PROYECTO FIN DE CARRERA

EXTRACCIÓN ROBUSTA DE PARÁMETROS PARA RECONOCIMIENTO AUTOMATICO DE HABLA INSPIRADA EN MODELOS COCLEARES Y DE ENMASCARAMIENTO DEL SISTEMA AUDITIVO HUMANO.

Autor: JESÚS TURIEL MERINO
Dirección: CARMEN PELAEZ MORENO

Universidad Carlos III de Madrid
Escuela Politécnica Superior
Ingeniería de Telecomunicación
Departamento de Teoría de la Señal y Comunicaciones

Septiembre 2011

DEDICATORIA Y AGRADECIMIENTOS.

Quiero agradecer y dedicar este PFC a mis padres por su paciencia y apoyo. A mi hermano por distraerme, y a la vez; alegrarme la redacción cada día que coincidíamos en la misma habitación. A Alma E. por sus ánimos y cariño incondicionales. Y por último, a la directora de este PFC, mi profesora Carmen P.; a la que quiero agradecer su enorme paciencia y empeño.

A todos vosotros, de nuevo; gracias.

RESÚMEN.

Como resumen de este PFC, diremos que evaluamos un Reconocedor Automático de Habla de tipo híbrido introduciendo tres tipos de extracciones de características de audio distintas. Dos de ellas siguen un modelo coclear avanzado (ERB, Seneff), mientras que la otra es una extracción clásica (MFCC) que nos sirve de referencia. A dichas extracciones se le aplicará además un tratamiento adicional al que hemos llamado "Procesado Morfológico", el cual básicamente simula el fenómeno de enmascaramiento en las extracciones citadas anteriormente. A este abanico de Extracciones de Características; más el post-Tratado del Procesado Morfológico, le realizaremos una serie de comparativas para ver cual es la combinación que mejores resultados nos ofrece a la hora de reconocer voz en diferentes situaciones ambientales (con o sin ruido añadido).

Para concluir el resumen, diremos que la B.D de dichos fonemas a reconocer por el bloque anterior viene dada ya por un Software determinado, y que la Red Neuronal que tiene el reconocedor en cuestión será Híbrida. Funcionará por tanto con Perceptrones Multicapa.

TABLA DE CONTENIDO

LISTA DE ILUSTRACIONES.

LISTA DE TABLAS.

1 MOTIVACIÓN.

El reconocimiento de habla está lejos de ser un problema resuelto a día de hoy. Si bien se han conseguidos buenos resultados en un determinado número de entornos, como son ambientes sin ruido de fondo, ambientes con micrófonos cercanos, y sistemas de vocabularios restringidos, etc. No se consiguió todavía esa misma eficiencia en transmisión no ideal y entornos con ruido ambiente severo.

Desde el punto de vista de la extracción de características, los parámetros clásicos, incorporan aspectos de los sistemas auditivo y fonador humanos de forma muy limitada debido a la capacidad de cómputo en la que se implementaban. Hoy en día, nos podemos permitir un coste computacional mucho mayor a la hora de hacer la extracción de características, por lo que podemos mejorar notablemente este proceso adaptando nuestra extracción de características a un comportamiento más real y natural. En consecuencia, un comportamiento más parecido al de nuestro sistema auditivo humano.

En nuestro PFC, lo que se pretende es mejorar los resultados de las extracciones clásicas frente al ruido ambiente utilizando modelos de extracciones más complejos y exactos con respecto al sistema auditivo humano. Para conseguir y evaluar lo anterior tomaremos un modelado basado en redes neuronales (Sistema Híbrido), que nos sirva para diferentes tipos de extracciones de forma equivalente. No utilizaremos; por tanto, el modelado acústico convencional ya que está ya muy optimizado para funcionar con los métodos clásicos y se adapta deficientemente a otros tipos de modelado.

Para centrarnos en el estudio anterior y poder evaluar nuestros modelos de extracción de características sin dificultades externas, no utilizaremos un modelo de lenguaje; simplificando así notablemente la estructura de nuestro reconocedor. Por ello utilizaremos una tarea de palabras aisladas que no necesite modelo de lenguaje. Utilizaremos por tanto la B.D. **[20]**

2 EL PROBLEMA DEL RECONOCIMIENTO DEL HABLA (RAH).

El **Reconocimiento Automático del Habla** (RAH) o **Reconocimiento Automático de Voz** tiene como objetivo permitir la comunicación hablada entre seres humanos y el entorno máquina. El problema que se plantea en un sistema de **RAH** es el de hacer cooperar un conjunto de informaciones que provienen de diversas fuentes de conocimiento (acústica, fonética, fonológica, léxica, sintáctica, semántica y pragmática), en presencia de ambigüedades, incertidumbres y errores inevitables para llegar a obtener una interpretación aceptable del mensaje acústico recibido.

Un sistema de reconocimiento de voz es una herramienta computacional capaz de procesar la señal de voz emitida por el ser humano y reconocer la información contenida en ésta, convirtiéndola en texto o emitiendo órdenes que actúan sobre un proceso. En su desarrollo intervienen diversas disciplinas, tales como la fisiología, la acústica, el procesado de señales, la inteligencia artificial y la ciencia de la computación. **[2]**

Un aspecto crucial en el diseño de un sistema de RAH es la elección del tipo de aprendizaje que se utilice para construir las diversas fuentes de conocimiento. Básicamente, existen dos tipos:

> **Aprendizaje Deductivo:** Las técnicas de Aprendizaje Deductivo se basan en la transferencia de los conocimientos que un experto humano posee a un sistema informático. Un ejemplo paradigmático de las metodologías que utilizan tales técnicas lo constituyen los Sistemas Basados en el Conocimiento y, en particular, los Sistemas Expertos.
>
> **Aprendizaje Inductivo:** Las técnicas de Aprendizaje Inductivo se basan en que el sistema pueda, automáticamente, conseguir los conocimientos necesarios a partir de ejemplos reales sobre la tarea que se desea modelar. En este segundo tipo, los ejemplos los constituyen aquellas partes de los sistemas basados en los modelos ocultos de Markov o en las redes neuronales artificiales que son configuradas automáticamente a partir de muestras de aprendizaje.

En la práctica se asume un compromiso deductivo-inductivo en el que los aspectos generales se suministran deductivamente y la caracterización de la variabilidad inductivamente. **[4]**

En nuestro caso, asumiremos un Reconocedor principalmente Inductivo, que utilizando un conjunto de entrenamiento a modo de ejemplo; es capaz de modelarse automáticamente para llegar a discriminar entre diferentes fonemas en dos ambientes diferentes; uno limpio y otro ruidoso. Para mejorar dicho

RAH le aplicaremos también una parte de Aprendizaje Deductivo a modo de bloques de simulación que, esperemos; sean para nuestro RAH una mejora razonable.

2.1 INTRODUCCION AL RAH.

La representación por bloques de un RAH sería la siguiente:

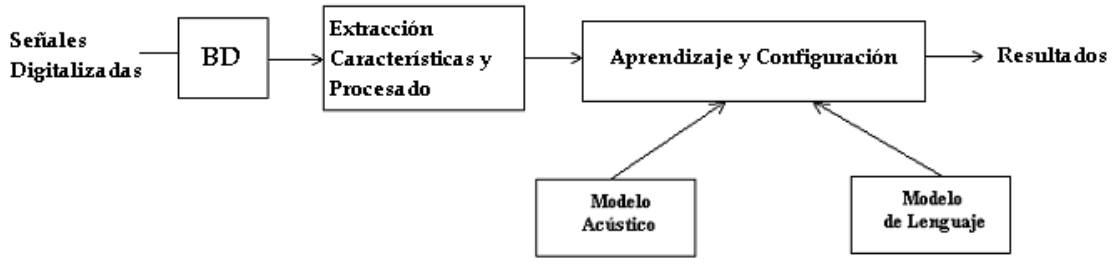


Ilustración 1 Ejemplo Bloques RAH.

Tal y como se puede ver en la figura anterior, en un primer momento nuestro RAH trabaja con un conjunto de señales digitalizadas que forman parte de una **B.D**. En nuestro caso, estas señales digitalizadas serán palabras (en nuestro caso, letras del alfabeto) grabadas por distintos locutores.

De éstas señales de entrada, se extraerán sus características más representativas reduciendo así el tamaño y la variabilidad propias de las ondas acústicas, convirtiéndolas en unos vectores de parámetros. Después estos vectores de parámetros serán procesados con la finalidad de mejorar la extracción de características anterior. Finalmente, nuestro RAH, realiza el **Aprendizaje y Configuración** utilizando para ello los denominados **Modelo Acústico y Modelo de Lenguaje**, y terminado esto, estaría preparado para correr en una simulación dándonos los llamados resultados. **[3]**

Así pues, antes de adentrarnos en cada bloque por separado, debemos citar que para obtener los Modelos Acústico y del Lenguaje es necesaria una etapa previa de *Entrenamiento supervisado*, en la que a partir de unas señales etiquetadas (Conjunto de Entrenamiento) el sistema compone dichos modelos. Una vez obtenidos, se lleva a cabo el *Reconocimiento y Resultados* propiamente dichos, es decir, la etapa donde se obtiene la palabra reconocida por nuestro sistema ante el conjunto de prueba; así como las estadísticas de acierto/fallo del sistema.

2.2 EXTRACCION DE LAS CARACTERISTICAS: Modelos Clásicos y Modelos Cocleares.

En nuestro PFC utilizaremos diferentes modelos de extracción de Características, estudiando por tanto ambos sistemas y comparando los resultados de cada uno de ellos. Utilizaremos como modelo clásico la extracción MFCC y como modelo coclear la extracción ERB y la extracción Seneff. **[1]**

Modelos Clásicos

Aunque existan más, el más utilizado y el que utilizaremos como referencia en nuestro PFC, será el método de extracción **MFCC (Mel-frequency cepstral coefficients)**. El cual corresponde a un método de extracción que presenta unos coeficientes basándose hasta cierto punto en la producción y percepción auditiva humanas con la ayuda de diferentes técnicas matemáticas como son:

La Transformada de Fourier: Utilizada para obtener las componentes frecuenciales de la señal y pasar la señal a un formato frecuencial más fácilmente manejable.

Un Banco de Filtros Mel: El cual es el encargado de modelar la respuesta auditiva humana (comportamiento del oído) espaciando las bandas de frecuencias de manera logarítmica al igual que hace la cóclea de nuestro sistema auditivo humano.

La DCT (Transformada Discreta del Coseno): Es la encargada de compactar la información para poder realizar la separación entre la envolvente espectral (procedente del tracto vocal) y la excitación mediante un **Liftering**. El cual corresponde al último proceso de la cadena de extracción MFCC. **[2]** **[3]**

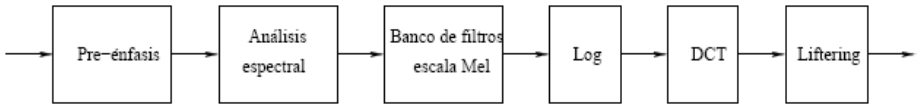


Ilustración 2 Esquema de bloques de la extracción MFCC.

Modelos Cocleares

Esperando una descripción más amplia del funcionamiento de la cóclea en el Capítulo 3 de este PFC, presentamos los modelos cocleares que hemos utilizado para la extracción de características.

ERB: Modelo de extracción que calcula los coeficientes del banco de filtros definidos por Patterson y Holdworth para la simulación de la cóclea. **[11]**

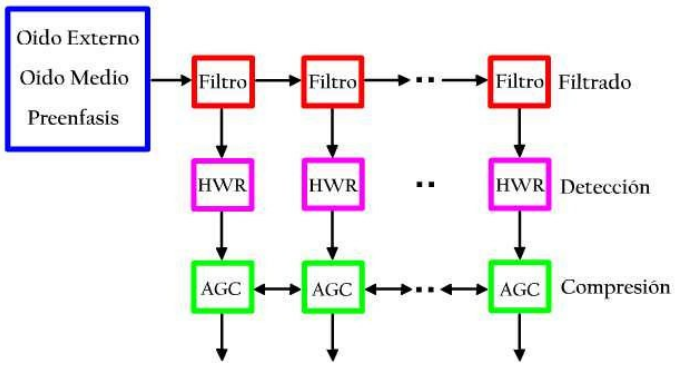


Ilustración 3 Modelado ERB del Sistema Auditivo Humano

El modelo utiliza un banco de filtros cuyas etapas están conectadas en cascada para simular el viaje de las ondas sonoras, una etapa de detección implementada con rectificadores de media onda (HWR) y una etapa de compresión implementada con circuitos de control de ganancia automático (AGC).

Su salida es un vector proporcional a la tasa de disparo de impulsos eléctricos en cada punto dentro de la cóclea, separando en frecuencia la energía de la onda acústica. El modelo por tanto captura la mayoría de los efectos cualitativos de una cóclea real.

A grandes rasgos el modelo está formado por dos etapas de filtros, una etapa de filtros conectados en cascada que emulan la propagación de las ondas a través de la cóclea, y un otra etapa en paralelo que emulan el movimiento de la membrana basilar, tomando el modelo original el nombre de cascada paralelo. Estos filtros son de orden cuatro y comparten los mismos "polos", pero que tienen diferentes "ceros", y ambos independientemente simulan la respuesta de las bandas separatistas de la cóclea a lo largo del espectro de audición. Ambos filtros serán por tanto independientes, y cada uno de ellos tendrá una respuesta para su banda frecuencial correspondiente la cual cogieron como entrada. La respuesta al impulso de cada uno de ellos sigue el siguiente patrón:

$$gt(t) = A^* t^{(N-1)} * \exp(-2b*t) \cos(Wr*t + \emptyset) \quad (t>0)$$

Respuesta al impulso de filtro ERB.

Seneff: El esquema de cálculo propuesto por *S. Seneff* para modelar el sistema auditivo humano trata de captar las características esenciales extraídas por la cóclea como respuesta ante las ondas sonoras de presión. **[10]** El sistema incluye tres bloques, los dos primeras de ellos están relacionados con las transformaciones que ocurren en la periferia de las primeras etapas del proceso de audición, mientras que el tercero intenta extraer la información correspondiente a la percepción como formantes e intenta la mejorar de la agudeza de de los segmentos de la señal. La señal de voz, es pre-filtrada y anulada en sus frecuencias de componente muy alta y/o baja. Luego se pasa a través del primer bloque, un *banco de filtros de 40 canales críticos individuales* diseñados con el fin de ajustar los datos fisiológicos de nuestro oído a la señal de entrada. El segundo bloque no lineal y tiene la intención de simular las características de la transformación de la membrana basilar, representando las propiedades de respuesta de las fibras del nervio auditivo. El tercer y último bloque es un bloque de unidad doble con dos salidas paralelas. La primera unidad es la denominada *Detector de Sincronía* (GSD), que implementa la conocida "fase de bloqueo" de las fibras nerviosas con el objetivo de mejorar los picos espectrales creados por el tracto vocal. La segunda unidad de esta tercera fase es la llamada *Detector de envolvente* (ED) y se encarga de calcular la envolvente de las señales de la etapa anterior del modelo, dándole más importancia a la captura de la dinámica cambiante de la palabra. Los resultados de esta unidad son por tanto más importantes en los sonidos transitorios.

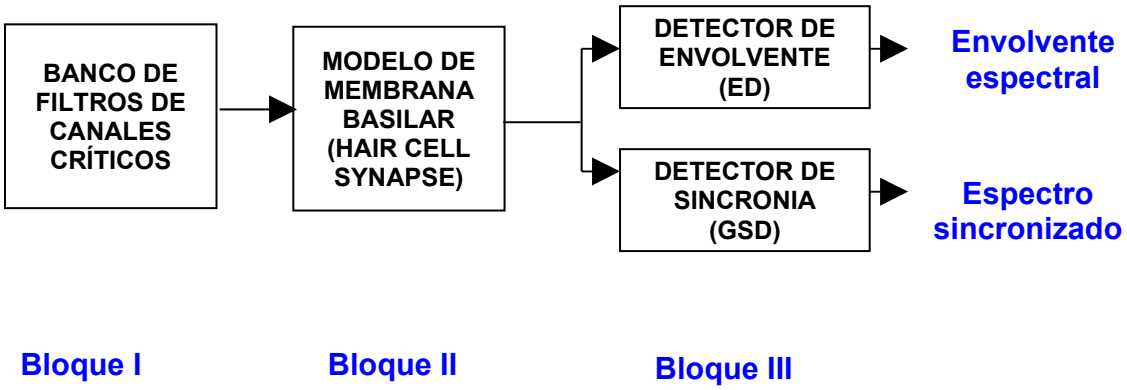


Ilustración 4 Diagrama bloques de la extracción Seneff

2.3 MODELADO ACÚSTICO: Modelos Markov y Redes Neuronales.

En los RAH más utilizados, el Modelado Acústico tradicionalmente se elabora a base de un Modelo de Oculto de Harkov (HMM) junto a una extracción de características del tipo MFCC (Véase 2.2). Viendo que en nuestro PFC utilizaremos diferentes extracciones (MFCC, ERB y Seneff), este modelado tradicional no nos es válido. Aparte de esto, hay que tener en cuenta que en nuestra tarea de RAH específica trabajamos únicamente con palabras aisladas

(Véase 2.1) en vez de, por ejemplo; habla continua. Por lo que nuestro sistema podrá ser adaptado a ello, simplificándolo para conseguir de él un cómputo más rápido.

Para que nuestro modelado acústico sea adaptativo y para que funcione igualmente con todas y cada una de nuestras diferentes extracciones, tendremos que implementarlo con lo que se conoce como Modelo Híbrido. El citado modelo consta de un Modelo Oculto de Markov tradicional junto con una Red Neuronal que le da esa capacidad de adaptación que tanto necesitamos. Pasamos a continuación, a explicar estos conceptos y cómo han sido implementados para adaptarlos a todas nuestras necesidades. **[6]**

Modelo Oculto de Harkov (HMM, 1960)

Los **Modelos Ocultos de Markov** son también una técnica de reconocimiento de plantillas o patrones, sin embargo, la diferencia estriba en que se utiliza un modelado estocástico o aleatorio de dichos patrones, ya que permiten una mayor flexibilidad a la hora de representar secuencias de duración variable. Un HMM es una máquina de estados finitos en la que el estado siguiente depende únicamente del estado actual, y asociado a cada transición entre estados se produce un vector de parámetros. A grandes rasgos, las plantillas se obtienen en función de los nodos recorridos en base a la probabilidad de que las unidades fonéticas sean observadas por los distintos Modelos Ocultos de Markov (HMM). **[7]**

Un Modelo Oculto de Markov (HMM) se puede representar como una máquina de estados finitos, donde la probabilidad de pasar al estado siguiente depende únicamente del estado actual (proceso de Markov), y asociada a cada transición entre estados se produce un vector de observaciones. La particularidad de los HMM, es que el paso de unos estados a otros de la cadena no es directamente observable, está oculto. Por tanto, se puede decir que un HMM está compuesto de 2 procesos estocásticos, el oculto, correspondiente a las transiciones entre estados, y el observable o no oculto, correspondiente a la generación del vector de observaciones que se produce en cada estado, y que representa la plantilla a reconocer. Además, cada estado tiene asociada una distribución de probabilidad sobre los posibles símbolos de salida, con lo que la secuencia de símbolos generada por un HMM proporciona cierta información sobre la sucesión de estados.

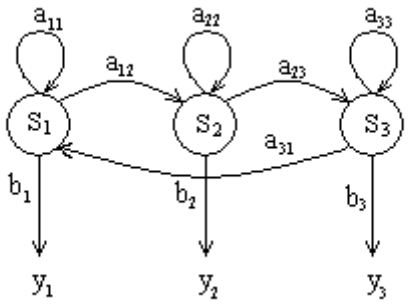


Ilustración 5 Ejemplo de HMM de 3 estados (*S* estados ocultos, *y* salidas observables, *a* probabilidades de transición, *b* distribuciones de probabilidad de salida)

Visto desde el lado del habla, cada nodo representaría una unidad acústica (fonema, difonema, trifonema o cualquiera otra) diferente, y cada sucesión de fonemas con su probabilidad de transición asociada; una palabra determinada. La cual vendría dada en su totalidad con su probabilidad de emisión del HMM.

En nuestro PFC, trabajamos como unidad acústica el fonema y por tanto, cada modelo de Markov sería un fonema diferente. Teniendo éste internamente probabilidades de transición entre sus estados y su correspondiente probabilidad de emisión que como versemos a continuación modelaremos con una red neuronal artificial. Igualmente no nos adentraremos más en el funcionamiento de nuestro conjunto de HMM ya que viene dado por el entorno de experimentación específico que estamos utilizando (Hidden Markov Model Toolkit) **[19]** estando siempre las probabilidades de transición entre estados prefijadas siendo solamente tratadas por nosotros las probabilidades de emisión; las cuales serán elaboradas por una **Red Neuronal (ANN).**

Redes Neuronales (ANN)

Las **Redes Neuronales Artificiales** (Artificial Neurals Networks, ANN) son estructuras de procesado paralelo, formadas por muchas unidades (nodos o neuronas) simples, conectadas entre sí con distintos pesos y agrupadas en diferentes capas.

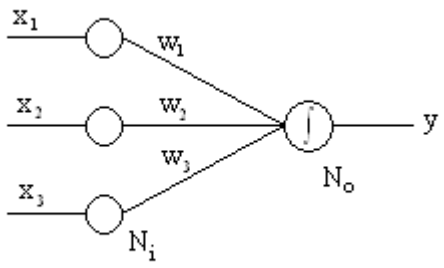


Ilustración 6 Ejemplo de Red Neuronal Artificial (w_i pesos, x_i vectores de entrada, y vector de salida, $\int$ función de activación, N_i capa de entrada, N_o capa de salida)

Con el paso del tiempo se han convertido en una técnica muy usada en reconocimiento de voz debido a su capacidad de clasificación y sobre todo a la capacidad que tienen para aprender una determinada tarea a partir de pares observación-resultado sin hacer ninguna suposición del proceso que modelan, pudiendo así encontrar un modelo adaptado a los datos independientemente de la naturaleza de los mismos. Sin embargo, presentan una serie de problemas, como son el desconocimiento a priori de la estructura de capas, un elevado tiempo de entrenamiento y el posible estancamiento en mínimos locales. Además las ANN, en general, sólo son capaces de procesar vectores de longitud fija, por lo que no suelen utilizarse de manera aislada en RAH, sino en combinación con otras técnicas, como son los HMM, que introducen el procesado temporal necesario para el reconocimiento de voz. Como se dijo antes, esta combinación es llamada **Modelo Híbrido (HMM/ANN).**

Citar por último que típicamente las ANNs se suelen construir en los Sistemas de Reconocimiento Automático de Voz con Perceptrones Multicapa (MLP, Multi-layered Perceptron). Son redes por tanto multicapa y de aprendizaje supervisado, es decir, que necesitan ser entrenadas con conjuntos de datos observación-resultado. Además son de alimentación hacia delante (feedforward), y su estructura se caracteriza por introducir entre la capa de entrada (N_i) y la de salida (N_o) una o más capas intermedias u ocultas (N_h), con capacidad de procesamiento, y sin otras conexiones con el exterior. **[8] [10]**

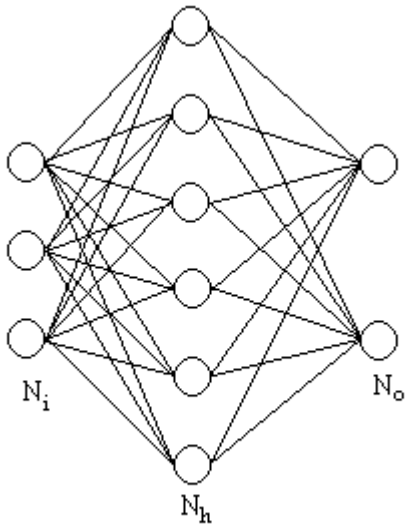


Ilustración 7 Ejemplo de Perceptrón Multicapa con una única capa oculta (N_h).

En nuestro PFC, nuestra Red Neuronal será entrenada con muestras formadas por los parámetros correspondientes a la trama actual extendida con un número de tramas anterior y posterior que denominamos contexto. Estas muestras deben estar previamente etiquetadas lo cual se puede hacer con un reconocedor convencional de tipo HMM[1]. Las secuencias de tramas serán posteriormente alineadas en fonemas los cuales son nuestra unidad acústicas a reconocer. Se entrena la red por tanto con una batería de tramas (representadas por sus parámetros) con sus etiquetas correspondientes (Ilustración 8) con el fin de que la red aprenda a distinguir señales y llegue a identificarlas. Estas etiquetas se expresan en un vector de dígitos binarios de dimensión igual al número de fonemas posibles considerados.

[1] En este proyecto se parte de una base de datos ya etiquetada mediante este procedimiento evitando entrar en este alineamiento inicial.

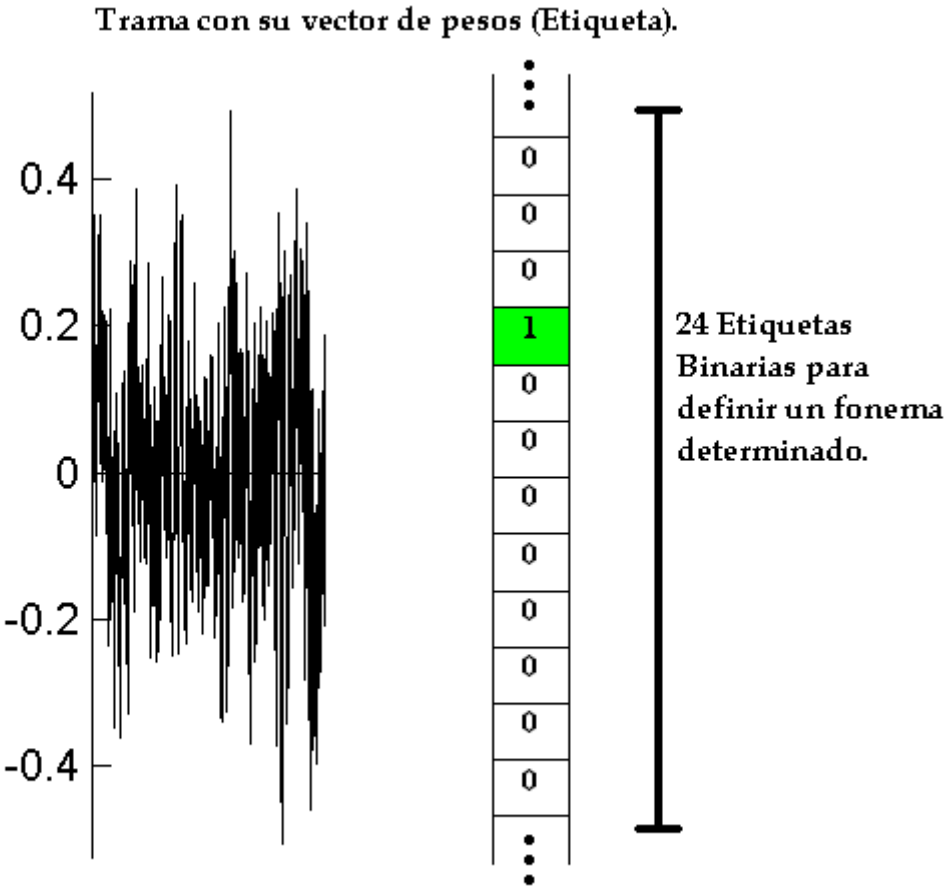


Ilustración 8: Unidad Pronunciación de entrenamiento de la ANN.

Una vez entrenada nuestra red para distinguir entre las diferentes fonemas las salidas de la red dejarán de ser binarias (véase la Ilustración 9) para presentar la probabilidad a posteriori de que cada trama individual pertenezca a cada uno de los fonemas. Estos son los valores que utilizaremos como probabilidades de emisión en el HMM encargado de realizar el alineamiento de estas tramas para formar los fonemas. Así pues, una vez entrenada la red, tendrá que recibir como entrada segmentos de fonemas (tramas) con el fin de identificarlos con su unidad de pronunciación correspondiente, hasta llegar a tener la cadena que formará el fonema. Estos segmentos serán extracciones cada 10ms de los fonemas de entrada; y la red trabajará sobre ellos con el fin de discernir a que unidad acústica pertenece cada uno.

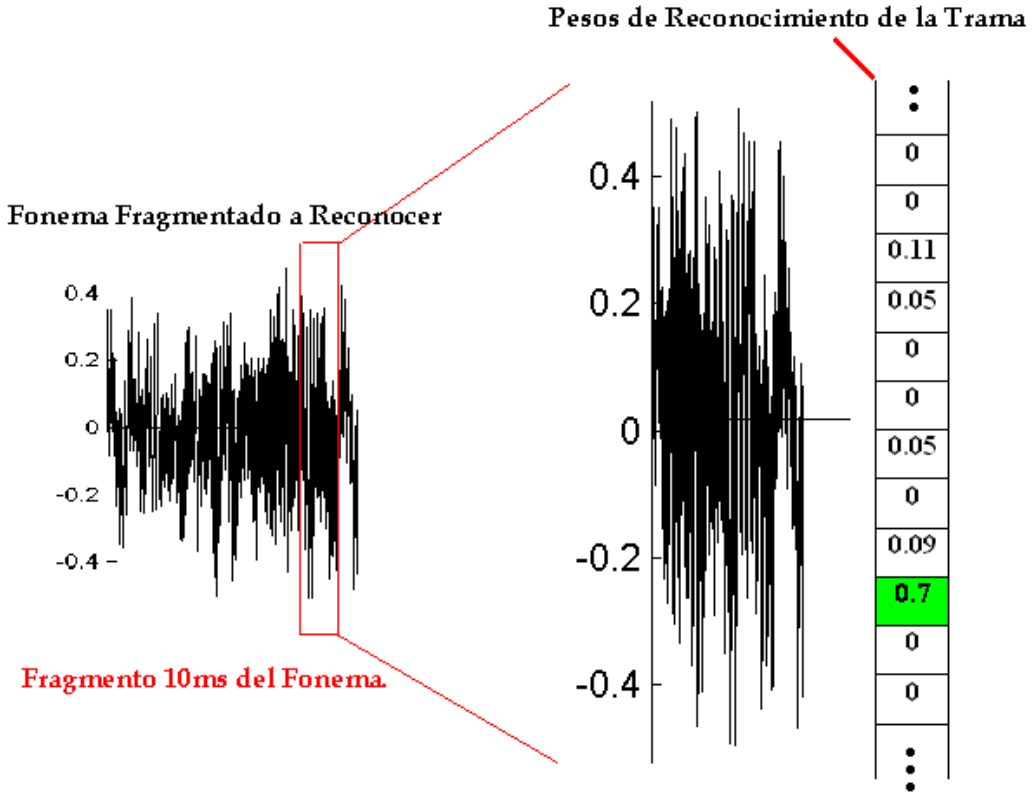


Ilustración 9 Ejemplo del Proceso de Reconocimiento de las Tramas

Dado que el entorno de experimentación utilizado en este PFC sólo considera el reconocimiento de las letras del alfabeto pronunciadas de forma aislada (es decir, palabras aisladas) deberemos aplicar sólo un diccionario de pronunciaciones de las diferentes letras del alfabeto evitando la utilización de un modelo de lenguaje propiamente dicho. No obstante, en la próxima sección describimos someramente el bloque de modelado del lenguaje que sería necesario en caso de abordar una tarea de reconocimiento más compleja.

2.4 MODELADO DEL LENGUAJE.

El **Modelo de Lenguaje** modela la probabilidad de aparición de una secuencia de palabras. Así, podemos decir que, el modelado lingüístico es uno de puntos indispensables para obtener un reconocedor de habla continua. Una de las técnicas más utilizadas en este ámbito son las basadas en N-gramas, que recogen de manera sencilla, a partir de aproximaciones, dada la complejidad del problema, las concatenaciones entre palabras. Así, el problema se reduce a calcular la probabilidad de las palabras en función de sus N predecesoras.

Como ya hemos mencionado, en el caso concreto de este proyecto, cuya tarea de reconocimiento se limita al reconocimiento de palabras aisladas (en particular, las letras del alfabeto) sin concatenación entre ellas, no nos hace falta un modelo de Lenguaje complejo de habla continua. Nosotros tenemos en

cambio un **Diccionario** de pronunciaciones con todas las letras del alfabeto y sus fonemas constituyentes en su orden correspondiente. Cada letra del alfabeto está identificada por la cadena de unidades fonéticas que lo forman. Este diccionario es el que guía la búsqueda de la palabra escogida como decisión final de reconocimiento para cada una de las formas de onda recogidas en cada archivo de la base de datos. Es importante tener en cuenta que nuestra tarea de reconocimiento tiene una estructura muy específica existiendo algunas letras que se confunden entre sí con mucha facilidad por compartir la vocal de la sílaba (el denominado conjunto de la E o E-set, por ejemplo) y otras muy fácilmente discernibles independientemente de la calidad del modelado acústico (por ejemplo, la letra 'W').

3 MÉTODOS DE EXTRACCIÓN DE CARACTERÍSTICAS BASADOS EN MODELOS COCLEARES AVANZADOS

Una vez revisadas las etapas fundamentales que conforman nuestro RAH, vamos a realizar una serie de procesados a las extracciones que se mencionaron anteriormente a fin de mejorarlas imitando mejor el comportamiento natural de nuestro oído humano. La finalidad de este procesado adicional no es más que acercar más nuestro RAH al comportamiento natural del oído mejorando así su funcionalidad y rendimiento final. Vamos a estudiar por tanto, los procesos que queremos simular con la citada etapa.

3.1 EL SISTEMA AUDITIVO HUMANO.

El oído humano puede dividirse en tres partes: *oído externo, medio e interno*.

Oído externo: Está constituido por el *pabellón auditivo* (oreja), el *conducto auditivo* y el *tímpano*. Las ondas sonoras son recogidas por el pabellón que las conduce a través del conducto auditivo hacia la membrana del tímpano.

Oído medio: Es una cavidad limitada por el *tímpano* por un lado, y por la base de la *cóclea* por el otro. En su interior hay tres huesecillos, llamados *martillo*, *yunque* y *estribo*. La cabeza del martillo se apoya sobre el tímpano y transmite vibraciones a través del yunque al estribo. A su vez éste último se apoya en una de las dos membranas que cierran la cóclea, la ventana oval. A grandes rasgos, es un adaptador de impedancias de un medio con poca densidad y membrana de amplitud grande (tímpano), a otro medio con mucha densidad y membrana mucho menor (membrana de la cóclea).

Oído interno: Es una cavidad hermética cuyo interior está anegado por un líquido llamado *linfa*. Consta de tres elementos: los *canales semicirculares*, el *vestíbulo* y la *cóclea (caracol)*. Los canales semicirculares no tienen relación directa con la audición, tienen que ver con el equilibrio. Las vibraciones de la ventan oval del vestíbulo son transformadas en la cóclea. Las señales de la cóclea son codificadas y transformadas en impulsos electroquímicos que se propagan por el nervio acústico hasta llegar al cerebro.

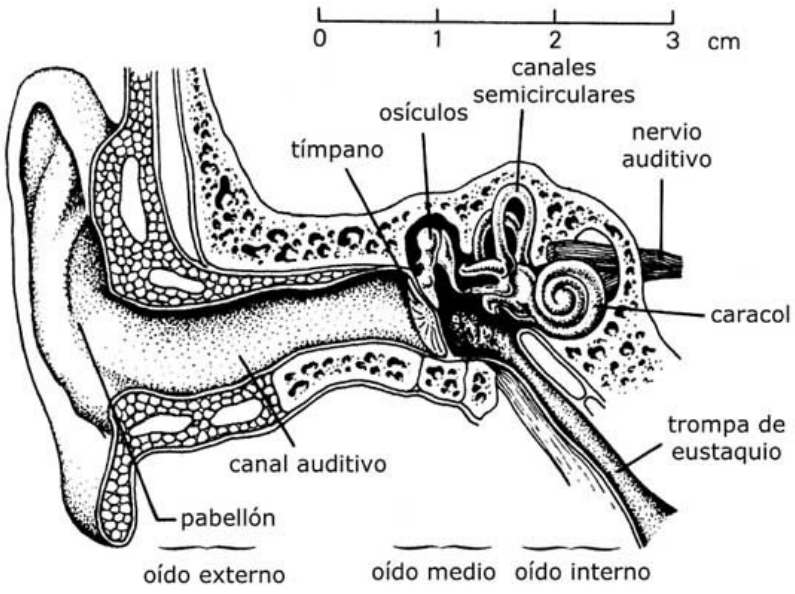


Ilustración 10 Imagen del Sistema Auditivo Humano.

En cuanto a nuestro PFC, lo que más nos interesa es la codificación de esas señales acústicas en impulsos nerviosos y de cómo el cerebro las interpreta. Por lo que nos centraremos en el funcionamiento del Oído interno. **[5]**

3.2 FUNCIONAMIENTO DEL OIDO INTERNO.

El conjunto coclear posee dos orificios (*ventanas oval y redonda*) tapados por sendas membranas. La ventana oval está unida al *estribo* y recibe de él sus vibraciones, la ventana redonda sirve para igualar las ondas de presión que entran gracias al conjunto anterior. La cóclea, a su vez; se divide longitudinalmente por la *membrana basilar*, sobre la que se asientan los filamentos terminales del nervio auditivo. Cuando el estribo empuja la ventana oval, se produce una sobrepresión en la parte superior de la cóclea que obliga a circular el *fluido linfático* hacia la cavidad inferior a través del *helicotrema*, mientras que la membrana basilar se deforma hacia abajo. Finalmente, la membrana elástica que cierra la ventana redonda cede hacia afuera. **[5]**

Cuando el estribo se mueve hacia la izquierda y la derecha, aumentando y disminuyendo la presión del líquido contenido encima de la membrana basilar, aparece una onda que se desplaza de izquierda a derecha a lo largo de la membrana. Esta onda puede visualizarse como un movimiento de traslación hacia arriba y hacia abajo de la membrana. Su velocidad de avance depende de la frecuencia y de las características de la membrana basilar. En algún punto de la cóclea la velocidad es cero. Cerca de ese punto, la membrana oscila hacia arriba y hacia abajo con mayor fuerza y absorbe la energía de la onda. Cada punto de la membrana basilar responde así a una determinada frecuencia. **[18]**

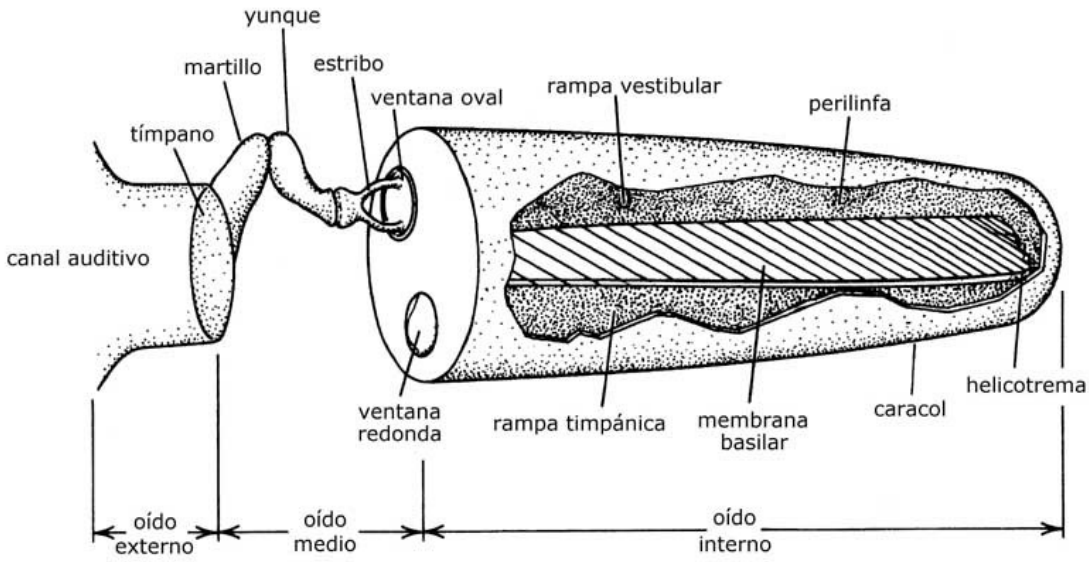
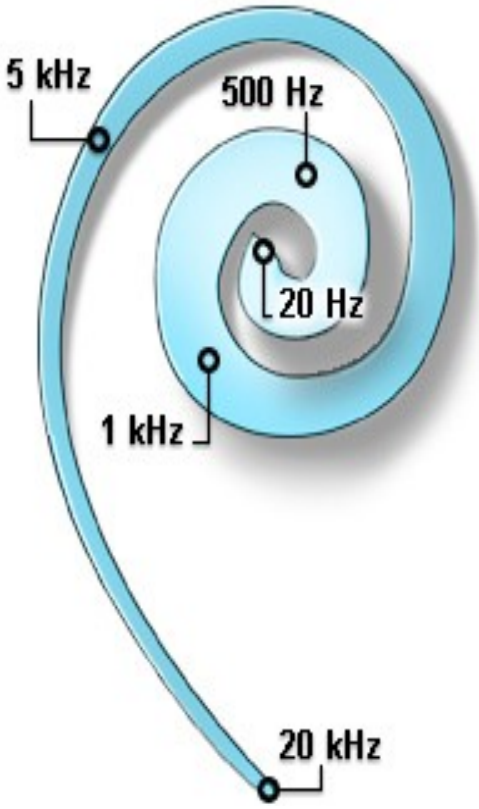


Ilustración 11 Interpretación longitudinal del Sistema Auditivo Humano.

Cuando el oído recibe un sonido con varias frecuencias, cada una de ellas excita un punto en la membrana basilar, de modo que el cerebro puede interpretar además de la altura del sonido su timbre, sin más que discernir qué terminaciones nerviosas fueron excitadas y con cuánta intensidad se excitaron. Es decir, el oído interno funciona como un analizador de sonidos cuyo grafico de muestreo es el siguiente:



band	fc(Hz)	BW(Hz)
1	50	80
2	150	100
3	250	100
4	350	100
5	450	110
6	570	120
7	700	140
8	840	150
9	1000	160
10	1170	190
11	1370	210
12	1600	240
13	1850	280
14	2150	320
15	2500	380
16	2900	450
17	3400	550
18	4000	700
19	4800	900
20	5800	1100
21	7000	1300
22	8500	1800
23	10500	2500
24	13500	3500

Ilustración 12 Muestreo Frecuencial del Oído Interno.

Como se aprecia en la Ilustración 12, cada una de las bandas de la membrana basilar es independiente. Ambas tienen a su vez diferentes anchos

de banda. Y cada una de ellas está centrada en una frecuencia central distinta. También hay que señalar que las bandas no presentan una relación lineal entre su frecuencia central y su ancho de banda, si no más bien presentan una relación exponencial entre ellas. Veamos un ejemplo con 6 bandas:

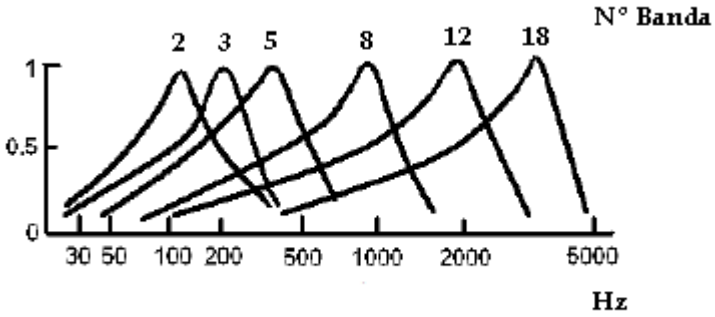


Ilustración 13: Ejemplo de 6 Bandas de la membrana basilar.

Por tanto, podemos afirmar que el Sistema Auditivo Humano responde como un **Banco de Filtros Independientes Paso Banda** que abarca un rango frecuencial entre los 20Hz y 20.000H y que se denominan **Bandas Críticas.** Una **Banda Crítica** constaría de un determinado número de frecuencias alrededor de una frecuencia central que activan la misma zona de la membrana basilar. Es decir, el conjunto de frecuencias colindantes que comparten el mismo estimulador nervioso de la membrana basilar dentro del conjunto de filtros pasabanda, con bandas superpuestas, al que corresponde el Oído Humano (*3.2*). Cada Banda Crítica correspondería en respuesta frecuencial como una GAUSSIANA centrada en una determinada frecuencia y tiene una determinada anchura frecuencial (una determinada varianza). El sistema auditivo humano sería por tanto un sistema de filtros Gaussianos que comparten entre ellos parte de sus espectros, que están centrados en diferentes frecuencias y que tienen diferentes varianzas entre ellos. Y además de esto, cada uno tiene su salida independiente correspondiente.

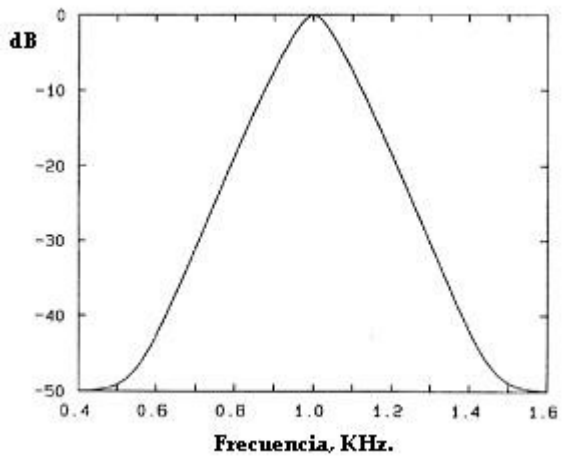


Ilustración 14 Ejemplo de Banda Crítica en centrada en 1KHz. Corresponde a uno de los estimuladores de la membrana basilar, y por tanto, será uno de nuestros filtros del conjunto que forman ese sistema de filtros pasabanda.

Este comportamiento puede modelarse de distintas maneras, desde las más sencillas utilizadas, por ejemplo, en la clásica extracción de parámetros MFCC o más elaboradas como las que utilizaremos en este proyecto y a las que nos referimos como "modelos cocleares". **[18]**

3.3 MODELOS COCLEARES ERB Y SENEFF. Inicios en Bancos Filtros MEL y Bark.

Tras demostrar como el Oído Humano se comporta como un Banco de Filtros Independientes colocados a lo largo de todo el espectro entre 20 y 20.000Hz, queda ya sólo pensar cómo se puede implementar éste para el uso de los RAH. Tiempo atrás, el cómputo de las máquinas limitaba dicha actividad, por lo que se utilizaban modelos de filtros uniformes y similares repartidos de igual manera por todo el espectro audible. Según avanzó la capacidad de cómputo de las máquinas, se produjeron los primeros tipos de Modelos Cocleares en cuanto a implementación se refiere, que poco a poco se van acercando más al comportamiento del Oído Humano.

Los primeros Modelos Cocleares que consideraremos son los modelos de filtros **MEL** y **Bark**. Ambos eliminan el carácter de filtros uniformes utilizado en un principio y comparten una semejanza entre ellos más que notable. Consisten en un modelo de filtros lineales con anchos de banda y frecuencias centrales repartidas de manera exponencial a lo largo del espectro, siendo los anchos de banda de estos filtros dependientes de la frecuencia en la que están centrados. Así pues, consiguen simular ese reparto de mayor número de bandas en las frecuencias bajas, con respecto a las altas, acercándose más así al comportamiento coclear del Oído Humano. Numerosos estudios, demuestran una mejoría notable con la utilización de estos filtros con modelo MEL y Bark con respecto a los uniformes utilizados en los inicios: **[24] [25]**.

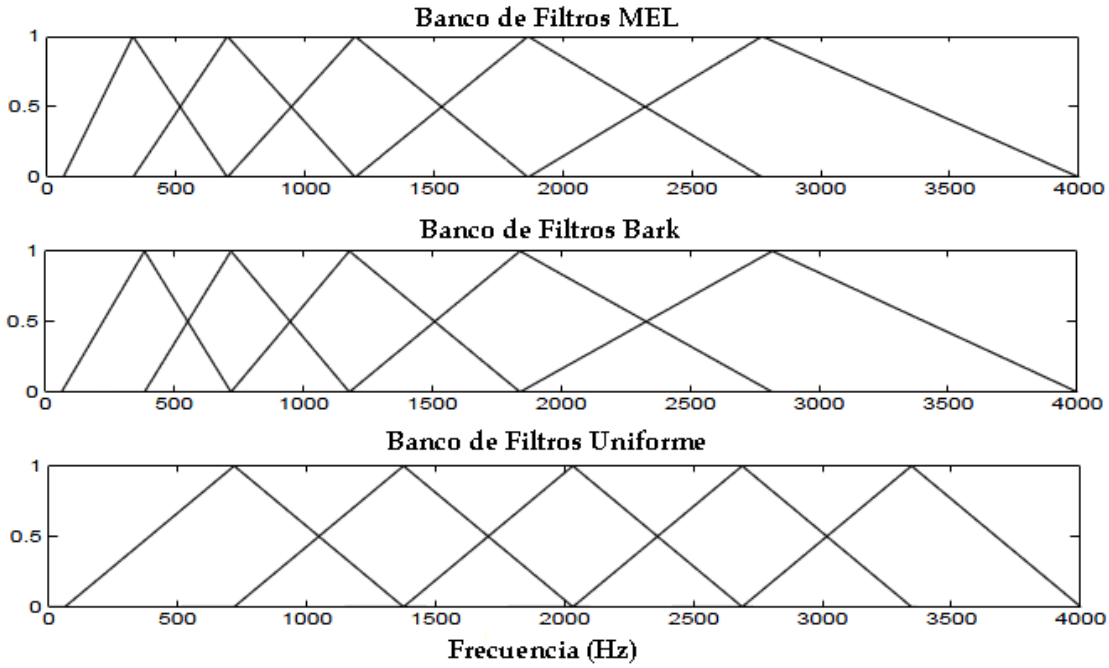


Ilustración 15 Bancos de Filtros MEL, Bark y Uniforme.

En nuestro PFC, la extracción **MFCC** utiliza un Banco de 40 Filtros MEL. Aún así, en nuestro PFC trabajamos con otras dos extracciones diferentes, **ERB** y **Seneff;** que utilizan unos modelos cocleares más cercanos al real utilizado por el oído humano. Ambos modelos, han podido desarrollarse y mejorarse gracias al gran avance computacional producido.

El Banco de Filtros **ERB** presenta una gran mejora en cuanto a modelo coclear ya que implementa unos filtros no similares entre sí y de caída no lineal. Así pues, los espectros de los filtros abarcan más rangos frecuenciales adyacentes a él, y las caídas de éstos tienen un comportamiento dependiente de la frecuencia central en la que se encuentren los mismos **[26]**. En nuestro PFC la extracción ERB utilizará un banco de 40 filtros.

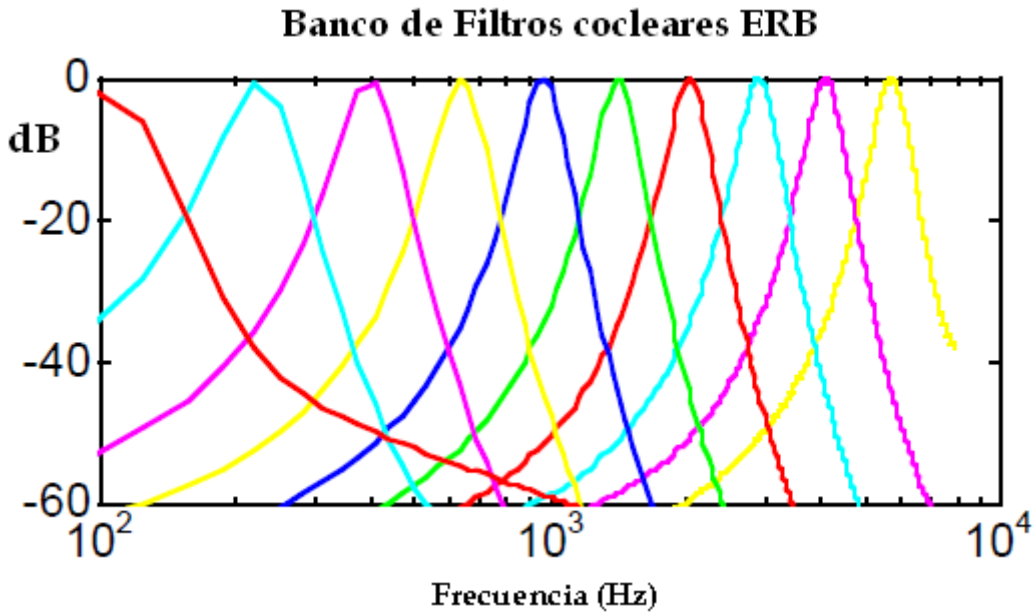


Ilustración 16 Banco de Filtros utilizado en extracción ERB

En cuanto al Banco de Filtros utilizado en la extracción **Seneff** se podría decir que es similar al ERB, con ligeras variaciones en caídas y frecuencias centrales de los filtros; pero que tiene como principal característica que los filtros son de amplitud dependiente de su frecuencia central. Seneff realza la facilidad y/o dificultad que tiene el oído humano al captar las diferentes frecuencias del espectro audible. Así pues, simula la facilidad de captación de sonidos en torno a 1Khz y 2Khz (Voz Humana), y la dificultad para captar frecuencias muy bajas (torno a 40Hz) o muy altas (torno a 10Khz). Esto se corresponde burdamente con las llamadas Curvas Isofónicas del Oído Humano (Ilustración 19), característica compleja de nuestro sistema auditivo que se introducirá en el próximo apartado. Citar que nuestra extracción Sennef utiliza igualmente un Banco de 40 Filtros.

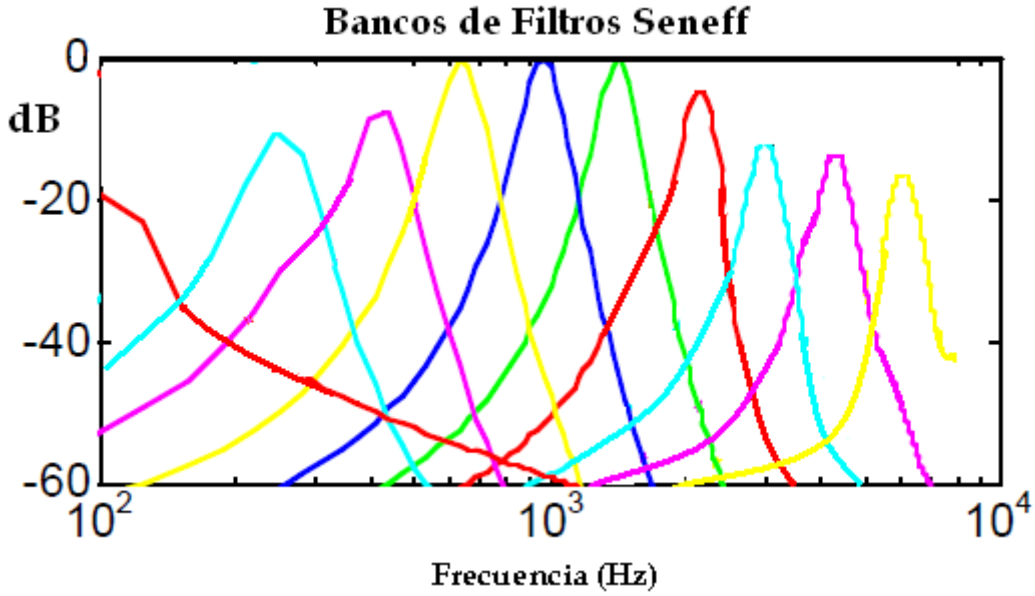


Ilustración 17 Bancos de Filtros Seneff

3.4 EL FENOMENO DEL ENMASCARAMIENTO SONORO.

Antes de ver el enmascaramiento en sí, debemos presentar una característica fundamental de la relación sonido/oído: **la Sonoridad.** Esta característica de la percepción humana, y sin entrar todavía a un nivel científico, puede ser expuesta en una gráfica a lo largo del espectro audible como se muestra en Ilustración 18. En ésta se indica el nivel de Sonoridad adecuado o aceptable para cada una de las frecuencias y representa en ella dos aspectos coloquiales como son "la voz" y "la música". Hay que explicar tan sólo que "Umbral de audibilidad" es el mínimo nivel de sonoridad necesario para oír cada una de las frecuencias, y que "Umbral de dolor" es el nivel máximo aceptado por nuestro sistema de audición sin sufrir daños irreparables en caso de exponernos mucho tiempo a éste. **[12]**

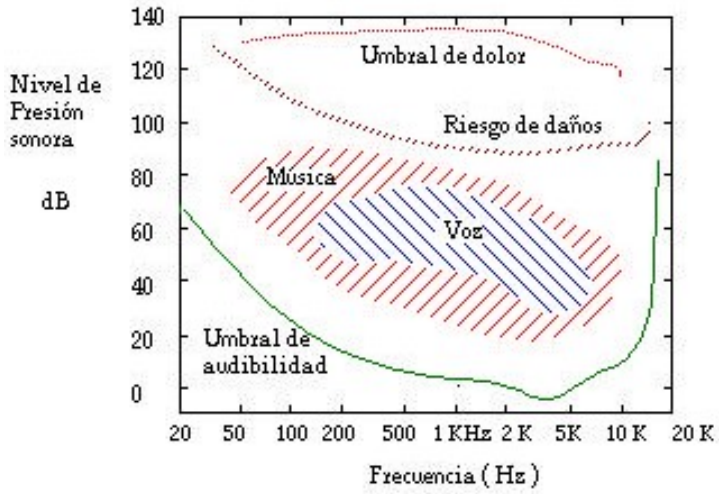


Ilustración 18 Margen audición correcta del Sistema Auditivo Humano.

Una forma práctica de abordar el problema de **la Sonoridad** es medir el nivel de sonoridad de nuestra señal (los dBs) y ver su influencia en la **Curvas Isofónicas** *(Ilustración 19)* del oído humano, es decir, determinar cuándo para nuestro oído un sonido es igual de fuerte que otro, y cuando no. Las Curvas Isofónicas del oído nos señalan cuando, un determinado sonido con su frecuencia e intensidad correspondiente, llega intrínsicamente a ser más audible por nuestro oído que otros en otras frecuencias; siendo, por tanto, susceptible de ser escuchado más fácilmente y mejor que el resto que suenan en ese mismo instante temporal. Para ello, deberemos tener en cuenta solamente tres variables, la intensidad de los sonidos, las frecuencias a la que están centrados, y la siguiente gráfica:

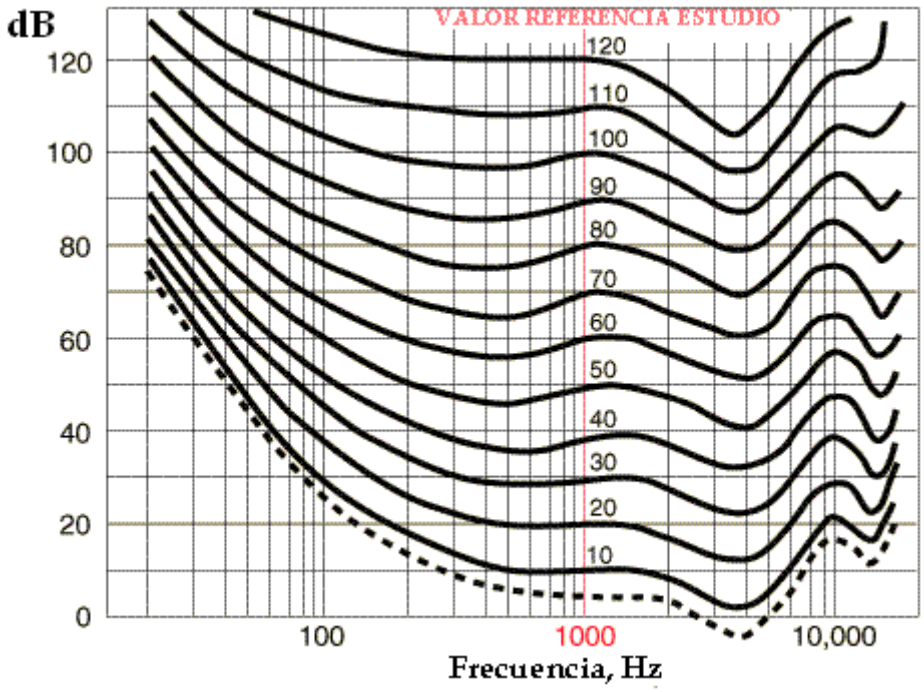


Ilustración 19 Curvas Isofónicas del Oído Humano.

La gráfica anterior muestra el comportamiento del oído ante sonidos a campo libre. Como se puede observar a simple vista, para que un sonido de frecuencia baja se escuche por encima de uno en torno a 1KHz tiene que tener obligatoriamente una intensidad sonora mayor (mayores dBs), también se puede ver fácilmente que el comportamiento del oído en las frecuencias altas no presenta uniformidad (a diferencia de las frecuencias bajas), y que; a mayores niveles sonoros, más plana es la curva de la respuesta sonora del oído. **[13]**

Así por tanto con lo anterior, ya podríamos y sabríamos decir con notable exactitud si un sonido centrado en determinada frecuencia es o no más fuerte que otro, o si tal o cual frecuencia al mismo nivel es más fuerte que la otra, quedando por tanto cerrado el tema de la Sonoridad y podemos seguir con la definición del **Enmascaramiento Sonoro Temporal**.

El **Enmascaramiento Sonoro Temporal** se produce cuando un sonido de una determinada frecuencia queda tapado por un sonido más fuerte de diferente frecuencia a lo largo del tiempo. El sonido fuerte se denomina *enmascarador*, y el débil *enmascarado* o *señal*. El enmascaramiento puede asimilarse a un defecto de audición: el enmascarador aumenta nuestro umbral de audición, es decir incrementa la intensidad que tiene que tener el sonido para que lo podamos oír correctamente. Físicamente el enmascaramiento tiene que ver con la distribución de los receptores a lo largo de la membrana basilar y al funcionamiento del oído interno. Sin entrar en más detalles físicos, definimos el enmascaramiento en el tiempo contando que, siempre que un sonido fuerte enmascara a otro más débil, se producen **3 fases de enmascamiento temporal** a lo **largo del tiempo** durante dicho proceso, las cuales son: **[14]**

Pre-Enmascaramiento: Aún cuando no sea fácilmente imaginable, sonidos que aún no existen pueden enmascarar sonidos ya existentes. Este proceso se reduce a lapsos sumamente reducidos, aproximadamente en el orden de los 20ms.
La explicación podría ser que los sistemas físicos no realizan saltos de tipo abrupto sino que realizan más bien transiciones continuas. De esa manera es posible pensar que el tiempo de ataque de un hecho sonoro puede ser del orden de esos 20ms.

Enmascaramiento Simultáneo (enmascaramiento frecuencial): Aunque coexistan los dos sonidos en el tiempo, nuestro oído sólo puede detectar el sonido fuerte ya que el otro ha quedado enmascarado por él. En este caso, influye la distribución frecuencial de ambos sonidos puesto que, como hemos visto, la sonoridad es fuertemente dependiente de la frecuencia.

Post-Enmascaramiento: Proceso ocurrido durante los primeros 150/200ms después de que se apague el sonido enmascarador y cuya característica es que prácticamente no se produce ninguna diferencia con

respecto al proceso del Enmascaramiento Simultáneo. Es decir, que en torno a los 150/200ms después de que el sonido enmascarador ha terminado, la sensación de enmascaramiento en el oído continúa. La caída/duración de este Post-Enmascaramiento es más variable que la del Pre-Enmascaramiento, y éstas dependen de los niveles de intensidades de los sonidos con los que estemos tratando. **[13]**

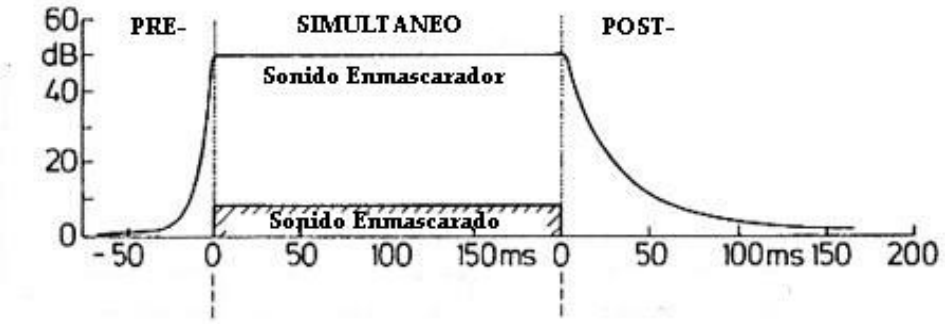


Ilustración 20 Enmascaramiento Sonoro en el Tiempo.

Como vemos, el procedimiento de enmascaramiento extiende el efecto del sonido enmascarador en dos dimensiones: la temporal y la frecuencial. Por ese motivo, en este proyecto, hemos considerado la representación bidimensional (espectrogramas y cocleogramas) como punto de partida, lo que nos permite aplicar métodos de procesado de imágenes para simular los efectos del enmascaramiento como el que presentamos en la próxima sección.

En relación con los modelos cocleares o de bandas críticas que antes mencionábamos, podemos afirmar que, para que un enmascaramiento se produzca como tal, aparte de tener un sonido más fuerte y otro/s débiles que coexistan en el tiempo, ambos deben también compartir la misma Banda Crítica, o como mínimo; bandas críticas colindantes.

Las **bandas críticas inmediatamente colindantes** pueden compartir espectros pequeños de frecuencia, a modo de que la misma "frecuencia enmascaradora" puede afectar a dos o más Bandas Críticas. Sabiendo de antemano que todas las Bandas Críticas tienen una salida totalmente independiente, y que cada una tiene su respuesta Gaussiana en su espectro a trabajar, podemos tener el caso que la misma componente frecuencial afecte a dos o más Bandas siendo ésta la componente enmascaradora de todas. Dándose por tanto, la posibilidad de que Bandas Críticas colindantes obtengan la misma componente enmascaradora en su espectro, pudiendo ser ésta incluso de la misma potencia para ambas en casos determinados.

Como se explicará posteriormente, en nuestro PFC, hemos desarrollado la simulación del concepto de **Enmascaramiento Temporal** y frecuencial y el concepto de **Banda Crítica**. No hemos desarrollado, sin embargo efecto de las **Curvas Isofónicas** *(Ilustración 19)* que nos obligaría en un principio, a utilizar máscaras o elementos estructurantes (véase la sección 3.5) dependientes de la

intensidad sonora; siendo ésta última simulación una posible Línea Futura de Investigación que mejorará el realismo y fidelidad de nuestros resultados, acercando nuestro RAH más al comportamiento real del Oído Humano.

3.5 PROCESADO MORFOLÓGICO.

En nuestro PFC hemos implementado el enmascaramiento haciendo un Procesado Morfológico que trata el proceso del Enmascaramiento Temporal y que se puede aplicar independientemente del método de extracción de Características que se haya utilizado, ya sea MFCC (Clásico); o sea ERB o Seneff (Cocleares). En nuestros métodos de extracción, siempre las características se sacan a modo de bandas independientes con diferentes Bancos de Filtros. Estos bancos de filtros pueden ser los MEL clásicos, como es el caso del MFCC, o pueden ir simulando de forma más fiel el comportamiento de la cóclea como es en caso del ERB y Seneff (véase la sección 2.2). La salida de estos bancos de filtros, será lo que posteriormente llamaremos Coeficientes de Extracción, son espectrogramas que indican la energía de las diferentes bandas frecuenciales de filtrado a lo largo del tiempo.

Por tanto, nosotros sólo tenemos que añadir a dichos Coeficientes de Extracción un SCRIPT que pueda simular todo el proceso del Enmascaramiento Sonoro Temporal que se explicó en la sección 3.4 con el cual conseguiremos mejorar nuestras extracciones acercándolas un poco más al verdadero funcionamiento del Oído Humano. Es de esperar; por tanto, que esta simulación adicional del enmascaramiento sonoro los modelos de extracción cocleares (ERB y Seneff), como para el modelo clásico (MFCC).

Resumiendo lo anterior, la simulación del proceso de enmascaramiento se producirá inmediatamente después de la extracción de los Coeficientes de Energías en Bandas y, básicamente; aplicaremos una variación al valor de estos Coeficientes de salida de la extracción tomando en cuenta los patrones de enmascaramiento citados anteriormente en la sección 3.4. Estos patrones serán aplicados a nuestra parametrización realizando los siguientes razonamientos:

> **Enmascaramiento Temporal Simultáneo:** Al hacer la simulación del enmascarado después del proceso de cálculo de los Coeficientes, no tendremos que preocuparnos por frecuencias de la misma Banda que coincidan en el tiempo; ya que, al calcular los Coeficientes en nuestras extracciones tenemos siempre la envolvente de la señal por bandas, y esto no es más que la frecuencia de mayor potencia para cada una de las bandas independientes. Es decir, nuestro Software de extracción de características ya nos asegura que tenemos la frecuencia que enmascara a todas las demás en cada Banda Crítica.

Enmascaramiento Temporal no Simultáneo: Un valor alto en potencia en una banda de Coeficientes determinada afectará un tiempo "*x*" por delante de la misma, y un valor temporal "*y*" por detrás. Siendo siempre "$x < y$". Esto simularía por tanto el enmascaramiento temporal que produce una frecuencia potente dentro de su propia Banda Crítica, y de cómo afecta en el Pre/Post- Enmascaramiento a las siguientes envolventes de dicha Banda.

Bandas Críticas Colindantes: Como se vio anteriormente, las bandas colindantes comparten parte de su espectro, y la misma componente frecuencial puede afectar a la salida de dos o más Bandas. Sin quitar que dichas salidas son completamente independientes, y de que cada banda tiene su propia respuesta Gaussiana en su espectro a tratar, tenemos que dos Bandas pueden tener la misma salida en cuanto a componente enmascaradora, e incluso, que ésta pueda ser de potencia similar por el simple hecho de que un sonido fuerte afecte a 2 o más bandas.
En nuestro caso, tomaremos que un valor muy alto en potencia en una banda determinada, puede afectar a los valores de los Coeficientes en sus Bandas inmediatamente colindantes; haciéndolas en este caso, dependientes entre ellas por dicho valor de gran potencia.

Visto desde el punto de vista del Software, para simular todo lo anterior tenemos que empezar a trabajar desde el espectrograma que contiene los Coeficientes de las extracciones de energías en bandas en el "*eje y*", y en el tiempo en el "*eje x*". Esta matriz; recordemos de nuevo, la conseguimos directamente de cada una de las diferentes extracciones, ya sea MFCC, ERB u Seneff. En esta Matriz existen esos Coeficientes con valores altos los cuales queremos que afecten en el tiempo tanto antes como después de su existencia, (pre- y post-Enmascarado), así como que también afecten de algún modo a los valores que se encuentran en las Bandas Colindantes de los mismos. Estos se pueden mostrar fácilmente en una gráfica de intensidades. He aquí un ejemplo:

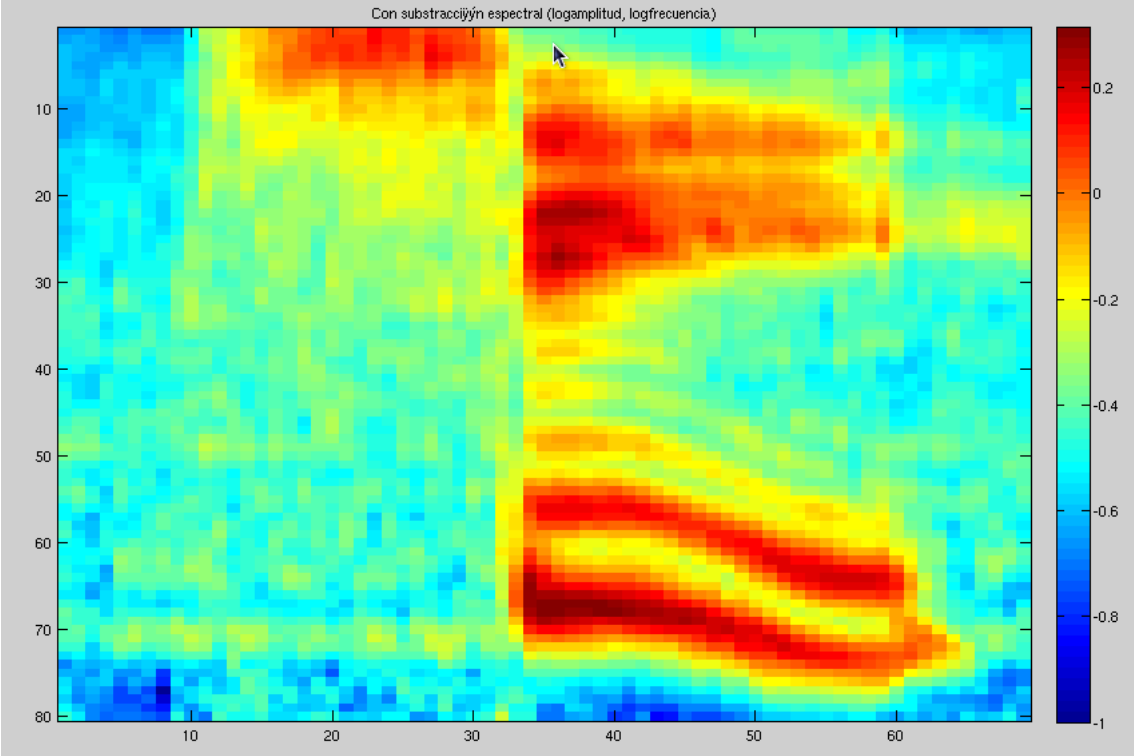


Ilustración 21 Ejemplo de espectrograma de una señal sin Procesado Morfológico.

Así pues, realizaremos un filtrado morfológico de la matriz de energías en bandas. Para ello partiremos de una **Matriz Binaria** a la que llamaremos **Elemento Estructurante o Máscara**. Este Elemento Estructurante o Máscara, alterará los valores contiguos al valor alto tanto en el "*eje x*" que corresponde al tiempo, como en el "*eje y*" el cual corresponde a las Bandas Críticas. Cada "*x píxel*" corresponderá con 10ms y cada "*y píxel*" con una Banda diferente. Dicha Máscara seguirá un patrón elegido por nosotros considerándolo de antemano coherente con la teoría del funcionamiento del Oído Humano. Su funcionalidad será la de que ese valor alto aislado en el espectrograma se convierta ahora en una "mancha" de valores altos que refleje su importancia tanto en el Tiempo como en las Bandas Colindantes a éste. Así pues, mostramos un ejemplo de una de nuestras máscaras elegidas:

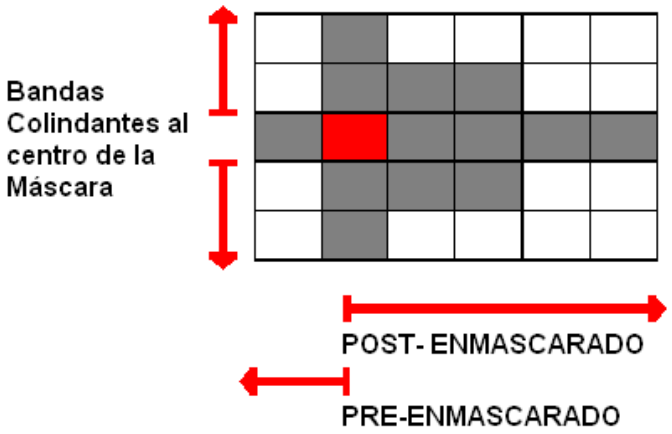


Ilustración 22 Ejemplo Máscara utilizada en el Procesado Morfológico.

Para realizar esta alteración a modo de "mancha" en los valores altos de la señal, utilizamos la llamada **dilatación de imagen**. La dilatación de imagen corresponde con un crecimiento de los valores elegidos con la forma del elemento estructurante o máscara que se utilice. He aquí un ejemplo en un caso binario:

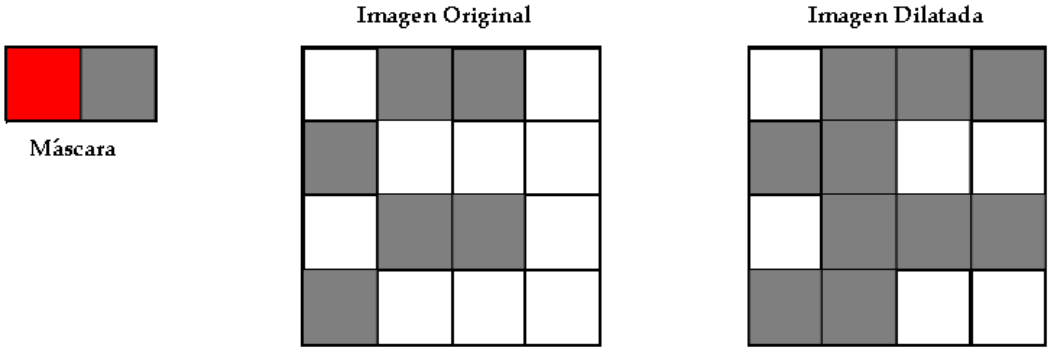


Ilustración 23 Ejemplo Dilatación Binaria.

En nuestro caso, no tendremos un caso binario ya que los valores de nuestra matriz de coeficientes son muy dispares; y los considerados valores elegidos serán los valores altos que rescatamos anteriormente. El proceso en este caso, viene dado por un algoritmo para la estructura Matlab fácilmente aplicable **[21].** El resultado de esta dilatación que expande a modo de Máscara los valores altos iniciales, tendrá como resultado una nueva Matriz que se sumará a la inicial para obtener nuestra matriz final. Esta matriz final será la que muestra a nuestro parecer, todo lo explicado en este Procesado Morfológico simulando así más fielmente el comportamiento del Oído Humano. Veamos los siguientes ejemplos gráficos en los que se verán, en primera instancia, cómo el procesado morfológico resalta las zonas de mayor intensidad de la señal, y de cómo luego esas zonas, enmascaran las zonas que tienen alrededor con la forma de la Máscara elegida en el proceso.

Ejemplo 1, con gráficas del proceso:

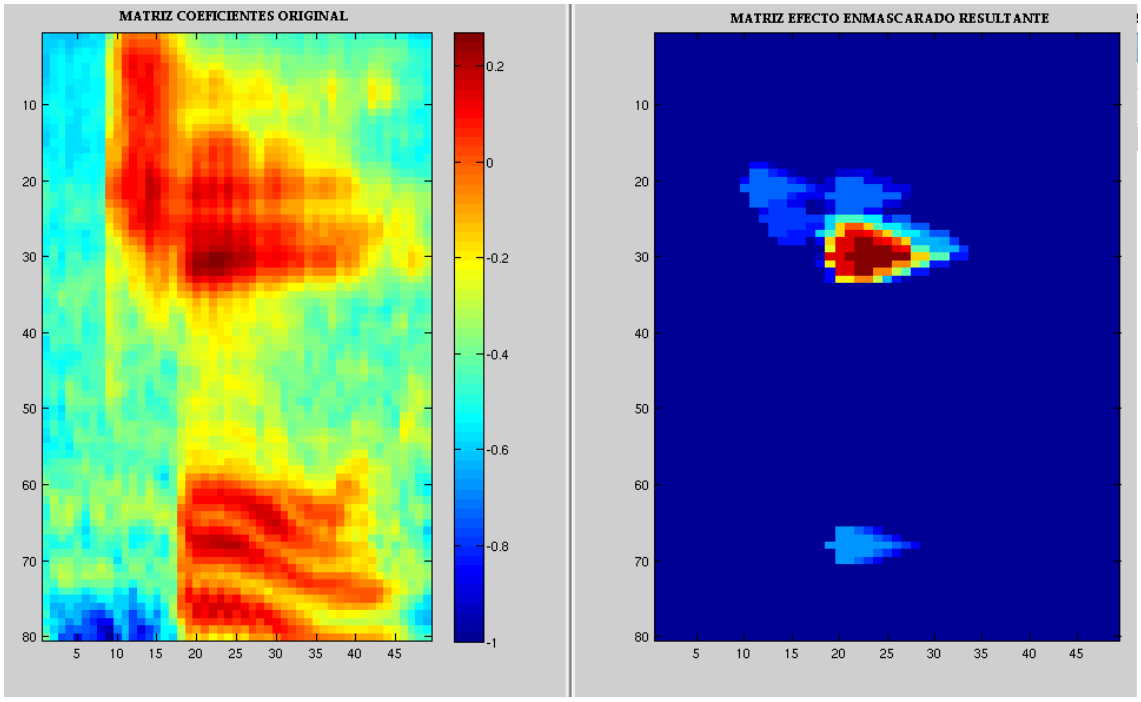


Ilustración 24 Matriz Coeficientes original y Matriz de datos enmascarados.

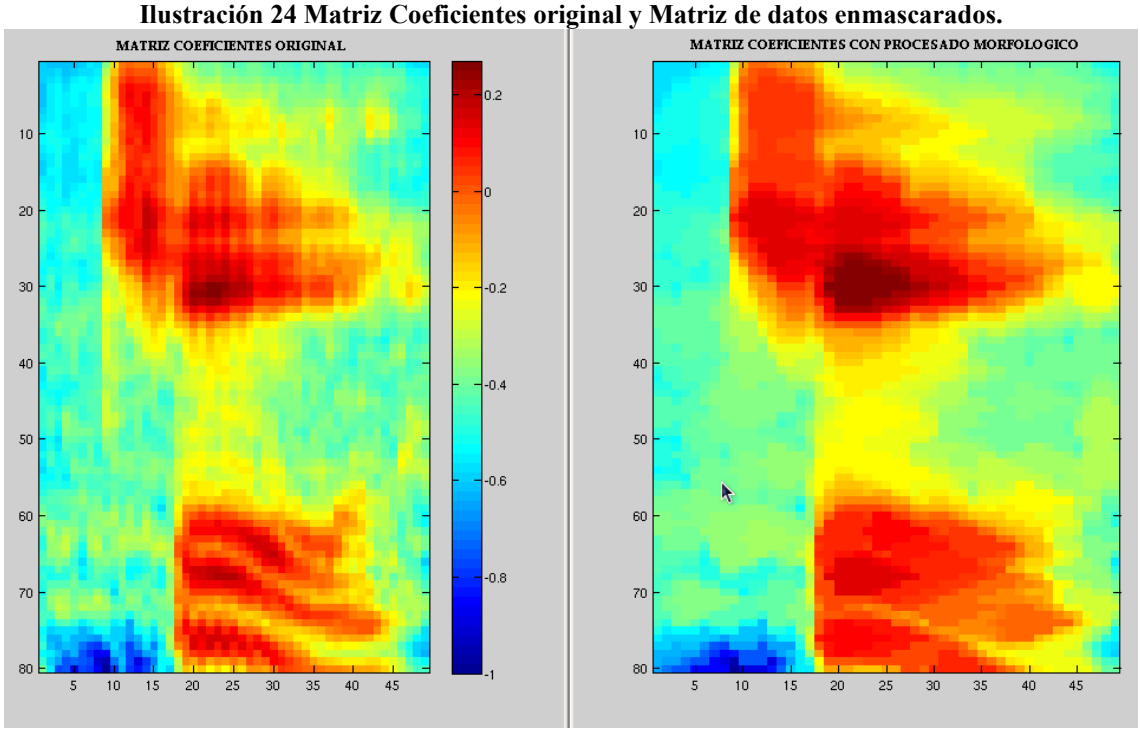


Ilustración 25 Matriz Coeficientes original y Matriz Coeficientes con Procesado Morfológico.

Ejemplo 2 con gráficas del proceso:

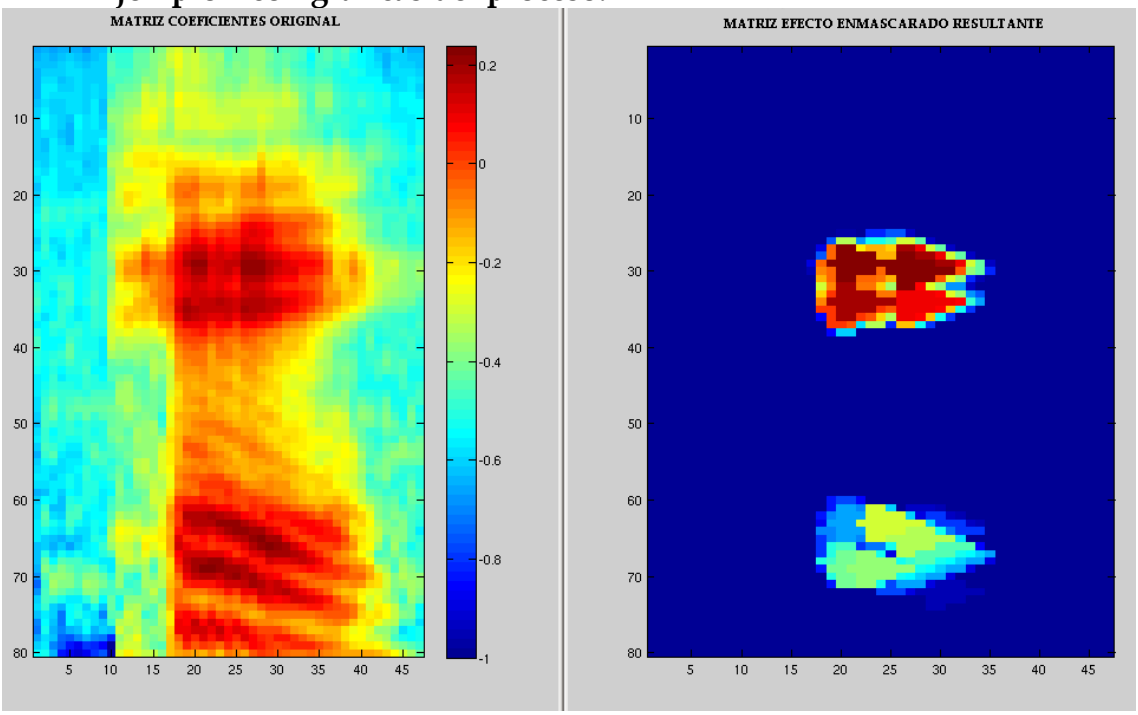


Ilustración 26 Otra Matriz Coeficientes original con su Matriz de datos altos enmascarados.

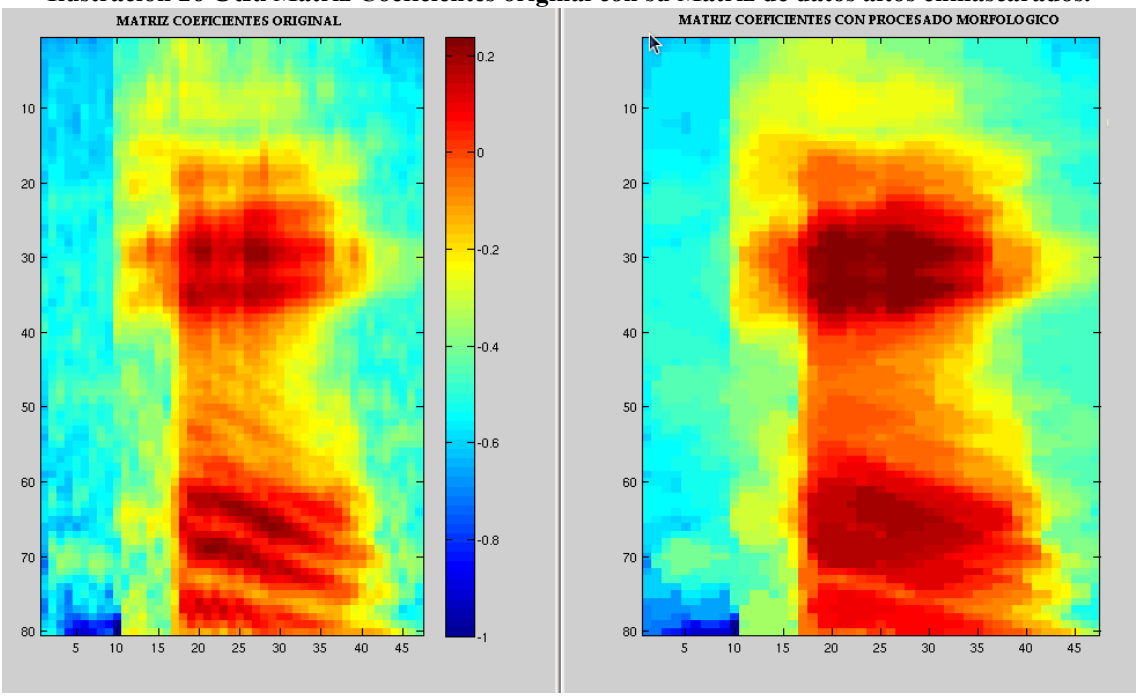


Ilustración 27 Otra Matriz Coeficientes original con su Matriz de Coeficientes con Procesado Morfológico.

3.6 MODELO DE SIMULACIÓN PROPUESTO.

Como conclusión; y después de haber visto el desarrollo de los apartados anteriores, presentamos el modelo de simulación propuesto en nuestro PFC y lo explicamos a continuación:

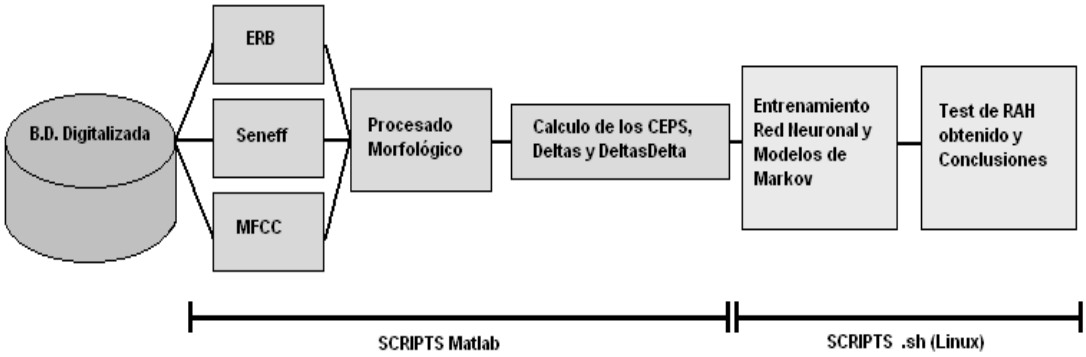


Ilustración 28 Esquema del Modelo Propuesto en nuestro PFC.

B.D Digitalizada

Base de datos en formato **.sph** denominada **Isolet** **[20]** que contiene un total de 7800 señales de voz en formato *limpio*, y de nuevo, las mismas 7800 señales tratadas con diferentes ruidos añadidos que serán el formato *contaminado o ruidoso*. Ambos conjuntos de 7800 señales serán todas utilizadas en nuestro PFC. Cada una de las 7800 señales fue grabada en dos ocasiones por cada uno de los 150 locutores que participaron en el proyecto. La B.D forma un total de 1,25 horas de grabación de habla continua, y todas las grabaciones se hicieron en laboratorio con un micrófono de cancelación de ruido para asegurar su fidelidad.

Extracción ERB, Seneff, MFCC

Ambas extracciones se realizan con un laborioso SCRIPT Matlab. La teoría y fundamento de todas ellas están explicados en la sección 2.2 de este PFC. Por tanto, solamente destacar de que de ambos SCRIPT sacamos los "Coeficientes cocleares" en el tiempo para cada una de las señales de nuestra B.D. Estos coeficientes cocleares serían las componentes de información que; según los modelos de extracción ERB, Seneff o MFCC, obtendría un oído humano ante las señales de entrada de la B.D. Estos coeficientes cocleares que se presentan como datos de salida, serán tratados y preparados para todas etapas posteriores del modelo propuesto. Dichas etapas posteriores recibirán; por tanto, unas Matrices de Coeficientes de extracción frente al tiempo para cada una las señales que le entraron al proceso.

Procesado Morfológico

SCRIPT Matlab que modela todo lo explicado en la sección 3.5. Dicho SCRIPT puede ser utilizado independientemente de la extracción elegida ya que recibe como entrada la matriz de los Coeficientes de extracción en el tiempo, y ofrece de salida esa misma matriz con los datos variados según el proceso de enmascaramiento.

Calculo de CEPS, Deltas y Deltadeltas

Obteniendo como salida los coeficientes de extracción ya tratados con el fenómeno del enmascaramiento, procedemos al cálculo de los **CEPS, Deltas y Deltadeltas** por medio de otro SCRIPT Matlab.

Los **CEPS (Coeficientes Cepstrum)** nos ofrecen mucha información frecuencial de la onda de entrada en un formato notablemente reducido, y se calculan simplemente aplicando la DCT a los coeficientes y diezmando la información obtenida. Éstos CEPS serán los que representen la información frecuencial de cada una de las señales a lo largo del tiempo. En este PFC trabajaremos únicamente 13 CEPS a lo largo del tiempo. Estos 13 CEPS se reparten siendo el primero de ellos (Co), el que contiene la información de la *Energía* de la señal de voz tratada, y los 12 CEPS siguientes, los que contienen la información frecuencial de la señal en sí. Viendo todas sus ventajas tomaremos ahora esa Matriz de Ceps en el tiempo como nuestra nueva unidad de información, siendo esta la que mejor representa la información de cada una de las señales de nuestra B.D.

Los **Deltas (Delta Ceps)** se ofrecen como una mejora significativa de la información obtenida de la señal en el dominio Cepstrum. Los Deltas ofrecen información sobre la V*elocidad* de la onda de voz tratada. Estos Deltas nos mostrarán su información siendo ellos únicamente 13 coeficientes en este PFC (al igual que los CEPS). El primer coeficiente Delta (Delt 0) contiene la información de la *Velocidad* de la señal de voz tratada, y los 12 siguientes, información de su fluctuación. Estos Deltas se calculan aplicando la siguiente fórmula a la matriz de CEPS en el tiempo obtenida anteriormente, siendo: **[17]**

$$DELTAS(t)=\frac{\sum_{x=1}^{\alpha} x((CEPS(t+x)-CEPS(t-x))}{2\sum_{x=1}^{\alpha} x^2}$$

α =VentanaDelta t = Trama Temporal de 13 Coef.

Ilustración 29 Fórmula de cálculo de Deltas.

Como se puede ver en la fórmula anterior, cuanto más grande sea la elección de la **VentanaDelta**, más tramas es necesario utilizar en cada cómputo pagando por ello un coste computacional más alto. Hay que encontrar la relación calidad/coste del cálculo ya que éste se hará para cada uno de los instantes temporales de cada una de las señales de la B.D. En nuestro caso hemos elegido una **VentanaDelta = 2**. Explicar también que, en los límites donde *(t+x)* o *(t-x)* se salen del número de tramas temporales de la señal, elegimos siempre réplicas de la trama frontera de dicho borde. La última o la primera respectivamente.

Y por último, el cálculo de los **Deltadeltas (DeltaDeltaCEPS, o DoubleDelta)**. Éstos se presentan como la segunda mejora significativa de la información obtenida de la señal. Los Deltadeltas ofrecen información sobre la *Aceleración* de la onda de voz tratada. Estos Deltadeltas nos mostrarán su información siendo ellos únicamente 13 coeficientes (al igual que los CEPS y los Deltas). El primer coeficiente Deltadelta (DeltaDelt 0) contiene la información de la *Aceleración* de la señal de voz tratada, y los 12 siguientes, información de su cambio a lo largo del tiempo. Estos Deltadeltas se calculan aplicando la fórmula de cálculo de los Deltas, a la matriz de Deltas propiamente dicha. **[*15*]**

Así pues, después de todo este proceso de extracción de la información frecuencial de la señal, tendremos que para cada una de estas 15600 señales de entrada tratadas, (7800 en formato limpio y 7800 contaminadas con ruido), hemos obtenido un total de **39 coeficientes** para cada uno de los momentos temporales de las mismas. Estos **39 coeficientes** se repartirán en cada instante temporal, para cada una de las señales, de la siguiente manera:

1 Coeficiente de información de *Energía* + 12 Coeficientes **CEPS.**
1 Coeficiente de información de *Velocidad* + 12 Coeficientes **Delta.**
1 Coeficiente de información de *Aceleración* + 12 Coeficientes **Deltadelta**.

Ilustración 30 Ejemplo de los Coeficientes obtenidos para cada instante temporal, para cada una de las señales de toda la B.D.

Entrenamiento Red Neuronal y Modelos de Harkov.

Proceso realizado por un único SCRIPT .sh de Linux, el cual recoge la información tratada de los 39 Coeficientes CEPS en el tiempo, de todas y cada una de las 7800 señales de la B.D, para llegar a entrenar la Red Neuronal

(mediante Quicknet) **[22]** y Modelo de Markov (utilizando SPRACHCORE) **[23].**

Como primer paso separamos las 7800 señales en cinco bloques de 1560 señales cada uno a los que denominaremos "Carpetas", y después comenzamos con todo el proceso de entrenamiento de la red. Debido a que la base de datos es pequeña, utilizaremos un procedimiento de **Validación Cruzada** (Jack-knife o leaving-one-out) con cinco particiones. Esto es que, de las 5 carpetas que tenemos con nuestros datos; tanto para señales en limpio como para ruidoso, el entrenamiento se hará siempre con las 4 carpetas diferentes a la que vamos a utilizar como datos de entrada en el test. Es decir, si utilizamos como datos de entrada la carpeta 1, el entrenamiento se hará con las carpetas 2, 3, 4 y 5. Y si hacemos lo propio con la carpeta 3, el entrenamiento se hará con las carpetas 1, 2, 4 y 5.

Este entrenamiento de validación cruzada se hará tanto para las señales en limpio, como para las contaminadas (o ruidosas), de forma que la red queda entrenada con sus cinco carpetas para ambas formas de manera independiente. Sin entrar en más detalles que no conciernen a nuestro PFC, diremos que el objetivo de este SCRIPT es conseguir que la red aprenda automáticamente las propiedades deseadas con dichos Conjuntos de Entrenamiento. Pudiendo de esta manera aprender a clasificar e identificar diferentes señales de los fonemas previamente etiquetados, extrayendo de ellas únicamente sus características discriminativas. Las variables de este entrenamiento Neuronal y formas de clasificación de pesos serán invisibles a nuestro nivel de estudio, y simplemente este SCRIPT es una "caja negra" la que hemos adaptado y modificado a nuestras necesidades a nivel de código SHELL (.sh). El resultado de salida de este proceso es un conjunto de megadatos Linux que serán utilizados posteriormente en el Test de nuestro RAH.

Test del RAH obtenido y Conclusiones.

Proceso realizado también por un único SCRIPT .sh de Linux. Dicho SCRIPT se encarga de Testear el modelo de Markov junto a su Red Neuronal obtenida. Como se dijo anteriormente, la red se entrena con la **Validación Cruzada** para cada una de las carpetas, tanto para la parte de datos limpios, como la de datos contaminados (o ruidosos). Así pues, un test para un caso determinado consistiría en pasar por dicha red los datos de todas las carpetas con el sistema de Validación Cruzada, y ver los resultados de aciertos y fallos para cada una de las señales de fonemas de voz que componen dichas carpetas. Los resultados de los test se reflejarán en una matriz en la que se indican los fallos y aciertos de forma trivial así como su Tasa de Error correspondiente.

Aquí tenemos un ejemplo de test de una extracción ERB para la Carpeta 3 de datos ruidosos:

Tasa de Error 20.3% (317 incorrectas de las 1560).

Matriz de Confusión

Cada fila de la Matriz es una palabra diferente hablada por un humano.
Cada columna de la Matriz es una palabra oída por el reconocedor.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	NADA
A	44	1	0	1	4	2	0	3	1	0	2	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
B	0	40	0	0	4	0	0	0	0	0	0	0	1	0	0	5	0	0	0	0	2	8	0	0	0	0	0
C	1	0	41	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	14	0
D	0	1	0	47	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	2	1	4	0	0	0	2	0
E	0	8	0	2	37	0	1	0	0	0	0	0	0	0	0	4	0	0	0	0	0	6	0	0	0	1	1
F	1	0	0	0	0	47	0	0	0	0	0	3	4	2	0	0	0	0	3	0	0	0	0	0	0	0	0
G	0	0	1	0	1	0	48	0	0	4	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	1	1
H	1	0	0	0	1	0	0	53	0	0	0	0	0	1	0	0	0	0	0	0	0	3	0	0	0	1	0
I	0	0	1	0	0	0	0	0	49	0	0	0	0	0	2	0	0	3	0	0	0	0	0	0	3	0	2
J	1	0	0	0	0	0	2	0	0	54	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0
K	3	0	1	0	0	0	2	0	0	6	43	0	0	0	0	2	0	0	0	0	0	0	0	0	0	3	0
L	0	0	0	0	0	4	0	0	0	0	0	49	4	0	2	0	0	0	0	0	0	0	0	0	1	0	0
M	4	0	0	0	0	7	0	0	0	0	0	3	34	11	1	0	0	0	0	0	0	0	0	0	0	0	0
N	5	0	0	0	1	10	0	1	0	0	0	0	5	38	0	0	0	0	0	0	0	0	0	0	0	0	0
O	0	0	0	0	0	2	0	0	0	0	0	1	0	0	57	0	0	0	0	0	0	0	0	0	0	0	0
P	1	1	0	2	1	0	0	0	0	1	0	0	0	0	0	43	0	0	0	1	1	8	0	0	0	1	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54	0	0	0	6	0	0	0	0	0	0
R	0	0	0	0	0	1	0	0	1	0	0	0	0	0	2	0	0	55	0	0	0	0	0	0	0	0	1
S	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	51	0	0	0	0	3	0	0	0
T	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	56	0	0	0	0	0	0	0
U	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	1	0	0	0	56	0	0	0	0	0	0
V	0	6	0	0	1	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	47	0	0	0	3	1
W	1	0	0	2	0	2	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	50	0	0	0	0
X	0	0	0	0	0	6	0	2	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	50	0	0	0
Y	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	56	0	0
Z	0	0	10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	44	0

Tabla 1 Ejemplo de resultados de un Test para extracción ERB en Carpeta 3.

Estos test se harán para cada una de las carpetas, en cada uno de los siguientes modelos de test:

- Test de Carpetas en formato **Limpio** con la Red Entrenada para datos en **Limpio (data Clean-Clean train).**

- Test de Carpetas en formato **Sucio** con la Red Entrenada para datos en **Ruidosos (data Noisy-Noisy train).**

- Test de Carpetas en formato **Sucio** con la Red Entrenada para datos en **Limpio (data Noisy-Clean train).**

Así pues, tenemos tres diferentes de modelos de Test con las siguientes características:

En el caso de **data Clean–Clean train** estamos ante un modelo ideal en el que entrenamos una red con datos de información limpia grabados en condiciones ideales, y después; una vez puesto nuestro diseño a correr, recibimos de nuestro canal de datos señales igual de limpias. Condición última que nunca se da en el mundo real de Telecomunicaciones.

En el caso de **data Noisy-Noisy train** estamos ante un sistema más verosimil. Esto se debe a que entrenamos nuestra red con datos corruptos y con ruido ya desde un principio, con la intención de que nuestro sistema funcione perfectamente en un ambiente realista. Si los diferentes ruidos con el que entrenamos la red son similares a los que nos encontraremos en el mundo real cuando pongamos el sistema a correr, nuestros resultados serán excelentes sin duda alguna. Claro está, lo difícil es llegar a "descubrir" o "identificar" los ruidos que nos encontraremos en el canal a la hora de correr el sistema, los cuales serán los elegidos a la hora de entrenar nuestra red. De ser mal elegidos estos mismos, el entrenamiento en ruidoso no hará más que complicarnos aún más las cosas empeorando notablemente todos nuestros resultados. En nuestro PFC, los datos corruptos o sucios vienen ya con el ruido añadido a ellos, por lo que nos hemos saltado todo el proceso de "identificar" los ruidos que nos encontraremos facilitándonos así las cosas, y centrándonos sólo en el estudio que concierne a este PFC. Aún así, podemos afirmar de antemano, que los resultados obtenidos en este modelo de Test serán siempre peores que los obtenidos en el modelo ideal **data Clean–Clean train.**

Y por último, en el caso de **data Noisy-Clean train** estamos ante un modelo en el que renunciamos a conocer el tipo y nivel del ruido bien porque esto no es factible o porque es muy variable. Para ello entrenamos nuestra red con datos puros grabados en condiciones ideales; y, cuando ponemos el diseño a correr en un mundo real las señales nos llegan al sistema con ruidos aleatorios y desconocidos. Es de esperar por tanto que los resultados de este tipo de Test sean peores que en los otros dos modelos. Aún así, es el modelo más general si desconocemos el ruido al que nos enfrentaremos. Por lo cual, este modelado de test tendrá una importancia mucho mayor que sus

compañeros y será vital a la hora de decidir que configuración interna será la mejor para nuestro RAH.

Por último, señalar que este Proceso de Test sigue siendo un SCRIPT Shell adaptado y preparado a todas nuestras necesidades, así como a nuestros datos y modo de representar los mismos. Por tanto, el código interno y variables de prueba se quedan a modo de Caja negra ante nuestros ojos por estar fuera del temario y temática de este PFC.

4 EXPERIMENTOS.

Iniciamos este apartado recordando nuestro modelo propuesto y explicado en la sección 3.5 separándolo en dos clases, **Bloques Constantes** y **Bloques Variables, como veremos a continuación.** Vamos a iniciar una serie de experimentos para intentar obtener los resultados mejores que podemos conseguir de nuestro RAH.

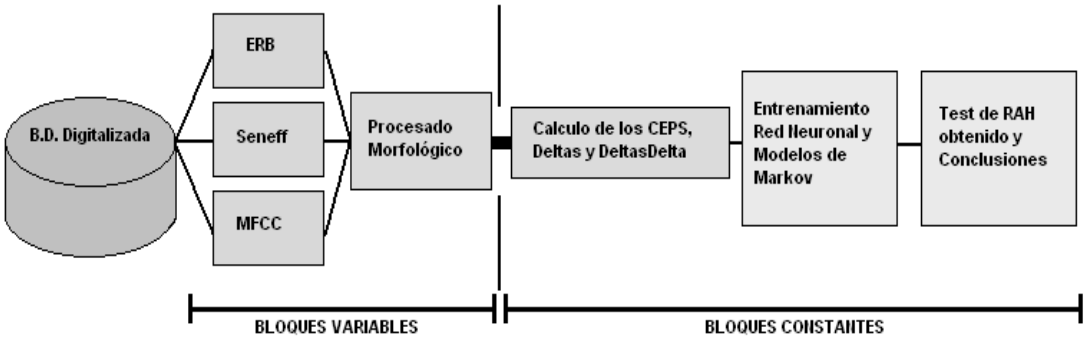


Ilustración 31 RAH propuesto reagrupado para los Experimentos del PFC.

Al centrarnos en el tema que nos ocupa en nuestro PFC, hay bloques del sistema (llamados **Bloques Constantes** en la Ilustración 31 RAH propuesto reagrupado para los Experimentos del PFC.) que no se tocarán ni variarán una vez programados y adaptados éstos a nuestras necesidades; mientras que, habrá otros (**Bloques Variables,** de nuevo reflejados en la Ilustración 31) cuyas variaciones nos permitan evaluar qué modificaciones resultan más ventajosas desde el punto de vista de tasa de reconocimiento. El centrarnos más en algunos bloques a la hora de estudiar nuestro modelo propuesto, es centrarnos más en la temática que abarca nuestro PFC, siendo el estudio exhaustivo de los llamados "Bloques Constantes" de nuestro Modelo Propuesto nuevas líneas de estudio para futuros PFC, apartado que; será tratado más adelante.

Bloques Constantes

Estos bloques de valores constantes a lo largo del PFC serán mayoritarios, siendo éstos:

Entrenamiento Red Neuronal y Test de RAH: Como se citó en la sección Error: Reference source not found, estos bloques una vez programados y adaptados, serán cajas negras para nuestro PFC.

Calculo de CEPS, Deltas y Deltadeltas: (Véase sección 3.6) Como numerosos estudios demuestran, el cálculo de los Deltas, y Deltadeltas, sobre las señales en dominio frecuencial (CEPS) es una mejora más que notable a la hora de discriminar señales de audio. El bloque de Cálculo de Deltas y Deltadeltas siempre mejorará los resultados, ya que siempre

añade nueva información a los CEPS iniciales. Esta información será muy útil a la hora de discriminar y diferenciar los distintos fonemas, que es el objetivo principal que concierne a nuestro PFC. Por tanto, una vez demostrado que nuestros SCRIPTS realizan correctamente el cálculo de nuestros Deltas y Deltadeltas, dicho bloque será fijo y constante para todos nuestros experimentos. Para ello, se realizan un total de **3 SCRIPTS de control** con los que quedará visualizado y demostrado que dicho proceso se hace correctamente, estos SCRIPTS se validan con unos datos de referencia a los que denominaremos **Linux** y que han sido extraídos utilizando la herramienta SPRACHCORE **[23]** y cada uno de ellos demuestra respectivamente lo siguiente:

1- **Cada Ceps, Delta o Deltadelta individualmente en el tiempo:** Gráficas que muestran la evolución en el tiempo de un determinado CEPS, Delta o Deltadelta de una señal determinada. El "eje x" representa el tiempo desde el inicio hasta el final de la señal, y el "eje y" representa el valor coeficiente elegido.

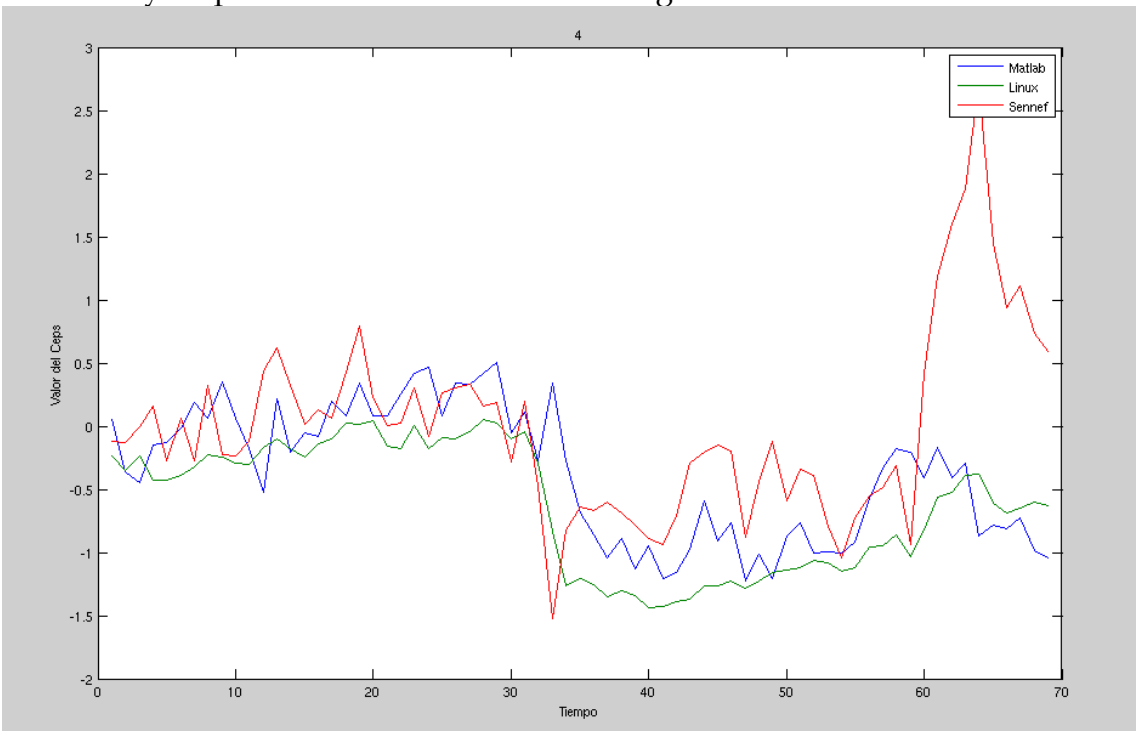


Ilustración 32 Ceps 4 de las extracciones Seneff, MFCC.

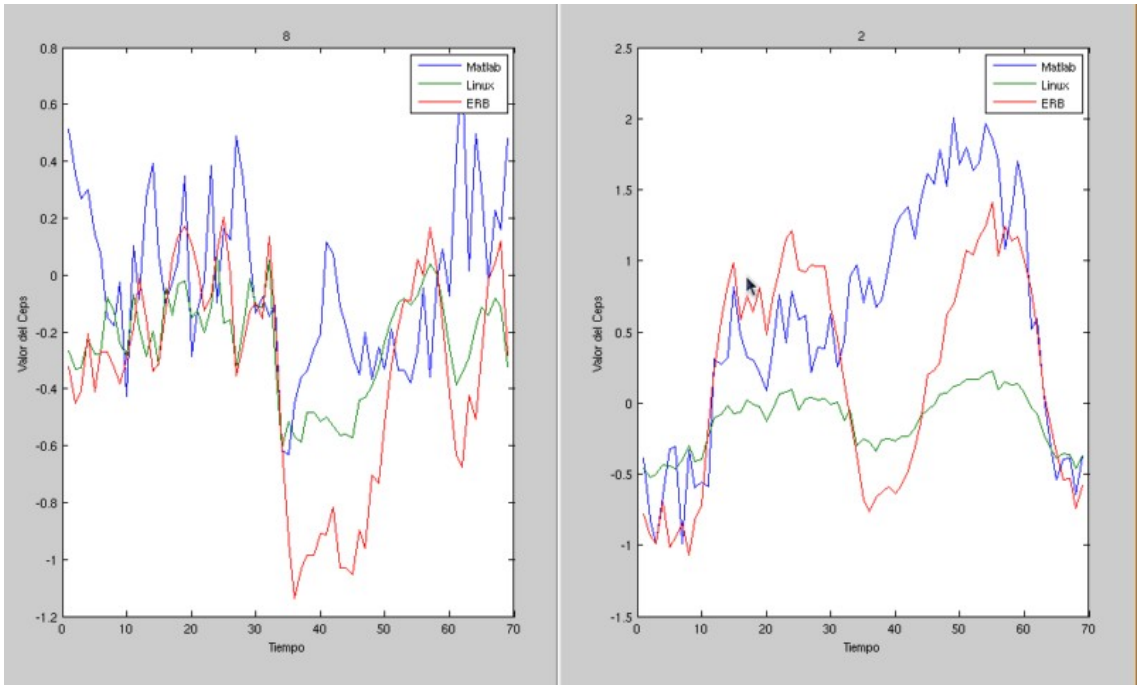


Ilustración 33 Ceps 8 y 2 de las extracciones ERB y MFCC.

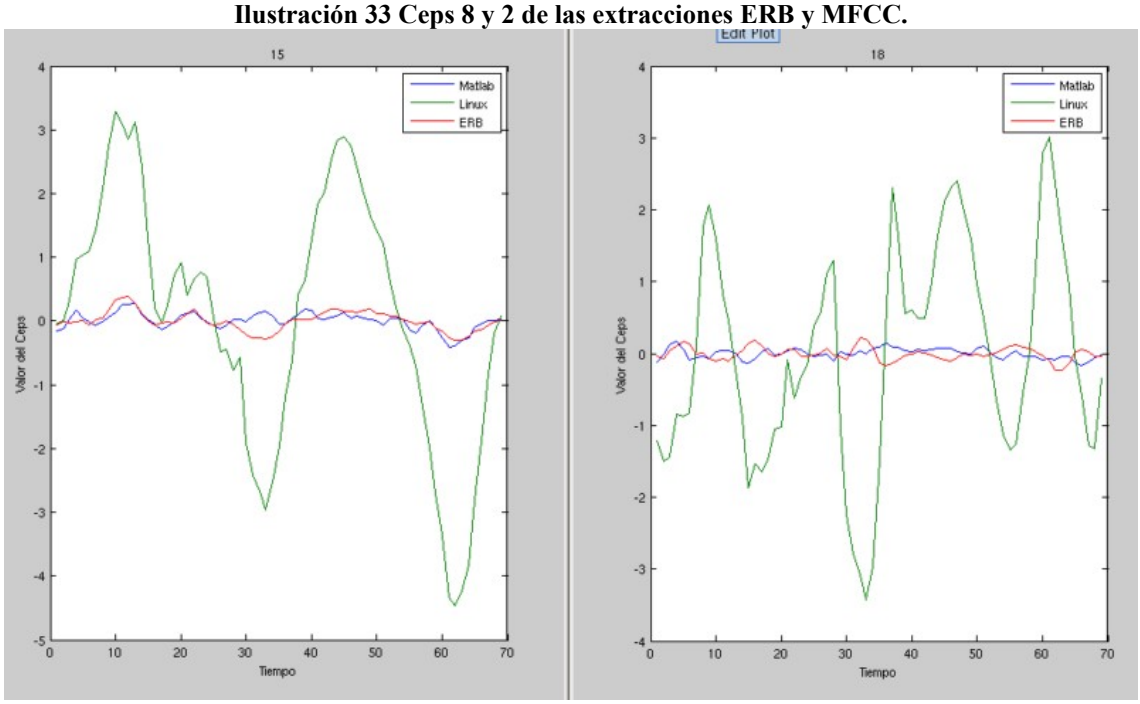


Ilustración 34 Deltas 2 y 5 de las extracciones ERB, MFCC.

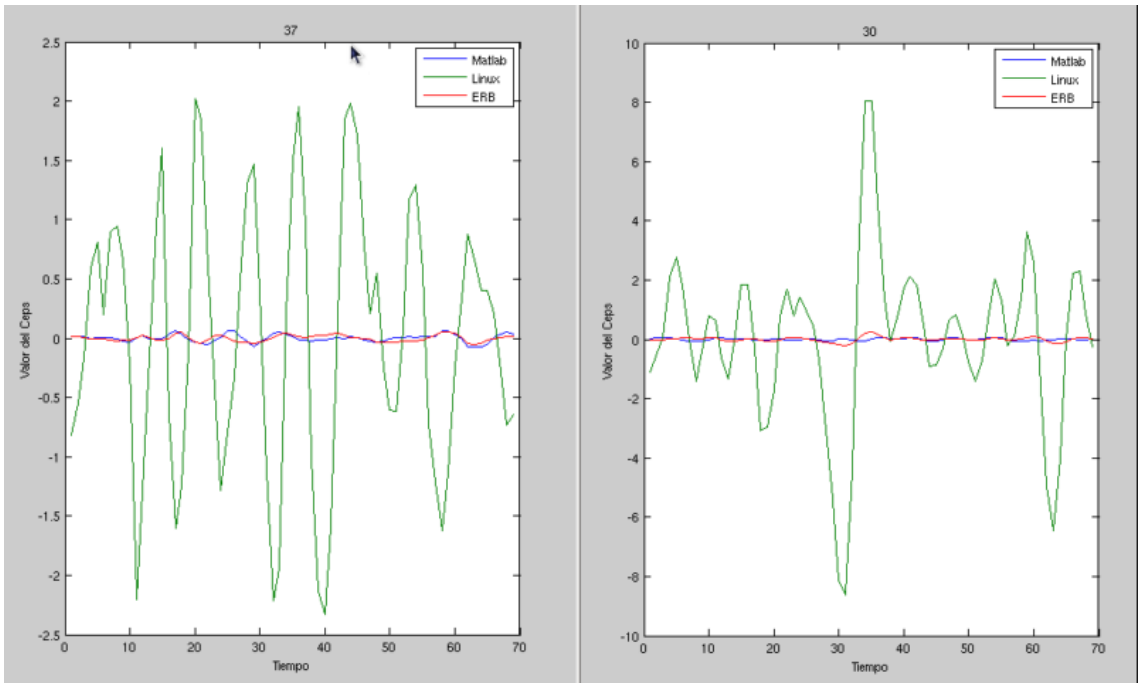


Ilustración 35 Deltadeltas 4 y 11 de las extracciones ERB y MFCC.

2- **<u>Todos los Ceps, Deltas y Deltadeltas en cada instante temporal:</u>** Gráficas que muestran el valor de todos los CEPS, Deltas y Deltadeltas en un instante temporal determinado. El "eje x" representa la 39 muestras (13 CEPS + 13 Deltas + 13 DeltaDeltas), y el "eje y" el valor de las mismas.

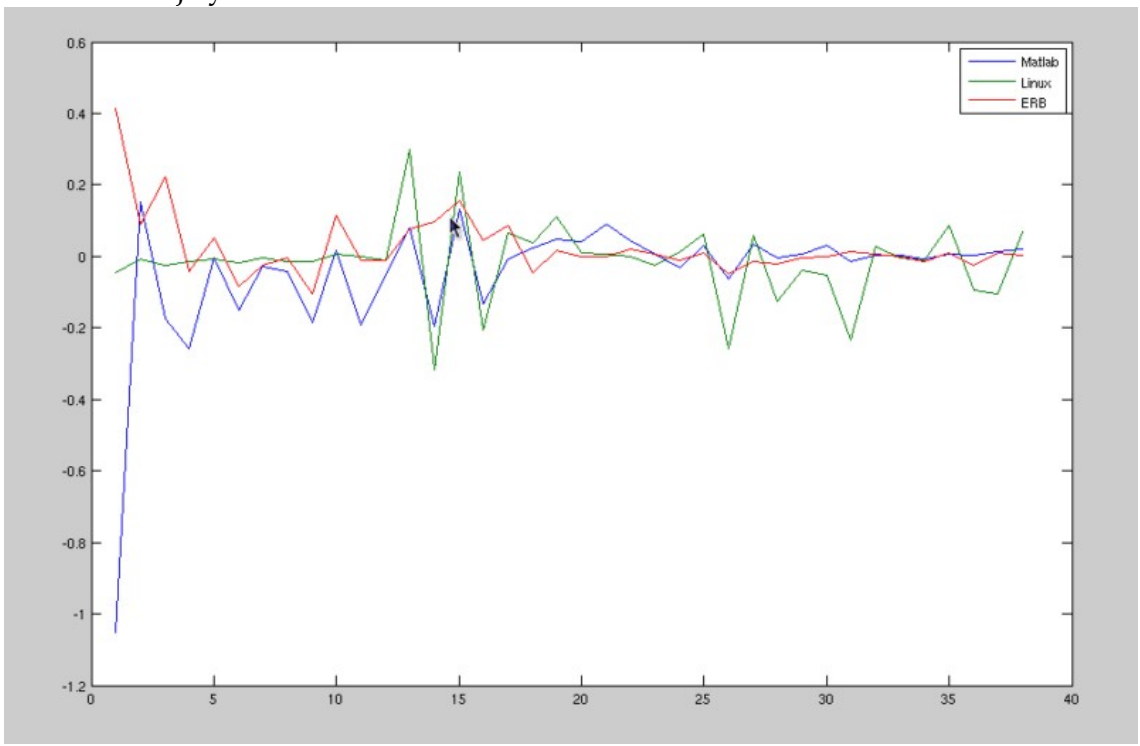


Ilustración 36 Muestra en instante temporal de las extracciones MFCC y ERB.

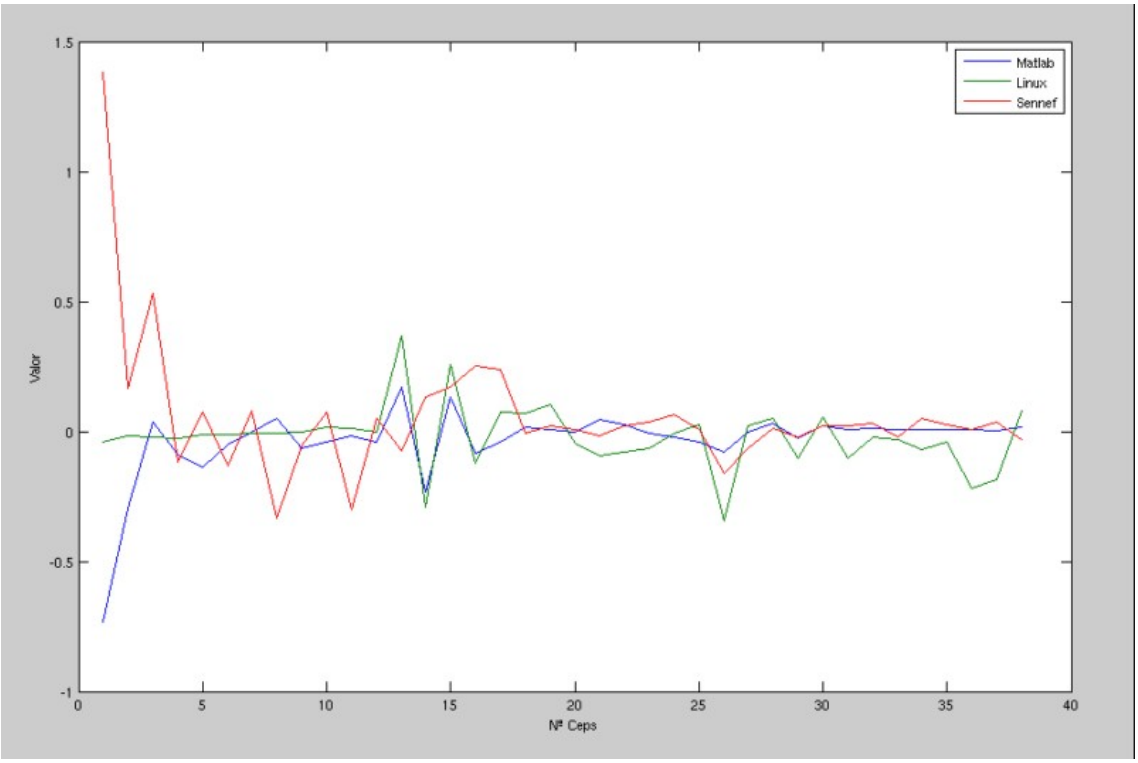


Ilustración 37 Muestra en instante temporal de extracción MFCC y Seneff.

3- **<u>Correcta lectura y trato de las señales de audio de nuestra B.D:</u>** Gráficas que muestran la carga y lectura de las señales de voz de nuestra B.D a nuestros SCRIPTS Matlab. Proceso clave en la unión de plataformas Linux con plataformas Matlab. El "eje x" representa el número de muestras en el tiempo que se tomaron y el "eje y" su valor.

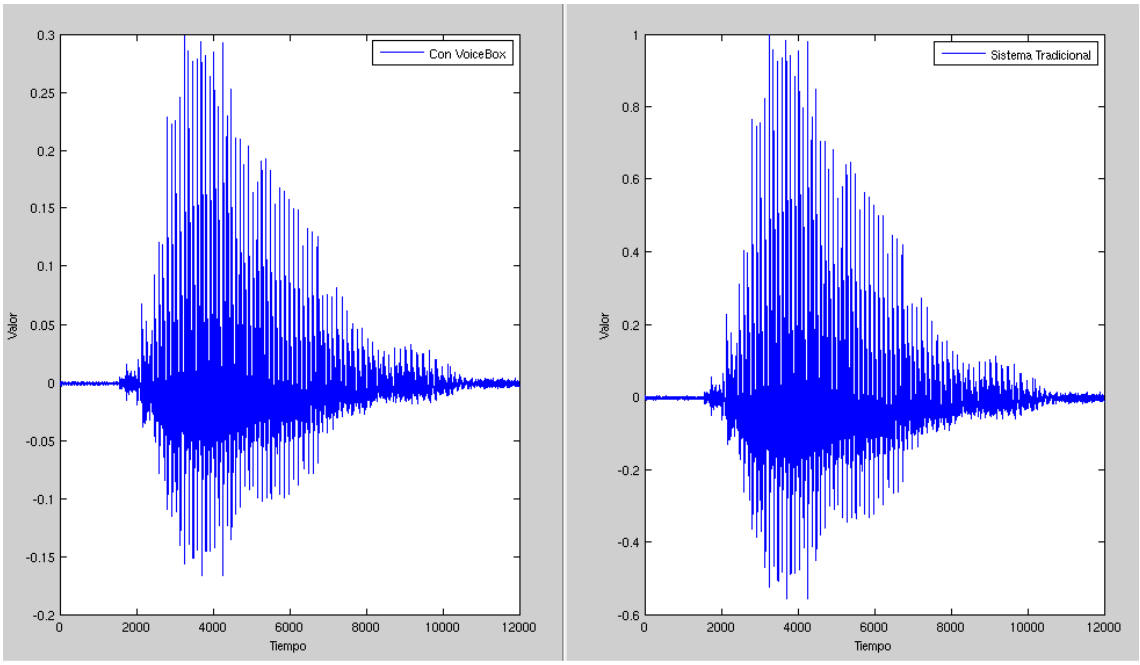


Ilustración 38 Muestra voz leída en crudo y leída con SCRIPTS del PFC.

Bloques Variables

Extracción características ERB, Seneff y MFCC: Como se pudo ver en la sección 2.2, el modelo MFCC de extracción de características está consolidado en cuanto a RAH se refiere y para su funcionamiento hemos elegido los parámetros estándar de configuración.

En cuanto a los dos restantes, aún es necesario elegir sus parámetros y variables para que los datos de salida sean lo más cercano a lo esperado. Además de esto, tenemos la dificultad de que este bloque será el responsable directo a la hora de unir correctamente los SCRIPTS realizados en Matlab con la parte de los SCRIPTS realizados en Linux (Véase la sección 3.6). Es el encargado de hacer que los bloques Linux reciban el número exacto de tramas en el tiempo que esperan por cada señal correspondiente. Y como dificultad, tenemos que el bloque de extracción de características tiende a calcular diferente número de tramas temporales en función de sus variables internas y longitudes de señales, siendo éstas a su vez independientes de los bloques que las preceden. Pues bien, nosotros hemos tenido que añadir también una **adaptación de eliminación y/o adición de tramas para cada señal** en función de lo que se esperan los bloques Linux relacionados con el Entrenamiento de la Red. Así pues, el bloque de extracción de características; aparte de sus variables internas como tal, llega a interactuar con los bloques Linux a modo que sus extracciones en el tiempo para cada señal concuerdan con lo que los bloques Linux esperan (Véase la sección 2.3) sobre las redes neuronales). Por tanto, es un bloque que depende e interactúa de una forma u otra, con los bloques Linux posteriores.

Esto es necesario debido a que las etiquetas fonéticas de cada una de las tramas, obtenidas mediante alineamiento forzado fuera de este PFC y con las que se entrenan las redes neuronales corresponden a unos instantes temporales que deben ser estrictamente respetados para evitar problemas de desalineamiento entre las nuevas parametrizaciones obtenidas y las etiquetas. Esta adaptación adicional; debida a su importancia, es controlada por un **SCRIPT de control o verificación**, el cual muestra con sus gráficas los valores temporales esperados por los bloques Linux para cada señal, y los valores temporales de salida del bloque de "Extracción de Características" para las mismas. Todo lo anterior se muestra con las Carpetas de 1560 señales explicadas en la sección Error: Reference source not found para mejor manejo y tratamiento de la información. Se muestra a continuación las gráficas que utiliza el citado SCRIPT a la hora de controlar los posibles problemas de desalineamiento. En el "eje x" están las 1560 muestras de la carpeta, y en el "eje y" el valor de tramas que espera cada una:

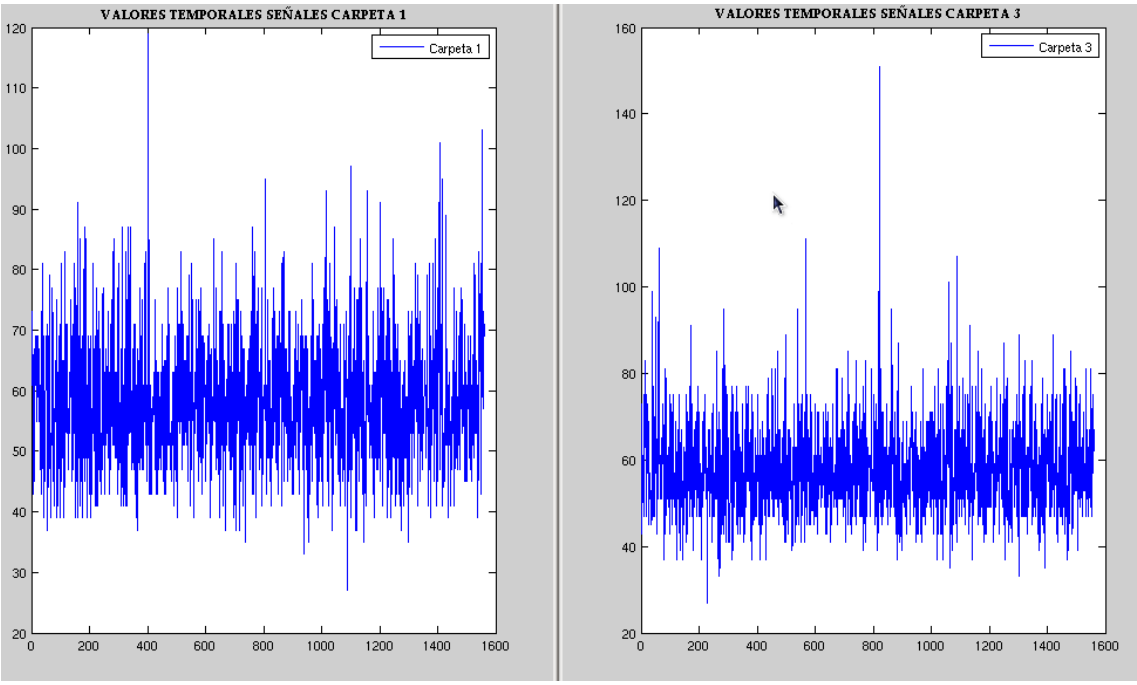


Ilustración 39 Valores Temporales de las señales de las carpetas 1 y 3.

Por último hay que explicar que dicha **adaptación de eliminación/adición temporal** se realiza teniendo en cuenta que tanto **la eliminación y adición de tramas** debe hacerse siempre en los periodos de silencio de la señal (primeras y últimas tramas temporales).

Procesado Morfológico:

Como se dijo anteriormente en la sección 3.5 el SCRIPT de Procesado Morfológico utiliza un elemento estructural o máscara para simular el Enmascaramiento Temporal y su efecto sobre la estructura de Bandas Críticas en el oído. Como era de esperar, dicha Máscara tiene que ser elegida por nosotros, gran cantidad de máscaras pueden ser teóricamente válidas, pero no por ello notablemente beneficiosas. Teniendo así que intentar encontrar la máscara válida que ofrezca los mejores resultados a nuestro RAH propuesto. En nuestro PFC; y siendo lo siguiente una característica que hace que dicho bloque de procesado sea del tipo "Variable", probaremos un total de **3 elementos estructurales o máscaras** válidas diferentes; con el fin de compararlas y quedarnos finalmente con la que ofrezca unos resultados mejores. Así pues, las máscaras elegidas y utilizadas serán las siguientes:

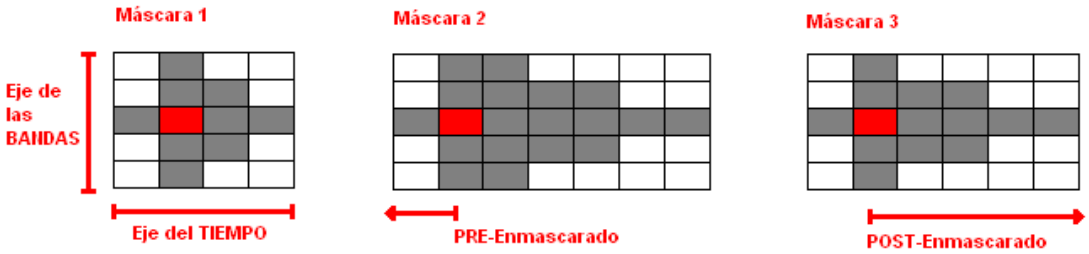


Ilustración 40 Elementos estructurales o máscaras propuestos.

Indicar para finalizar que, no podemos asegurar que alguna de nuestras 3 máscaras elegidas sea la optima de todas las máscaras válidas posibles. Abriendo así otra línea futura de estudio para un posible PFC que estudie en profundidad este bloque de Procesado Morfológico, y que llegue a encontrar la máscara que mejore nuestro RAH notablemente. En especial, queremos llamar la atención sobre el hecho de que estas máscaras son binarias y que fácilmente pueden elegirse otras con valores continuos que, por una parte, añadirían una mayor flexibilidad para modelar las características auditivas pero que, por otra, multiplican el espacio de posibilidades de exploración empírica notablemente.

4.1 RAH SIN PROCESADO MORFOLOGICO.

Los resultados obtenidos en nuestro PFC serán representados en tablas que muestran las **Tasas de Error en** % conseguidos por nuestro RAH en el reconocimiento. Dichas tasas están separadas según el tipo de tests realizado (Clean-Clean, Clean-Noisy y Noisy-Noisy (Véase capítulo 3.5)), y según el método de extracción de características utilizado (MFCC, ERB y Seneff (Véase capítulo 2.2)). Los datos se muestran siempre en grupos de cinco para que cada carpeta de 1560 muestras tenga su representación individual (Véase capítulo 3.6). Los datos coloreados en verde representan el mejor resultado obtenido comparando entre tipos de extracción.

TASA ERROR %	SIN PROCESADO		
CLEAN-CLEAN	4.2	3.5	7.9
	4.4	3.2	7.1
	5.1	3.9	6.6
	4.3	3.9	7.1
	5.2	5.4	9.7
CLEAN-NOISY	45.7	56.3	49.1
	46.6	55.1	48.4
	48.6	57.6	48.9
	46.7	55.9	48.4
	45.6	51.5	49.9
NOISY-NOISY	18.2	17.9	23.7
	18.4	17.7	22.9
	19.5	19.7	25.6
	16.8	19.2	20.8
	17.5	16.3	23.3
	MFCC	ERB	Seneff

Tabla 2 Tabla resultados de Tasa Error en % sin Procesado Morfológico.

Como se puede apreciar en la Tabla 2 que muestra los resultados **sin el Procesado Morfológico**, la extracción que obtiene mejores resultados en nuestro RAH es la extracción **MFCC**, la cual está seguida muy de cerca por la extracción **ERB**. Siendo ERB una extracción notablemente buena en el caso de data Clean-Clean y MFCC en la de data Clean-Noisy, siendo ambas casi similares en caso data Noisy-Noisy. Aún estando empatadas desde su lado numérico, y dando al caso más real de test una importancia mayor (data Clean-Noisy, véase la sección 3.6), elegiremos como extracción más adecuada para el caso de Sin Procesado Morfológico la extracción MFCC. Nos llama la atención, por otra parte, que la extracción Seneff, en general peor que las otras dos en los

casos clean-clean y noisy-noisy, presente mejores prestaciones que la ERB en el caso clean-noisy situándose en segundo lugar con resultados intermedios.

Veamos pues, en la siguiente tabla la evolución de la Tasa de Error para los diferentes tipos de test, en cada una de las carpetas, para las diferentes extracciones:

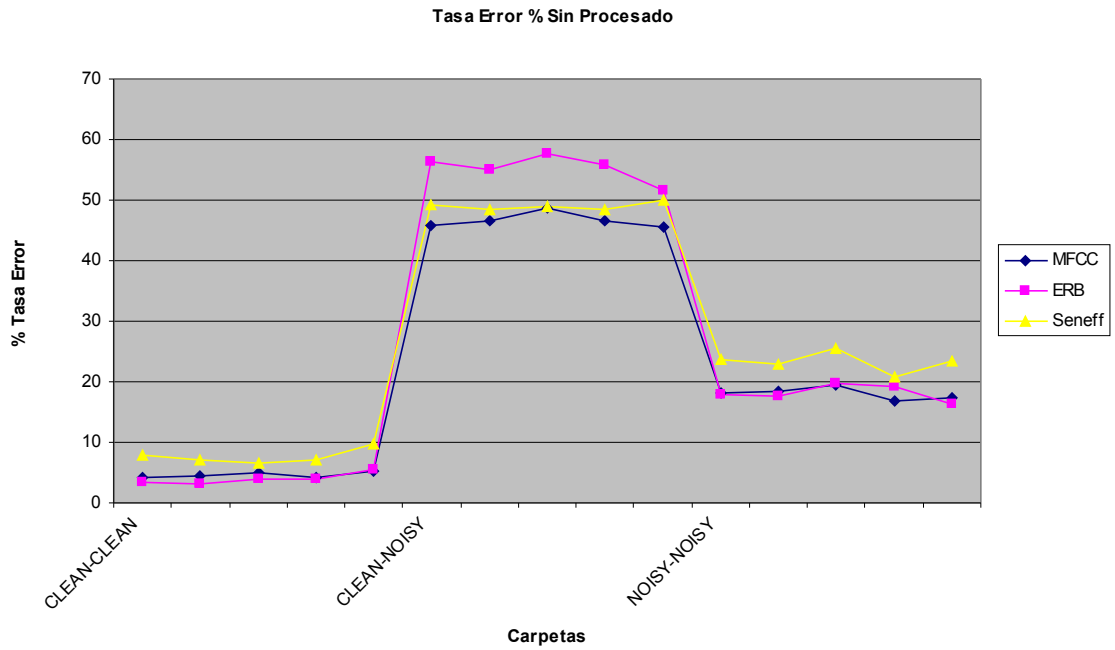


Ilustración 41 Evolución Tasa Error en % sin Procesado Morfológico.

4.2 RAH CON PROCESADO MORFOLOGICO MÁSCARA1.

TASA ERROR %	MASCARA1		
CLEAN-CLEAN	5.1	3.4	10.5
	4.6	3.9	10.8
	4.4	4.4	9.5
	4.2	4.2	10.2
	5.5	6.2	11.5
CLEAN-NOISY	41.1	48.1	44.2
	40.8	49.4	43.9
	42.8	50.6	42.2
	41.7	53.5	42.3
	40.3	48.8	42.8
NOISY-NOISY	19.7	18.8	27.6
	19.6	19.7	24.7
	21.3	20.3	27.9
	19.8	18.1	24.0
	20.0	19.9	26.9
	MFCC	ERB	Seneff

Tabla 3 Tabla resultados de Tasa Error en % con Procesado Morfológico y Máscara1.

Como se puede apreciar en la tabla que muestra los resultados **con Procesado Morfológico Máscara1**, de nuevo ERB y MFCC obtienen notablemente mejores prestaciones que Seneff. Siendo ERB una extracción notablemente buena en el caso de data Clean-Clean y data Noisy-Noisy; mientras que MFCC gana en la de data Clean-Noisy de nuevo. Si bien es cierto que MFCC gana en la extracción de un canal con ruido desconocido (data Clean-Noisy), la extracción ERB ha tenido esta vez muy buenos resultados en los otros dos tipos de testeo. Por ello; no se puede afirmar con rotundidad cual es la extracción más adecuada en este caso.

En todo caso, podemos afirmar que la Máscara1 tiene efectos positivos en el data Clean-Noisy, y negativos en el caso de data Clean-Clean y Noisy-Noisy; teniendo por tanto su parte positiva y negativa. Cabe destacar como cambio significativo que la extracción Seneff mejora notablemente en la extracción Clean-Noisy con la Máscara1 llegando casi a empatar en resultados a la ganadora.

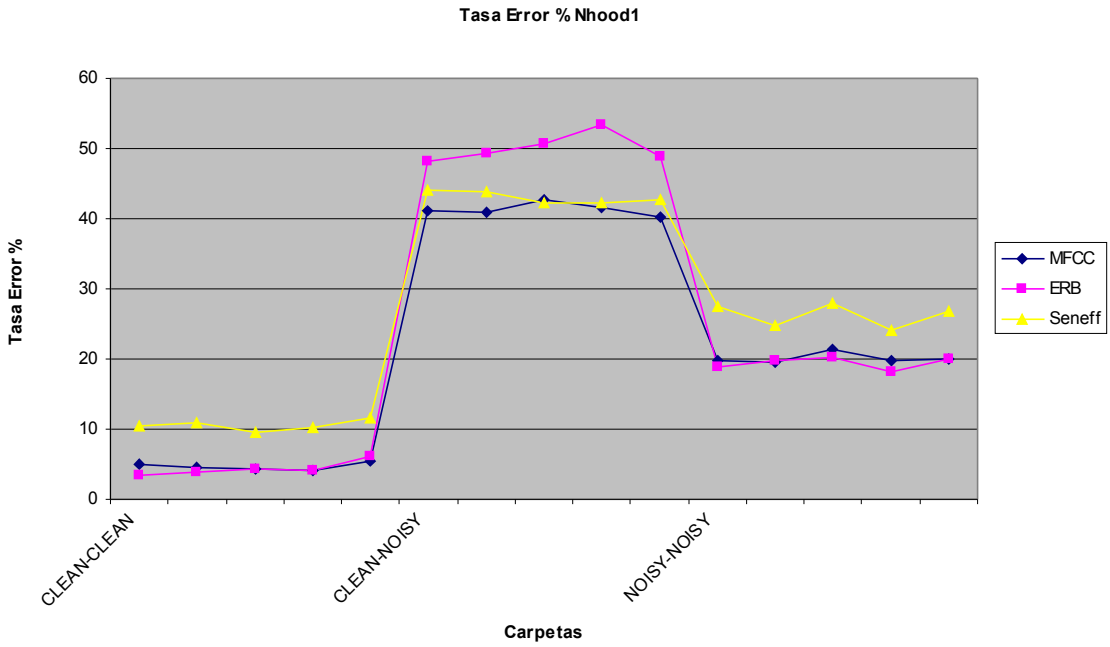


Ilustración 42 Evolución Tasa Error en % con Procesado Morfológico y Máscara1.

4.3 PROCESADO MORFOLOGICO CON MÁSCARA2.

TASA ERROR %	MASCARA2		
CLEAN-CLEAN	4.7	5.5	9.2
	5.1	4.6	10.3
	4.6	4.2	9.2
	4.4	4.2	8.5
	6.0	6.4	11.1
CLEAN-NOISY	42.6	47.6	44.6
	42.2	49.8	45.8
	43.0	50.6	43.5
	41.6	54.0	42.5
	41.0	49.7	44.0
NOISY-NOISY	20.4	20.9	28.1
	20.3	20.0	26.8
	22.6	23.0	27.8
	20.2	19.7	24.9
	19.7	19.6	28.1
	MFCC	ERB	Seneff

Tabla 4 Tabla resultados de Tasa Error en % con Procesado Morfológico y Máscara2.

Como se puede apreciar en la tabla que muestra los resultados **con Procesado Morfológico Máscara2**, la extracción que obtiene mejores resultados en nuestro RAH es la extracción **MFCC**, la cual está seguida muy de cerca por la extracción **ERB** de nuevo. Siendo ERB una extracción notablemente buena en el caso de data Clean-Clean y data Noisy-Noisy; mientras que MFCC gana en la data Clean-Noisy de nuevo. Si bien es cierto que MFCC gana en la extracción de un canal con ruido desconocido (data Clean-Noisy), la extracción ERB ha tenido esta vez muy buenos resultados en los otros dos tipos de test. Por ello; no se puede afirmar con rotundidad cual es la extracción más adecuada en este caso. Quedando por tanto la extracción ERB y MFCC empatadas de nuevo para la configuración de Procesado Morfológico Máscara2.

Destacar de nuevo, que la extracción Sennef sin ser la que mejores resultados da en general, para el caso de testeo más importante (data Clean-Noisy) vuelve a mejorar notablemente gracias a esta Máscara2, llegando casi a empatar a la extracción ganadora. Mas adelante, compararemos todos y cada uno de los resultados obtenidos con cada Máscara entre ellos.

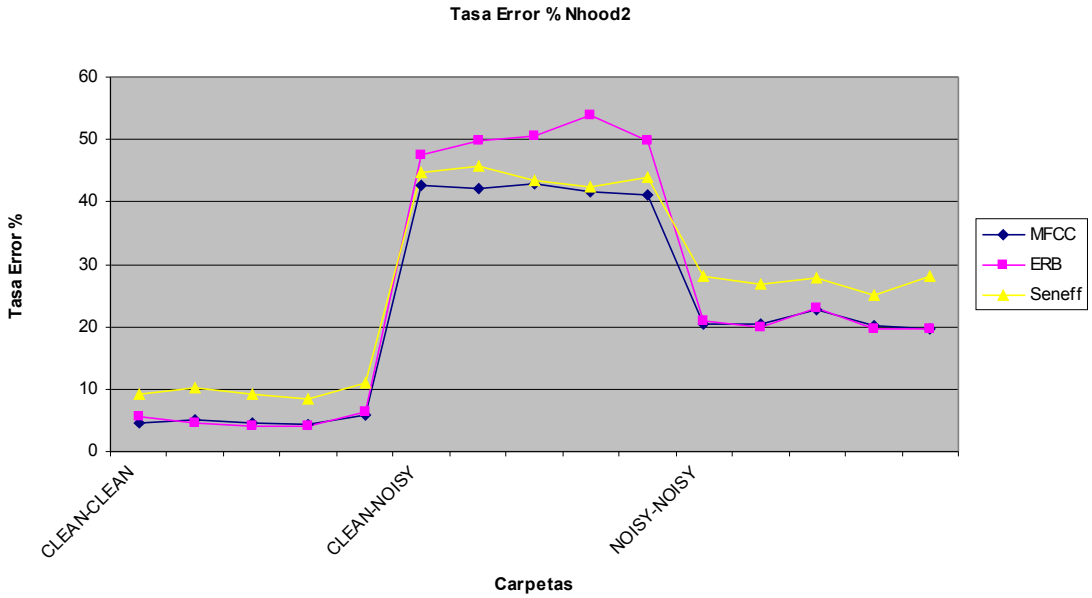


Ilustración 43 Evolución Tasa Error en % con Procesado Morfológico y Máscara2.

4.4 PROCESADO MORFOLOGICO CON MÁSCARA3.

TASA ERROR %	MASCARA3		
CLEAN-CLEAN	4.6	6.0	8.9
	5.0	4.9	9.2
	4.6	4.3	9.8
	4.3	5.1	7.4
	5.6	6.3	8.8
CLEAN-NOISY	41.0	50.4	40.7
	41.5	49.9	41.4
	43.6	51.5	41.0
	41.7	56.1	39.7
	40.9	51.2	41.8
NOISY-NOISY	20.0	19.9	27.4
	21.1	21.2	23.9
	21.3	21.2	27.6
	20.1	19.2	21.9
	18.9	19.0	25.8
	MFCC	ERB	Seneff

Tabla 5 Tabla resultados de Tasa Error en % con Procesado Morfológico y Máscara3.

Observando la tabla que muestra los resultados **con Procesado Morfológico Máscara3**, sorprendentemente vemos un empate en caso data Clean-Clean y data Noisy-Noisy entre nuestras dos anteriores mejores extracciones, **MFCC** y **ERB**, mientras que, en la extracción de un canal con ruido desconocido (data Clean-Noisy)**,** la parametrización que obtiene mejores resultados en nuestro RAH es la **Seneff.** Y no sólo eso, sino que ha conseguido los mejores resultados en data Clean-Noisy obtenidos a lo largo de nuestro estudio con todas nuestras diferentes configuraciones del RAH. Por ello, Seneff queda como la extracción más adecuada en el caso de Máscara3, y abre una nueva línea de estudio en la que se puede pensar en conseguir mejores resultados en el test data Clean-Noisy, gracias a otras Máscaras, y a la extracción que mejora siempre con todas ellas, Seneff. Por otra parte, tenemos que advertir que, este procesado tiene consecuencias perniciosas especialmente graves en el caso de Clean-Clean.

Ilustración 44 Evolución Tasa Error en % con Procesado Morfológico y Máscara3.

5 CONCLUSIONES Y LÍNEAS FUTURAS.

5.1 CONCLUSIONES.

Como se explicó y se vio en el capítulo 3.6, las extracciones que mejores resultados cualitativos en general nos dieron en términos globales, para nuestro RAH es sus diferentes configuraciones fueron:

Sin Procesado Morfológico: **MFCC.**
Con Procesado Morfológico Máscara1: **MFCC y ERB.**
Con Procesado Morfológico Máscara2: **MFCC y ERB**.
Con Procesado Morfológico Máscara3: **Seneff.**

Aún así vamos a continuar nuestro estudio representando todos los datos obtenidos en nuestros experimentos más claramente en otra gráfica. En la cual representaremos en color verde las mejoras obtenidas gracias al bloque de **Procesado Morfológico** frente a nuestro RAH sin dicho bloque. Y representaremos en verde oscuro el mejor dato obtenido gracias a las Máscaras del Procesado Morfológico, siempre y cuando éste haya sido anteriormente una mejora frente al RAH sin dicho bloque (color verde). Y además, una vez conseguidas los mejores datos obtenidos gracias a las Máscaras (verde oscuro) subrayaremos cada uno de los mejores datos en cada configuración de testeo mejorada (color verde). O explicado de otra forma:

Verde Claro = Mejora los resultados de SIN PROCESADO.
Verde Oscuro = Mejor Máscara para ese tipo de extracción.
Subrayado = Mejor resultado obtenido entre todos.

Así pues, tenemos la siguiente gráfica:

TASA ERROR %	SIN PROCESADO			CON PROCESADO MORFOLOGICO								
				Máscara1			Máscara2			Máscara3		
CLEAN-CLEAN	4.2	3.5	7.9	5.1	3.4	10.5	4.7	5.5	9.2	4.6	6.0	8.9
	4.4	3.2	7.1	4.6	3.9	10.8	5.1	4.6	10.3	5.0	4.9	9.2
	5.1	3.9	6.6	4.4	4.4	9.5	4.6	4.2	9.2	4.6	4.3	9.8
	4.3	3.9	7.1	4.2	4.2	10.2	4.4	4.2	8.5	4.3	5.1	7.4
	5.2	5.4	9.7	5.5	6.2	11.5	6.0	6.4	11.1	5.6	6.3	8.8
CLEAN-NOISY	45.7	56.3	49.1	41.1	48.1	44.2	42.6	47.6	44.6	41.0	50.4	**40.7**
	46.6	55.1	48.4	**40.8**	49.4	43.9	42.2	49.8	45.8	41.5	49.9	41.4
	48.6	57.6	48.9	42.8	50.6	42.2	43.0	50.6	43.5	43.6	51.5	**41.0**
	46.7	55.9	48.4	41.7	53.5	42.3	41.6	54.0	42.5	41.7	56.1	**39.7**
	45.6	51.5	49.9	**40.3**	48.8	42.8	41.0	49.7	44.0	40.9	51.2	41.8
NOISY-NOISY	18.2	17.9	23.7	19.7	18.8	27.6	20.4	20.9	28.1	20.0	19.9	27.4
	18.4	17.7	22.9	19.6	19.7	24.7	20.3	20.0	26.8	21.1	21.2	23.9
	19.5	19.7	25.6	21.3	20.3	27.9	22.6	23.0	27.8	21.3	21.2	27.6
	16.8	19.2	20.8	19.8	18.1	24.0	20.2	19.7	24.9	20.1	19.2	21.9
	17.5	16.3	23.3	20.0	19.9	26.9	19.7	19.6	28.1	18.9	19.0	25.8
	MFCC	ERB	Seneff	MFCC	ERB	Seneff	MFCC	ERB	Seneff	MFCC	ERB	Seneff

Tabla 6 Tabla de resultados obtenidos por nuestro RAH en diferentes configuraciones.

Como se aprecia a simple vista, la nube de mejoras siempre se mantiene dentro del testeo data Clean-Noisy, el cual, como ya se dijo en la sección 3.6 es el modelo más realista de los estudiados al simular una distorsión de canal desconocida a priori; y por tanto, le daremos una importancia mayor. Por ello, podemos afirmar que: "Nuestro Procesado Morfológico adicional mejora cualitativamente los resultados obtenidos en nuestro RAH en todas sus extracciones para el caso clean-noisy de Test". Este es el caso del test data Clean-Noisy; en el que centraremos un poco más nuestro estudio obteniendo que:

> La extracción **MFCC** es la que tiene el comportamiento más independiente a la Ventana del Procesado Morfológico utilizada, ya que sus resultados son similares en cada una de ellas. Mejores que en el caso Sin Procesado Morfológico, pero similares para cada una de las Ventanas. Además la extracción MFCC tiene los mejores resultados en el test data Clean-Noisy para las Máscaras 1 y 2, seguida muy de cerca por la extracción Seneff. Veamos la evolución de esta extracción con las diferentes máscaras utilizadas:

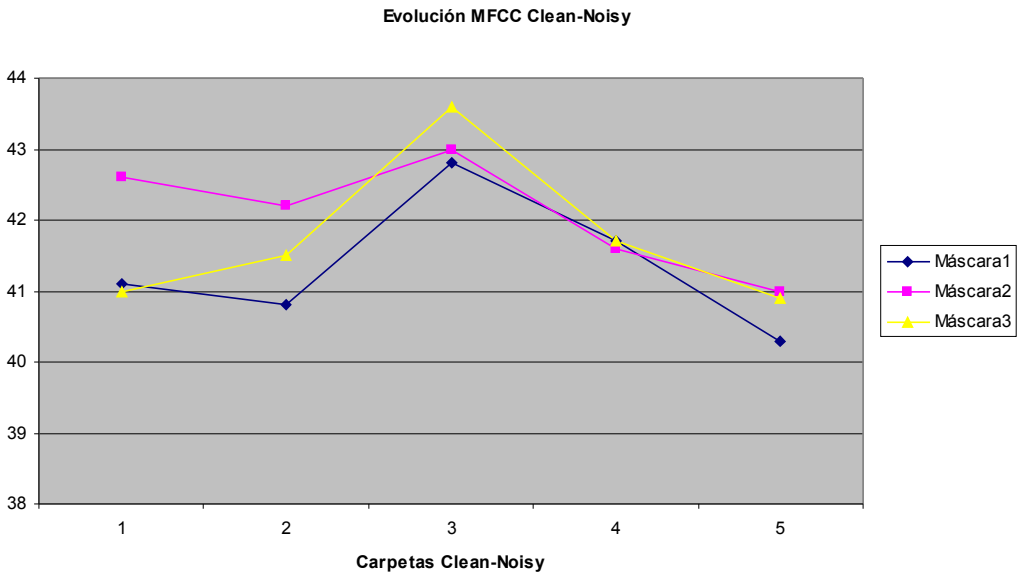


Ilustración 45 Evolución MFCC en carpetas Clean-Noisy

La extracción **ERB** tiene un comportamiento más dependiente a la Ventana de Procesado Morfológico que el MFCC. Sus mejores datos se distribuyen entre las Máscaras 1 y 2, siendo estos peores en la Máscara3. Recordemos que, en estos casos, quedó empatada en eficiencia general con el MFCC, por lo que no era de extrañar este resultado. Hay que tener en cuenta también que, como se vio anteriormente, el ERB funciona siempre peor que el MFCC y Seneff en el caso del test Clean-Noisy. Veamos también la evolución con las diferentes máscaras utilizadas:

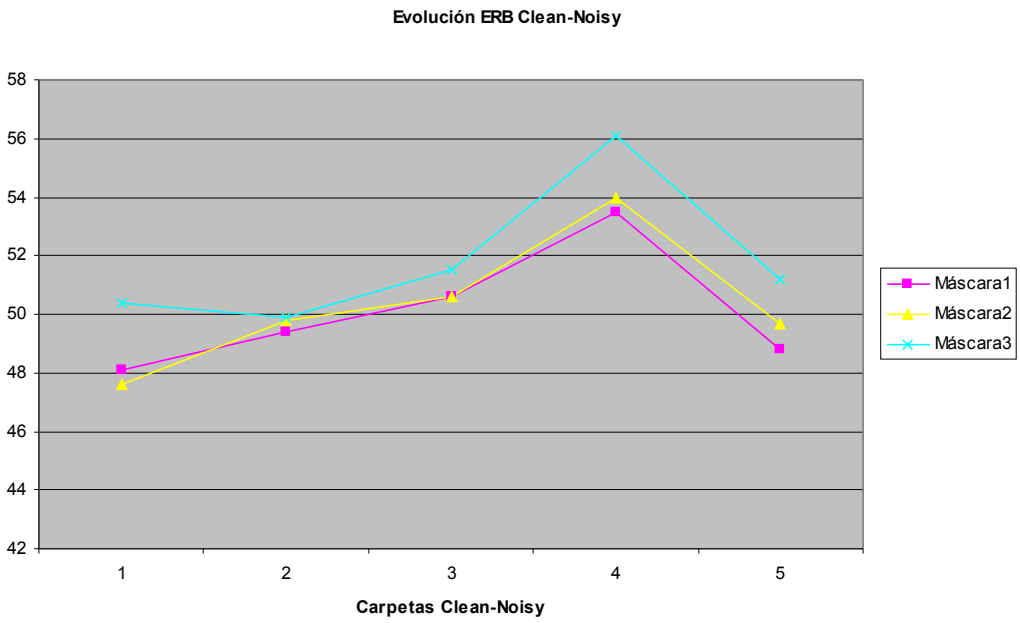


Ilustración 46 Evolución ERB en carpetas Clean-Noisy.

La extracción **Seneff** tiene el comportamiento más dependiente y variable de todos. Tiene los peores resultados en general, excepto en el Clean-Noisy. En dicho test llega casi a empatar a la extracción más eficiente en las Máscaras 1 y 2, la MFCC. Aparte de esto, se ve que el Procesamiento Morfológico mejora más agresivamente sus resultados y que, diferentes ventanas llegan a conseguir que Seneff varíe notablemente sus resultados en el test data Clean-Noisy. Llegando dicha extracción a conseguir los mejores resultados en el test Clean-Noisy obtenidos en este PFC para nuestro RAH. Éstos se consiguen con la máscara que queda por citar, la Máscara3. Veamos su evolución en este Clean-Noisy para las diferentes máscaras:

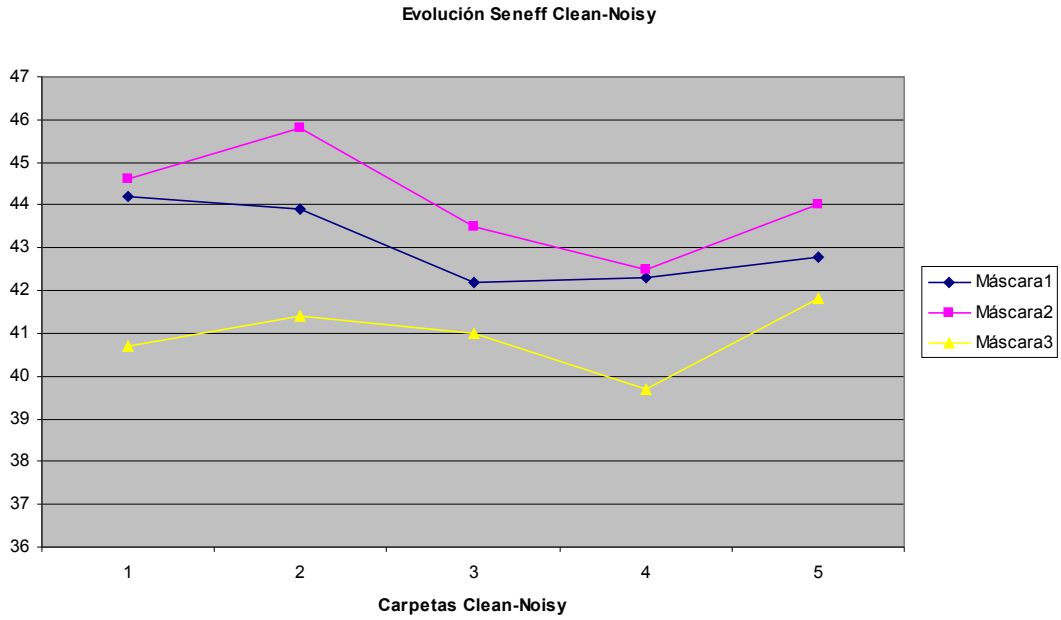


Ilustración 47 Evolución Seneff en carpetas Clean-Noisy.

Con todo lo citado y expuesto en capítulo 4 de este PFC, podemos recomendar las siguientes configuraciones de nuestro RAH como las mejores a la hora de correrlo en un canal de ruido desconocido a priori:

-RAH con extracción del tipo **MFCC** y **Procesado Morfológico en Máscara1**.
-RAH con extracción del tipo **Seneff** y **Procesado Morfológico en Máscara3**.

Todo lo anterior se ve mejor reflejado en la siguiente gráfica, en la que se muestra la Tasa de Error Promedio de cada una de las Extracciones utilizadas, en cada uno de los tipos de experimentos, para el test data Clean-Noisy:

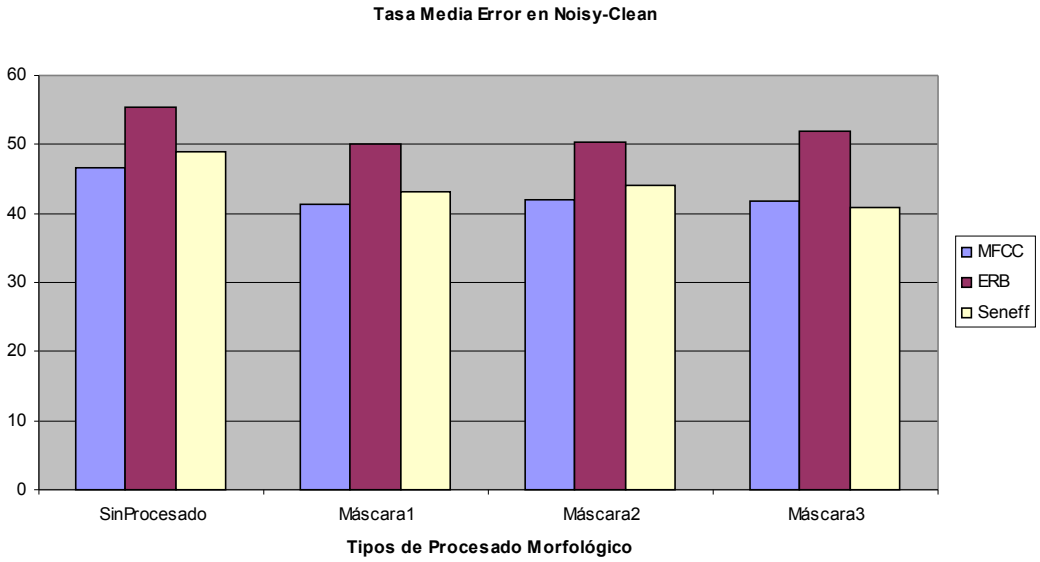


Ilustración 48 Tasa de Error Promedio en Clean-Noisy

5.2 LINEAS FUTURAS.

Como se vio en el capítulo 4 de nuestro modelo propuesto existen una serie de bloques que, una vez adaptados y modificados a nuestras necesidades; sus variables de entorno permanecerán constantes durante el periodo de experimentos y pruebas. Esta situación nos hace afirmar, que muy posiblemente; un estudio más exhaustivo a dichos bloques constantes durante el proceso de experimentos puede hacernos conseguir unos resultados más adaptados, y por tanto; mejorar nuestro RAH significativamente. Una línea de estudio futura, puede ser por tanto, la experimentación de resultados con un cálculo de Deltas y Deltadeltas con diferentes **VentanaDelta** (en todo nuestro PFC VentanaDelta=2), y la experimentación de resultados con diferentes números de Deltas y Deltadeltas calculados en cada extracción (en todo nuestro PFC son 13 ambos –véase 3.6).

Otra línea a tratar en un futuro sería el estudio exhaustivo de las máscaras del bloque de nuestro Procesado Morfológico (sección 3.4 y capítulo Error: Reference source not found), de forma que la batería de resultados del futuro PFC no sea únicamente con las 3 mascaras elegidas en nuestro PFC. Así pues, esa línea futura puede llegar a conseguir una ventana válida de procesado cuyos resultados sean notablemente mejores que los obtenidos por las 3 nuestras. Como se vio en el capítulo 4 la elección de una u otra ventana es determinante a la hora de elegir la extracción que mejor resultados da, y a su vez, es determinante a la hora de mejorar éstos mismos.

Otra línea de investigación que mejoraría notablemente el realismo y fidelidad al oído de nuestro RAH sería el desarrollo de un bloque dentro de nuestro Procesado Morfológico que simulase las **Curvas Isofónicas del Oído Humano** utilizando máscaras dependientes de la intensidad y frecuencia de la señal donde trabajan(sección 3.2 e Ilustración 19). Si es cierto que la extracción Seneff simula estas curvas de un modo tosco, una mejora externa a dicha simulación acercaría más todavía la extracción al comportamiento del Oído Humano. Queda abierta ahí por tanto una línea de estudio notablemente laboriosa; que es el añadir a nuestro sistema RAH un bloque nuevo que trabaje esas Curvas Isofónicas y simule los enmascaramientos que se producen intrínsecamente por las mismas. Así como estudiar los resultados del RAH que se obtienen por ello.

6 PRESUPUESTO DEL PROYECTO.

COSTES DIRECTOS PERSONAL

Apellidos y nombre	N.I.F. (no rellenar - solo a titulo informativo)	Categoría	Dedicación (hombres mes) [a)]	Coste hombre mes	Coste (Euro)
Turiel Merino, Jesús	52979488-F	Ingeniero Junior	131,25	1.200,00	6.000,00
Peláez Moreno, Carmen	43642368-E	Dirección	21,875	3.900,00	3.250,00
				Total	9.250,00

COSTES DIRECTOS EQUIPO

Descripción	Coste (Euro)	% Uso dedicado proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable [d)]
Ordenador doble Núcleo	650,00	100	5	60	54,17
Licencia Matlab 2009b	4.000,00	100	5	60	333,33
Licencia SPRACHCORE	0,00	100	5	60	0,00
Licencia B.D ISOLET	100,00	100	5	60	8,33
				Total	395,83

[d)] Fórmula de cálculo de la Amortización:

$$\frac{A}{B} \text{xCxD}$$

A = nº de meses desde la fecha de facturación en que el equipo es utilizado
B = periodo de depreciación (60 meses)
C = coste del equipo (sin IVA)
D = % del uso que se dedica al proyecto (habitualmente 100%)

Tabla 7 Costes Directos Presupuesto.

Presupuesto Costes Totales	Presupuesto Costes Totales
Personal	9.250,00
Amortización	396
Subcontratación de tareas	0
Costes de funcionamiento	0
Costes Indirectos (20% del Total)	1.929,2
IVA (18% del Total)	2083,5
Total	**13.659**

Tabla 8 Resumen Costes Totales.

7 REFERENCIAS.

[1] Navarro Mesa J.L., ***"Procesador Acústico: El bloque de extracción de características".*** Universidad de Las Palmas de Gran Canaria, Departamento de Canales y comunicación.

[2] Rufiner H.L., Milone D., ***"Sistema de reconocimiento automático del habla".*** Ciencia Docencia y Tecnología. Concepción del Uruguay, v.15, n.28, p.149 - 178, 2004.

[3] Vicente Peña J., ***"Contribuciones al Reconocimiento Robusto de Habla".*** Universisdad Carlos III de Madrid, Departamento de Teoría de la Señal y Comunicaciones, Leganés 2007.

[4] Web **http://es.wikipedia.org/wiki/Reconocimiento_del_habla** (15 abr 2011, a las 21:40).

[5] Curso de Acústica creado por GA **http://www.ehu.es/acustica/espanol/fisiologia1/siaues/siaues.html**

[6] Martínez Bernardo de Quirós C., ***"Fundamentos básicos del Reconocimiento de Voz"***. Tutorial de Complementos de Sonido y Audiofrecuencia, www.adictosaltrabajo.com, 2005.

[7] Rabiner L., Juang B., ***"An introduction to hidden Markov models"***. ASSP Magazine, IEEE Publicación, v.3, issue 1, part 1, p.4-16, 1986.

[8] Molano Martín D., ***"Introducción al manejo de las herramientas pfile en el ámbito del reconocimiento de voz".*** Universidad Carlos III de Madrid, Departamento de Teoría de la Señal y Comunicaciones, Estudio Tecnológico, 2006

[9] Misra H., ***"Multi-stream Processing for Noise Robust Speech Recognition".*** Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland 2006.

[10] Miriam Cordero Ruiz, **"Seneff's Auditory Model".** SONY Advanced Technology Center Stuttgart, 2002.

[11] ERB filters compilation: "**Schofield**", 1985; "**Patterson and Moore**", 1986; "**Patterson, Holdsworth, Nimmo-Smith and Rice**", 1988. "**Glasberg and Moore**", 1990.

[12] **"Sistema Auditivo Humano, Tratamiento Sonoro y capacidades",** Francisco J. García Castillo.

[13] **"Audio Box"** Joanne Bengert, Allen Upward, 2009.Web: **www.lpi.tel.uva.es/~nacho/docencia/ing_ond_1/trabajos_06_07/io1/public_html/enascaramiento.htm.**

[14] Web: **http://www.labc.usb.ve/paginas/EC4514/AUDIO/PSICOACUSTICA/Enmascaramiento_sonoro.html, "Enmascaramiento Sonoro",** Sin Autor**.**

[15] **http://cmusphinx.sourceforge.net/sphinx4. "A speech recognizer written entirely in the Java™ programming language"** 1999-2008 Carnegie Mellon University.

[16] "**Tratamiento digital de señales. Prentice Hall".** Y **"PROAKIS",** John G. y MANOLAKIS, Dimitris G (1997).

[17] Web **http://www.bme.ogi.edu/~hynek/, "Robust automatic recognition of speech",** Hynek Hermansky.

[18] **"Estudio de Acústica Musical",** Daniel Maggiolo (2002).

[19] Foro científico para Modelos Ocultos Markov http://www.icsi.berkeley.edu/Speech/papers/gelbart-ms/hybrid-testbed/

[20] Web descarga B.D Isolet:

http://www.icsi.berkeley.edu/speech/papers/eurospeech05-onset/isolet/

[21] Foro de funciones Matlab: http://www.engr.uky.edu/~lgh/soft/dilate.m

[22] Página Web con contenidos descargables (Quicknet):

http://www.icsi.berkeley.edu/Speech/qn.html

[23] Página Web con contenido descargable (SPRACHCORE):

http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html

[24] **"*Spoken Language Processing: A guide to theory, algorithm, and system development*"**. Prentice Hall, Inc., 2001,

[25] B. Moore, **"*Hearing*".** Academic Press, Inc., 1995.

[26] **"Auditori Toolbox Tech Report"** de Malcolm Slaney.