UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de Telecomunicación

TRABAJO FIN DE GRADO

Caracterización de hablantes mediante extracción de información de cualidad vocal

Adrián García Cantalapiedra
Tutor: Joaquín González Rodríguez

Julio 2015

Caracterización de hablantes mediante extracción de información de cualidad vocal

Adrián García Cantalapiedra
Tutor: Joaquín González Rodríguez

Biometric Recognition Group - ATVS
Departamento de Tecnología Electrónica y de las Comunicaciones Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio 2015

Agradecimientos

No han sido pocas las horas dedicadas a este trabajo, ni tampoco lo ha sido el esfuerzo. Ahora ya solo queda un último empujón para cerrar otra etapa más en mi vida, y solo tengo palabras de agradecimiento para todos aquellos que han estado mostrándome su apoyo en los momentos buenos y en los no tan buenos.

En primer lugar me gustaría dar las gracias a Joaquín, que me ofreció la oportunidad de llevar a cabo este proyecto y me ha dado todas las pautas necesarias para sacarlo adelante. También a todo el grupo ATVS que siempre que he necesitado algo, se ha mostrado encantado de ofrecérmelo de forma desinteresada.

A todos los compañeros que he tenido durante estos 4 años, y que dentro de poco dejarán de ser compañeros para ser únicamente amigos. Nosotros mejor que nadie conocemos todo lo que hemos vivido en este tiempo, sería imposible llegar a buen puerto sin el apoyo mutuo que nos ofrecemos. No me puedo olvidar tampoco de todos esos amigos ajenos a la universidad, que en todo momento han entendido mis dificultades para encontrar el tiempo que se merecen, que por otro lado, es mucho.

Gracias Gema, compañera, amiga y mucho más. Gracias por saber entenderme como poca gente y apoyarme con las palabras y gestos que justo he necesitado en cada momento

Y por último, y sí por ello más importante, gracias a mi familia, que ha logrado soportarme en mis días malos y de estrés, yo personalmente no hubiera sido capaz. En especial, quiero agradecer la paciencia y el apoyo mostrado por mi hermana y mi madre en los momentos más duros.
Y gracias a ti, papá, gracias por haber sido la mejor referencia que jamás habría podido tener.

Resumen

Hoy en día el reconocimiento biométrico casi forma parte de nuestra vida cotidiana, y cada vez son más comunes los sistemas que consiguen detectar a personas por diferentes características físicas o de comportamiento del individuo, ya sea para aumentar la seguridad restringiendo el acceso a cierto servicio, o para ajustar las preferencias a las necesidades de cada usuario. Las características en las que se basan también han aumentado considerablemente en las últimas décadas, como pueden ser la huella dactilar, o como en el caso de este trabajo: la voz.

Este trabajo de fin de grado va a llevar a cabo una caracterización de los hablantes por medio de parámetros basados en la cualidad vocal, a diferencia de otros métodos más habituales, basados únicamente en información espectral.

Para ello, se explicará brevemente el funcionamiento del sistema fonador, del mismo modo que se hará una revisión de los métodos empleados para poder extraer las diferentes características de forma automática. Estos algoritmos se aplican con facilidad mediante el uso de un repositorio de códigos para el procesamiento de señal de voz denominado COVAREP. Seguidamente se repasarán las técnicas elegidas a la hora de realizar pruebas de identificación de hablantes, basadas en un sistema GMM-UBM. Por último antes de comenzar con las pruebas llevadas a cabo, se comentará el método de clustering aglomerativo o hacia arriba.

Cuando ya se tengan todas las bases teóricas, se mostrarán todos los resultados obtenidos para poder comprobar la bondad de los parámetros escogidos y su capacidad para caracterizar a los hablantes. En este documento también se recogen pruebas de identificación empleando estos nuevos parámetros, los mismos que se emplearán para crear nuevos grupos de diferentes tipos de locutor. Por último, se buscará combinar los resultados de identificación obtenidos, con otros basados en un sistema más convencional, empleando para ello la fusión de scores.

Palabras clave

Caracterización de locutor, GMM-UBM, NIST, pulso glotal, GCI, COVAREP, clustering aglomerativo, fusión de scores

Abstract

Nowadays, biometric recognition is almost part of our daily lives, and are becoming more common systems which detect people by different physical or behavioral characteristics of the individual, either to increase security by restricting access to certain service, or to adjust the preferences to the needs of each user. The characteristics on which are based, have also increased significantly in recent decades such as fingerprint, or as in case of this work: the voice.

This TFG will conduct a characterization of the speakers through parameters that are based on the vocal quality, unlike other more conventional methods, which are only related to spectral information.

To do this, we briefly explain the functioning of the vocal system, just as there will be a review of the methods used to extract the different characteristics automatically. These algorithms are implemented easily by using a collaborative repository for speech processing, called COVAREP. Then, the chosen techniques to the speaker identification tests will be explained. Finally, before beginning the test carried out, it will be discussed the method of agglomerative clustering, or from bottom to top.

When you already have all the theoretical knowledge, all the results are displayed, to check the goodness of the chosen parameters and their ability to characterize the speakers. In this document identification tests are also collected using these new parameters, the same that will be used to create new groups of different types of speaker. Finally, we seek to combine the identification results obtained with other methods based on a conventional system, employing score fusion.

Keywords

Speaker characterization, GMM-UBM, NIST, glottal source, GCI, COVAREP, agglomerative clustering, score fusion

Glosario:

COVAREP:	Collaborative voice analysis repository
EE:	Excitation strength.
EER:	Equal Error Rate
EM:	Expectation-maximization
F0:	Frecuencia fundamental
FIR:	*Finite Impulso Response*
GCI:	*Glottal Closure Instant*
GMM:	*Gaussian Mixture Model*
GOI:	*Glottal Opening Instant*
H1-H2:	Relación entre los dos primeros armónicos
HRF:	Harmonic Richness Factor
IAIF:	*Iterative Adaptative Inverse Filtering*
IIR:	*Infinite Impulso Response*
LPC:	*Linear Predictive Coding*
MAP:	*Maximum a posteriori*
MDQ:	*Maxima Dispersion Quotient*
MFCC:	*Mel Frequency Cepstral Coefficients*
NAQ:	*Normalized Amplitude Quotient*
NIST-SRE:	US National Institute of Standards and Technology – Speaker Recongnition Evaluations
OQ:	Open quotient
PS:	Peak slope
PSP:	*Parabolic Spectral Parameter*
QOQ:	*Quasi-Open Quotient*
RA:	Dynamic leakage
SRH:	*Summation of Residual Harmonics*
UBM:	Universal Background Model

Índice de contenidos

Índice de figuras

Índice de tablas

1 Introducción

1.1 Motivación y objetivos

En la actualidad los sistemas clásicos que buscan caracterizar al locutor y poder realizar identificación de hablantes a través de señales de voz se basan en propiedades del tracto vocal, como es el caso de los Mel Frequency Cepstral Coefficients (MFCCs), basados en la percepción del sistema auditivo de los humanos.

La principal motivación de este proyecto es trabajar con parámetros relacionados con la excitación del tracto, totalmente diferentes a los habituales. Cada persona tiene una forma diferente de producir sonidos, lo que denominamos cualidad vocal, y estas diferencias no se deben únicamente a la contribución del tracto vocal. Las cavidades intraglóticas, la forma en la que una persona hace vibrar sus cuerdas vocales, o la velocidad con la que lo hace, entre otros factores, también aportarán una información muy valiosa para poder caracterizar a un locutor.

Al contrario de lo que ha ocurrido con los coeficientes cepstrales, históricamente ha sido complicado extraer características glotales a través de una señal de voz, siendo necesario recurrir a aparatos eléctricos para obtener, por ejemplo, una estimación del pulso glotal producido por una persona al hablar. Esto era de poca utilidad a la hora de desarrollar un sistema que busca caracterizar a los hablantes de forma automática a partir de una serie de grabaciones.

Esto ha cambiado recientemente, y se han desarrollado diversos algoritmos capaces de detectar características glotales a partir de señales de voz. Estos algoritmos han sido recogidos en un repositorio denominado COVAREP (Collaborative voice analysis repository) elaborado de forma conjunta por centros de investigación muy prestigiosos en este campo. Gracias a esta recolección de métodos y técnicas, se facilita la extracción de características glotales en una señal de voz, y será una herramienta de vital importancia a lo largo de este trabajo de fin de grado.

Para llevar a cabo las pruebas que nos indicarán la bondad de estos parámetros en la caracterización del locutor, se emplearán los mismos entornos de evaluación que los utilizados en sistemas basados en características del tracto vocal, los cuales están muy desarrollados y consolidados. Serán llevadas a cabo en audios con independencia del texto, es decir, sin tener conocimiento de lo que está diciendo el hablante, y en conversaciones telefónicas. No se marca como objetivo principal igualar o superar los resultados obtenidos con los parámetros habituales, pero sí comprobar hasta dónde pueden llegar los parámetros glotales por sí solos, y por otro lado determinar si pueden ser combinados de alguna forma con los sistemas empleados hasta ahora.

Además de las pruebas de identificación, se buscará clasificar a los locutores en diferentes grupos empleando métodos de clustering. Esta tarea puede servir para poder hacer una discriminación a la hora de aplicar la identificación con los sistemas más eficaces hasta el momento, de modo que si dos locutores pertenecen a grupos diferentes, se sabrá que no pueden ser el mismo locutor, siempre y cuando la separación en grupos sea lo suficientemente buena. Esto resultará de especial utilidad en los casos en los que gracias a los parámetros glotales se detecten diferencias entre locutores que no son posibles apreciar en otro tipo de características.

Por último, también se quiere comprobar si se pueden combinar los resultados de identificación de locutores obtenidos con los parámetros glotales con los obtenidos con métodos tradicionales como es la parametrización mediante MFCCs.

1.2 Estructura

Este documento está estructurado de la siguiente forma:

Motivación, objetivos y estructura. Explicación de los motivos por los que se ha llevado a cabo este trabajo y se detalla cómo se organiza la información.

Estado del arte. Se aporta la información necesaria sobre el sistema fonador y diferentes algoritmos para obtener parámetros relacionados con este sistema y que servirán para caracterizar a los locutores. Explicación de las técnicas usadas para modelar locutores y para contrastar diferentes modelos de locutor. Descripción del método de agrupación empleado.

Diseño y desarrollo. Se detalla el tipo de audios empleados para llevar a cabo las pruebas, se hace una evaluación de los parámetros que pueden ser de interés y se describen los métodos empleados para conseguir los resultados de la sección siguiente.

Pruebas y resultados. Se aportan datos y gráficas sobre las pruebas llevadas a cabo.

Conclusiones y trabajo futuro. Breve análisis de los procesos y los resultados conseguidos, y repaso a las posibles líneas de avance tras este trabajo.

2 Estado del arte

Para poder llevar a cabo una correcta compresión de los procesos explicados en este trabajo, será fundamental conocer los elementos básicos del sistema gracias al cual los seres humanos podemos hablar, la forma de comunicación por excelencia y distintiva de nuestra especie.

2.1 Aparato fonador

El aparato fonador, no solo nos permite comunicarnos, sino que debido a las diferencias entre distintos locutores, nos permite diferenciar e identificar las fuentes. Se divide en los órganos relacionados con la producción de voz y aquellos que participan modificando ésta mediante resonancias y modulaciones generando diferentes sonidos. En la siguiente figura se puede ver esta división:

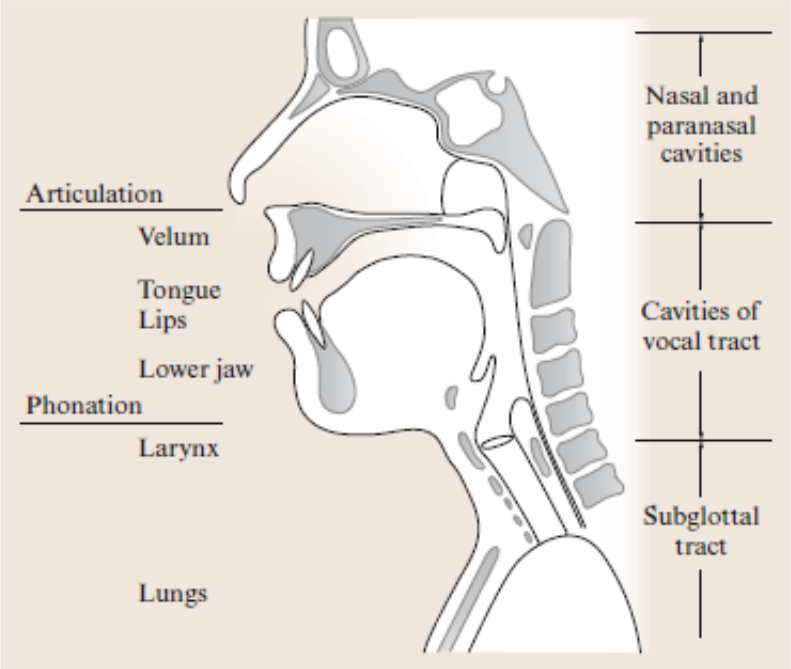


Figura 2.1. Representación del aparato fonador [1]

Los sonidos que podemos emitir los humanos gracias a este sistema, se pueden dividir en sonidos sordos y sonoros. Los sordos se caracterizan entre otras cosas por no presentar estructura armónica, y la excitación se podría modelar por ruido blanco. Dentro de este grupo de sonidos encontramos fonemas como /S/ o /f/. Por otro lado, los sonoros sí que presentan estructura armónica y en esta categoría entran todas las vocales y otros fonemas como /m/.

Uno de los componentes del sistema más importantes para este documento pertenece a los órganos relacionados con la producción de voz, la laringe. Está situada encima de la tráquea y su función principal es la de evitar la entrada de cualquier cuerpo extraño a los pulmones. Está formada por varias estructuras rígidas como la epiglotis.

En este órgano podemos encontrar las cuerdas vocales, cuya oscilación convierte el aire expulsado por los pulmones en pulsos intermitentes. El hueco entre las cuerdas vocales se denomina glotis, que continuamente cambia de forma mientras se produce la fonación, estrechándose y ensanchándose, y cuyo tamaño es variable, en función de factores como el sexo.

Durante la producción de sonidos sonoros, las membranas de las cuerdas vocales entran en contacto en un movimiento repetitivo y periódico de apertura y cierre de la glotis. [1] Este proceso puede dividirse en 4 fases, como se ve en la siguiente figura:

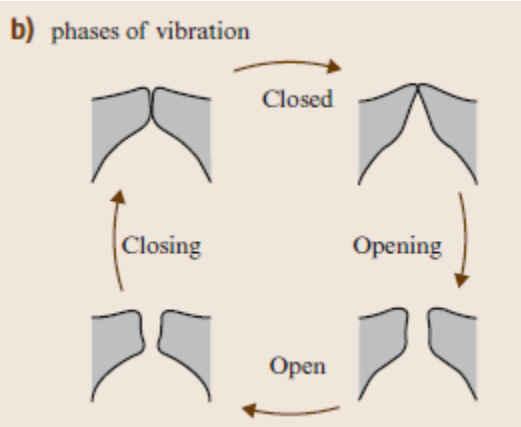


Figura 2.2. Fases de apertura y cierre glotal [1]

Los diferentes tipos de sonidos sonoros se generan por la diferente disposición del tracto vocal, donde encontramos varias cavidades que generan frecuencias de resonancia y dan lugar a máximos en la envolvente espectral en dichas frecuencias, denominadas formantes. En la figura Figura 2.1 se muestran estas cavidades del tracto.

En contraste con los sonidos sonoros, en los sonidos sordos, la glotis permanece abierta y el sonido se produce por obstrucciones y colusiones en el tracto vocal, manteniéndose la glotis abierta, permitiendo el paso del aire proveniente de los pulmones hacia el tracto vocal.[14]

Por lo tanto, el aparato fonador puede ser descrito como un sistema donde el flujo de aire que atraviesa la glotis, es modelado por el tracto vocal que actúa como un filtro acústico, de modo que una misma señal de excitación puede dar lugar a diferentes sonidos en función de la disposición del tracto. En la siguiente figura se observa este fenómeno para señales ideales, donde con la misma señal de excitación se obtienen dos vocales diferentes [u] e [i]:

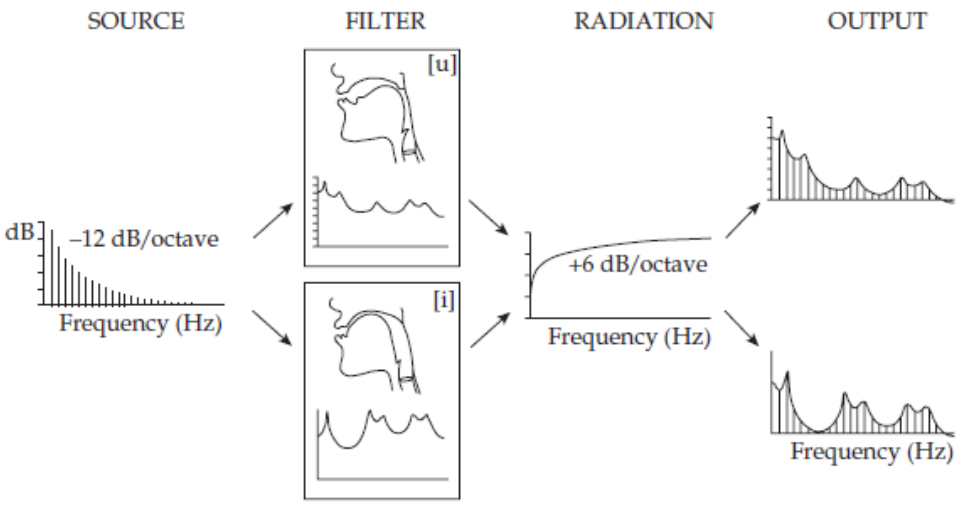


Figura 2.3 Representación esquemática del proceso de producción de las vocales (u) e (i) [21]

2.1.1 Cualidad vocal

La forma de hablar característica de cada individuo viene dada por la cualidad vocal, que puede ser definida como aquellas características que están presentes en mayor o menor medida mientras una persona está hablando. Estas características pueden ser divididas en aquellas producidas en la laringe, y las que lo son en zonas superiores a ésta. Un ejemplo sería la disposición del tracto vocal a la hora de emitir cierto sonido, o la velocidad con la que se abre y se cierra la glotis.

Típicamente, los fonetistas han estado más interesados en las actividades que ocurren en las áreas supralaríngeas que en aquellas relacionadas en la excitación, ya que el número de características lingüísticas presentes en el tracto vocal es mayor que la cantidad de diferentes tipos de fonación. No obstante sí que se han diferenciado tipos de excitación, que pueden ser de utilidad para caracterizar a hablantes.

Para poder hablar de diferentes configuraciones fonatorias, es necesario establecer antes un modo neutral, que pueda considerarse como el normal, y a partir de éste observar las diferencias con el resto. Este modo ha sido denominado voz modal, y se caracteriza por tener una vibración de las cuerdas vocales periódica, eficiente y sin ninguna fricción audible. El resto de configuraciones fontarios incumplen una o varias de las propiedades de la voz modal, como es el caso de la voz breathy (aspirada), donde la excitación es ineficiente y con una ligera fricción. Es importante mencionar, que estas formas de excitación características de cada locutor pueden variar también a lo largo del tiempo en función de las situaciones.[20]

2.1.2 Proceso fonatorio

En este trabajo nos vamos a centrar en encontrar las diferencias entre locutores, en este proceso de fonación, por lo que veremos más en detalle cómo se produce la excitación en las señales de voz. Las tareas de conseguir características de la excitación glotal han sido llevadas a cabo tradicionalmente gracias a aparatos eléctricos como es el electroglotógrafo, que detecta el contacto de las cuerdas vocales durante la vibración gracias al uso de electrodos sobre la piel capaces de detectar las variaciones de las cuerdas vocales, y así predecir los instantes en los que se cierran y abren, como se ve en la siguiente figura, donde también se puede observar imágenes reales del proceso de cierre y apertura glotal:

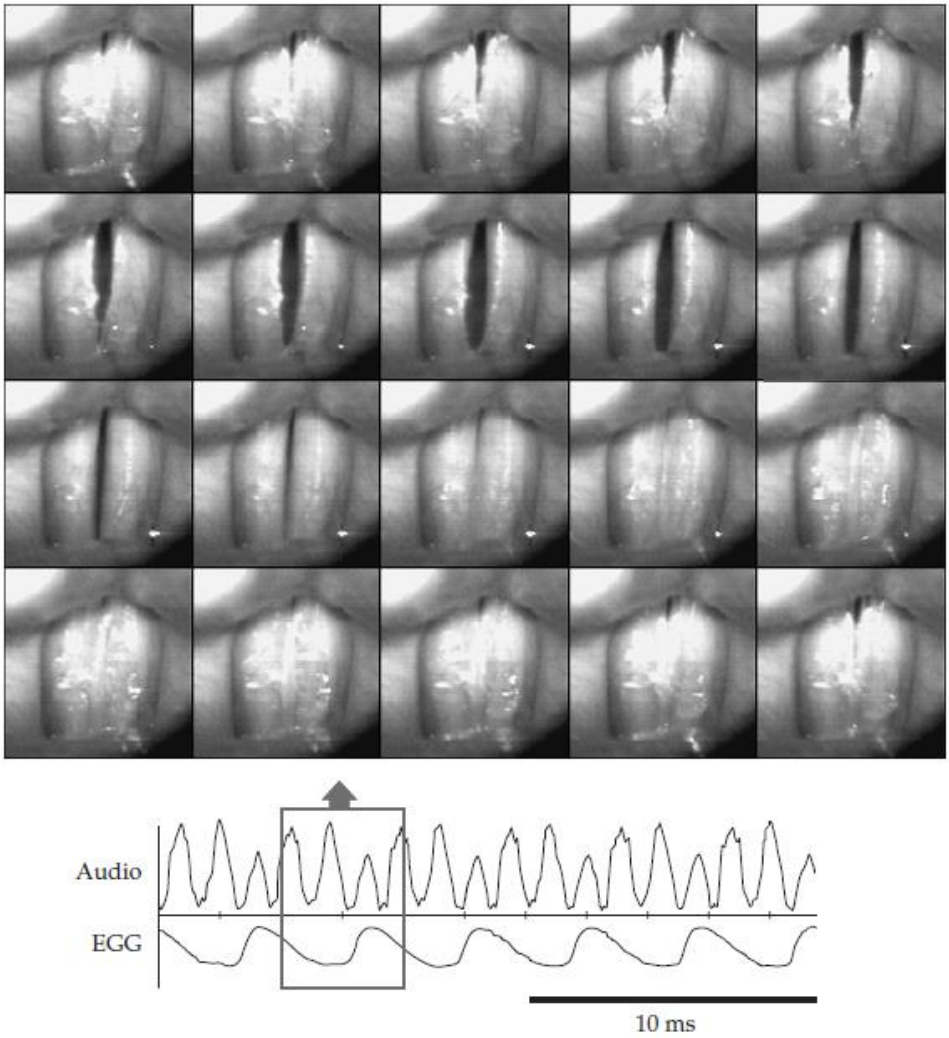


Figura 2.4 Imágenes de la laringe en el proceso de producción de la vocal /e/ y la representación de la señal de voz y el resultado obtenido con un electroglotógrafo

En el caso que nos ocupa, será necesario poder estimar el pulso glotal a través únicamente de la señal de voz. La mayoría de las técnicas que consiguen esto se basan en el filtrado inverso. Con estos métodos, la señal de voz se pasa por un filtro cuya función de transferencia es el

inverso de la función de transferencia supraglotal. Este filtro inverso debe cancelar el efecto de las resonancias del tracto vocal manifestadas en el espectro de la señal como formantes en cualquier instante. Esta tarea será especialmente crítica en el primer formante, donde un error causará gran distorsión en la estimación de pulso glotal.

Además de cancelar el efecto del tracto vocal, también será necesario hacer lo mismo con el efecto causado por los labios en la radiación, lo cual se lleva a cabo mediante la integración de la señal.

Por lo tanto, será esencial conseguir una función de transferencia del tracto vocal lo más precisa posible para conseguir mejores estimaciones del pulso glotal. Muchas de las técnicas empleadas para ello recurren al proceso LPC (Linear predictive coding), como es el caso de nuestro trabajo, donde se ha empleado el algoritmo "Iterative Adaptive Inverse Filtering" (IAIF) y cuyos detalles se explicarán más adelante.

En la siguiente figura se muestra este proceso explicado para poder obtener la excitación a través de la señal de voz mediante la función de transferencia del tracto vocal, tanto en tiempo como en frecuencia:

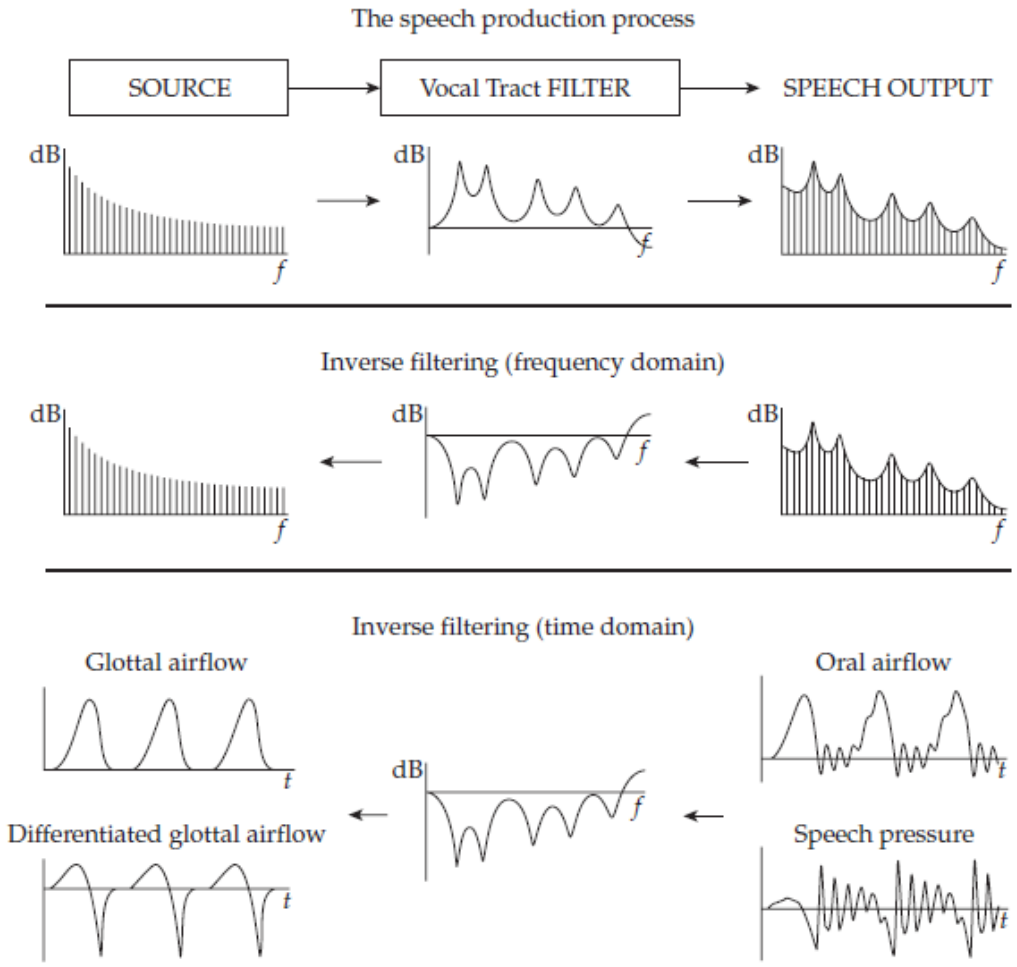


Figura 2.5 Representación esquemática del filtrado inverso para obtener el pulso glotal en el dominio temporal y frecuencial [21]

2.1.3 Modelo Liljencrants-Fant (LF)

Para poder estudiar el proceso fonatorio, se han desarrollado diferentes modelos que intentan representar cómo se lleva a cabo la excitación y se ha ido incrementando el número de parámetros que pueden variar para así adaptarse a los diferentes pulsos glotales que caracterizan a cada locutor. El modelo más popular actualmente, es el Liljencrants-Fant (LF), el cual tiene control sobre la frecuencia fundamental y 4 parámetros más sobre el pulso glotal.

Para hacer una explicación más precisa de los diferentes parámetros de este modelo, recurriremos a la Figura 2.6. En la parte superior se muestran dos pulsos glotales de este modelo, mientras que la figura inferior se corresponde con su derivada. Las expresiones matemáticas que permiten obtener estas señales, y que se muestran en la misma figura, proporcionan los valores de la derivada, y mediante integración podemos obtener los pulsos glotales.

El modelo está formado por dos segmentos determinados por las ecuaciones presentes en la figura. El primero de ellos es sinusoidal con amplitud creciente desde el instante de la apertura de la glotis t_0, hasta el instante de la excitación principal t_e. La forma de este segmento viene dada por 3 parámetros: $w_g=2\pi F$, siendo F la frecuencia del seno; α, que determina la tasa de aumento de la amplitud; y E_0, que es un factor de escala.

El segundo segmento es una función exponencial que modela el flujo desde el instante de máxima excitación hasta el momento que se produce el cierre glotal, t_c. Esta parte del periodo glotal es denominada fase de retorno, y determina un flujo residual tras la máxima excitación. En el modelo LF se controla esta fase mediante el parámetro TA, una medida de la duración efectiva de la fase de retorno.

Con este modelo se asume que los instantes de cierre y de apertura son el mismo, es decir: $t_o=t_c$, por lo que se perdería la fase en la que la glotis permanece cerrada. Además de los 4 parámetros de variación mencionados, se establece que el área negativa de la derivada del pulso glotal debe ser igual a la positiva, lo que implicaría conservar las características principales entre dos pulsos consecutivos.

Con este modelo de la excitación, es posible tomar medidas específicas sobre las formas de ondas glóticas. Algunos de estos parámetros son los siguientes:

Frecuencia fundamental. Se define como $1/T_0$, siendo T_0 la duración del periodo glotal, definido por el tiempo entre los instantes de máxima excitación de dos pulsos consecutivos.

Excitation strength, EE. Amplitud negativa de la excitación principal, que ocurre en el instante de máxima discontinuidad del pulso glotal. Suele calcularse como el mínimo valor negativo de la derivada del pulso glotal en cada ciclo.

Dynamic leakage, RA. Es el flujo residual durante la fase de retorno, que tiene lugar desde el instante de la excitación hasta el cierre glotal. Mide la duración efectiva de la fase de retorno, TA, normalizada al periodo fundamental.

Open quotient, OQ. Es la proporción de tiempo en la que la glotis permanece abierta. Otro parámetro relacionado con éste, es UP, el pico del pulso glotal como se ve en la Figura 2.6.

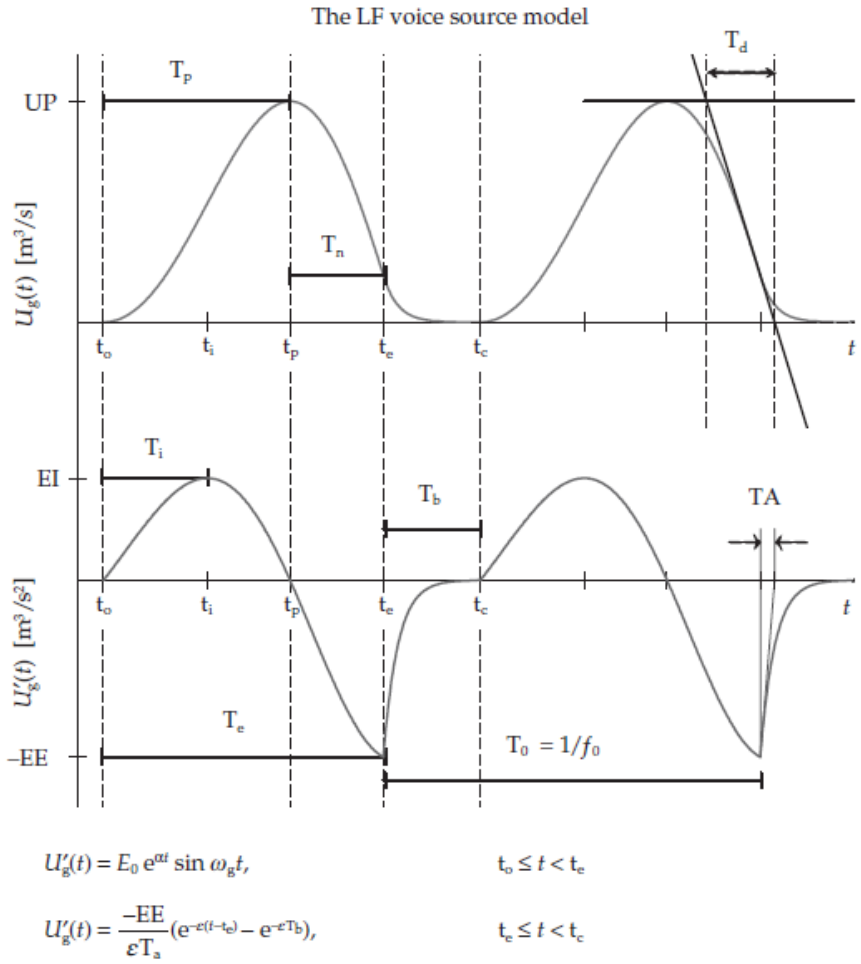


Figura 2.6 Pulsos glotales y su correspondiente derivada extraídos a partir del modelo LF [21]

Además de estos y otros parámetros, también es importante la estabilidad pulso a pulso para determinar la cualidad vocal. Para ello se han empleado medidas como el jitter, que determina la variación de f_0, y el brillo, que equivale a las fluctuaciones en amplitud entre pulsos.

Todas las características mencionadas hasta ahora, eran relativas al dominio temporal del pulso glotal, pero también es posible extraer características a partir de su espectro. Una de las más comunes es el cálculo de la pendiente de este espectro.

También se pueden llevar a cabo medidas sobre el espectro de la señal de voz. Una característica comúnmente usada es la comparación de la amplitud del primer armónico con el nivel de otra componente frecuencial. Estas técnicas son fácilmente aplicables, pero hay que tener en cuenta que en este espectro también se verán reflejados factores ajenos a la excitación. [21]

2.2 Extracción de características con COVAREP (Collaborative voice analysis repository for speech technologies)

Tradicionalmente el cálculo de características relativas a la laringe, como puede ser el pulso glotal, no podían llevarse a cabo de forma eficiente a través de únicamente la señal de voz. Los algoritmos de procesamiento de señales de voz están en continuo desarrollo, pero en muchas ocasiones a costa de alta complejidad. Hay muchos métodos de cálculo diferentes para una misma tarea, y muchos de ellos basados en conceptos y procesos complicados. Para facilitar el acceso a las nuevas técnicas de procesado de voz, varios laboratorios han desarrollado de forma colaborativa un conjunto de códigos abiertos, los cuales la comunidad científica puede usar y corregir si fuera necesario. Este repositorio recibe el nombre de COVAREP y ha sido la herramienta empleada para poder extraer las diferentes características definidas en las secciones anteriores, directamente desde la señal de voz mediante los algoritmos que se irán explicando a lo largo de esta sección. [7]

2.2.1 Detección de voz y estimación de frecuencia fundamental:

La frecuencia fundamental es el primer componente armónico de los sonidos sonoros, y se corresponde con la frecuencia de vibración de las cuerdas vocales. Es un valor que puede ser característico de cada individuo, aunque no se mantiene fijo a lo largo del tiempo y está relacionado con factores como el tamaño de las cuerdas vocales: para frecuencias fundamentales altas son más estrechas y largas. [1]

En la siguiente figura podemos observar 100 milisegundos de un sonido sonoro real, tanto en el dominio del tiempo como en el frecuencial:

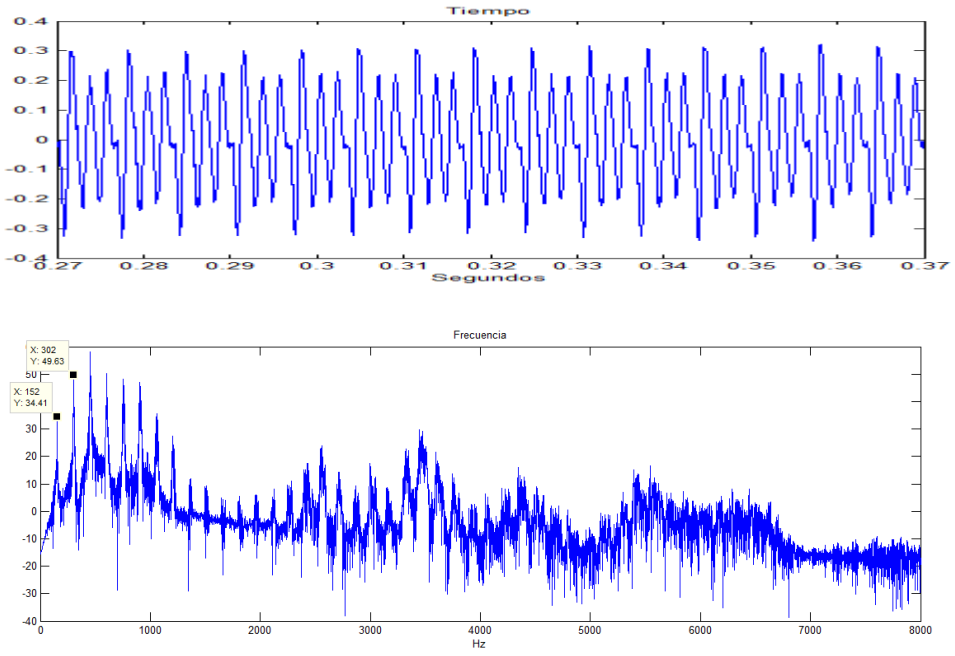


Figura 2.7. Señal de voz en el tiempo y en la frecuencia

En la primera de las gráficas se puede observar el carácter quasi-periódico de la señal de voz a corto plazo. En la segunda, observamos los armónicos, que al definirse como múltiplos de la frecuencia fundamental nos indican el valor de ésta, y por tanto la frecuencia fundamental de vibración de las cuerdas vocales. También se puede apreciar la envolvente espectral, causada por el sonido concreto que se está reproduciendo.

En la investigación acerca de las características de la voz, se han desarrollado varios métodos para estimar la frecuencia fundamental de señales de voz sonora, tanto en el dominio temporal como en el espectral. Debido a que las señales de voz son quasi-periódicas, en lugar de ser estacionarias y periódicas puras, no todos los métodos de estimación obtienen los mismos resultados. En el caso de COVAREP, se ha empleado el método basado en los armónicos del espectro de la señal residual del proceso LPC (Linear predictive coding). Para poder comprender mejor esta técnica, se explicará brevemente qué es la señal residual.

2.2.1.1 Señal residual

Denominamos señal residual a la señal de error obtenida al predecir la señal de voz empleando un sistema de predicción lineal de un determinado orden. [8]

Mediante el análisis LPC, se pueden obtener los coeficientes de un filtro IIR que se corresponde con la envolvente del espectro de la señal. Esta envolvente se adaptará mejor o peor en función del número de coeficientes de predicción lineal, y por lo tanto del orden del filtro, lo cual se puede observar en la siguiente figura:

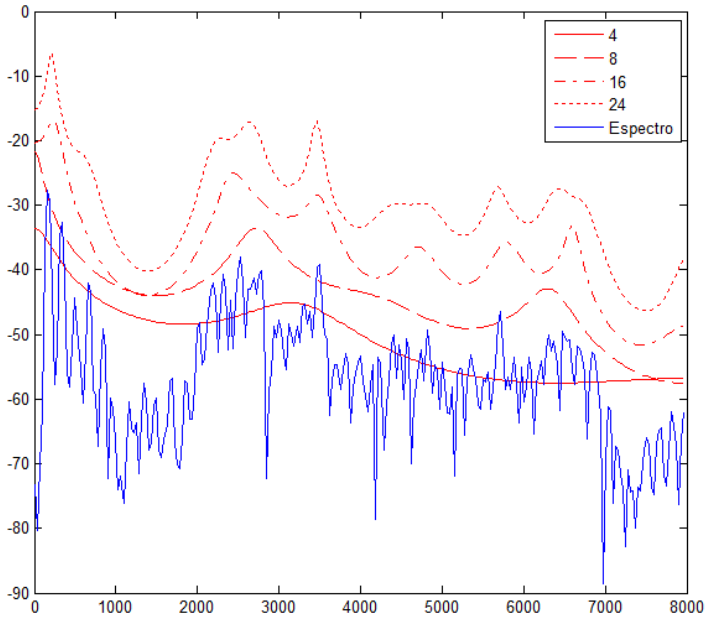


Figura 2.8. Espectro de señal de voz y diferentes envolventes espectrales en función del orden LPC

En la figura se pueden observar tanto el espectro, como envolventes de diferente orden. Para evitar el solape de estas curvas, para órdenes mayores, se han representado aumentando sus valores en todo el rango de frecuencia.

A medida que aumentamos el orden, el grado de detalle es mejor ya que se aumenta el número de polos del filtro, y por lo tanto se incrementan los máximos que presenta la envolvente.

Una vez se ha calculado la envolvente espectral, se obtiene el error de predicción (señal residual) restando a la señal original la que se ha predicho mediante el análisis LPC. El espectro del error es de aspecto blanqueado con valores similares en todo el ancho de banda. En la siguiente figura se muestran algunas de las señales comentadas:

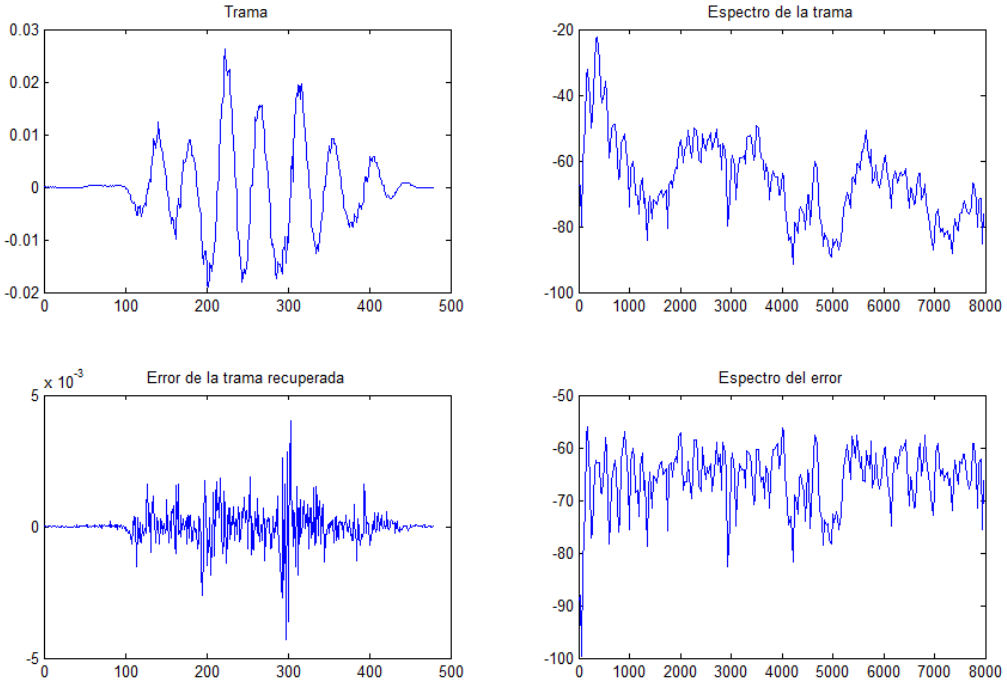


Figura 2.9. Representación temporal y espectral de una señal de voz real y de la señal residual

2.2.1.2 SRH (Summation of Residual Harmonics)

Gracias al blanqueado que se consigue al trabajar con el error de predicción, se elimina los efectos de las resonancias del tracto. El sumatorio de los armónicos residuales se calcula de la siguiente forma:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} \left[E(k \cdot f) - E\left(\left(k - \frac{1}{2} \right) \cdot f \right) \right]$$

(2.1)

Donde *E(f)* es el espectro de la señal residual. Esta expresión es evaluada para un rango de frecuencias determinado, y se obtendrá un valor máximo en la frecuencia fundamental. El motivo de que esto ocurra es que en las frecuencias múltiplo de la fundamental, denominadas armónicos, encontramos máximos locales, de modo que el primer término del sumatorio toma valores más altos para la frecuencia fundamental y los armónicos. El segundo término tiene la finalidad de atenuar el valor de la función para las frecuencias de los armónicos pares, ya que en estos casos la expresión $\left(k - \frac{1}{2} \right) \cdot f$ será un múltiplo de la frecuencia fundamental y el valor a restar será el nivel espectral de un armónico. [6]

En la siguiente figura se puede observar el valor de la función SRH para un tramo de sonido sonoro en un rango de frecuencias de entre 80 y 400 Hz:

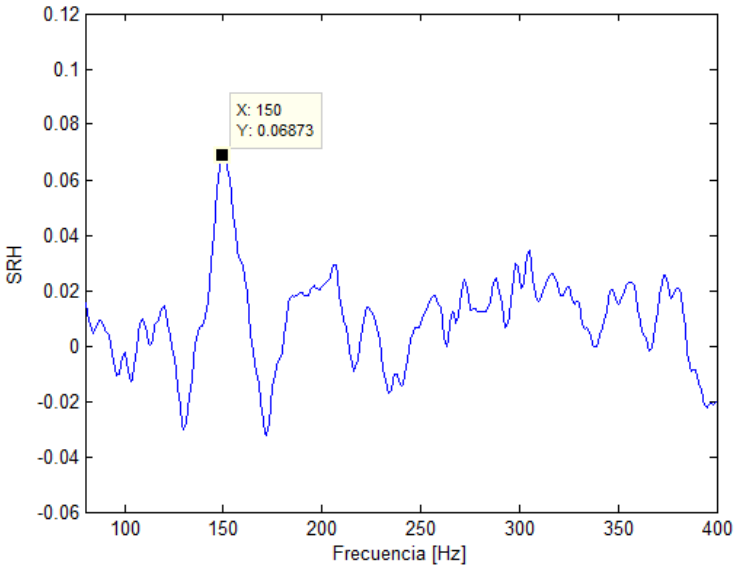


Figura 2.10. Función SRH en un instante determinado de un sonido sonoro

Se puede observar claramente el máximo en 150 Hz, lo que según lo explicado anteriormente, nos indica que la frecuencia fundamental o pitch se encuentra en ese valor para ese instante.

Por otro lado, también se muestra la evolución del SRH a lo largo del tiempo para la misma locución:

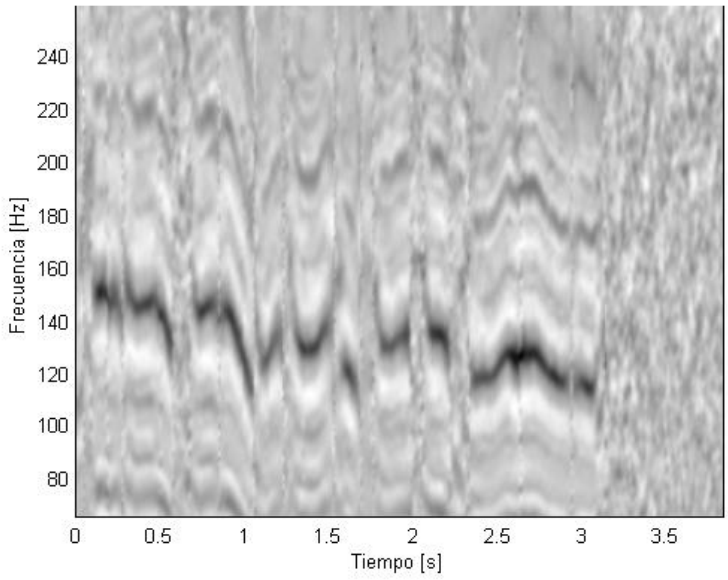


Figura 2.11. Evolución temporal del SRH en una locución

En la gráfica se representa para cada instante los valores de SRH para un rango de frecuencias determinado. De modo que en un momento dado, los valores que presentan mayor SRH, tendrán valores más oscuros.

Se puede seguir con facilidad el pitch de esta trama de audio, que se encuentra siempre en valores entre 120 y 160 Hz.

La misma función, también puede ser empleada para determinar la presencia o ausencia de voz sonora. Para ello, se establece un determinado umbral, y en caso que el SRH para la supuesta frecuencia fundamental obtenida lo supere, se considerará que el tramo contiene voz sonora. Es importante recalcar que será necesario normalizar en energía el espectro de la señal residual para cada trama.

En la siguiente figura se pueden observar algunos de los detalles comentados tras aplicar este algoritmo a un audio de la base de datos TIMIT, la cual ofrece audios de corta duración, con poco ruido y con etiquetas tanto a nivel de fonema como de palabras. De esta forma, podemos comprobar por ejemplo las diferencias entre tramos de voz sorda y sonora, estando seguros de la naturaleza de los sonidos.

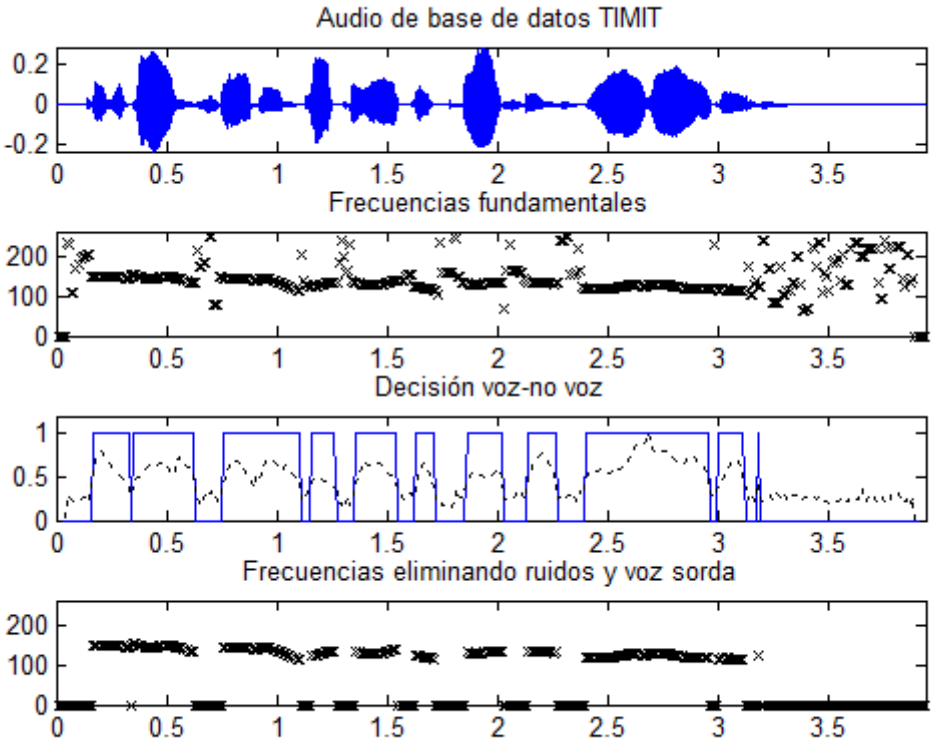


Figura 2.12. Señal de voz. Frecuencias fundamentales y decisión voz/no voz calculado mediante SRH.

La primera gráfica representa la grabación de un audio a lo largo del tiempo. El audio pertenece a la base de datos TIMIT y en él se dice lo siguiente:

"Kindergarten children decorate their classrooms for all holidays".

En la segunda gráfica se presenta una estimación de la frecuencia fundamental en cada instante de cálculo mediante el algoritmo SRH. Estos valores han sido calculados a lo largo de toda la duración del archivo, con independencia del contenido. Por este motivo, hay tramos donde el pitch presenta gran inestabilidad, con valores muy por encima o por debajo de lo que podría considerarse como normal. Esta irregularidad es porque pertenecen a segmentos de voz sorda o a silencios.

La siguiente representación muestra la decisión voz/no voz a lo largo del tiempo determinada por los algoritmos proporcionados por COVAREP, y el valor de la función SRH para la frecuencia fundamental del instante en cuestión. Esta decisión no solo discrimina los silencios sino que además también los fragmentos que contienen voz sorda. Gracias a ello podremos trabajar únicamente con los segmentos que son de interés.

En la última gráfica se aplica la decisión voz/no voz a la estimación de pitch mediante SRH, a modo de máscara: todos los instantes estimados como ruido o sonidos sordos, se les otorga valor 0 de frecuencia fundamental. Es interesante comparar las estimaciones de pitch antes y después de aplicar la "máscara": en el segundo caso desaparecen las irregularidades, y el pitch siempre se mantiene en valores cercanos.

2.2.2 Estimación del pulso glotal

De acuerdo con lo explicado anteriormente, los sonidos sonoros podrían ser modelados como la salida de un filtro donde el tracto vocal filtra la señal proveniente de la glotis y finalmente es radiada al exterior por los labios. Se ha demostrado que esta aproximación, a pesar de algunas limitaciones, es de gran utilidad para numerosas aplicaciones del procesado de voz. [17]

El principal objetivo de estimar la onda glotal es poder separar las señales de voz en dos componentes. Esta tarea no resulta trivial, por lo que se han desarrollado varios métodos para obtener estimaciones cada vez más precisas y correctas.

En este trabajo se ha empleado la técnica de "Glottal inverse filtering", cuyo objetivo es eliminar la influencia del tracto vocal para obtener el pulso glotal. Hay varias ténicas englobadas en esta categoría, nos centraremos en una de ellas: Iterative Adaptative Inverse Filtering (IAIF).

Este algoritmo parte de dos suposiciones:

- La componente del pulso glotal es la responsable de la pendiente general del espectro de una señal de voz
- La onda del flujo glotal se puede representar por un filtro con solo polos de orden bajo.

El proceso consiste en calcular estimaciones cada vez más precisas de las influencias tanto del tracto vocal como de la glotis en la señal de voz, mediante filtrados inversos. Se siguen los siguientes pasos:

Se hace un filtrado inverso de señal de voz con el filtro resultante en el análisis LPC de primer orden, para eliminar de la señal de voz el efecto del tracto glotal más notable: la atenuación de la señal de voz en las altas frecuencias.

Una vez se ha eliminado el efecto del tracto glotal de la señal de voz, se estima el efecto del tracto vocal mediante un filtro de predicción lineal de orden 20, y es eliminado de la señal de voz nuevamente mediante un filtrado inverso, lo que da lugar a una primera estimación del pulso glotal.

Estas dos primeras estimaciones de los efectos tanto del tracto vocal como del tracto glotal, son refinadas mediante el uso nuevamente del filtrado inverso. El orden de los filtros LPC empleados en estos dos últimos pasos, es de 8 para estimar el efecto del tracto glotal, y de 20 para determinar el efecto del tracto vocal.

Las 4 estimaciones mencionadas se muestran en la siguiente figura:

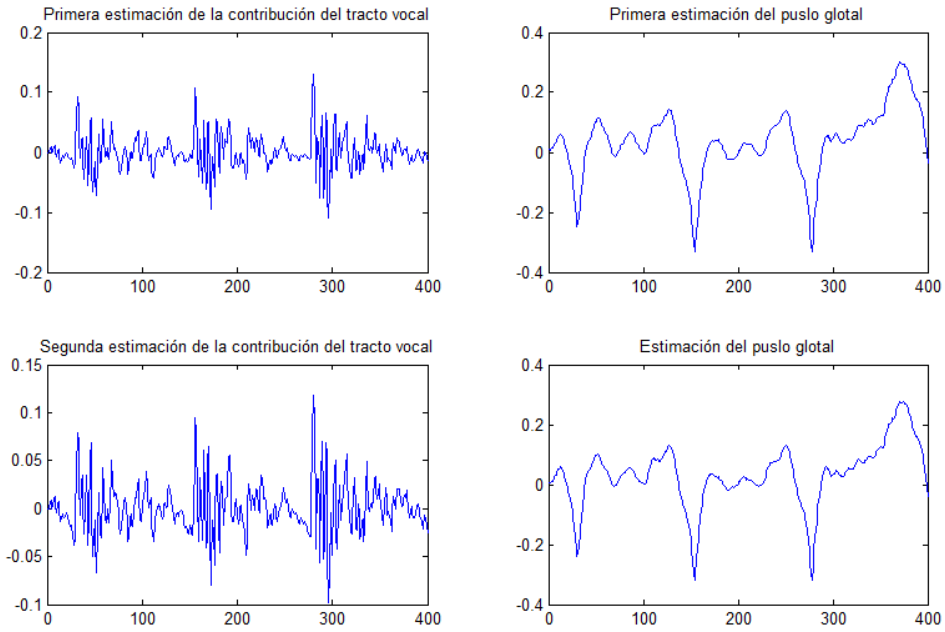


Figura 2.13 Estimaciones de la contribución en la señal de voz tanto del tracto vocal como la de glotis, para obtener el pulso glotal mediante el algoritmo IAIF

Para conseguir el pulso glotal de una señal de voz, ésta es enventanada y se lleva a cabo el proceso comentado anteriormente en cada una de las ventanas. Gracias a la técnica de solapamiento y suma se pueden obtener todos los pulsos de una locución.

En la siguiente figura se muestra tanto el pulso glotal como la derivada de éste de un sonido sonoro de 60 milisegundos:

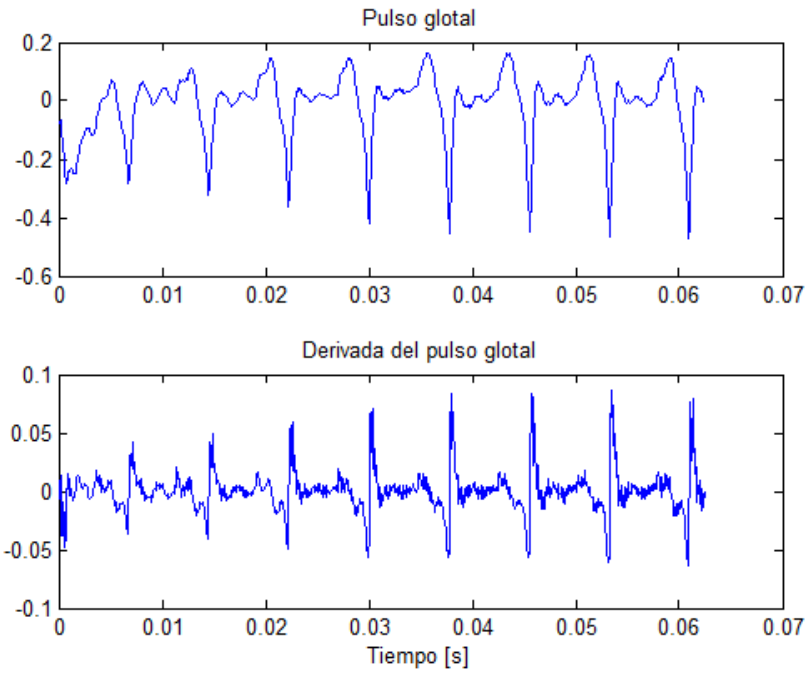


Figura 2.14. Pulso glotal y su derivada

2.2.3 Detección de GCIs y GOIs (Glottal Closure and Opening Instants):

Ya se ha explicado anteriormente que para la producción de sonidos sonoros, las cuerdas vocales llevan a cabo movimientos repetitivos de apertura y cierre. En esta sección nos centraremos en explicar cómo es posible detectar estos instantes, tanto de apertura como de cierre, en una grabación de voz. [3]

En [22] los autores defienden que hay discontinuidades en la excitación en todo el ancho de banda espectral, incluyendo en frecuencia 0. Por este motivo, se hizo un análisis en la señal promedio calculada de la siguiente forma:

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m)$$

(2.2)

Donde $s(n)$ es la señal de voz y $w(m)$ una función de enventanado.

Esta señal no puede proporcionar con exactitud los instantes de apertura ni cierre de la glotis, pero tras diversas observaciones, se definen regiones temporales donde es probable que van a encontrarse dichos acontecimientos:

- Los GCIs tendrán lugar entre el mínimo de la señal promedio y el siguiente cruce por cero.
- Los GOIs estarán situados entre los máximos y el cruce por cero hacia valores negativos.

El siguiente paso del método consiste en determinar con mayor exactitud los instantes tanto de apertura como de cierre. Para ello, se empleará la ya mencionada anteriormente señal residual. Es razonable pensar que un cambio tan importante en la excitación se va a ver reflejado en la señal de error de predicción, por lo que para cada intervalo donde se predijo que habría un cierre glotal, se busca el máximo valor en la señal residual y se toma como el GCI de ese ciclo, y del mismo modo con las aperturas.

Todo este proceso es aclarado con las siguientes gráficas:

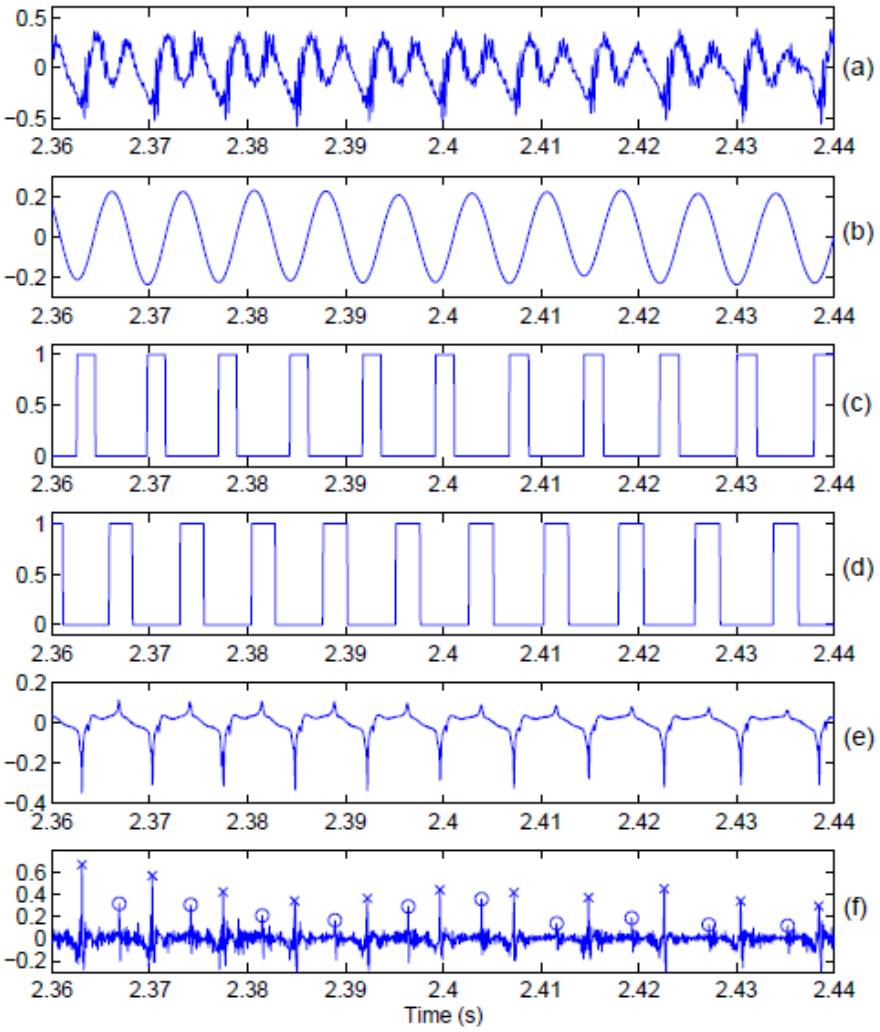


Figura 2.15. Detección de GCI y GOI. (a). Señal de voz. (b). Señal promedio resultante de (2.2). (c). Intervalo de presencia de GCI. (d). Intervalo de presencia de GOI. (e). Resultado obtenido con un electroglotógrafo (f). Señal residual con los GCIs (x) y GOIs (o) detectados en los máximos. [3]

Comparando las dos últimas figuras, se comprueba la bondad de los resultados obtenidos con este sistema de detección de GCIs y GOIs, ya que coinciden los instantes detectados con los máximos y mínimos obtenidos con el electroglotógrafo.

2.2.4 Parámetros

En esta sección se describirán brevemente, algunos de los parámetros empleados para caracterizar a los locutores y cómo gracias a la herramienta empleada podemos calcularlos a partir de la señal de voz. Cabe mencionar, que además de los parámetros explicados en las siguientes líneas, también se utilizarán otros de los que ya se ha hablado anteriormente, como puede ser la frecuencia fundamental.

Paralelamente, se muestran varias gráficas referentes a los pulsos glotales que serán de gran ayuda para conseguir una mejor compresión del modo de cálculo de los diferentes parámetros.

En primer lugar se muestran los pulsos glotales y su derivada de un sonido sonoro, donde cada línea vertical refleja un GCI, por lo que entre dos líneas encontramos un único pulso.

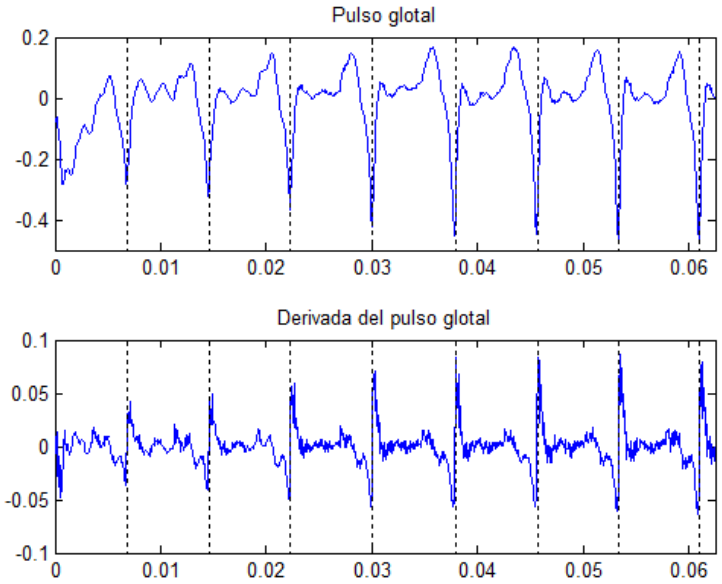


Figura 2.16 Pulso glotal y su derivada, dividiendo ambas gráficas en los diferentes pulsos glotales mediante líneas verticales

NAQ (Normalized Amplitude Quotient): es un parámetro que describe la fase del cierre glotal en la producción de sonidos sonoros, y es una forma de parametrizar este proceso. Se calcula como el cociente entre la máxima amplitud del pulso glotal y la máxima amplitud negativa de su derivada. El resultado es normalizado con el periodo fundamental de la trama. Se ha comprobado que este parámetro es más robusto frente a distorsiones como el ruido en la medición que otros parámetros en el dominio temporal como el "closing quotient". [9]

En la Figura 2.16, para cada uno de los pulsos glotales, hay que determinar el máximo del pulso, y el mínimo valor de su derivada. Se lleva a cabo su cociente, y se normaliza el resultado con el periodo fundamental.

Para cada pulso glotal, se extraería un valor de NAQ, y si aplicamos el algoritmo a una locución entera y representamos todos los valores obtenidos el resultado es el siguiente:

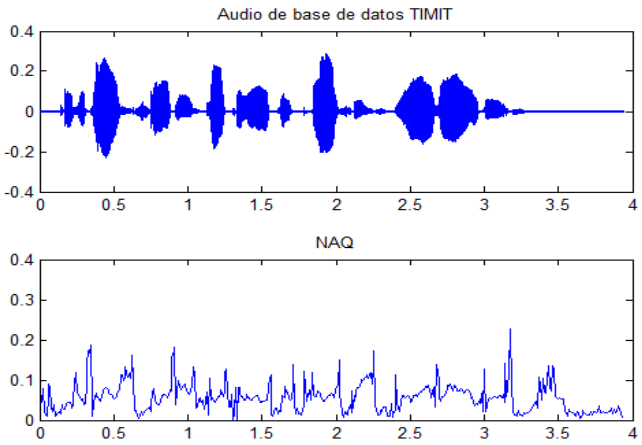


Figura 2.17 Señal de voz y valores de NAQ obtenidos para cada pulso glotal

Quasi-Open Quotient (QOQ): es un parámetro que describe el porcentaje de tiempo en el que la glotis permanece abierta en cada periodo glotal. Se calcula como el tiempo en el que el flujo glotal está por encima del 50% del máximo, normalizado por el periodo fundamental. [15]

Recurriendo nuevamente a la Figura 2.16, para cada pulso habría que determinar el valor máximo del pulso glotal, y calcular la cantidad de tiempo en la que el pulso glotal toma valores mayores al 50% de dicho máximo. Estos tiempos son normalizados por el periodo fundamental, obteniendo así un valor de QOQ para cada pulso.

En la siguiente figura se muestran todos los valores de QOQ obtenidos para una locución:

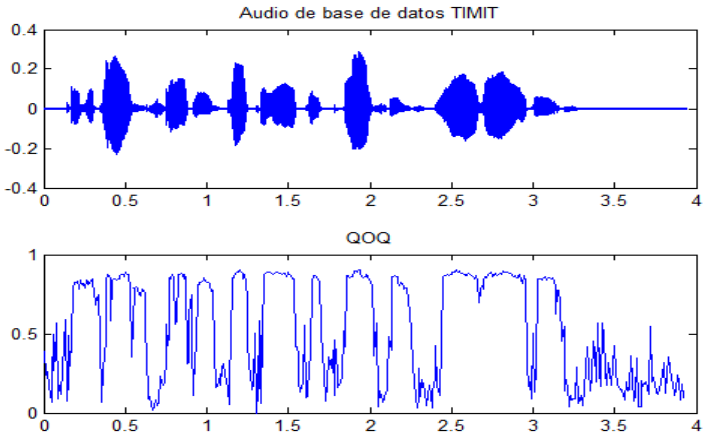


Figura 2.18 Señal de voz y valores de QOQ obtenidos para cada pulso glotal

Como se ha comentado anteriormente en la sección 2.1.3, algunas características son extraídas a partir del espectro de los pulsos glotales. Se muestra a continuación el espectro calculado a partir de un pulso glotal:

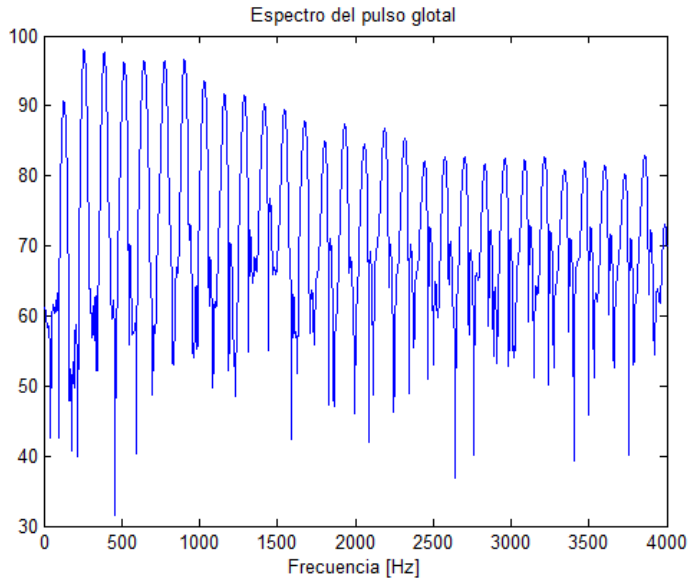


Figura 2.19 Espectro de pulso glotal

H1-H2 ratio (H1-H2): relación entre las amplitudes espectrales del primer armónico, es decir, de la frecuencia fundamental, y el segundo armónico de la señal de voz. [15]

Para su cálculo habrá que determinar el espectro de cada uno de los pulsos glotales de una señal de voz y localizar los dos primeros armónicos. Una vez se conocen sus valores, se resta la amplitud del segundo a la del primero. Cabe mencionar que este valor no siempre va a ser positivo, ya que como se comprueba en el ejemplo de la Figura 2.19, el segundo armónico tiene mayor amplitud.

Si calculamos este parámetro para todos los pulsos glotales se obtiene el siguiente resultado:

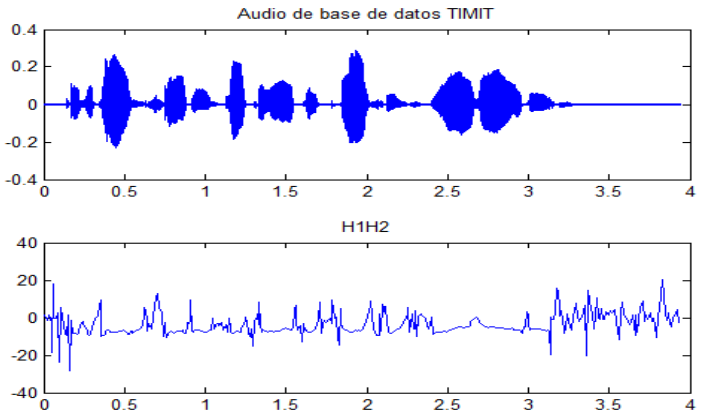


Figura 2.20 Señal de voz y valores de H1-H2 obtenidos para cada pulso glotal

Harmonic Richness Factor (HRF): Cuantifica la cantidad de armónicos en el espectro. Se calcula como el cociente entre la suma de las amplitudes espectrales de los armónicos, y la amplitud de la frecuencia fundamental. [10]

Nuevamente habrá que calcular el espectro de cada pulso glotal y llevar a cabo la detección de todos los armónicos y sus valores de amplitud. Para cada pulso se realizará el sumatorio de los valores de amplitud espectral de todos los armónicos excepto el primero, y se dividirá por la amplitud espectral en la frecuencia fundamental, obteniendo un valor de HRF por cada pulso. Realizando el proceso para todos los pulsos de una locución se obtiene el siguiente resultado:

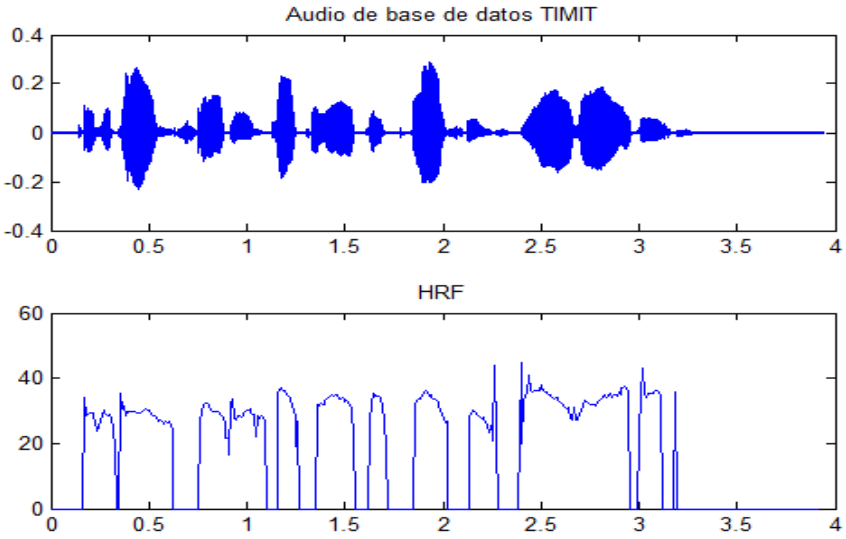


Figura 2.21 Señal de voz y valores de HRF obtenidos para cada pulso glotal

PSP (Parabolic spectral parameter): Es un parámetro que trata de cuantificar la caída espectral en las altas frecuencias del espectro del pulso glotal. Está basado en ajustar una función parabólica a la parte de baja frecuencia del espectro del pulso glotal estimado. Se refleja en un único valor que describe cómo el espectro de un determinado pulso glotal decae en comparación con el límite de caída teórico. [11]

La forma de calcular este parámetro es ajustar una función parabólica definida como $Y(k) = ak^2 + b$, al espectro del pulso glotal. Del mismo modo, se buscará otra función del mismo tipo que se ajuste a una señal de amplitud unidad y de longitud igual al periodo fundamental, que representará el límite de caída teórico. Para obtener el valor de PSP de cada pulso glotal se realiza el cociente entre los coeficientes cuadráticos, a, obtenidos para ambas funciones. En la siguiente figura se representan todos los valores de PSP para una locución:

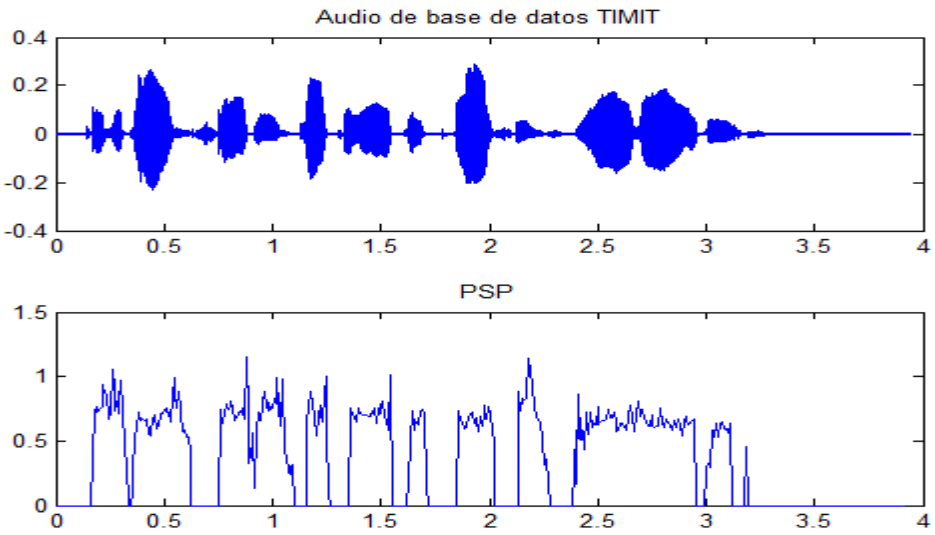


Figura 2.22 Señal de voz y valores de PSP obtenidos para cada pulso glotal

Por último, los parámetros MDQ (Maxima Dispersion Quotient) y Peak Slope, se centran en la diferenciación e identificación de tipos de voz, como podría ser voz "breathy" (aspirada) o "tense" (tensa). No se entrará en más detalle con estos parámetros ya que no han sido de utilidad para este trabajo.

2.3 Sistema de modelado y de identificación de locutores

Una vez ya hemos explicado qué características pueden resultarnos interesantes, y la forma de calcularlas a partir de la señal de voz, en esta sección nos centraremos en definir los métodos que se van a emplear para comprobar la bondad de nuestros parámetros en tareas de reconocimiento de locutor.

Para afrontar tanto el modelado de locutores, como para proporcionar una medida de verosimilitud entre dos locuciones diferentes, se ha empleado un sistema basado en GMMs (Gaussian Mixture Model), un método eficiente computacionalmente muy empleado en el reconocimiento de locutor independiente del texto, ya que no depende de factores temporales.

2.3.1 GMM

Se denomina mezcla de gaussianas, a una suma ponderada de las densidades de componentes gaussianas. Este tipo de modelos han sido utilizados habitualmente como modelos de la distribución de probabilidad en sistemas basados en características biométricas, como es el caso que nos atañe en este trabajo: caracterización de locutor a través de información de cualidad vocal.

Si nuestro modelo está definido por vectores de características de D dimensiones, x, la función de densidad de probabilidad se define como:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i \, p_i(x)$$

(2.3)

Donde M es el número de gaussianas, w_i es el peso de las mezclas, cuya suma debe ser igual a la unidad, y

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'\left(\sum_i\right)^{-1}(x-\mu_i)\right\}$$

(2.4)

Siendo $\sum i$ una matriz de covarianza de $D\,x\,D$ dimensiones, y μ_i un vector $D\,x\,1$ de medias. [5]

En nuestro desarrollo se emplearon matrices de covarianza diagonal por tres motivos principales:

- Más eficientes computacionalmente.
- Se pueden lograr iguales resultados con matrices diagonales de mayor tamaño que empleando matrices de covarianza completa.
- Empíricamente se ha observado que las matrices diagonales de GMM mejorar a las matrices completas.

2.3.2 UBM, Universal Background Model:

Podemos definir un UBM como un gran GMM cuya intención es reflejar todas las posibles alternativas de habla a la hora de enfrentarnos al reconocimiento de locutor. Por ejemplo, en caso de conocer a priori, que en nuestras pruebas de identificación solo van a participar mujeres por medio de conversaciones telefónicas, nuestro modelo UBM tendrá que ser entrenado únicamente con este tipo de grabaciones.

En el sistema GMM-UBM, se obtiene el modelo de cada locutor adaptando los parámetros del UBM con las locuciones de entrenamiento. Este proceso consigue modelos más independientes unos de otros, y por otro lado se consigue una técnica de puntuación más rápida.

En la siguiente figura, se puede observar esta adaptación del UBM para cada locutor:

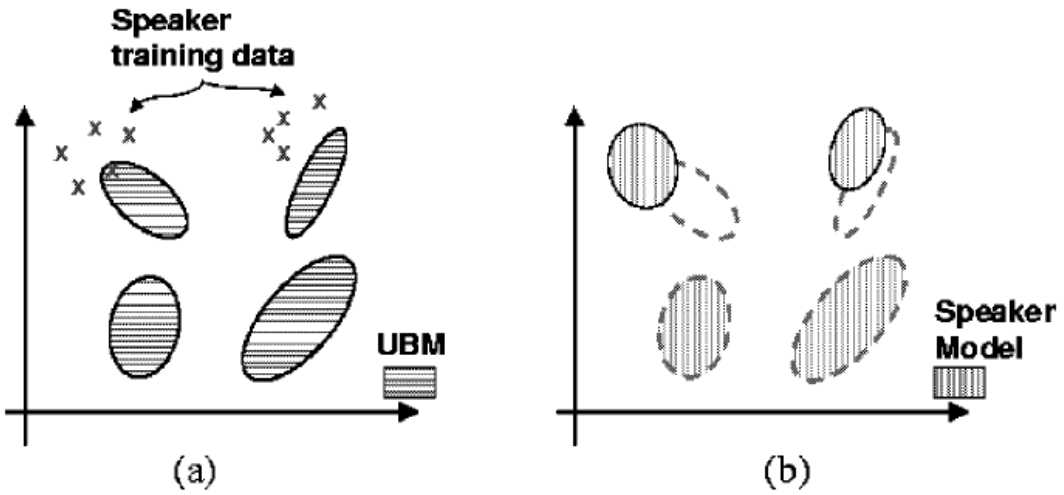


Figura 2.23 Adaptación del modelo de locutor en un sistema GMM-UBM. (a) Los datos de entrenamiento son asignados a las mezclas del UBM. (b) El modelo final del locutor se calcula variando los datos del UBM entrenado [5]

En nuestro caso, el UBM es entrenado empleando el algoritmo EM (expectation-maximization). Se trata de un algoritmo iterativo que incrementa el ajuste del modelo a los vectores de entrenamiento, de modo que la probabilidad de pertenencia de un vector dado al modelo elaborado, aumenta de una iteración a la siguiente:

$$p\left(X \left| \lambda^{(k+1)} \right.\right) > p(X|\lambda^k),$$

(2.5)

Donde X sería el vector de entrenamiento, y λ^k el modelo entrenado en la interacción k. Por lo general, cinco iteraciones bastan para conseguir la convergencia.

Por otro lado, para adaptar cada uno de los modelos de locutor con nuestro UBM previamente entrenado se recurre a la técnica MAP (maximum a posteriori). El mismo algoritmo se emplea a la hora del reconocimiento de locutores, donde la probabilidad que tiene una locución, con vector de características *X,* de pertenecer a un determinado locutor hipotético, se calcula como:

$$\Lambda(X) = \log p(X|\lambda_{hip}) - \log p(X|\lambda_{UBM})$$

(2.6)

Siendo λ_{hip} el modelo del locutor ya adaptado, y λ_{UBM} el modelo del UBM. Este valor de probabilidad, será el denominado como "score", y nos permitirá tomar la decisión en el reconocimiento.

Esta forma de calcular medidas de verosimilitud, en lugar de emplear únicamente modelos entrenados, se debe principalmente a dos causas:

1. Cuando un GMM se avalúa para un vector de características, sólo algunas mezclas contribuyen al valor probabilístico final.
2. Los componentes de un modelo, guardan relación con las mezclas del UBM, por lo que esos vectores cercanos a una determinada mezcla, también estarán relacionados con la mezcla del modelo correspondiente al locutor

2.3.3 Z-norm:

Es una técnica de normalización de los scores que se han obtenido en las pruebas de verificación, que permite establecer un umbral de decisión independiente del locutor, ya que alinea la distribución de puntuaciones de locutores impostores para conseguir que tenga media 0 y desviación típica uno:

$$y_j^Z = \frac{y - \mu_j^I}{\sigma_j^I}$$

(2.7)

Donde y es el score obtenido, y σ y μ la media y la desviación de los scores de impostor respectivamente. Los súper índices I indican que se tratan de valores estadísticos obtenidos a partir de scores de impostor, mientras que el súper índice Z indica que en esos scores ya ha sido aplicado z-norm. [16]

La forma de obtener estas medidas estadísticas de scores non-target es enfrentar el modelo de locutor en cuestión a un conjunto de modelos los cuales se conoce a priori que no pertenecen al mismo hablante. Esto se puede llevar a cabo con los archivos de entrenamiento.

2.4 Agrupación de locutores

Otro de los objetivos del trabajo, es comprobar si las cualidades vocales podrían servir para establecer una clasificación de tipos de locutores.

La clasificación en diferentes grupos de locutores con propiedades semejantes se llevaría a cabo mediante técnicas de clustering. Dichos métodos consiguen identificar grupos y establecer una clasificación en un conjunto de datos.[2] Han sido en diferentes áreas de investigación con tres propósitos principales:

Analizar la estructura de los datos.

Relacionar aspectos de los datos entre sí.

Establecer una clasificación.

A la hora de conseguir una agrupación adecuada de los locutores se escogió el método de clustering aglomerativo.

2.4.1 Clustering aglomerativo

El clustering aglomerativo, o clustering hacia arriba, parte de la idea de considerar a cada uno de los locutores como un único clúster, definido por sus características, y sigue los siguientes pasos:

1. Se determinan los dos clusters que presentan mayor similitud, y se unen en una única agrupación.
2. Se repite el paso anterior hasta que todos los locutores hayan sido agrupados en dos clusters, o se llegue a otro criterio de parada

El criterio de parada de esta técnica puede deberse a dos motivos: Cuando la distancia entre todos clusters es demasiado elevada como para que se produzcan más agrupaciones de forma efectiva, o cuando se llega a un número de clusters definido al principio del algoritmo.

Podemos definir la distancia entre dos clusters de 4 formas:

- Single Linkage: Mínima distancia entre dos componentes cualquiera de los dos clusters.
- Complete Linkage: Máxima distancia entre cualquier componente de ambos clusters.
- Average Linkage: Media de la distancia entre todos los componentes de ambas clases.
- Centroid Method: Distancia entre los dos centroides.

Con esta técnica, el sumatorio de las distancia de los locutores al centroide el que pertenecen, irá aumentando a medida que aumentamos el número de clases, ya que partimos de distancia cero.

3 Diseño y desarrollo

3.1 Base de datos empleada

Para realizar las pruebas, y aplicar los conocimientos explicados en la sección anterior, se ha recurrido a bases de datos de NIST-SRE (US National Institute of Standards and Technology – Speaker Recongnition Evaluations). Durante los últimos años este instituto ha llevado a cabo evaluaciones de reconocimiento de locutor independiente de texto que han servido para comprobar y compartir con el resto de comunidad científica, los avances en este campo.

Concretamente, se empleará la base de datos con locuciones de la evaluación del año 2006. Esta base de datos está formada por grabaciones de conversaciones telefónicas, en su mayoría de habla inglesa. En este trabajo únicamente se han empleado locuciones de habla inglesa. [13] [12]

En nuestras pruebas se emplearán dos conjuntos de locutores, una con 3 locuciones de 6 locutores diferentes y la otra con 5 locuciones de 30 locutores diferentes. En ambos casos con mismo número de hombres que de mujeres:

Conjunto	Nº Locuciones	Nº Locutores	Locuciones por locutor	% Mujeres	% Hombres
Reducido	18	6	3	50	50
Extendido	150	30	5	50	50

Tabla 1 Especificación de los conjuntos de locutores empleados en las pruebas

La primera labor para poder llevar a cabo los experimentos fue adaptarse a los audios de la base de datos que se iba a utilizar. Las locuciones tienen una duración aproximada de 5 minutos. Cada una de ellas contiene una conversación entre dos locutores diferentes siendo únicamente posible escuchar a uno de ellos. Por este motivo, será fundamental poder decidir para cada muestra del audio si contiene voz o no. Gracias a esto podremos analizar únicamente segmentos que no contengan silencios.

Además de eliminar los silencios, también nos interesará eliminar sonidos sordos, ya que en ellos parámetros como el NAQ, relacionado con el cierre glotal, no tendrían validez. Para hacer esta segmentación, se recurre al método de SRH (Summation of Residual Harmonics), que como se ha explicado, permite hacer una estimación del pitch, y determinar la presencia de sonidos sonoros gracias al sumatorio de los armónicos de la señal residual.

En la figura siguiente se muestra un histograma de la proporción de tiempo donde encontramos voz sonora frente a la duración total de cada locución, en el conjunto de locutores Extendido:

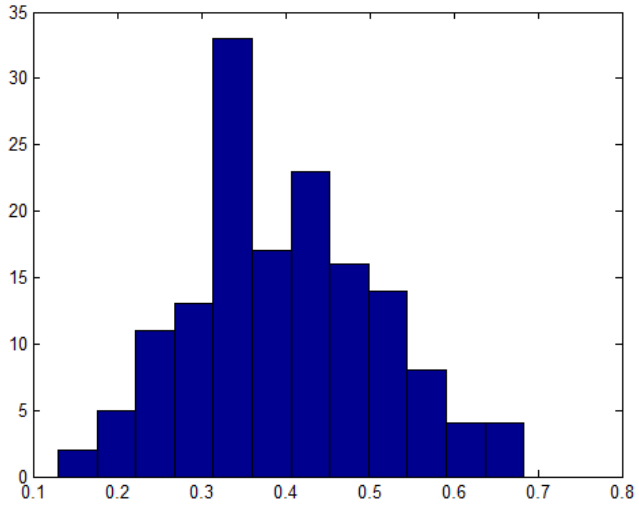


Figura 3.1 Histograma del porcentaje de tiempo hablado en las locuciones del conjunto Extendido

Como era de esperar, gran parte de los audios presentan algo menos de la mitad del tiempo voz sonora. Esto se debe a que solo contienen el audio perteneciente a uno de los dos locutores participantes, y dentro del tiempo en el que se habla también encontramos silencios y tramos con voz sorda.

También es interesante comprobar cómo el porcentaje de tiempo hablando es independiente del locutor. Para ello, se representa el porcentaje de duración de cada una de las 150 locuciones agrupadas por locutor, de modo que cada región comprendida entre dos líneas verticales corresponde a un único hablante. En cada una de estas zonas se representan 5 cruces que indican el porcentaje de tiempo hablando de cada una de las locuciones por locutor de las que dispone la base de datos. Por lo general es un valor muy dispar entre todas las locuciones de un mismo hablante:

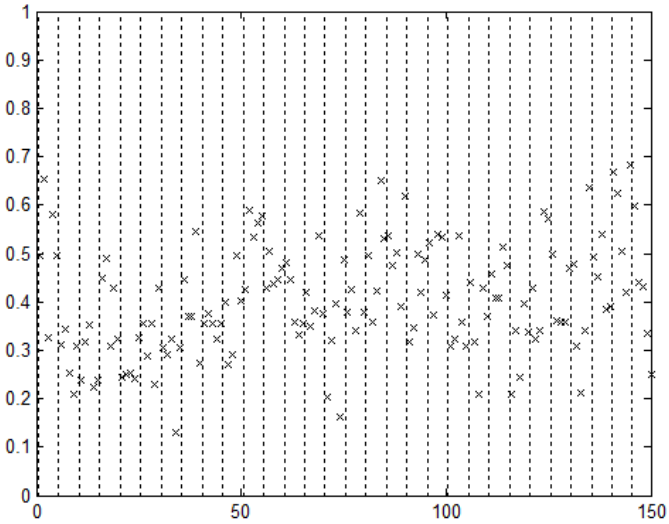


Figura 3.2 Porcentaje del tiempo hablado en las locuciones del conjunto Extendido, donde las líneas verticales separan a diferentes locutores, y cada cruz es el porcentaje de una locución determinada.

3.2 Parametrización

Para poder obtener parámetros, los segmentos de voz sonora deben tener una duración mínima de 150 milisegundos, ya que algunos algoritmos empleados requieren duraciones mínimas para poder obtener resultados adecuados.

Una vez que se tienen bien localizados los segmentos con voz en cada uno de los audios, se lleva a cabo un enventanado con solape de 2/3 con una ventana de tipo Hanning de 150 milisegundos, y por lo tanto desplazamiento de 50 ms. Para cada una de las ventanas se llevará a cabo una extracción de parámetros independiente de la del resto de ventanas. La siguiente figura representa esta segmentación de forma visual:

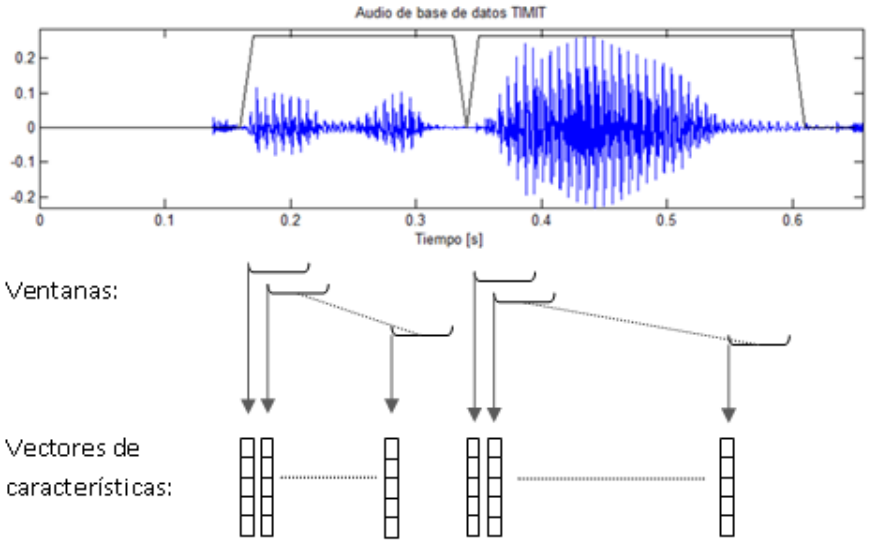


Figura 3.3 Señal de voz en el dominio temporal junto con la decisión voz/no voz y proceso de parametrización de dicha señal

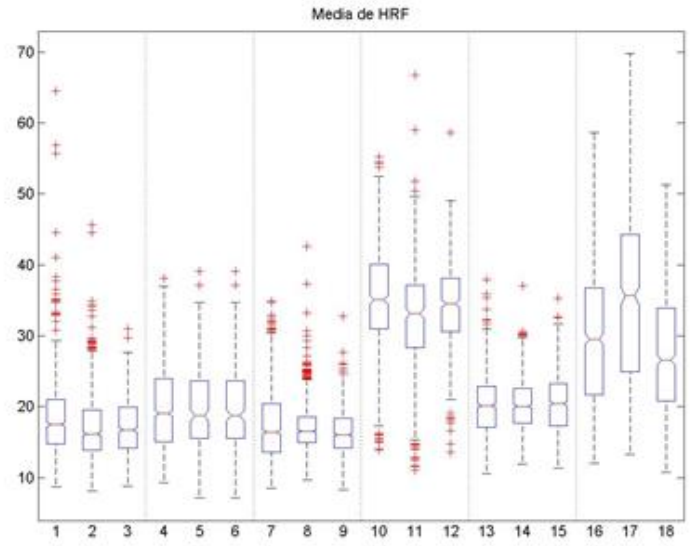
Es importante recalcar que este proceso es diferente al empleado en otros sistemas, en los que sin hacer ningún estudio previo de la señal de voz, se extraen vectores de características de forma continua. En nuestro caso, como se ha comentado anteriormente, primero será esencial localizar los tramos donde haya sonidos sonoros, y de esta forma extraer vectores de características únicamente en estas zonas. En la figura se puede comprobar que en los instantes donde no se tiene voz sonora no se extraen vectores de características.

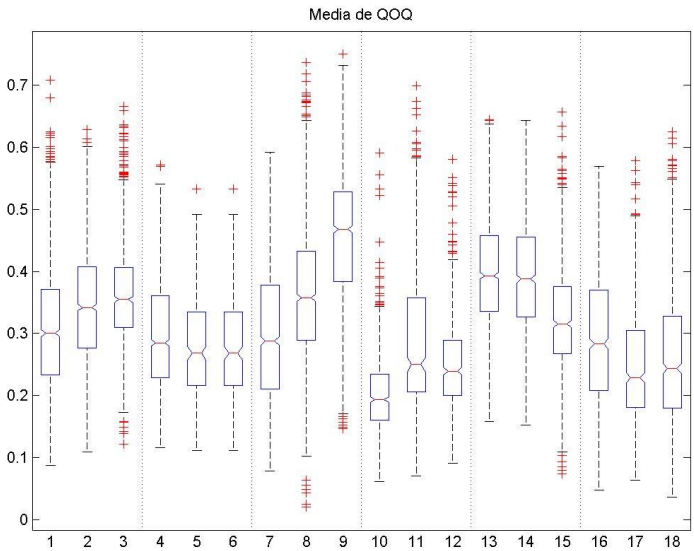
Para todos los parámetros relacionados con el pulso glotal, en cada ventana se obtendrán tantos valores como periodos glotales tengan lugar en ese intervalo de tiempo, un valor que no será fijo en diferentes ventanas. Para poder obtener un único valor por cada característica y ventana, se aplicarán operaciones estadísticas, como la media y la desviación típica, a los valores obtenidos de todos los parámetros en cada ventana. De modo que un parámetro queda definido en cada uno de nuestros vectores por su media y/o su desviación típica.

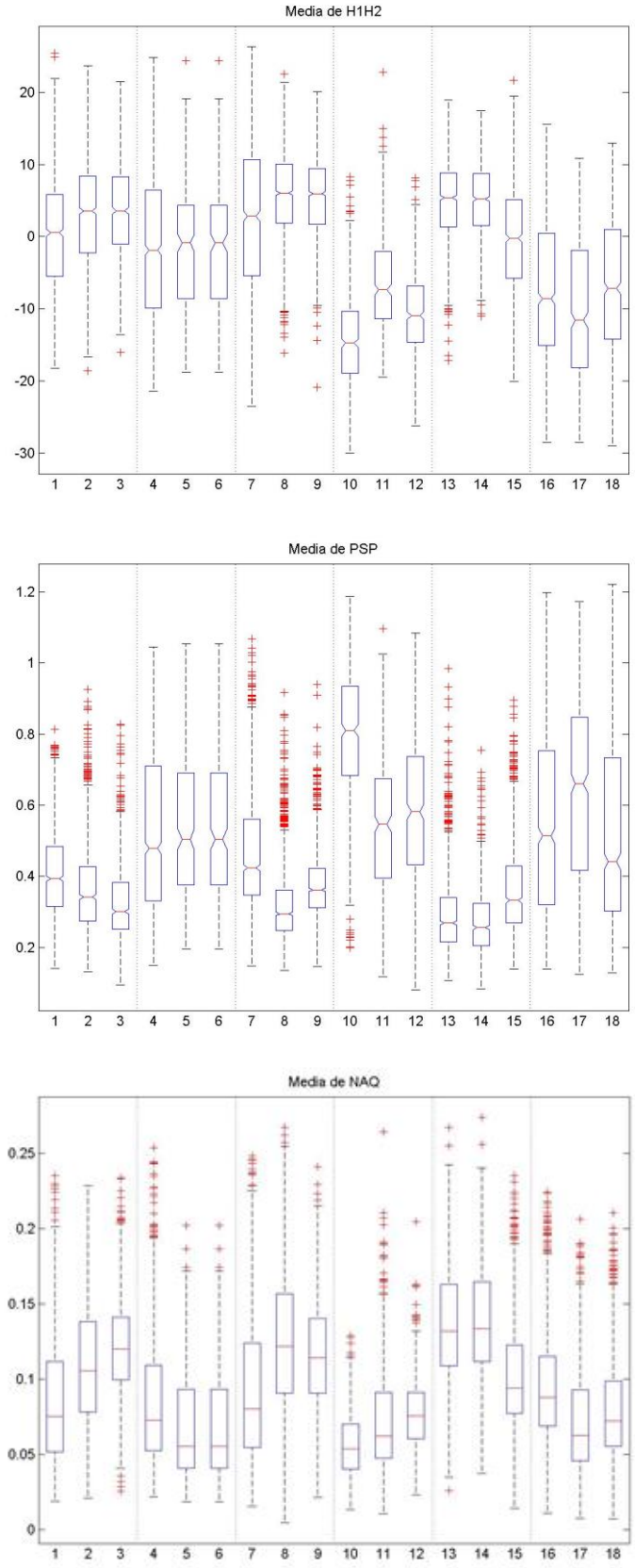
3.2.1 Características

Los parámetros que caractericen a los hablantes serán de interés siempre que se mantengan estables para un mismo locutor y sean diferentes a los presentes en otros hablantes. De este modo se conseguirá caracterizar a cada uno de los locutores de forma eficaz. Es decir, un parámetro de gran interés será aquel que tenga estabilidad intra-locutor, y mucha variación inter-locutor.

Para poder comprobarlo, se hizo una primera prueba con pocos locutores, recurriendo para ello al grupo de locutores "Reducido" definido en la Tabla 1. En este caso se representa el valor medio de cada parámetro para cada una de las ventanas. Los resultados obtenidos para algunos de los parámetros más importantes en este trabajo se representaron mediante diagramas de cajas:



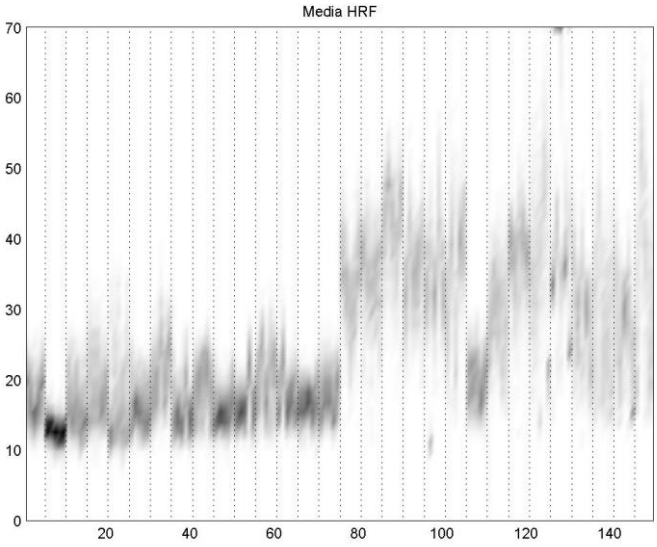


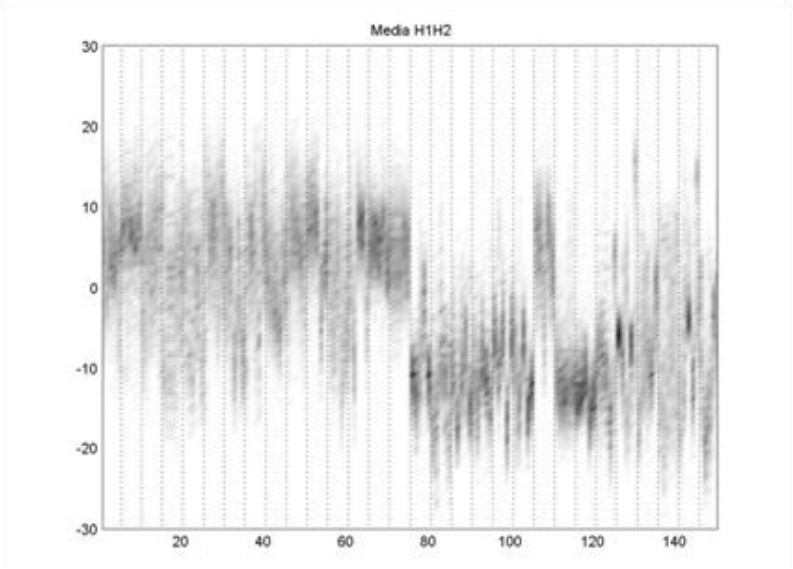


3.4Figura 3.3 Diagrama de cajas de los valores medios de cada ventana de NAQ, HRF, H1-H2, PSP y QOQ del conjunto de locutores "Reducido". Cada diagrama representa una locución, y las líneas verticales separan diferentes locutores

Como se comprueba, sí que hay diferencias entre locutores diferentes y se observa cierta estabilidad dentro del mismo locutor para diferentes grabaciones. Por este motivo será interesante observar qué ocurre aumentando el número de locutores, empleando para ello el conjunto de locutores "Extendido" definido en la Tabla 1.

En las siguientes figuras se muestran histogramas de un parámetro dado de todas locuciones. Cada línea vertical punteada representa divisiones entre diferentes hablantes, por lo que dentro de estas regiones encontramos los 5 histogramas correspondientes a las 5 locuciones, mostrándose con tonos más oscuros los valores más altos. Además, para conseguir apreciar la diferencia entre sexos, se muestran al principio todas las características correspondientes a mujeres y al final las de los hombres.





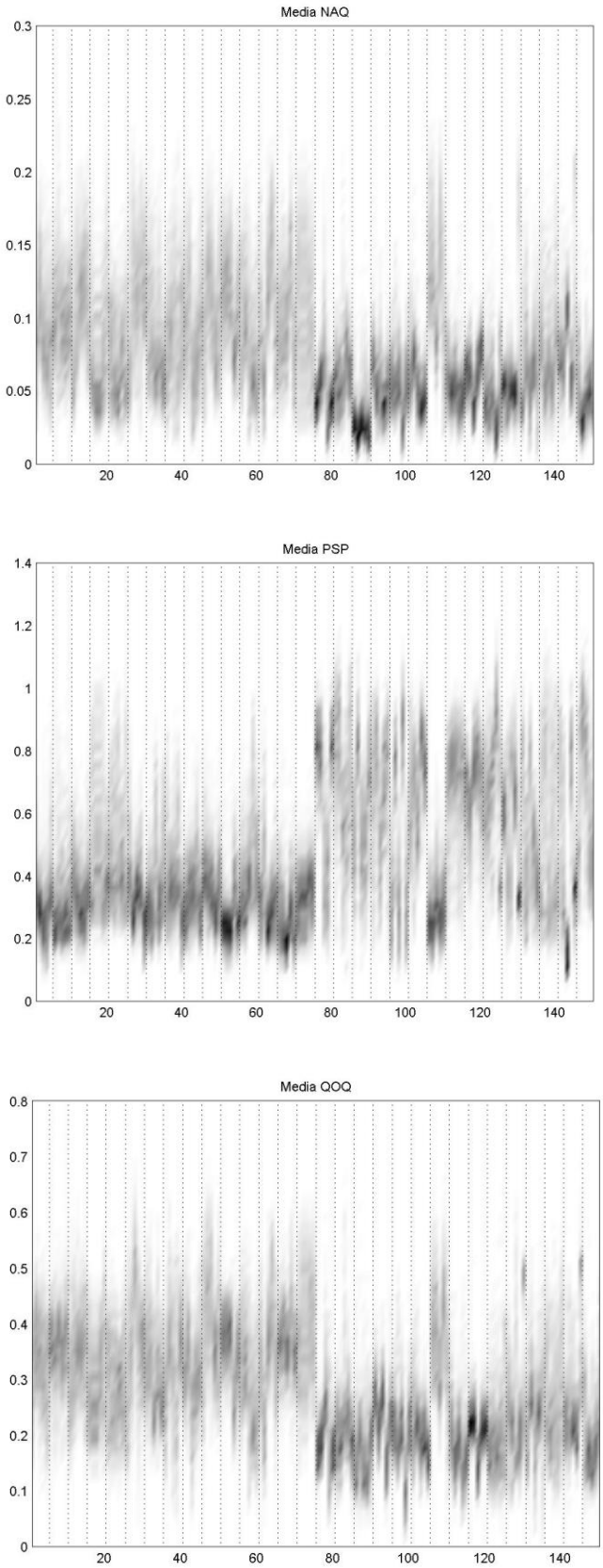


Figura 3.5 Histogramas de los valores medios de NAQ, HRF, H1-H2, PSP y QOQ en cada ventana del conjunto de locutores "Extendido". Las líneas verticales separan diferentes locutores, y entre dos líneas encontramos los 5 histogramas de cada locutor

Siguiendo el mismo modelo de representación, se muestra una comparativa entre la frecuencia fundamental estimada y el número de cierres glotales por segundo:

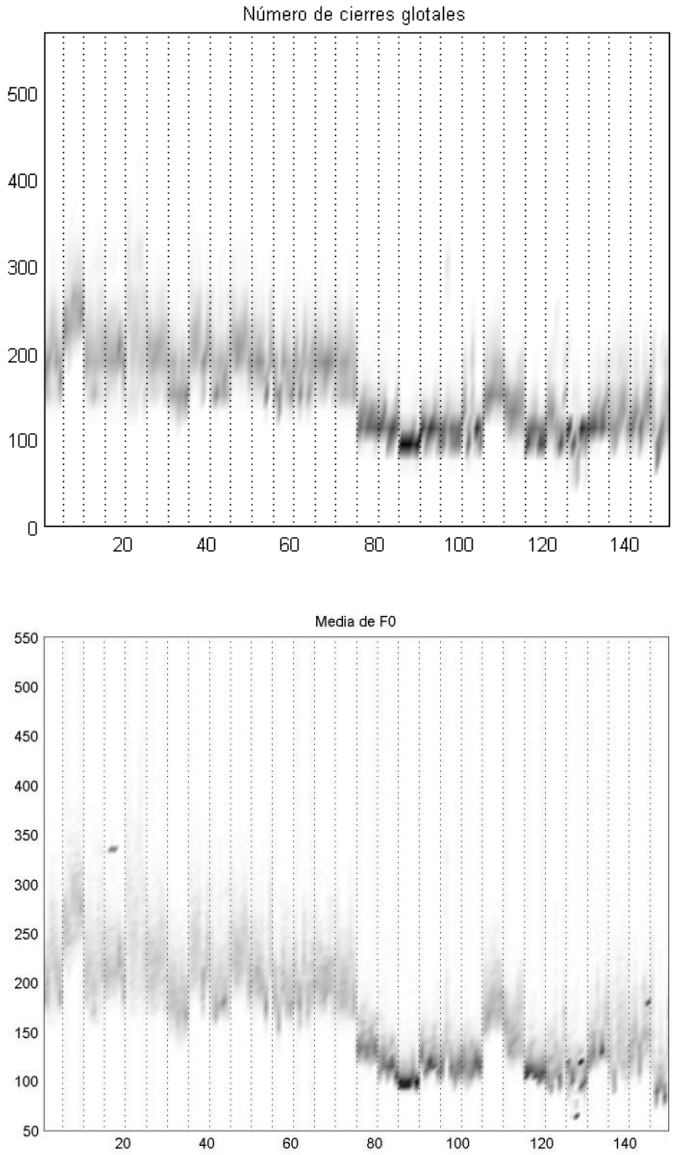


Figura 3.6 Histogramas de los valores medios de F0 en cada ventana y del número de cierres glotales por segundo del conjunto de locutores "Extendido". Las líneas verticales separan diferentes locutores, y entre dos líneas encontramos los 5 histogramas de cada locutor

La frecuencia fundamental viene dada por la velocidad con la que las cuerdas vocales se cierran y se abren, por lo que era de esperar la gran similitud de estas dos gráficas.

3.2.2 Valores instantáneos

Anteriormente todas las pruebas realizadas se habían llevado a cabo con una parametrización por ventana, de modo que para cada ventana se obtenía un valor por cada característica. El siguiente paso fue realizar algunas comprobaciones con valores instantáneos. En las siguientes dos figuras se muestran con el mismo tipo de representación ya empleada, histogramas de dos de los parámetros para el conjunto de locutores "Extendido":

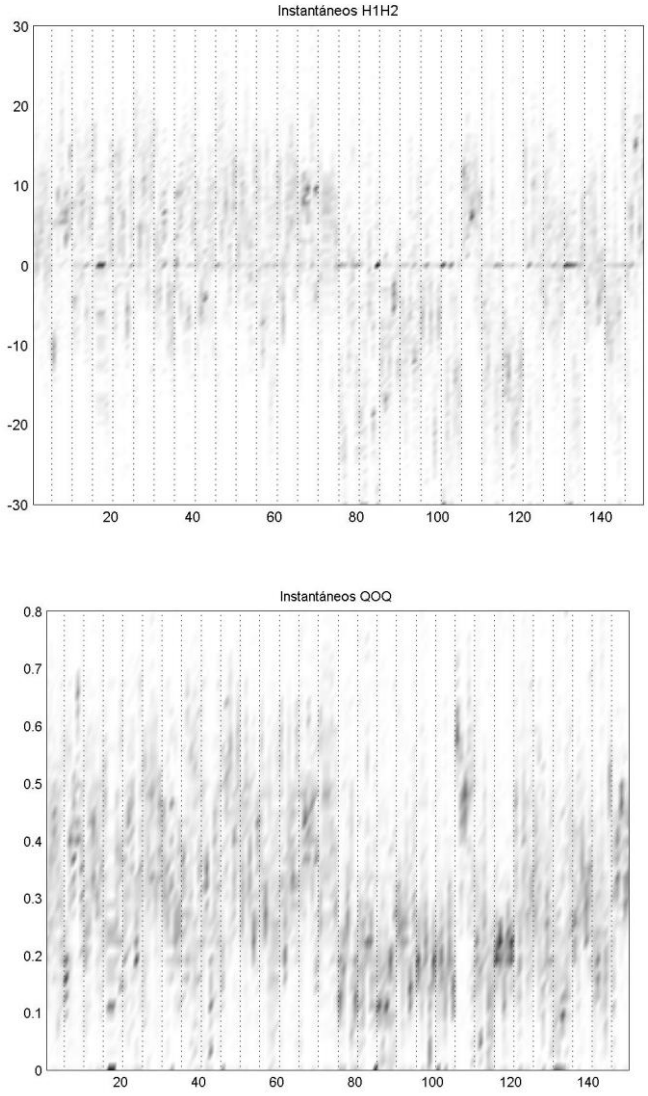


Figura 3.7 Histogramas de los valores instantáneos de H1-H2 y QOQ del conjunto de locutores "Extendido". Las líneas verticales separan diferentes locutores, y entre dos líneas encontramos los 5 histogramas de cada locutor

En estas gráficas se comprueba que los parámetros presentan valores muy inestables dentro de un mismo locutor en diferentes grabaciones, por lo que se estima que este tipo de parametrización no será de ayuda.

Por otro lado, caracterizar las locuciones de este modo tiene otro problema añadido: no se obtiene el mismo número de valores de todos los parámetros. Esto limitaría en gran medida poder añadir nuevos parámetros a nuestro vector de características, ya que para que el sistema pueda funcionar correctamente el vector debe tener siempre la misma dimensión.

3.3 Técnicas de modelado y entorno experimental

Con todos los audios parametrizados, comenzamos las pruebas de identificación. Para ello se empleará el ya explicado sistema GMM-UBM, y Z-norm para conseguir una normalización de los resultados.

El UBM se entrenará con las características de los 150 audios de los que dispone el conjunto de locutores "Extendido", que aunque no es la forma óptima de llevar a cabo el algoritmo, sí que servirá para comprobar de qué medida los parámetros empleados son de utilidad para caracterizar a un determinado locutor.

La base de datos de 150 locutores se ha separado en archivos de train y archivos de test, teniendo para cada locutor 2 audios de train y 3 de test. El modelo de cada hablante ha sido elaborado con los dos ficheros de train. A la hora de elaborar las pruebas de identificación se decidió enfrentar todos los modelos con todas las locuciones de test, de modo que en cada caso tendremos 3 scores target, y 87 non-target.

En cuanto a Z-norm, como se ha explicado en la sección 2.3.3, se han empleado los archivos de train para conseguir valores estadísticos de locuciones de las que se conoce a ciencia cierta que pertenecen a locutores impostor.

3.4 Medidas de error y tipos de gráficas empleadas

A la hora de cuantificar el error que se está cometiendo con nuestro sistema, hay que tener en cuenta dos tipos diferentes de errores:

- Errores de inserción o falsos positivos: considerar que dos locuciones de locutores diferentes pertenecen al mismo hablante.
- Errores de borrado o falsos negativos: considerar que dos locuciones de un mismo locutor se corresponden a hablantes diferentes.

Como medida para caracterizar la eficacia del sistema, emplearemos la tasa de error conocida como Equal Error Rate (EER). Con este valor, podremos determinar si nuestra caracterización es adecuada, teniendo en cuenta los dos tipos de error que nos vamos a encontrar. Cuanto menor sea el EER mayor será el valor de los scores target frente a los valores non-target. [13]

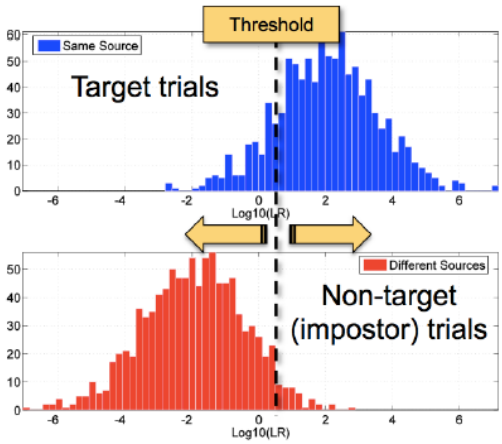


Figura 3.8 Histogramas de scores target y non-target [13]

En el ejemplo de la figura anterior, se pueden ver los dos tipos de errores que nos encontraremos. Un sistema óptimo sería aquel en el que no hubiera ningún score de impostor por encima del umbral, y a la vez ningún score del mismo locutor por debajo.

Para poder visualizar todos los scores obtenidos para cada uno de los locutores, se recurrirá a faunagramas: cada fila de este tipo de gráfica representa un solo locutor, y en ella se reflejan los scores target y non-target de forma que puedan ser diferenciados.

También se emplearán curvas DET para comparar diferentes sistemas. En este tipo de gráficas se representa en el eje de ordenadas la probabilidad de falso negativo, y en el eje de abscisas la probabilidad de falso positivo. De esta forma es muy sencillo comprobar la variación del tipo de error en función del otro tipo. La intersección de la curva DET con la diagonal principal coincide con el EER.

Conjuntamente a estas dos tipos de representaciones, también se mostrará la evolución del EER para diferentes números de mezclas, e histogramas de scores target y non-target.

4 Pruebas y resultados

Ya hemos definido cómo van a ser las bases de datos empleadas, las características con las que podemos caracterizar a los locutores y el sistema empleado para poder realizar pruebas de identificación de locutores. En esta sección vamos a mostrar algunos de los resultados más relevantes de los que se han obtenido.

4.1 Pruebas de identificación con valores medios y varianzas

El primer vector propuesto contenía 11 parámetros que a priori se estimó podían ser de utilidad. A lo largo de esta sección este vector será identificado como vector Completo. Para cada ventana de todos los audios, se obtiene un valor de cada una de las siguientes características:

μ y σ de NAQ	Normalized Amplitude Quotient
μ y σ de QOQ	Quasi-Open Quotient
μ y σ de PSP	Parabolic spectral parameter
μ y σ de H1H2	H1-H2 ratio
μ y σ de HRF	Harmonic Richness Factor
Nº de GCIs	Glottal Closure Instant

Tabla 2 Composición del vector "Completo"

Los resultados obtenidos son los siguientes:

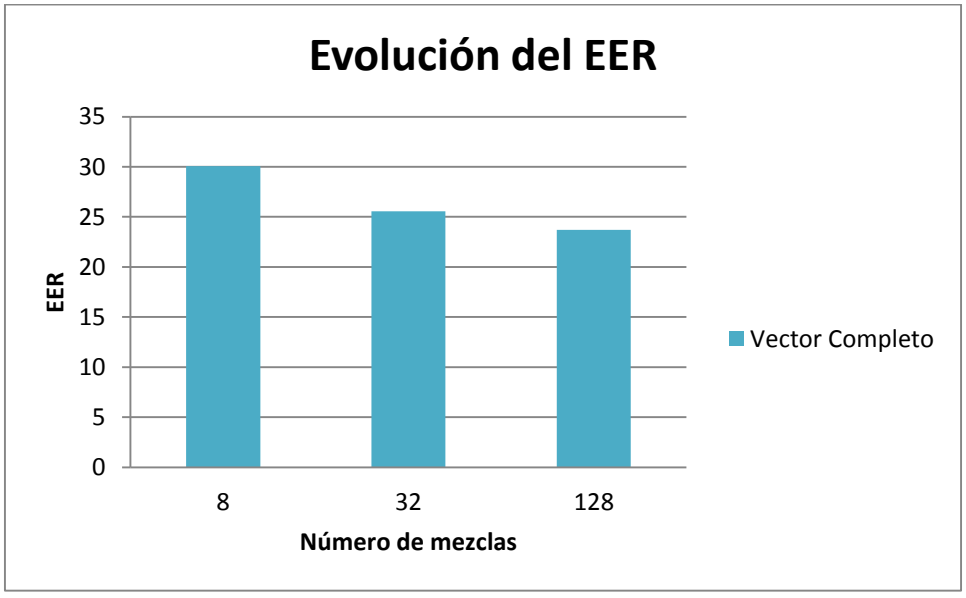


Figura 4.1 Evolución de EER para diferente número de mezclas con el vector Completo

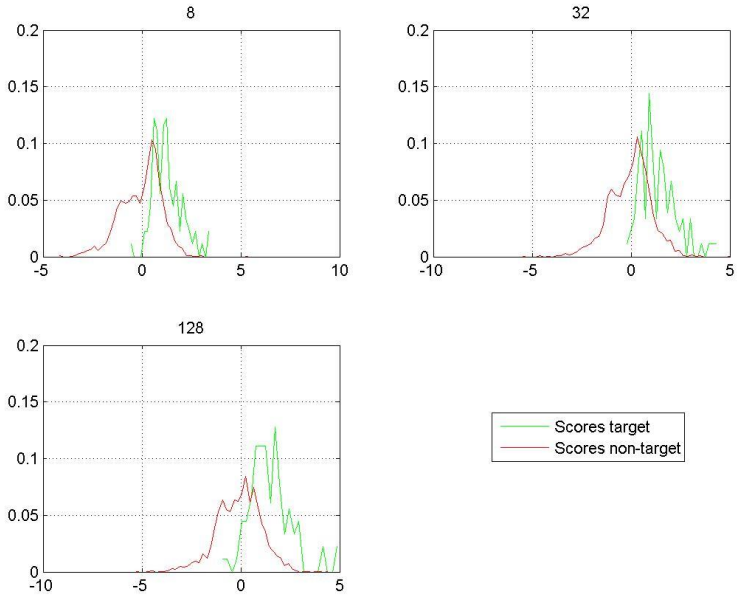


Figura 4.2 Histogramas de scores obtenidos con el vector Completo para diferente número de mezclas

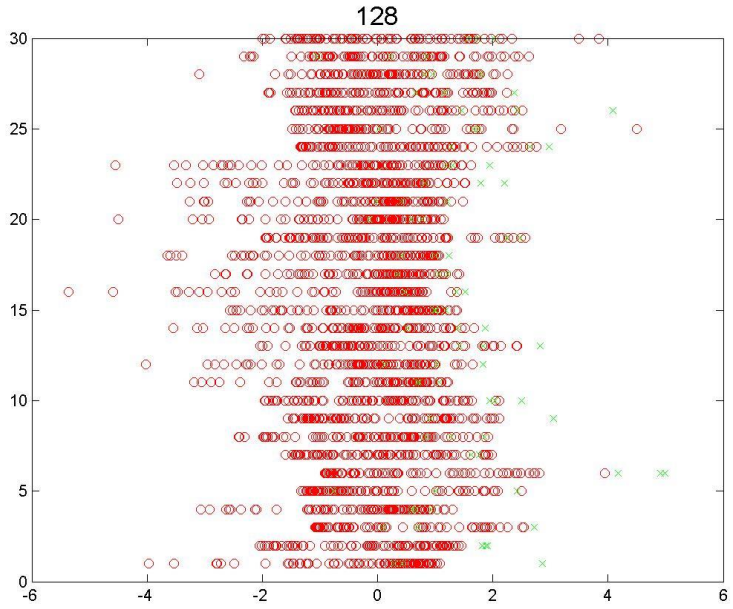


Figura 4.3 Faunagrama obtenido para el vector Completo con el número de mezclas que mejor resultados ofrece (128)

Para seguir mejorando los resultados se propusieron nuevos vectores añadiendo algunos parámetros como son PS y MDQ, y quitando aquellos que se pudieran considerar que afectan negativamente.

El vector con el que se consiguieron mejores resultados, y además es un vector simplificado respecto a otros que fueron testeados, un vector únicamente de 5 valores por cada ventana. Nos referiremos a este vector como vector Simplificado:

μ de NAQ	Normalized Amplitude Quotient
μ de QOQ	Quasi-Open Quotient
μ de PSP	Parabolic spectral parameter
μ de H1H2	H1-H2 ratio
μ de HRF	Harmonic Richness Factor

Tabla 3 Composición del vector "Simplificado"

Los resultados obtenidos con este parámetro son los siguientes:

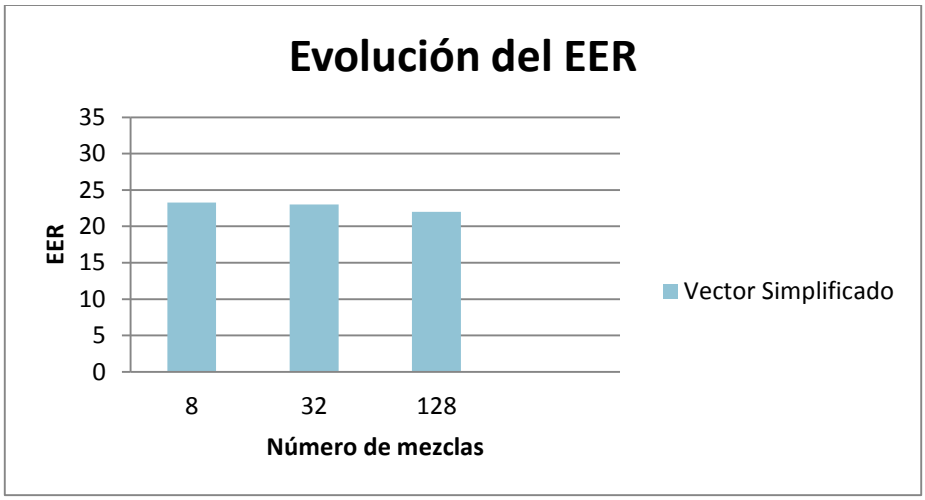


Figura 4.4 Evolución de EER para diferente número de mezclas con el vector Simplificado

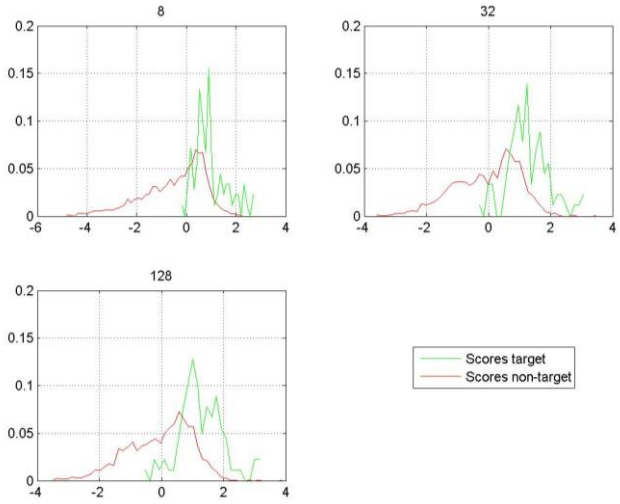


Figura 4.5 Histogramas de scores obtenidos con el vector Simplificado para diferentes números de mezclas

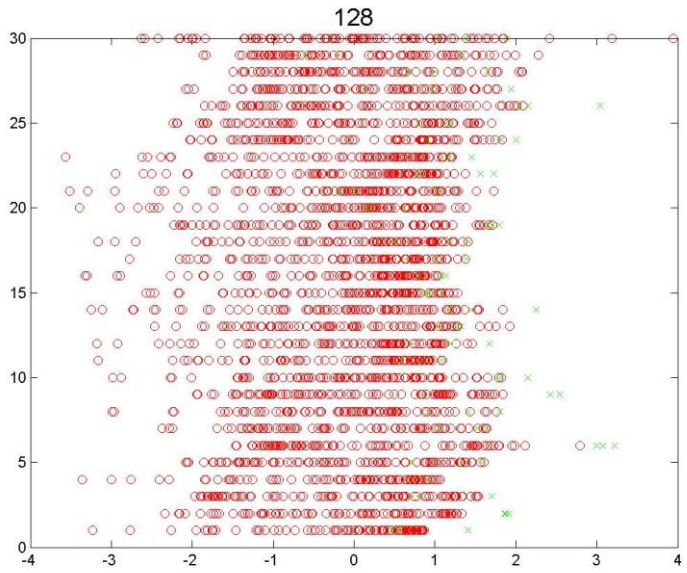


Figura 4.6 Faunagrama obtenido para el vector Simplificado con el número de mezclas que mejor resultados ofrece (128)

En la siguiente figura se muestra una comparación de las evoluciones de EER para los dos vectores comentados:

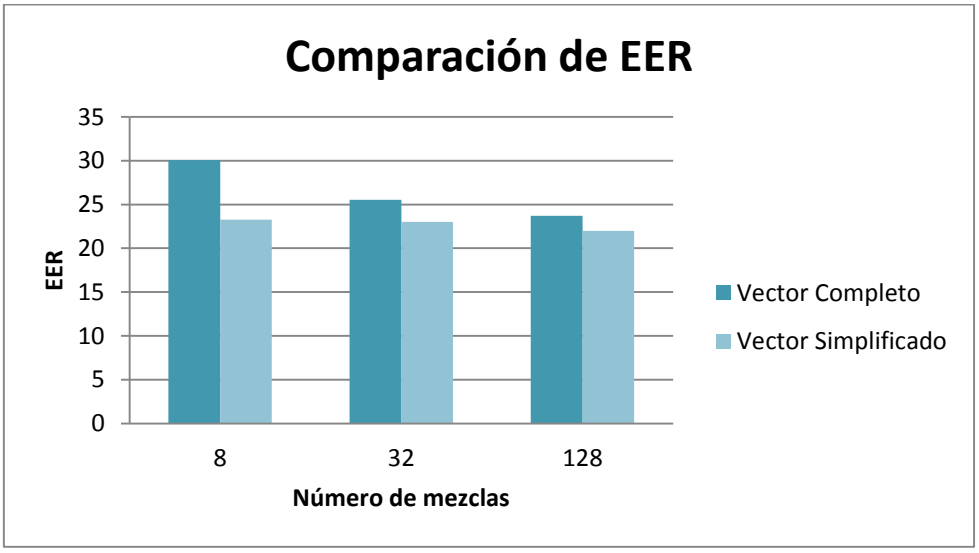


Figura 4.7 Comparativa de la evolución de EER para los vectores Completo y Simplificado

En todos los resultados anteriores fue aplicado el algoritmo Z-norm para normalizar los scores, explicado en le sección 2.3.3. Gracias a ello se consigue que los valores de score de impostor para cada locutor tengan media cero y desviación típica uno.

A continuación se muestran dos faunagramas de un mismo experimento, el primero con aplicación de Z-norm, y el segundo sin él:

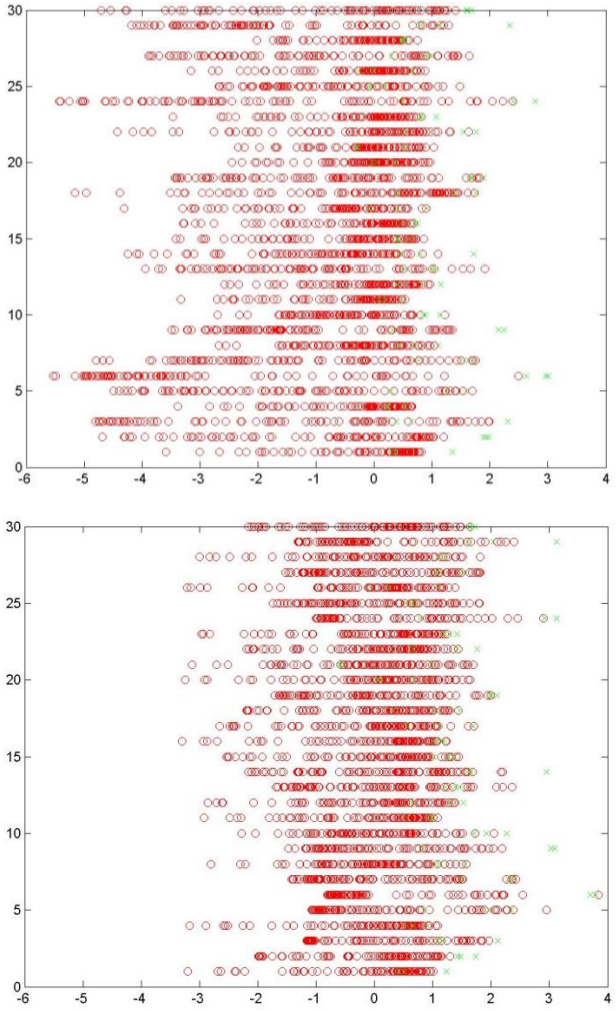


Figura 4.8 Faunagramas de scores sin aplicar Z-norm (imagen superior), y sin aplicarlo (imagen inferior)

Comparando ambas gráficas, se comprueba que en la segunda las distribuciones de scores de impostor están centradas en cero, lo que no ocurre en el primer caso. Gracias a ello es posible establecer un umbral único independiente del locutor.

Para finalizar esta sección, se hará un pequeño resumen de los resultados obtenidos con diferentes configuraciones. Para ello se recurre a la siguiente tabla:

Nº de mezclas	Vector Completo		Vector Simplificado	
	Sin Z-norm	Con Z-norm	Sin Z-norm	Con Z-norm
8	31.6	30.1	25.2	23.1
32	27.9	25.5	24.4	23.0
128	24.7	23.7	23.9	22.1

Tabla 4 Resumen de las tasas de error obtenidas para diferentes configuraciones tanto para el vector completo como para el simplificado

Se ha comprobado que la técnica Z-norm ofrece una mejoría en los resultados, del mismo modo que el vector simplificado, que no solo reduce el coste computacional en el proceso de modelado e identificación, sino que también obtiene tasas de error menores. En cuanto al número de mezclas, en los experimentos llevados a cabo se reduce el error a medida que es aumentado hasta llegar a 128, a partir del cual los valores de EER dejan de mejorar.

Por lo tanto, el mejor valor obtenido se consigue con el vector simplificado, aplicando Z-norm y con 128 mezclas.

4.2 Agrupación de locutores

En esta sección, buscaremos similitudes entre varios locutores de forma que puedan agruparse. Esto podría ser de ayuda en un futuro, ya que si bien agrupando no conseguimos determinar si dos locuciones pertenecen al mismo hablante, sí que podríamos ser capaces de discriminar en caso de pertenecer a diferentes grupos.

4.2.1 Matrices de distancia

Para ver posibles semejanzas se recurrió a matrices de distancias tanto entre las locuciones, como en los locutores.

La matriz de distancias entre locuciones, será una matriz $N_l \, x \, N_l$, donde N_l es el número de locuciones. Para cada posición i, j de la matriz, se calcula la distancia entre la locución i y la locución j. El resultado obtenido es el siguiente:

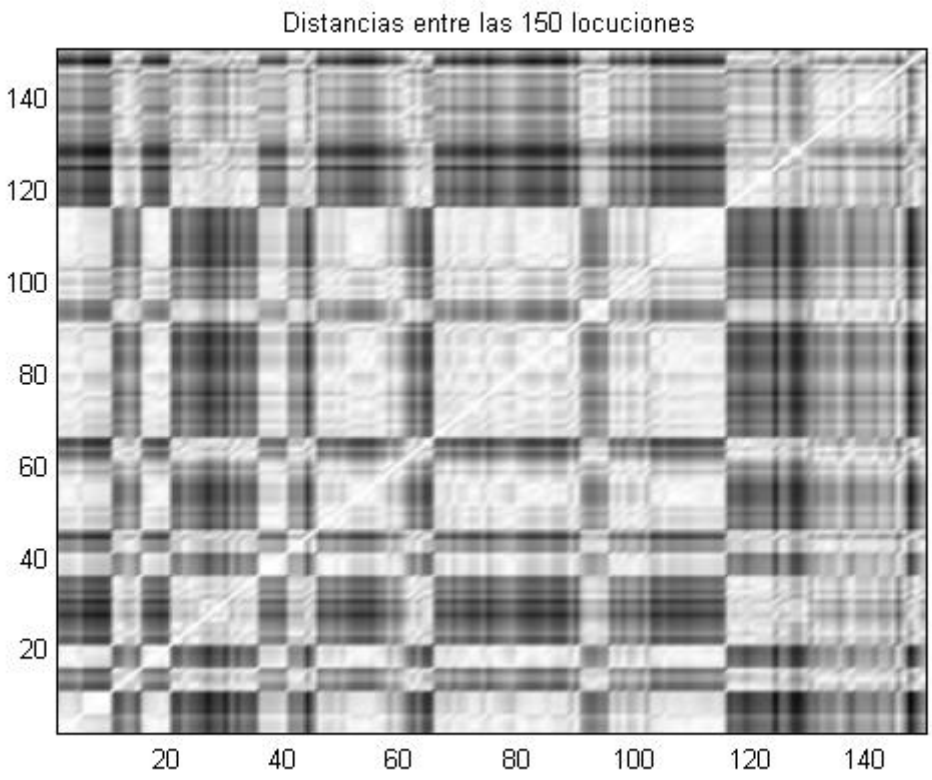


Figura 4.9 Matriz de distancias entre las locuciones del conjunto de locutores Extendido. La posición i,j de la matriz, refleja la distancia entre las locuciones i y j, mostrándose con tonos más claros distancias pequeñas

En esta matriz los valores más claros son aquellos que presentan distancias menores entre locuciones, por lo que la línea blanca que encontramos en la diagonal se debe a enfrentamientos de dos locuciones iguales, lo que implica distancia cero. También es interesante comprobar cómo en esa misma diagonal, se forman pequeñas áreas cuadradas con tonos claros, causadas por distancias entre locuciones diferentes de un mismo locutor.

Además de esta matriz de distancias entre locuciones, se calcularon matrices de distancias entre locutores de dimensión $n_l \; x \; n_l$, siendo n_l el número de locutores. Se plantearon dos métodos para computarlas:

1. Generar un único modelo por locutor, y calcular las distancias entre los modelos de locutor calculados.
2. A partir de la matriz de distancias entre locuciones, generar las distancias medias de los enfrentamientos entre todos los locutores.

Los resultados obtenidos son los siguientes:

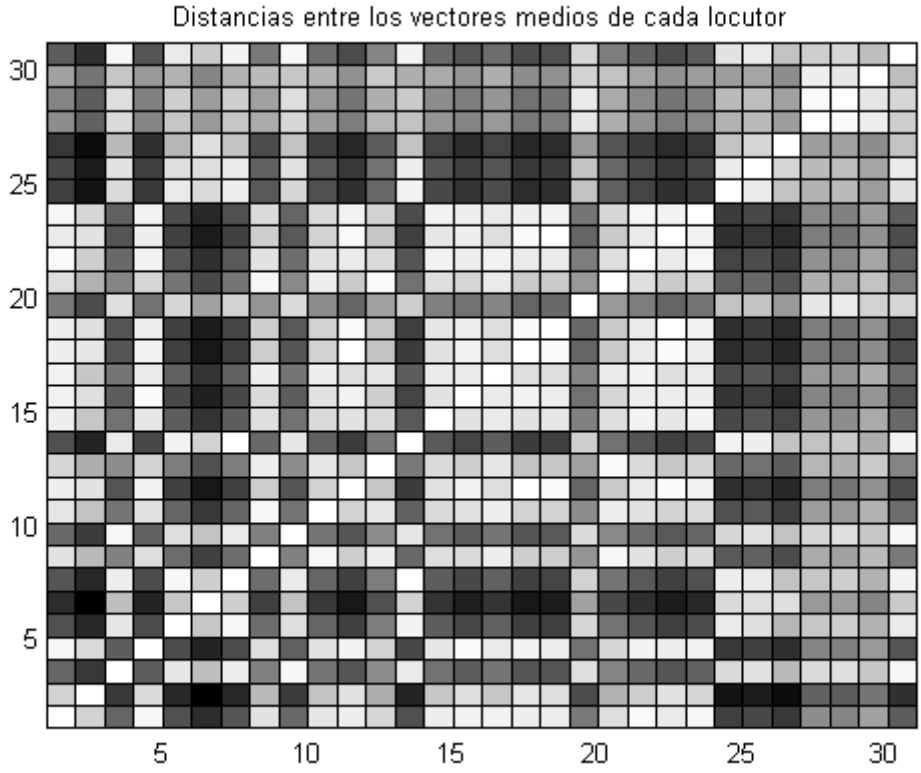


Figura 4.10 Matriz de distancias entre los locutores del conjunto Extendido, calculada a partir de modelos de locutor. La posición i,j de la matriz, refleja la distancia entre los locutores i y j, mostrándose con tonos más claros distancias pequeñas

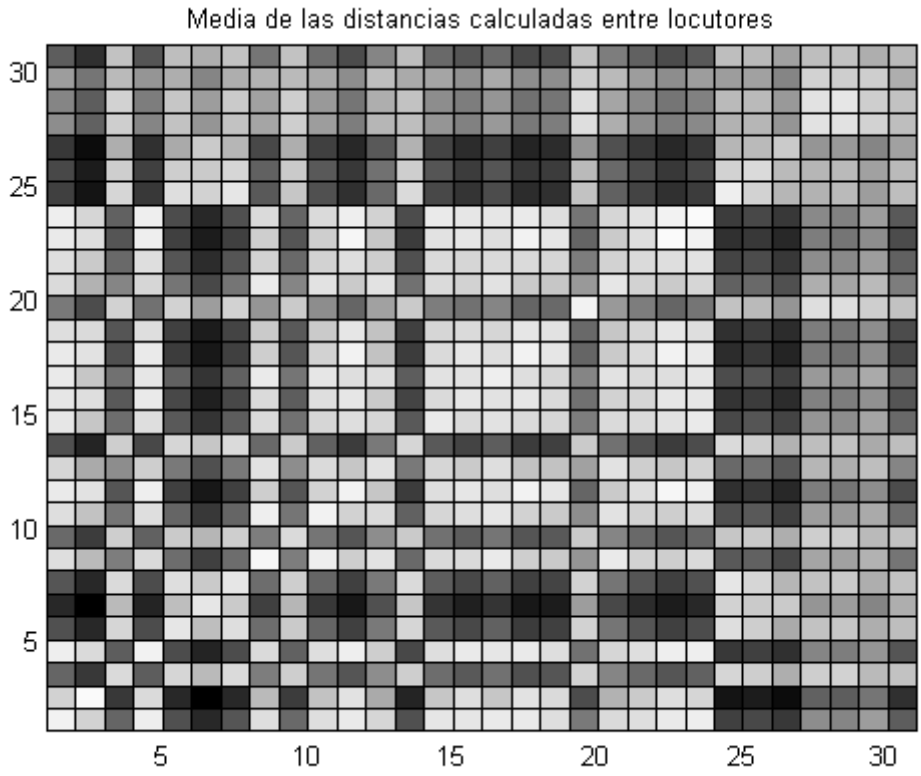


Figura 4.11 Matriz de distancias entre los locutores del conjunto Extendido, calculada a partir de las distancias entre locuciones. La posición i,j de la matriz, refleja la distancia entre los locutores i y j, mostrándose con tonos más claros distancias pequeñas

Estas dos gráficas son muy similares, donde la única diferencia a comentar es que con el primer tipo, los valores de la diagonal siempre serán de distancia cero, mientras que en el segundo caso no será distancia cero, pero sí serán valores bajos.

4.2.2 Clustering aglomerativo

Una vez tenemos calculadas las matrices de distancias entre modelos de locutores, emplearemos la técnica de clustering aglomerativo, o clustering hacia arriba. Para las medidas de distancias entre clusters se ha escogido la técnica basada en calcular la distancia entre centroides (Centroid Method).

Como ya se ha explicado, este método consiste en iterativamente buscar los dos locutores que más se parecen y agruparlos en una misma clase que para las iteraciones sucesivas representará a todos los hablantes asociados.

En la siguiente gráfica, se muestra la distancia entre clusters, y el sumatorio de las distancias que separan a cada modelo de locutor hasta el centroide del cluster al que pertenece:

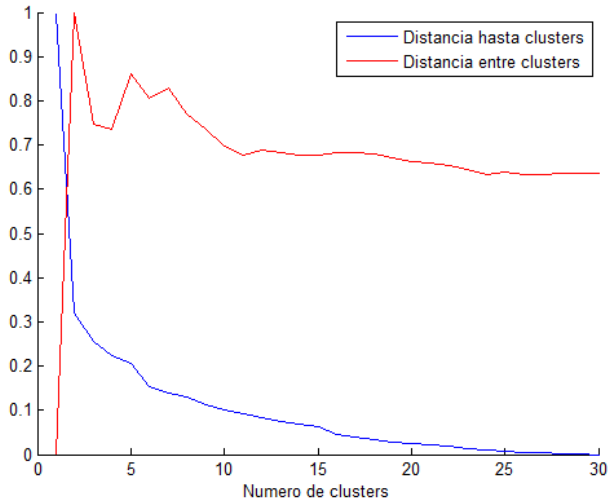


Figura 4.12 Distancia entre clusters y distancia entre el modelo de locutor y el centroide al que pertenece

Ambas curvas han sido normalizadas para poder ser visualizadas en una única figura. Además, la distancia entre cetroides ha sido dividida por el número de relaciones entre clusters para cada caso.

Para comprender bien la evolución de la curva azul, hay que tener presente que con un número de clusters igual al número de locutores, la distancia hasta el centroide siempre va a ser cero; y en cambio, a medida que aumentamos el número de clases, hay menos clases en las que solo esté asociado un hablante.

Para poder visualizar las agrupaciones que se llevan a cabo, se representa la matriz de distancias entre los 30 locutores, pero completando dicha matriz de forma ordenada de acuerdo a los grupos de los hablantes. En la siguiente figura se muestra un ejemplo con 3 grupos:

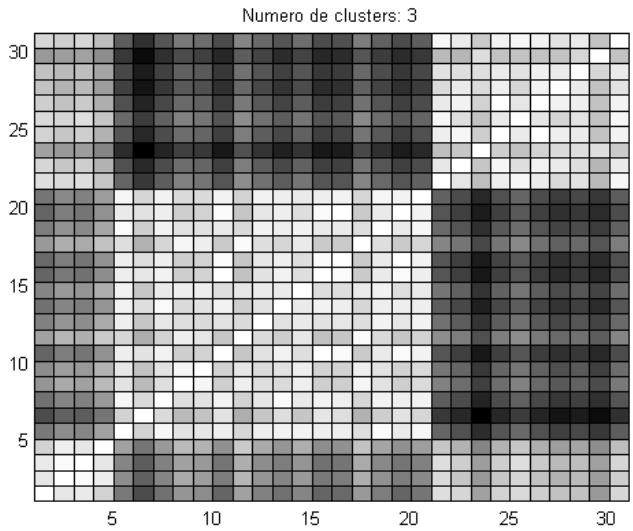


Figura 4.13 Matriz de distancias entre los locutores del conjunto Extendido, representando los locutores de forma ordenada en función del grupo asociado

En esta nueva matriz se pueden distinguir los 3 grupos creados. En cada caso, los locutores con los que tienen mayor semejanza siempre están en el área cercana donde solo hay hablantes del mismo tipo.

Tras esto, es interesante recuperar uno de los histogramas de scores target-nontarget, y comprobar si dentro de los scores non-target, los que tenían mayor puntuación y hacían bajar la eficiencia del sistema se corresponden con locutores del mismo grupo.

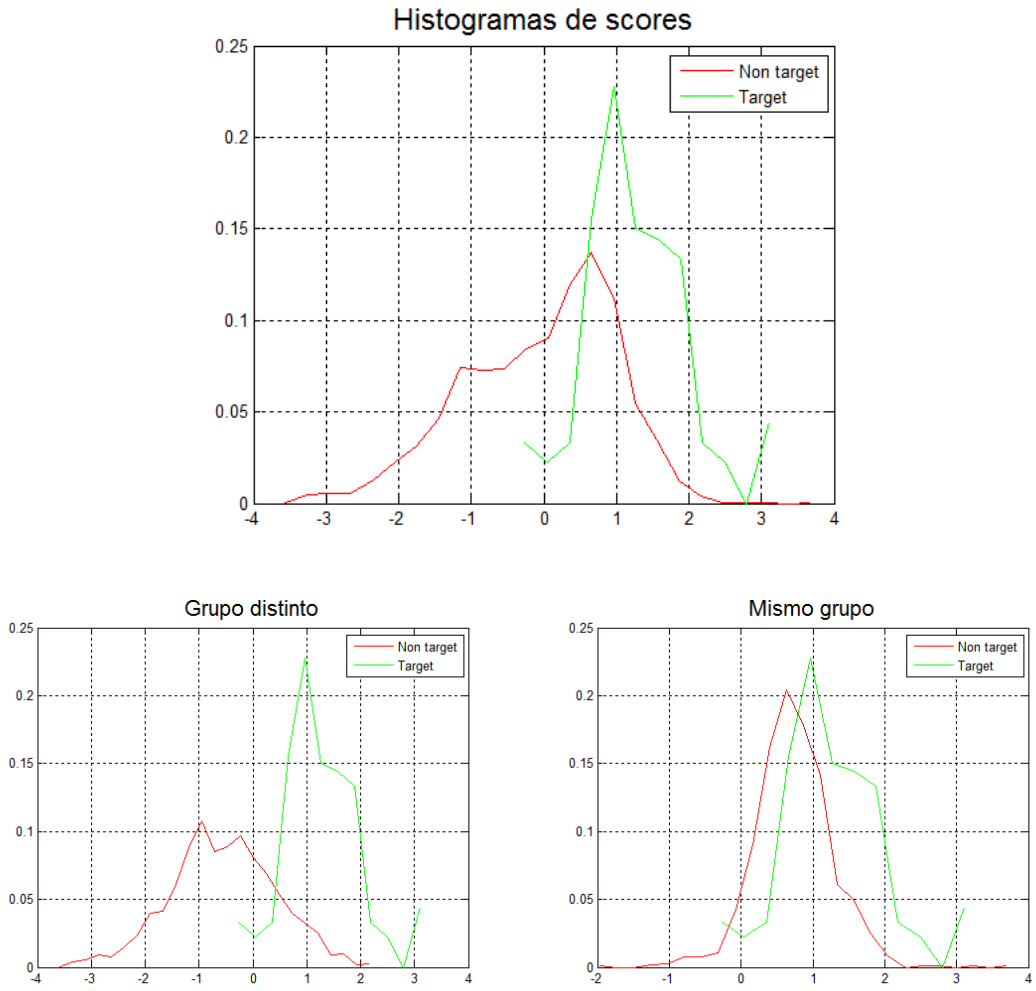


Figura 4.14 Comparación de histogramas de scores separando los scores non target entre aquellos que pertenecen al mismo grupo y los que no

En la primera figura se presentan los histogramas originales, mientras que en las siguientes dos gráficas se solo se han tenido en cuenta los scores pertenecientes a dos locutores de grupos distintos (esquina inferior izquierda), o al mismo grupo (esquina inferior derecha). Contrastando las 3 imágenes, se puede comprobar que cada uno de los dos picos que se observan en el histograma de scores de impostor de la primera gráfica, corresponden a los dos histogramas reflejados en las figuras inferiores.

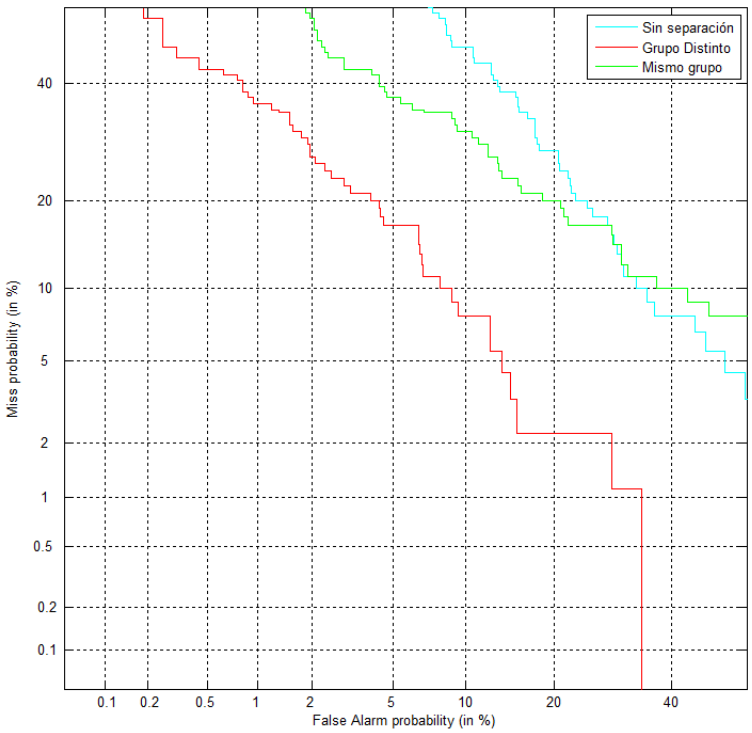


Figura 4.15 Curvas DET sin separar en grupos, enfrentando locutores del mismo grupo y enfrentando locutores de grupos diferentes

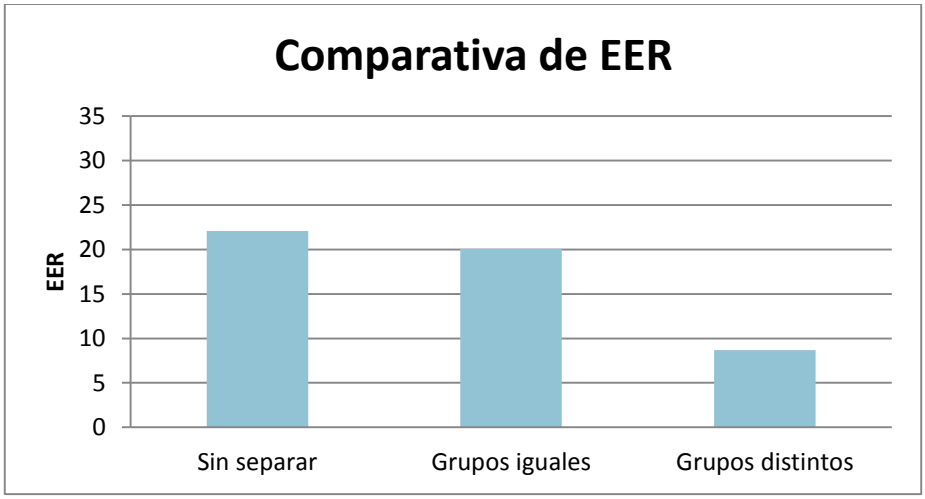


Figura 4.16 Comparación de EER sin separar en grupos, enfrentando locutores del mismo grupo y enfrentando locutores de grupos diferentes

En estas gráficas se comprueba que el porcentaje de error cometido con locutores pertenecientes a grupos distintos es drásticamente menor al que se comete en el caso de no hacer una separación por grupos, o cuando enfrentamos sólo a locutores de un mismo grupo.

4.3 Fusión de scores con técnicas basadas en MFCCs

En esta última sección de pruebas, se ha querido comprobar si estos parámetros podrían servir de complemento para un sistema basado en MFCCs, los usados convencionalmente para identificación de locutores.

4.3.1 Sistema de referencia GMM-MFCC

Para ello, se ha elaborado un sistema muy básico de parametrización MFCC y se han combinado los resultados con los que se habían obtenido con los parámetros glotales.

La parametrización con estos parámetros se va a llevar a cabo en los mismos segmentos y con el mismo tipo de enventanado que el realizado para obtener las características glotales, para tener dos sistemas cuyos resultados son obtenidos a partir de la misma cantidad de datos. Adicionalmente, para la extracción de MFCCs se ha empleado la técnica Cepstral mean and variance normalization (CMVN), la cual proporciona robustez a esta parametrización. Consiste en calcular la media y varianza de cada MFCC a lo largo de toda la señal de voz, y restar por la primera y dividir por la segunda, para así conseguir que la distribución de todos los MFCCs para el audio tenga media cero y desviación típica uno. Gracias a ello se consigue eliminar la distorsión en los coeficientes cepstrales que provoca el ruido estacionario. El vector resultante con estos parámetros será denominado como vector MFCC.

El faunagrama obtenido para un número de mezclas igual a 128 y con aplicación de Z-norm es el siguiente:

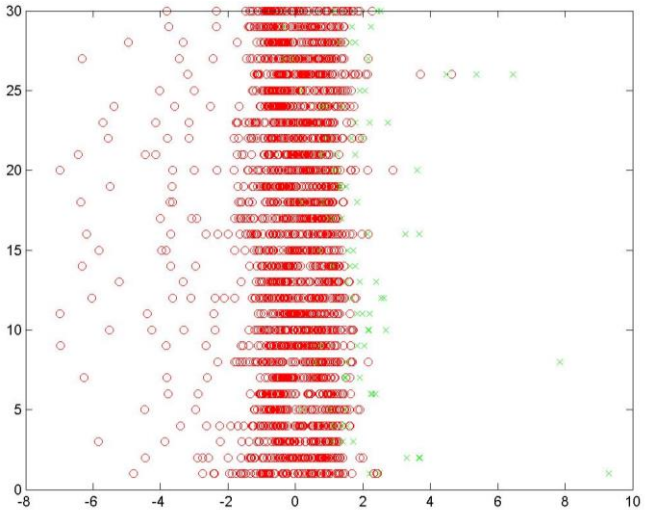


Figura 4.17 Faunagrama obtenido para el vector MFCC con el número de mezclas que mejor resultados ofrece (128) y con aplicación de Z-norm

4.3.2 Resultados de la fusión

Para hacer la combinación de ambos métodos, se recurre al método de fusión de scores, en el que aprovechando que todos los valores de verosimilitud están normalizados con z-norm, bastará con hacer una suma de los resultados obtenidos con los dos sistemas que se quieren fusionar.

Los resultados obtenidos son los siguientes:

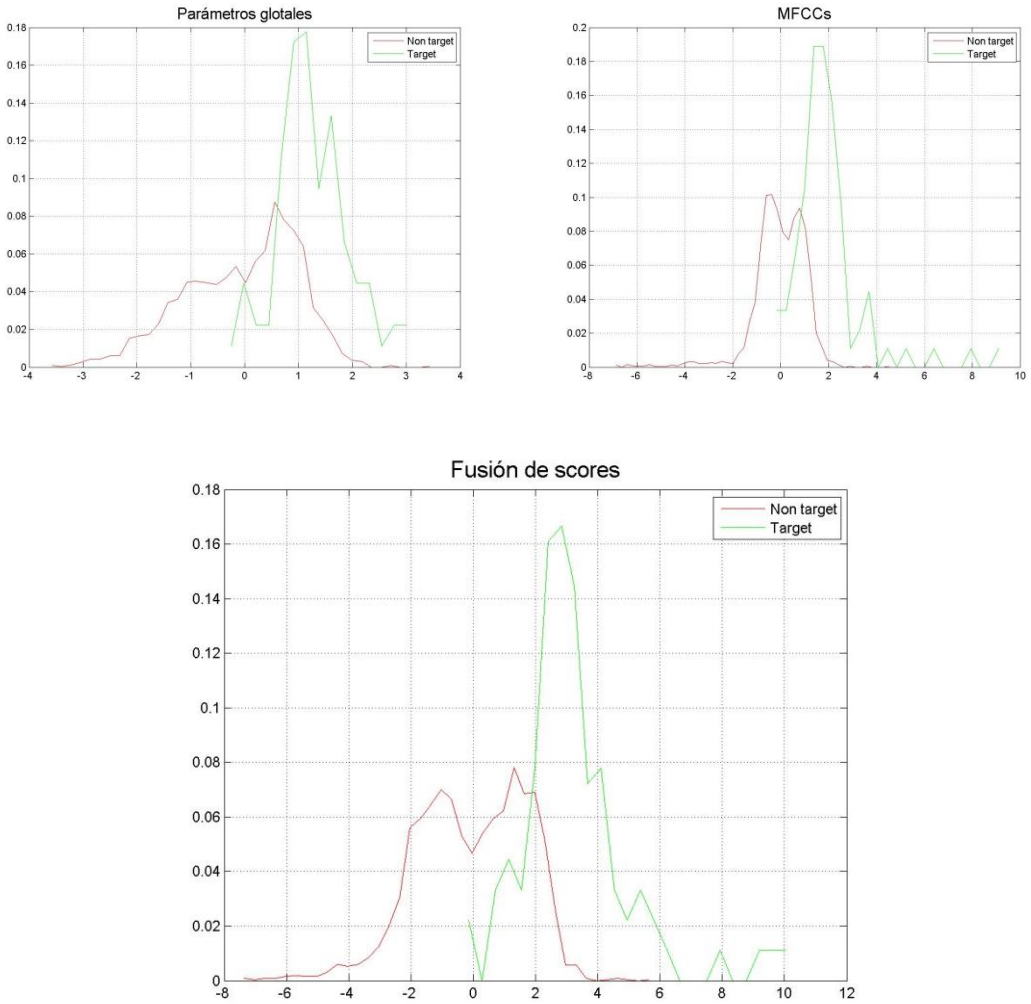


Figura 4.18 Comparativa de los histogramas de scores obtenidos con el vector Simplificado, MFCC y tras la fusión de scores

En cuanto a tasas de error, se partía de un 15.7% con la parametrización con MFCCs, y tras la fusión se consigue bajarlo hasta un 14.8%, lo que implica una mejora del 5.7%.

Conjuntamente a estos histogramas de scores que son de utilidad para comparar los diferentes resultados obtenidos, se muestran las curvas DET y la evolución de EER para los 3 casos:

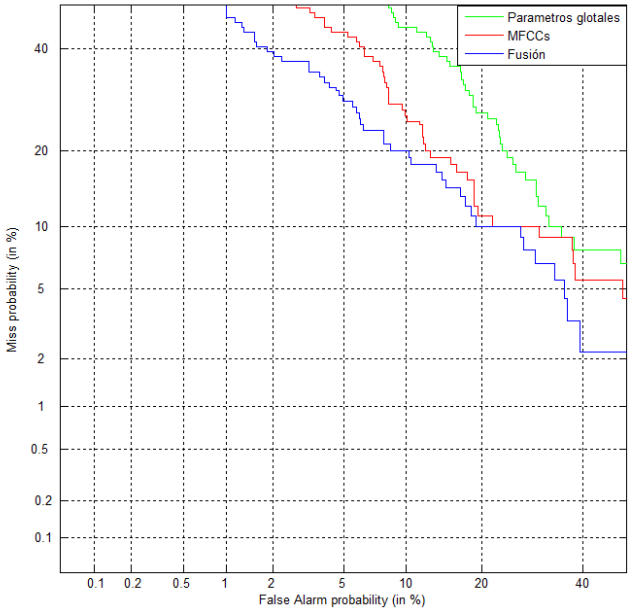


Figura 4.19 Curvas DET para vector Simplificado, MFCC y tras la fusión de scores

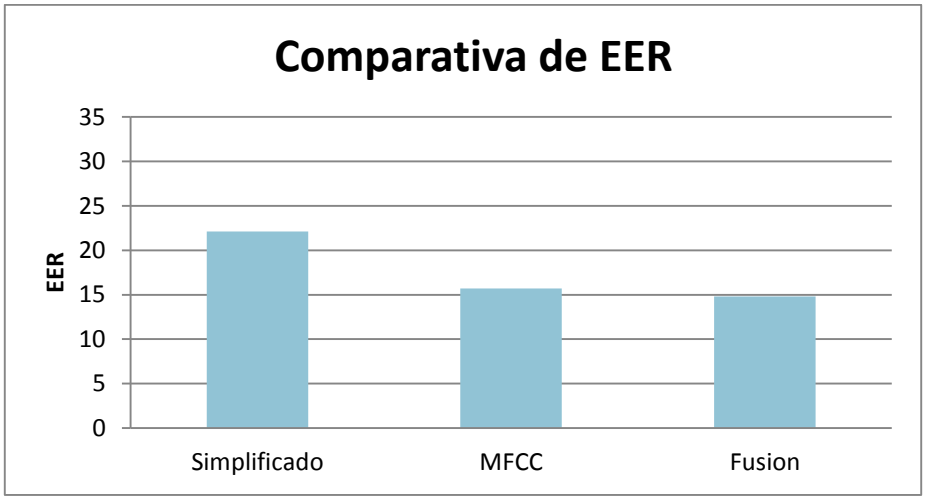


Figura 4.20 Comparación de EER para vector Simplificado, MFCC y tras la fusión de scores

Estos resultados pueden indicar que los parámetros glotales y los cesptrales están ciertamente incorrelados, ya que en caso contrario, no se obtendría ninguna mejoría a la hora de combinar los resultados.

Por último, se muestran los histogramas de scores tras la fusión haciendo la misma separación de grupos que en la sección anterior:

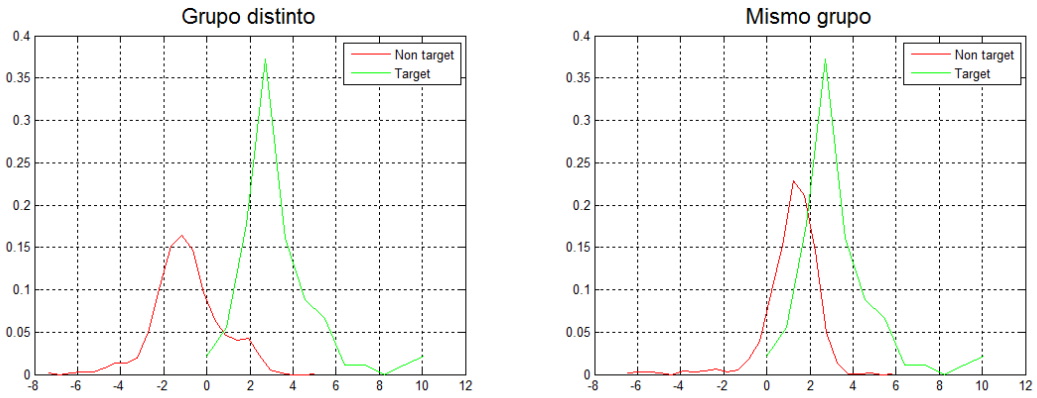


Figura 4.21 Comparación de histogramas de scores tras la fusión, separando los scores non target entre aquellos que pertenecen al mismo grupo y los que no

En este caso, en el histograma de scores de impostor tras la fusión, se observan dos picos muy acentuados. Tras hacer la separación en grupos se comprueba que el pico situado en la zona de scores más altos, corresponde a locutores del mismo grupo.

5 Conclusiones y trabajo futuro

5.1.1 Conclusiones

En este Trabajo de Fin de Grado se han ido desarrollando todo tipo de pruebas para ir cumpliendo los objetivos que se marcaron al principio. Tras todas las pruebas realizadas se pueden extraer las siguientes conclusiones:

- Se demuestra que es posible obtener buenas estimaciones de características relativas al proceso fonatorio a partir de la señal de voz, lo cual siempre había resultado un proceso complicado.

- Se ha conseguido caracterizar al hablante con los parámetros glotales, como se comprueba en la sección *3.2 Parametrización,* donde en las gráficas representadas se pueden comprobar diferencias entre locutores distintos para todos los parámetros mostrados.

- Los resultados de verificación obtenidos con los parámetros glotales en locuciones con independencia del texto no son óptimos, pero si son lo suficientemente buenos para poder ser combinados con otro tipo de características para mejorar los resultados que obtienen éstas.

- Las agrupaciones realizadas han servido para poder identificar que la mayoría de los errores en las pruebas de verificación vienen dadas por enfrentamientos de locutores que pertenecen a mismos grupos, lo cual podrá servir de ayuda en un futuro para poder hacer una discriminación en caso de tener enfrentamientos de locutores sobre los que se tenga la certeza que no pertenecen al mismo grupo.

- Se han conseguido combinar sistemas de MFCCs y el creado para este trabajo mediante la técnica de fusión de scores. Partiendo de un sistema de MFCCs muy simple, se reduce su tasa de error en la base de datos empleada. Esto es de gran interés ya que nos indica cierta incorrelación entre ambos parámetros, lo que puede ser de utilidad para mejorar sistemas más avanzados.

5.1.2 Trabajo futuro

- En un futuro será muy interesante ampliar la base de datos para poder hacer funcionar mejor métodos como el UBM. Además de esto, los resultados obtenidos con un mayor número de locuciones serán más realistas a los obtenidos con 30.

- Para mejorar esta caracterización, se puede llevar a cabo un estudio más detallado de los diferentes parámetros glotales y la relación que hay entre ellos, para así conseguir vectores con mejor rendimiento y de la menor dimensión posible eliminando aquellas características que presenten una gran correlación con otras.

- También sería de interés realizar la fusión de scores con un sistema de MFCCs más eficiente, ya que con las pruebas realizadas se ha mostrado ambos tipos de caracterización están en cierta medida incorreladas.

- Seguir estudiando la división en grupos, de modo que se pueda conseguir la solidez suficiente para poder asegurar que dos locutores nos pertenecen al mismo grupo y así llevar a cabo la discriminación.

6 Referencias

[1] Jacob Benesty, M. Mohan Sondhi, Yiteng Huang. (2008). Springer Handbook of Speech Processing: Springer.

[2] Boris Mirkin. (1996). MATHEMATICAL CLASSIFICATION AND CLUSTERING. Dordrecht: Kluwer Academic Publishers.

[3] Thomas Drugman, Thierry Dutoit. Glottal Closure and Opening Instant Detection from Speech Signals. Mons, Belgium.

[4] D. Reynolds y R. Rose. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing. VOL. 3 (1995).

[5] D. A. Reynolds, T. F. Queatieri and R. B. Dunn: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19-41 (2000).

[6] Drugman, A.Alwan, Joint Robust Voicing Detection y Pitch Estimation Based on Residual Harmonics.

[7] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, Stefan Scherer. COVAREP – A COLLABORATIVE VOICE ANALYSIS REPOSITORY.

[8] P. P. Vaidyanathan. (2008). The Theory of Linear Prediction. California Institute of Technology: Morgan & Claypool.

[9] P. Alku, T. B¨ackstr¨om, y E. Vilkman, Normalized amplitude quotient for parameterization of the glottal flow, J. Acoust. Soc. Am., vol. 112, no. 2, pp. 701–710, 2002.

[10] D. Childers and C. Lee, Voice quality factors: Analysis, synthesis and perception, J. Acoust. Soc. Am., vol. 90, no. 5, pp. 2394–2410, 1991.

[11] P. Alku, H. Strik, y E. Vilkman, Parabolic spectral parameter – A new method for quantification of the glottal flow, Speech Commun., vol. 22, no. 1, pp. 67–79, 1997.

[12] NIST. (Marzo, 2006). The NIST Year 2006 Speaker Recognition Evaluation Plan.

[13] J. Gonzalez-Rodriguez, "Evaluating Automatic Speaker Recognition systems: An overview of the NIST Speaker Recognition Evaluations (1996-2014)", Loquens, CSIC, Vol. 1, n. 1, pp. 1-15, January 2014.

[14] Javier Ortega García. Tratamiento de Señales de Voz y Audio. Escuela Politécnica Superior Universidad Autónoma de Madrid.

[15] Thomas Drugman, Thierry Dutoit. Glottal-based Analysis of the Lombard Effect. TCTS Lab, University of Mons, Belgium.

[16] Messaoud Bengherabi, Farid Harizi, Norman Poh, Elhocine Boutellaa, Abderrazek Guessoum y Mohamed Cheriet. IMPROVING BIOMETRIC VERIFICATION SYSTEMS BY FUSING Z-NORM AND F-NORM.

[17] Haoxuan Li. (Febrero, 2013). Glottal Source Parametrisation by Multi-estimate Fusion. Dublin City University

[18] J. Kane y C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," IEEE Trans. on Audio, Speech, and Lang. Proc., vol. 21, no. 6, pp. 1170–1179, 2013.

[19] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in Proc. Interspeech, 2011, pp. 177–180.

[20] John Laver. (1980). The phonetic description of voice quality. New York: Cambridge.

[21] William J. Hardcastle, John Laver, y Fiona E. Gibbon. (2010). The Handbook of Phonetic Sciences, segunda edición. Wiley-Blackwell.

[22] Murty, K. y Yegnanarayana, B., "Epoch Extraction From Speech Signals", IEEE Trans.Audio Speech Lang. Processing, vol. 16, pp. 1602-1613, 2008.