número de mezclas también aumenta el tiempo del cálculo computacional. Se realizó por tanto una prueba para determinar cuánto varía la taza de error al variar el número de mezclas de gausianas. Como se puede observar en las figura 36. Al procurar entrenar el modelo con valores inferiores a 1024 el error es considerablemente superior que el deseado.

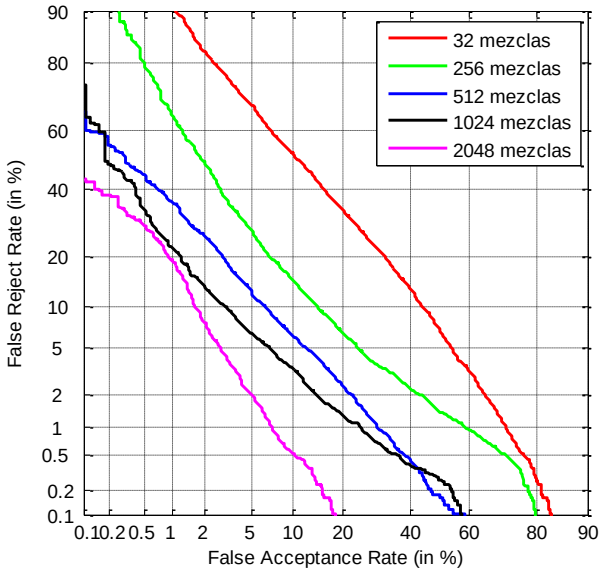


Figura 36. Mejora del error medio obtenido en un sistema de identificación del habla variando el numero de mezclas de gausianas.

Reynolds obtuvo resultados similares [13], al comparar el error obtenido en función del número de mezclas, la figura 37, muestra como hay una mejoría del error obtenido en función del orden utilizado para el entrenamiento. La razón por la que se escoge 1024 mezclas y no otro valor, obedece a las características extraídas del programa de AGNITIO, el cual usa estos valores.

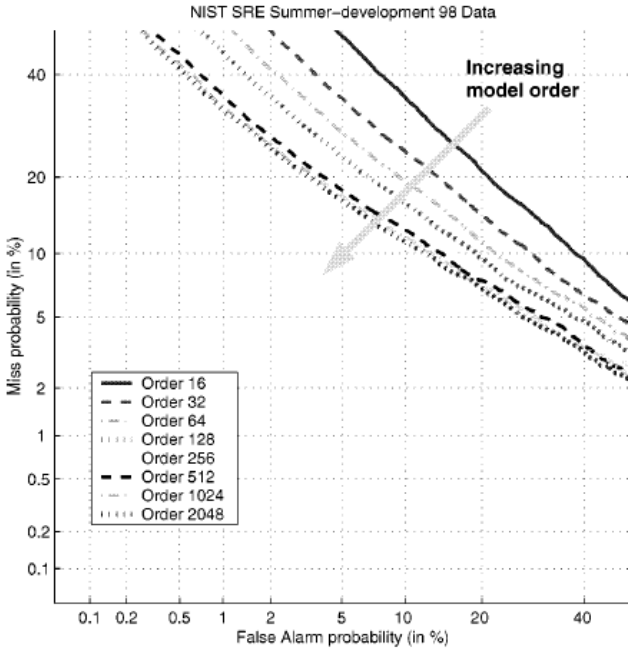


Figura 37. Curvas DET para un sistema basado en UBM con 16-2048 mezclas. [5]

Pruebas realizadas con la inclusión de la curva melódica:

Para incluir al modelo la curva melódica se ha intentado lo siguiente:
Dado que como la intención de la curva melódica es clasificar los sonidos sonoros y apartarlo de los sonidos sordos (ver figura 39), lo que intentamos fue escoger solo el segmento sonoro de los audio tanto de entrenamiento como el de prueba, y utilizarlos para poder entrenar y comparar los modelos. Los resultados obtenidos mejoraron el error obtenido para una persona usando el mismo audio de entrenamiento y el mismo audio de test, pero sin involucrar los sonidos sordos, mejoraron el sistema de 5.3% a 2.0% de EER

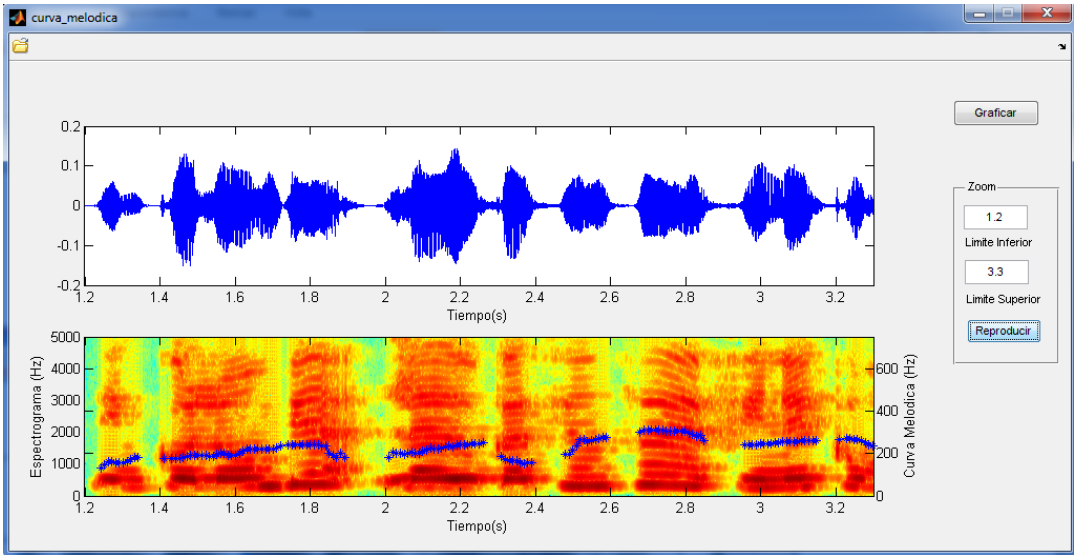


Figura 38. Extracción de la curva melódica para un segmento de voz

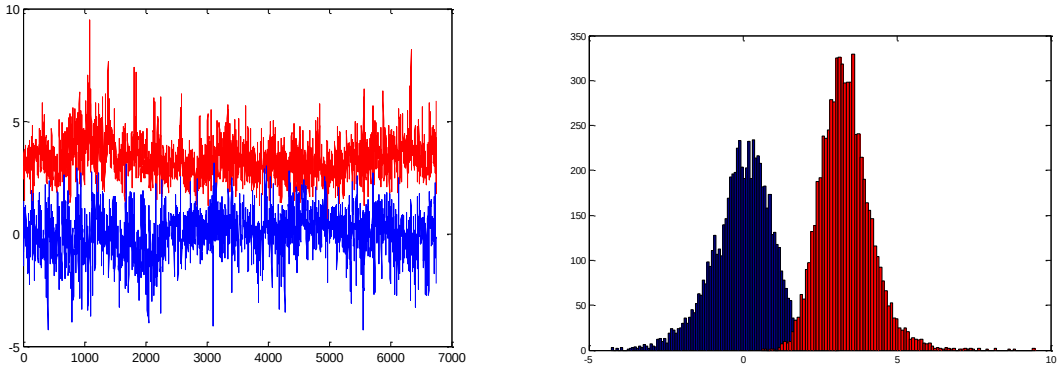


Figura 39. Puntuaciones (score) de un audio de test, en Azul, puntuación de falso rechazo, en rojo puntuación de falsa aceptación para un sistema de identificación del habla donde se incluye el valor de la curva melódica junto con sus derivadas.

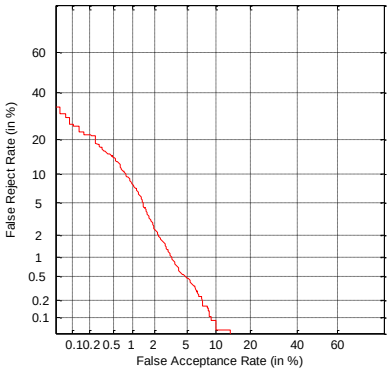


Figura 40. EER=2.0% para un test incluyendo curva melódica y sus derivadas

6. CONCLUSIONES.

- La estimación de la curva melódica, a partir del análisis en el dominio cepstral arroja buenos resultados, aunque es importante tener en cuenta que el enventanado que restringe el paso de las primeras componentes cepstrales (energía de la señal), debe ser muy selectivo, ya que una ligera atenuación a la componente de la frecuencia fundamental, ocasiona errores en las medidas.

- La curva melódica nos permite realizar una clasificación del sonido sonoro, diferenciándolo del sonido no sonoro o sordo, en el análisis se planteo la hipótesis que dada la naturaleza de los sonidos sordos, estos no contienen información de la identidad de la persona, ya que no depende de ninguna sección del tracto vocal, por tanto el eliminar este sonido, mejora significativamente el sistema tanto en el entrenamiento por GMM, como en la verificación con el audio de test, esto porque no existirán componentes similares a los demás audios.

- Al realizar pruebas con los 57 coeficientes MFCCS, y al extraer de los audios los sonidos sordos, se obtuvo un error medio de EER=2.0%. Después de esto quisimos agregar al modelo 3 coeficientes adicionales, correspondientes al valor de la curva melódica junto con sus dos derivadas, el sistema arrojo datos muy similares a los obtenidos sin incluir el valor de la curva melódica, por tanto, concluimos que el clasificar los sonidos sordos y eliminarlos para el entrenamiento, mejora el sistema, mientras que incluir el valor de la curva melódica no.

- Pese a que el sistema arroja un error comparable al desarrollado por el software AGNITIO, vale la pena notar que todos los audios escogidos se encuentran en las mejores condiciones.

7. LIMITACIONES Y TRABAJO FUTURO

- Pese a que el presente trabajo de investigación denominado ***"IDENTIFICACION DEL HABLANTE EMPLEANDO CEPSTRO Y CURVA MELODICA"***, pretendía realizar una mejora de la tasa de error basándose en la inclusión de la curva melódica como característica fundamental en el proceso de identificación del hablante, esta no se pudo realizar, sin embargo, si se pudo mejorar la tasa de error al 2%, aunque basándonos en el hecho que existen segmentos del habla que no aportan información al sistema, en otras palabras, todos los segmentos del habla donde la frecuencia fundamental este presente y bien definida, es un segmento valido para el sistema.

- El presente trabajo, se fundamento en la obtención del sistema de identificación del hablante a partir del estudio del software BATVOX de AGNITIO, y se resolvió trabajar con todas las consideraciones que allí se plantean, sin embargo, no se ha encontrado una sustentación teórica del por qué BATVOX trabaja con 1024 mezclas de gaussianas en su modelo, tampoco existe documentación de la incidencia del error del sistema en función de las normalizaciones (Feature Warping, T- Norm y D - Norm)

- Inicialmente, se tenían problemas utilizando el método de entrenamiento individual, ya que para cada modelo gastábamos 30 minutos aproximadamente por modelo en entrenamiento, por eso se tomó la decisión de trabajar con el modelo universal UBM, aunque no puede considerarse universal porque está conformado por audio limitado (solo 40 audios), como trabajo futuro se plantea la posibilidad de generar un modelo universal con por lo menos 3 horas de audio de diferentes características de habla. Para determinar su efecto sobre el sistema.

8. BIBLIOGRAFIA

[1] JOSEPH P. CAMPBELL, JR. Speaker Recognition: A Tutorial. PROCEEDINGS OF THE IEEE, VOL. 85, NO. 9, SEPTEMBER 1997

[2] AGNITIO CORPORATION, Batvox 3.0 Basic (Manual De Usuario), 2009

[3] FURUI, S. Speaker-independent isolated word recognition using dynamic features of speech spectrum, Acoustics, Speech and Signal Processing, IEEE Transactions on (Volume:34 , Issue: 1), February 1986

[4] LETICIA RUEDA ROJO, Mejoras en reconocimiento del habla basadas en mejoras en la parametrización de la voz, Universidad Autónoma de Madrid, 2011

[5] GASTÓN SALAMANCA Y STEPHEN MARLETT. Curso básico de fonética general, 2001 SIL International

[6] FLETCHER H. (1940). Auditory patterns. Rev. Mod. Phys. 12, 47-65.

[7] Hugo Fastl, Eberhard Zwicker, Psychoacoustics: Facts and Models, Springer Science & Business Media, 2007

[8] Castellanos Cesar Germán, Castillón Gómez Omar Danilo, "Comparación de algoritmos de estimación del pitch en el análisis acústico de la voz normal y patológica", VII Simposio de tratamiento de Señales, Imágenes y Visión Artificial, Bucaramanga. Colombia (2002).

[9], Henry Stark, John Woods Probability, Statistics, and Random Processes for Engineers, Hardcover 2011.

[10] Sergio Cruces Álvarez (2006). Apuntes de la asignatura "Tratamiento Digital de la Voz". Departamento de Teoría de la Señal y Comunicaciones (Universidad de Sevilla).

[11] PABLO AGUILERA BONET, Reconocimiento De Voz Usando HTK, Universidad De Sevilla.

[12] Douglas A. Reynolds. Speaker identification and verification using Gaussian mixture speaker models, Speech Communication 17 (1995) 91-108

[13] Douglas A. Reynolds, Speaker verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10, 19-41 (2000)

[14] Universidad Nacional Autónoma de México. http://descargacultura.unam.mx/app1

[15] ESTRUCH, M.- GARRIDO, J.M.- LLISTERRI, J.- RIERA, M. (1999) “Técnicas y procedimientos para la representación de las curvas melódicas”. Unpublished Ms. Grup de Fonètica, Departament de Filologia Española, Universitat Autònoma de Barcelona

[16] M. Bosi, R.E. Goldberg, Introduction To Digital Audio Coding and Standards, Kluwer Academic Publishers, Boston, 2003.

[17] Aldebaro Klautau, The MFCC, Departamento de Ciência da Computação, Nov 2005, http://www.cic.unb.br/~lamar/te073/Aulas/mfcc.pdf