# UNIVERSIDAD AUTÓNOMA DE MADRID

## ESCUELA POLITÉCNICA SUPERIOR


# -PROYECTO FIN DE CARRERA-

## IDENTIFICACIÓN DE LOCUTOR A PARTIR DE INFORMACIÓN GLOTAL EN UNIDADES LINGÜÍSTICAS

**Ignacio Rodríguez Ortega**

**Mayo 2014**

# IDENTIFICACIÓN DE LOCUTOR A PARTIR DE INFORMACIÓN GLOTAL EN UNIDADES LINGÜÍSTICAS

**AUTOR: Ignacio Rodríguez Ortega**
**TUTOR: Joaquín González Rodríguez**


**ATVS Grupo de Reconocimiento Biométrico**
**(http://atvs.ii.uam.es)**
**Dpto. de Ingeniería Informática Escuela Politécnica Superior Universidad Autónoma de Madrid**
**Mayo 2014**

# *Agradecimientos*

En tan solo unas horas terminaré la carrera de Ingeniero Superior de Telecomunicación tras 6 años como alumno en la Universidad Autónoma de Madrid, donde he pasado momentos fantásticos pero también he tenido que trabajar duro, pero ya se sabe las cosas que merecen la pena no las regalan. Es un momento muy emocionante en mi vida ya que al fin podré decir eso de soy 'ingeniero'.

Quiero dar mi más sincero agradecimiento a Joaquín González Rodríguez, por haberme dirigido este proyecto de fin de carrera. Por todas las atenciones, por el tiempo que ha perdido conmigo, y sobre todo por su apoyo. Mi tutor ha sabido comprender mis necesidades en cada momento y ha puesto todo de su parte en ayudarme en otros asuntos fuera del proyecto, así que no puedo estar más satisfecho con el trato recibido.

También me gustaría destacar la paciencia que han tenido para conmigo Fernando Espinoza, sufridor constante de mis preguntas para la resolución de los distintos problemas de código y formas de representación de datos.

En general, a todos aquellos profesores y alumnos de la laboratorio ATVS que de una u otra forma me han ayudado a realizar este trabajo, que aunque no les mencione de forma explícita, no les puedo negar un sincero agradecimiento. La verdad es que en este laboratorio se respira muy buen ambiente, y todo el mundo está abierto a ayudar en lo que fuera necesario.

Y finalmente, no puedo dejar de agradecer la comprensión de mis padres y mi novia, puesto que ellos han sido los que más me han tenido que aguantar en mis malos momentos, enfados y demás sucesos que me han acaecido a lo largo de la elaboración de este proyecto.

A todos muchas gracias.

# *Índice de contenidos*

# *Lista de figuras*

# *Lista de tablas*

# Acrónimos

| | |
|---|---|
| CPIF: | Closed Phased Inverse Filtering |
| CCD: | Complex Cepstrum Descompotion |
| DAP: | Discrete All Pole |
| DTFT: | Discrete time Fourier Transform |
| EE: | Excitation Strength |
| EER: | Equal Error Rate |
| FFE: | F0 Frame Error |
| FIR: | Finite Impulso Response |
| GCI: | Glottal Closure Instant |
| GMM: | Gaussian Mixture Model |
| GOI: | Glottal Opening Instant |
| GPE: | Gross Pitch Error |
| HRF: | Harmonic Richness Factor |
| IAIF: | Iterative Adaptative Inverse Filtering |
| LPC: | Linear Predictive Coding |
| MDQ: | Maxima Dispersion Quotient |
| NAQ: | Normalized Amplitude Quotient |
| OQ: | Open Quotient |
| PCA: | Principal Component Analysis |
| PS: | Peak Slope |
| PSP: | Parabolic Spectral Parameter |
| QOQ: | Quasi-Open Quotient |
| RESKEW: | Residual Excitation Skewness |
| SEDREAMS: | The Speech Event Detection using the Residual Excitation And a Mean-based Signal |
| SNR: | Signal Noise Ratio |
| SRH: | Summation of Residual Harmonics |
| TSNE: | Stochastic Neighbor Embedding |
| VDE: | Voice decision error |
| ZZT : | Zeros of the Z-Transform |

# *Sección 1*

# *Introducción*

## 1.1 Resumen

En este estudio pretendemos obtener información glotal que caracterice a los locutores según su cualidad vocal. El primer paso es obtener el pulso glotal a partir de la señal de voz ya segmentada en fonemas, para lo cual explicaremos y escogeremos entre distintos algoritmos con el criterio de hacer nuestro sistema lo más robusto y estable. Para esta tarea recurriremos a contribuciones de distintos investigadores. La estimación del pulso glotal sobre lenguaje hablado implica una serie de algoritmos complejos y que si no se realiza de manera conveniente va a repercutir en la extracción de características glotales.

Más tarde extraeremos de la señal glotal ciertos parámetros siguiendo el modelo LF (Liljencrants Fant). El siguiente paso consistirá en realizar un análisis inter/intra variabilidad con 200 locutores de la base de datos TIMIT con el fin de proponer un vector con la media y la varianza de los parámetros LF que más nos convengan.

Por último con el motor de scoring GMM-UBM-MAP aplicado a los 630 locutores de TIMIT podremos cuantificar si estas características glotales diferencian suficientemente a unos locutores de otros.

## 1.2 Estructura del proyecto

El proyecto consta dos partes fundamentales, la parte donde se describe el sistema que saca las características glotales para un segmento de voz dado, y los experimentos diversos que se realizan posteriormente para comprobar la fiabilidad y el alcance que tienen esta características a la hora de caracterizar a un locutor diferenciándolo de otros.

Dentro de la primera parte se incluirían las secciones 1, 2, 3 y 4 dentro de las cuales se hace una comparativa de los diferentes algoritmos explicando de manera más exhaustiva el algoritmo escogido. La segunda parte la sección 5 muestra distintos experimentos sobre la base de datos TIMIT.

## 1.3 Motivación

La motivación de este proyecto es como bien dice el título extraer información glotal de las unidades lingüísticas que más nos convengan para caracterizar a un locutor. Con esta información no pretendemos competir con los métodos de identificación tradicionales, pero sí va a lograr clasificar a los locutores en una serie de grupos. Poniendo un ejemplo con el tipo de sangre que dos personas tengan el mismo tipo de sangre no significa que sean la misma persona, pero ya vale para descartar a mucha gente que no tenga tu tipo de sangre. En este caso, podremos incluir gracias a estas características obtenidas a un locutor en voz tensa, carrasposa, modal... las llamadas cualidades vocales (voice qualities). Sin embargo las voice qualities no son grupos perfectamente limitados, puede haber mezclas. Gente que esté en la frontera de dos tipos de cualidad vocal. Esta nueva identificación tiene la ventaja de que sacar información de carácter distinto a los métodos tradicionales de identificación.

Este proyecto pretende además ser utilizado con el sistema que está actualmente en funcionamiento en el laboratorio del grupo de investigación ATVS que segmenta unidades lingüísticas de las frases.

En los pocos trabajos que se han realizado hasta la actualidad sobre la aplicación de características glotales a la identificador de locutor se basaban en pasar al sistema toda la frase que producía el locutor y sobre ella extraer los parámetros glotales. Sin embargo esto no resulta del todo conveniente ya que aunque en principio pensamos que lo único que varía de pronunciar una 'a' de una 'e' es el tracto vocal, esto no es del todo cierto ya que la glotis y el tracto vocal forman parte de un conjunto que varía dinámicamente, que hacen diferentes esas características glotales que pretendemos extraer. Otra diferencia con esa investigación es que se hace una revisión de los nuevos algoritmos que han surgido en estos años para estimar el pulso glotal con más precisión, ya que cuanto mejor lo extraigamos más fácil y mejores características de él podremos obtener.

La novedad de aplicarlo a unidades lingüísticas tiene la ventaja de poder usarlo a lenguaje hablado. El lenguaje hablado es el que se produce de manera habitual entre dos personas cuando se comunican, a diferencia de otros estudios en los que el sistema del que disponen solo puede usarse para vocales sostenidas, en este proyecto se pueden extraer características glotales en situaciones de habla real, lo que permite aplicarlo a todo tipo de grabaciones de voz. El lenguaje hablado es más preciso a la hora de sacar características como acabo de comentar pero tiene el inconveniente de la corta duración temporal que nos va a acarrear problemas.

Por último como posibles trabajos futuros está la mejora de identificación de locutor a través de parámetros glotales hasta tasas de error similares a las que se obtienen con los métodos tradicional para que ambos sistemas de búsqueda se puedan fusionar.

# *Sección 2*

# *Conceptos básicos*

## 2.1 Introducción

La voz humana es quizás el mecanismo para comunicarse más potente y ubicuo que existe. Es usada por un amplia variedad de funciones en la interacción hablada, desde señalizar la prominencia en una pronunciación hasta la expresión de estados afectivos o actitudes. Durante varias décadas se ha tratado de construir sistemas de procesado de voz automáticos que puedan recrear o incluso sobrepasar la habilidad de los seres humanos para extracción de información y producción proveniente de la voz. Mientras que se han encontrado resultados de éxito en algunos tipos de información como por ejemplo en el contenido fonético, en otros campos como la cualidad vocal permanecen difíciles de analizar y reproducir.

Desde un punto de vista de la producción de voz , las propiedades de las señales de voz pueden estar divididas principalmente en dos categorías: efectos del tracto vocal, que están relacionadas con la forma de la cavidad nasal y oral en un determinado instante de tiempo y los efectos del pulso glotal, que están relacionados con los patrones característicos de la corriente de aire que atraviesa la glotis mientras las cuerdas vocales se abren y se cierran.
Estos dos componentes de la producción de voz tienen distintos papeles a la hora de recopilar distintos tipos de información contenida en las señal de voz. La identidad fonética, por ejemplo, está transmitida en muchos casos por los formantes que son una serie de resonancias que están controlados por la forma que adopte el tracto vocal. Para otros tipos de contenido en la voz, tales como identidad y cualidad vocal, ambos tracto vocal y componentes glotales contienen información relevante y complementaria, mientras que en el caso de estados de ánimo y emociones existe un gran interés en su relación con las variaciones del pulso glotal.

Aunque la naturaleza complementaria del tracto vocal y el pulso glotal han sido bien definidos, en la práctica, estos componentes a menudo no son explícitamente analizados como entidades separadas.
Una razón de esto es que estos componentes no se observan por separado, sino en combinación en la señal de voz, y la separación de dos señales desconocidas de una señal conocida permanece a pesar de suponer ciertas simplificaciones como un problema de deconvolución a ciegas. De hecho, la mayoría de las aplicaciones más utilizadas como reconocimiento de voz o identificador de locutor se han fijado no en el modelo de producción sino en uno perceptual, adoptando una figurada y suave representación espectral de la señal de voz que es afectada por ambos componentes tracto vocal y glotis y de la cual la información discriminante se extrae vía análisis de patrones.

Sin embargo, en parte debido a las aparentes limitaciones de las representaciones convencionales del espectro están emergiendo aplicaciones como análisis del estado de la voz , pero también debido al interés en comprender el comportamiento de la glotis, existe una nueva corriente de investigación que pretende caracterizar las componentes vocales en las señales de voz.

Este proyecto se centra en el análisis de herramientas para el modelado y la caracterización de ciertos aspectos deseados de la voz. Concretamente se fija en el efecto del pulso glotal que emana de la vibración de las cuerdas vocales, y su impacto en la señal de voz. Además, este proyecto observa las contribuciones de las variaciones dinámicas del pulso glotal y su efecto percibido en la cualidad vocal. Hay que tener en cuenta que este estudio no se extiende a voces cantadas, en cambio a lo que se dedica es a la efectiva caracterización de la variación del pulso glotal y la cualidad vocal en hablas sin patologías.

Una precisa caracterización acústica y modelado del pulso glotal es deseable para un gran rango de aplicaciones entre las que se encuentran herramientas de análisis, como características acústicas usadas para la tecnología de voz (por ejemplo: síntesis de habla, modificaciones de voz, identificación de locutor). Sin embargo el potencial de estas características acústicas relativas al pulso glotal y a la cualidad vocal no están completamente explotadas en estas áreas. Esto se debe principalmente a la falta de robustez en los métodos automáticos de análisis.

Por tanto un primer paso de este proyecto es recopilar los distintos algoritmos propuestos por distintos investigadores centrados en el tema que resulten más eficientes para nuestra tarea.

# 2.2 Modelo de producción de voz

Para el desarrollo de aplicaciones de procesamiento de voz útiles, es necesario entender cómo se produce la voz humana. La anatomía de nuestro mecanismo vocal determina la generación de los diferentes sonidos. En la figura 2.2.1 podemos observar a qué altura esta la glotis y en la figura 2.2.2 vemos una glotis vista desde arriba. El aire que exhalan los pulmones sube por la tráquea hasta atravesar la glotis que es el orificio. Las cuerdas vocales que delimitan a la glotis vibran a gran velocidad abriendo y cerrando este orificio. Como estas cuerdas vocales cierran y abren la glotis, y la forma de esta parte del cuerpo humano ya puede ser característico de cada persona. Por ejemplo hay personas que consiguen cerrar mas la glotis durante la vibración y en otras nunca llega a cerrarse del todo son personas que tenderían a hablar de manera más aspirada que como vamos a ver próximamente será la cualidad vocal llamada *breathy*. Podría haber diferencias en el tamaño, masa de esta glotis, así como también en cada locutor puede variar la forma de abrirse y de cerrarse (hay gente que se le cierra como una cremallera). Otro distintivo de cada locutor podría ser ese frecuencia de apertura y cierra que se conoce como 'pitch' tono, y que en los hombres esta alrededor de 100Hz y en las mujeres alrededor de 200Hz. La epiglotis es un cartílago que actúa como tabique para evitar que la comida se vaya al sistema respiratorio.

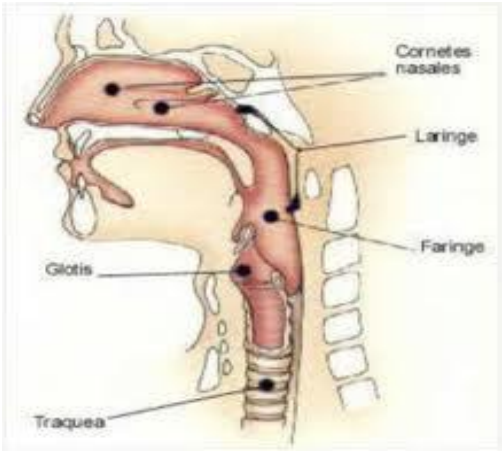



***Figura 2.2.1:*** Cartílago y órganos que intervienen en el proceso de producción

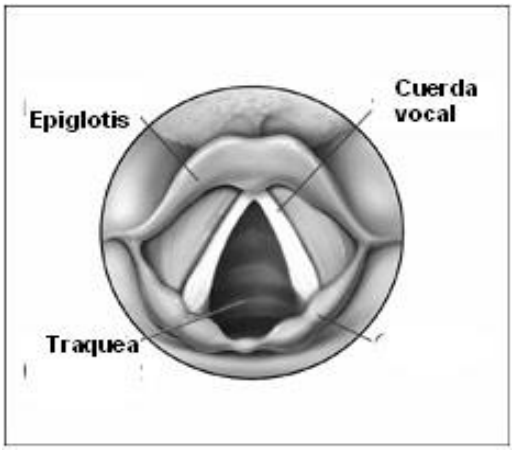



***Figura 2.2.2:*** Corte transversal a la altura de la glotis

La figura 2.2.3 muestra una representación esquemática del proceso de producción de voz para dos vocales 'u' e 'i'. El espectro del pulso glotal es idéntico en ambos casos: Contiene todos los componentes harmónicos y tiene una pendiente constante de -12 dB por octava .

Esto significa que la amplitud de los harmónicos disminuye monótonamente con el aumento frecuencia , de modo que para cada vez que duplicamos la frecuencia , la amplitud se ha reducido en 12 dB . Debemos tener en cuenta que este es el caso ideal, el verdadero espectro del pulso glotal no tiene una pendiente constante , y puede presentar valles dependiendo de la forma del pulso glotal debido a distintos locutores, fonemas, estados de ánimo....

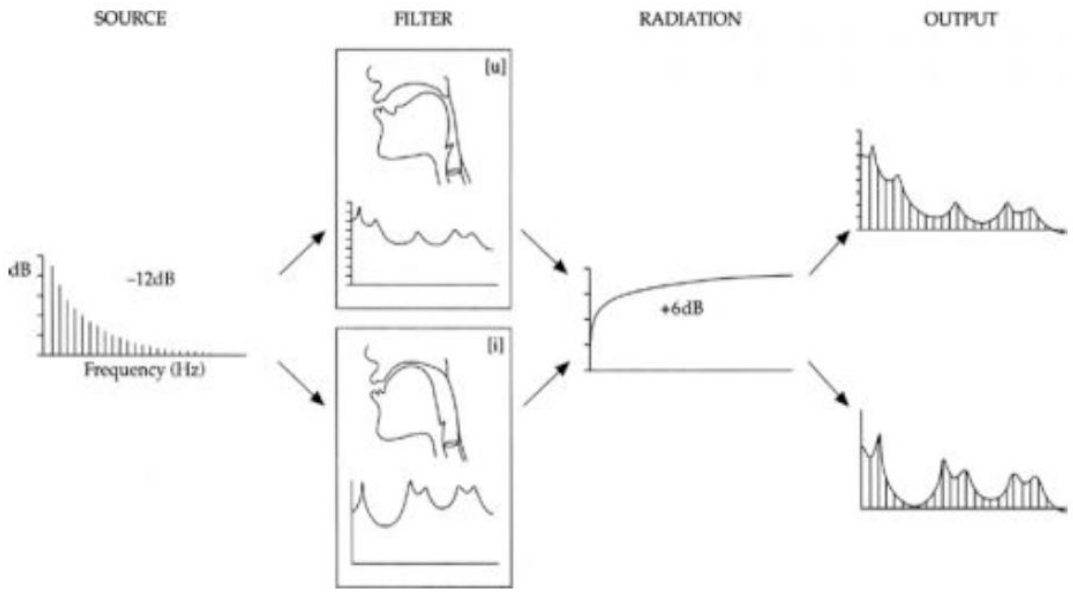



***Figura 2.2.3:*** Representación esquemática del modelo de producción de voz, extraído de [Will 10]

El efecto de filtrado del tracto vocal (referida como la función de transferencia) es bastante diferente para los dos vocales, debido a las diferentes posiciones adoptadas por la boca, la lengua y la mandíbula. Los harmónicos del pulso glotal que caen cerca de los picos de la función de transferencia serán amplificados por el filtro. Por otra parte, 5 los harmónicos que caen en los valles se atenúan. En consecuencia , la salida del filtro , es decir , la señal de voz final, tiene un espectro que presenta picos y valles en contra posición con el espectro del pulso glotal que tiene una caída más uniforme. El filtro del tracto vocal que queda determinado por los formantes y marcan la diferencia en este caso la diferencia entre [ u] y [ i]) .

Por último la influencia de los labios en el modelo, los labios son como un filtro que tiene un espectro con una pendiente de aproximadamente 6 dB por octava, es decir tendría una respuesta en frecuencia que amplifica las altas frecuencias, se trataría de un derivador de primer orden en el tiempo.

El efecto de los labios siempre puede ser añadido o cancelado , metiendo un filtro con un cero en cero para derivar o con un polo en cero para integrar respectivamente. La ilustración de la figura 2.2.3 es de una fuente idealizada que se supone que es constante para los dos vocales.

En el habla real, el pulso glotal varía dinámicamente de manera que refleja la configuración de la glotis, el grado de cualquier tensión en la laringe que puede estar presente y que se utiliza el esfuerzo respiratorio. El pulso glotal puede variar en el curso del habla. La variación se produce si el hablante elige para cambiar entre las diferentes cualidades vocales como por ejemplo voz carrasposa, como es a menudo en la señalización paralingüística de la emoción y la actitud. Diferentes locutores pueden también variar considerablemente en función del tipo cualidad vocal que utilizan habitualmente.

En cuanto al pulso glotal, se conoce bastante acerca de la variación f0, y cómo varía en función de la entonación, el tono y el estrés. Se sabe relativamente poco acerca de otros aspectos de la fuente de la voz y la forma en que varía en el habla. Por supuesto, hay también muchos estudios sobre la variación de la intensidad de la señal de voz. Aunque la amplitud de la salida de voz, en cierta medida refleja la amplitud de la del pulso glotal, uno debe tener en cuenta que la amplitud total de la salida de voz es una función del pulso glotal y del filtro.

Por tanto suponiendo este modelo se puede separar la influencia de tracto vocal de la del pulso glotal. La mayoría de los estudios experimentales del pulso glotal se han basado en el filtrado inverso . Esta técnica es efectivamente una inversión del proceso de la producción de voz. La señal de voz se pasa a través de un filtro cuya función de transferencia es la inversa de la función de transferencia del tracto vocal. En principio, esto produce el pulso glotal ya que se cancela el efecto de filtrado del tracto vocal . La figura 2.2.4 ilustra este proceso en el dominio de la frecuencia y en el dominio del tiempo. La cancelación de la radiación de labios no se muestra aquí , por razones que se explican a continuación.

El filtro inverso debe contener una especificación de las frecuencias y anchos de banda de los antiresonadores necesarios para cancelar los formantes (polos complejos conjugados ) de la función de transferencia del tracto vocal en un momento dado en el tiempo. Es importante para obtener el número de polos adecuado para el ancho de banda determinado por la frecuencia de muestreo. La separación media entre los polos se determina por la longitud del tracto vocal, para un hombre típico con un tracto vocal de 17,5 cm podemos esperar uno de formantes de media por 1000 Hz . La especificación de la frecuencia precisa y ancho de banda es muy crítico para los

formantes inferiores, especialmente F1. Cualquier error aquí se traducirá en una distorsión del pulso glotal. Los errores menores en los formantes superiores tienen poco efecto sobre la forma del pulso principal o su correspondiente espectro de frecuencia.

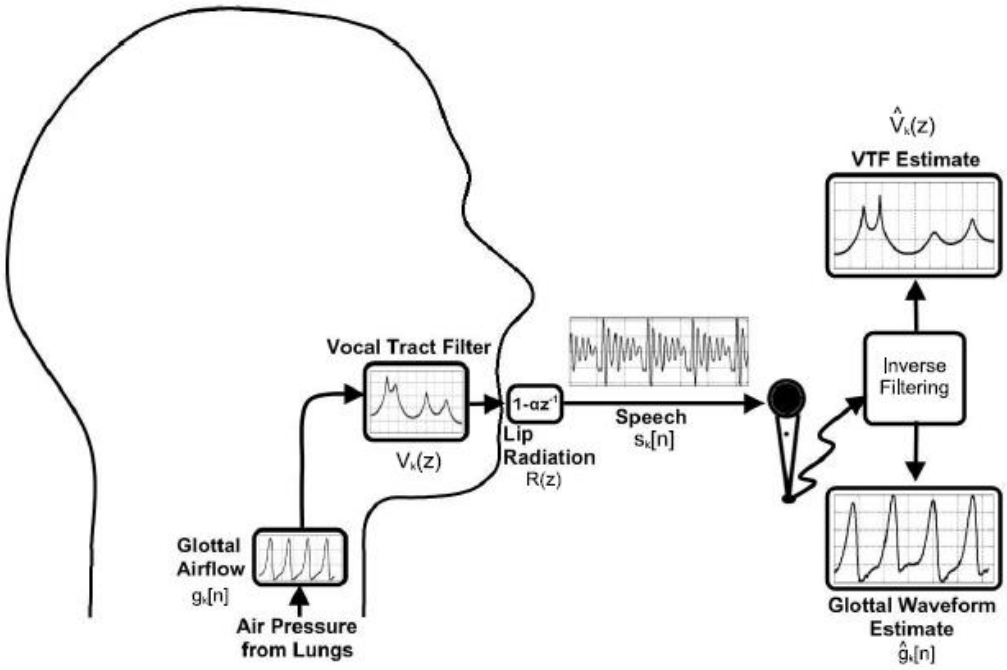



***Figura 2.2.4:*** Modelado del proceso de producción de voz, extraído de [Kan PhD 12]

# 2.3 Cualidades vocales

Se puede definir como ``la configuración del sistema vocal habitual combinada con los cambios dinámicos en el sistema utilizados para propósitos comunicativos´´. La clasificación que establecemos en este trabajo es la siguiente:

Voz modal (*modal*), voz susurrada (*whispery*), voz tensa (*tense*), voz carrasposa (*'creaky'*), voz relajada (*lax*), voz aspirada (*Breathy*), voz farseto (*falsetto*) y voz áspera (*harsh*).

Las cualidades vocales se pueden clasificar en grupos. El primer grupo lo formarían *modal* y *falsetto*, en este grupo se pueden combinar de manera individual con miembros de otros grupos pero no entre sí. El segundo grupo, *whispery* y *'creak'*, pueden aparecer solos o como un tipo compuesto *whispery 'creak'*. Y el tercer grupo está formado por *harsh, tense y breathy*, en este grupo estas cualidades vocales siempre aparecen junto a otras y nunca de manera individual.

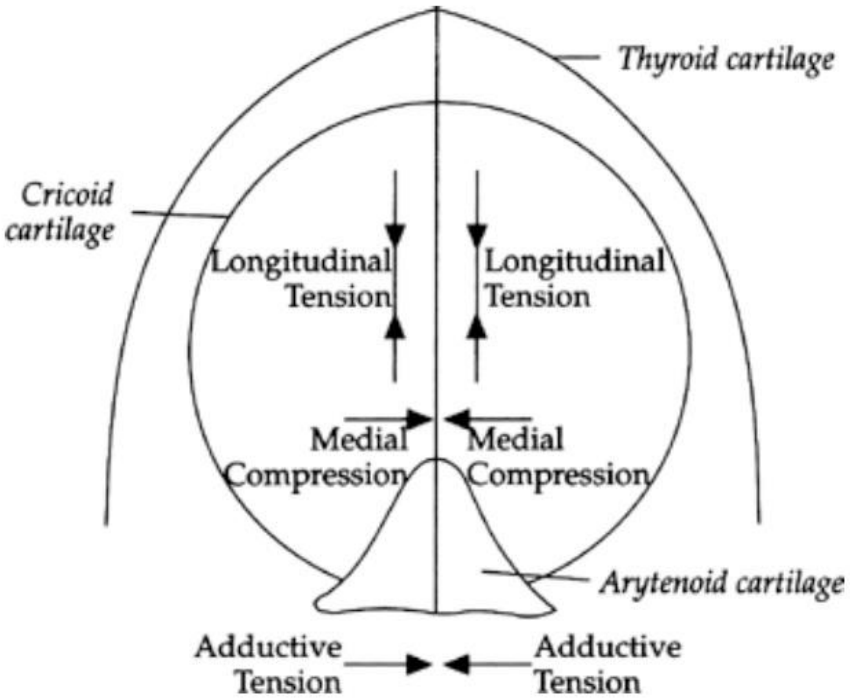



***Figura 2.3.1:*** Esquema de los músculos que intervienen en el proceso de producción de voz, extraído de [Laver 09]

En la figura 2.3.1 se pueden observar las descripciones fisiológicas aquí están en función de tres parámetros hipotéticos de la tensión muscular, la *'adductive tension*', *'medial compression*' y la *'longitudinal tension'.* Éstas determinan los valores de configuración y la tensión de las cuerdas vocales, e interactúan con los factores aerodinámicos relacionados con la presión subglótica y el flujo de aire glotal para producir una variedad de cualidades vocales.

*'Adductive tension'* se define como la fuerza por la cual las aritenoides se atraen, de manera que se produce la aducción del glotis cartilaginoso. Es controlado por los músculos del interaritenoidea .

*'Medial compression'* se define como la fuerza por la cual la glotis se cierra , a través de la aproximación de los procesos de vocales de los aritenoides. Está controlada principalmente por el músculo cricoaritenoideo lateral, pero el músculo tiroaritenoideo externa también puede estar involucrado.

*'Longitudinal tension'* es la tensión de las cuerdas vocales, y está controlada principalmente por la contracción de los músculos cricotiroideos, cuya función principal es el control de pitch.

A continuación se explicarán brevemente cada uno de los tipos de cualidad vocal comentada sus características primordiales.

**Voz modal ('*modal'*)**

Voz modal es el tipo de cualidad vocal normal. Se describe como la forma más eficiente de cualidad vocal ya que utiliza una *'adductive tension'*, *'medial compression*' y la *'longitudinal tension'* moderada. Aumento de la tensión longitudinal se utiliza principalmente para aumentar el pitch. La vibración de las cuerdas vocales son a menudo casi-periódica, con fricación mínima y cierre completo. Otros tipos de fonación se describirán en referencia a la presente descripción de voz modal.

**Voz susurrada*('whispery')***

La cualidad vocal de '*Whispery*' (susurro) se produce cuando la glotis se abre triangularmente con forma de Y invertida. Logra esta forma ya que tiene una baja 'adductive *tension*' y una *'medial compression'* de moderada a alta. Tanto el flujo de aire constante a través de esta forma triangular como la turbulencia resultante da esa sensación de susurro a la voz.

**Voz aspirada ('*breathy'*)**

La cualidad vocal *breathy* es, para que nos hagamos una idea, una voz aspirada. Durante esta cualidad vocal las cuerdas vocales vibran de manera mucho menos eficiente acompañada de una vibración con una audible fricción. En términos de tensión muscular existe una mínima *'adductive tension'* y una baja *'medial compression*'. La glotis no llega a cerrarse completamente por lo que provoca una constante turbulencia del paso de aire a través de la glotis.

**Voz tensa ('*tense'*)**

En la voz tensa todas las tensiones musculares tienen un carácter elevado durante todo el periodo que dure esta cualidad vocal. Por este motivo se la puede confundir con el tipo '*Harsh*', sin embargo la voz tensa no llega a valores tan extremos ni está caracterizada por las irregularidades propias de la cualidad vocal de 'Harsh'. En comparación con la voz modal tiene unas tensiones más elevadas.

**Voz áspera ('*harsh'*)**

Este tipo de cualidad vocal se produce cuando hay una tensión excesiva en las cuerdas vocales que viene acompañada de un pitch muy bajo. Pero hay otras cosas que la determinan como son sus irregularidades en términos de frecuencia y amplitud entre pulsos bastante cercanos. Estas fluctuaciones de amplitud es lo que le dan esa sensación al oído de voz áspera.

**Voz carrasposa (*'creaky'*)**

La cualidad vocal de *'creaky'* (voz carrasposa) tiene una *'adductice tension'* y *'medial compression'* fuertes y una baja *'longitudinal tension'* comparada con la voz modal. Típicamente se produce con una frecuencia fundamental menor y durante esta cualidad vocal solo la parte anterior de las cuerdas vocales vibra mientras que la parte posterior las cuerdas vocales se mantienen juntas. La sensación que da esta cualidad vocal es como que existen impulsos adicionales en la voz.

**Voz faseto ('*falsetto'*)**

falsetto es un tipo de cualidad que implica tensiones más fuertes de *'adductive tension'*, *'medial compression'* y *'longitudinal tension'* en comparación con la voz modal. Quizás debido a la mayor *'longitudinal tension'*, las cuerdas vocales se vuelven

delgados y menor de nivel de presión sub-glótica es aplicado en comparación con la voz modal. Esta cualidad vocal se asocia con elevados valores de pitch que se da en pulsos glotales muy cortos.

Además de la configuración que tenga habitualmente esta puede variar por los distintos factores: lingüísticos, paralingüísticos, sociolingüísticos y extralingüísticos.

Lingüísticos, la cualidad vocal debido a la posición del fonema dentro de la frase así como fonemas que puede tener delante o detrás, por ejemplo mucha gente al final de la frase hace *breathy* aspiración.

El factor paralingüístico es debido a los diferentes estados de ánimo y humor que puede tener en un momento determinado el locutor, este cambia la entonación, aunque de manera universal tiende a utilizarse *breathy* para momentos íntimos, *whispery* para temas confidenciales y *harsh* para momentos de enfado. Luego hay diferencias en función de la cultura por ejemplo en la cualidad vocal de *'creaky'* está asociada para ciertos locutores ingleses como aburrida, mientras que en Tzeltal (una lengua maya) está relacionada con una queja.

La cualidad vocal también puede tener una función sociolingüística, que sirve para diferenciar entre los grupos lingüísticos, regionales y sociales. Como cualquier persona que tenga enseñar un idioma extranjero tendrá constancia de que las diferencias entre idiomas en la cualidad vocal son un aspecto importante de un acento convincente, pero difícil de enseñar, ya que casi nunca se describen en la lingüística o aplican literatura lingüística. Esto puede llevar a errores de percepción culturales, ya que el hablante nativo es probable que interprete la cualidad vocal del extranjero en términos de su propio sistema paralingüísticos. Por poner un ejemplo, en Edimburgo la voz chirriante se asocia a un estatus social más alto, mientras que las susurrantes y duras están vinculados a un estatus social más bajo.

Hay otros factores, extralingüísticos, que determinan la cualidad vocal, muchos de ellos fuera del control del locutor. Las diferencias en el tamaño, la forma y el tono muscular de las estructuras de la laringe juegan un papel importante. Las voces de hombres, mujeres y niños en su mayoría reflejan diferencias anatómicas. También está afectada por la salud mental y física de la persona.

# 2.4 Definición y fases del pulso glotal

En la figura 2.4.1 vemos como sería un pulso glotal ideal en la parte de arriba y abajo su derivada, ya podemos incluir o no el efecto de los labios. Se pueden apreciar tres fases en la que está el pulso glotal *openig, closing y return*.

La fase de apertura (*openning)* se define como la parte del pulso glotal en que las cuerdas vocales se abren y se incrementa la corriente de aire que atraviesa la glotis. Esto llega hasta el punto 'Tp' donde las cuerdas vocales están en su máxima amplitud.

Después comienzan a cerrarse en la fase de cierre (*closing phase*) que llega hasta el GCI (*glottal closure instant*) que es instante donde las cuerdas donde se cierran más rápidamente. Por eso en la derivada del pulso glotal coincide con el instante temporal en que la función tiene un mínimo. Hay que recordar este punto ya que durante todo el proyecto se hablará de él, es un punto muy característico, en el grafico se corresponde con Te.

Por último la fase de retorno es la que continua desde el GCI hasta que se cierra la glotis completamente.

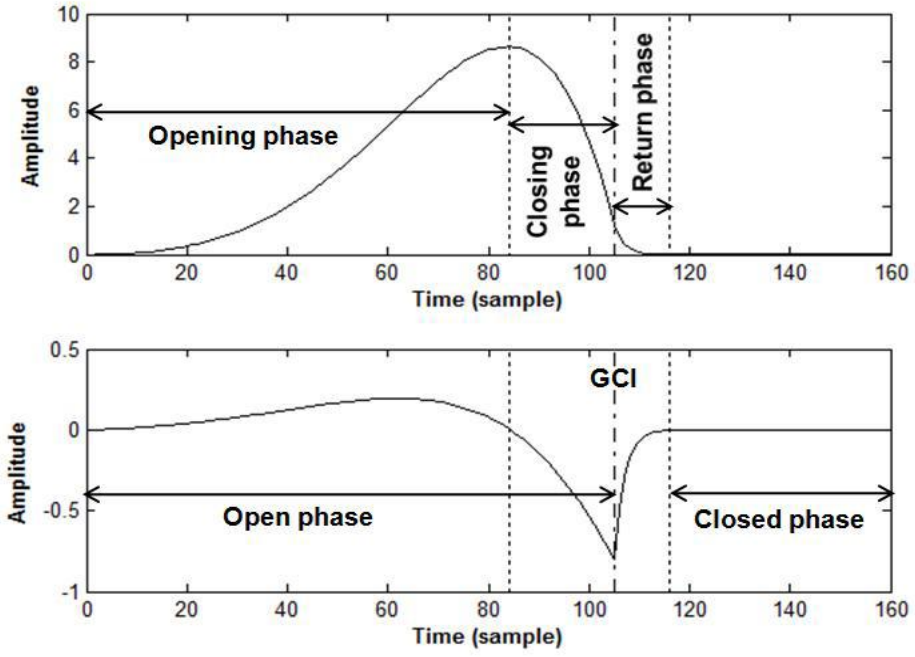



***Figura 2.4.1:*** Pulso glotal y derivada del pulso glotal según el modelo LF, extraído de [Drug 12]

Muchos investigadores optan por trabajar con esta señal en lugar del verdadero pulso glotal. El énfasis de las frecuencias más altas en 6 dB por octava permite un modelado

más preciso de la pendiente espectral de la señal fuente. También es conveniente para los propósitos de resíntesis ya que al tener unido el pulso glotal a la influencia de los labios no se necesita eliminarla y para luego reintroducirlo. Otros de los motivos por lo que se utiliza la derivada del pulso glotal son que se aprecia más fácilmente el GCI, el modelo con el que vamos a trabajar del pulso glotal tiene muchos parámetros definidos sobre esta derivada y además el énfasis sobre las altas frecuencias que supone la derivada permite un modelado más preciso de la pendiente del espectro del pulso glotal. Por lo que ahora en adelante, cuando digamos pulso glotal nos referiremos indistintamente a él o a su derivada, aunque en general lo que se va a obtener con más frecuencia y con lo más se va a trabajar será la derivada del pulso glotal.

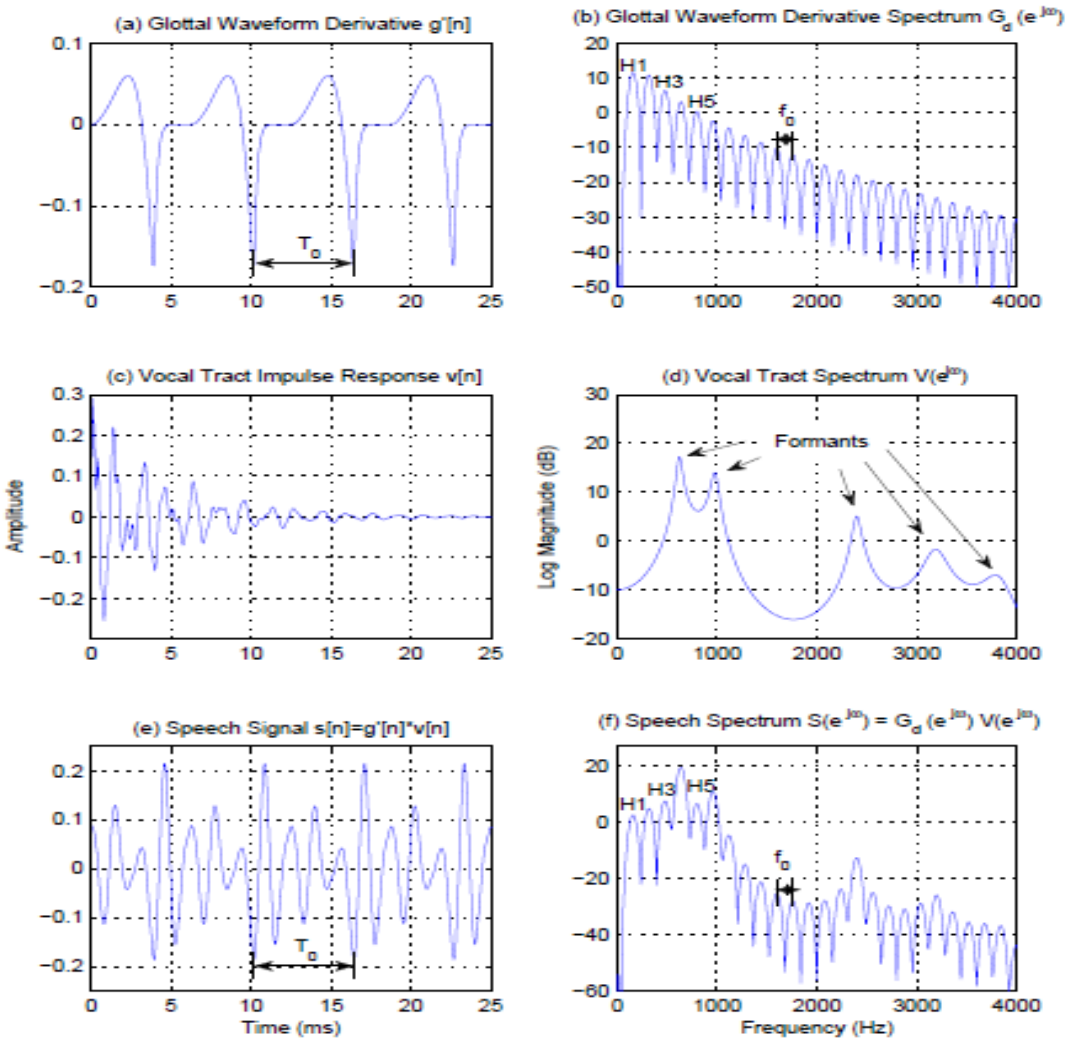



***Figura 2.4.2:*** Derivada del pulso glotal en tiempo y frecuencia, diferenciando entre la influencia de la glotis y del tracto vocal, extraído de [Haou 13]

# 2.5 Señal residual

Antes de comenzar a hablar del pulso glotal tenemos que comentar en qué consiste la siguiente herramienta matemática que de ahora en adelante saldrá a menudo. Esta señal no es el pulso glotal pero si contiene información glotal que puede ser utilizada para sacar el pulso glotal. En la primera imagen la figura 2.5.1 vemos la señal temporal del fonema 'aa' y debajo su espectro en la 2.5.2. Con un LPC(lineal predictive coding) de orden entre 12 y 18 sacamos la envolvente del espectro, si ponemos un orden menor la curva de la envolvente la forzaríamos a ser casi recta y no se ajustaría tan bien, queremos extraer una envolvente. El LPC saca los coeficientes de un filtro FIR (solo ceros) para sacar estos coeficientes solo tiene en cuenta la muestra actual y las anteriores. Por tanto con el LPC y orden adecuado obtenemos la envolvente que se muestra en la figura 2.5.2 como la función en rojo.

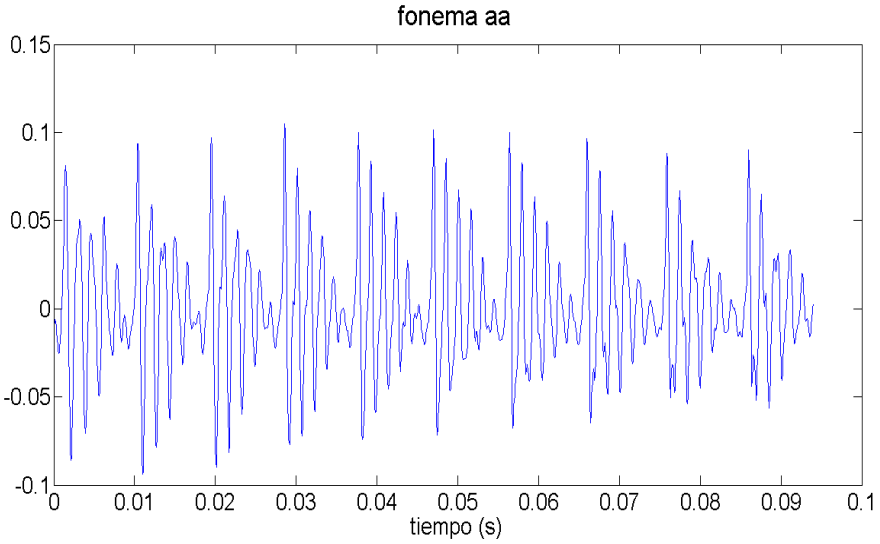



***Figura 2.5.1:*** Señal de voz del fonema “aa” en el tiempo

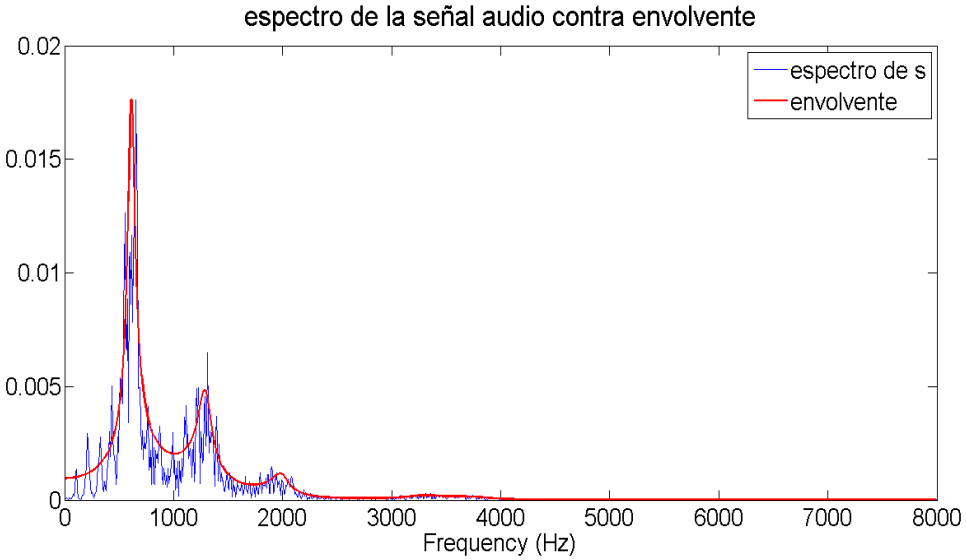



***Figura 2.5.2:*** Espectro del fonema “aa”

Con estos coeficientes hacemos el filtro inverso (solo polos) y pasamos el filtro a la señal cancelando la envolvente de su espectro. En la figura 2.5.3 se ve la envolvente del espectro sacado con el LPC y el filtro inverso en color rojo.

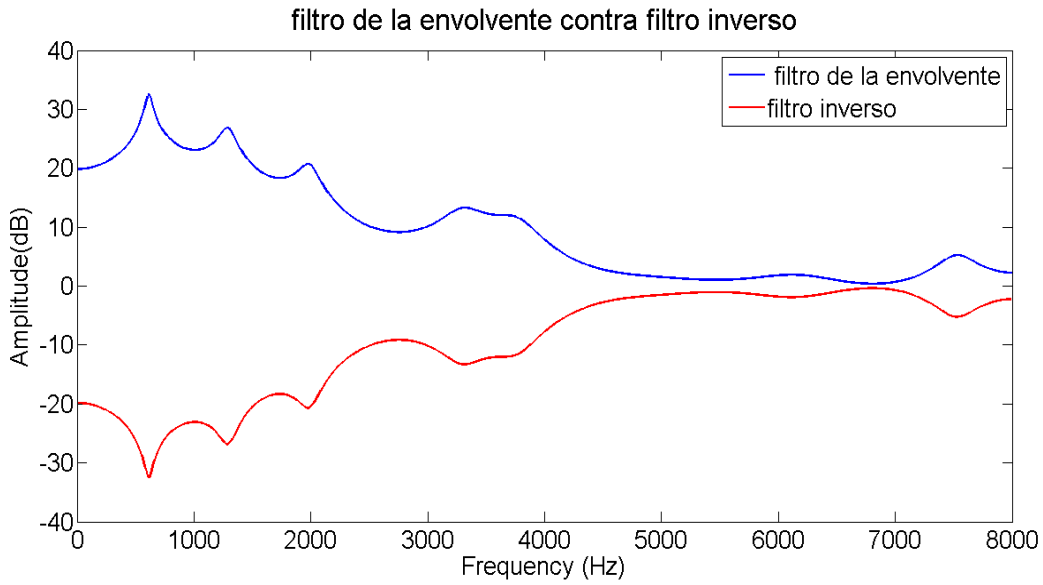



***Figura 2.5.3:*** Envolvente del espectro de la señal de voz y del filtro inverso

En la figura 2.5.4 vemos la señal residual t en tiempo. ¿Por qué esto no es la señal glotal? Porque en esta envolvente del espectro hay contribución tanto del tracto vocal como del pulso glotal, y cancelando esa envolvente también cancelamos parte del pulso glotal. Si pintáramos el espectro de la señal residual tendría una envolvente plana ya que eso es lo que hemos cancelado de la señal original. Estos picos positivos en la señal residual muestran discontinuidades en la señal de voz que se van a corresponder con los instantes en los que se encuentran los GCI, aunque habrá que utilizar un algoritmo más avanzado para localizarlos de manera más precisa.

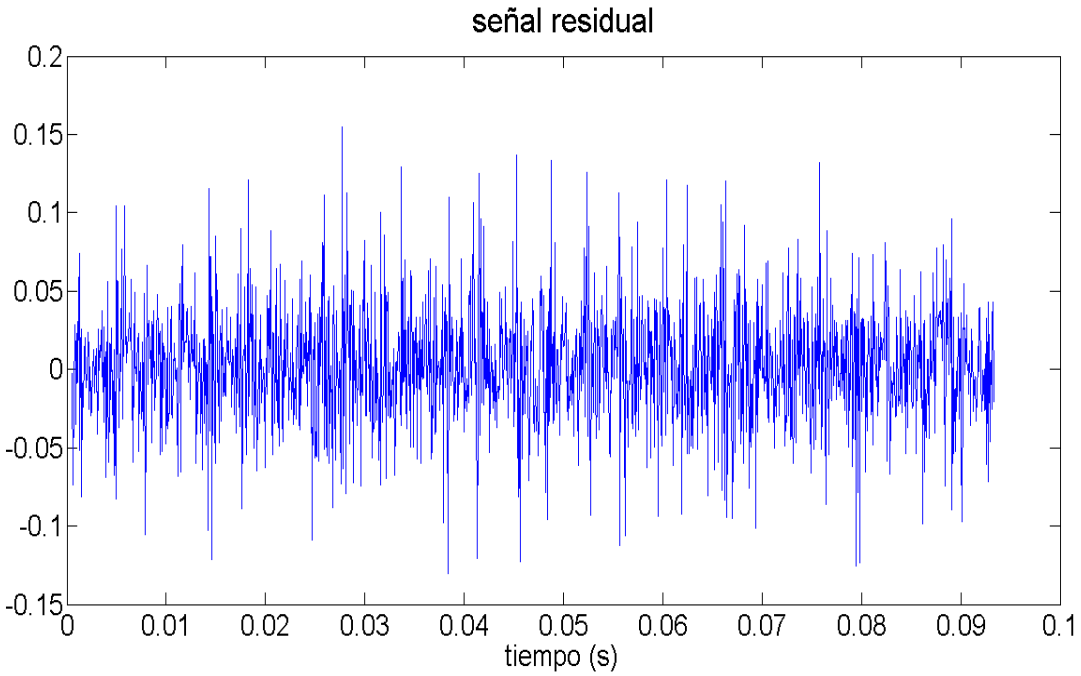



***Figura 2.5.4:*** Señal residual en el tiempo

# *Sección 3*

# *Estado del arte sobre la estimación del pulso glotal*

## 3.1 Introducción

En esta sección hablaremos sobre como estimar el pulso glotal de una señal de voz para lo cual necesitaremos una serie de pasos y requerimientos previos. Como se observará más adelante, esto no es una tarea trivial para lo cual utilizaremos tanto código libre como distintas publicaciones de los investigadores que están actualmente centrados en este tema, los cuales son Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio y Stefan Scherer.

La sección empieza viendo el esquema de todo el sistema. Más tarde desarrolla la siguiente organización en cada una de las etapas: explicación de la utilidad de la etapa, una comparación de los distintos algoritmos existentes hasta la fecha, y por último una breve explicación del algoritmo escogido.

# 3.2 Esquema del sistema

En este esquema se muestran las diferentes etapas por las que debe de pasar la señal de audio para la estimación del pulso glotal. Estas etapas incluyen una segmentación para la obtención de fonemas individuales, una detección de actividad de voz, ya que para sonidos sordos no se pueden sacar características glotales, un detector de pitch, un detector de polaridad para orientar la señal de modo adecuado y un detector de instantes temporales donde la glotis permanece cerrada. Estos instantes le permitirán al último algóritmo, el cual realiza la estimación del pulso glotal en sí, una estimción síncrona, es decir pulso a pulso.

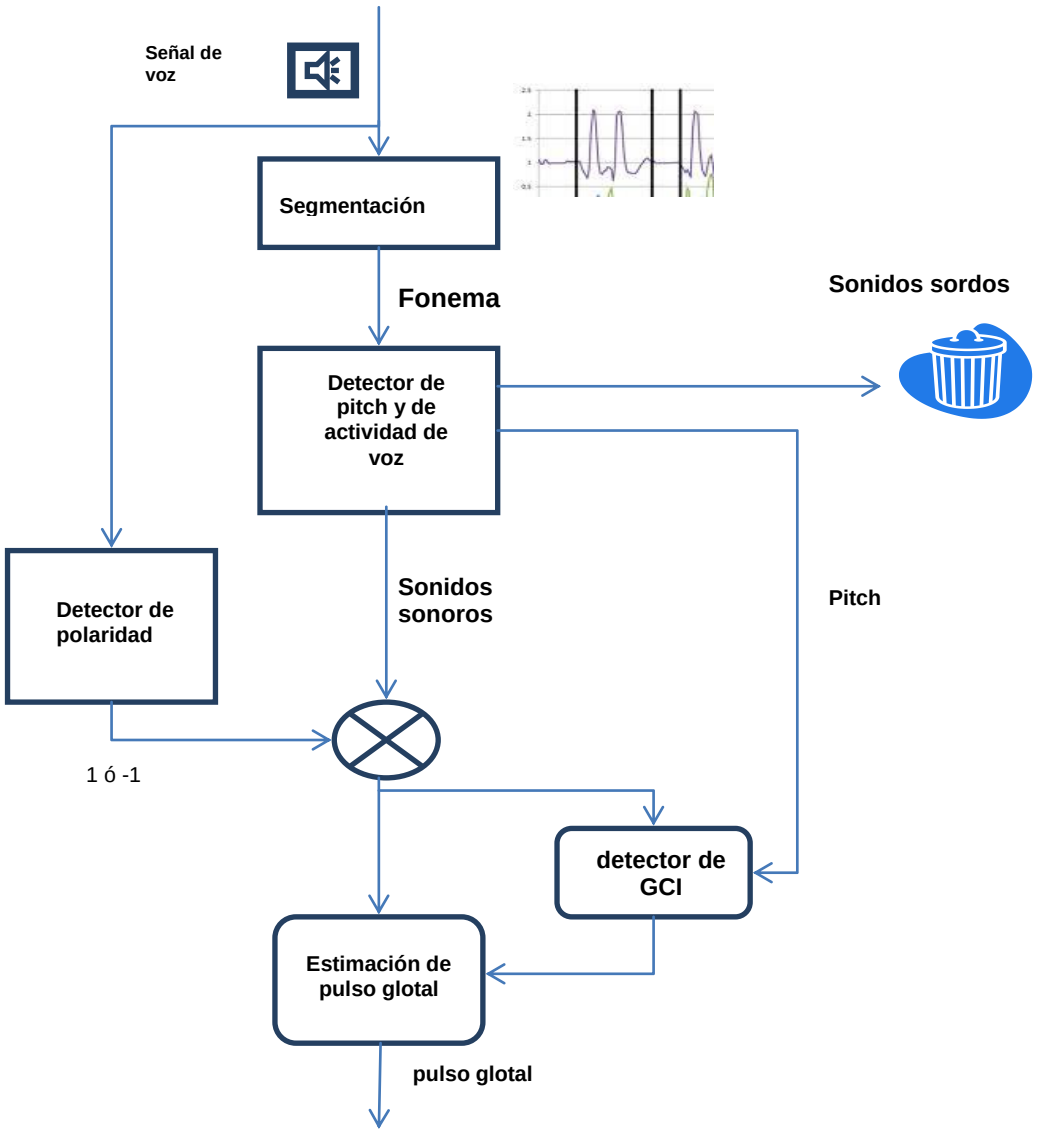



***Figura* 3.2.1:** Esquema del sistema de estimación del pulso glotal

# 3.3 Segmentación

Aunque la extracción de características glotales en principio se puede hacer para toda la frase, es más correcto realizarlo sobre unidades lingüísticas, ya que estos parámetros glotales que se sitúan dentro de una cualidad vocal pueden variar dentro de una misma conversación incluso dentro de una misma frase. Para ser todavía más precisos en la extracción de estas características nos centraremos en la extracción de fonemas de sonidos vocales ya que son lo que más duración y estabilidad van a tener.

El laboratorio ATVS cuenta con un sistema que segmenta, es decir, divide las locuciones en fonemas diciendo cuales son. Con añadir este sistema al principio del nuestro ya podremos aplicarlo a la unidad lingüística que más nos convenga en bases de datos sin etiquetar.

# 3.4 Detector de pitch y de actividad de voz

## *3.4.1. Utilidad*

Antes de estimar el pulso glotal hay que tener en cuenta dos cuestiones. La primera como es evidente es que vibren las cuerdas vocales para que se puedan extraer características glotales, por lo tanto tenemos que distinguir entre sonidos sordos y sonidos con voz, una forma de hacer esto es mirando la etiqueta del fonema, así fonemas como 'k' ,'t' serán sordos , sin embargo estaremos pasando por alto que fonemas como 'aa' pueden decirse de forma que no haya vibración en las cuerdas vocales por lo tanto hay que usar un pitch detector que distinguirá entre sonidos sordos y sonidos con voz, de manera que los sonidos sordos los descartemos.

Además el Pitch Detector nos dará un valor de pitch (frecuencia fundamental del pulso glotal) que se utilizará en posteriores funciones para calcular la longitud de sus ventanas en concreto en el algoritmo que se utiliza para posicionar los GCI que veremos en esta sección más adelante.

## *3.4.2. Comparativa*

Antes de empezar a comparar vamos a definir una medidas de error. VDE (*Voice decision error* ) es la proporción de tramas sobre las que se produce un error a la hora de decidir si es un sonido sordo o no. GPE (*Gross Pitch Error*) es la proporción de

tramas dentro de las que se decide sonidos sonoros de las cuales se comente un error relativo en la detección del pitch de más de un 20%. Por último FFE (*F0 Frame Error*) es la proporción de tramas sobre las que se produce un error ya sea por el VDE o por el GPE. FFE puede ser visto como una medida de evaluación del rendimiento global de *Pitch detector-Voice decision*.

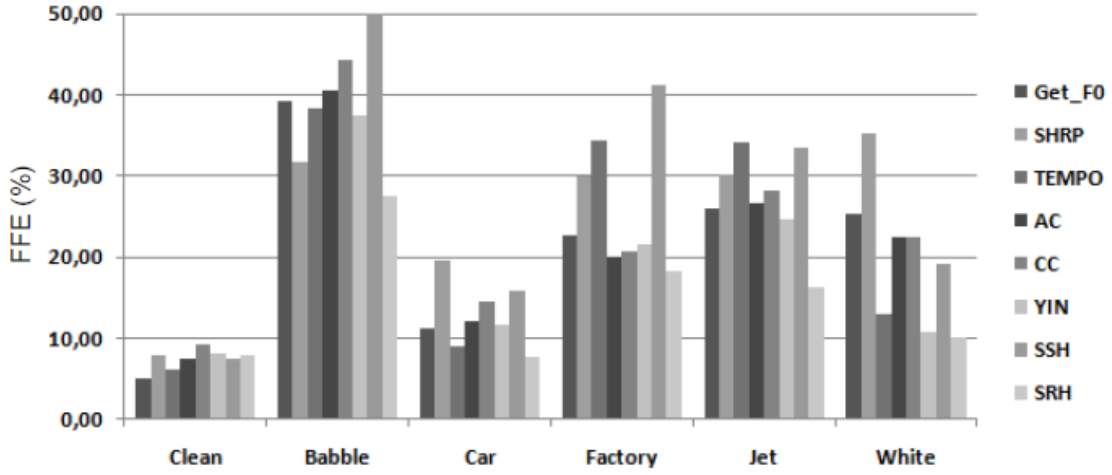



***Figura* 3.4.2:** Comparativa que muestra el porcentaje de error global de distintos pitch detector,extraído de [Drug 10]

En esta figura se puede observar la justificación del detector de pitch que he escogido llamado SRH (*Summation of Residual Harmonics*). Cada una de las barras se corresponde con un algoritmo distinto etiquetado en la derecha de la imagen siendo el ultimo el SRH el que se corresponde con la barra de más a la derecha en cada uno de los distintos ruidos. Como se ve en la comparativa en condiciones sin ruido están todos bastante a la par mientras que con distintos tipos de ruido el que mejor resultado da es el SRH. Como el propósito final de esta investigación es que estas características se puedan extraer de habla real, los criterios para seleccionar cada uno de los algoritmos son que sea lo más robusto y estable frente al ruido.

## *3.4.3. SRH (Summation of Residual Harmonics)*

El SRH usa armónicos de la señal residual, E(f) es el espectro de la señal residual. La frecuencia donde la función SRH alcanza su máximo valor es el pitch, y el valor máximo de esta función se compara con un umbral para determinar si el sonido es sordo o no.

SRHmax ⟹ Decide entre sonidos sordos o sonoros

SRH(f0)max ⟹ f0 pitch

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

*E(f) espectro de la señal residual*

Considerando sólo el término E (k · f) en la suma, esta ecuación toma la contribución de los primeros armónicos $N_{harm}$ en cuenta. Entonces se podría esperar que esta expresión alcanzara un máximo para f = F0. Sin embargo, esto también es cierto para los armónicos presentes en el rango [F0 min, F0 max]. Por esta razón, la sustracción por E ((k-1/2) · f) permite reducir significativamente la importancia relativa de los máximos de la función SRH en los armónicos pares.

El detector funciona en su punto más óptimo usando ventanas de 100ms. Entonces podemos hacer dos cosas no usar fonemas de menos de 100ms o permitir en el sistema fonemas de duraciones inferiores y asumir el error. Yo he decidido usar la primera opción aunque haya fonemas que no puedan ser utilizados. Aunque en experimentos posteriores pruebo a reducir esta ventana para cuantificar qué error se produce en la decisión final.

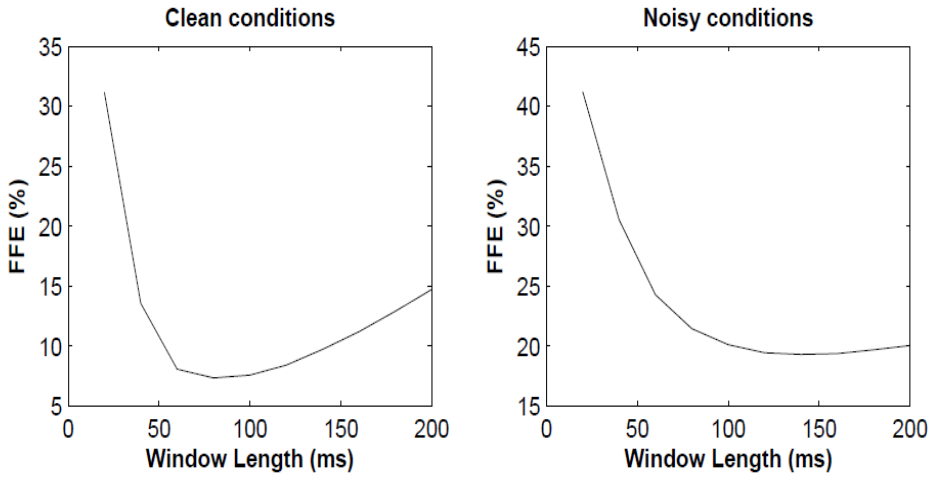



***Figura* 3.4.3:** Porcentaje del error global SRH en función de la longitud de la ventana, extraído de [Drug 10]

# 3.5 Detector de polaridad

## *3.5.1. Utilidad*

La detección de la polaridad de un señal de voz de manera correcta es un paso previo necesario para varias técnicas de procesamiento de voz. Un error en su determinación podría tener un impacto negativo en su rendimiento. Dado que los sistemas actuales tienen que lidiar con gran cantidades de datos procedentes de múltiples dispositivos, la detección automática de la polaridad del habla se ha convertido en un problema crucial. Para este fin, se propone aquí un algoritmo muy sencillo basado en la asimetría de las dos señales.

Cuando un micrófono es usado en la grabación de un locutor la inversión de conexiones eléctricas en el micrófono puede causar la inversión de la polaridad. El oído humano es insensible a este cambio de polaridad sin embargo si no se halla bien puede tener consecuencias graves en este procesado de audio. El origen de la polaridad en el pulso glotal se deriva de la asimetría del pulso glotal al excitar las resonancias del pulso glotal. Durante la producción de sonidos sonoros la derivada del pulso glotal muestra una discontinuidad que se produce de manera periódica llamada *Glottal Closure Instant* (GCI). Como se describe en los modelos de pulso glotal se dice que la polaridad es positiva si la discontinuidad presenta un pico negativo en el GCI, de lo contrario si presenta un pico positivo en el GCI la polaridad es positiva.

Por tanto la polaridad puede tener dos valores -1 y 1, y tras obtenerla se multiplica por la señal. Cuando la señal se multiplica por -1 la señal se invierte con respecto del eje x, y si se multiplica por +1 se queda tal y como está. El objetivo de esta multiplicación es que en los GCI de la derivada del pulso glotal siempre se dé un mínimo. En este ejemplo particular en la gráfica se ve la derivada de un pulso glotal de la misma señal, en la primera gráfica se supone polaridad -1, por lo que la derivada del pulso glotal ha sido multiplicada por -1 , y en la segunda imagen se supone polaridad +1 por lo que la señal se ha quedado como estaba. Por lo que de las dos imágenes se puede deducir que para esta señal concreta la polaridad es -1 porque es cuando en los GCI se da un mínimo. Si no se saca la polaridad correcta en el posterior algoritmo de extracción de parámetros obtendrá parámetros erróneos.

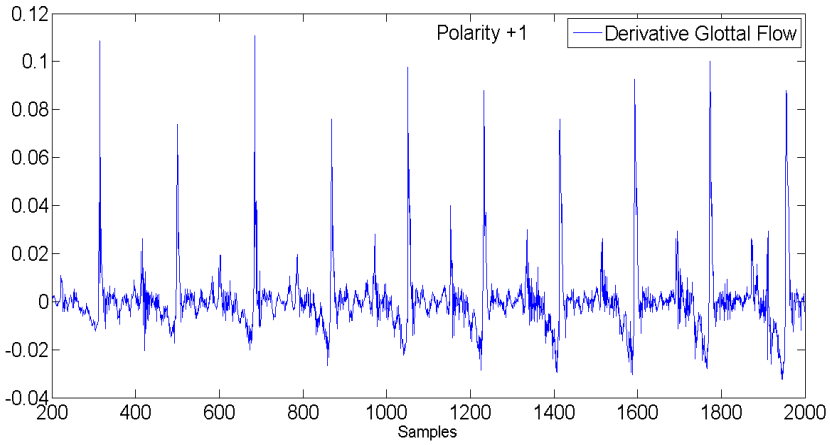



***Figura* 3.*5.1:*** Derivada del pulso glotal con polaridad mal extraída

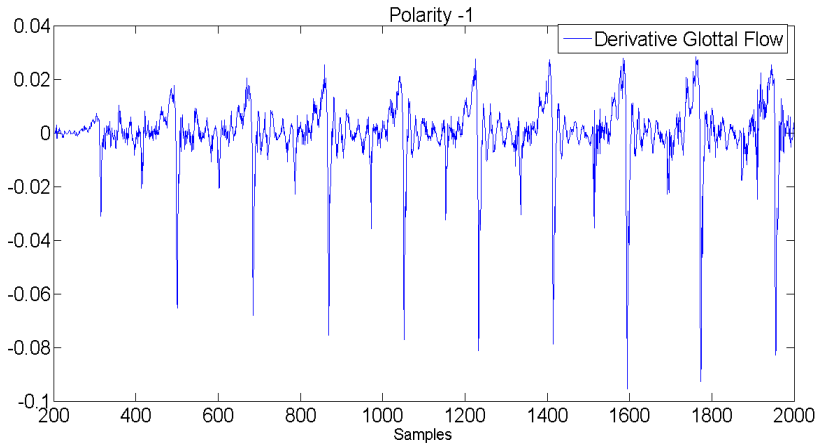



***Figura* 3.*5.2:*** Derivada del pulso glotal con polaridad bien extraída

## *3.5.2. Comparativa*

Ahora en la figura 3.5.3 vemos una comparación de los distintos algoritmos que existen. El que mejor rendimiento da es el que está marcado con una línea roja llamado RESKEW(*residual excitation skewness*). También es el más novedoso de 2013 y vuelve a utilizar la señal residual. "*skewness*" en inglés se puede traducir como el grado de asimetría de la función de densidad de probabilidad que su signo sea positivo significa que la cola de la derecha de la función de densidad de probabilidad es más largo que el de la izquierda.

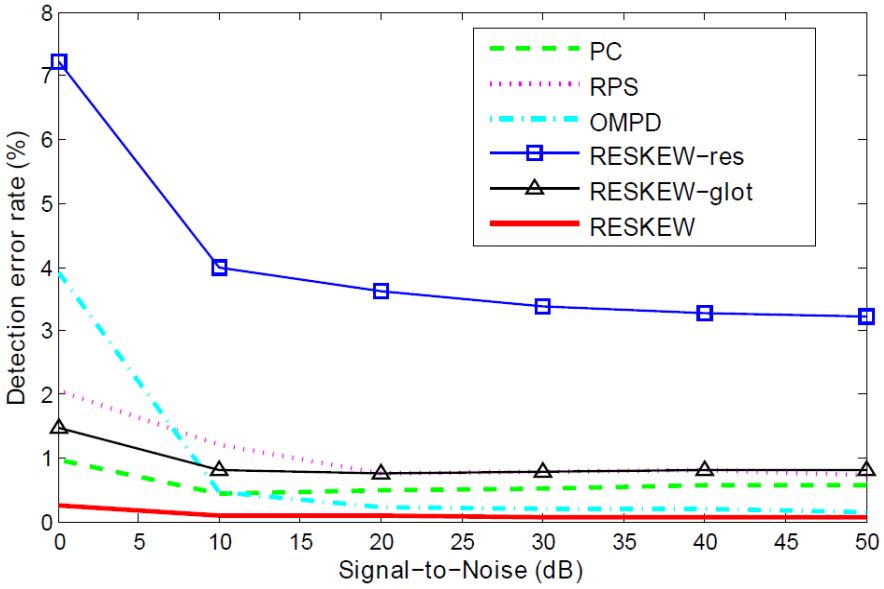



***Figura* 3.5.3:** Tasa de error de los distintos métodos que extraen la polaridad, extraído [Drug 13]

## *3.5.3. RESKEW*

Entonces para este algoritmo miramos el signo de la asimetría de la señal residual y de una burda aproximación de la derivada del pulso glotal. Con lo cual sabemos que si la asimetría de la señal residual es positiva la polaridad es 1 o que si la asimetría de la derivada del pulso glotal es -1 la polaridad es 1. Para que ambos criterios coincidan en su decisión ambas asimetrías tienen que tener signo contrario, si se diera el caso de que ambas tienen el mismo signo te quedas con la de mayor valor absoluto, ese es el motivo de que en la formula final aparezcan ambas asimetrías restadas.

$$\gamma_1(x) = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2)^{\frac{3}{2}}},$$

- r(n) señal residual.
- g´(n) burda aproximación de la derivada del pulso glotal

**Polaridad** = signo de $\gamma_1(r(n)) - \gamma_1(g'(n))$.

Esta g´(n) no cuenta ni con detectores de GCI y de sonidos sonoros ni nada de eso, ya que se presupone que extraer la polaridad de la señal es un paso previo a todos estos y tan solo es necesaria una aproximación lo más simple posible del pulso glotal. La forma de obtener g´(n) es la siguiente: primero filtra paso alto la señal de voz y

después halla los coeficientes del filtro FIR mediante un análisis LPC y filtra la señal original de manera inversa con estos coeficientes. La idea clave de g'(n) es que en el filtrado paso alto pasen la mayor contribución posible a los formantes del tracto vocal, que es lo que más tarde cancelas y que la influencia de la señal glotal que normalmente esta en bajas frecuencias, por eso del filtrado paso alto, no se cancele.

Algo a tener en cuenta muy importante, que aparece reflejado en el esquema general del sistema es que la entrada del *polarity detector* debe ser la locución completa y no un fonema. Además de ser lo que dice expresamente en las especificaciones de la función estuve realizando pruebas con distintos fonemas de distintos tipos de duración y la solución más estable se daba cuando la entrada estaba constituida por la locución entera. Además como sabemos que la polaridad está relacionada con las conexiones eléctricas del micrófono tiene sentido que la polaridad no cambie en toda la locución.

## 3.6 Detector de GCI (Glottal Closure Instants)

### *3.6.1. Utilidad*

Un aumento del interés del procesado glotal de voz de manera síncrona ha dado lugar al incremento de la demanda de técnicas fiables en la detección de los GCI tanto en habla limpia como en habla con ruido y reverberación.

La gran ventaja de obtener los GCI antes del pulso glotal, es que con los GCI se puede obtener el pulso glotal de manera síncrona, es decir, pulso a pulso. Existen otras técnicas en la estimación del pulso glotal en la que la señal de entrada es toda la señal de voz. Sin embargo, cuando contamos de manera previa con los GCI se puede estimar el pulso glotal de un periodo fundamental, más tarde del siguiente y así sucesivamente lo que se ha demostrado que logra extraer un pulso glotal que se asemeja más adecuadamente a la realidad

### *3.6.2. Comparativa*

Vemos una comparativa en la que sean utilizado seis bases de datos, donde se ha evaluado tomando como referencia los GCI tomados por un electroglotografo. La

gráfica muestra la distribución del error en tiempo, evidentemente nos conviene que esto todo en el cero o lo más próximo a él. Solo he añadido esta grafica comparativa aunque hay otras que lo que concluyen es lo siguiente:

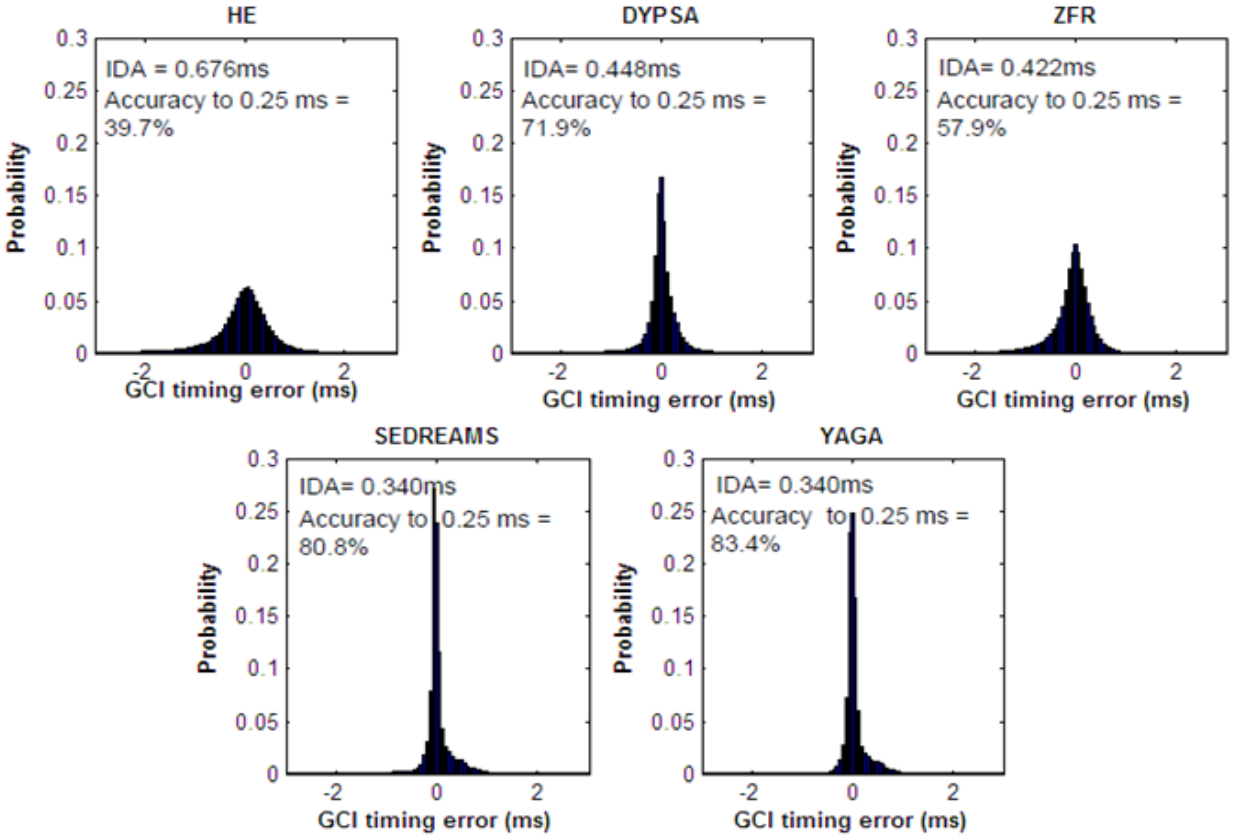



***Figura* 3.6.2:** Comparativa de los distintos algoritmos que implementan el GCI detector, extraído de [Drug 09]

En otra comparativa sobre la complejidad computacional de cada método mostramos la tabla en la cual cada número que aparece representa el tiempo relativo de computación que es un porcentaje de tiempo que ha tardado el método correspondiente con respecto al que tardaron el conjunto de los métodos. El resultado se ha promediado para las 6 bases de datos.

| Method | Male | Female |
|---|---|---|
| HE | 35.0 | 31.8 |
| Fast HE | 7.6 | 7.8 |
| DYPSA | 19.9 | 19.4 |
| ZFR | 75.7 | 74.9 |
| SEDREAMS | 27.8 | 27.1 |
| Fast SEDREAMS | 5.4 | 6.9 |
| YAGA | 28.6 | 28.3 |

***Tabla 3.6.2:*** Comparativa de los diferentes algoritmos que implementan del GCI, extraido de [Drug 09]

Las conclusiones que se pueden sacar realizados por distintos experimentos que evalúan las diferentes técnicas de detección del GCI son las siguientes:

- Da los mejores resultados junto con YAGA en entornos sin ruido.
- Con ruido da el mejor rendimiento.
- Presenta junto con ZFR la mayor robustez en entornos con reverberación.
- Es el método más adecuado en aplicaciones en tiempo real.

## *3.6.3. SEDREAMS*

*The Speech Event Detection using the Residual Excitation And a Mean-based Signal* (SEDREAMS) es un algoritmo recientemente propuesto que localiza fiablemente los GCI y GOI (*glottal openning instant*) de una señal de voz. Para nuestro proyecto sólo es necesario la obtención de los GCI, con lo cual cómo hallar los GOI se omite.
El suavizado de la señal residual es algo conveniente ya que las resonancias del tracto vocal, el ruido y la reverberación son atenuadas mientras que la periodicidad de la señal permanece invariable. La desventaja radica en la ambigüedad de las precisión en situar el instante temporal donde se encuentran los GCI. Por esta razón la señal residual se usa de manera combinada con la señal suavizada para lograr una localización más exacta.

El algoritmo consta de dos pasos: 1) Localización de intervalos donde la existencia de los GCI está garantizada y 2) refinamiento de estas posiciones dentro de los intervalos. Ahora paso a describir de manera más detallada cada uno de estos dos pasos.

***1) Determinación de los intervalos de presencia usando la señal promedio:***

Una discontinuidad en la excitación de la señal se refleja sobre toda la banda del espectro incluida la frecuencia cero. Derivada de esta observación surge el análisis centrado en la utilización de la señal promedio. Denotando la señal de voz como s(n) la señal promedio y(n) se define como:

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

donde w(m) es una función de enventanado de longitud 2N+1. Mientras que la elección de la forma de la ventana no es algo crítico (para este estudio vamos a utilizar

la ventana de Blackman) se ha demostrado que su longitud que influye en la respuesta al impulso que tiene la operación de filtrado, puede afectar en la fiabilidad del método. Se observa interesantemente que la señal promedio oscila de manera cercana con el periodo fundamental (pitch). Si la ventana es demasiado corta ocasiona la aparición de espurios en la señal residual que dan lugar a falsas alarmas. Por el contrario si la ventana es demasiado larga puede que se nos pase algún instante donde se encuentre un GCI. Se ha observado que el mejor rendimiento se da cuando la longitud de la ventana está entre 1.5 y 2 veces la duración de la media del periodo fundamental $T_0$ del locutor considerado. En este proyecto se ha utilizado 1.75 veces $T_0$.

Sin embargo la señal promedio no es suficiente por sí sola para una obtención de los GCI precisa. De hecho, considerando la figura 3.6.3.2 donde para 5 locutores diferentes se muestra la posición actual del GCI dentro de un ciclo normalizado de la señal promedio. Resulta que los GCI se dan en una posición relativa no constante dentro del ciclo. Sin embargo, una vez que el valor mínimo y máximo de la señal promedio están localizados es inmediato extraer intervalos donde con mucha seguridad aparezca un GCI. Más precisamente, como se observa en la figura 3.6.3, estos intervalos están definidos como el periodo de tiempo que empieza en el mínimo de la señal promedio y dura hasta una longitud de 0.35 veces la duración del periodo fundamental.

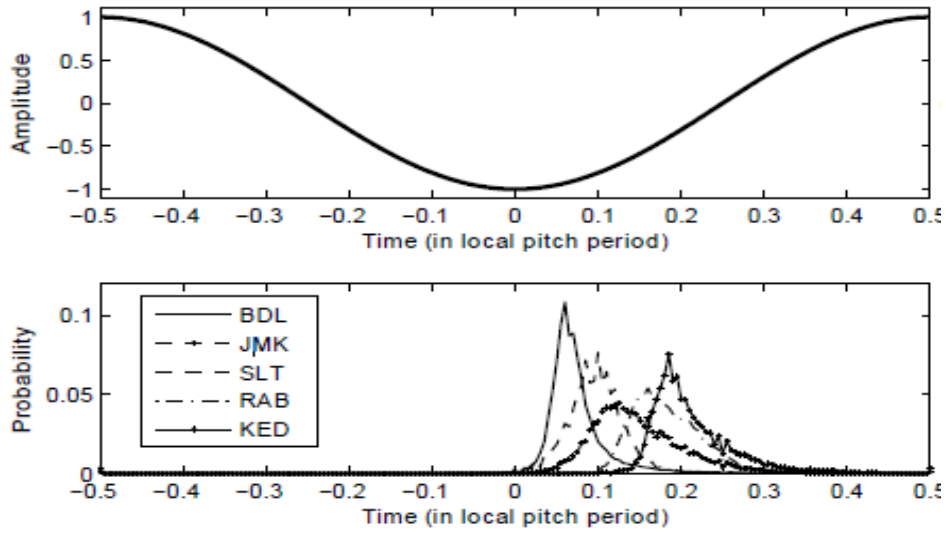



***Figura* 3.6.3.2:** Figuras que muestran en dónde se sitúan la mayoría de los GCI, extraído de [Drug 09]

### *2) Refinamiento de la localización de los GCI usando la señal residual:*

Los intervalos de presencia obtenidos en el paso anterior dan una idea difusa de dónde se localizan los GCI. El objetivo del siguiente paso es afinar la localización de la

ocurrencia de los GCI dentro de este intervalo. Para lo cual la señal residual es hallada, asumiendo que una gran discontinuidad de la señal se hace presente en ella y se corresponde con el lugar preciso de existencia de los GCI. Se ve claramente que la

combinación de intervalos de presencia extraídos por la señal promedio con los picos que marca la señal residual permite una precisa y fiable ubicación de los GCI.

Existe un algoritmo propuesto para la detección de los GCI más avanzado que SEDREAMS llamado SE-VQ el cual utilizan los investigadores que forman el grupo de COVAREP, y que podemos encontrar en su página al igual que todos estos algoritmos. La única diferencia de SE-VQ con respecto a SEDREAMS es que SE-VQ aplica ligeras modificaciones al algoritmo cuando el segmento de voz que se está tratando es de cualidad vocal *''creak''*. Por tanto el detector de *'creak'* que disponemos debe situarse antes que este algoritmo para poder localizar los GCI.

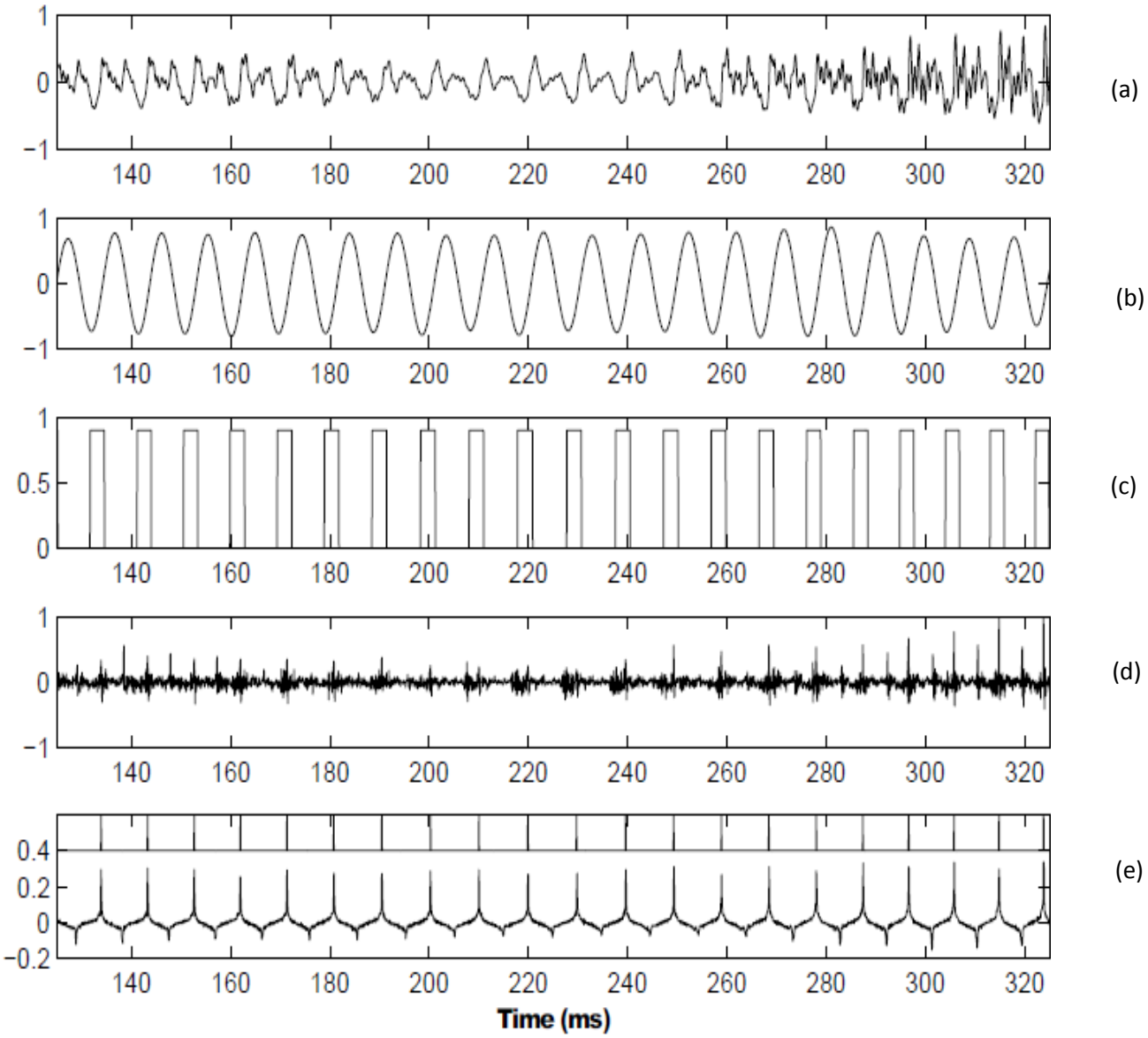



***Figura* 3.*6.3.1:*** Esquema de funcionamiento del algoritmo SEDREAMS, extraído de [Drug 09]

(a) Señal de voz

(b) Señal promedio

(c) Intervalos de presencia de los GCI derivados de la señal promedio

(d) Señal residual

(e) Localización final de los GCI basada en la combinación de la señal residual con la señal promedio

# 3.7 Estimación del pulso glotal

## *3.7.1. Introducción*

La voz resulta de filtrar el pulso glotal a través de las cavidades del tracto vocal y pasando por último a través de la influencia de los labios que actúan como una derivada. En muchas aplicaciones de voz es importante separar la contribución de la glotis y del tracto vocal. El logro de esta deconvolución puede llevar caracterización del modelado de estas dos componentes y a un mejor entendimiento de la producción del habla humano. Esta descomposición es un primer paso fundamental en el estudio de parámetros glotales que nos permitan observar la máxima inter-variabilidad y mínima intra-variabilidad. Limitamos el alcance de nuestro proyecto a métodos que realizan las estimación del pulso glotal directamente de la señal de audio. Aunque algunos dispositivos como electroglotografos o laringografos que miden la impedancia entre las cuerdas vocales (pero no el pulso glotal en sí) pueden contener información muy relevante sobre el comportamiento de la glotis. En la mayoría de los casos supone un inconveniente ya que no se dispone de los aparatos y para el análisis sólo se cuenta con la señal de voz. Este es el típico problema de separación a ciegas ya que ni el tracto vocal ni la contribución de la glotis son observables. Esto también implica que no es posible hacer una comparación cualitativa del rendimiento de estas técnicas de estimación de la señal glotal sobre habla real ya que no contamos con la señal de referencia (el pulso glotal que realmente produce el locutor).

Uno de los retos o problemas básicos en el procesado de audio ha sido siempre la estimación del pulso glotal, existen muchos investigadores que se han dedicado a este tema y muchas técnicas disponibles en la literatura. Sin embargo la diversidad de estos algoritmos y el hecho de que no disponemos de la señal glotal de referencia hacen muy cuestionable todas estas técnicas en habla real.

### *3.7.2. Métodos existentes*

Tales técnicas pueden ser separadas en dos clases de acuerdo a la manera en la que realizan la separación tracto vocal pulso glotal. La primera categoría se basa en el filtrado inverso mientras que la segunda se basa en las propiedades de mezclas de fase de la voz. Los métodos que consisten en el filtrado inverso primero estiman la parametrización del tracto vocal para después obtener el pulso glotal cancelando la contribución del tracto vocal vía filtrado inverso. Los distintos métodos se diferencian en la manera que consiguen la estimación del tracto vocal.

#### *Closed Phased Inverse Filtering (CPIF)*

*Closed phase* se refiere a la fase temporal durante la cual la glotis está cerrada. Durante este periodo los efectos de las cavidades subglotales son mínimas, dado como resultado el momento óptimo para calcular la función de transferencia del tracto vocal. Por lo tanto este método obtiene la envolvente del espectro del tracto vocal en la correspondiente estimación en la que la glotis se encuentra en la fase cerrada. La principal contra radica en dificultad determinar con exactitud los periodos temporales donde la glotis se encuentra cerrada. Distintas aproximaciones han sido propuestas en la literatura para resolver el problema. En (Plumpe et al., 1999), se propone sacar la fase en que la glotis está cerrada analizando la modulación de la frecuencia de los formantes entre las fases abierta y cerrada. Además de este problema de determinar de manera exacta la fase cerrada, puede ocurrir que el periodo fundamental sea demasiado corto (voces con pitched alto) en los que no se tienen suficientes muestras disponibles para una estimación fiable. Fue más tarde propuesto en (Brookes and chan, 1994 ) una técnica multiciclo que usaba LPC, donde un pequeño número de ciclos glotales colindantes dotan de suficientes muestras para una estimación del pulso glotal concisa.

En este proyecto la técnica para realizar el CPIF que se usa está basada en un función de transferencia del tracto vocal DAP (*discrete all pole*). Con el objetivo de lograr un mejor ajuste espectral a la función de transferencia del tracto vocal, la técnica DAP, propuesta por Jaroudi y Makhoul en 1991, tiene como fin optimizar la autoregresión de los parámetros del modelo minimizando la distancia Itakura-Saito en vez de el error cuadrado temporal hallado por el LPC tradicional. La distancia de Itakura-Saito está

justificada ya que es una medida de distorsión espectral que surge de la percepción del oído humano. Destacar que para voces con un pitch alto se utilizan dos ventanas para analizar el pulso glotal tal y como sugiere (Brookes and Chan, 1994), (Yegnanarayana and Veldhuis, 1998) y (Plumple eta al. 1999). Para los experimentos que aparecen en la posterior comparación se ha fijado el orden del DAP a 18, aunque este valor no es crítico siempre que se encuentre dentro del rango 12 a 18.

## Descomoposición de mezcla de fase *(Mixed-Phase descomposition)*

De acuerdo con este modelo la voz está compuesta por una fase mínima (causal) y una fase máxima (anticausal). Mientras que la respuesta al impulso del tracto vocal y la fase de retorno del pulso glotal pueden ser consideradas señales de fase mínima, el periodo donde la glotis está abierta se corresponde con una señal de fase máxima. Además se ha comprobado en el estudio de (Gardner and Rao, 1997) que los modelos de descomposición en *mixed-phase* representan de manera adecuada la naturaleza de la excitación glotal. Este estudio mostraba que la utilización de un filtro anticausal sólo polos es necesario para extraer información en magnitud y fase de manera correcta. La idea clave de este método es descomponer la voz en sus componentes de fase mínima y máxima siendo la última debida solo a la contribución glotal.

Una cuestión crucial en la descomposición *mixed-phase* es el peso de la ventana que se le aplica a la señal de voz para análisis de corta duración. De hecho, ya que la descomposición se basa en las propiedades existentes en la fase de la señal de voz, el enventanado tendrá una gran influencia. Se ha probado que tanto una correcta localización de los GCI como que se den ciertas restricciones en la longitud de la ventana son esenciales para garantizar una correcta descomposición. Para este algoritmo se utiliza una ventana Blackman de duración dos veces el periodo fundamental y centrada en el GCI. Existen dos aproximaciones para utilizar este método una es la técnica basada en los ceros de la transformada-Z (ZZT) y la otra propuesta por Drugman en 2009 basada en la descomposición en cepstrum compleja(CCD).

*Complex Cepstrum Decomposition* (CCD) se define según las siguientes ecuaciones donde los cepstrum complejos $\hat{x}$ (n) de una señal discreta x(n) se define como:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}$$

$$\log[X(\omega)] = \log(|X(\omega)|) + j\angle X(\omega)$$

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log[X(\omega)]e^{j\omega n}d\omega$$

donde la primera ecuación se corresponde con una DTFT , la segunda con un logaritmo de una función compleja y por último con una inversa de la DTFT. Esta descomposición en el dominio de los cepstrum surge por el hecho de que los coeficientes cepstrum complejos de una señal anticausal (causal) son cero para todos los n positivos (negativo). Quedándote sólo con los índices negativos de estos coeficientes cepstrum complejos hace posible la estimación del pulso glotal.

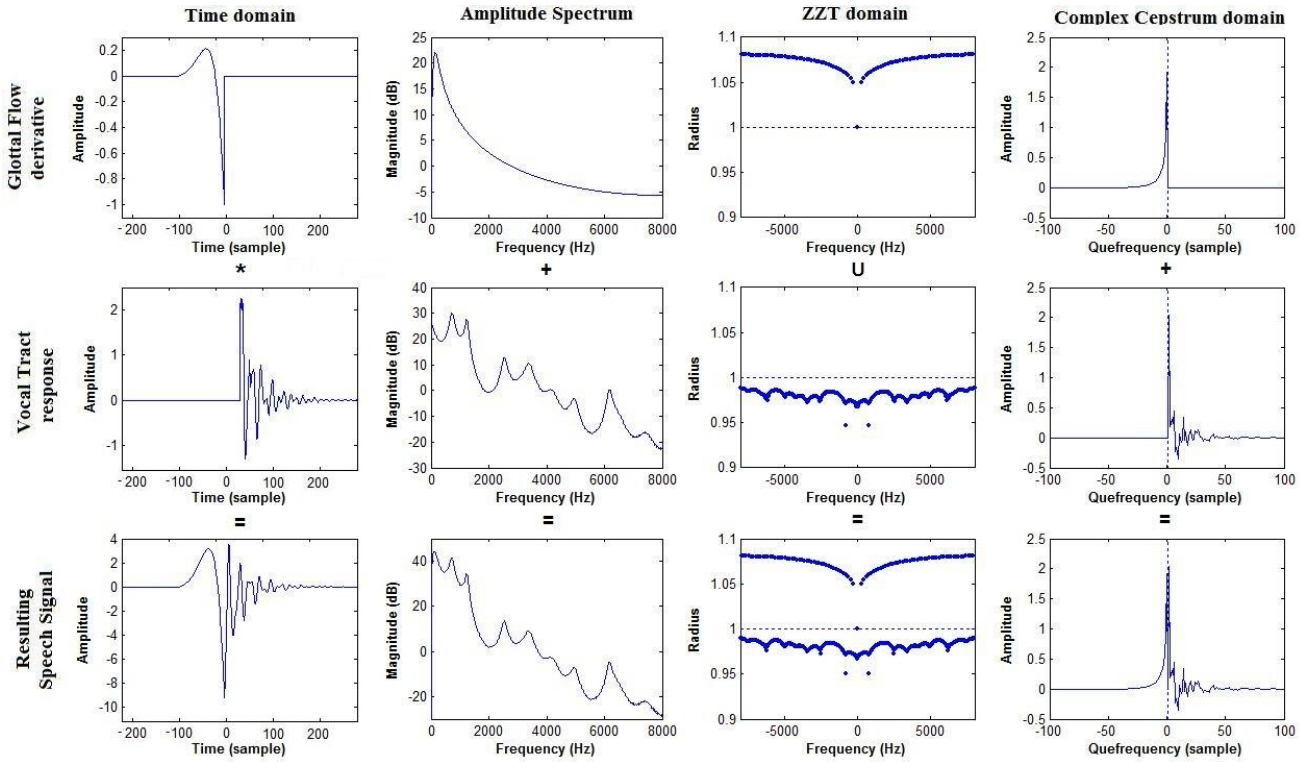



***Figura* 3.*7.2.1:*** Distintas funciones que explican la diferencia entre ZZT y CCD (métodos de estimación del pulso glotal), extraído de [Kan PhD 12]

Ilustración de la descomposición de fase mixta. Filas: respectivamente representan las siguientes señales: la derivada del pulso glotal (arriba), la respuesta del tracto vocal (en el centro), y la señal de voz resultante (abajo). Cada columna corresponde a un dominio de la representación de estas señales: el dominio del tiempo (columna primera), amplitud del espectro (segunda columna), la representación ZZT en coordenadas polares (tercera columna), y dominio cepstrum complejo (cuarta columna).

## *IAIF (Iterative Adaptative Inverse Filtering)*

Presentado y desarrollado por Alku [Alku and Laine, 1989; Alku 1992], *Iterative Adaptative Inverse Filtering* (IAIF) se basa en la suposición de que la pendiente del espectro de la señal puede ser atribuida al pulso glotal, y la forma de onda del pulso glotal puede ser representada por modelo sólo polos de orden bajo. En IAIF, la forma principal del pulso glotal son estimadas repetidamente gracias a un análisis LPC de orden bajo, y el efecto de este pulso glotal es cancelado de la señal original a través de un filtrado inverso. Posteriormente una estimación mas concisa del tracto vocal puede ser obtenida gracias a un análisis LPC de orden más alto al anterior que actúe sobre la señal de voz sin la influencia correspondiente al pulso glotal. En el último paso, la señal de voz original es inversamente filtrada por los coeficientes del filtro del tracto vocal , y como resultado se obtiene el pulso glotal.

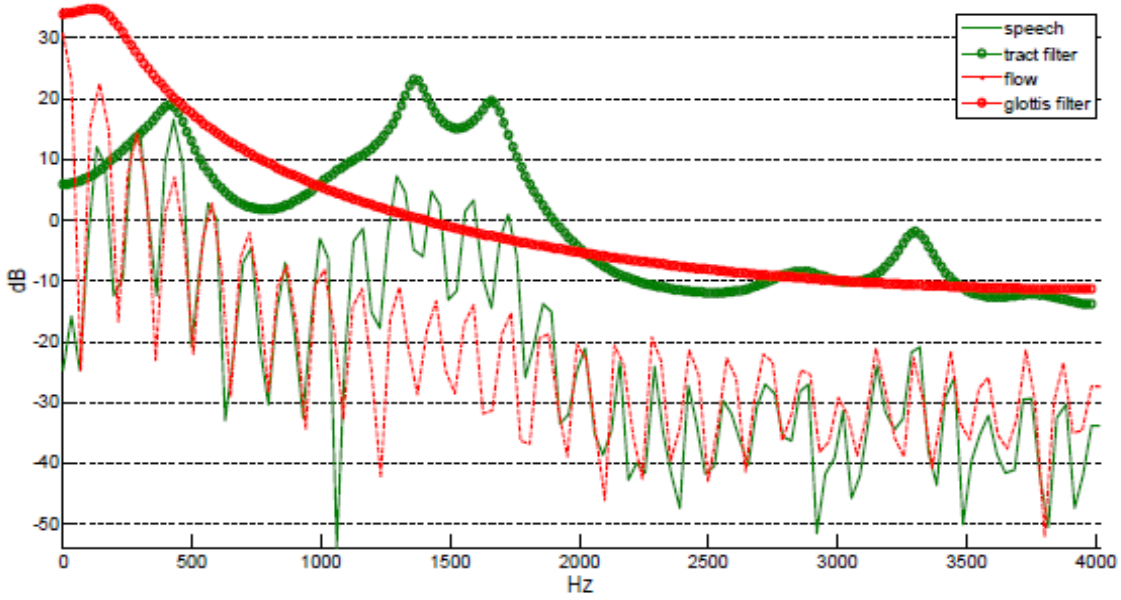



***Figura* 3.*7.2.2:*** Espectro de cómo contribuyen los distintos componentes del modelo de voz en la formación del espectro de la señal de voz final, extraido de [Haou 13]

En un trabajo posterior [Alku and Vilkman, 1994], Alku propone usar un análisis DAP (*Dicrete all pole*) para estimar mejor estos coeficientes que caracterizan al tracto vocal en vez de usar el convencional LPC. Un diagrama esquemático del algoritmo IAIF es mostrado en la figura 3. 7.3. Primero , como muestra el bloque 1 la señal de voz original es filtrada paso alto por un filtro FIR para eliminar la contribución de las componentes frecuenciales por debajo de 50 Hz.

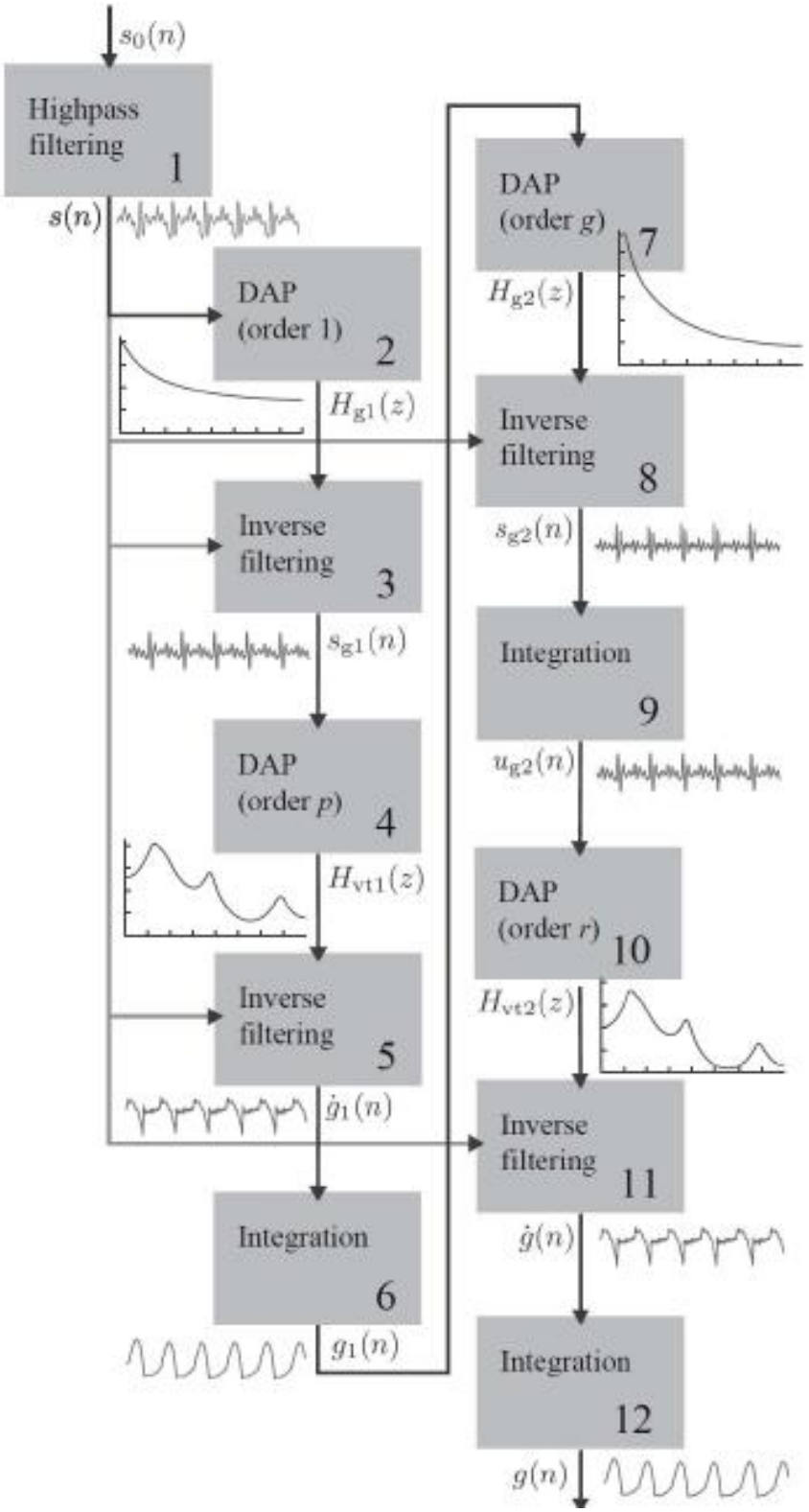



***Figura* 3.*7.2.3:*** Esquema explicativo del algoritmo IAIF, extraido de [Haou 13]

La primera estimación del pulso glotal es implementada de los bloques 2 a 6. En el bloque 2, un DAP de primer orden es aplicado al modelo del efecto de la combinación del pulso glotal a la influencia de los labios en el espectro de la voz, después de la cual la señal de voz es filtrada de manera inversa para eliminar dichos efectos en el bloque 3. En el bloque 4 un análisis DAP de orden p (que doble el número de formantes así que generalmente escogemos p = Fs/1000 + 2) es aplicado para extraer la primera respuesta al impulso del tracto vocal. Los resultados son usados en el bloque 5 para filtrar de manera inversa la señal de voz y obtener una burda estimación de la derivada del pulso glotal . La derivada del pulso glotal es integrada y filtrada paso alto con una frecuencia de corte de de 50 Hz para eliminar las bajas frecuencias y extraer una primera estimación de la forma de onda del pulso glotal en el bloque 6, la cual será utiliza para posteriores análisis dentro del mismo algoritmo.

Los bloques del 7 al 12 repiten lo mismo constituyendo una segunda fase del algoritmo IAIF en el que se consiguen versiones tanto del pulso glotal como del tracto vocal más aproximadas a las reales. En el bloque 7, la primera burda aproximación del pulso glotal es analizada por un DAP de orden g (2 a 4) para obtener una nueva estimación de la contribución del pulso glotal al espectro de la señal de voz. Para los bloques restantes del 8 al 12, el proceso es el mismo que en la primera fase (generalmente el orden de p se fija igual que el de g , pero puede ser ajustado manualmente para aumentar el rendimiento) y finalmente obtenemos una rigurosa estimación de la forma de onda del pulso glotal.

Un ejemplo de aplicar IAIF para un segmento de voz masculina se muestra en la figura 3.7.2. La figura presenta el espectro de la señal de voz, el espectro del pulso glotal finalmente extraído, el filtro estimado que supone la glotis y el filtro estimado que supone el tracto vocal.

La limitación de IAIF radica en las señales de voz que presentan un formante a muy bajas frecuencias, donde el formante glotal(el pico del espectro del pulso glotal) se superpone con el primer formante del tracto vocal. En estas situaciones es complicado eliminar el efecto de la señal al glotal del espectro de la señal de voz y los coeficientes estimados para determinar la función de transferencia del filtro del tracto vocal no son del todo fiables.

*Ejemplos de derivadas del pulso glotal obtenidas.*

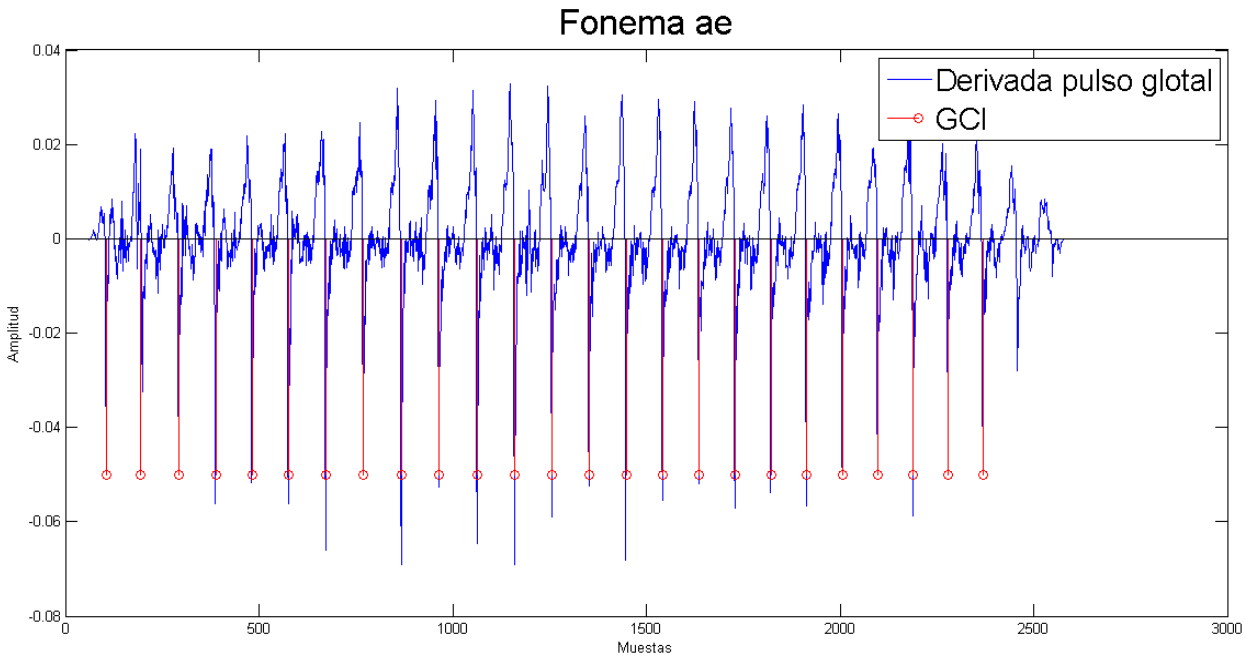



***Figura* 3.*7.2.4:*** Ejemplo de derivada de pulso glotal extraída

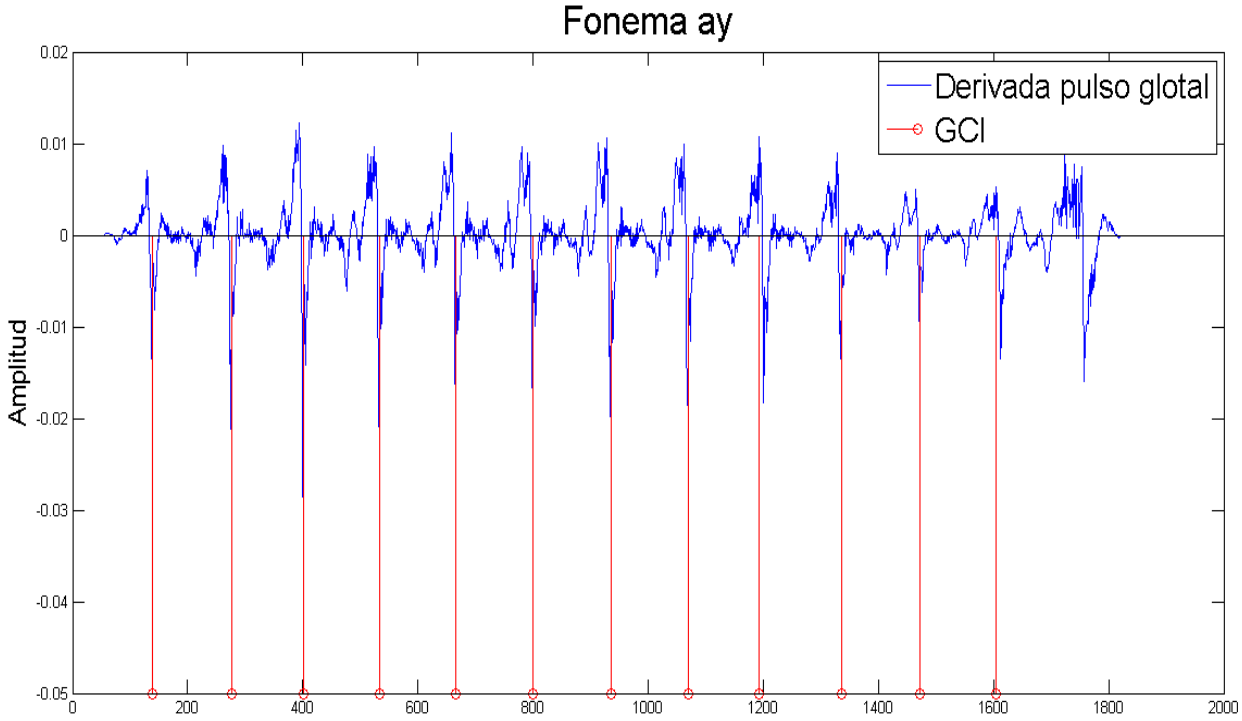



***Figura* 3.*7.2.5:*** Ejemplo de derivada de pulso glotal extraída

Los instantes temporales marcados en rojo coinciden con los GCI encontrados por SEDREAMS. Como está claro estos instantes se localizan antes de sacar la derivada del pulso glotal para poder hacerlo de manera síncrona (pulso a pulso). Calcular los GCI con la derivada del pulso glotal ya extraída sería muy fácil, sin embargo sí que es cierto que dentro del posterior algoritmo de extracción de parámetros glotales sí que se refina la localización de los GCI colocando dentro un mínimo cercano de la función.

Ese ligero rizado que presenta la derivada del pulso glotal se debe a que IAIF no ha conseguido cancelar la influencia de los formantes del tracto vocal por completo.

## *3.7.3. Comparativa*

Los tres algoritmos antes descritos para estimar el pulso glotal son ahora comparados. Para evaluar su rendimiento se definen dos medidas objetivas. Como ya dijimos en apartados anteriores tenemos el problema de no tener una señal de referencia con la que comparar, para lo cual en estos experimentos se realizan comparaciones con habla sintética donde sí se conoce a priori este pulso glotal. El efecto del ruido y de la frecuencia fundamental en esas medidas es estudiado a continuación:

**Tasa de error en NAQ y QOQ:**

NAQ y QOQ son dos parámetros que se extraen del pulso glotal. Hallar un error en estos parámetros tras la descomposición en pulso glotal y tracto vocal será penalizado. Un ejemplo de la distribución del error relativo en QOQ en condiciones con apenas ruido (SNR = 80dB) es mostrado en la figura 3.7.3.1 Hay muchas formas de evaluar estas medidas de error relativo representadas por el histograma nosotros usamos es la proporción de segmentos que tienen un error relativo superior a un umbral del 20%. Cuanto menor tasa de error tengan sobre esta medida mejor funcionamiento proporcionará el algoritmo.

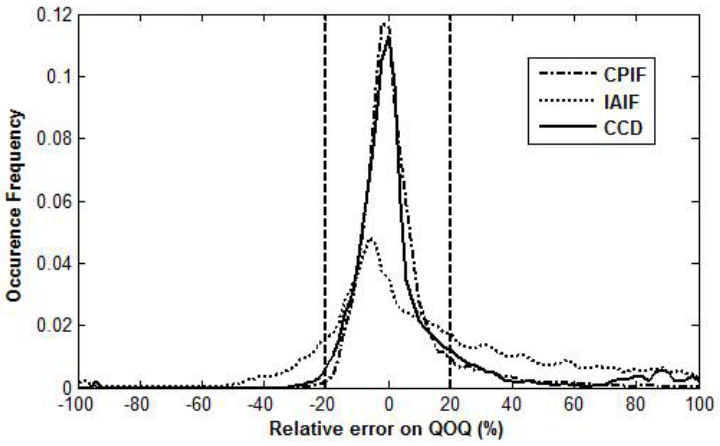



***Figura* 3.*7.3.1:*** Error relativo que cometen los diferentes algoritmos en la extracción de NAQ y QOQ , extraído de [Drug 12]

**Distorsión espectral:**

Muchas de las medidas en el dominio de la frecuencia para cuantificar la distancia entre dos segmentos de voz x e y surgen de la literatura de codificación del habla. Idealmente la sensibilidad subjetiva del oído debe ser formalizada por la incorporación de sonidos psicoacusticos tales como enmascaramientos. Sin embargo emplearemos una simple y relevante medida llamada distorsión espectral (SD) definida como (Nordin y Eriksson,2001)

$$SD(x,y) = \sqrt{\int_{-\pi}^{\pi} (20\log_{10}|\frac{X(\omega)}{Y(\omega)}|)^2 \frac{\mathrm{d}\omega}{2\pi}}$$

Donde X(w) e Y(w) denotan los espectros de ambas señales como función de frecuencia angular normalizada. En (K. Paliwal, 1993), los autores argumentan que la diferencia de 1 dB (con una frecuencia de muestreo de 8kHz) se percibe cuantiosamente. Para tener en cuenta este apunte, usamos la siguiente medida entre los espectros de las señales estimada y de referencia:

$$SD(Estimated, Reference) \approx \sqrt{\frac{2}{8000}\int_{20}^{4000} (20\log_{10}|\frac{S_{Estimated}(f)}{S_{Reference}(f)}|)^2 \mathrm{df}}$$

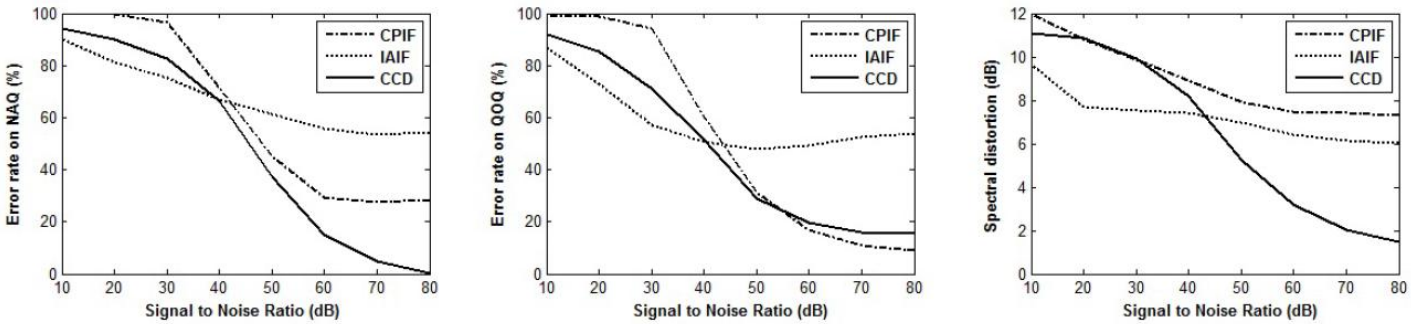



***Figura* 3.*7.3.2:*** Error relativo que se comete sobre los distintos parámetros en función del pitch del locutor, extraído de [Drug 12]

En esta figura representa la tasa de error antes definida para cada uno de estos tres parámetros cuando variamos la adición de un ruido Gaussiano blanco.

Para ver la robustez que cada método tiene ante el ruido se añade ruido Gaussiano blanco a la señal para tener varios niveles de SNR. Este ruido es usado como un pobre sustituto del ruido de grabación. Los resultados de acuerdo al rendimiento de cada método son mostrados en la figura 3.7.3.2 Como era de esperar todos los algoritmos empeoran cuando se añade más ruido. Particularmente CCD resulta ser el más afectado. Esto puede ser explicado por el hecho de que una ligera presencia puede afectar dramáticamente a la información que va en la fase de la señal y consecuentemente en la calidad de descomposición del algoritmo en cuestión. El rendimiento de CPIF también es degradado cuando se empieza a incrementar los niveles de ruido. Este puede deberse al hecho de que el ruido también modifica cuantiosamente la envolvente del espectro calculada durante la fase en la que la glotis permanece cerrada y, por tanto, la resultante estimación del pulso glotal es errónea. Por el contrario, a pesar del hecho de que IAIF no resulta tan eficiente en condiciones de bajo ruido, destaca por encima de los otros algoritmos cuando se dan condiciones adversas (por debajo de 40 dB). Una posible explicación de esta robustez radica en la naturaleza iterativa del algoritmo. Aunque la primera iteración puede ser altamente afectada por el ruido (como en el caso de CPIF) esta perturbación se vuelve más débil cuanto tras varias iteraciones converge.

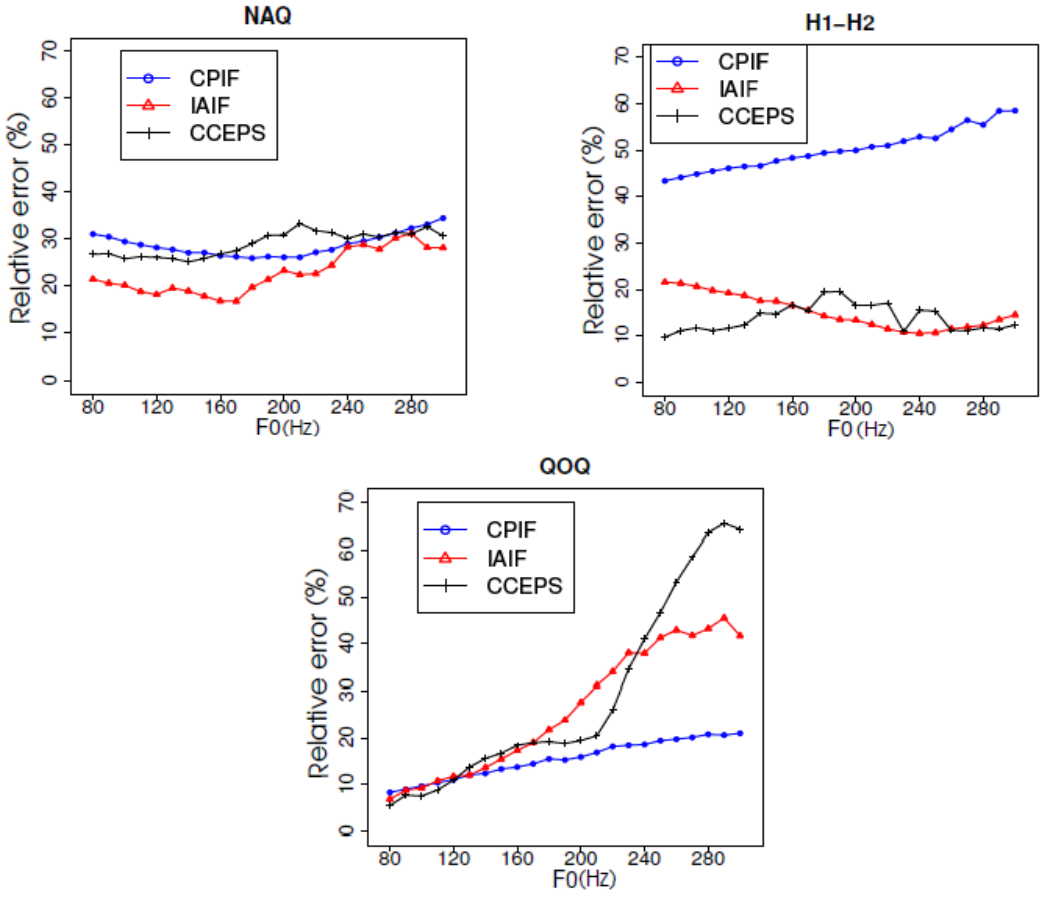



***Figura* 3.*7.3.3:*** Media del eror relativo para los parámetros glotales NAQ, QOQ y H1H2, extraído de [Kan PhD 12]

La figura 3.7.3.3 muestra la media del error relativo para los parámetros glotales NAQ, QOQ y H1H2 en función del pitch f0. El parámetro NAQ parece ser bastante poco sensible a las variaciones del tono del locutor(f0). Por debajo de frecuencias de 240Hz en el pitch del locutor el algoritmo de IAIF produce el menor error relativo, pero sin embargo a partir de este punto los tres algoritmos producen resultados similares.

Para QOQ el método *closed-phase inverse filtering* (CPIF) da los errores relativos más bajos. Esto es particularmente notable para altas frecuencias , con os algoritmos IAIF y CCPES mostrando significativos aumentos en el error en frecuencias alrededor de 200 Hz.

Para el caso de H1H2, sin embargo CPIF da los errores relativos más altos. Aparentemente puede observarse de este análisis que, a pesar del hecho de que la extracción en el dominio temporal del pulso glotal que realiza CPIF, da parámetros temporales adecuados, no ocurre lo mismo para parámetros frecuenciales que da un error relativo bastante mayor.

## *3.7.4. Conclusión*

Un estudio previo exponía que los algoritmos CCEPS (Drugman et al,2009) y CPIF(Yegnanarayana y Veldhuis, 1998) destacaban en IAIF en ciertos experimentos. Sin embargo CCEPS no separa la fase de retorno que tiene el pulso glotal de la contribución del tracto vocal y la fase de retorno es un aspecto importante de donde se puede obtener información y que hay que considerar. Esto se debe a que ambas tracto vocal y fase de retorno están contenidos en el componente de la fase mínima de la señal de voz. CPIF es también muy sensible a pequeños errores en la posición de los GCI y GOI . De hecho en comparaciones realizadas en la tesis de John Kane sugieren que el rendimiento de IAIF es mucho más comparable a los otros métodos de lo que antes se había expuesto. Además recientes métodos de síntesis de habla que utilizan el pulso glotal han optado por utilizar IAIF (Cabral et al 2011, Raitio et al 2011).

También he utilizado el criterio de comparar visualmente el pulso glotal estimado con el pulso glotal ideal representado en la literatura. Estuve haciendo pruebas tanto con CCEPS como con IAIF, y la forma de onda que más se asemejaba tanto en condiciones adversas como favorables la producía el algoritmo de IAIF.

Por lo tanto como el criterio que caracteriza nuestro estudio es que el sistema de obtención de características glotales sea lo más robusto frente al ruido y frente a distintas adversidades, la elección del algoritmo se decanta por IAIF. Además la mayoría de los investigadores actuales también eligen este algoritmo frente a los otros en todos sus estudios y experimentos.

# *Sección 4*

# *Extracción de parámetros*

## 4.1 Introducción

En esta sección hablaremos de tres cosas, por un lado del modelo elegido para sacar parámetros del pulso glotal el cual va a ser el modelo LF que es un modelo clásico, por otro lado hablaremos de una serie de parámetros que aunque se extraen directamente de la señal de voz permiten distinguir entre distintas cualidades vocales. Por último en esta sección expondremos un algoritmo novedoso propuesto por John Kane en su tesis de 2012 en la que propone un nuevo método para extraer los parámetros clásicos del modelo LF a partir de un pulso glotal obtenido con más exactitud que los métodos que existían hasta el momento.

## 4.2 Modelo LF

El modelo LF ( Fant et al. 1985) es un modelo de 5 parámetros (incluyendo f0 y asumiendo $t_c = T_0$) de la derivada del pulso glotal que surge de modelos previos desarrollados por el mismo autor. Además de $f_0$, el modelo LF cuenta con parámetros que se derivan de tres puntos temporales: $t_p$ , $t_e$, $t_a$ así como un valor de amplitud, EE.

El modelo está compuesto por dos partes, la fase en que la glotis está abierta y la fase de retorno en que la glotis se está terminando de cerrar y es calculado:

$$g'_{LF}(t) = \begin{cases} E_0 e^{\alpha t} sin\omega_g t & \text{for} \quad t_o \leq t \leq t_e \quad \text{open-phase} \\ \frac{-EE}{\epsilon T_a}(e^{-\epsilon(t-t_e)} - e^{-\epsilon T_b}) & \text{for} \quad t_e < t < t_c \quad \text{return-phase} \end{cases}$$

donde $w_g$ es $\frac{\pi}{Tp}$, $T_b = t_c - t_e$ , α y $E_0$ son requeridos para conseguir un equilibrio entre áreas (el área absoluta de los dos segmentos tiene que ser siempre la misma) y Є se saca usando:

$$\epsilon = \frac{1}{T_a} \cdot \left(1 - e^{-\epsilon T_b}\right)$$

Más detalles sobre cómo encontrar los valores de α y $E_0$ son proporcionados en Gobl(2003) y Fant et al. (1985). La optimización de ambas partes de la ecuación para resolver los valores de Є, α y $E_0$ para que cumpla el equilibrio entre áreas se consigue utilizando el método Newton-Raphson.

La forma del pulso del modelo LF también puede ser caracterizada por una serie de parámetros que ahora vamos a describir uno a uno:

### *La frecuencia fundamental, f0:*

La frecuencia fundamental = $1/T_0$ , donde $T_0$, el periodo fundamental, es el tiempo entre dos excitaciones consecutivas.

### *Excitation strength, EE:*

El EE es la amplitud negativa en el instante temporal de máxima discontinuidad de la derivada del pulso glotal. Normalmente se produce en la pendiente máxima de la rama descendente del pulso glotal, el cual típicamente precede el cierre total de la glotis. Teniendo en cuenta la producción del sonido está determinado por la velocidad en que se cierran las cuerdas vocales y el aire que pasa a través de ellas. EN el nivel acústico se corresponde con la intensidad global de la señal. Este parámetro es uno de los más similares a la amplitud del impulso que caracterizaba a ese modelo tradicional del pulso glotal.

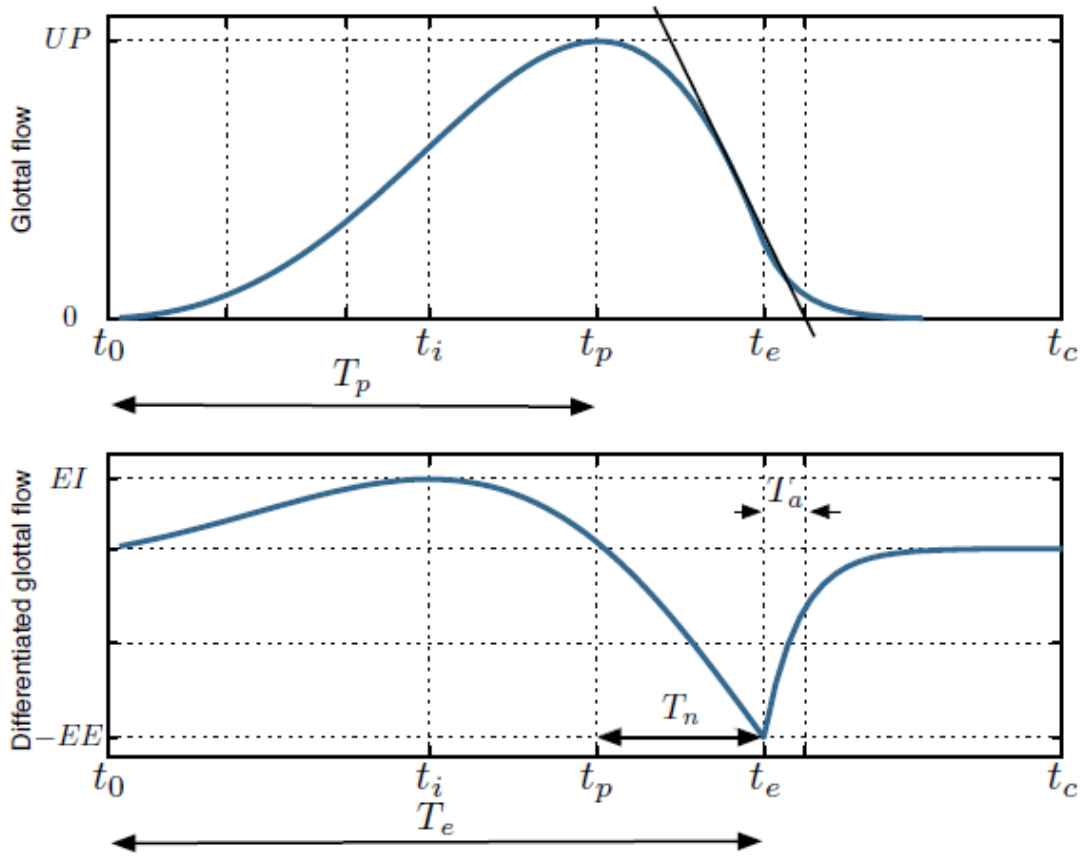



***Figura 4.2.1:*** *Ejemplo del pulso glotal del modelo LF (arriba), derivada del pulso glotal abajo,* extraído de [Kan PhD 12]

## *Dynamic leakage, Ra*

$R_a$ es la corriente de aire residual que se produce durante la fase de retorno y que se produce desde que se da el EE hasta que se cierra por completo la glotis. En términos del pulso glotal, la fase de retorno aparece como la esquina redondeada de la rama en la que el pulso glotal se está cerrando. En términos del modelo LF, $R_a$ es igual a $TA/T_0$ donde Ta se puede hallar como el tiempo que transcurre desde el $t_e$ en la figura 4.2.1 hasta el que una recta tangente en ese punto corte al eje x. $R_a$ se refiere a la manera de cerrar las cuerdas vocales, es decir, si lo haces de manera instantánea o más gradual. Las diferencias en $R_a$ son importantes acústicamente ya que afectan a la pendiente del espectro de la señal.

$$Ra = \frac{T_a}{T_0}$$

### *UP*

En la figura 4.2.1 se puede observar que se corresponde con la máxima amplitud alcanzada por el pulso glotal, que para el modelo se da en el instante $t_p$.

### *Open Quotient, OQ*

El OQ parámetro (originalmente propuesto en Timcke et al., 1958), se cree que es útil para discriminar entre las *voice quality* de *breathy* y *tense* (Henrich et al, 2001.; . Hanson y otros, 2001), también puede ser derivado del modelo LF:

$$OQ = \frac{T_e}{T_0} = \frac{1 + Rk}{2Rg}$$

### *Glottal frequency, FG*

FG definido como 1 /2$t_p$ (Fant 1979) es una frecuencia determinada por el periodo en el que la señal del pulso glotal está en la fase de abierta. Una expresión más práctica de este parámetro es $R_g$, es esencialmente lo mismo que FG pero normalizado por f0. $R_g$ tiende a variar inversamente con OQ y UP. Consecuentemente, un $R_g$ alto se encuentra con valores atenuados del espectro para altas frecuencias. Un alto $R_g$ junto con un segundo armónico bastante elevado suele ser característico de cualidad vocal tense. Bajos valores de $R_g$ son encontrados cuando el valor de UP es grande, que contribuye a ensalzar la amplitud del primer armónico.

$$Rg = \frac{T_0}{2T_p}$$

### *Glottal symmetry/skew, Rk*

A la asimetría del pulso glotal es una cuestión sobre la que se ha prestado mucho interés. El típico pulso glotal esta desplazado a la derecha, es decir, la fase en la que la

glotis se está abriendo tiende a ser más larga en comparación con la fase en la que la glotis se está cerrando. Las consecuencias acústicas de esta asimetría so algo complejas. Afecta principalmente a la parte baja del espectro del pulso glotal de manera que cuanto más simétrico es el pulso glotal más se ensalzan las amplitudes de los armónicos bajos. Sin embargo, el grado de asimetría también determina la profundidad de los valles en el espectro ( falta de armónicos ). En el espectro del pulso glotal cuanta más simetría tenga el pulso más profundos serán estos valles espectrales. La asimetría está altamente correlada con el EE, y su contribución perceptual es fácilmente confundible con el EE. Este riesgo de confusión es particularmente alto cuando sobre el modelo del pulso glotal primero no se establece un control sobre el EE, en otras palabras el parámetro EE es más importante que la asimetría.

$$Rk = \frac{t_e - t_p}{t_p}$$

## $R_d$

Otro R-parameter que se puede añadir lo constituye $R_d$. $R_d$ fue desarrollado para tener un único parámetro que capturara la mayor variación posible del resto de parámetros del modelo LF. $R_d$ se calcula usando:

$$Rd = 1000 \cdot \left(\frac{UP}{EE}\right) \cdot \left(\frac{f_0}{110}\right)$$

Los otros R-parameters ($R_g$,$R_a$,y $R_k$) pueden ser predichos de $R_d$ utilizando la siguiente regresión descrita en Fant et al. (1995), usando estas ecuaciones:

$$Ra_p = (-1 + 4.8Rd)/100$$

$$Rk_p = (22.4 + 11.8Rd)/100$$

$$Rg_p = \frac{Rk_p}{4(\frac{0.11Rd}{0.5+1.2Rk_p} - Ra_p)}$$

## *Otros parámetros del modelo LF*

Medidas del dominio del tiempo del pulso glotal se derivan principalmente de cocientes del periodo glotal. Pueden dar información muy relevante, como por ejemplo OQ caracteriza la duración relativa de fase en que la glotis está abierta. Otra medida es el CLQ (*closing quotient*) que simboliza la misma idea que el OQ pero en este caso sobre la fase en que la glotis se está cerrando. Sin embargo la localización de estos puntos temporales se sabe que es problemática (Alku et al., 2002) y como resultado parámetros que se derivan de amplitudes y cocientes entre ellas han demostrado mejorar la resistencia ante el ruido otros factores adversos. El cociente de amplitud normalizada (NAQ, Alku et al., 2002), por ejemplo es calculada como:

$$NAQ = \frac{f_{ac}}{d_{peak} \cdot T_0}$$

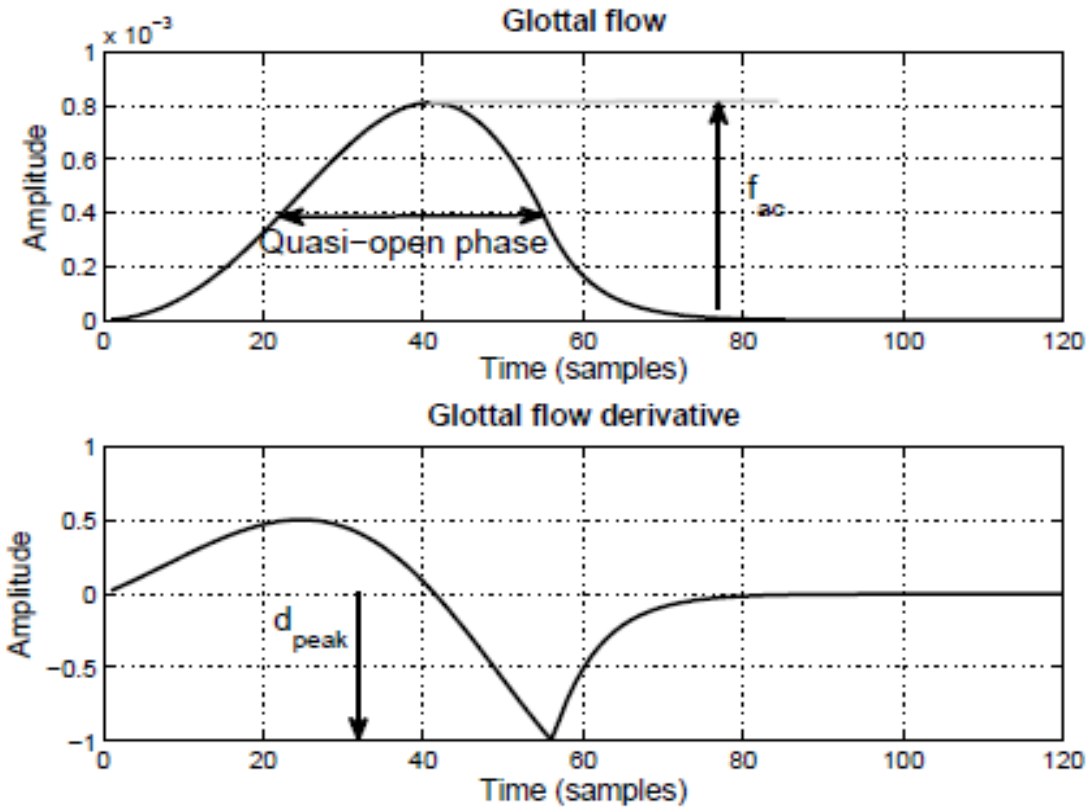



***Figura 4.2.2:*** *Pulso glotal (arriba) y derivada del pulso glotal (abajo), con las medidas necesarias para sacar NAQ y QOQ resaltados,* extraído de [Kan PhD 12]

Para entender mejor la formula mirar la figura 4.2.2, donde $f_{ac}$ es la máxima amplitud del pulso glotal, lo que también conocemos como UP y $d_{peak}$ es la máxima amplitud negativa de la derivada del pulso glotal, lo que también hemos visto como EE. Se ha demostrado que está altamente correlado con ClQ y ha sido usado para analizar la cualidad vocal en condiciones de grabación no ideales. Sin embargo algunos de los experimentos realizados en (Gobl and Ní Chasaide, 2003) sugieren que la habilidad de NAQ para separar entre voz *'tense'* y voz *'breathy'* es reducida cuando existe una alta variabilidad en f0.

El QOQ (*'quasi-open quotient'* , Hacki 1989) fue desarrollado como medida más robusta en relación con el 'open-quotient'. Es calculado detectando el pico de máxima amplitud en el pulso glotal para después encontrar el punto previo y posterior que ha descendido un 50% de esta máxima amplitud. La duración del intervalo entre estos dos puntos es dividida por el periodo fundamental para dar como resultado el QOQ.

Otros parámetros basados en medidas en el dominio de la frecuencia ha sido desarrollados. La diferencia de amplitudes entre los dos primeros armónicos (H1-H2) en la derivada del pulso glotal es uno de ellos. Otro parámetro espectral es el factor de riqueza de armónicos (HRF, Chiledrs and LEE, 1991) que es la suma de amplitudes de armónicos por encima del primero dividido entre la amplitud del fundamental. Un parámetro más seria el parámetro espectral parabólico (PSP, Alku et al, 1997) que fue propuesto para modelar las características frecuenciales del pulso glotal. PSP se calcula ajustando una parábola a las bajas frecuencias del espectro del pulso glotal. Destacar que sin embargo, la efectividad de esta medida puede verse seriamente afectada cuando no hay una cancelación completa del formante a más baja frecuencia.

Los parámetros descritos en esta sección se basan en una estimación fiable del pulso glotal. Se sabe que el filtrado carece de resistencia en ciertos tipos de regiones de voz y como consecuencia estos parámetros se verán negativamente afectados.

## 4.3 Otros parámetros sacados directamente de la señal

Son parámetros sacados directamente de la señal de voz y no del pulso glotal, pero que nos permiten también distinguir entre distintos tipos de *voice quality*. Estos son el *peak slope*, la *maxima dispersion quotient* y por ultimo un detector de *'creak'*.

## *4.3.1 Peak Slope*

Este parámetro es útil para diferenciar entre cualidades vocales desde *breathy* hasta tense en un segmento de voz dado. Técnicas que nos dan información robusta para diferenciar entre cualidades vocales para segmentos de voz nos pueden ayudar tanto en tareas de análisis como en etiquetado de la cualidad vocal en grandes bases de datos. Para sacar este parámetro de una señal de voz hay realizar una descomposición en ondiculas (wavelet) en octavas y después realizar una regresión lineal que una la máxima amplitud de las diferentes escalas. La pendiente de esta regresión lineal es finalmente el parámetro *'peak slope'* con el cual mostraremos su habilidad para diferenciar entre cualidades vocales y la ventaja que supone sobre otros parámetros en esta tarea. El *'peak slope'* muestra resistencia al ruido de burbuja añadido a la señal con SNR tan bajos como 10 dB. Además, proporciona una mejor entre cualidades vocales de '*breathy*' hasta *'tense'* que otros parámetros tanto en vocales como en segmentos en medio de la frase.

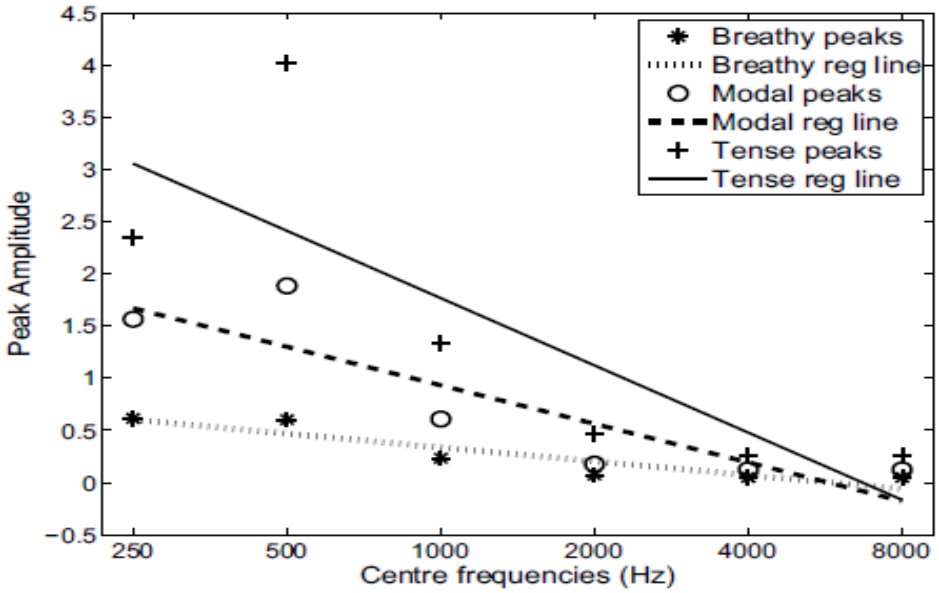



***Figura 4.3.1.1: A****mplitudes de los picos de la descomposición wavelet junto con su regresión lineal pronunciados para la vocal /o/ de un locutor masculino con cualidades vocales desde 'breathy' hasta 'tense',* extraído de [Kane Gob 13]

En la figura 4.3.1.2 observamos la distribución de los cuatro parámetros para las cualidades vocales de '*breathy*', *'modal'* y *'tense'* para una base de datos de vocales. Se proponen tres umbrales para identificación ( a estas vocales se les ha añadido ruido blanco)de cualidad vocal que aparecen con una línea discontinua.

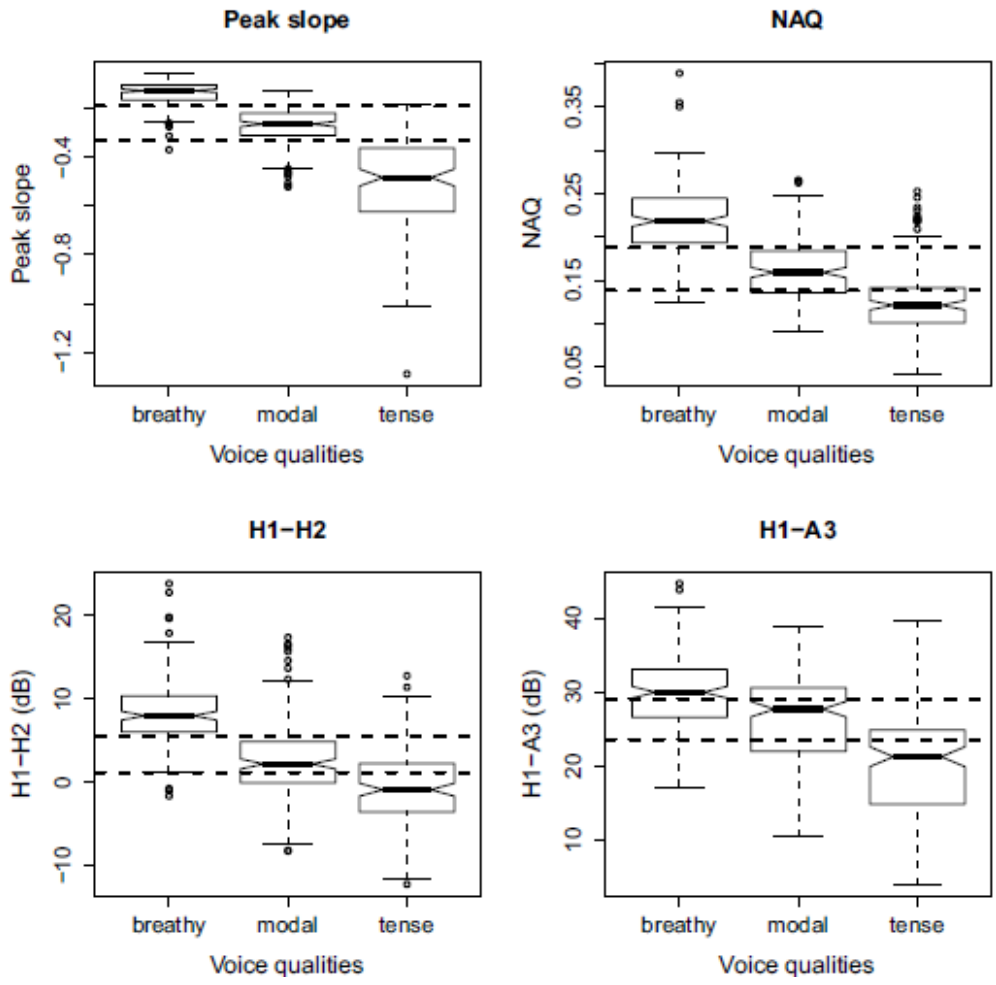



***Figura* 4.3.1.2:** Ventaja para discriminación aspirada, modal y tensa que tiene peak slope sobre otros parámetros, extraído de [Kane Gob 13]

## *4.3.2 Maxima Dispersion Quotient (MDQ)*

Los máximos hallados por descomposición en ondículas se suelen usar para detectar bordes en procesados de imágenes, donde la localización de este máximo se situaba en la vecindad del borde. De manera similar para la cualidad vocal de 'tense' la que normalmente tiene unas características abruptas en la fase de cierre de la glotis. Por tanto estos análisis de máximos se encontrarán en la vecindad de los GCI. Según la cualidad vocal se desplaza desde *'tense'* hasta *'breathy'* se observa que estos máximos aparecen cada vez más dispersos. El parámetro MDQ está diseñado para cuantificar la extensión de esta dispersión de máximos y también es un parámetro que da unos resultados favorables para distinguir entre cualidades vocales especialmente para habla real. MDQ ha resultado ser un parámetro que presenta resistencia al ruido hasta niveles por debajo de los 10 dB de SNR.

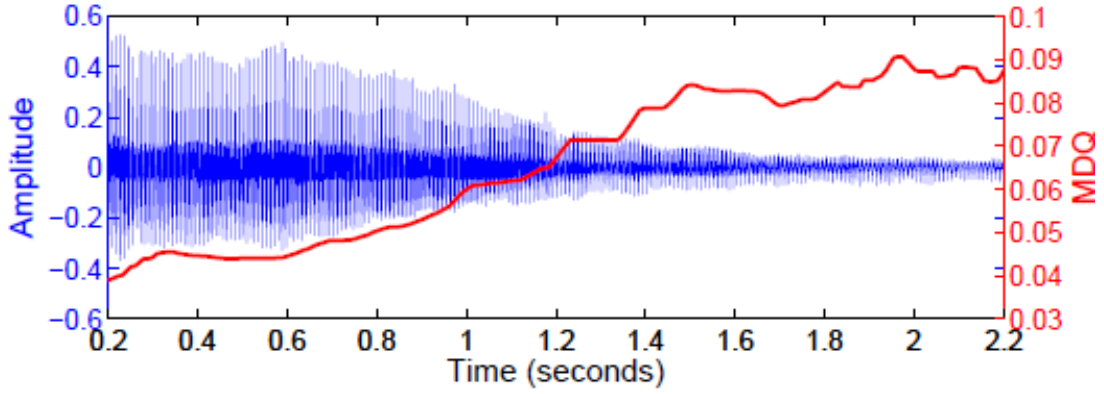



***Figura 4.3.2.1:*** Forma de onda de la pronunciación de una /a/ por un locutor masculino que varía gradualmente de cualidad vocal de *'tense'* a *'breathy'*. También muestra como está relacionado con el MDQ, extraído de [Kane Gob 13]

Para hallar el MDQ de una señal de voz se deben realizar los siguientes pasos: 1. Estimar los GCI de la señal de voz. 2. Calcular la señal residual con un análisis LPC. 3. Llevar a cabo la descomposición en ondículas de esta señal residual. 4. Definir el intervalo de búsqueda relativo a los instantes de localización de los GCI. 5 .Encontrar la localización de estos máximos, $m_i$, para las diferentes escalas en las regiones de búsqueda. 6. Medir la distancia entre estos máximos $m_i$, y los GCI antes localizados. Calcular la media de estas distancias normalizadas por el periodo glotal.

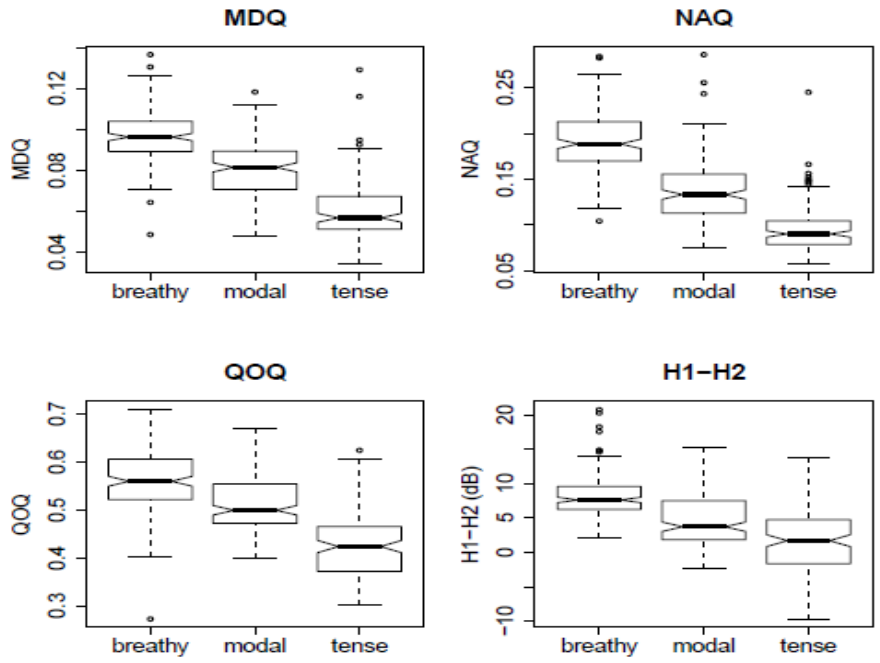



***Figura 4.3.2.2:*** Distribución de los parámetros MDQ,NAQ,QOQ y H1H2 como función de las distintas cualidades vocales, extraído de [Kane Gob 13]

### *4.3.3 Detector de 'creak'*

Primero el algoritmo intenta encontrar dos características propias de las señales de voz con cualidad vocal *'creaky'*, para más tarde cada una de estas dos características es usada como entradas de un árbol de decisión binaria que por fin tomara la determinación de si es *'creaky'* o no. Destacar que estas dos características intentan describir diferentes aspectos de la señal residual en regiones *'creaky'*.

Para sacar por tanto estas características primero hay que sacar la señal residual de la señal de voz como ya hemos visto (análisis LPC...). La primera particularidad que presenta la señal residual para que estemos en un segmento con cualidad vocal *'creaky'* es la ocurrencia de picos secundarios (incluso terciarios) que preceden al pico principal el cual se corresponde con el GCI. Este hecho está ilustrado en la figura 4.3.3.1 donde grandes picos en la señal residual se encuentran antes que los correspondientes a los GCI.

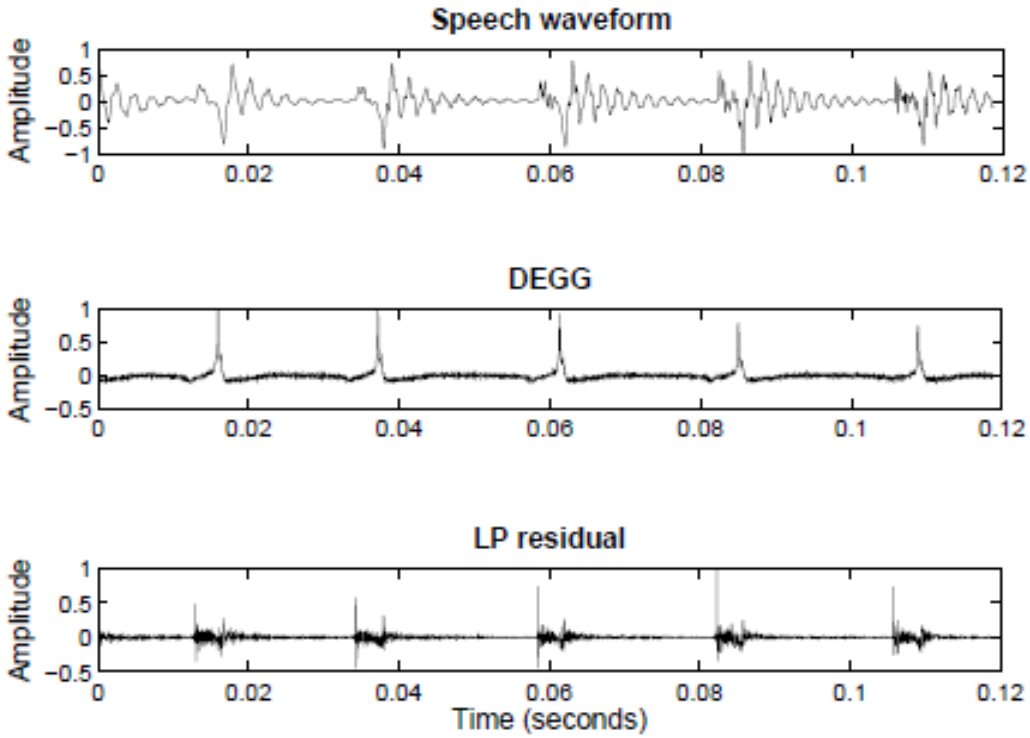



***Figura 4.3.3.1:*** Señales de una región *'creaky'* pronunciadas por un locutor masculino, extraído de [K&D 13]

Se presupone que estos picos secundarios se corresponden con los momentos en los que la glotis se abre bruscamente para después continuar con una fase de cierre bastante larga. Sin embargo, a veces excitaciones laríngeas fuertes también pueden causar estos picos secundarios en la señal residual. La primera componente del

algoritmo está diseñada para estudiar estos picos secundarios en la señal residual.

No obstante análisis posteriores de señales residuales en regiones ''*creaky*' han revelado que aunque la presencia de estos picos secundarios es muy frecuente, a veces en una región *'creaky'* pueden no aparecer. La segunda componente del algoritmo está diseñada específicamente para capturar esta particularidad y además sabiendo que las regiones *'creaky'* producen pulsos glotales de duración mayor. La segunda componente del algoritmo está diseñada para encontrar picos prominentes en la señal residual. Estos picos prominentes surgen como consecuencia de un cierre brusco de las cuerdas vocales por altos niveles de tensión aductiva. Inicialmente estos picos eran medidos directamente de la señal residual pero exámenes posteriores que esto era muy sensible ante la presencia de ruido. En su lugar se utilizo un resonador. Las características que se le pusieron fueron un ancho de banda de 1kHz que es el adecuado para destacar la prominencia de estos picos principales de la señal residual, sin verse demasiado perjudicado por los picos secundarios. De cada trama el máximo pico absoluto que sale del resonador es identificado y la trama se desplaza para quedar centrada en este pico. Según muestra la figura 4.3.3.2 la derecha una región *'creaky'* (a la derecha) con su poco en el centro de la trama. Para la región modal (izquierda), sin embargo, existen picos próximos dentro de la trama.

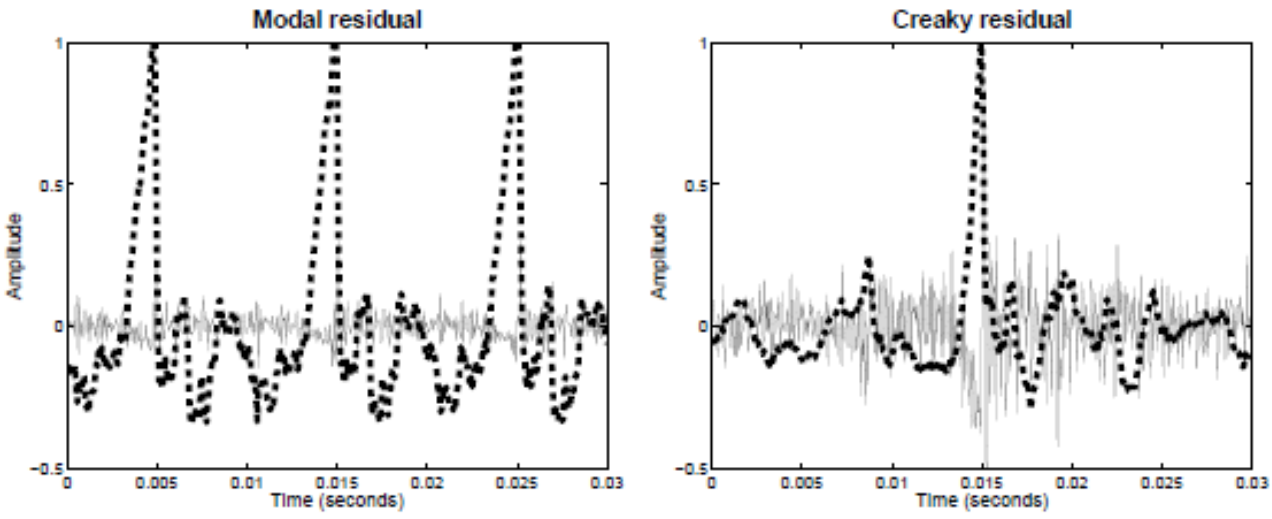



***Figura 4.3.3.2:*** 30ms de una señal residual centrada en su pico (línea fina) y salida del resonador (línea gruesa discontinua), extraído de [K&D 13]

Midiendo la diferencia de amplitudes entre el pico máximo (situado en el centro de la trama) y el siguiente pico más alto uno puede obtener un parámetro que diferencia las regiones 'modal' de las *'creaky'*. para no cometer el error de seleccionar valores muy próximos al pico central, la búsqueda del siguiente pico más alto se realiza más allá de una distancia de 6ms del centro de la trama.

Por último para detectar las regiones de *'creaky'* estos dos parámetros anteriormente descritos se usan como entradas de un árbol de decisión binaria. La separación de las dos cualidades vocales (es decir *'creaky'* y no *'creaky'*) se hace usando una aproximación que va de arriba a abajo donde ambas cualidades vocales empiezan situadas en la raíz después se realizan una serie de preguntas binarias ( que tienen que ver con las entradas) y para cada pregunta un nuevo nodo es creado. Esto crea un árbol de decisión en cuyo final hay nodos hoja.

## 4.4 Extracción de parámetros

Debido a la falta de robustez de métodos que analizaran el pulso glotal de manera automática se deicidio optar por métodos de que requerían ajuste manual pulso a pulso, sin embargo esto tiene el problema de no poder ser aplicable a grandes bases de datos. para afrontar el problema John Kane propone un nuevo método, para sacar los parámetros LF de un pulso glotal, en su tesis de 2012. El algoritmo pasa por tres fases: búsqueda exhaustiva, programación dinámica y métodos de optimización.

Para sacar f0 y EE, como el algoritmo cuenta con los GCI, simplemente los desplaza hasta una posición cercana donde se dé la mínima amplitud, este valor constituye el EE y f0 es determinado como la inversa del tiempo entre GCI continuos.

### *4.4.1 Búsqueda exhaustiva de Rd*

Los métodos estándares que extraen los parámetros LF del pulso glotal en el dominio del tiempo típicamente implican unas estimaciones iniciales de los parámetros para después refinarlos utilizando un método de optimización. Un problema común es que estas estimaciones iniciales suelen producir medidas iniciales bastantes pobres, esto se debe a la inconsistencia sobre donde colocar el punto en el que la glotis se abre $t_0$. Consecuentemente más tarde el proceso de optimización difícilmente resuelve estos fallos iniciales. Para resolver esto se propone una búsqueda exhaustiva la cual implica variar un amplio rango de posibilidades de combinaciones de los parámetros LF y guardar las configuraciones que dan valores mínimos en una determinada función de coste. Sin embargo cubrir el amplio rango de posibles combinaciones de los parámetros LF tendría un alto coste computacional. Por lo tanto en vez buscar todas las combinaciones posibles se simplifica variando solo un nuevo parámetro llamado Rd, para después generar de él Ra,Rk y Rg.

Se define una función de coste que tiene en cuenta el tiempo y la frecuencia:

$$\text{total\_cost} = \frac{\text{spec\_err} + \text{time\_err}}{2} \quad \in [0, 1]$$

Por cada Rd se genera un segmento de tres pulsos y el error entre los dos sets de armónicos es:

$$\text{spec\_err} = \{1 - |cor\{h_U(m), h_{LF}(m)\}|\} \cdot w_s \quad 1 \leq m \leq N \quad \in [0,1]$$

Donde $h_U$ y $h_{LF}$ son las amplitudes de los armónicos de la derivada del pulso glotal y de la derivada del pulso glotal sintético generada a partir de un Rd respectivamente. N es el numero de armónicos, $w_s$ es un peso y cor{.} es coeficiente de correlación de Pearson.

Por otra parte la función de coste temporal es:

$$\text{time\_err} = \{1 - |cor\{g_{LF}, g'(t)\}|\} \cdot w_t \quad \in [0,1]$$

Donde g'(t) es la derivada del pulso glotal y $g_{LF}$ es la derivada del pulso glotal sintética generada a partir de un Rd. $w_t$ es un peso y cor{.} es el coeficiente de correlación de Pearson.

Por cada pulso glotal te quedas con los 5 Rd que minimizan la función de coste para pasar al siguiente paso.

## *4.4.2 Programación dinámica*

La programación dinámica es usada para elegir el camino optimo formado por distintos valores de Rd que tienen en cuenta toda la señal de voz que le entra al algoritmo.

La función de coste d(i,j) está definida como el valor de error calculado en la función de búsqueda exhaustiva para cada Rd candidato en cada pulso glotal de análisis, donde $1 \leq j \leq N_{cand}$ (este valor suele ser 5) , $1 \leq i \leq M$ y m el número de GCIs (es decir en número de pulso glotales implicados en el análisis). La función de coste de transición puede está definida como:

$$\delta_{i,j,k} = \{1 - cor\{seg_{i,j}, seg_{i-1,k}\}\} \cdot w_{tr} \cdot ss \quad \in [0,1]$$

donde $seg_{i,j}$ se refiere a un único pulso glotal generado usando los parámetro R predichos por el candidato j de Rd en el pulso i. $seg_{i-1,k}$ se refiere al pulso generado usando el candidato k de Rd en el pulso previo i-1. Esta función de coste de transición se basa en el hecho de que, como en el tracto vocal, el pulso glotal debe variar suavemente en periodos temporales cortos(20 ms).

La función de coste de transición también está modulada dinámicamente por el factor ss ( medida de la estabilidad del espectro usada en Talkin, 1995):

$$ss = \frac{0.2}{\text{itakura}(f_i, f_{i-1}) - 0.8} \quad \in [0, 1]$$

donde itakura(.) es la medida de distorsión de Itakura de la derivada del pulso glotal de la trama $f_i$ (centrada en GCI, y con una duración de tres veces el periodo del pulso glotal) y la trama centrada en el GCI anterior, $f_{i-1}$. ss tiende a 1 cuando las características espectrales de tramas contiguas son similares y a cero cuando muestran gran diferencia. Este factor afecta a la función de coste de transición de manera que en los tramos donde hay cambios rápidos, por ejemplo en cambios vocal consonante, la función de coste de transición tiene menos efecto es decir se permite cambios más bruscos entre Rds mientras que en región estables, por ejemplo centro de vocales, la función de coste de transición tiende a mantener un contorno de variación de Rds más suave.

$$D_{i,j} = d_{i,j} + \min_{k \in N_{cand}} \{D_{i-1,k} + \delta_{i,j,k}\}, \quad 1 \leq j \leq N_{cand}$$

Inicializada como:

$$D_{o,j} = 0, \quad 1 \leq j \leq N_{cand}$$

Una ilustración del camino optimo de Rds es mostrado en la figura 4.4.2 con los 5 Rds candidatos por pulso glotal ( uno por cada GCI ). Se puede observar que en la región que va desde 0.9 segundos hasta 1 segundo, justo antes de la oclusión de la consonante, que ocurre un cambio bastante rápido en la señal y como resultado los Rds también cambian de manera considerable. Esto es debido a la baja estabilidad espectral en esta región que minimiza el efecto de la función de coste de transición. Sin embargo, para las regiones estables (por ejemplo de 0.7 a 0.85 segundos) la alta estabilidad espectral de pulsos glotales sucesivos asegura que la función de coste de

transición tiene un fuerte efecto y como resultado los valores de Rds permanecen muy estables.

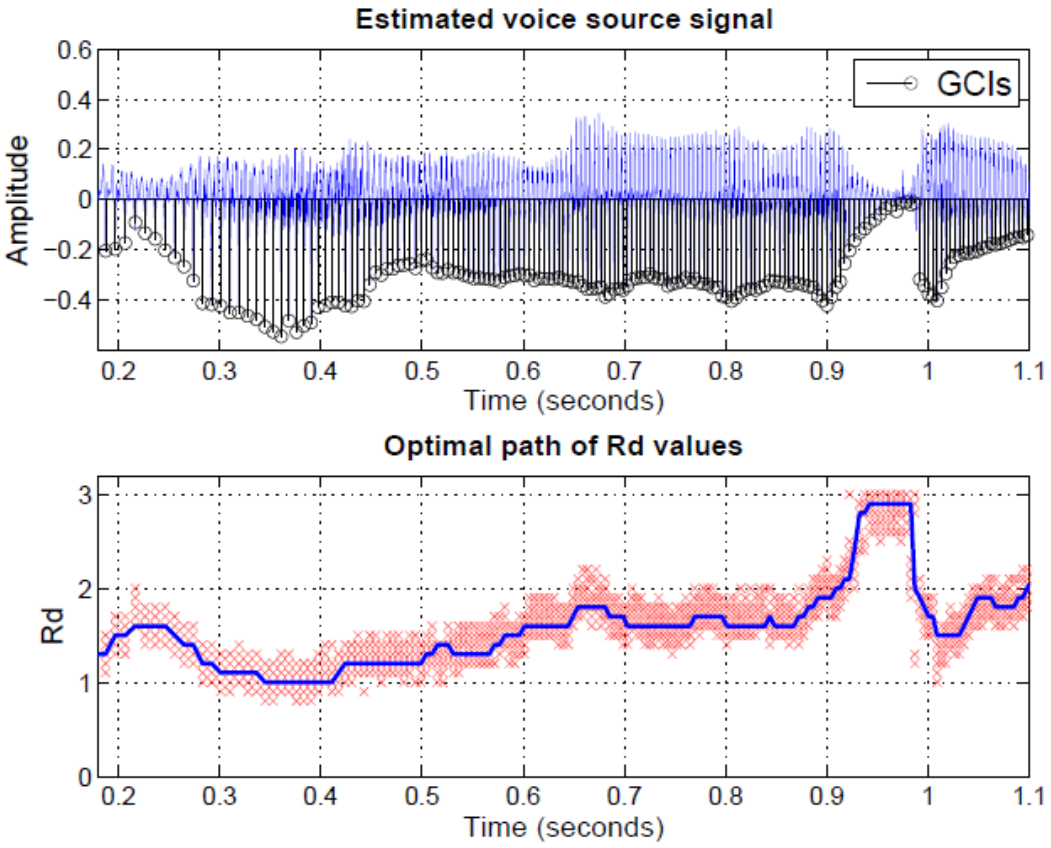



***Figura 4.4.2:*** Candidatos Rds (x's) y trayectoria optima de Rds (línea) representada en el tiempo (figura de abajo), extraído de [Kan PhD 12]

## *4.4.3 Optimización*

Aunque el parámetro Rd puede ser usado para caracterizar muchos de los tipos de pulsos glotales que surgen de diferentes tipos de cualidades vocales, es probable que algunos pulsos glotales quedan fuera de las restricciones de Rd. Además no era la intención limitar los grados de libertad del modelo. Para resolver este inconveniente, los valores de los distintos parámetros son refinados utilizando un método de optimización. Para cada pulso glotal Ra, Rk y Rg son derivados del valor de Rd elegido del paso anterior de programación dinámica. Un simple método (Nelder y Mead, 1965) es usado, el cual permite una optimización multivariable ilimitada. Los 3 parámetros R pueden variar para minimizar el mismo error de la función de coste del paso de búsqueda exhaustiva.

## *Ejemplos*

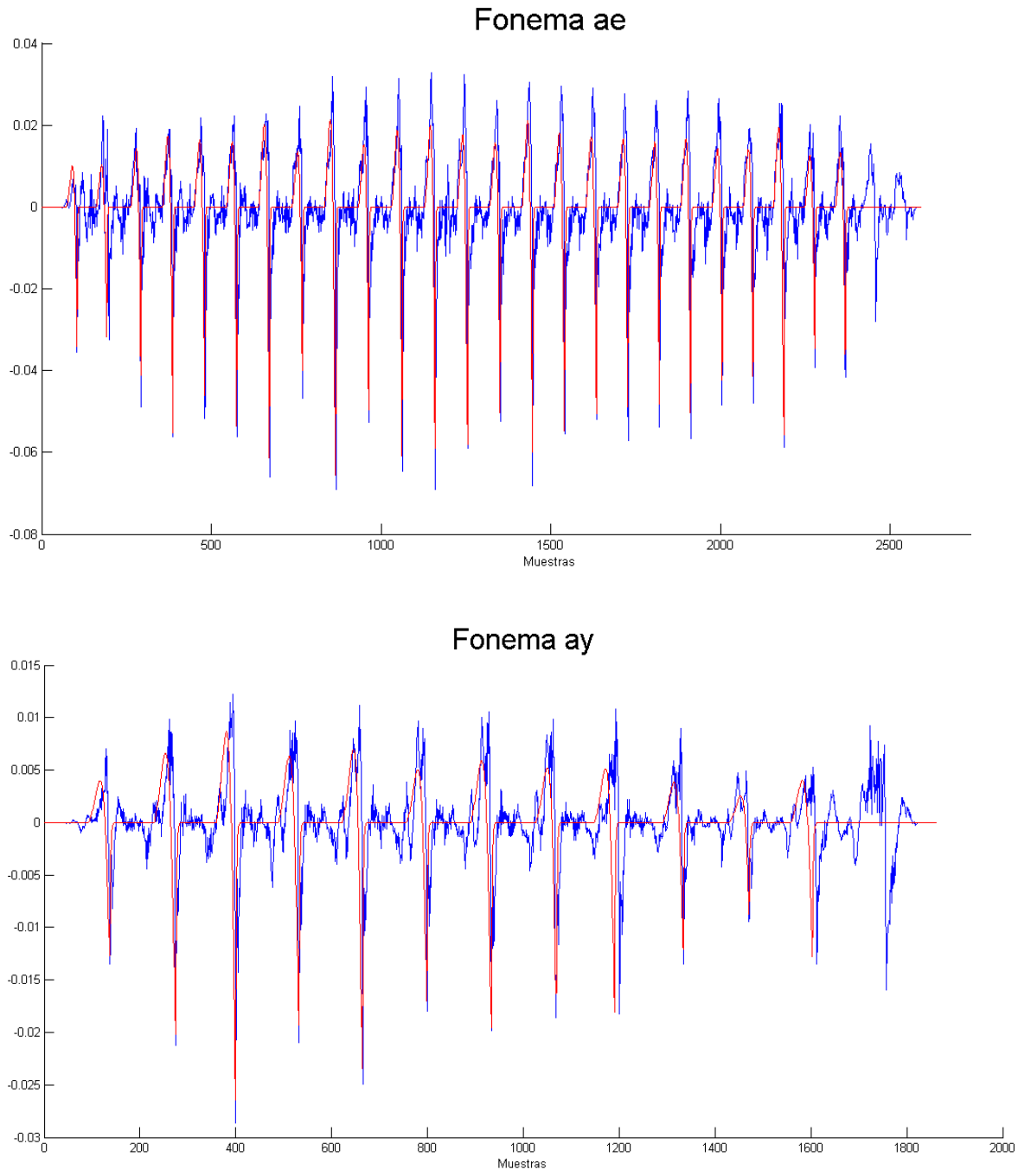



***Figura* 4.4.3:** Derivada del pulso glotal estimado (en azul) y derivada del pulso glotal resonstruida a partir de los parámetros LF extraídos de la derivada del pulso glotal estimado (en rojo)

En cada uno de estos ejemplos se observa dos señales: la derivada del pulso glotal estimado, y la derivada del pulso glotal reconstruida a partir de los parámetros LF extraídos de la derivada del pulso glotal estimado.

# *Sección 5*

# *Experimentos con TIMIT*

Lo primero que realizaremos en esta sección es describir brevemente la base de datos que vamos a usar, TIMIT, que resultará ser un entorno idóneo para un primer análisis de todas estas características glotales en distintos locutores.

Más tarde veremos algunos estudios básicos sobre los parámetros que extrae el algoritmo como son la correlación de los parámetros y algunos diagramas de cajas donde se observa la inter-variabilidad e intra-variabilidad que proporcionan estos parámetros para un número reducido de locutores (10 a 15). Después definiremos unas distancias para averiguar cómo están funcionando los parámetros también a nivel de intra/inter-variabildad sobre el conjunto de locutores, siempre diferenciando, claro está, entre hombres y mujeres.

Con estos primeros estudios ya descartaremos algunos parámetros que no proporcionan apenas información y ya estaremos listos para proponer un primer vector con distintas combinaciones de estos parámetros con los que entrenar un UBM.

También haremos una serie de estudios para evaluar como distinguen los distintos algoritmos y parámetros entre las principales cualidades vocales.

# 5.1. Entorno experimental

Hacemos un pequeño paréntesis para hablar de la base de datos que vamos a utilizar, TIMIT está diseñado para proporcionar datos de voz para los estudios acústicos - fonéticos y para el desarrollo y evaluación de sistemas de reconocimiento de voz automático. TIMIT contiene grabaciones de banda ancha de 630 hablantes de ocho dialectos de Estados Unidos. Cada locutor pronuncia diez locuciones ricas fonéticamente. TIMIT incluye transcripciones alineadas en el tiempo ortográficas, fonéticas y de palabras, a una frecuencia de muestreo de 16 kHz, 16 - bit para cada enunciado. El diseño fue un esfuerzo conjunto entre el Instituto de Tecnología de Massachusetts (MIT) , SRI International (SRI) y Texas Instruments, Inc. (TI). El discurso fue grabado en TI , que se transcribe en el MIT y verificado y preparado para producción de CD- ROM por el Instituto Nacional de Estándares y Tecnología ( NIST) .

Las transcripciones corpus TIMIT han sido verificadas a mano. Los subconjuntos de *'train'* y *'test'* han sido equilibrados para una máxima cobertura fonética . Se incluye tanto Información propia para la realización de las pruebas automáticas, así como la documentación escrita.

Esta tabla puede aclarar cómo están distribuidos los distintos locutores en la base de datos.

| Dialect Región(dr) | Male | Female | Total |
|---|---|---|---|
| 1 | 31 (63%) | 18 (27%) | 49 (8%) |
| 2 | 71 (70%) | 1 (30%) | 102 (16%) |
| 3 | 79 (67%) | 23 (23%) | 102 (16%) |
| 4 | 69 (69%) | 31 (31%) | 100 (16%) |
| 5 | 62 (63%) | 36 (37%) | 98 (16%) |
| 6 | 30 (65%) | 16 (35%) | 46 (7%) |
| 7 | 74 (74%) | 26 (26%) | 100 (16%) |
| 8 | 22 (67%) | 11 (33%) | 33 (5%) |
| **Total** | **438 (70%)** | **192 (30%)** | **630 (100%)** |

***Tabla 5.1.1*** Distribución de los locutores en base de datos TIMIT

Por tanto TIMIT nos proporciona un entorno perfecto, de ruido, etiquetas, para realizar pruebas donde sabes que las carencias que se produzcan están debida a nuestro sistema y no a la base de datos. Aunque el propósito final de analizar estas características glotales es analizar habla real.

Para las primeras pruebas utilizo 200 locutores de los cuales son 130 hombres y 70 son mujeres.

## 5.2. ¿Sobre qué unidades lingüísticas aplicar el sistema?

De la multitud de unidades lingüísticas que disponemos lo primero que nos tenemos que preguntar es sobre qué fonemas nos va a convenir más aplicar todo esto. Ya hemos discutido antes que nos conviene aplicar el sistema a fonemas ya que las características glotales pueden variar de uno a otro debido a que el tracto vocal y glotis forman un conjunto que varia dinámicamente por lo que no podemos suponer que la glotis siempre vibra de la misma manera. Incluso utilizando un mismo fonema puede haber diferencias en estas características glotales por el cambio de cualidad vocal.

Aunque en principio vale para cualquier fonema nos centramos en las vocales por varios motivos Son los fonemas que más energía van a tener. Tienen mayor duración por lo que se va extraer más información ya que hay más pulsos glotales y además es más fácil de extraer en esos periodos. En las vocales es también de donde vamos a obtener información de manera más estable.

## 5.3 Coeficientes de correlación

Calculamos el coeficiente de correlación entre parámetros para el set de 200 locutores (130 hombre y 70 mujeres ) que se muestra en la tablas 5.3.1 y 5.3.2 Como vemos los dos parámetros últimos están muy poco correlados con los otros como era de esperar ya que son parámetros que no se sacan del pulso glotal. Buscamos parámetros con poca covarianza con F0 es decir que sean independientes del pitch y este estudio también tiene como fin no poner dos parámetros muy correlados ya que si proponemos más adelante un vector con algunos de estos parámetros tendríamos información redundante.

| | FO | EE | Ra | Rk | Rg | OQ | UP | NAQ | QOQ | H1H2 | HRF | PSP | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FO | 1 | | | | | | | | | | | | |
| EE | 0,277 | 1 | | | | | | | | | | | |
| Ra | 0,027 | 0.035 | 1 | | | | | | | | | | |
| Rk | 0,045 | 0.057 | 0,099 | 1 | | | | | | | | | |
| Rg | 0,014 | 0.045 | 0,520 | 0,559 | 1 | | | | | | | | |
| OQ | 0,014 | 0,026 | 0,496 | 0,593 | 0,969 | 1 | | | | | | | |
| UP | 0,178 | 0,532 | 0,335 | 0,632 | 0,739 | 0,743 | 1 | | | | | | |
| NAQ | 0,125 | 0,104 | 0,004 | 0,085 | 0,101 | 0,121 | 0,008 | 1 | | | | | |
| QOQ | 0,250 | 0,323 | 0,031 | 0,007 | 0,074 | 0,061 | 0,206 | 0,289 | 1 | | | | |
| H1H2 | 0,376 | 0,263 | 0,030 | 0,066 | 0,018 | 0,046 | 0,120 | 0,232 | 0,229 | 1 | | | |
| HRF | 0,701 | 0,412 | 0,04 | 0,020 | 0,019 | 0,005 | 0,235 | 0,272 | 0,286 | 0,525 | 1 | | |
| PSP | 0,08 | 0,140 | 0,004 | 0,030 | 0,006 | 0,005 | 0,079 | 0,347 | 0,543 | 0,130 | 0,058 | 1 | |
| M | 0,122 | 0,073 | 0,010 | 0,060 | 0,023 | 0,023 | 0,053 | 0,073 | 0,029 | 0,016 | 0,139 | 0,078 | 1 |

***Tabla 5.3.1:*** Coeficientes de correlación entre distintos parámetros glotales para mujeres

| | FO | EE | Ra | Rk | Rg | OQ | UP | NAQ | QOQ | H1H2 | HRF | PSP | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FO | 1 | | | | | | | | | | | | |
| EE | 0,048 | 1 | | | | | | | | | | | |
| Ra | 0,035 | 0,059 | 1 | | | | | | | | | | |
| Rk | 0,062 | 0,019 | 0,297 | 1 | | | | | | | | | |
| Rg | 0,163 | 0,095 | 0,477 | 0,456 | 1 | | | | | | | | |
| OQ | 0,131 | 0,073 | 0,578 | 0,692 | 0,866 | 1 | | | | | | | |
| UP | 0,208 | 0,615 | 0,321 | 0,498 | 0,410 | 0,532 | 1 | | | | | | |
| NAQ | 0,126 | 0,103 | 0,066 | 0,153 | 0,146 | 0,178 | 0,036 | 1 | | | | | |
| QOQ | 0,183 | 0,111 | 0,080 | 0,139 | 0,189 | 0,194 | 0,036 | 0,633 | 1 | | | | |
| H1H2 | 0,306 | 0,214 | 0,036 | 0,054 | 0,143 | 0,111 | 0,139 | 0,190 | 0,299 | 1 | | | |
| HRF | 0,679 | 0,197 | 0,040 | 0,071 | 0,184 | 0,152 | 0,222 | 0,250 | 0,351 | 0,537 | 1 | | |
| PSP | 0,008 | 0,068 | 0,050 | 0,030 | 0,057 | 0,067 | 0,073 | 0,084 | 0,016 | 0,004 | 0,006 | 1 | |
| M | 0,040 | 0,076 | 0,048 | 0,060 | 0,065 | 0,065 | 0,018 | 0,048 | 0,103 | 0,151 | 0,010 | 0,060 | 1 |

***Tabla 5.3.2:*** Coeficientes de correlación entre distintos parámetros glotales para hombres

# 5.4 Análisis de parámetros

## *5.4.1 Diagramas de cajas*

En una primera observación propusimos unos diagramas de cajas. Para hacer este análisis hay que diferenciar entre géneros, en el proyecto solo muestro diagramas de cajas para el caso de 'male' ya que los resultados son muy similares. Cada color separado muestra un locutor distinto, dentro del cual cada una de las cajitas es un fonema encontrado en una locución. Buscamos la máxima variabilidad interlocutor y la mínima intralocutor. Si observamos el parámetro Ra es un ejemplo de parámetro con una peor intervariablidad entre locutores. Como anteriormente he dicho, este parámetro pertenece al grupo R-parameters que se basan en coger puntos temporales muy precisos y que por tanto suelen dar peores resultados que otros parámetros como NAQ,QOQ que se basan en cocientes de amplitudes.

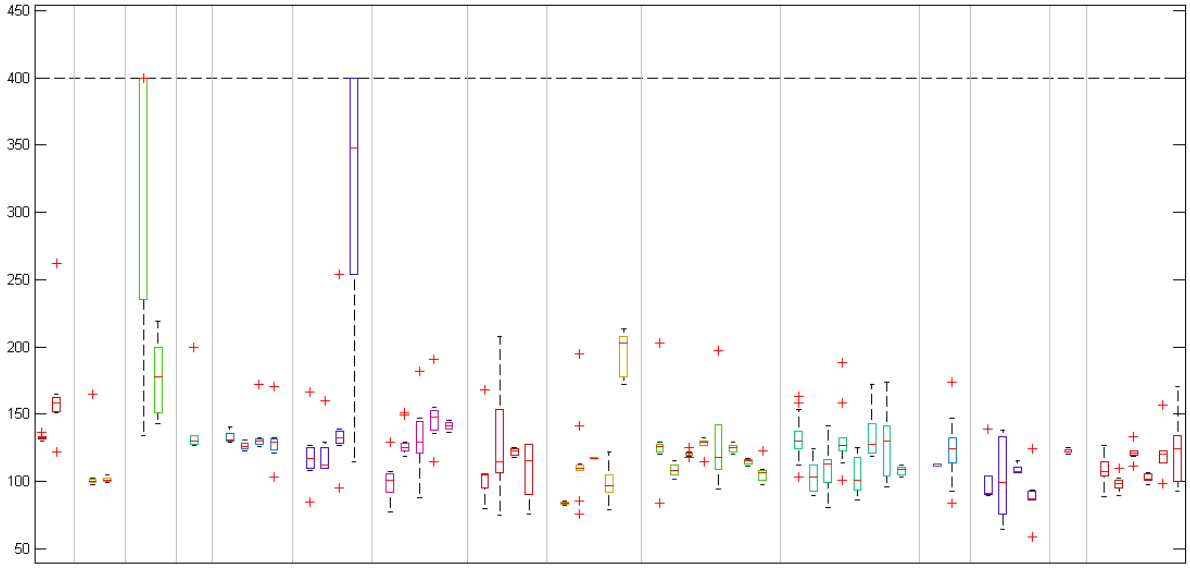



***Figura* 5.4.1.1*:*** Diagrama de cajas de F0 para hombre y para el fonema “AO”. Cada grupo entre líneas verticales es un locutor y cada columna es la distribución para una locución. En este experimento buscamos distribuciones parecidas dentro de un mismo locutor, y diferentes entre locutores.

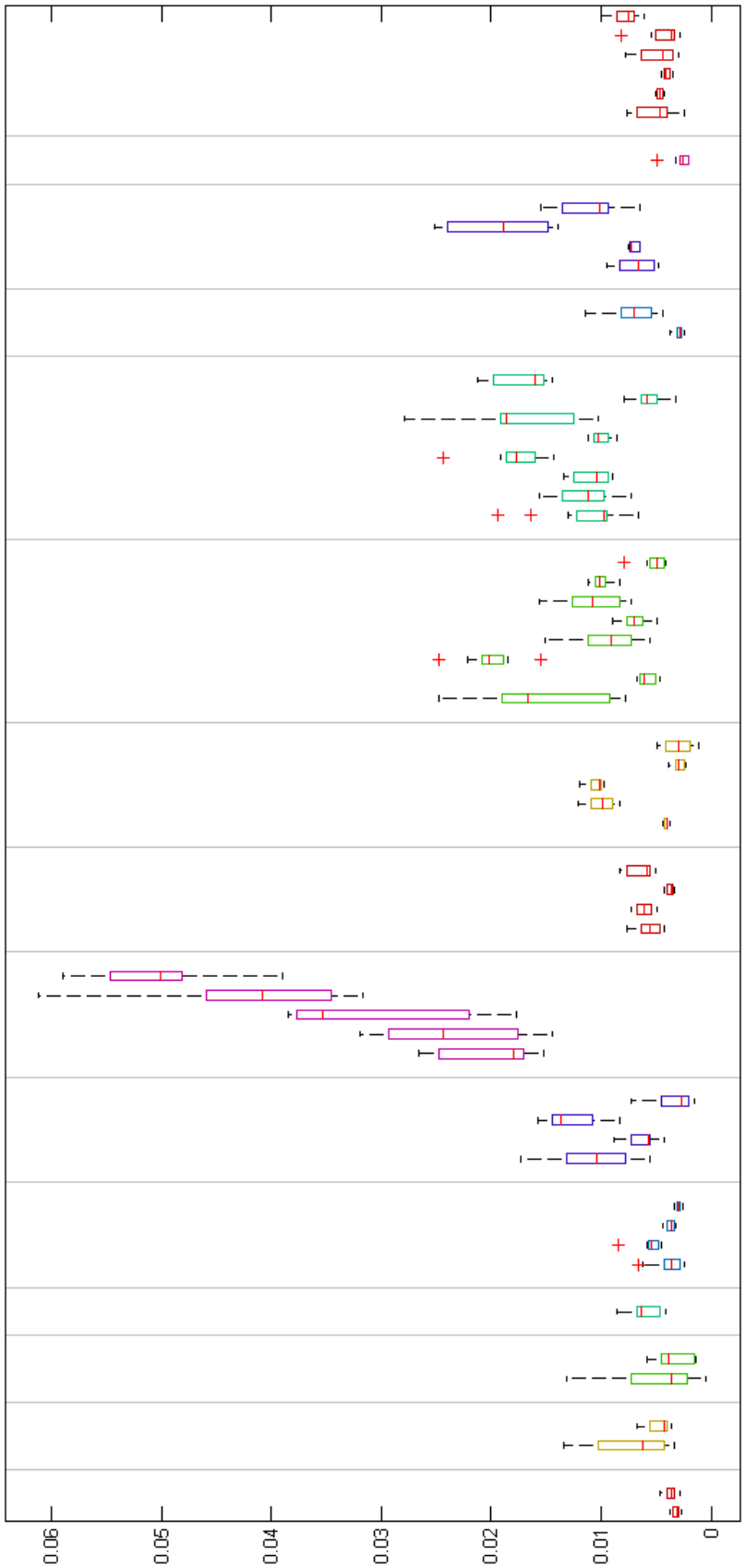



***Figura* 5.4.1.2:** Diagrama de cajas de EE para hombre y para el fonema "AO"

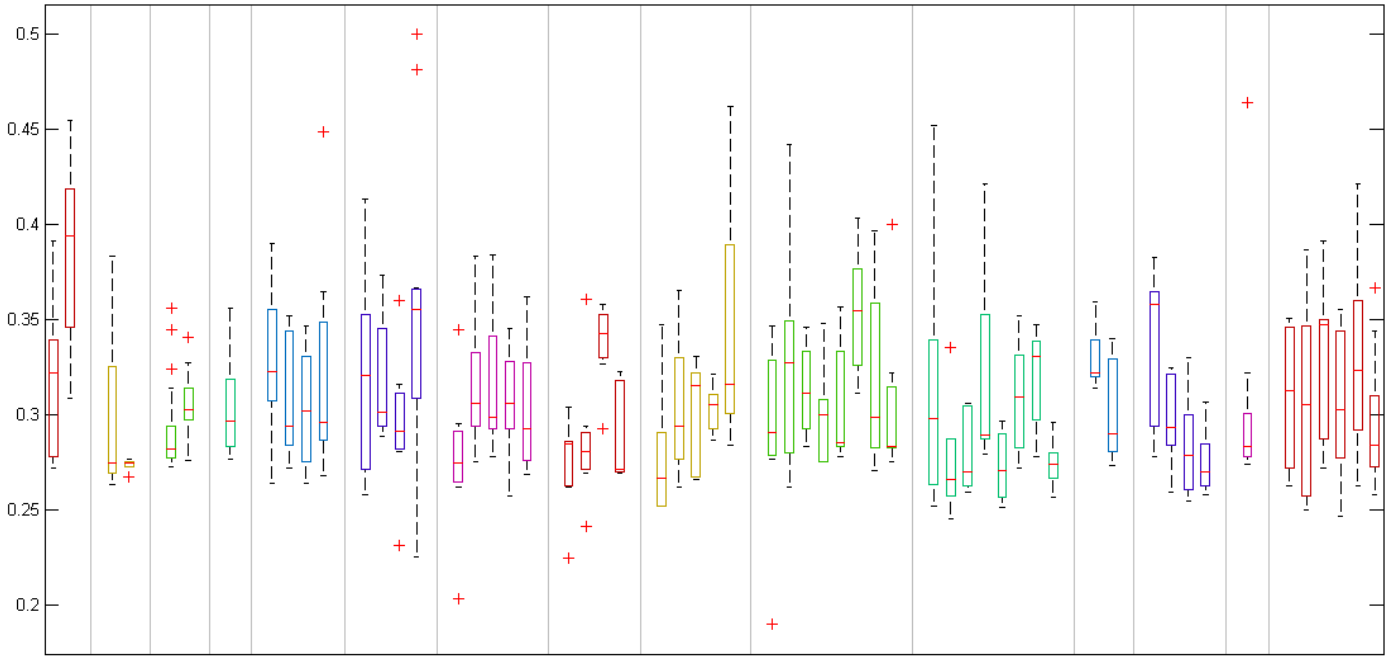



***Figura* 5.4.1.3*:*** Diagrama de cajas de Rk para hombre y para el fonema “AO”

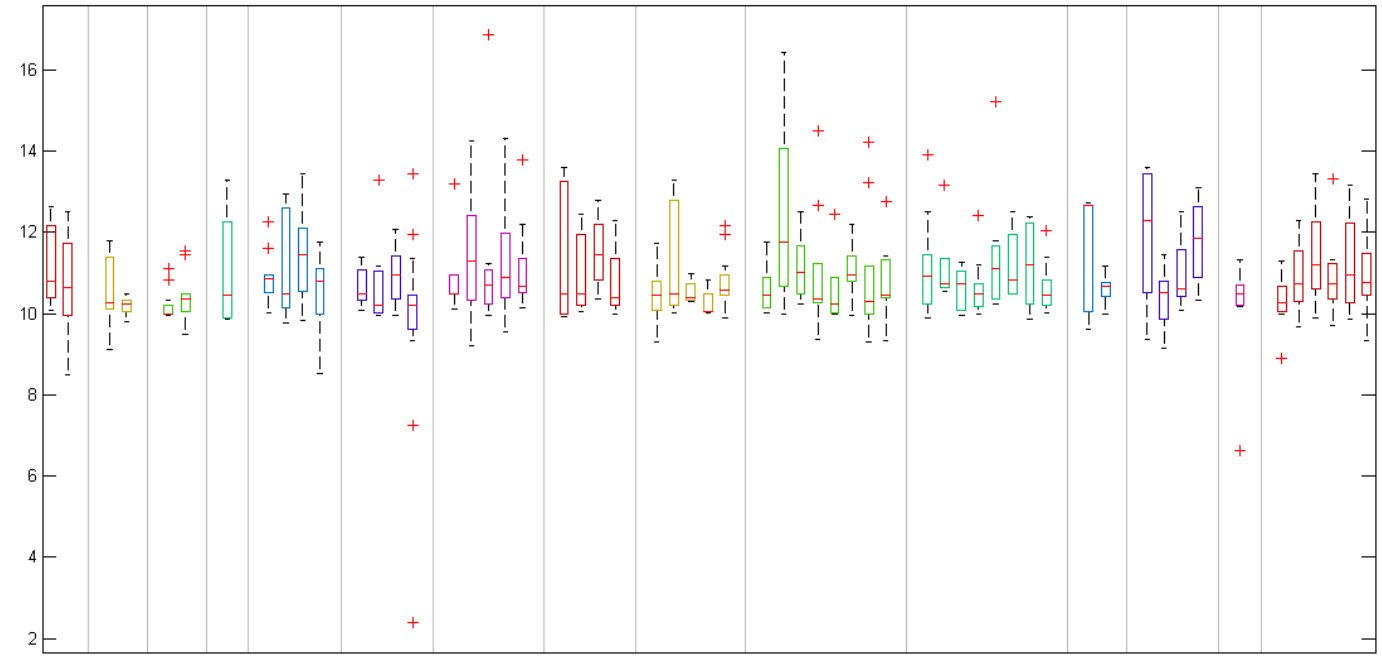



***Figura* 5.4.1.4*:*** Diagrama de cajas de Ra para hombre y para el fonema “AO”

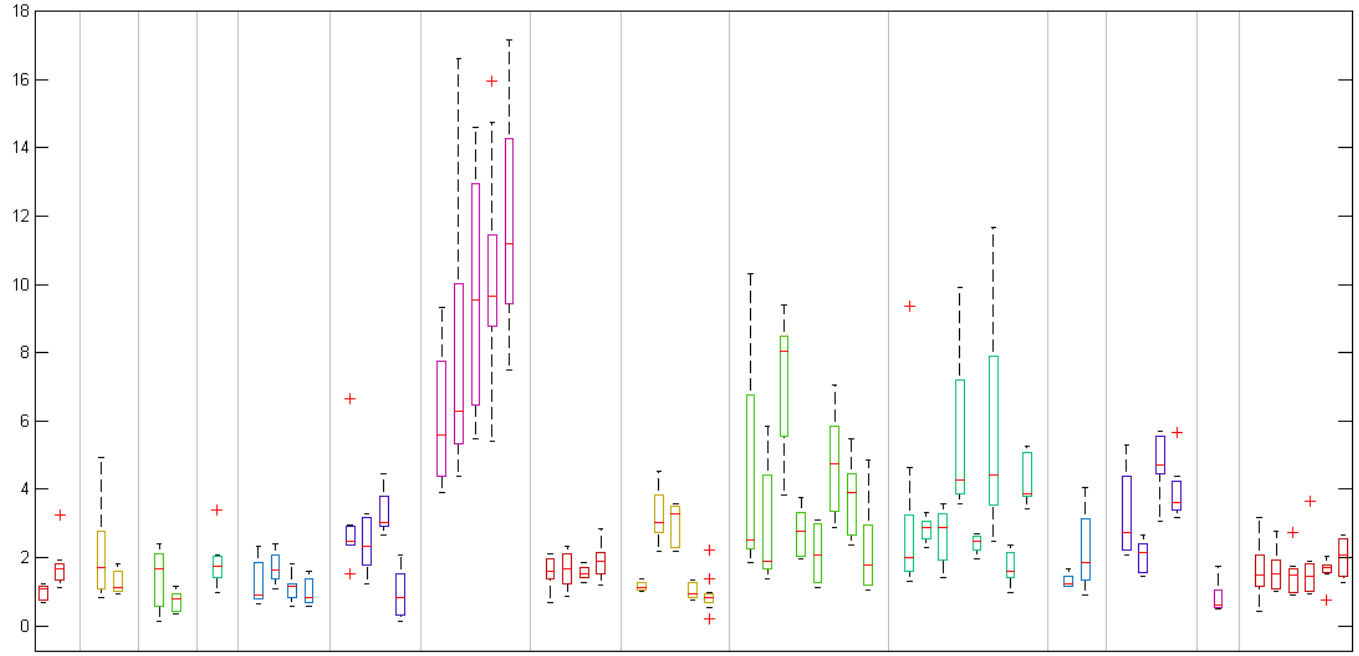



***Figura* 5.4.1.5*:*** Diagrama de cajas de UP para hombre y para el fonema “AO”

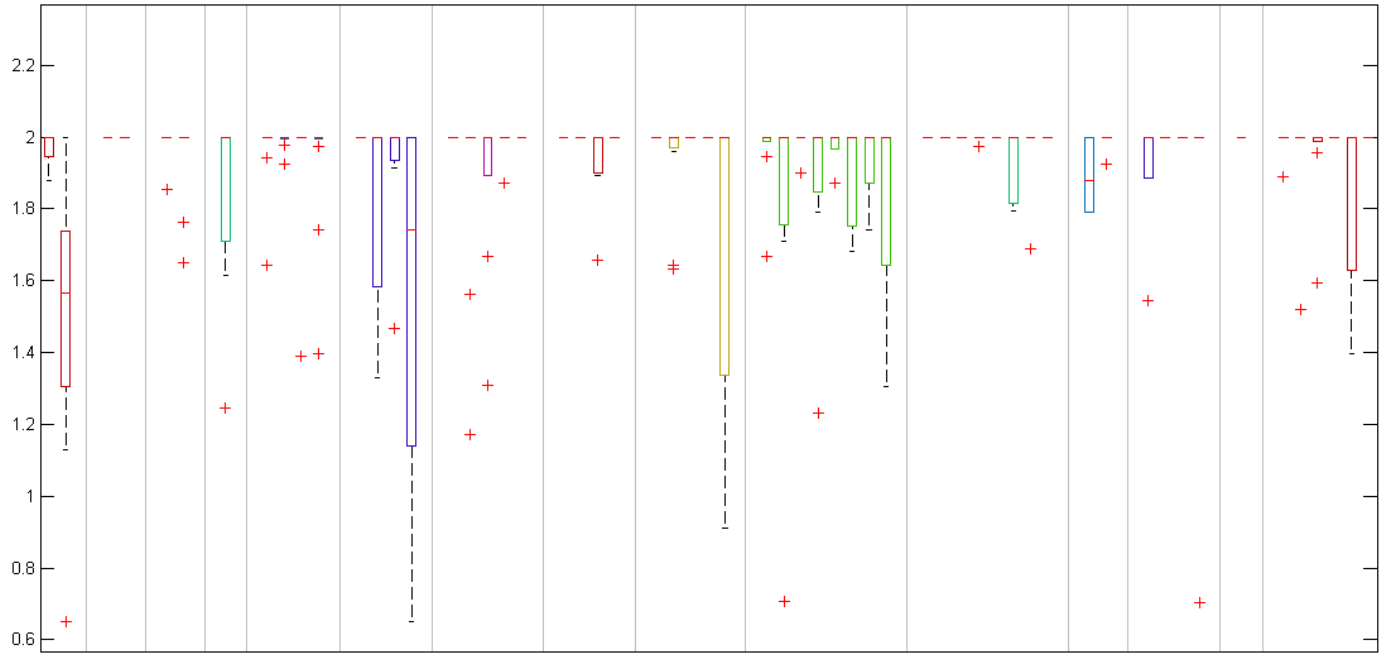



***Figura* 5.4.1.6*:*** Diagrama de cajas de Rg para hombre y para el fonema “AO”

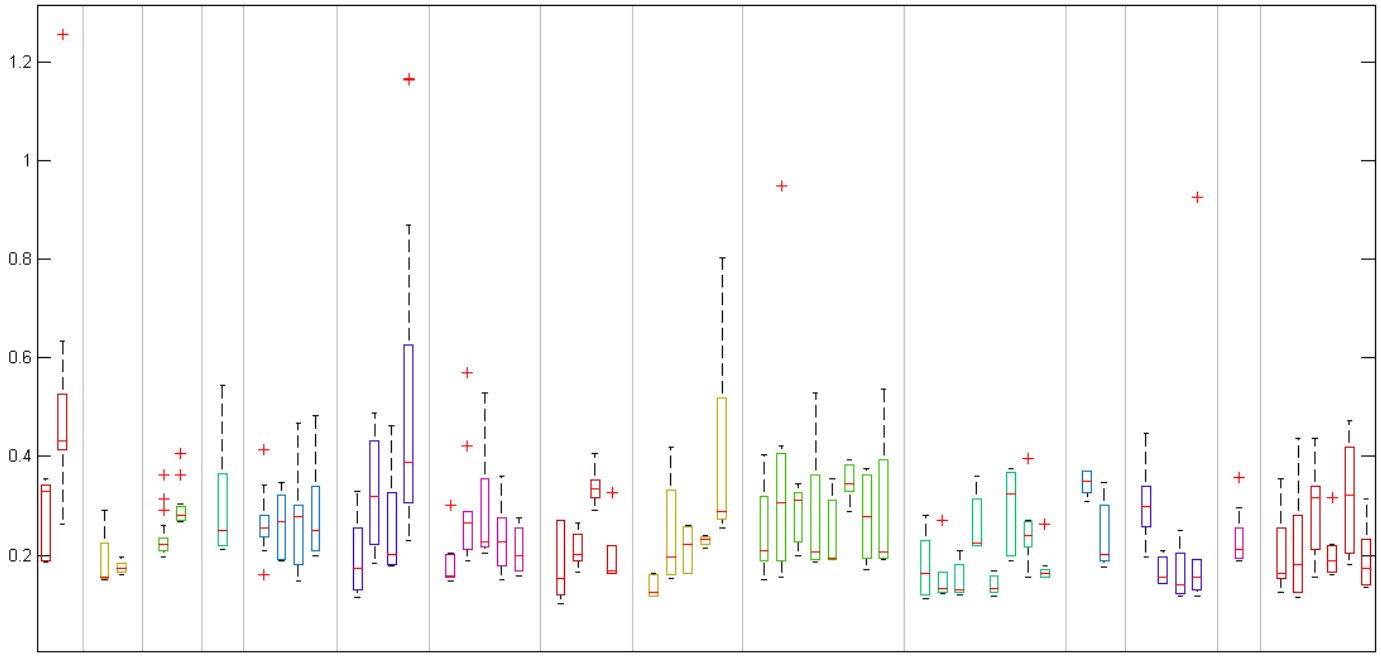



***Figura* 5.4.1.7*:*** Diagrama de cajas de OQ para hombre y para el fonema “AO”

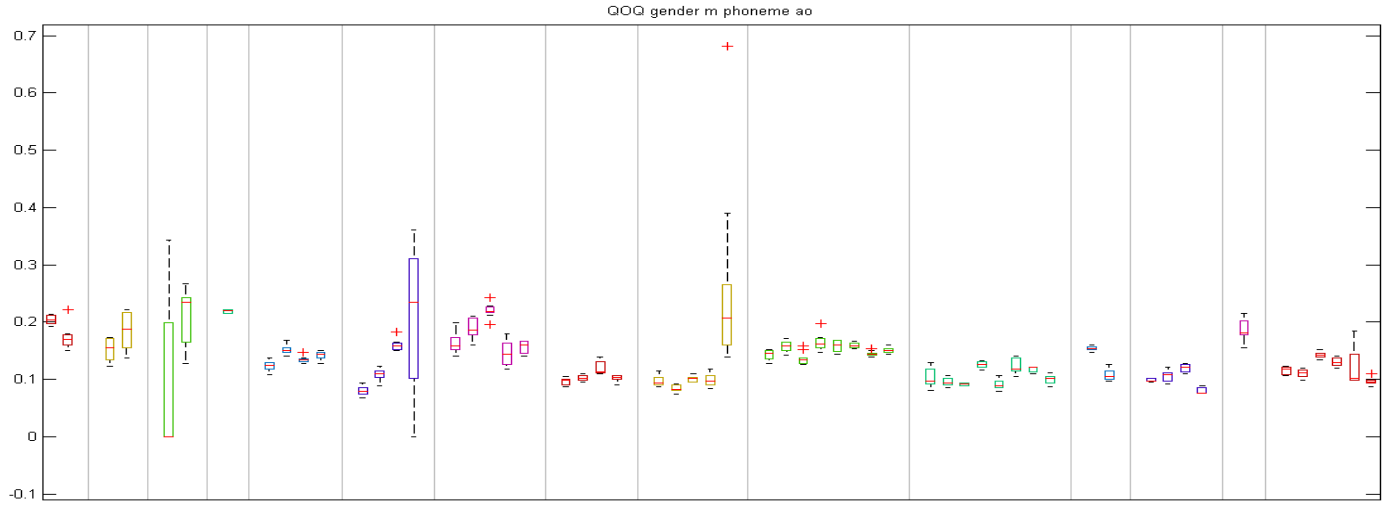



***Figura* 5.4.1.8*:*** Diagrama de cajas de QOQ para hombre y para el fonema “AO”

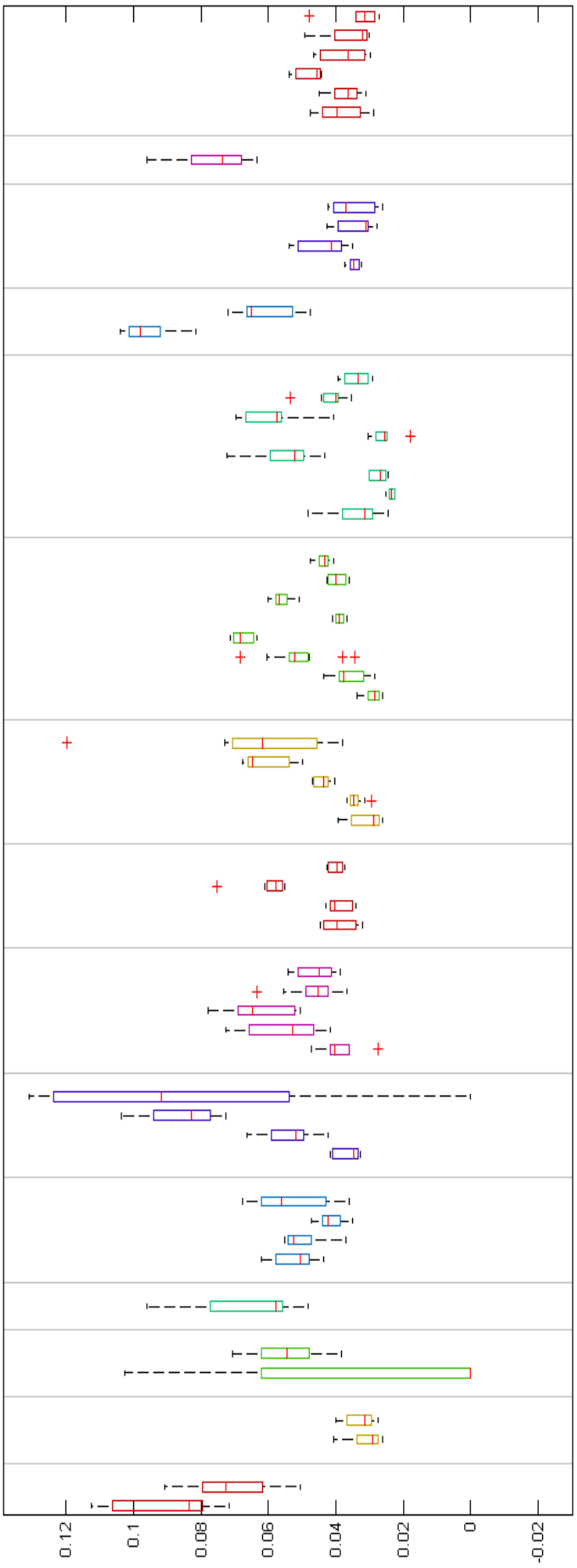



***Figura* 5.4.1.9:** Diagrama de cajas de NAQ para hombre y para el fonema "AO"

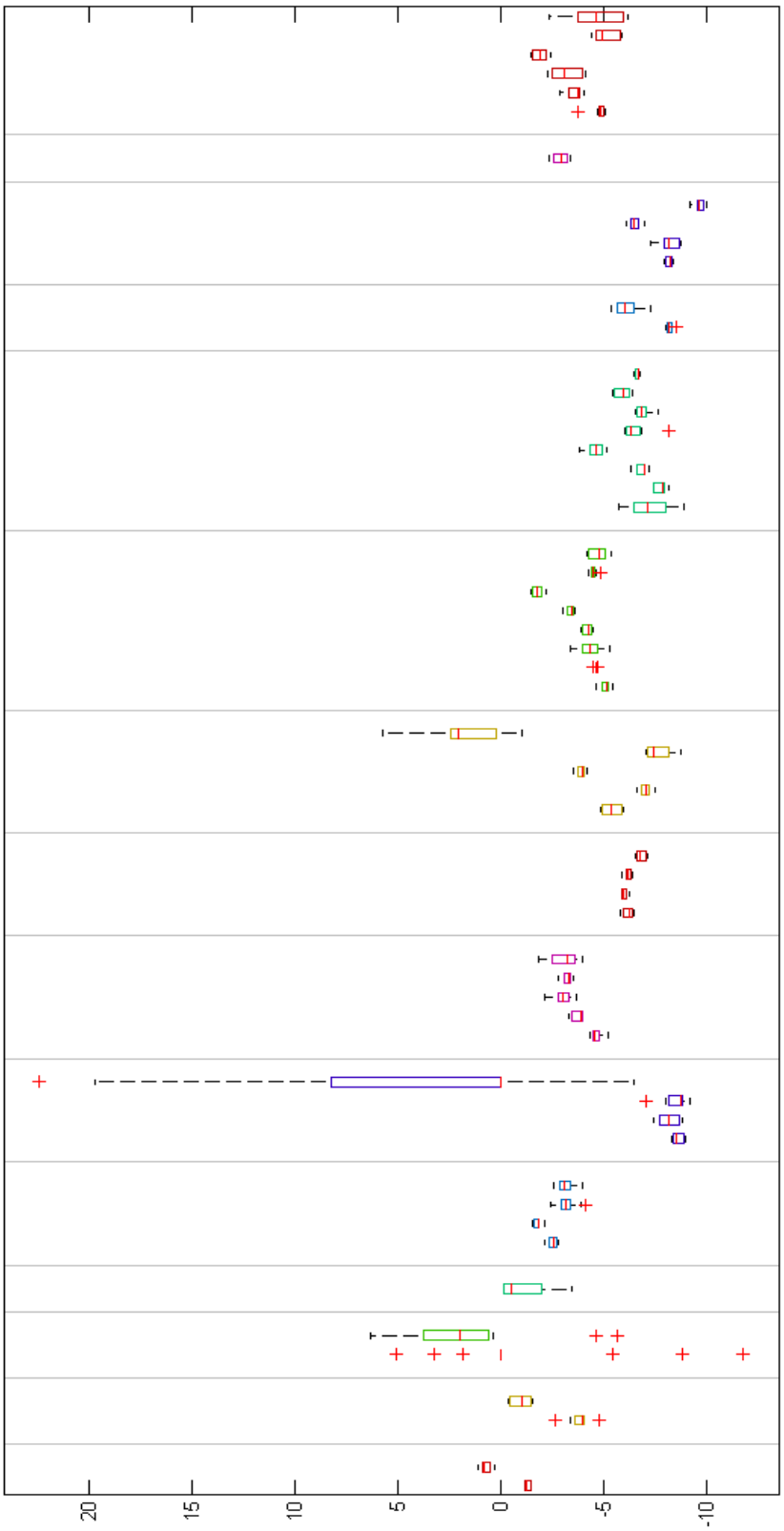



***Figura* 5.4.1.10:** Diagrama de cajas de H1H2 para hombre y para el fonema "AO"

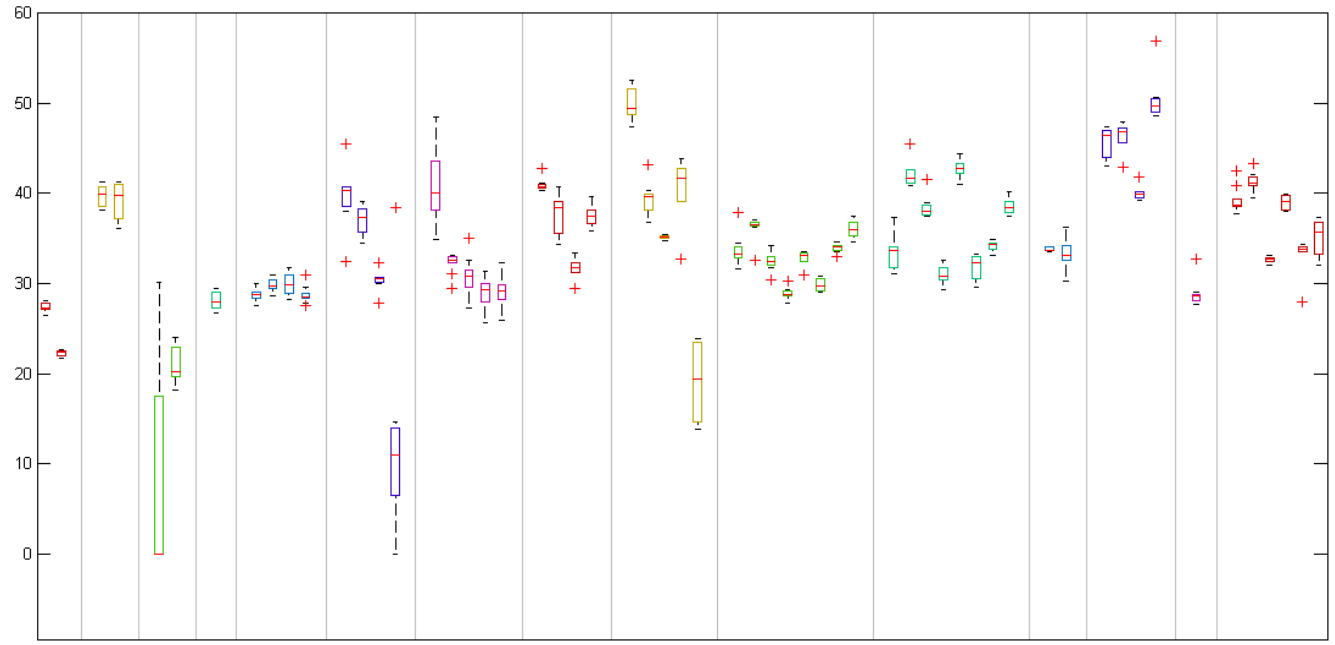



***Figura* 5.4.1.11*:*** Diagrama de cajas de HRF para hombre y para el fonema “AO”

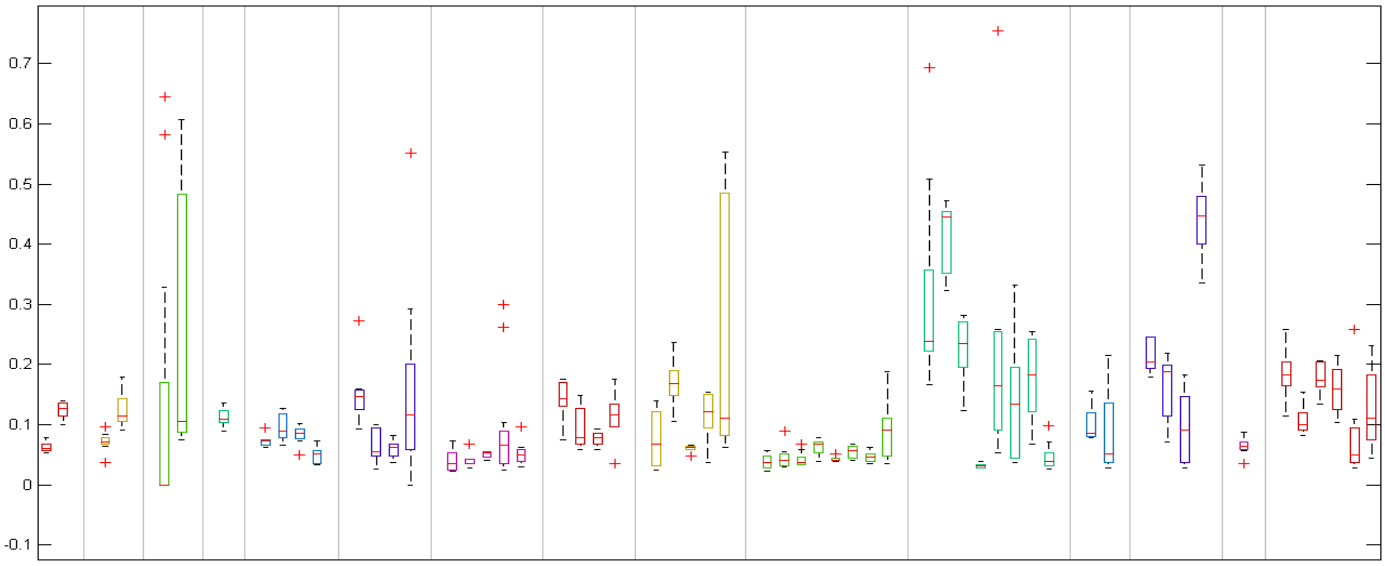



***Figura* 5.4.1.12*:*** Diagrama de cajas de PSP para hombre y para el fonema “AO”

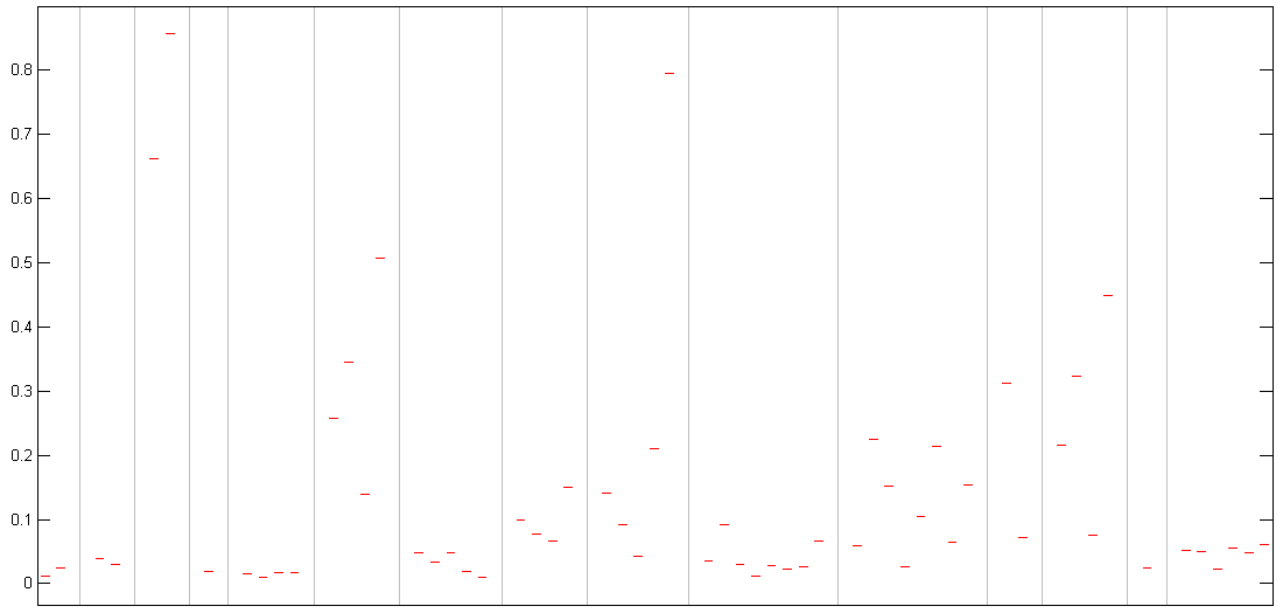



***Figura* 5.4.1.13*:*** Diagrama de cajas de *'creak'* para hombre y para el fonema “AO”

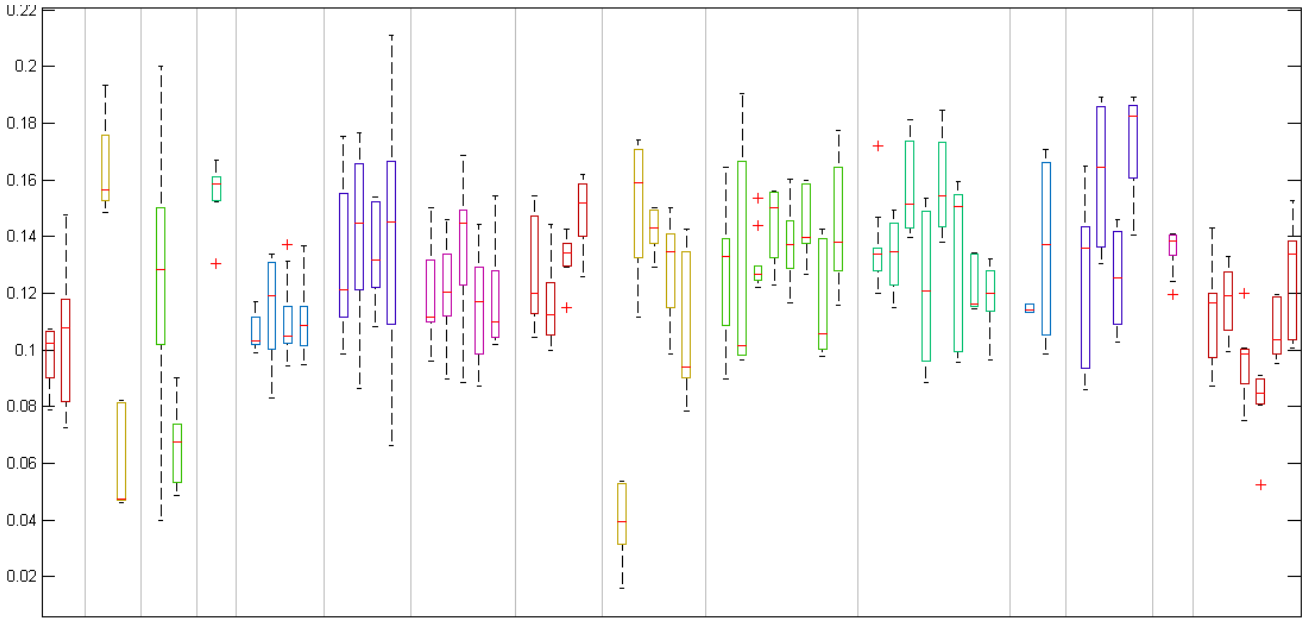



***Figura* 5.4.1.14*:*** Diagrama de cajas de MDQ para hombre y para el fonema “AO”

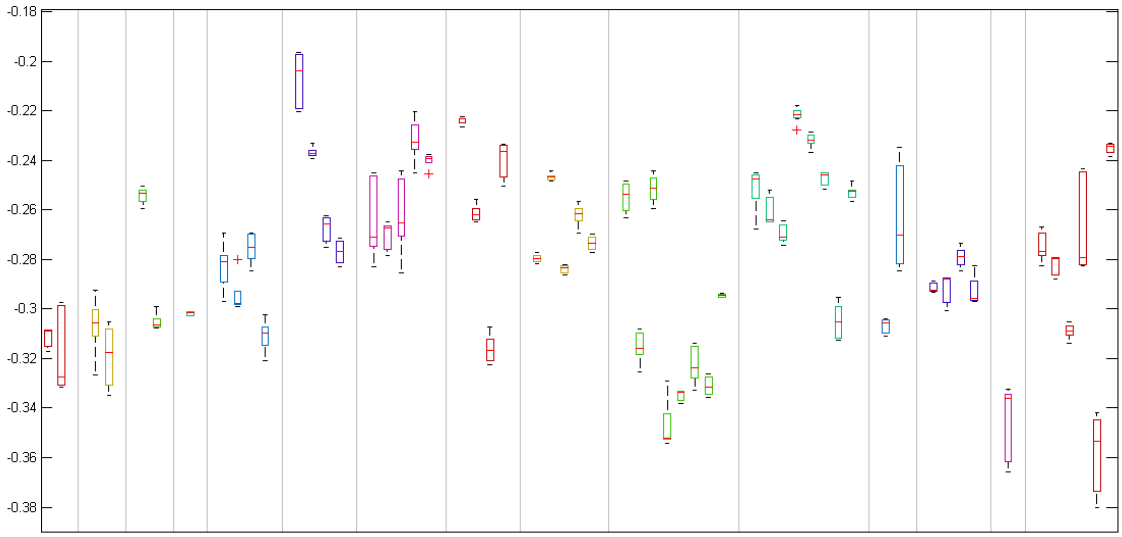



***Figura* 5.4.1.15*:*** Diagrama de cajas de Peakslope para hombre y para el fonema "AO"

Los diagramas de cajas anteriormente expuestos dan una primera referencia de la variabilidad proporciona cada uno de los parámetros para distintos locutores de manera visual. Sin embargo para hacer un estudio más detallado sobre la eficacia de los distintos parámetros que englobe un número de locutores mucho mayor tenemos que proponer una medida que podamos cuantificar. Por lo que ahora vamos a ver proponer dos medidas la distancia entre gaussianas y cociente de varianzas. En este análisis es para 200 locutores 70 mujeres y 130 hombres.

## Cociente de varianzas

El cociente de varianzas consiste en una división en la que en el numerador aparece la varianza de un parámetro en un fonema para un locutor dado y en el cociente juntamos todos los valores de ese mismo parámetro de ese mismo tipo fonema pero para todos los locutores y hacemos la varianza. Por último, hacemos el promedio de estos cocientes de varianzas, ya que tendremos uno por locutor, para conseguir finalmente una medida por parámetro para cada clase de fonema.

Por si no ha quedado claro pondré un ejemplo: primero escogemos un tipo de fonema sobre el que queremos sacar los distintos cocientes de varianzas (uno por parámetro ). Para sacar por ejemplo el cociente de varianzas del parámetro EE cogeremos todos los valores de EE producidos por el locutor 1, y dividiremos esto entre el conjunto de valores de EE producido por todos los locutores (para este primer estudio 130 hombres, 70 mujeres), por último hacemos el promedio de esta medida ya que tendremos una por locutor y queremos tener tan solo una.

**"AA"**

| | Male | | Female | |
|---|---|---|---|---|
| | **100%** | **50%** | **100%** | **50%** |
| **FO** | 0,844 | 0,990 | 0,731 | 0,557 |
| **EE** | 0,468 | 0,382 | 0,461 | 0,339 |
| **Ra** | 0,996 | 0,977 | 0,975 | 0,980 |
| **Rk** | 0,921 | 0,898 | 0,904 | 0,866 |
| **Rg** | 0,904 | 0,855 | 0,903 | 0,857 |
| **OQ** | 0,927 | 0,988 | 0,889 | 0,823 |
| **UP** | 0,566 | 0,486 | 0,675 | 0,612 |
| **NAQ** | 0,722 | 0,524 | 0,656 | 0,520 |
| **QOQ** | 0,853 | 0,621 | 0,710 | 0,503 |
| **H1H2** | 0,741 | 0,490 | 0,714 | 0,515 |
| **HRF** | 0,700 | 0,563 | 0,624 | 0,479 |
| **PSP** | 0,835 | 0,743 | 0,841 | 0,602 |
| ***'creak'*** | 0,672 | 0,672 | 0,485 | 0,485 |
| **MDQ** | 0,947 | 0,763 | 0,892 | 0,591 |
| **Peak slope** | 0,424 | 0,368 | 0,395 | 0,336 |

***Tabla 5.4.2.1:*** Cociente de varianzas para el fonema "aa": Parámetro del locutor y parámetro para todos los locutores

Muestro sólo un ejemplo para el fonema 'aa' pero ya es bastante representativo para observar lo que estábamos buscando. Las columnas que aparecen con un 50% quiere decir que si el fonema tiene, por ejemplo, 10 pulsos glotales nos hemos quedado con los 5 del medio. El resultado de esta medida es mejor cuanto más pequeño es el valor ya que esto representa que la varianza de un parámetro para un locutor es más pequeña que la varianza de ese mismo parámetro para el conjunto de locutores. Se vuelve a corroborar lo que deducimos del diagrama de cajas sobre qué parámetros nos van a convenir más para formar nuestro vector, y estos parámetros son aquellos que no se basan en localizar instantes temporales precisos para ser obtenidos sino en cocientes de amplitudes que son además más robustos ante el ruido.

Como ya he comentado en el cociente de varianzas nos conviene un valor lo más pequeño posible y cómo podemos observar en el caso que cogemos el 50% de los valores de la ventana en general dan números menores. A priori pensamos que esto podía ser debido a que el algoritmo produce valores más estables en regiones centrales de la ventana ya que tardaba en estabilizarse, pero en estudios posteriores donde iba variando el porcentaje de valores con los que me quedaba en cada ventana me di cuenta que siempre que reducía el porcentaje quedaba un número de cociente de covarianzas menor, y la explicación es que al ser vectores de dimensiones tan pequeñas (de 7 a 15) eliminar elementos va suponer una reducción de la varianza casi siempre aunque realmente la estabilidad de los valores del vector sea similar en todas las regiones de él. Por tanto, finalmente después de darnos cuenta de esto y después de ver multitud de derivadas de pulsos glotales extraídos y ver que no existe ninguna preferencia regiones del pulso glotal para valores estables decidimos quedarnos con todos los valores ya que es más información.

## Distancia de gaussianas

La idea es la misma que en el apartado anterior pero la nueva medida es una distancia entre gaussianas utilizando la distancia Hellinger. La distancia Hellinger al cuadrado viene dada por esa fórmula (1) donde f y g son funciones de densidad de probabilidad. La distancia tiene un rango de cero a uno, siendo uno la máxima distancia y cero la mínima. La formula de abajo representa la particularización para cuando las dos distribuciones de densidad de probabilidad son gaussianas. Esta distancia está entre cero y uno porque resolviendo la segunda igualdad de la fórmula (1) a la segunda igualdad resolviendo el paréntesis y sabiendo que la integral de una función de densidad de probabilidad en todo su dominio da. Entonces fijándose en la última igualdad de la fórmula (1) cuando las funciones de densidad de probabilidad no coinciden nada el resultado de la raíz cuadrada da cero y la distancia Hellinger da uno que representa la máxima distancia entre dos gaussianas. Por otro lado cuando f y g son exactamente iguales el resultado tanto de la raíz cuadrada como de la integral es uno y la distancia Hellinger es cero representando la mínima distancia entre gaussianas

### *Distancia Hellinger*

$$H^2(f,g) = \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx = 1 - \int \sqrt{f(x)g(x)}\,dx \quad (1)$$

$$H^2(P,Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\, e^{-\frac{1}{4}\frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}} \quad (2)$$

$$0 \le H(P,Q) \le 1$$

Por tanto los resultados que se muestran en la tabla se basan en lo mismo que en el caso de cociente de varianzas. En este caso una gaussiana se forma con los valores de un parámetro para un locutor y la otra gaussiana con los valores de ese mismo parámetro pero en todos los locutores. Vamos a tener una distancia entre gaussianas por locutor, pero lo que mostramos en la tabla es promedio para todos los locutores.

**Distancia entre Gaussianas**

| | "aa" | | "ao" | |
|---|---|---|---|---|
| | **male** | **female** | **male** | **female** |
| **FO** | 0,507 | 0,496 | 0,517 | 0,617 |
| **EE** | 0,432 | 0,434 | 0,413 | 0,472 |
| **Ra** | 0,108 | 0,073 | 0,120 | 0,110 |
| **Rk** | 0,142 | 0,116 | 0,125 | 0,119 |
| **Rg** | 0,145 | 0,113 | 0,164 | 0,125 |
| **OQ** | 0,194 | 0,128 | 0,206 | 0,156 |
| **UP** | 0,369 | 0,311 | 0,355 | 0,347 |
| **NAQ** | 0,268 | 0,266 | 0,273 | 0,353 |
| **QOQ** | 0,276 | 0,248 | 0,285 | 0,312 |
| **H1H2** | 0,269 | 0,242 | 0,270 | 0,363 |
| **HRF** | 0,334 | 0,322 | 0,324 | 0,417 |
| **PSP** | 0,317 | 0,717 | 0,307 | 0,194 |
| ***'creak'*** | 0,535 | 0,717 | 0,544 | 0,722 |
| **MDQ** | 0,123 | 0,142 | 0,150 | 0,192 |
| **Peak slope** | 0,368 | 0,381 | 0,348 | 0,415 |

***Tabla 5.4.2.2:*** Distancia entre gaussianas (distancia Hellinger) del experimiento que busca la mímina intravariabilidad y máxima intervariabilidad entre locutores

En este caso nos conviene los valores más próximos a uno ya que representan más separación con respecto al total de locutores. La conclusión de estos primeros análisis sobre parámetros es que los parámetros que nos van a interesar van a ser F0, EE, UP NAQ, QOQ, H1H2, HRF, PSP, *'creak'*, MDQ y Peak slope.

## 5.5 Proposición del vector

Lo primero que hacemos es proponer un vector basado en conclusiones sacadas del apartado anterior. El vector propuesto tiene una dimensión de 17 y cuenta con los siguientes campos:

[ media(EE) var(EE) media(NAQ) var(NAQ) media(QOQ) var(QOQ) media(H1H2) var(H1H2) media(HRF) var(HRF) media(PSP) var(PSP) media(*'creak'*) media(MDQ) var(MDQ) media(ps) var(ps) ]

Como TIMIT no cuenta con excesivos datos para realizar un estudio diferenciando dentro de cada fonema vocal, de ahora en adelante por locutor se producirán una serie de vectores que corresponderán indistintamente a los fonemas aa, iy, eh, ey, ae, aw, ay, oy, ow, er y uw. Como digo, lo más correcto sería hacer un estudio distinto para cada vocal. Pero contamos tan solo con unas 50 vocales por locutor y 5 vocales por locución de media con más de 100 ms de duración.

## 5.6 Representación de datos multidimensionales usando TSNE

Ahora utilizamos una herramienta llamada TSNE que sirve para representar datos multidimensionales, le pasamos los vectores de 10 y 15 locutores para que se puede apreciar visualmente las agrupaciones que se realizan para vectores de un mismo locutor en distintos fonemas. En aquellas regiones donde estén las agrupaciones es donde el UBM colocará los centro de sus gaussianas.

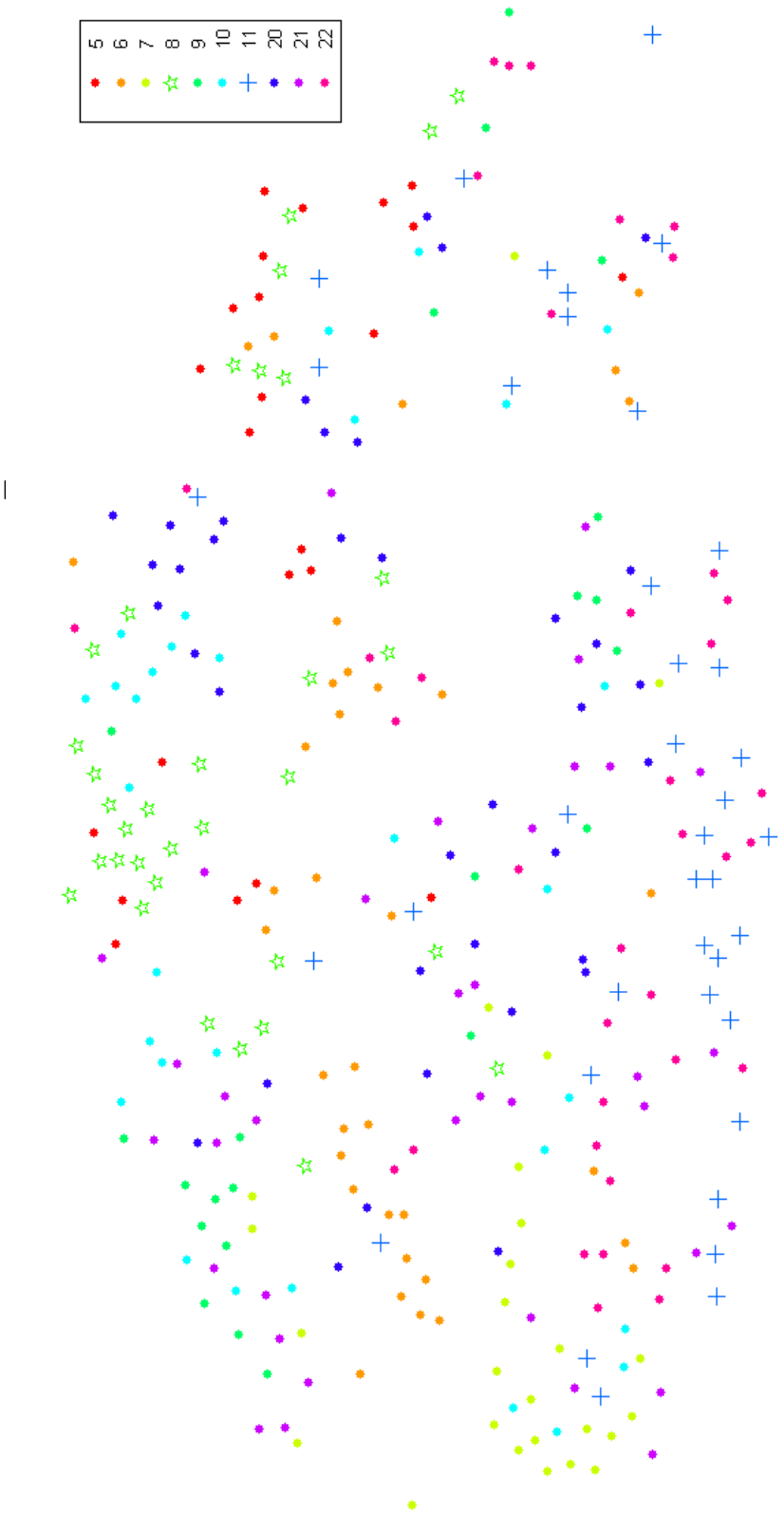



***Figura* 5.6.1:** Representación de datos multidimensionales utilizando TSNE para 10 locutores

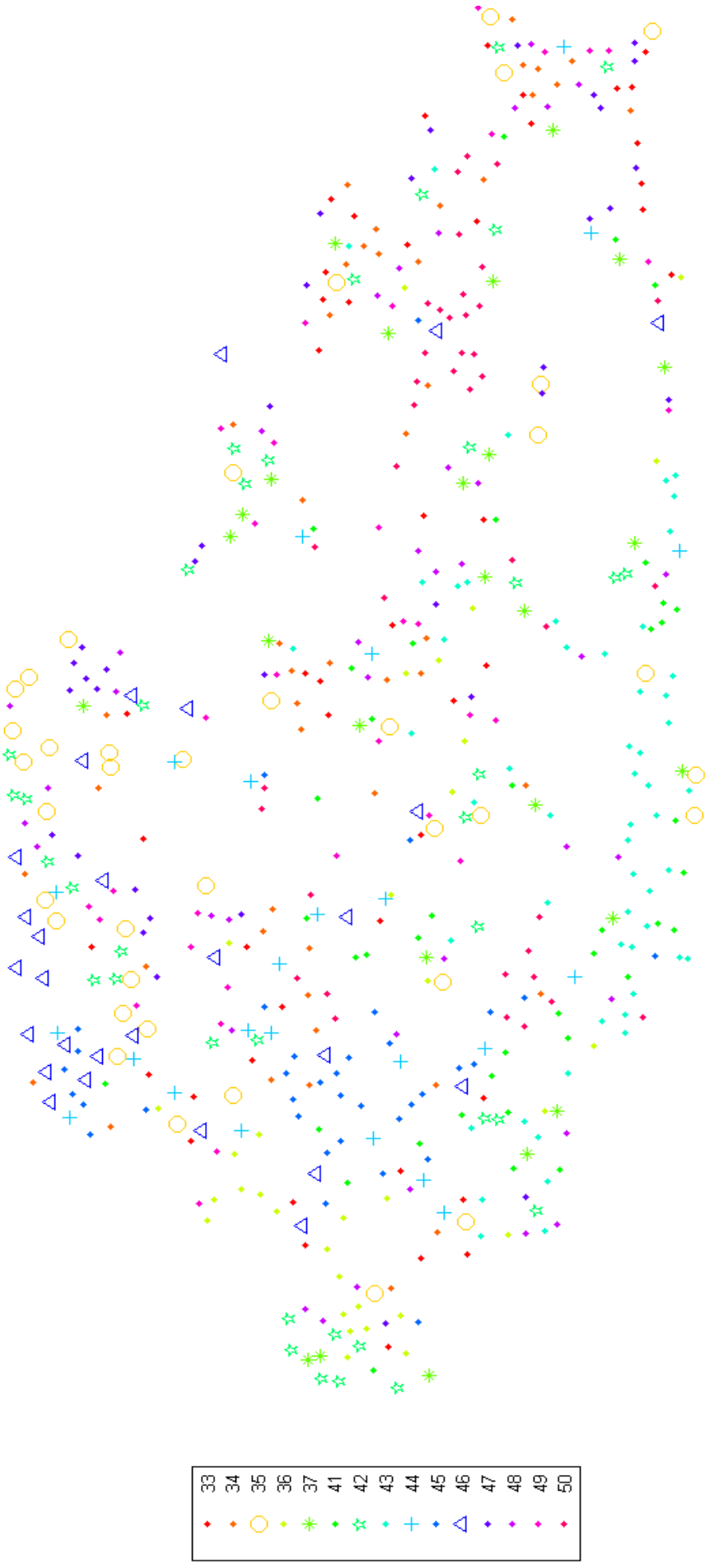



***Figura 5.6.2:*** Representación de datos multidimensionales utilizando TSNE para 15 locutores

# 5.7 Scoring mediante la técnica UBM-GMM-MAP

Para representar numéricamente cuan bueno es nuestro lo sistema a la hora de caracterizar locutores a través de parámetros glotales utilizamos la técnica UBM-GMM-MAP. Esta técnica consta de tres pasos fundamentales: en el primero de ellos se entrena un UBM con gran cantidad de datos de locutores que no volverán a usarse para sacar resultados. Un UBM es una mezcla de gaussianas multidimensionales, si datos entre distintos locutores son bastante diferentes esto estará representado por gaussianas separadas en el espacio multidimesional. Cada una de estas dimensiones la forma un coeficiente del vector que proponemos para caracterizar a un locutor. El siguiente paso es sacar todos los datos provenientes de un locutor distinto de los que hemos usado para crear el UBM y ver donde se encajan sus datos en el UBM antes formado. Por último se calculan las similitudes entre el modelo del locutor ajustado en el UBM y vectores del mismo locutor (*scoreTarget) y* similitudes entre el modelo del locutor ajustado en el UBM y vectores de otro locutor (*scoreNontarget*). El resultado más favorable que se puede obtener es que los *scoreTarget* sean valores los más altos posibles con respecto a los *scoreNontarget*.

## 5.7.1 División de los locutores

Se ha realizado la siguiente división de locutores en estos tres modulos:

| | UBM | Development | Evaluation |
|---|---|---|---|
| **Porcentaje** | 50% | 25% | 25% |
| **Hombres** | 219 | 110 | 109 |
| **Mujeres** | 96 | 48 | 48 |

Cada locutor cuenta con diez locuciones y en cuanto a la repartición de las locuciones para *train* y *test*, distinguimos para los primero experimentos entre dos casos:

1. **50% *train* y 50 % para test.**
2. **70% *train* y 30% para test.**

## 5.7.2 Proposición del score

El score es un valor númerico que representa la similitud existente entre el conjunto de datos entrenados de un locutor y una serie de vectores, de manera que cuanto más alto sea este valor más se parecen los dos conjuntos.Para nuestro sistema definimos el siguiente:

$$\text{Score} = \text{ML}(\text{vectortest}_{\text{locutor i}}, \text{MixUser}_{\text{locutor i}}) \,/\, \text{ML}(\text{vectortest}_{\text{locutor i}}, \text{UBM})$$

Donde 'MixUser' es el modelo adaptado que saca la función MAP por locutor cuando le entra todos los vectores que ese locutor tiene para *train*, y 'ML' es la máxima similitud. Por tanto la interpretación de este score es la siguiente: en el numerador nos conviene el número más alto posible ya que eso significará que el modelo adaptado del usuario es muy parecido a sus vectores de test con los que como está claro no se ha construido el modelo adaptado. Por tanto, cuanto mayor que uno sea el score mejor va a ser para nosotros ya que tendrá como consecuencia que el vector de test se parece más al modelo adaptado del locutor correspondiente que al UBM entrenado con todos los locutores.

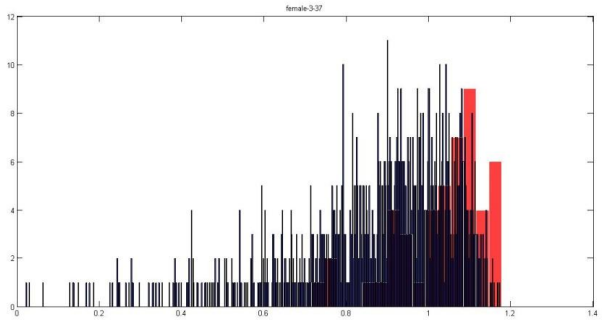
## 5.7.3 Distribuciones de los scores target y non-target

Para nuestros primeros experimentos, el número de centros de gaussianas para male es 32 y para female 16. dentro de la división 70% locuciones para *train* y 30% para test distinguiremos dos casos, donde se tratan las 3 locuciones de test como un bloque conjunto, por lo tanto solo se produce un scoreTarget por locutor, y donde cada locución se trata de manera individual de manera que produce un scoreTarget distinto por cada locución. Las gráficas que se muestran a continuación muestran las distribuciones de los scoresTarget (en rojo bloques más anchos) y scoresNonTarget (en azul y bloques más finos) para los casos antes descritos.

1. 70% *train*, 30% test, bloques de 3 locuciones(1 scoretarget por locutor)

notación male/female 3-37

**Female** **Male**



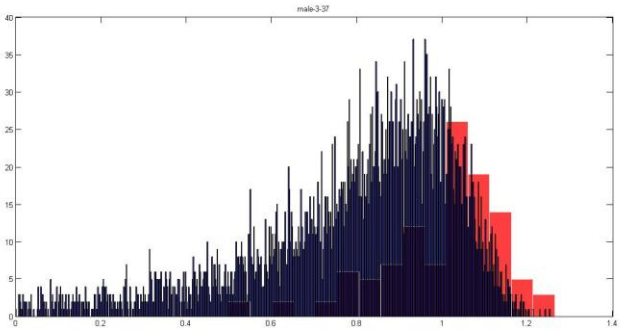
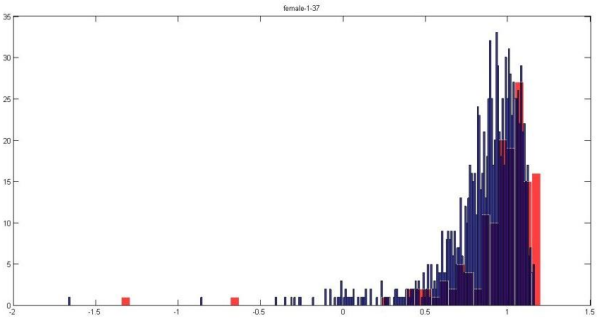




***Figura* 5.7.3.1*:*** Distribuciones de los scores, target y nontarget para buscar la mayor distancia entre dichas distribuciones .

2. 70% *train*, 30% test, bloques de 1 locución (1 scoretarget por locución)

notación male/female 1-37

**Female** **Male**



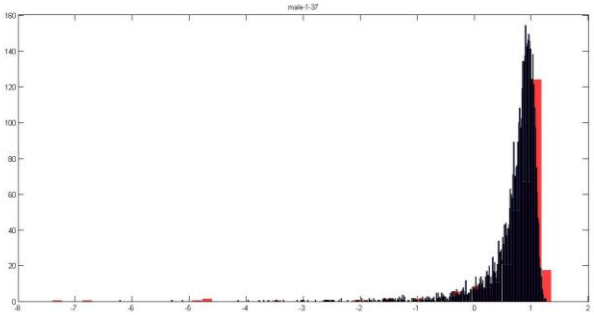
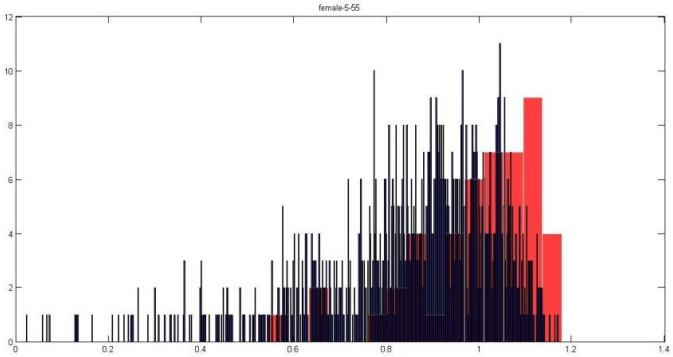




***Figura* 5.7.3.2*:*** Distribuciones de los scores , target y nontarget para buscar la mayor distancia entre dichas distribuciones .

3. 50% *train*, 50% test, bloques de 5 locuciones (1 scoretarget por locutor),

notación male/female 5-55

**Female** **Male**



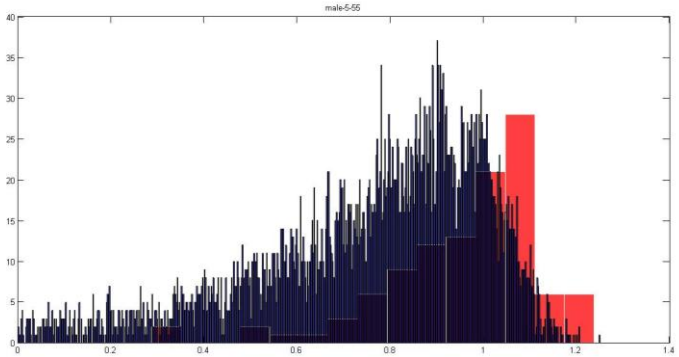




***Figura* 5.7.3.3*:*** Distribuciones de los scores, target y nontarget para buscar la mayor distancia entre dichas distribuciones .

En los tres casos se observa cierta discriminación. Los scores target quedan con una media por encima de uno quedando por encima de la media de los nonTarget que está sobre 0.9.

De ahora en adelante en vez de mostrar para cada caso la distribución de las dos gaussianas ( scores target y nonTarget ) se dará el EER ( *Equal Error Rate* )que es un valor que representa el error mínimo cuanto intentamos discriminar entre ambas distribuciones y que se da cuando la tasa de falsos positivos y falsos negativos es la misma.

Si por ejemplo tenemos las dos distribuciones que muestra la figura, tenemos que colocar un umbral, de manera que las muestras que queden por debajo de él queden asignadas a la distribución uno y todas las muestras que queden por encima del umbral queden asignadas a la distribución dos. El umbral donde se comete menos error es el que iguala la tasa de falsos positivos a falsos negativos, y el error que se comete tomando ese umbral es el EER.

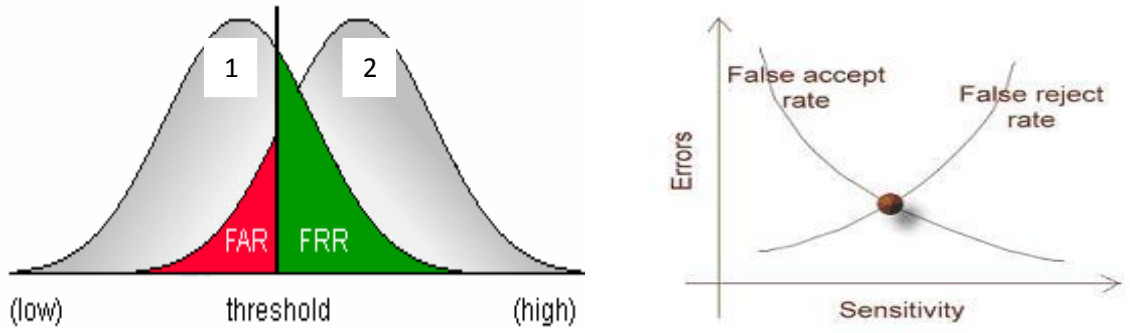



***Figura* 5.7.3.4** Figura de la izquierda: distribución de scores target y non-target y acepatciones y rechazos en función del umbral. A la derecha: representación de FA y FR en función del umbral

| | male-3-37 | female-3-37 | male-5-55 | female-5-55 | male-1-37 | female-1-37 |
|---|---|---|---|---|---|---|
| EER | 28,97 | 28,41 | 30,39 | 31,34 | 41,91 | 37,59 |

***Tabla 5.7.3:*** EER para las distintas agrupaciones de datos con el factor Reynolds a 16 y 20 centros. (3-37: bloques de 3 locuciones con 30% para test y 70% para train, 5-55: bloques de 5 locuciones con 50% para test y 50% para train, 1-37: bloques de 1 locución con 30% para test y 70% para train)

## 5.7.4 Faunagramas

Un faunagrama es una figura que sirve para analizar el comportamiento de un sistema cuya salida son las diferentes puntuaciones de las comparaciones entre el rasgo introducido y los almacenados en la base de datos. Este tipo de figura muestra el comportamiento de todos los scores de cada experimento, haciendo sencilla la visualización del alineamiento de los scores entre las distintas comparaciones. Cada fila del gráfico se corresponde con una lista de candidatos para una búsqueda distinta, y en cada una de las filas aparecen los scores correspondientes a la muestra genuina en forma de círculo azul, y los scores de muestras no genuinas en forma de cruz roja. El EER se muestra para cada comparación con un asterisco negro.

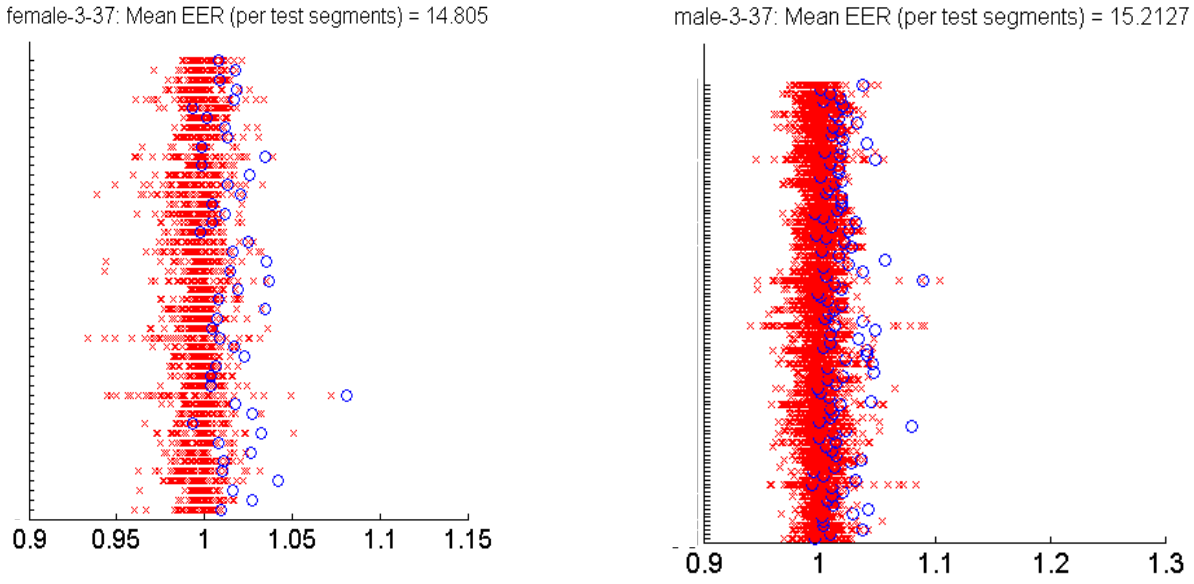



***Figura* 5.7.4.1-6*:*** Faunagramas para las distintas agrupaciones de datos

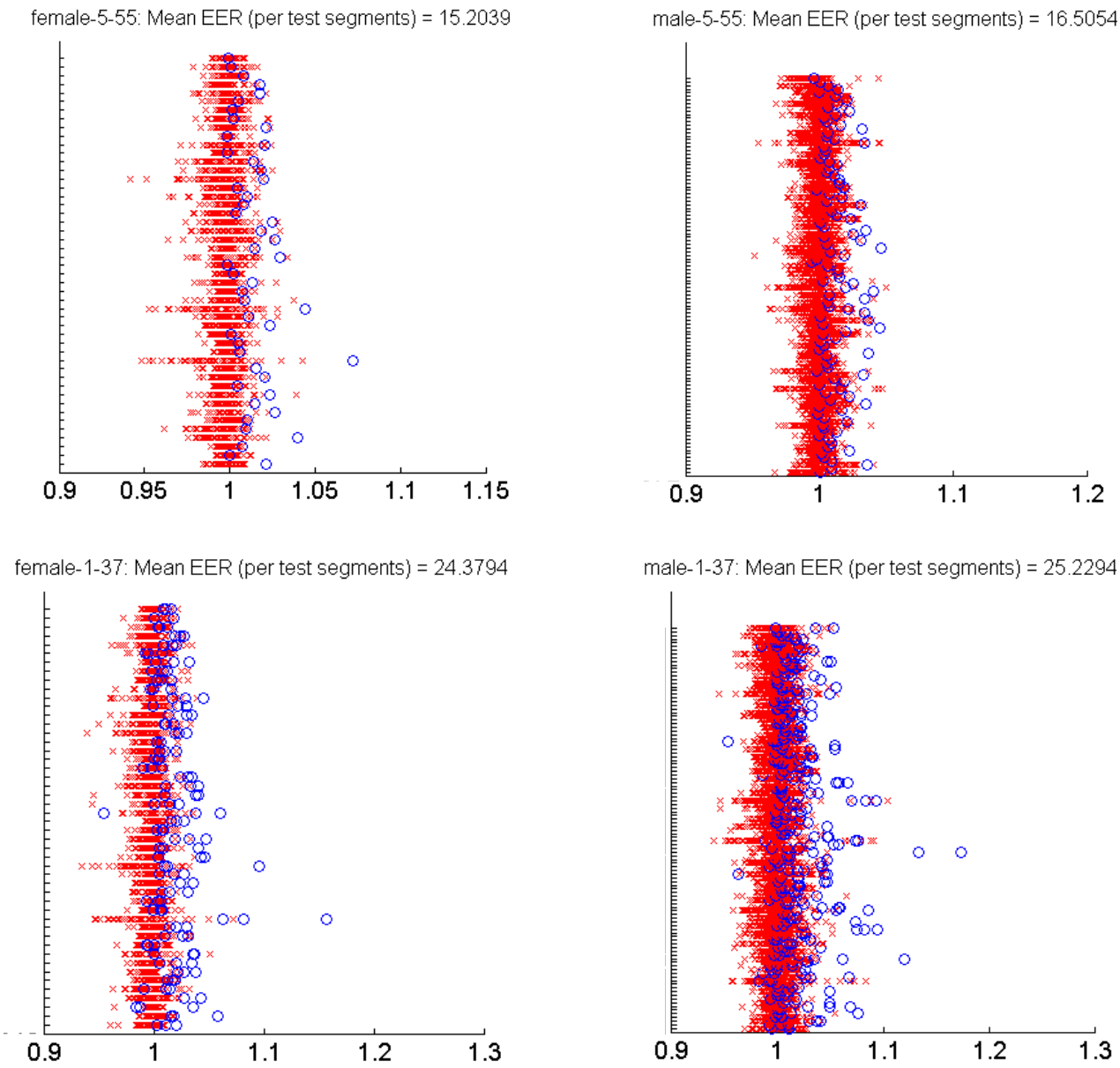



***Figura* 5.7.4.1-6*:*** Faunagramas para las distintas agrupaciones de datos

Como el EER medio por locutor es bastante menor que el EER global que obtenía anteriormente he pasado a normalizar los scores, que se explica en el siguiente apartado.

## 5.7.5 Normalización de scores

La normalización de scores es una técnica ampliamente aceptada para reducir el desalineamiento de los rangos de scores de salida de un sistema biométrico para diferentes comparaciones (entradas) . Aunque existen muchas técnicas de normalización de scores, la más extendida es la llamada “impostor-centric”, en la que los parámetros para la normalización se estiman a partir de los scores obtenidos de las comparaciones de impostor o non-target, es decir, scores de comparaciones en la que ambas muestras no pertenecen a la misma fuente. El motivo de la popularidad de esta técnica es que para una determinada base de datos, existirán muchas más comparaciones non-target que comparaciones target, por lo que la normalización será mucho más robusta.

De entre todas las técnicas de normalización “impostor-centric”, en este proyecto vamos a describir el método T-Norm, o '*Test Normalization*'. Este método trata de alinear los scores de distintas comparaciones ajustándolos a una gaussiana de media cero y desviación típica uno. Para ello, se calcula la media y la desviación típica de un conjunto de entrenamiento de scores non-target calculados con la muestra biométrica de test y una cohorte de impostores, denotadas por μTnorm y σTnorm . A partir de estos datos se puede alinear cualquier conjunto de scores generados con dicha muestra de test aplicando la siguiente expresión:

$$S_{Tnorm} = \frac{S - \mu_{Tnorm}}{\sigma_{Tnorm}}$$

Donde STnorm será el score normalizado y S es el score sin normalizar.
Cuando no se dispone de datos suficientes para tener un conjunto de entrenamiento y un conjunto de test diferenciados, podemos usar el conjunto de scores que queremos normalizar para calcular la distribución. Para ello, cada vez que se normaliza un score, debe calcularse μTnorm y σTnorm a partir de todos los demás conjuntos de scores non-target obtenidos de las comparaciones de las otras muestras. Esta técnica es útil cuando no se dispone de un conjunto de scores de entrenamiento, sin embargo la gran correlación entre los conjuntos de test y de entrenamiento hacen que la normalización sea más ideal que real.

Paso a comparar la mejoría sobre el EER global que se obtiene normalizando.

| EER | male-3-37 | female-3-37 | male-5-55 | female-5-55 | male-1-37 | female-1-37 |
|---|---|---|---|---|---|---|
| sin normalizar | 28,97 | 28,41 | 30,39 | 31,34 | 41,91 | 37,59 |
| normalizado | 20,09 | 18,84 | 21,00 | 21,23 | 25,05 | 20,92 |

***Tabla 5.7.5.1:*** ERR para las distintas agrupaciones de datos con normalizar y sin normalizar, con el factor Reynolds a 16 y 20 centros

A continuación mostramos las DETs para los casos 3-37 normalizados y sin normalizar para ver como varían los falsos positivos ( *False alarm probability* ) en función de los falsos negativos ( *Miss probability* )

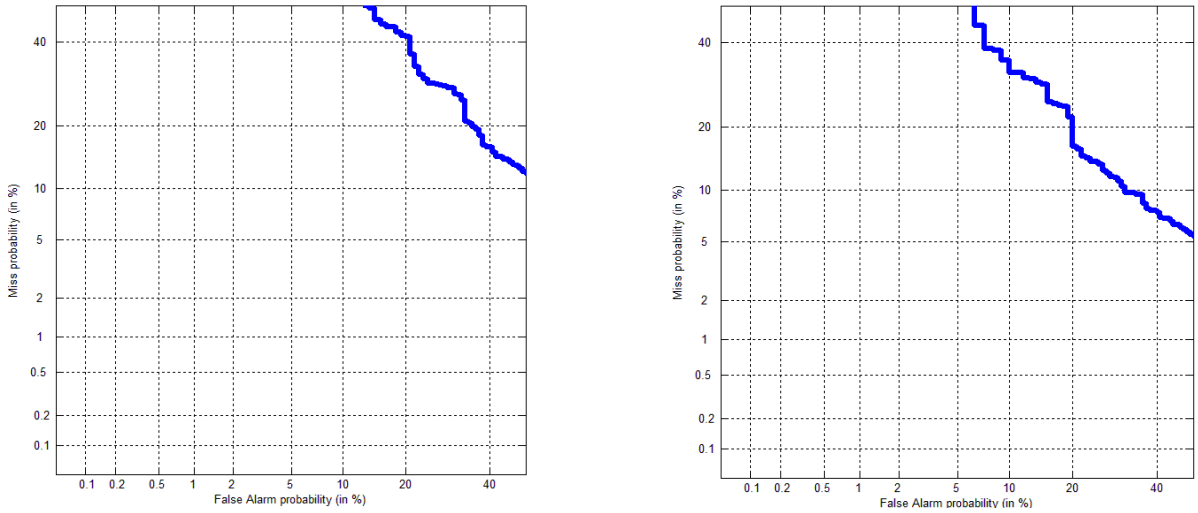



***Figura* 5.7.5.1*:*** DETs para el caso 3-37 para hombres. La figura de la izquierda muestra el resultado sin la normalización de scores y el de la derecha con normalización

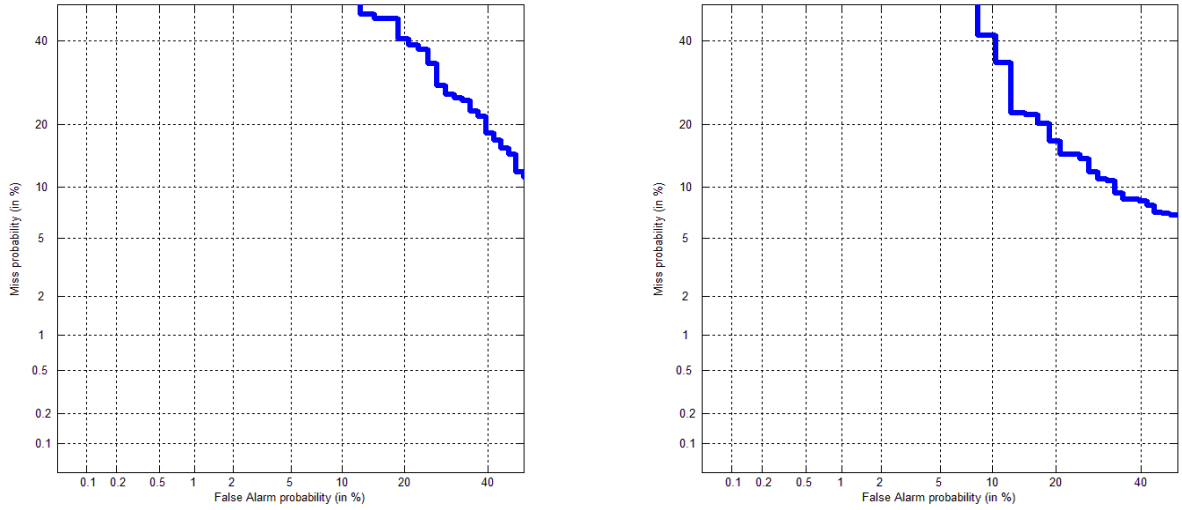



***Figura* 5.7.5.1*:*** DETs para el caso 3-37 para mujeres. La figura de la izquierda muestra el resultado sin la normalización de scores y el de la derecha con normalización

Como el mejor caso lo daba el caso de 30%-70% con bloques de 3 locuciones, para este caso he variado el número de centros y el factor de Reynolds para ver si conseguíamos un EER aún menor y de ahora en adelante solo haré pruebas con ese caso.

*Female:*

| | Reynolds = 2 | Reynolds = 4 | Reynolds = 8 | Reynolds = 12 |
|---|---|---|---|---|
| ncentres = 4 | 20,26 | 20,30 | 20,39 | 20,17 |
| ncentres = 8 | 18,84 | 18,84 | 17,07 | 18,26 |
| ncentres = 12 | 20,48 | 18,84 | 18,84 | 18,31 |
| ncentres = 16 | 23,01 | 23,01 | 20,92 | 19,55 |
| ncentres = 18 | 23,01 | 20,92 | 20,92 | 20,92 |
| ncentres = 20 | 20,92 | 19,55 | 18,84 | 18,17 |

| | Reynolds = 16 | Reynolds = 18 | Reynolds = 20 | Reynolds = 24 | Reynolds = 32 |
|---|---|---|---|---|---|
| ncentres = 4 | 20,17 | 20,39 | 20,48 | 20,74 | 20,79 |
| ncentres = 8 | 18,66 | 18,84 | 18,84 | 18,84 | 19,55 |
| ncentres = 12 | 17,95 | 18,04 | 18,13 | 18,22 | 18,09 |
| ncentres = 16 | 18,84 | 18,84 | 18,84 | 18,26 | 18,26 |
| ncentres = 18 | 20,92 | 20,74 | 20,35 | 18,84 | 18,84 |
| ncentres = 20 | 17,46 | 16,76 | 16,76 | 16,76 | 17,07 |

***Tabla 5.7.5.2:*** Obtención ERR cuando variamos el número de centros y el factor de Reynolds del UBM para female

*Male:*

| | Reynolds = 2 | Reynolds = 4 | Reynolds = 8 | Reynolds = 12 | Reynolds = 16 |
|---|---|---|---|---|---|
| ncentres = 4 | 19,18 | 20,09 | 20,09 | 20,09 | 20,09 |
| ncentres = 8 | 19,48 | 19,28 | 19,18 | 19,18 | 19,30 |
| ncentres = 12 | 19,18 | 19,48 | 19,18 | 19,48 | 19,48 |
| ncentres = 16 | 16,41 | 16,09 | 16,46 | 16,46 | 15,54 |
| ncentres = 18 | 16,76 | 17,02 | 16,76 | 17,16 | 16,93 |
| ncentres = 20 | 19,18 | 19,28 | 20,09 | 19,48 | 19,02 |
| ncentres = 22 | 18,27 | 17,27 | 17,16 | 16,46 | 16,46 |
| ncentres = 24 | 16,46 | 16,46 | 17,36 | 16,98 | 17,36 |
| ncentres = 28 | 21,00 | 21,70 | 20,09 | 19,62 | 19,18 |
| ncentres = 32 | 21,91 | 21,00 | 20,09 | 20,09 | 20,09 |
| ncentres = 36 | 21,10 | 19,09 | 19,18 | 19,18 | 19,88 |
| ncentres = 40 | 18,27 | 17,36 | 18,27 | 18,97 | 19,06 |
| ncentres = 42 | 20,09 | 19,18 | 18,27 | 18,07 | 17,36 |

| | Reynolds = 18 | Reynolds = 20 | Reynolds = 22 | Reynolds = 24 | Reynolds = 32 |
|---|---|---|---|---|---|
| ncentres = 4 | 20,09 | 20,19 | 20,79 | 20,81 | 20,73 |
| ncentres = 8 | 19,09 | 19,02 | 18,84 | 18,82 | 18,93 |
| ncentres = 12 | 19,48 | 19,67 | 19,61 | 19,48 | 19,18 |
| ncentres = 16 | 15,55 | 15,55 | 15,55 | 16,34 | 16,21 |
| ncentres = 18 | 16,91 | 17,15 | 17,16 | 17,15 | 17,21 |
| ncentres = 20 | 19,02 | 18,90 | 18,83 | 18,99 | 18,97 |
| ncentres = 22 | 16,06 | 15,85 | 15,85 | 15,85 | 15,85 |
| ncentres = 24 | 16,46 | 16,46 | 17,36 | 17,36 | 17,36 |
| ncentres = 28 | 19,18 | 19,18 | 18,97 | 18,97 | 18,97 |
| ncentres = 32 | 20,09 | 19,98 | 19,18 | 19,18 | 19,48 |
| ncentres = 36 | 19,65 | 19,77 | 19,99 | 19,84 | 19,18 |
| ncentres = 40 | 19,18 | 19,18 | 19,48 | 19,67 | 20,09 |
| ncentres = 42 | 17,36 | 17,36 | 17,36 | 17,36 | 17,36 |

***Tabla 5.7.5.3:*** Obtención ERR cuando variamos el número de centros y el factor de Reynolds del UBM para male

## 5.7.6 Estudio de fonemas y pitch

En esta tabla se muestra el porcentaje de fonemas que superan la duración marcada por la primera columna. Por ejemplo el 59% de todos los fonemas de vocales superan una duración de 100ms.

| | total | iy | eh | ey | ae | aa |
|---|---|---|---|---|---|---|
| 100ms | 59% | 27% | 31% | 76% | 90% | 70% |
| 90ms | 69% | 39% | 43% | 85% | 93% | 81% |
| 80ms | 78% | 53% | 58% | 92% | 96% | 89% |
| 70ms | 86% | 70% | 73% | 96% | 98% | 94% |
| 60ms | 93% | 84% | 85% | 98% | 99% | 97% |
| 50ms | 97% | 93% | 94% | 99% | 99% | 99% |
| 40ms | 99% | 98% | 98% | 99% | 99% | 99% |
| 30ms | 99% | 99% | 99% | 100% | 100% | 99% |
| 20ms | 100% | 100% | 100% | 100% | 100% | 100% |

| | aw | ay | ao | oy | Ow | er | uw |
|---|---|---|---|---|---|---|---|
| 100ms | 92% | 75% | 68% | 96% | 74% | 65% | 47% |
| 90ms | 95% | 85% | 78% | 98% | 84% | 75% | 55% |
| 80ms | 97% | 92% | 86% | 99% | 91% | 83% | 68% |
| 70ms | 99% | 97% | 92% | 99% | 96% | 92% | 78% |
| 60ms | 100% | 99% | 96% | 100% | 98% | 97% | 86% |
| 50ms | 100% | 99% | 98% | 100% | 99% | 99% | 92% |
| 40ms | 100% | 100% | 99% | 100% | 100% | 99% | 97% |
| 30ms | 100% | 100% | 99% | 100% | 100% | 100% | 99% |
| 20ms | 100% | 100% | 100% | 100% | 100% | 100% | 100% |

***Tabla 5.7.6.1:*** Porcentaje de fonemas de duración mayor o igual a la ventana que se especifica a la izquierda

En esta tabla he representado el percentil del error absoluto del pitch, considerando el error absoluto: abs((f0 obtenido con otra ventana - f0 obtenido con ventana de 100ms))

Por ejemplo, para ventanas de 90ms en el 92% de los casos error es 1Hz o menor.

| | 90ms | 80ms | 70ms | 60ms | 50ms |
|---|---|---|---|---|---|
| 1Hz | 92% | 89% | 81% | 73% | 62% |
| 2Hz | 94% | 92% | 88% | 83% | 75% |
| 5Hz | 95% | 93% | 91% | 87% | 82% |
| 20Hz | 97% | 96% | 93% | 90% | 85% |
| 50Hz | 98% | 97% | 95% | 93% | 90% |
| 100Hz | 99% | 98% | 97% | 96% | 94% |

***Tabla 5.7.6.2:*** Percentiles del error (diferencia entre pitch extraído con ventanas de 100 ms y los ms especificados en cada columna) en valores absolutos.

El pitch se utiliza para que en el algoritmo que calcula los GCI pueda estimar la duración de sus ventanas. Según los investigadores la duración de esta ventana es bastante crítica para que el algoritmo funcione, sin embargo no se qué error en el pitch ellos consideran grave.

De todas formas quiero remarcar que si por ejemplo consideramos un error un fallo en 5Hz y pasamos fonemas de hasta 70 ms, el error no sería 8% (100%-91%) ya que la ventana del pitch detector sería adaptable a la duración del fonema que venga. Es decir, si viene un fonema de más de 100ms utilizamos una ventana de 100ms que es la óptima, si viene un fonema de 72 ms utilizamos una ventana de 72ms y si viene un fonema de menos de 70 ms no lo utilizamos en el sistema.

La conclusión de este estudio sobre la elección de distintas longitudes de ventana es que depende del compromiso entre error y cantidad de datos que entren en el sistema. Si queremos el error mínimo en la detección del pitch tendremos que usar ventanas de 100 ms aunque pasen por el sistema menos fonemas.

## 5.7.8 Variación del vector de características

Nuestra tarea ahora consiste en reducir lo máximo posible el EER. Para asegurarnos de que el vector está correctamente escogido hemos quitado o añadido medias y varianzas de distintos parámetros.

Los mínimos siempre se daban para un factor de Reynolds de 16 o 18, así que para las pruebas solo lo variaba entre esos dos valores, en cambio el número de centros sí lo variaba entre 8 ,12, 14, 16 ,18 20 y 22 centros, ya que el EER mínimo cada vez se daba para un número de centros distintos. Las pruebas están hechas solo para male. Otra cuestión que sorprende de esta prueba es que al añadir casi cualquier otra característica incluido el pitch el EER empeora.

También se distingue dos pruebas una en la que todos los fonemas tienen una duración de 100ms o más y otra para fonemas de 70ms o más dando en este último unos EER mayores.

### *Estructura del vector propuesto en la primera prueba(vector original)*

[EE(med,var), NAQ(med,var), QOQ(med,var), H1H2(med,var), HRF(md,var) PSP(med,var), *'creak'*(med), MDQ(med,var), ps(med,var)]

| Vector | 100 ms | 70 ms |
|---|---|---|
| **Vector original + F0, Ra, Rk, Rg, OQ, UP** | 17.12 | 18.32 |
| **Vector original** | 15.04 | 16.45 |
| **Vector original + F0** | 19.03 | 20.73 |
| **Vector original - NAQ,QOQ,H1H2,HRF,PSP,*'creak'*,MDQ,ps + Ra, Rk, Rg, UP, OQ** | 29.43 | 33.12 |
| **Vector original + F0 - PSP, *'creak'*, MDQ, ps** | 20.65 | 21.43 |
| **Vector original + Ra, Rk, Rg, OQ, UP** | 17.50 | 18.39 |
| **Vector original + F0, UP, OQ** | 20.41 | 21.72 |
| **Vector original + F0, UP** | 18.49 | 19.31 |
| **Vector original + UP** | 17.74 | 18.42 |
| **Vector original + OQ** | 15.84 | 16.79 |
| **Vector original + UP, OQ** | 15.74 | 16.84 |
| **Vector original + OQ, Rk** | 16.50 | 17.53 |
| **Vector original - EE** | 15.54 | 16.04 |
| **Vector original - NAQ** | 17.33 | 18.61 |
| **Vector original - QOQ** | 15.11 | 16.92 |
| **Vector original - H1H2** | 17.44 | 18.82 |
| **Vector original - HRF** | 20.07 | 21.63 |
| **Vector original - PSP** | 18.17 | 19.31 |

***Tabla 5.7.8:*** EER que produce la variación del vector de parámetros introducidos al sistema UBM-GMM-MAP

El resumen de esta prueba sería que siempre se obtienen valores entre 15 y 22%, siendo el mejor vector el que se propuso en la primera prueba con un 15.04%.

## 5.7.9 Vectores instantáneos y PCA

En el siguiente experimento en vez de probar con medias y varianzas de los distintos parámetros glotales, se forman vectores con los valores instantáneos. Es decir, si por ejemplo en un fonema encontramos diez pulsos glotales, extraeremos diez vectores cada uno con los valores instantáneos de cada parámetro glotal. En este experimento no se realiza un promedio, por lo que los fonemas más largos producirán más vectores y por lo tanto tendrán más influencia al determinar el EER.

Por otro lado también se prueba a pasar los datos por un PCA que tiene la función de ortogonalizar los datos con la posibilidad de reducir su dimensionalidad. La ortogonalización de los datos beneficia el correcto posicionamiento de los núcleos de gaussianas que hace el posterior GMM.

En las siguientes tablas mostramos los EER que se obtienen solo con los vectores instantáneos y más tarde los que se obtienen con los vectores instantáneos tras pasar los datos por un PCA.

### EER con vectores instantáneos

*Male*

| | Reynolds = 12 | Reynolds = 16 | Reynolds = 18 | Reynolds = 24 | Reynolds = 30 |
|---|---|---|---|---|---|
| ncentres = 8 | 16,24 | 16,10 | 16,01 | 15,85 | 15,55 |
| ncentres = 12 | 17,36 | 17,14 | 17,06 | 17,36 | 17,36 |
| ncentres = 16 | 18,27 | 18,76 | 18,57 | 18,27 | 18,27 |
| ncentres = 20 | 17,98 | 17,36 | 17,29 | 16,96 | 17,26 |
| ncentres = 24 | 15,97 | 16,25 | 15,85 | 16,46 | 16,56 |
| ncentres = 28 | 15,34 | 15,28 | 15,40 | 15,55 | 15,55 |
| ncentres = 32 | 16,76 | 17,16 | 17,04 | 16,76 | 16,46 |
| ncentres = 36 | 15,37 | 14,64 | 14,43 | 14,18 | 14,55 |
| ncentres = 44 | 17,23 | 17,20 | 16,96 | 17,28 | 16,93 |
| ncentres = 54 | 16,46 | 14,64 | 14,66 | 14,64 | 14,64 |

***Tabla 5.7.9.1:*** Distintos ERR para vectores instantáneos cuando variamos el número de centros y el factor de Reynold del UBM (hombre)

*Female*

| | Reynolds = 12 | Reynolds = 16 | Reynolds = 18 | Reynolds = 24 | Reynolds = 30 |
|---|---|---|---|---|---|
| ncentres = 8 | 14,98 | 14,67 | 14,67 | 14,67 | 14,67 |
| ncentres = 12 | 14,67 | 14,67 | 14,67 | 14,67 | 14,67 |
| ncentres = 16 | 16,76 | 16,76 | 16,76 | 16,76 | 16,76 |
| ncentres = 20 | 16,76 | 16,76 | 16,76 | 16,76 | 16,76 |
| ncentres = 24 | 16,76 | 16,76 | 16,76 | 16,76 | 16,76 |
| ncentres = 28 | 15,98 | 16,81 | 16,81 | 16,82 | 16,82 |
| ncentres = 32 | 17,98 | 17,82 | 16,60 | 16,90 | 17,82 |

***Tabla 5.7.9.1:*** Distintos ERR para vectores instantáneos cuando variamos el número de centros y el factor de Reynold del UBM (hombre)

EER con vectores instantáneos con PCA

*Male*

| | Reynolds = 12 | Reynolds = 16 | Reynolds = 20 | Reynolds = 24 | Reynolds = 30 |
|---|---|---|---|---|---|
| ncentres = 2 | 17,31 | 17,26 | 17,24 | 17,24 | 17,40 |
| ncentres = 3 | 15,52 | 15,52 | 15,52 | 15,29 | 15,17 |
| ncentres = 4 | 18,81 | 19,16 | 19,20 | 19,20 | 19,20 |
| ncentres = 5 | 17,26 | 17,56 | 17,73 | 17,61 | 17,51 |
| ncentres = 6 | 18,95 | 19,63 | 19,58 | 19,05 | 19,05 |
| ncentres = 8 | 17,81 | 17,81 | 17,81 | 18,32 | 17,73 |
| ncentres = 12 | 19,08 | 19,08 | 19,08 | 19,77 | 19,77 |
| ncentres = 16 | 22,59 | 22,59 | 22,59 | 22,12 | 22,19 |
| ncentres = 20 | 20,54 | 21,18 | 20,93 | 21,59 | 21,83 |
| ncentres = 24 | 23,44 | 23,49 | 23,49 | 23,49 | 23,99 |

***Tabla 5.7.9.2:*** Distintos ERR variando el número de cenros y el factor de Reynold utilizando vectores instantáneos con PCA (hombre)

*Female*

| | Reynolds = 12 | Reynolds = 16 | Reynolds = 18 | Reynolds = 24 | Reynolds = 30 |
|---|---|---|---|---|---|
| ncentres = 2 | 13,74 | 13,74 | 13,74 | 13,74 | 13,74 |
| ncentres = 3 | 17,51 | 17,46 | 17,46 | 17,24 | 17,07 |
| ncentres = 4 | 14,45 | 14,32 | 14,32 | 14,45 | 14,67 |
| ncentres = 5 | 14,72 | 14,72 | 14,72 | 14,72 | 14,72 |
| ncentres = 6 | 14,32 | 14,32 | 14,32 | 14,32 | 14,32 |
| ncentres = 8 | 19,77 | 19,73 | 19,59 | 19,59 | 19,55 |
| ncentres = 12 | 21,85 | 22,47 | 22,52 | 22,78 | 22,74 |
| ncentres = 16 | 20,21 | 20,21 | 20,21 | 20,21 | 20,21 |
| ncentres = 20 | 23,71 | 23,98 | 24,07 | 24,07 | 23,01 |
| ncentres = 24 | 17,33 | 19,46 | 19,46 | 18,93 | 17,07 |

***Tabla 5.7.9.2:*** Distintos ERR variando el número de cenros y el factor de Reynold utilizando vectores instantáneos con PCA (mujer)

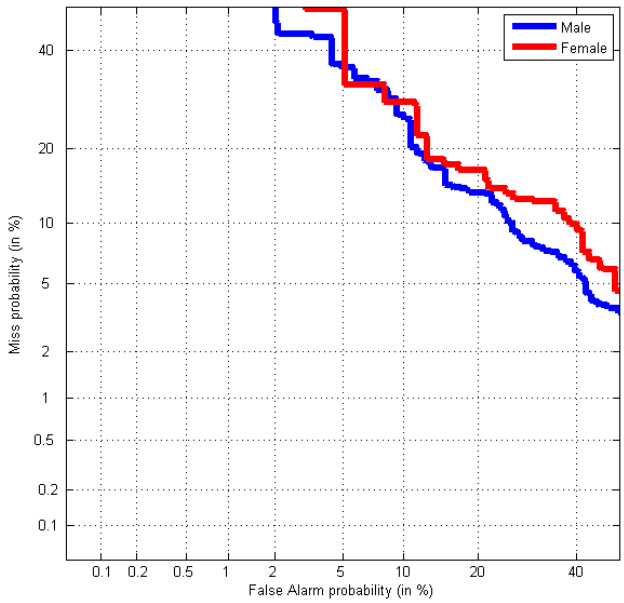



***Figura 5.7.9:*** DET para el mejor caso común entre hombre y mujer (3 centros y nº Reynolds 30)

## 5.8 Análisis de voz carrasposa

El principal problema que tenemos es no disponer para este proyecto de bases de datos clasificadas en sus distintas cualidades vocales, sin embargo contamos con un fichero con unas cuantas muestras de distintas cualidades de voz pronunciadas por John Laver.

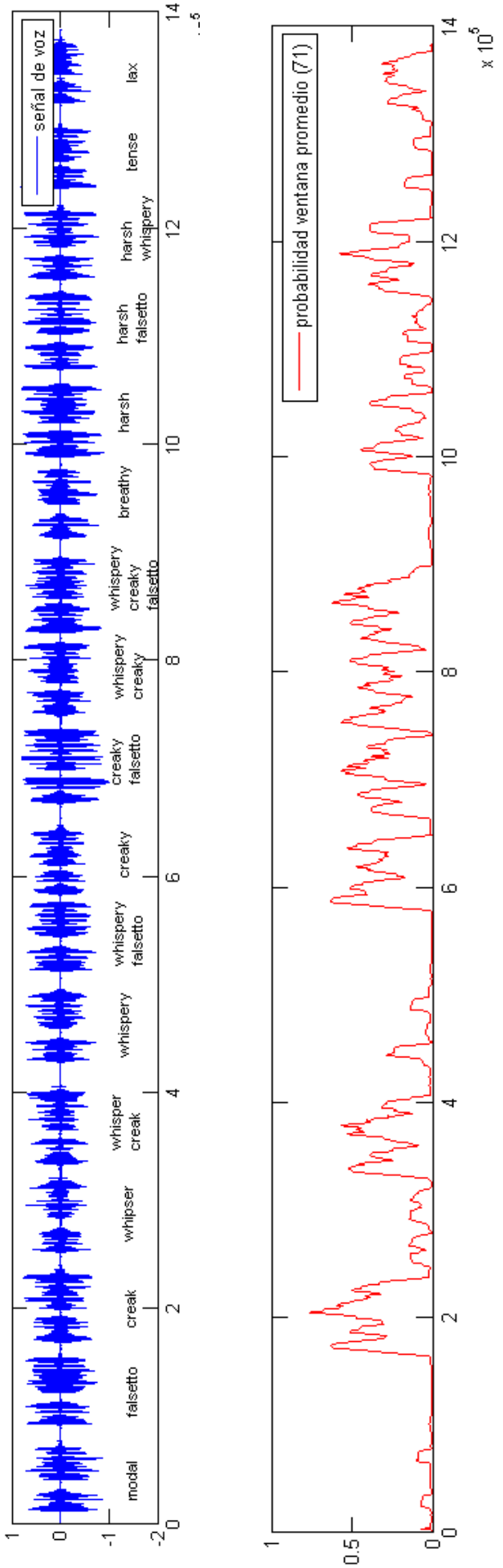



***Figura* 5.8.1.1:** Detección de '*creaky*' en un segmento de audio con distintas cualidades vocales. Arriba la señal de voz y abajo la probabilidad suavizada de `creaky '

En este primer análisis parece que la función detecta correctamente aquellas cualidades vocales en las que hay algún rastro de *'creaky'*. Más tarde probamos con algunas locuciones de habla real obtenida de la base de datos NIST.

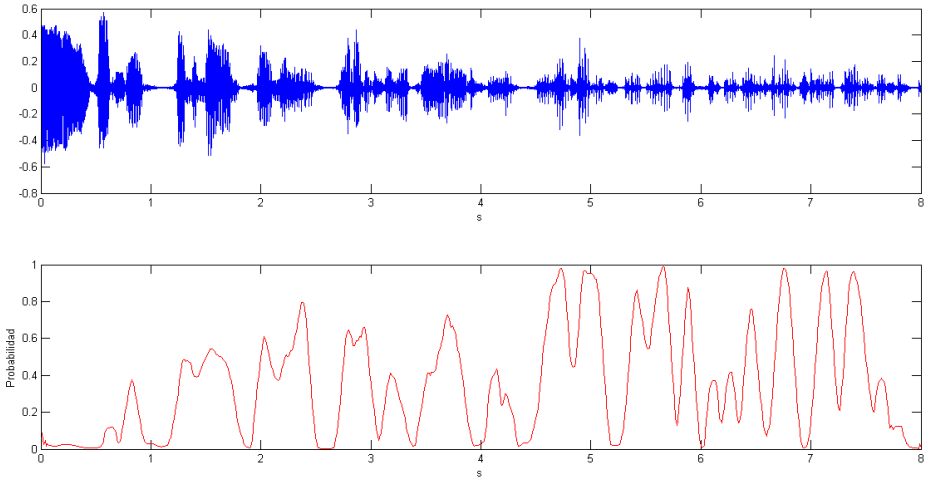



***Figura 5.8.1.2*** Ejemplo de locutor con voz carrasposa. Arriba señal de voz y abajo probabilidad suavizada de "creaky"

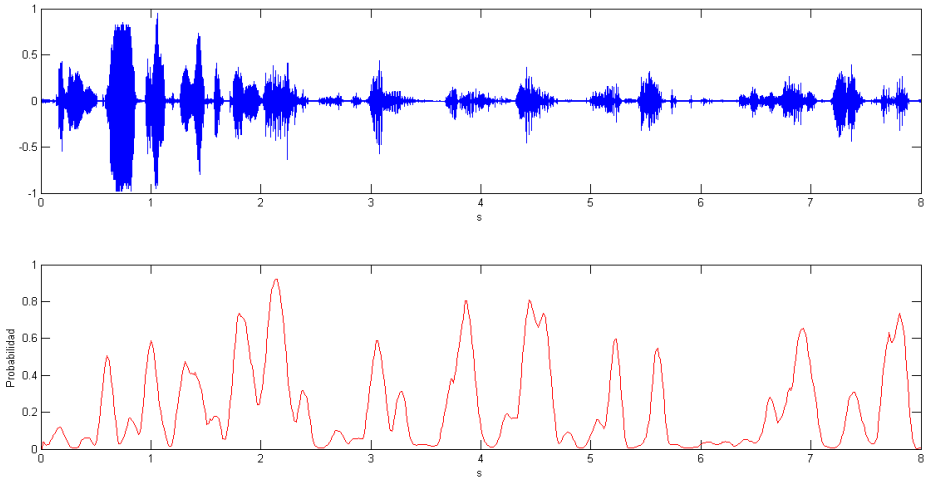



***Figura 5.8.1.3:*** Ejemplo de otro locutor con voz carrasposa. Arriba señal de voz y abajo probabilidad suavizada de "creaky"

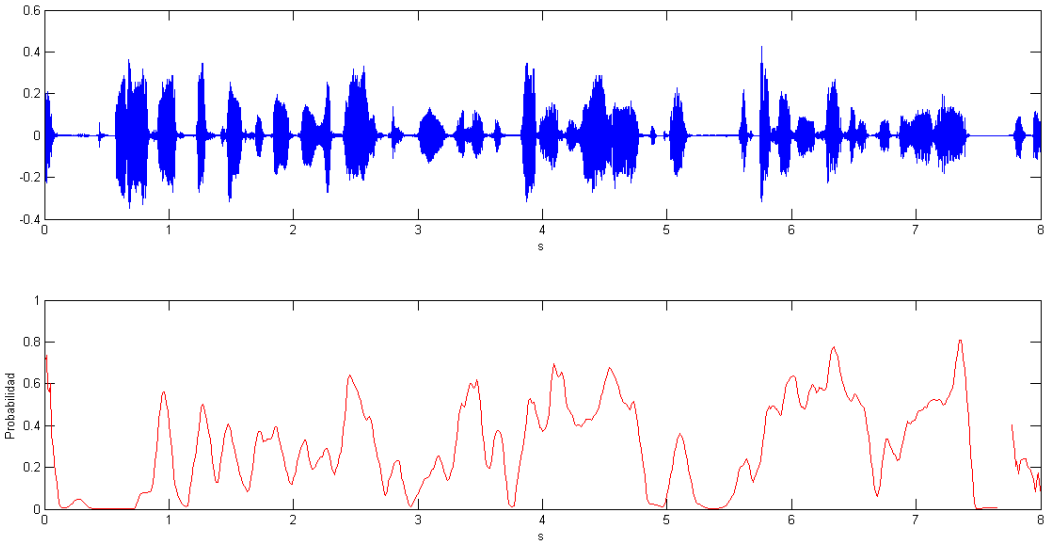



***Figura 5.8.1.4:*** Ejemplo de locutor con voz modal. Arriba señal de voz y abajo probabilidad suavizada de "creaky"

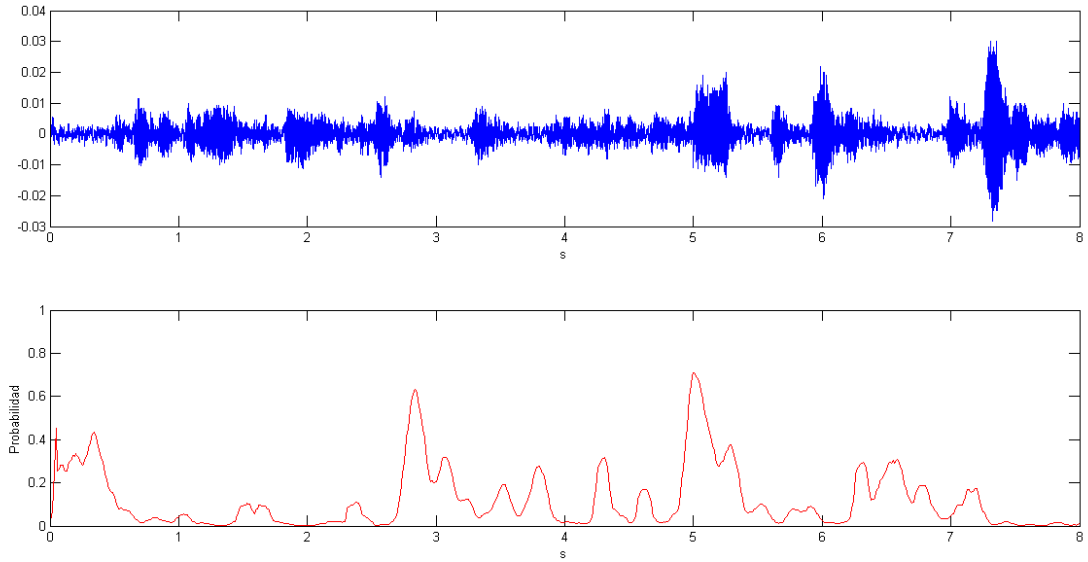



***Figura 5.8.1.5:*** Ejemplo de otro con voz modal. Arriba señal de voz y abajo probabilidad suavizada de "creaky"

# 5.9 Estimación de parámetros sobre voz NIST

En este apartado realizamos algunas estimaciones de derivadas de pulsos glotales sobre la base de datos NIST, para tener una primera referencia de cómo todas estas técnicas pueden funcionar con habla real que es el propósito final de este estudio de características glotales. La base de datos NIST cuenta con dos inconvenientes que no tenía TIMIT y que van a dificultar la correcta extracción de características glotales para caracterizar a cada locutor, que son un nivel mayor de ruido y que las etiquetas no están realizadas manualmente por un lingüística.

A continuación muestro algunas de las derivadas de pulsos glotales obtenidos de un par de locutores de NIST, a primera vista y comparando con las derivadas de pulsos glotales extraídas de TIMIT observamos que tienen un rizado mayor y una forma de onda que dista más de la mostrada en el modelo LF. Sin embargo para determinar los resultados finales que puede tener nuestro sistema en habla real son necesarios más experimentos con esta base de datos.

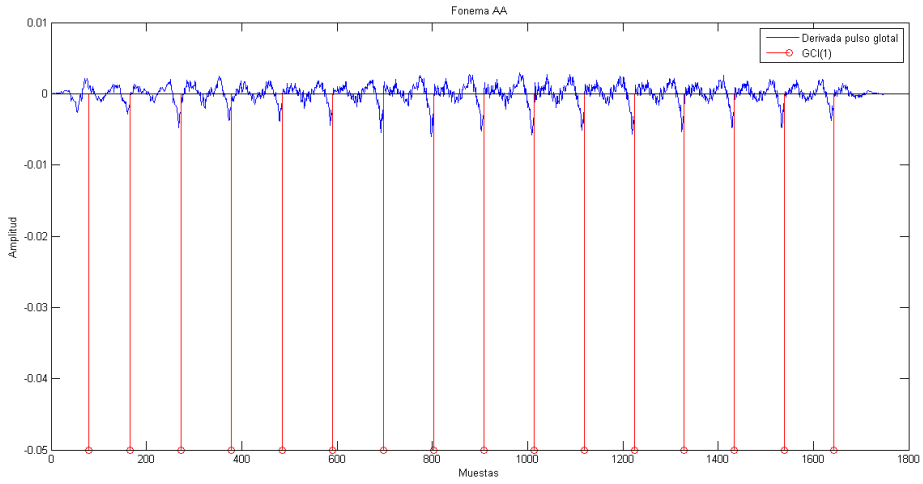



***Figura* 5.9.1*:*** Derivadas de pulsos glotales de NIST (Locutor 1)

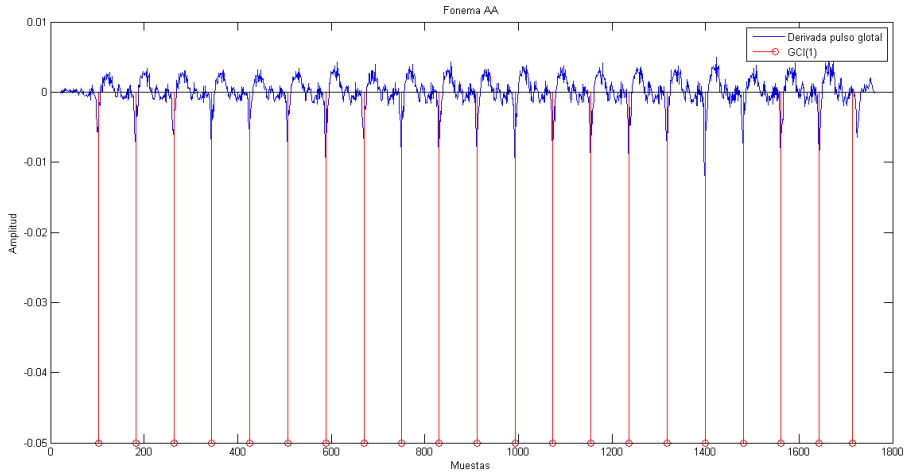



***Figura* 5.9.1*:*** Derivadas de pulsos glotales de NIST (Locutor 2)

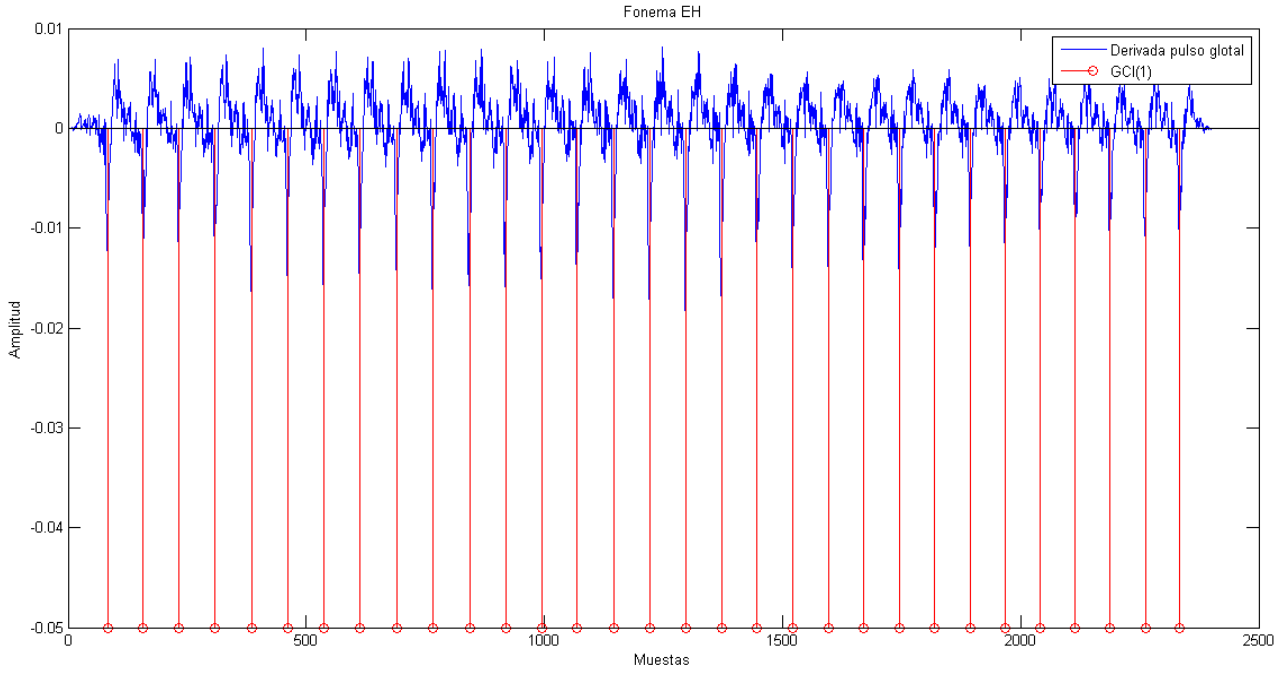



***Figura* 5.9.1*:*** Derivadas de pulsos glotales de NIST (Locutor 3)

# Sección 6

# Conclusiones y trabajo futuro

## 6.1 Conclusiones

En este capítulo se enumeran clasificadas Las conclusiones que pueden extraerse a partir de los resultados, así como se especifica el trabajo a futuro de continuación de este proyecto.

- Es posible obtener información glotal característica de locutor y que además lo clasifique en un tipo de cualidad vocal. El principal problema a la hora de discriminar totalmente entre locutores es la correcta estimación del pulso glotal, por eso el algoritmo elegido es IAIF al ser el más robusto al ruido, criterio que nos conviene ya que el objetivo final de este estudio es poder aplicar todas estas técnicas en habla real.

- Como ya predecíamos los parámetros de los cuales se consigue obtener una mayor diferenciación entre locutores son los que se basan en cocientes de amplitudes ya sean temporales o espectrales. Esto se debe a que cuando extraemos un pulso glotal de un segmento de voz hablada, éste presenta un cierto rizado que hace muy difícil obtener los parámetros que se basan en localizar instantes temporales muy precisos. Pero en aquellos parámetros que se basan en cocientes de amplitudes temporales o frecuenciales, este rizado al no presentar gran amplitud con respecto a la del pulso glotal no va ser tan crítico, de hecho en algunas funciones se elige el máximo de una zona próxima con lo que no influye el rizado que haya en las regiones próximas.

- El conjunto experimental analizado es de un tamaño muy pequeño como para poder sacar conclusiones estadísticamente fiables ya que uno de nuestros objetivos es hacer todas las pruebas presentadas anteriormente para un mismo tipo de fonema, pero esto resulta imposible en TIMIT ya que cada locutor se extrae en promedio 50 fonemas vocales de 13 tipos distintos, por lo que si quisiéramos hacer los experimentos para un tipo de fonema vocal cada locutor contaría tan solo con 4 ó 5 muestras . Sin embargo, los resultados obtenidos tiene un gran valor de cara a dirigir las futuras investigaciones cuando el conjunto de resultados se amplíe.

- La función que disponemos para extraer el pitch da su valor óptimo cuando utiliza ventanas de 100 ms de duración, una duración tanto menor como mayor supone un error que asciende gradualmente.

## 6.2 Trabajo futuro

A partir de este trabajo se abren nuevas líneas de investigación. Las más interesantes se detallan a continuación:

- Repetición de los experimentos con un nuevo y más rico conjunto experimental. Como ya se ha explicado, la fiabilidad estadística de los resultados obtenidos en este proyecto es algo escasa. Esto es debido a no tener suficiente número de fonemas de un mismo tipo por locutor, lo que también supone tan solo contar con un score target por locutor.

- Se propone la repetición de los experimentos en bases de datos de habla real como puede ser NIST. El objetivo final de esta línea de investigación es aplicar y mejorar estas técnicas hasta conseguir tasas de error que puedan ser comparables a las que obtienen los métodos tradicionales para posteriormente fusionarlas y lograr superar las tasas de error que se están consiguiendo en la actualidad.

- Experimentos con base de datos que estén clasificadas en distintas cualidades vocales. Nuestro sistema obtiene distintos parámetros glotales que permiten distinguir a que cualidad vocal pertenece cada locutor pero al no disponer de dicha base de datos no se ha podido realizar las pruebas correspondientes. Exceptuando la detección de segmentos que presentaban alguna rasgo característico de la cualidad vocal de *'creaky'*, que como se ha observado a través de las distintas pruebas da un resultado bastante aceptable aunque queda de alguna manera cuantificarlo.

- Otra tarea pendiente es mejorar o proponer algún algoritmo nuevo para conseguir estimar un pulso glotal más fiable que facilitará una mejor extracción de los parámetros glotales y repercutirá en los resultados finales que produce el sistema. Así como la proposición de algún parámetro glotal nuevo que sea más robusto y fiable que los existentes.

# *Referencias*


[ Alku 92 ] Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Communication,1992.

[Drug 11] Drugman , A.Alwan, Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics, Interspeech, pp. 1973-1976, 2011.

[Drug 09] Drugman, T. and Dutoit, T. Glottal closure and opening instant detection from speech signals. Proceedings of Interspeech, Brighton, UK, pages 2891{2894, 2009.

[Drug 12] Drugman, B.Bozkurt, T.Dutoit, A Comparative Study of Glottal Source Estimation Techniques, Computer Speech & Language, Elsevier, vol.26, issue 1, pp. 20-34, January 2012

[Drug 10] Drugman, T.Dutoit, A Comparative Evaluation of Pitch Modification Techniques, 18th European Signal Processing Conference (EUSIPCO10), Aalborg, Denmark, 2010.

[Drug 13] T.Drugman, Residual Excitation Skewness for Automatic Speech Polarity Detection, IEEE Signal Processing Letters, vol. 20, issue 4, pp. 387-390, 2013

[Drug Boz 09] Drugman, B.Bozkurt, T.Dutoit, Complex Cepstrum-based Decomposition of Speech for Glottal Source Estimation, Interspeech09, Brighton, U.K, 2009.

[Fant 85] Fant, G., Liljencrants, J., Lin, Q., 1985. A four-parameter model of glottal ow. STLQPSR 26, 1.

[Deg 14] G.Degottex, J.Kane, T.Drugman, T.Raitio, S.Scherer, COVAREP - A Collaborative Voice Analysis Repository for Speech Technologies, IEEE International Conference on Audio Speech and Signal Processing (ICASSP), Florence, Italy, 2014.

[Haou 13] Haoxuan Li: Glottal Source Parametrisation by Multi-estimate Fusion, Dublin City University,February 2013.

[Kane Gob 13] Kane, J., Gobl, C., (2013) `Wavelet maxima dispersion for breathy to tense voice discrimination', IEEE Transactions on Audio Speech and Language Processing, 21(6), pp. 1170-1179.

[K&D 13] Kane, J., Drugman, T., Gobl, C., (2013) `Improved automatic detection of *'creak*", 27(4), pp. 1028-1047, Computer Speech and Language.

[Kan PhD 12] Kane, John PhD, Trinity College Dublin 2012.

[Laver 09] Laver J. The phonetic description of voice quality, Cambridge University Press, 12/2/2009

[Plumpe 99] Plumpe, M., Quatieri, T., and Reynolds, D. Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Transactions on Speech and Audio Processing,1999.

[Reynolds 95] Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17 (1995)

[Will 10] William J. Hardcastle (Editor), John Laver (Editor), Fiona E. Gibbon (Editor),The Handbook of Phonetic Sciences, 2nd Edition, February 2010, Wiley-Blackwell

# A

# Presupuesto

| | |
|---|---|
| 1) Ejecución Material | 2.200 € |
| Compra de ordenador personal (Software incluido) | 2.000 € |
| Alquiler de impresora láser durante 6 meses | 50 € |
| Material de oficina | 150 € |
| 2) Gastos generales | 330 € |
| 15% sobre la "Ejecución Material" | 330 € |
| 3) Beneficio industrial | 220 € |
| 10% sobre la "Ejecución Material" | 220 € |
| 4) Honorarios Proyecto | 12.800 € |
| 640 horas a 20 €/hora | 12.800 € |
| 5) Material fungible | 380 € |
| Gastos de impresión | 80 € |
| Encuadernación | 300 € |
| Subtotal Presupuesto (1+2+3+4+5) | 15.930 € |
| IVA 21% s/subtotal | 3.345 € |
| *Total Presupuesto* | *19.275 €* |

Madrid, mayo 2014

El Ingeniero Jefe de Proyecto

Fdo.: Ignacio Rodríguez Ortega
Ingeniero de Telecomunicación

# B

# Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización de este proyecto. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema.

Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

# *Condiciones generales*

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con

arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados.
Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna
clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de
obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas.
Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del
contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al

contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado ”Presupuesto de Ejecución Material” que hoy designa otro concepto.

## *Condiciones particulares*

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.