UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

REDUCCIÓN DE RUIDO EN GRABACIONES DE AUDIO

Ingeniería de Telecomunicación

Guillermo González Caravaca

Julio 2011

REDUCCIÓN DE RUIDO EN GRABACIONES DE AUDIO

AUTOR: Guillermo González Caravaca

TUTOR: Doroteo Torre Toledano

Grupo de Reconocimiento Biométrico - ATVS
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
JULIO 2011

Resumen

El contexto de este proyecto es el conjunto de técnicas orientadas a la reducción de ruido en grabaciones de voz, tomadas en entornos embarcados, como es un vehículo en marcha. El estudio comienza con el análisis de uno de los filtros de audio más utilizados, el filtro de Wiener, desgranando todas sus particularidades. Se probará una implementación del mismo y se estudiaran los resultados para comprobar su eficacia, con objeto de determinar posible mejoras que serán sometidas a prueba.

Su objetivo es lograr una mejora sustancial de las implementaciones del filtro de Wiener ya existentes, poniendo a prueba la hipótesis que se plantea al inicio; lograr una versión del filtro de Wiener con dependencia fonética, que consiga mejores resultados en la reducción de ruido.

Palabras Clave

Filtro de Wiener, reducción de ruido, distorsión de voz, clase amplia fonética, detector de actividad de voz, adaptabilidad.

Abstract

The context of this project is a set of techniques aimed at reducing noise in voice recordings, taken on board environments, such as a moving vehicle. The study begins with the analysis of one of the most used filters in noise reduction, the Wiener filter, reeling off all its peculiarities. It will test an implementation of it and study the results for effectiveness, to determine possible improvements to be tested.

Its aim is to achieve a substantial improvement of the Wiener filter implementations already exist, testing the hypothesis posed at the beginning, a phone-conditioned Wiener filter that can get better results in noise reduction.

Key Words

Wiener filter, noise reduction, speech distortion, broad phonetic classes, voice activity detector, adaptability.

Agradecimientos

En primer lugar quiero agradecer a mi tutor, Doroteo Torre, la oportunidad que me ha brindado de poder realizar este proyecto bajo su consejo. Su paciencia y determinación son dignas de elogio.

Asimismo, quisiera agradecer a todos los miembros del ATVS por el buen ambiente dentro de este y por la ayuda prestada ante cualquier problema que ha podido surgir.

En el plano personal, me gustaría comenzar dando las gracias a mis padres, sin ellos, no sería la persona que soy hoy. A mi hermana por animarme siempre a realizar nuevos proyectos, y a terminar este, y a mi hermano, por haberme inculcado desde pequeño la pasión por la ciencia y la tecnología.

Gracias a Laura por haberme apoyado durante todos estos años y haber hecho que los buenos momentos lo fueran aún mejor. Este proyecto y lo que ello culmina también te pertenece.

Gracias a Alberto y Pablo por haber hecho de mi estancia en la universidad, una de las mejores épocas de mi vida. Y en especial a Bruno, porque sin él, los laboratorios no habrían sido lo mismo.

Por último, gracias a mis amigos de toda la vida, que han sabido compartir conmigo lo mejor de sí. Este proyecto también va dedicado a vosotros.

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1

Introducción

1.1. Motivación del proyecto

La penetración de las tecnologías del habla en la sociedad actual es cada vez mayor. Un claro ejemplo de ello es el creciente uso de la telefonía móvil, que permite que millones de usuarios a la vez, puedan mantenerse en contacto desde prácticamente, cualquier lugar. La domótica es otro referente de este hecho, desde hace varios años, una persona es capaz de subir y bajar las persianas de su casa, con una sola orden vocal. Con la proliferación de este tipo de sistemas, la necesidad de que el intercambio de información sea fiable y sin distorsiones es cada vez mayor. Para mantener la integridad de la señal que se desea transmitir, almacenar o procesar, es necesario dotar a estos sistemas de mecanismos de defensa frente al ruido, distorsiones u otro tipo de señales interferentes que hagan, que la calidad de la señal vocal de origen se vea mermada, y por tanto, de lugar a un fallo en la transmisión del mensaje.

A causa de la gran diversidad de aplicaciones que se le puede dar a este tipo de sistemas, es necesario un análisis más específico, en función del entorno donde se desee aplicar, dado que tanto la naturaleza de las fuentes de ruido y las señales vocales, como la posterior aplicación que se les pueda dar, van a requerir unas condiciones optimas de filtrado distintas, y por tanto necesitan algoritmos y métodos de reducción de ruido adaptados a cada entorno.

1.2. Objetivos

El objetivo de este proyecto es estudiar y analizar el comportamiento de los distintos métodos de reducción de ruido aplicado a señales de voz que más se utilizan en la actualidad, para poder utilizarlos sobre locuciones obtenidas a bordo de vehículos, y en el caso de ser posible, proponer y desarrollar las mejoras que sean necesarias para optimizar el proceso de reducción de ruido.

Asimismo, se ahondará en la temática de los sistemas de reconocimiento automático del habla, que dependen directamente de la reducción de ruido, y en la tecnología de detección de actividad de voz. Ambos elementos juegan un papel trascendental en el procesado de voz, y veremos cómo pueden afectar en la mejora de los sistemas de reducción de ruido.

1.3. Estructura de la memoria

La estructura de este PFC se organiza como sigue:

En el *capítulo 2* se presenta una revisión de los conceptos básicos necesarios para abordar con soltura los métodos de reducción de ruido y su evaluación, como son los modelos de señales empleados, o los factores de definen la calidad del filtrado. Asimismo, se profundiza en el área de la reducción de ruido con el filtro de Wiener y sus posibles aplicaciones. También se introduce la notación que se utilizará durante todo el proyecto.

En el *capítulo 3* se desarrolla la idea central de este proyecto, la propuesta de los cambios necesarios para mejorar los resultados del filtro de reducción de ruido estudiado en el *capítulo 2*. Con el planteamiento de dichas mejoras, se realiza el diseño detallado de los sistemas de reducción de ruido que van a ser sometidos a test y posteriormente evaluados. Al final del *capítulo 3*, podremos encontrar una descripción de las herramientas software utilizadas para la consecución de todas las pruebas.

En el *capítulo 4* se detallan las distintas pruebas que se han llevado a cabo para la evaluación de las propuestas de mejora, y se presentan los resultados obtenidos a partir de las mismas.

En el *capítulo 5* se detallan las conclusiones del proyecto y las posibles líneas de trabajo futuro.

Capítulo 2

Estado del arte

En este capítulo vamos a familiarizarnos con los conceptos básicos necesarios para abordar la temática de reducción de ruido. Comenzaremos revisando las definiciones de ruido, y como éste afecta a las señales de información, como podemos modelarlo matemáticamente y cuantificarlo.

Con los modelos de señal planteados, y unos breves conceptos previos, ahondaremos en las técnicas de reducción de ruido, en concreto con el Filtro de Wiener, y su desarrollo matemático, para hacernos una idea de cómo funciona y que aplicaciones tiene.

Finalizaremos el capítulo profundizando en la teoría de detección de voz aplicada al filtro de Wiener, y sobre su importancia en la reducción de ruido.

2.1. El Ruido

En el ámbito de las comunicaciones, existen dos tipos de elementos perturbadores de una señal, estos son el ruido y la distorsión. Mientras que la distorsión es una modificación de la señal producida, por ejemplo, por las no linealidades del canal, el ruido es un elemento independiente de la señal, pero que como consecuencia puede acarrear la degradación de la calidad y la inteligibilidad de la misma, o en su caso, al procesamiento y/o almacenamiento de dicha señal. Si esto lo aplicamos en el campo de las señales de voz, los efectos del ruido pueden llegar a ser muy perjudiciales. Para intentar reducir al máximo dichos efectos, y mejorar la calidad de las comunicaciones, se han desarrollado diversas técnicas de procesamiento de señal, que ayudan a mejorar la calidad de la voz, eliminando de la manera más óptima todo el ruido que sea posible.

Antes de abordar el problema de cómo reducir o eliminar ese ruido, es necesario definirlo, caracterizarlo y clasificarlo. En este contexto podemos definir el ruido como "toda señal no deseada, que interfiere en la comunicación, procesamiento o medida de otra señal portadora de información" [1]. La simplicidad de esta idea nos puede ayudar a acercarnos al problema desde una perspectiva general, no obstante, no es una definición que permita abordar el problema de una forma técnica, por lo que es necesario realizar una clasificación menos generalista de los distintos tipos de ruido, y su procedencia. Hay que destacar además que con esta definición no estamos

excluyendo los efectos provenientes de la distorsión, que aun no siendo ruido, se consideran como tal al tratarse de parte de señal no deseable.

De este modo, podemos definir lo siguientes tipos de ruido/distorsión:

- ***Ruido aditivo***
  En este caso, el ruido aditivo se puede considerar todo aquel ruido procedente de distintas fuentes que coexisten en el mismo entorno acústico.

- ***Señales interferentes***
  En el caso de señales de voz, se considera señal interferente a toda aquella que proceda de otros locutores, que no sean objeto de interés.

- ***Reverberación***
  Producida por la propagación multitrayecto que se da en los entornos acústicos cerrados o semi cerrados. No se trata exactamente de ruido, sino de una forma de distorsión.

- ***Eco***
  Producido generalmente por el acoplamiento entre los micrófonos y los altavoces. Al igual que en el caso anterior, se trata de una forma de distorsión.

Cada una de estas subclases de ruido/distorsión representa un campo de investigación distinto, y es por ello que en los últimos años se han desarrollado avanzadas técnicas de procesamiento de señal de voz, cada una de estas técnicas orientada a suprimir los efectos negativos antes mencionados. En el caso de este proyecto, vamos a ahondar en el área de la reducción de ruido aditivo, y suponer que no nos vemos afectados por el resto de casos.

2.1.1. Ruido Aditivo

Para precisar en la definición de ruido aditivo, podemos considerar que una señal de voz está formada por la superposición de la voz limpia y del ruido. De esta manera, la reducción de ruido llevará a cabo la tarea de separar ambas partes de la forma más óptima posible.

Una de la propuestas iniciales que se plantean a la hora de abordar el filtrado, es plantear la eliminación del ruido como un problema de estimación de parámetros, donde la estimación optima de la voz limpia puede llevarse a cabo bajo el criterio de optimización de, por ejemplo, el factor MSE (*Mean Squared Error*) o de la SNR (*Signal to Noise Ratio*) de la estimación de la voz limpia frente al audio original.

Desafortunadamente, este criterio de optimización, en algunos casos, no coincide en la realidad con lo que el oído humano percibe como la mejor calidad, y es que tenemos enfrentados parámetros subjetivos y parámetros objetivos a la hora de evaluar las técnicas de filtrado. Es por ello que se hace necesario replantear el problema del ruido aditivo, estableciendo nuevos objetivos, que tengan en cuenta esta nueva situación:

- Mejorar criterios objetivos, MSE, SNR, etc.
- Mejorar la calidad que se percibe de la señal restaurada.
- Como paso previo a otros procesamientos de señal de voz, aumentar la robustez de otros sistemas (codificación de voz, reconocimiento de voz, etc.) frente al ruido.

Dependiendo de qué objetivo deseemos cumplir, la complejidad y la dificultad del filtrado puede variar tremendamente, pero en general, el número de micrófonos (o canales) utilizados para obtener las grabaciones, será determinante. En este caso, cuantos más canales haya disponible, más opciones se abren para mejorar la calidad de la voz. Por ejemplo, supongamos que tenemos varios micrófonos disponibles, situados uno de ellos cerca del locutor, y el resto a cierta distancia, captando el sonido ambiente. Si consideramos los últimos micrófonos como nuestra referencia de ruido, la obtención del canal de voz limpio se simplifica, dado que no es necesario aplicar complejos algoritmos al estimar el ruido. En base a esto, podemos afirmar que cuantos más micrófonos estén disponibles, mayores posibilidades hay para el filtrado de la voz.

En la realidad, esta situación de un "array" de micrófonos no es la más común. Un ejemplo muy sencillo de esto sería el teléfono móvil, el cual sólo dispone de un micrófono, por el cual son captados voz y ruido ambiente por igual. En este caso, estaríamos hablando de un sistema mono canal, y la reducción de ruido se complica, dado que no tenemos ninguna referencia de ruido, y por tanto, tendremos que hacer uso de técnicas más complejas para realizar el filtrado.

Los trabajos en el filtrado de señales de voz en sistemas mono canal comenzaron hacia 1958, por el profesor Manfred R. Schroeder, en el que proponía por vez primera una implementación analógica de la substracción espectral. Quince años más tarde, se haría lo mismo pero en el campo de las señales digitales. En el año 1979, los investigadores Jae S. Lim y Alan V. Oppenheim, en sus trabajos sobre voz ruidosa, realizaron un análisis de las técnicas existentes hasta el momento en el campo de la mejora de las señales de voz, y concluyeron que la reducción de ruido no solo tenía efectos beneficiosos sobre la calidad de la voz recuperada, sino también sobre la calidad e inteligibilidad de la codificación lineal predictiva (sus siglas en inglés LPC), útil en los sistemas de codificación y reconocimiento de voz.

Las técnicas desarrolladas hasta ahora pueden englobarse en tres grandes grupos, en función de cómo se realice la reducción de ruido:

1. Filtrado lineal adaptativo.
2. Substracción espectral.
3. Basado en modelo.

La base del filtrado lineal adaptativo, como su propio nombre indica, es hacer pasar a la señal ruidosa a través de un filtro lineal que se adapta al ruido a eliminar, atenuando así la componente de ruido, dejando la señal de voz sin distorsionar, en la medida de lo posible. Los filtros de Wiener estarían dentro de esta categoría. En su lugar, los métodos de substracción espectral, realizan la reducción de ruido a través de una estimación del espectro de la señal de voz, a partir de la señal original ruidosa. El algoritmo más conocido de esta categoría seria MMSE (*Minimum-Mean-Squared-Error*). Los métodos de reducción basados en modelos, tratan la reducción de ruido como un problema de estimación de parámetros, donde se hace uso de diversos modelos matemáticos de la generación de la voz. Técnicas como LP-Kalman (*Linear Prediction*) son representativas de este grupo.

2.2. Modelo de señal empleado

La reducción de ruido que se pretende llevar a cabo, está basada en recuperar la señal de voz de interés *x(n)* de la señal ruidosa observada

$$y(n) = x(n) + v(n) \quad (1)$$

donde *v(n)* es la señal de ruido que se pretende eliminar, asumiendo que es un proceso aleatorio de media cero e incorrelado con la señal de voz. Podemos considerar la señal *y(n)* como un vector de la forma

$$y(n) = [y(n)\, y(n-1) \ldots y(n-L+1)]^T \quad (2)$$

que incluye las *L* muestras más recientes, donde *x(n)* y *v(n)* están definidas de forma similar. De esta manera, el problema de la reducción de ruido se basa en la estimación de *x(n)* a partir de la señal *y(n)* original.

Aplicando una transformada de Fourier discreta (DFT) sobre los *L* puntos definidos, podemos decir que la señal observada *y(n)*, en el dominio de la frecuencia quedaría de la forma

$$Y(n, \mathrm{i}\omega_k) = X(n, \mathrm{i}\omega_k) + V(n, \mathrm{i}\omega_k) \quad (3)$$

donde tenemos que

$$Y(n, \mathrm{i}\omega_k) = \sum_{l=0}^{L-1} w(l) y(n - L + l + 1) \mathrm{e}^{-\mathrm{i}\omega_k l} \quad (4)$$

es la DFT de la señal *y(n)* ruidosa en el instante n-ésimo, *w(l)* es la función de enventanado escogida (por ejemplo, ventana de Hamming), $X(n, \mathrm{i}\omega_k)$ y $V(n, \mathrm{i}\omega_k)$ son las señales de voz y ruido respectivamente, definidas de la misma manera que $Y(n, \mathrm{i}\omega_k)$. Ahora en el dominio de la frecuencia, podemos decir que la reducción de ruido se basa en la estimación de $X(n, \mathrm{i}\omega_k)$ a partir de $Y(n, \mathrm{i}\omega_k)$.

2.3. Evaluación de la reducción de ruido

El principal objetivo de la reducción de ruido, en nuestro caso, es eliminar el ruido de fondo de la muestra de audio, e intentar evitar que la señal de voz se vea afectada, distorsionándola o produciendo algún otro efecto no deseado. Para verificar esto último, cuando hemos llevado a cabo un filtrado, necesitamos algún criterio en el que basarnos para comprobar el rendimiento de la operación.

Existen dos categorías en las que clasificar estos criterios, y son:

- ***Medidas subjetivas***

Las medidas subjetivas hacen referencia a un test realizado por un grupo de personas, escuchando la muestra de audio, y asignando una calificación a éste, o realizando una comparación con otros audios de las mismas características. Se podría decir que en este caso se realiza un examen cualitativo del resultado del filtrado.
Existen varios test en este sentido, como pueden los test MOS (*Mean Opinion Score*) o los test CE (*Categorical Estimation*).

- ***Medidas objetivas.***

Al contrario que las medidas subjetivas, las medidas objetivas se obtienen a partir de los resultados del filtrado, atendiendo así a aspectos cuantitativos de la señal, siendo independientes de criterio humano alguno.

Atendiendo a estas dos categorías a la hora de comprobar los resultados obtenidos, siendo coherentes con los objetivos propuestos al principio, deberíamos dar más importancia a las medidas subjetivas, ya que éstas están basadas en el juicio de la persona que escucha, y por tanto, el usuario final. En la práctica, realizar este tipo de medidas es de una gran complejidad y coste, por el tiempo empleado en realizar las medidas y la escasa uniformidad, dependiendo siempre del criterio de una persona. Es por ello que gracias a su simplicidad y rapidez en los cálculos, las medidas objetivas son las más usadas en esta área. En esta línea, varios algoritmos de medidas objetivas han sido desarrollados, siendo las más comunes la medida de SNR o la medida de la distancia *Itakura-Saito* [2].

- Medidas objetivas de calidad subjetiva: son intentos de aproximación de las medidas subjetivas mediante medidas objetivas, y que por tanto tienen las ventajas de ambos métodos, aunque no son tan fiables como las medidas subjetivas si el experimento está correctamente diseñado y cuenta con un número suficiente de personas. Ejemplo de este tipo de medidas son las definidas en la serie de recomendaciones UIT-T P.800 [3].

2.3.1. Conceptos previos

- ***Relación Señal a Ruido (Signal To Noise Ratio)***

La relación señal a ruido (SNR – *Signal to Noise Ratio*) es uno de las medidas más utilizadas en el campo de la reducción de ruido, cuantificando como de ruidosa es una señal en referencia a los niveles de voz y ruido. Esta relación está definida como la intensidad de la señal de interés (en nuestro caso, la voz) relativo a la intensidad de señal del ruido de fondo, y generalmente se representa en decibelios (dB). Con el modelo de señal de *y(n)* presentado anteriormente, podemos definir la SNR como:

$$SNR \triangleq \frac{\sigma_x^2}{\sigma_v^2} = \frac{E[x^2(n)]}{E[v^2(n)]} \tag{5}$$

donde el operador *E[]* representa la esperanza estadística de una señal dada.

Podemos definir, como hemos realizado anteriormente con otras señales, la SNR en el dominio de la frecuencia, haciendo uso del Teorema de Parseval de la forma

$$SNR = \frac{\int_{-\pi}^{\pi} P_x(\omega) d\omega}{\int_{-\pi}^{\pi} P_v(\omega) d\omega} \tag{6}$$

donde $P_x(\omega)$ y $P_v(\omega)$ son, respectivamente, las densidades espectrales de potencia de las señales temporales *x(n)* y *v(n)*, y donde ω es la frecuencia angular.

En el ámbito de reducción de ruido, esta medida se suele utilizar como SNR a priori (SNR_{in}) y SNR a posteriori (SNR_{out}), y se suele considerar que cuanto más alto sea el valor de SNR, mejor es la calidad del audio. En este sentido, para establecer la calidad de un filtrado, se pueden comparar ambas SNR definidas anteriormente (SNR_{out} y SNR_{in}). La diferencia de ambas se le denomina SNR de mejora, y cuanto más alta sea ésta, podemos decir que mejor son los resultados de la reducción de ruido.

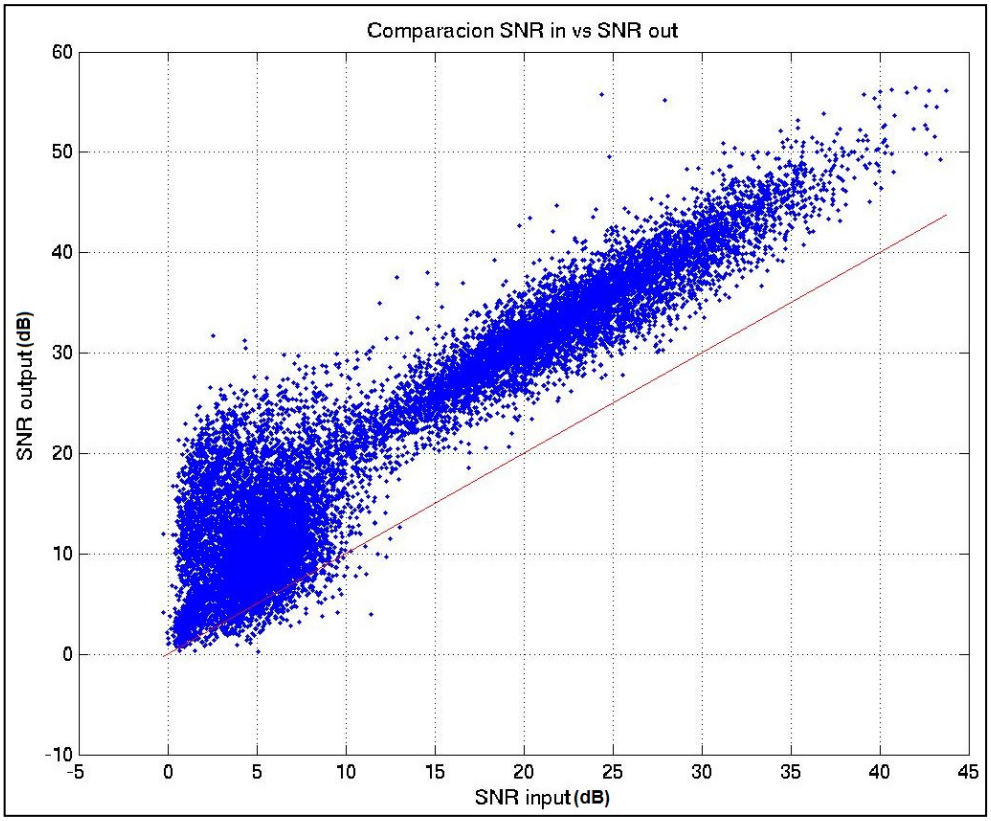

Figura 1. Ejemplo de grafica de resultados de SNR. La recta diagonal representa los puntos en los que la $\boldsymbol{SNR_{in}}$es igual a la $\boldsymbol{SNR_{out}}$.

- ***Factor de reducción de ruido***

Antes de definir otros parámetros que pueden ser objeto de medida en el área de la reducción de ruido, tenemos que definir un término que haga referencia a cuanto ruido es eliminado o atenuado en una determinada muestra, y para ello hacemos uso del *factor de reducción de ruido.* Este se obtiene a partir de la relación entre la intensidad de ruido inicial en la muestra, y la intensidad del ruido remanente tras el filtrado. Teniendo en cuenta la señal *y(n)* definida anteriormente, llamamos $v_r(n)$ al ruido residual, siendo el factor de reducción de ruido tal que

$$\xi_{nr} \triangleq \frac{E[v^2(n)]}{E[v_r^2(n)]} \tag{7}$$

Si el resultado del filtrado es satisfactorio, este factor será mayor que la unidad. También se puede comprobar que cuanto más alto sea ξ_{nr}, mejor será la calidad de la reducción de ruido.

A la hora de evaluar este tipo de medidas, hay que tener en cuenta que las señales acústicas suelen tener gran cantidad de fluctuaciones, y esto es igual para las señales de voz y de ruido. Es por ello que no se debe entender el factor de reducción de ruido como un valor absoluto, de forma que este ha de ser calculado como una ponderación media de la reducción de ruido en todos los instantes de la señal acústica (esperanza matemática). Además de las variaciones temporales, hay que tener en cuenta las que se dan en el dominio de la frecuencia, las cuales, generalmente, no son uniformes, por lo que habría que hacer medidas en cada banda para saber exactamente como se ha comportado el ruido tras el filtrado.

Para hacernos una idea más general del comportamiento del ruido en frecuencia, podemos hacer uso nuevamente del factor de reducción de ruido definido para densidades espectrales, de forma que

$$\Psi_{nr}(\omega) \triangleq \frac{E[|V(i\omega)|^2]}{E[|V_r(i\omega)|^2]} = \frac{P_v(\omega)}{P_{v_r}(\omega)} \tag{8}$$

donde $V(i\omega)$, $V_r(i\omega)$, $P_v(\omega)$ y $P_{v_r}(\omega)$ son los espectros de Fourier y las densidades espectrales de potencia de *v(n)* y $v_r(n)$ respectivamente. De esta forma, la función de ganancia de reducción de ruido es dependiente de la frecuencia.

- ***Índice de distorsión del habla***

Hasta ahora, solo se ha tenido en cuenta los efectos del filtrado sobre la señal de ruido, pero no se ha comprobado el resultado sobre la señal de voz. Es posible que esta se vea afectada, y es un factor que habrá que tener en cuenta, dado que cuanto más agresivo es el filtrado, es lógico pensar que mayor será la distorsión generada en la señal de voz. Es vital entonces, realizar la operación de reducción de ruido con toda la información de la que dispongamos a priori de la señal, para intentar disminuir la degradación de la voz. Por este motivo, definimos el **Índice de distorsión de la voz** (SDI, *speech distortion index*) como una medida para cuantificar la distorsión que genera el algoritmo de reducción de ruido utilizado.

A partir de los modelos de señal definidos anteriormente, si tenemos que $\hat{x}_{nr}(n)$ es la componente estimada de la voz limpia en el algoritmo, podemos decir que el índice de distorsión de la voz es

$$\varphi_{sd} \triangleq \frac{E\{[x(n)-\hat{x}_{nr}(n)]^2\}}{E[x^2(n)]} \tag{9}$$

El rango de valores de φ_{sd} es de cero a uno, correspondiendo cero con un valor de distorsión nula, y uno como un valor alto con gran distorsión. Por tanto, nuestro objetivo para mantener un compromiso de calidad con la señal de voz será mantener φ_{sd} tan bajo como nos sea posible. Es importante remarcar que para calcular el índice de distorsión de la voz es necesario disponer de una referencia de la voz sin ruido $x(n)$, lo que limita en gran medida sus posibilidades de aplicación.

Para medir la distorsión de la voz en el espacio de la frecuencia, tenemos que hacer uso del concepto de **distorsión de atenuación de frecuencias** o simplemente **distorsión de atenuación** usado en teoría de comunicación. La distorsión de atenuación es una medida que fue desarrollada para evaluar como un canal telefónico es capaz de mantener la fidelidad de una señal de voz. Está definida a partir de la variación de amplitud de la señal transmitida sobre la banda de frecuencias de voz.

Adaptando este concepto, podemos definir la distorsión de atenuación de frecuencias como

$$\Phi_{sd}(\omega) \triangleq \frac{E\,[|X(i\omega)|^2-|\hat{X}_{nr}(i\omega)|^2]}{E[|X(i\omega)|^2]} = \frac{P_x(\omega)-P_{\hat{x}_{nr}}(\omega)}{P_x(\omega)} \tag{10}$$

donde $X(i\omega)$ y $P_x(\omega)$ son el espectro y la densidad espectral de potencia de la señal limpia *x(n)*, $\hat{X}_{nr}(i\omega)$ y $P_{\hat{x}_{nr}}(\omega)$ son respectivamente, el espectro y la densidad espectral de potencia de la componente de voz de la señal filtrada. No se puede decir que $\Phi_{sd}(\omega)$ sea el equivalente en frecuencia de φ_{sd}, ya que no existe una correspondencia directa entre ellas, aunque si estén relacionadas con respecto a lo que cuantifican. Al igual que con la medida anterior, esta medida requiere también disponer de la voz limpia, lo que también limita sus posibilidades de aplicación.

2.4. Reducción de ruido a través del filtrado

Una vez revisados los conceptos básicos sobre la reducción de ruido, vamos a pasar a ver las técnicas y algoritmos más utilizados. Para empezar, vamos a analizar las técnicas de filtrado. Estas se basan en el diseño de un filtro lineal o transformación de forma que, cuando hacemos pasar la señal ruidosa (voz y ruido aditivo) a través del filtro, la componente de ruido es atenuada. Los algoritmos más representativos en esta categoría (en el dominio temporal y frecuencial) son el *filtro de Wiener*, y el *filtro de Wiener paramétrico*.

2.4.1. Filtro de Wiener en el dominio temporal

El filtro de Wiener es una de las aproximaciones básicas a la reducción de ruido y tiene la particularidad de que es óptimo de acuerdo con el error cuadrático medio (*MSE*) entre la señal limpia y la señal obtenida por el proceso de filtrado.

El filtro de Wiener se puede formular en el dominio del tiempo y en el de la frecuencia. La formulación del filtro en el tiempo se obtiene minimizando el error cuadrático medio (*MSE*) entre una señal de interés y su estimación. Con los modelos de señal planteados anteriormente, la estimación de la componente de voz limpia se puede obtener haciendo que la señal *y(n)* pase a través de un filtro *FIR* especificado en el dominio temporal de la forma

$$\hat{x}(n) = \boldsymbol{h}^T \boldsymbol{y}(n) \tag{11}$$

donde

$$\boldsymbol{h} = [h_0\ h_1 \dots h_{L-1}] \tag{12}$$

representa la respuesta al impulso finita de longitud *L*. La señal de error entre la señal de voz limpia y su estimación en el instante *n* es definida como

$$e_x(n) \triangleq x(n) - \hat{x}(n) = x(n) - \boldsymbol{h}^T \boldsymbol{y}(n) \tag{13}$$

Según lo expuesto antes, nuestro objetivo es minimizar el MSE, por lo que la función MSE del filtro planteado sería

$$J_x(h) \triangleq E[e_x^2(n)] \tag{14}$$

Considerando el siguiente filtro

$$h_1 = [1 \quad 0 \quad \dots \quad 0]^T \tag{15}$$

si hiciésemos pasar la señal *y(n)* sobre este filtro, la salida sería idéntica a la entrada (no hay reducción de ruido). Para este caso, la función MSE correspondiente sería

$$J_x(h_1) = E[v^2(n)] = \sigma_v^2 \tag{16}$$

La estimación optima $\hat{x}_0(n)$ de la señal limpia de voz *x(n)* tiende a contener menos ruido que la señal observada *y(n)*, por lo que podemos decir que el filtro óptimo que forma $\hat{x}_0(n)$ es un filtro de Wiener óptimo, obtenido de la forma

$$h_0 = \arg\min J_x(h) \tag{17}$$

En principio, para el filtro óptimo h_0 tenemos que

$$J_x(h_0) < J_x(h_1) = \sigma_v^2 \tag{18}$$

lo que indica que el filtro óptimo de Wiener debe ser capaz de reducir el nivel de ruido en la señal ruidosa *y(n)*. De la ecuación (17), podemos obtener la ecuación de Wiener-Hopf

$$R_y h_0 = r_{yx} \tag{19}$$

donde

$$R_y = E[y(n)y^T(n)] \tag{20}$$

es la matriz de correlación de la señal observada *y(n)* y

$$r_{yx} = E[y(n)x(n)] \tag{21}$$

es el vector de la correlación cruzada de la señal observada y de la señal de voz limpia.

Se puede comprobar de las ecuaciones anteriores, que para obtener el filtro de Wiener óptimo h_0 es necesario conocer de antemano R_y y r_{yx}. La matriz de correlación R_y se puede obtener directamente a partir de la señal *y(n)*, pero dado que no tenemos acceso directo a la señal *x(n)*, el cálculo de r_{yx} complica la tarea de la obtención del filtro ideal.

Haciendo uso de la ecuación (17), obtenemos el valor de $x(n)$

$$x(n) = y(n) - v(n)$$

que podemos utilizar para poder calcular r_{yx} utilizando (21) de la siguiente manera

$$r_{yx} = E[y(n)x(n)] = E\,[y(n)\{y(n) - v(n)\}] \tag{22}$$

simplificando y utilizando (1) nuevamente para sustituir $y(n)$ tenemos

$$r_{yx} = E\,[y(n)y(n) - y(n)\,v(n)\}] = E[y(n)y(n)] - E[y(n)v(n)] = \\ r_{yy} - E[\{x(n) + v(n)\}v(n)] = E[x(n)v(n)] + E[v(n)v(n)] \tag{23}$$

Como asumimos al principio del capítulo, las señales $x(n)$ y $v(n)$ están incorreladas entre sí, de forma que anulando el termino sobrante en la ecuación anterior podemos decir que

$$r_{yx} = E[y(n)y(n)] - E[v(n)v(n)] = r_{yy} - r_{vv} \tag{24}$$

Ahora, r_{yx} depende de dos vectores de correlación, r_{yy} y r_{vv}. El vector r_{yy} no es más que la primera columna de la matriz R_y, y puede ser obtenida directamente de *y(n)*. El vector r_{vv} se puede obtener a través de la observación de *y(n),* en los tramos en los que la voz no está presente y sólo hay ruido. Con esta nueva información, podemos reescribir la ecuación de Wiener-Hopf

$$R_y h_0 = r_{yy} - r_{vv} \tag{25}$$

Si asumimos que la matriz R_y es invertible, como ocurre en la mayoría de las ocasiones, el filtro de Wiener se puede obtener resolviendo las ecuaciones Wiener-Hopf planteadas

$$h_0 = R_y^{-1} r_{yx} = R_y^{-1} r_{yy} - R_{yy}^{-1} r_{vv} = h_1 - R_y^{-1} r_{vv} \tag{26}$$

donde h_1 es el filtro definido en (15).

Si ahora definimos dos matrices de correlación normalizadas

$$\tilde{R}_x \triangleq \frac{R_x}{\sigma_x^2} , \qquad \tilde{R}_y \triangleq \frac{R_v}{\sigma_v^2} \tag{27}$$

donde R_x y R_v, son, respectivamente, la matrices de correlación del habla limpia y del ruido, y que están definidas de forma similar que R_y, el filtro de Wiener se puede expresar de forma

$$h_0 = \left[I - R_y^{-1} R_v\right] h_1 = [I - (SNR\tilde{R}_x + \tilde{R}_v)^{-1} \tilde{R}_v] h_1 \tag{28}$$

donde *I,* representa la matriz identidad y SNR es la relación señal a ruido de la señal de voz. Si ahora hacemos que la SNR tienda a infinito (condiciones óptimas), tenemos

$$\lim_{SNR \to \infty} h_0 = h_1 \tag{29}$$

Este hecho era de esperar, ya que en las condiciones de SNR descritas, no sería necesaria la reducción de ruido, y por tanto el filtro óptimo sería aquel que no variase la señal de entrada. Si en lugar de buscar condiciones optimas de SNR de entrada, hacemos que esta tienda a cero, tenemos que

$$\lim_{SNR \to 0} h_0 = 0 \tag{30}$$

donde el vector **0** tiene el mismo tamaño en muestras que h_0, y todos sus valores son cero. En este caso, cuando la señal de entrada al filtro carece de señal de voz (*SNR = 0*), el filtro no deja pasar nada, eliminando todo sonido.

Una vez formulado, podemos pasar a analizar como el filtro de Wiener puede reducir el nivel de ruido, como en un principio se espera. Para ello, vamos a echar un vistazo al

factor de reducción de ruido (definido en apartados anteriores). Sustituyendo nuestro filtro óptimo h_0 en las ecuaciones, obtenemos la estimación ideal de voz

$$\hat{x}_0(n) = h_0^T y(n) = h_0^T x(n) + h_0^T v(n) \tag{31}$$

Se puede observar que los términos de lado derecho de la ecuación anterior, son $h_0^T x(n)$, siendo esta la componente de voz limpia filtrada con el filtro de Wiener, y $h_0^T v(n)$ la componente de ruido residual. Por lo tanto, el factor de reducción de ruido, según la ecuación (7) puede ser descrito como

$$\xi_{nr}(h_0) = \frac{E\{[h_1^T v(n)]^2\}}{E\{[h_0^T v(n)]^2\}} = \frac{h_1^T R_v h_1}{h_0^T R_v h_0} \tag{32}$$

Sustituyendo $h_0 = R_y^{-1} r_{yx} = R_y^{-1} r_{xx} = R_y^{-1} r_x h_1$ en (32) llegamos a que

$$\xi_{nr}(h_0) = \frac{h_1^T R_v h_1}{(h_1 R_x R_y^{-1})^T R_v (R_y^{-1} R_x h_1)} \tag{33}$$

que es una función que depende de de las tres matrices de correlación R_x, R_v y R_y. Utilizando la descomposición de autovalores [4], podemos descomponer estas tres matrices de correlación de la siguiente manera:

$$\begin{aligned} R_x &= B^T \Lambda B \\ R_v &= B^T B \\ R_y &= B^T (I + \Lambda) B \end{aligned} \tag{34}$$

donde B es una matriz cuadrada invertible y

$$\Lambda = diag(\lambda_1\ \lambda_2\ \ldots\ \lambda_L) \tag{35}$$

que es una matriz diagonal que cumple que $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_L \geq 0$. Sustituyendo (34) en la ecuación (33), se deduce que

$$\xi_{nr}(h_0) = \frac{\sum_{i=1}^{L} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2} \tag{36}$$

donde $b_{i1} = 1, 2, \ldots, L$ es la primera columna de la matriz B y que hace $\sum_{i=1}^{L} b_{i1}^2 = \sigma_v^2$.

Nuevamente, a partir de la ecuación (34), podemos calcular la SNR de la señal observada de forma

$$SNR = \frac{h_1^T R_x h_1}{h_1^T R_v h_1} = \frac{\sum_{i=1}^{L} \lambda_i b_{i1}^2}{\sum_{i=1}^{L} b_{i1}^2} \tag{37}$$

Usando la ecuación (36) que acabamos de obtener, podemos recalcular el factor de reducción de ruido de la forma

$$\xi_{nr}(h_0) = \frac{1}{SNR} \frac{\sum_{i=1}^{L} \lambda_i b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2} =$$

$$\frac{1}{SNR} \frac{\sum_{i=1}^{L} \frac{(1+\lambda_i)^2}{(1+\lambda_i)^2} \lambda_i b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2} = \frac{1}{SNR} \left(\frac{\sum_{i=1}^{L} \frac{\lambda_i + \lambda_i^3}{(1+\lambda_i)^2} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2} + 2 \right) \tag{38}$$

Teniendo en cuenta que $\lambda_i + \lambda_i^3 \geq \lambda_i^3$ se puede deducir de la ecuación (38) que

$$\xi_{nr}(h_0) \geq \frac{1}{SNR} \left(\frac{\sum_{i=1}^{L} \frac{\lambda_i^3}{(1+\lambda_i)^2} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2} + 2 \right) \tag{39}$$

Teniendo en cuenta las consideraciones realizadas en el Anexo A, si hacemos que $\mu = 1$ y $q_i = b_{1i}$, podemos obtener la siguiente inecuación

$$\frac{\sum_{i=1}^{L} \frac{\lambda_i^3}{(1+\lambda_i)^2} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2} \geq \frac{\sum_{i=1}^{L} \lambda_i b_{i1}^2}{\sum_{i=1}^{L} b_{i1}^2} = SNR \tag{40}$$

que pasa a convertirse en ecuación si y solo si todos los λ_i correspondientes a los b_{1i} distintos de cero son iguales, donde $i = 1, 2, \dots, L$. De esto podemos deducir que

$$\xi_{nr}(h_0) \geq \frac{SNR+2}{SNR} \tag{41}$$

Por tanto, podemos decir que el factor de reducción de ruido descrito en (41) será siempre mayor que 1, teniendo en cuenta que solo se consideran valores de SNR

positivos. Esto demuestra que la reducción de ruido es siempre posible con el filtro de Wiener. De (41) se puede demostrar que el factor de reducción de ruido es una función decreciente, acotada inferiormente. Su valor tiende a infinito cuando la SNR se aproxima a 0, y tiende a 1, cuando el valor de la SNR crece. Esto nos indica que se producirá una mayor reducción de ruido con valores de SNR bajos, lo cual es preferible y deseable, puesto que hay más cantidad de ruido a ser eliminado.

El índice de distorsión del habla para el filtro de Wiener, teniendo en cuenta su definición en (9), se puede reescribir de la siguiente manera

$$\varphi_{sd} = \frac{(h_1 - h_o)^T R_x (h_1 - h_0)}{h_1^T R_x h_1} \tag{42}$$

Como ya se dijo anteriormente, el valor del índice de distorsión del habla siempre cumple que

$$\varphi_{sd} \geq 0 \tag{43}$$

Sustituyendo (34) en (42) tenemos que

$$\varphi_{sd} = \frac{\sum_{i=1}^{L} \frac{\lambda_i^2}{(1+\lambda_i)^2} b_{i1}^2}{\sum_{i=1}^{L} \lambda_i b_{i1}^2} \leq \frac{\sum_{i=1}^{L} \frac{\lambda_i}{1+2\lambda_i} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i + 2\lambda_i^2}{1+2\lambda_i} b_{i1}^2} \leq \frac{1}{2SNR+1} \tag{44}$$

donde hemos hecho uso de la siguiente inecuación

$$\frac{\sum_{i=1}^{L} \frac{\lambda_i^2}{1+2\lambda_i} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i}{1+2\lambda_i} b_{i1}^2} \geq \frac{\sum_{i=1}^{L} \lambda_i b_{i1}^2}{\sum_{i=1}^{L} b_{i1}^2} = SNR \tag{44}$$

pudiendo ser probada teniendo en cuenta las consideraciones realizadas en el Anexo A.

Podemos concluir entonces que el filtro de Wiener siempre será capaz de lograr reducción de ruido dada una señal de entrada, a costa de distorsionar la señal de voz en cierta medida.

Llegados a este punto, solo nos queda comprobar cuales serán los efectos del filtro de Wiener sobre la SNR del audio de entrada, y verificar si a pesar de la distorsión introducida por el filtro, este es capaz de mantenerla o incluso mejorarla.

De la ecuación (37) sabemos que la SNR de la señal de entrada se define como

$$SNR = \frac{\sum_{i=1}^{L} \lambda_i b_{i1}^2}{\sum_{i=1}^{L} b_{i1}^2} \tag{45}$$

Tras el filtrado, tenemos que la SNR de la señal de salida es

$$SNR_0 = \frac{h_0^T R_x h_0}{h_0^T R_v h_0} \tag{46}$$

Si sustituimos en (46) el filtro $h_0 = R_y^{-1} R_x h_1$ tenemos que

$$SNR_0 = \frac{h_1^T R_x R_y^{-1} R_x R_y^{-1} R_x h_1}{h_1^T R_x R_y^{-1} R_v R_y^{-1} R_x h_1} \tag{47}$$

Haciendo uso de la descomposición matricial vista en (34), podemos deducir que

$$SNR_0 = \frac{\sum_{i=1}^{L} \frac{\lambda_i^3}{(\lambda_i+1)^2} b_{i1}^2}{\sum_{i=1}^{L} \frac{\lambda_i^2}{(\lambda_i+1)^2} b_{i1}^2} \tag{48}$$

Si revisamos la desigualdad vista en (44), podemos comprobar que

$$SNR_0 \geq SNR \tag{49}$$

Es decir, en las condiciones planteadas en este desarrollo, el filtro de Wiener siempre es capaz de aumentar la SNR de la señal de entrada al filtro, o lo que es lo mismo, siempre es capaz de lograr la reducción de ruido.

2.4.2. Filtro de Wiener subóptimo

Del análisis anterior, se ha podido concluir que el filtro de Wiener en el dominio del tiempo logra la reducción de ruido en todos los casos, en detrimento de la calidad de la señal de voz limpia, dado que ésta se ve distorsionada por el filtro. Este hecho nos hace

plantearnos los requisitos previos de nuestro filtro, y añadir un parámetro más, el cual es lograr la máxima reducción de ruido, sin poner en peligro la señal de voz, y por tanto, nuestro diseño ha de cumplir un compromiso de equilibrio entre la reducción de ruido y la distorsión del habla. Colocando ambos requisitos en una balanza, si le damos mayor importancia a uno de ellos, el otro se verá mermado, y viceversa. Por lo tanto es necesario establecer un control en el filtro, que nos permita inclinar la balanza hacia un lado u otro. Para ello, vamos a definir un filtro de Wiener subóptimo.

El filtro descrito en (26) tiene una interpretación física muy intuitiva: se compone de la suma de dos filtros, h_1 y $-R_y^{-1}r_{vv}$, donde cada uno de ellos tiene un propósito distinto. El primer filtro es el encargado de crear una réplica de la señal original de entrada, mientras que el segundo realiza la estimación (y supresión) del ruido. En esta línea, podríamos decir que el filtro de Wiener trabaja en dos pasos: crea una estimación óptima del ruido, para luego restarla de la señal de entrada. Si fuéramos capaces de introducir un parámetro que sea capaz de controlar la cantidad de ruido a eliminar, podríamos, acudiendo a la metáfora anterior, controlar de qué lado se inclina la balanza, y por tanto, mantener el compromiso entre reducción de ruido y distorsión de la voz.

Por lo tanto, ahora vamos a crear el siguiente filtro

$$h_{sub} = h_1 - \alpha R_y^{-1} r_{vv} \tag{50}$$

donde $\alpha \geq 0$ es un número real. Hay que destacar, que el filtro h_{sub} no es una solución óptima de acuerdo conel criterio MMSE visto anteriormente, por lo que podemos llamarlo filtro subóptimo.

Sustituyendo h_{sub} en (14) podemos hallar la función MSE correspondiente al filtro subóptimo

$$J_x(h_{sub}) = E\{[x(n) - h_{sub}^T y(n)]^2\} = \sigma_v^2 - \alpha(2-\alpha) r_{vv}^T R_y^{-1} r_{vv} \tag{51}$$

Para lograr la reducción de ruido con este nuevo filtro, el factor α ha de ser escogido, de forma que $J_x(h_{sub}) < J_x(h_1)$, de lo que se deduce que

$$0 < \alpha < 2 \tag{52}$$

El factor de reducción de ruido, para el caso del filtro subóptimo seria:

$$\xi_{nr}(h_{sub}) = \frac{E\{[h_1^T v(n)]^2\}}{E\{[h_{sub}^T v(n)]^2\}} \tag{53}$$

Gracias a la descomposición matricial vista en (34) , podemos reformular el factor de reducción de ruido de manera que

$$\xi_{nr}(h_{sub}) = \frac{\sum_{i=1}^{L} b_{i1}^2}{\sum_{i=1}^{L} \frac{(\lambda_i+1-\alpha)^2}{(1+\lambda_i)^2} b_{i1}^2} \tag{54}$$

De forma similar al factor de reducción de ruido, podemos reescribir el índice de distorsión de la voz

$$\varphi_{sd}(h_{sub}) = \frac{E\{[x(n)- h_{sub}^T x(n)]^2\}}{\sigma_x^2} = \alpha^2 \frac{\sum_{i=1}^{L} \frac{\lambda_i}{(1+\lambda_i)^2} b_{i1}^2}{\sum_{i=1}^{L} \lambda_i b_{i1}^2} = \alpha^2 \varphi_{sd}(h_o) \tag{55}$$

Por lo que la relación entre los índices de distorsión de voz correspondientes a los dos filtros, h_{sub} y h_0 solo depende del parámetro α.

Para lograr tener menor distorsión de voz en el caso del filtro subóptimo h_{sub} que en el caso del filtro de Wiener h_0, debemos encontrar un valor de α que haga que se cumpla

$$\varphi_{sd}(h_{sub}) < \varphi_{sd}(h_0) \tag{56}$$

A partir de (55) se comprueba que esta condición se satisface cuando α cumple $-1< \alpha <1$. Haciendo uso de (52), podemos determinar finalmente, que para los valores de α que $0 < \alpha < 1$, el filtro subóptimo h_{sub} reduce el nivel de ruido presente en la señal observada *y(n)* sin que la señal de voz se vea tan distorsionada como en el caso del filtro óptimo h_0. Para los casos extremos de α, cuando $\alpha = 0$, se tiene que $h_{sub} = h_1$, en el que no hay reducción de ruido, pero tampoco hay distorsión de voz. Para el caso que $\alpha = 1$, se tiene que $h_{sub} = h_0$, donde la reducción de ruido y la distorsión, son máximas.

La SNR a la salida del filtro subóptimo, viene dada por

$$SNR_{sub} = \frac{h_{sub}^T R_x h_{sub}}{h_{sub}^T R_v h_{sub}} \quad (57)$$

Para el caso de la SNR, mientras que α siga cumpliendo $0 < \alpha < 1$, se tiene que

$$SNR_{sub} < SNR \quad (58)$$

lo que quiere decir que el filtro subóptimo es capaz de mejorar la SNR de la señal de entrada al filtro, pero esta será siempre más baja o igual que la SNR del filtro de Wiener óptimo.

2.4.3. Filtro de Wiener definido en el dominio de la frecuencia

El filtro de Wiener también puede ser formulado en el dominio de la frecuencia. Una forma de obtener las ecuaciones correspondientes es aplicando directamente la transformación sobre las ecuaciones del filtro de Wiener definido en el tiempo. En ese caso, ambos filtros (temporal y frecuencial) presentan el mismo rendimiento. En otros casos, el filtro en el dominio frecuencia puede ser obtenido estimando directamente el espectro de la voz limpia a partir del espectro de la voz ruidosa.

Si se realiza a partir de este método, aparecen dos diferencias principales con respecto al filtro en el dominio del tiempo:

- ✓ El filtro en el dominio temporal es causal, mientras que el correspondiente al dominio frecuencial, no lo es.
- ✓ El filtro temporal trabaja sobre toda la banda de una vez, mientras que el frecuencial lo hace por porciones, haciendo que cada filtro sea independiente del resto.

Considerando el modelo de señal planteado en (3), podemos derivar el filtro de Wiener en el dominio de la frecuencia,

$$H_0(i\omega_k) = \arg\min_{H(i\omega_k)} J_X[H(i\omega_k)] \quad (59)$$

donde

$$J_X[H(i\omega_k)] = E[|X(n, i\omega_k) - H(i\omega_k)Y(n, i\omega_k)|^2] \quad (60)$$

es el error cuadrático medio (*MSE*) entre el espectro de la voz y su estimación a la frecuencia ω_k. Sustituyendo con [59,] y despejando, obtenemos que el filtro de Wiener es:

$$H_0(i\omega_k) = \frac{E[|X(n,i\omega_k)|^2]}{E[|Y(n,i\omega_k)|^2]} = \frac{P_x(\omega_k)}{P_y(\omega_k)} \tag{61}$$

donde

$$P_x(\omega_k) = E[|X(n,i\omega_k)|^2], \quad P_y(\omega_k) = E[|Y(n,i\omega_k)|^2] \tag{62}$$

son las densidades espectrales de potencia (*PSD*) de *x(n)* e *y(n)* respectivamente. De esta expresión cabe destacar que el filtro de Wiener en el dominio de la frecuencia presenta siempre valores positivos y reales, por lo que mantiene la componente de fase de la señal intacta.

De (61) podemos ver que para poder obtener el filtro de Wiener, es necesario conocer las densidades espectrales de potencia de las señales limpia y ruidosa. Para el caso de la señal ruidosa, el cálculo es directo ya que es la señal observada, pero la señal limpia *x(n)* no es accesible antes de la salida del filtro, lo que complica el cálculo de su densidad espectral de potencia.

Considerando que la señal de voz *x(n)* y la señal de ruido vista en (1) están incorreladas, podemos relacionar sus densidades espectrales de potencia de forma que

$$P_y(i\omega k) = P_x(i\omega_k) + P_v(i\omega_k) \tag{63}$$

Con ello, podemos reescribir la ecuación del filtro, obteniendo

$$H_0(i\omega_k) = \frac{P_y(i\omega_k) - P_v(i\omega_k)}{P_y(i\omega_k)} \tag{64}$$

Ahora si podemos tener acceso a todas las señales involucradas en la ecuación del filtro. A partir de la señal *y(n)* observada obtenemos de forma directa $P_y(i\omega k)$, y examinando los intervalos en los que no se detecta actividad de voz, obtenemos $P_v(i\omega k)$.

La estimación óptima del espectro de voz limpia haciendo uso del filtro anterior es

$$\hat{X}_0(n, i\omega_k) = H_0(i\omega_k)Y(n, i\omega_k) = H_0(i\omega_k)X(n, i\omega_k) + H_0(i\omega_k)V_0(n, i\omega_k) \quad (65)$$

Aplicando la transformada discreta de Fourier inversa sobre la señal anterior, se obtiene la estimación óptima de las muestras de voz $\hat{x}_0(n)$. La potencia de $\hat{x}_0(n)$ puede ser calculada haciendo uso del Teorema de Parseval de forma

$$\begin{aligned} E\left[\hat{x}_0{}^2(n)\right] &= \sum_{k=0}^{L-1} \frac{1}{L} E\left[|\hat{X}_0(n, i\omega_k)|^2\right] = \sum_{k=0}^{L-1} H_0^2(i\omega_k) P_y(i\omega_k) \\ &= \sum_{k=0}^{L-1} \frac{P_x^2(i\omega_k)}{P_y^2(i\omega_k)} P_x(i\omega_k) + \sum_{k=0}^{L-1} \frac{P_x^2(i\omega_k)}{P_y^2(i\omega_k)} P_v(i\omega_k) \end{aligned} \quad (66)$$

que es la suma de dos términos, donde el primero es la potencia de la voz limpia filtrada y el segundo es el ruido residual filtrado.

Si el ruido no es nulo (consideración inicial), podemos calcular el factor de reducción de ruido del filtro de Wiener en el dominio de la frecuencia basándonos en (7), por lo que

$$\xi_{nr}[H(i\omega_k)] = \frac{\sum_{k=0}^{L-1} P_v(i\omega_k)}{\sum_{k=0}^{L-1} \frac{P_x^2(i\omega_k)}{P_y^2(i\omega_k)} P_v(i\omega_k)} \quad (67)$$

considerando que $P_x^2(i\omega_k) \leq P_y^2(i\omega_k)$, se verifica que

$$\xi_{nr}[H(i\omega_k)] \geq 1 \quad (68)$$

Lo que nos dice que el filtro de Wiener puede reducir el nivel de ruido siempre que este no sea nulo. De igual manera se puede demostrar que la potencia de la señal de voz filtrada es menor que la potencia de la señal de voz original, como pasaba en el caso del filtro de Wiener definido en el tiempo en (26), por lo que la reducción de ruido se lleva a cabo asumiendo la distorsión en la señal de voz.

También se puede demostrar, como en el caso del filtro definido en el dominio temporal, que en este caso, el filtro de Wiener definido en frecuencia puede mejorar la SNR de la señal observada.

2.4.4. Filtro de Wiener Paramétrico

Para realizar la implementación del filtro de Wiener definido en el apartado anterior (dominio frecuencial), es necesario aplicar una serie de aproximaciones, ya que la densidad espectral de potencia de la señal ruidosa y el propio ruido han de ser estimados. Una forma de hacerlo es aplicando el teorema de Parseval , teniendo que

$$H_s(n, i\omega_k) = \frac{|Y(n,i\omega_k)|^2 - |V(n,i\omega_k)|^2}{|Y(n,i\omega_k)|^2} \tag{69}$$

Con este filtro, la estimación de la voz limpia se define entonces como

$$\hat{X}_0(n, i\omega_k) = H_s(n, i\omega_k) Y(i\omega_k) \tag{70}$$

Para proporcionar mayor flexibilidad al filtro, y poder mantener así el equilibrio entre reducción de ruido y distorsión, con lo cual el filtro de Wiener deja de ser óptimo y vuelve a ser subóptimo, la definición de $H_s(n, i\omega_k)$ se modifica de forma que se obtiene el filtro de Wiener que se suele denominar paramétrico

$$H_{PW}(n, i\omega_k) = \left[\frac{|Y(n,i\omega_k)|^p - \eta|V(n,i\omega_k)|^p}{|Y(n,i\omega_k)|^p}\right]^q \tag{71}$$

donde p y q son ambos números positivos reales no nulos, y η es un parámetro introducido para controlar la cantidad de ruido a ser eliminado. Con valores de $\eta > 1$ se realiza un filtrado muy agresivo, pero esto provoca mayor distorsión en la señal de voz. Si por el contrario se pretende que dicha distorsión sea mínima, es necesario escoger valores tal que $\eta < 1$. Las configuraciones más típicas de (p, q, η) suelen ser (1,1,1), (2,1,1) o (2,1/2, 1).

Por ello, la voz limpia estimada con el filtro paramétrico de Wiener se obtiene a través de:

$$\hat{X}_0(n, i\omega_k) = H_{PW}(n, i\omega_k) Y(n, i\omega_k) \tag{72}$$

Hay que destacar que las configuraciones más frecuentes de los parámetros del filtro detallados anteriormente no representan los valores óptimos de los mismos, aunque ello no limita las posibilidades del filtro, dada la sencillez de su implementación y su rápida adaptación a través de los parámetros a las condiciones de filtrado.

2.5. Detección de actividad de voz

Un detector de actividad de voz (*voice activity detector,* VAD) se encarga de clasificar segmentos de una señal de voz, como fragmentos de voz, si se ha detectado su presencia, o como fragmentos de no voz, si solo se ha encontrado ruido. Podemos asumir los mismos modelos de señal vistos hasta ahora, donde la componente de voz $x(n)$ se ha visto afectada por una señal de ruido aditivo $v(n)$, encontrándose ambas señales incorreladas. De esta manera, forman la señal $y(n)$, que será con la que se pretende trabajar, como ya se definió en la ecuación (1)

$$y(n) = x(n) + v(n)$$

2.5.1. Fundamentos de un detector de actividad de voz

Como se ha visto anteriormente, un VAD, o *detector de actividad de voz*, es el encargado de segmentar y etiquetar un audio en fragmentos clasificados como *voz* o *no voz*, indicando en cuales de estos fragmentos de la señal se ha detectado voz. Esta herramienta es de gran utilidad en el procesado de señales de voz, como puede ser codificación, reconocimiento, transmisión discontinua, etc. Con este propósito, se han desarrollado diversos algoritmos, que se adaptan a cada entorno acústico, optimizando diversos parámetros que definen un VAD como son el retardo, la sensibilidad del VAD, la precisión o el coste computacional.

El principal problema al que se enfrenta un VAD a la hora de decidir que es voz y que no lo es, se encuentra en el ruido de fondo presente en la señal observada, donde la variedad de su naturaleza e intensidad puede dificultar la tarea a la hora de procesar una señal. Es por ello, que en el momento de decidir qué tipo de algoritmo vamos a utilizar, debemos de tener en cuenta las condiciones acústicas del medio, para poder escoger la

opción que más se acerque a nuestros requisitos, aunque esta decisión también dependerá del tipo de procesado al que se vaya a someter la señal de voz.

2.5.2. Esquema básico de funcionamiento

El funcionamiento de un VAD básico trata de extraer diversas características o medidas realizadas sobre la señal observada y comparar estos valores con una serie de umbrales preestablecidos, seleccionados normalmente en función de las características del ruido y de la voz. La decisión sobre que es voz y que no lo es, se realiza cuando las medidas o características extraídas superan los umbrales anteriormente mencionados. En muchas ocasiones, en las que el ruido presenta un comportamiento no estacionario, los valores umbral han de ser actualizados constantemente para que la detección sea correcta.

Aunque este esquema de funcionamiento es el más habitual por su relación entre resultados y coste computacional, también hay que indicar que es posible realizar VADs más complejos en los que se emplean modelos más complejos y que pueden dar mejores resultados (por ejemplo de mezclas de Gaussianas, modelos ocultos de Markov, redes neuronales, support vector machines (*SVM*) o cualquier otro paradigma de aprendizaje automático).

De forma generalizada, podemos decir que el algoritmo de un VAD básico puede descomponerse en dos partes:

1.- Cálculo de umbrales y toma de medidas y/o extracción de características.
2.- Aplicación de la regla de decisión en función de los umbrales.

Independientemente del método utilizado en cada implementación de un VAD, el principal compromiso a mantener es no identificar fragmentos de ruido como voz y viceversa, o en su caso, reducir al mínimo estos fallos. El poder cumplir los objetivos planteados por este compromiso de calidad hace que la tarea de detección de voz sea más compleja en entornos altamente ruidosos, donde las señales presentan valores de SNR muy bajos, siendo difícil de distinguir la voz frente al ruido.

Para resolver este problema, es necesario que el VAD sea robusto frente al ruido, ya que de esta forma estamos asegurando su funcionamiento en una amplia variedad de

condiciones acústicas. Podremos decir que un VAD es robusto frente al ruido cuando aporta resultados similares, tanto para señales de voz limpia, como con señales de voz ruidosa. De esta forma cuanto más robusto sea el VAD, menores serán los errores de detección.

2.5.3. Evaluación de un VAD

El rendimiento o la calidad de un VAD se puede medir en términos de la cantidad de errores cometidos, detectando ruido como voz o viceversa, y en la agresividad a la hora de decidir la duración de los fragmentos detectados como voz. Dicho rendimiento se evalúa en función de cinco parámetros básicos, comparando los resultados que proporciona el VAD en estudio con los que nos aportaría un VAD ideal (habitualmente un etiquetado manual). Los parámetros objetivos utilizados en la evaluación son:

- *Front End Clipping* (FEC): recorte generado al pasar de fragmentos clasificados como ruido a los que han sido clasificados como voz.
- *Mid Speech Clipping* (MSC): recortes debidos a fragmentos de voz clasificados erróneamente como ruido.
- OVER: ruido contiguo a un fragmento de voz que ha sido clasificado como voz dentro del mismo fragmento.
- *Noise Detected as Speech* (NDS): ruido interpretado como voz en un periodo de silencio.
- *Correct VAD decision:* Decisiones que han sido realizadas de forma correcta.

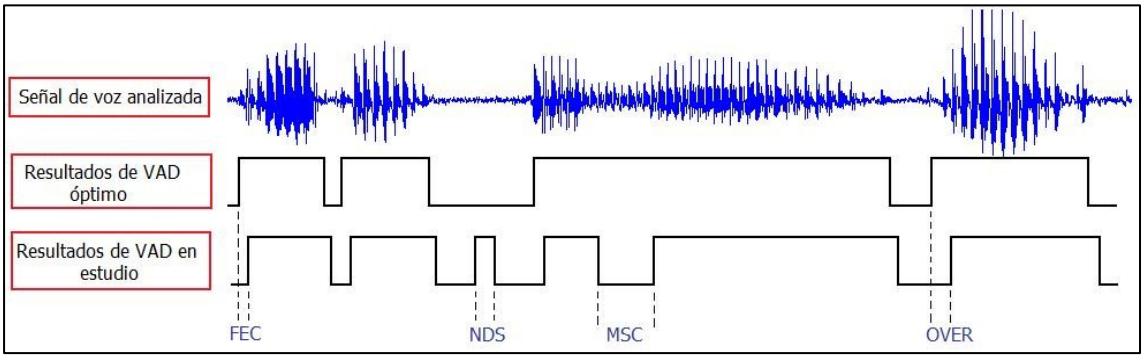

Figura 2. Representación de los parámetros de evaluación de un VAD sobre una muestra de audio de ejemplo.

Aunque con los parámetros expuestos podemos ser capaces de realizar una evaluación sobre cualquier VAD, no hay que olvidar que al tratarse de procesado de voz, también tenemos que tener en cuenta parámetros que sean capaces de medir de forma subjetiva la calidad de la clasificación. En este caso, el procedimiento es más complejo, dado que se necesitan un número mínimo de auditores que determinen aspectos clave del resultado de la clasificación por parte del VAD. En este caso, las medidas subjetivas que se lleven a cabo han de tener en cuenta:

- Calidad.
- Comprensibilidad.
- Efecto de los recortes generados

Una vez obtenidos una serie de resultados atendiendo a estos parámetros, y tras haber sido analizados un conjunto de muestras de audio procesadas con el VAD, las calificaciones resultantes de la prueba subjetiva son analizadas y ponderadas para obtener una estimación global del comportamiento del VAD. Aunque los métodos de evaluación objetivos son de gran utilidad en las fases iniciales de un análisis, los métodos subjetivos son más significativos. Aunque la aplicación de los métodos subjetivos requiere la participación de varias personas durante varios días evaluando las locuciones, solo suelen ser utilizados en los procesos de estandarización, como pueden ser los VAD's utilizados en telefonía GSM para transmisión discontinúa.

2.5.4. Evolución hasta la actualidad

En la actualidad, el desarrollo de los VAD's ha estado impulsado por su necesidad en las fases previas de todo tipo de reconocedores lingüísticos, sistemas de mejora de calidad de voz, así como para los procesos de transmisión discontinua de la voz utilizados en telefonía móvil. Esta variabilidad en cuanto a los entornos de trabajo de los VAD´s obliga a adaptar a cada una de estas aplicaciones los algoritmos propuestos.

Diversos tipos de algoritmos han sido propuestos desde que en 1959 los laboratorios Bell comenzaran a publicar sus trabajos sobre "*Time-assignment speech interpolation"* (TASI) [5], en los que aplicaban la detección de voz para realizar una multiplexación en

el tiempo del canal telefónico. La mayoría de las técnicas desarrolladas hasta ahora hacen uso de alguno de los siguientes parámetros:

- Análisis de energía en cortos periodos de señal.
- Cruces de señal por ceros.
- Análisis del tono.
- Duración de la señal.
- Codificación lineal predictiva (*Lineal predictive coding*, *LPC*)

Actualmente, estas técnicas han sido mejoradas, aunque sirven de base para los desarrollos más novedosos. El uso de LPC sigue estando muy extendido, y es la base de muchos de los algoritmos desarrollados en la actualidad. Estos nuevos desarrollos utilizan técnicas como

- Análisis de características cepstrum.
- Aplicación de la transformada wavelet.
- Modelos estadísticos de señal.
- Análisis de los coeficientes de verosimilitud (*Likelihood ratio test*, *LRT*)

A pesar de la gran cantidad de técnicas existentes desarrolladas para la detección de voz, existen muy pocos algoritmos que hayan sido estandarizados para su uso comercial. Un ejemplo de esta estandarización es la recomendación G.729 de la Unión Internacional de Telecomunicaciones (*International Telecommunication Union, ITU)*, pensada para codificación de voz en telefonía fija, en la cual, en su anexo B, se describe el VAD utilizado que da soporte a la transmisión discontinua de la voz [6]. Dicho VAD se encuentra desactualizado, y trabajos posteriores han tratado de mejorarlo. Otro organismo, en este caso el Instituto Europeo de normas de Telecomunicación *(European Telecommunications Standards Institute, ETSI),* desarrolló y estandarizó un detector de voz pensado para realizar transmisiones de tasa binaria variable adaptativa sobre canales de tráfico de voz (ETSI-AMR) [7]. Posteriormente, este mismo organismo, estandarizó otro VAD para ser aplicado sobre sistemas de reconocimiento de voz distribuido (ETSI-AFE) [8].

Capítulo 3

Diseño y Desarrollo

El principal objetivo de la reducción de ruido es eliminar la componente de señal ruidosa en la señal observada, y como ya se ha visto anteriormente, evitar en la medida de lo posible la degradación de la calidad de la señal de información, en nuestro caso, una señal de voz. Como se vio en la parte teórica, un proceso de filtrado puede provocar, no solo la reducción de ruido, sino además la aparición de distorsión en la señal, una distorsión no deseada, y que degrada la calidad de las señales.

En este capítulo, vamos a hacer hincapié en este aspecto, tratando de introducir nuevas técnicas o parámetros en la reducción de ruido, que hagan posible reducir al máximo la distorsión, y mantener el compromiso entre reducción de ruido y distorsión.

En la actualidad, las técnicas de filtrado más comunes ya han sido optimizadas hasta el límite, dentro de las posibilidades de cada una de ellas. Estas tienen en común el funcionamiento básico, que trata de observar la señal a filtrar, obtener una serie de parámetros de la misma, y aplicar el filtrado según el algoritmo. La información utilizada a la hora de decidir qué tipo de filtrado se va a aplicar es generalmente siempre la misma, por ejemplo, duración de la señal, amplitud, frecuencias, y en los casos más actuales, valores de SNR.

Por tanto, para lograr mejorar estas técnicas de filtrado, es necesario la búsqueda de otras fuentes de información, que nos aporten datos para obtener mejores resultados en la reducción de ruido. Esta nueva información, se transforma en nuevos parámetros que nos permitirán mejorar las técnicas y los algoritmos de filtrado actuales.

En concreto, en este capítulo vamos a plantear el uso y la mejora de uno de los filtros adaptativos más utilizados en el campo del procesamiento de imagen y sonido, el filtro de Wiener, que como ya hemos visto en el capítulo de fundamentos teóricos, posee una serie de características especiales que hace que pueda modificarse para lograr que la adaptación del mismo a la fuente de información sea mayor, logrando mejores resultados a la hora de aplicar la reducción de ruido.

El planteamiento que vamos a presentar consiste en realizar la adaptación del filtro de Wiener en función del contenido fonético del audio a tratar. El principal problema al que nos enfrentamos en la reducción de ruido es la distorsión provocada. Si logramos establecer qué fonemas o grupos fonéticos tienen una mayor robustez a esta distorsión,

podremos aplicar esta característica en el filtro de Wiener. Para ello analizaremos el comportamiento a nivel fonético de las muestras de audio tras el filtrado, estableceremos una regla de decisión a la hora de aplicar el filtro, y comprobaremos si los resultados de esta adaptación mejoran cualitativa y cuantitativamente la señal.

3.1. Estructura básica del filtro de Wiener

En el apartado introductorio de esta memoria hemos visto las ecuaciones a partir de las cuales se deriva el filtro de Wiener para procesado de sonido, y las posibles modificaciones/adaptaciones que pueden llegar a mejorarlo, pasando de la versión del filtro de Wiener óptimo a la versión subóptima. Si observamos con detenimiento la ecuación que define el filtro:

$$h_0 = h_1 - R_y^{-1} r_{vv}$$

podemos comprobar que éste está compuesto por dos partes bien diferenciadas. En la primera parte se genera una réplica de la señal original, mientras que en la segunda, lo que se obtiene es una estimación del ruido presente en dicha señal. De tal forma, combinando ambas partes obtenemos una versión filtrada del audio de entrada.

Es fácil entender, que la calidad del filtrado en este caso, dependerá directamente de la estimación del ruido que se realice. Si ahora observamos la definición del filtro subóptimo, el cual se supone una mejora sobre el filtro inicial

$$h_{sub} = h_1 - \alpha R_y^{-1} r_{vv}$$

lo que pretende es regular por medio del factor de reducción de ruido α la cantidad de ruido a eliminar en la señal original. Hay que tener en cuenta, que primero es necesario haber realizado una estimación de la cantidad de ruido presente en la señal, para después poder decidir que nivel de ruido vamos a eliminar.

Es por ello, que la primera fase, y una de las más importantes, es la estimación de la cantidad de ruido presente en el audio. De la calidad y precisión de esta, dependerá el resultado final.

3.1.1. Estimación de ruido

A partir de una señal de voz afectada por ruido, podemos ser capaces de realizar una estimación de la cantidad de ruido que hay presente en dicha señal. Todos los estimadores de ruido conocidos hasta la actualidad funcionan siguiendo una estructura básica:

1.- Uso de un detector de voz/no voz (VAD, *voice activity detector*).

2.- Estimación del ruido a partir de los segmentos detectados como no voz por el VAD.

Este esquema de trabajo para la estimación de ruido, goza de mucha flexibilidad, dado que generalmente los resultados arrojados por el VAD, se basan en un etiquetado o segmentación de la señal, indicando que fragmentos de la misma contienen información que puede ser considerada como voz. A este modelo de clasificación se le denomina comúnmente *"voz/no voz"*. A partir de esta segmentación, la estimación de ruido se basa en medir la cantidad y características del ruido que hay presente en los fragmentos que han sido etiquetados como *"no voz"*.

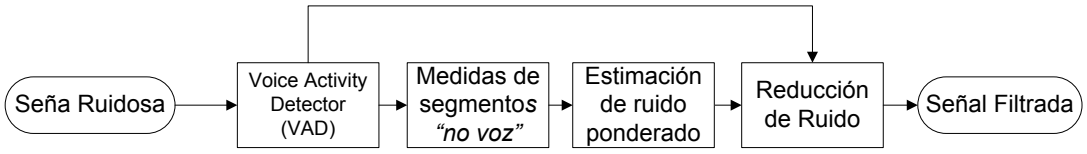

Figura 3. Esquema básico del sistema de reducción de ruido basado en VAD.

Por tanto, podemos decir que la calidad de la estimación de ruido, siguiendo este patrón de trabajo, depende directamente de la robustez de la segmentación de la voz, y de la capacidad del VAD de detectar voz en condiciones de ruido elevado.

3.2. Posibilidades de mejora del filtro de Wiener

Según se ha visto anteriormente, el papel del *detector de actividad de voz* en cualquier sistema de procesado de señales de voz es vital para su correcto funcionamiento, y la mejora de este, conlleva una mejora completa del sistema. Desde este punto de vista, y teniendo en cuenta el esquema de la Figura 3, para poder mejorar

un sistema de reducción de ruido basado en un filtro de Wiener, tenemos, entre otras, dos opciones:

- Mejorar la segmentación que realiza el VAD.
- Mejorar la estimación de ruido presente en el audio.
- Mejorar el proceso de reducción de ruido.

En la propuesta de mejora que vamos a plantear, estos son los tres aspectos que vamos a considerar más importantes. Cada uno de ellos depende directamente del anterior, por lo que es posible mejorar el rendimiento del sistema completo, optimizando cualquiera de estos tres puntos.

El primer objetivo que nos planteamos, siguiendo el orden lógico del esquema propuesto en la Figura 3, es lograr que la segmentación y clasificación de la voz realizada por el VAD se aproxime lo mayor posible a la realidad. Para esto es, es necesario incrementar la sensibilidad y robustez del VAD, y para ello, vamos a proponer un cambio sustancial en el mismo: sustituirlo por un reconocedor de voz, capaz de identificar los fragmentos de la locución que contienen voz, y discriminar aquellos en los que no se detecta su presencia. El objetivo de esta primera modificación será por tanto, mejorar la segmentación que realiza el VAD. Este cambio afecta de forma directa a la fase de estimación de ruido, haciendo que ésta sea más precisa.

El segundo punto sobre el cual vamos a basar nuestra propuesta de mejora es en el proceso de reducción de ruido. En esta fase, el mayor problema al que nos enfrentamos es la distorsión de la señal de voz: cuanto más ruido eliminamos, mayor es la distorsión provocada. Es por ello que la mejora que vamos a proponer y con la que esperamos mejores resultados, no se basa solamente en la cantidad de ruido a eliminar, sino también en cómo se elimina ese ruido.

Nos basamos en que la distorsión no afecta de igual manera a todas las señales de voz, ni siquiera en la misma locución. Vamos a suponer que ésta depende del contenido fonético de la señal de voz, es decir, que determinados fonemas o grupos de fonemas son más sensibles a los efectos de la distorsión, mientras que otro grupo de fonemas presentan una mayor robustez. Teniendo esto en cuenta, si podemos conseguir aplicar distintos niveles de reducción de ruido en función del contenido fonético de la señal de

voz, podremos adaptar el filtro de Wiener, y por tanto, mejorar los resultados finales del sistema de reducción de ruido.

A continuación vamos a analizar en profundidad cada una de estas mejoras, y su posible aplicación, así como una propuesta de desarrollo de las mismas.

3.2.1. Sustitución del VAD por un reconocedor fonético

Uno de los principales inconvenientes que presentan los VAD's convencionales es su escasa robustez al ruido y su mal funcionamiento con valores bajos de SNR. En condiciones de ruido agresivo, el comportamiento del detector comienza a ser irregular, perdiendo precisión, y etiquetando fragmentos de ruido como si fueran voz. Para poder solventar este problema, necesitamos que el VAD disponga de más información de la señal de voz a partir de las características de la misma, en lugar de basarse en la medida de los niveles medios de energía, como ocurre en los VAD básicos.

Una solución que vamos a plantear en este PFC es la sustitución del VAD por un *reconocedor de voz*. En este caso, el reconocedor etiqueta cada fonema pronunciado, así como las pausas, o segmentos no reconocidos. Analizando los resultados obtenidos a partir del reconocedor, etiquetamos cada fonema reconocido como voz, y el resto de fragmentos como silencios. De esta forma tendríamos un audio analizado y etiquetado con sus correspondientes fragmentos de voz y no voz, de la forma que hemos descrito anteriormente, tal como queríamos.

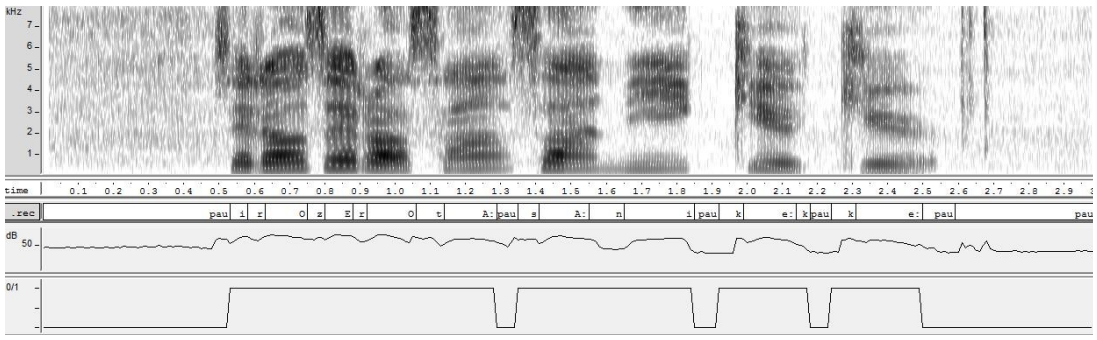

Figura 4. Espectrograma, transcripción, forma de onda, y segmentación voz/no-voz de un audio de ejemplo

Atendiendo al tipo de reconocedores que existen, vamos a plantear el uso de un reconocedor fonético, dado que aporta gran cantidad de información para su procesado

posterior, y además, permite que el sistema sea razonablemente independiente del idioma, el principal inconveniente a la hora de utilizar un reconocedor de voz como un detector de voz.

La información obtenida con el reconocedor fonético en el VAD podrá ser utilizada posteriormente, para aplicar la propuesta de mejora en la fase de reducción de ruido que se ha planteado en el punto anterior.

3.2.2. Efectos negativos del filtrado: la distorsión

En la definición del filtro de Wiener que hemos realizado anteriormente, hemos podido comprobar cómo éste, independientemente de la implementación que utilicemos, distorsiona la señal de voz al llevar a cabo la reducción de ruido. Para poder cuantificar la distorsión que se genera en la señal de voz, el parámetro denominado **distorsión de atenuación en frecuencias** (*AFD, attenuation frequency distortion*)

$$\Phi_{sd}(\omega) \triangleq \frac{E\,[|X(i\omega)|^2 - |\hat{X}_{nr}(i\omega)|^2]}{E[|X(i\omega)|^2]} = \frac{P_x(\omega) - P_{\hat{x}_{nr}}(\omega)}{P_x(\omega)}$$

nos permite obtener un valor numérico que indica como de distorsionada está una señal de voz que previamente ha sido filtrada. Para poder hacer uso del AFD y evaluar cómo de agresiva es la distorsión para nuestro filtro, tendremos que hacer una serie de consideraciones prácticas previas a su implementación.

En primer lugar, si observamos la ecuación (10) a partir de la cual se obtiene el valor del AFD para un audio dado, necesitamos disponer de las señales $X(i\omega)$ y $P_x(\omega)$ que son el espectro y la densidad espectral de potencia del canal de voz limpio *x(n)* y las señales $\widehat{X}_{nr}(i\omega)$ y $P_{\hat{x}_{nr}}(\omega)$ que representan el espectro y la densidad espectral de potencia de la componente de voz de la señal filtrada. En un entorno real, realizar la medida del AFD es tarea bastante compleja, dado que no contamos con la señal original, y por tanto no podemos realizar la comparación con la señal filtrada.

En un entorno de test, como es el nuestro, la solución a ese problema es relativamente sencilla. En este caso, la base de datos de audios con la que vamos a realizar las pruebas cuenta con grabaciones obtenidas de forma simultánea a través de

dos tipos distintos de micrófonos; un micrófono de habla cercana (*close talking* o **CT**) y otro micrófono manos libres (*hands free* o **HF**). Podemos aprovechar esta concurrencia para asumir que las grabaciones obtenidas a través del micrófono "*close talking"* son las señales originales y las obtenidas a través del micrófono *hands free* son las señales ruidosas, sobre las cuales será necesario aplicar la reducción de ruido, para posteriormente, evaluar el factor AFD.[1]

La segunda consideración a tener en cuenta antes de proceder a realizar las medidas de los factores AFD es cómo vamos a evaluar los audios, de manera que los resultados aporten información significativa a la hora de plantear mejoras. De esta manera se plantean diversas alternativas, que aportan resultados distintos.

- Aplicar el algoritmo de medida sobre la duración completa del audio.
- Segmentar el audio en fonemas, aplicando el algoritmo a cada fonema de forma independiente.

Cada una de estas alternativas analiza las grabaciones desde un nivel de profundidad distinto, y sus resultados han de ser interpretados de forma totalmente distinta. Veamos qué ventajas e inconvenientes presenta cada uno de los métodos de cálculo.

La primera de las opciones nos da una idea general de cómo de distorsionado está un audio. Este dato no es útil para los objetivos que nos hemos planteado, dado que lo que se pretende es poder estimar un patrón de comportamiento de la distorsión tras el filtrado y poder predecirlo para contrarrestarlo en todos los audios, no exclusivamente en uno de ellos, pero si puede ser utilizado para determinar que opción de filtrado funciona mejor a nivel global, comparando valores para distintos filtros.

La segunda opción de cálculo del factor de AFD profundiza aún más en el análisis del audio, acudiendo directamente al nivel fonético. En este caso, lo que se trata es comparar los fragmentos del audio fonema a fonema. De esta forma, podemos crear un diccionario fonético con los valores correspondientes de AFD y ser capaces de predecir cómo va a ser la distorsión de la señal de voz tras el filtrado. A simple vista, podría

[1] Hay que destacar que esta consideración solo la tendremos en cuenta para el cálculo del AFD, dado que se trata de una medida comparativa. En las pruebas de filtrado, se considera que todos los audios han sido perturbados en mayor o menor medida por el ruido aditivo, independientemente de que se traten de grabaciones simultáneas.

parecer que este método no es independiente del idioma, ya que idiomas distintos presentan un conjunto de fonemas distinto, pero comparten una serie de características a nivel fonológico que es posible explotar para lograr que esta forma de comparar los audios sea razonablemente independiente del idioma, y que posteriormente procederemos a estudiar.

3.3 Estudio de la distorsión a nivel fonético

De todas las opciones que hemos planteado anteriormente para analizar los efectos de distorsión que introduce el filtrado, el estudio del AFD a nivel fonético es el que más ventajas nos aporta, dado que será a partir de este análisis desde donde podremos comprobar, según nuestra hipótesis inicial, que el valor del AFD es distinto para cada fonema o grupo fonético, y además detectar que fonemas o grupos de fonemas presentan mayor sensibilidad a la distorsión. Aparte de esto hay que tener en cuenta que este método es más flexible en cuanto al idioma, dadas las similitudes existentes a nivel fonético entre distintos idiomas.

Para poder calcular el AFD a nivel fonético dentro del sistema de reducción de ruido que hemos planteado, lo primero que tenemos que hacer es definir el diccionario de fonemas sobre el cual vamos a trabajar. Para poder definirlo de forma correcta y ajustada, tenemos que tener en cuenta dos aspectos muy importantes:

- ✓ El idioma de las locuciones de origen.
- ✓ El idioma del reconocedor fonético utilizado.

En este caso, como veremos posteriormente donde se describe el entorno experimental de este PFC, tenemos que el idioma de las locuciones de origen es el japonés y el idioma del reconocedor fonético es el húngaro. Lograr una compatibilidad entre ambos idiomas a la hora de calcular el AFD no sería posible sin la segmentación a nivel fonético que hemos planteado, y la agrupación de los distintos fonemas en clases amplias fonéticas, dado que a ese nivel, las diferencias entre idiomas diferentes son mínimas.

Para poder equiparar a nivel fonético ambos idiomas, vamos a utilizar el alfabeto fonético internacional, que define de forma independiente al idioma, todos los fonemas

existentes que se utilizan en la comunicación oral. De esta manera, y utilizando un único alfabeto, podremos comparar el húngaro y el japonés de forma bidireccional.

3.3.1. El Alfabeto Fonético Internacional

El Alfabeto Fonético Internacional (o *International Phonetic Alphabet, IPA*) es una herramienta creada por la *International Phonetic Association* (*IPA*) [9] para promover el estudio de la ciencia fonética y su uso como apoyo a otras ciencias. Su objetivo es representar de forma consistente los distintos sonidos que componen el lenguaje hablado, de forma escrita. Con este alfabeto somos capaces de representar la pronunciación de cualquier palabra, y de cualquier idioma.

CONSONANTES © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	N		
Trill	B			r					R		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Tabla 1. Tabla fonética del Alfabeto Fonético Internacional, indicando el modo de articulación y el punto de articulación de cada fonema de carácter consonántico.

CONSONANTES (NO PULMONARES)

Chasquidos		Implosivas		Eyectivas	
ʘ	Bilabial	ɓ	Bilabial	’	Ejemplos:
ǀ	Dental	ɗ	Dental/alveolar	pʼ	Bilabial
ǃ	(Post)alveolar	ʄ	Palatal	tʼ	Dental/alveolar
ǂ	Palatoalveolar	ɠ	Velar	kʼ	Velar
ǁ	Alveolar lateral	ʛ	Uvular	sʼ	Alveolar fricative

Tabla 2. Tabla fonética del Alfabeto Fonético Internacional con los sonidos consonánticos no pulmonares.

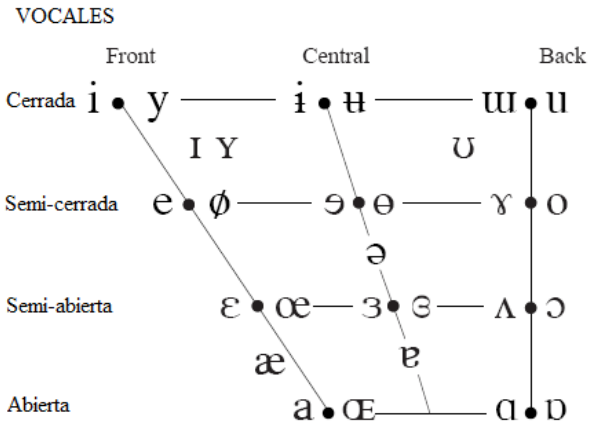

Figura 5. Diagrama del punto de articulación de los sonidos vocálicos definidos por el IPA.

La clasificación que establece la IPA en este alfabeto, ordena los distintos fonemas según el modo de articulación y el punto de articulación de dicho sonido en el tracto vocal humano. Asimismo, los distintos sonidos existentes se clasifican en función de su naturaleza consonántica o vocálica.

Los fonemas consonánticos

En la articulación de los sonidos consonánticos siempre hay un obstáculo más o menos grande que impide salir el aire desde los pulmones al exterior. Según las circunstancias que rodean esta salida del aire, existen ciertos factores que debemos tener en cuenta a la hora de clasificarlos:

- ✓ ***Zona o punto de articulación.*** Es el lugar donde toman contacto los órganos que intervienen en la producción del sonido. Por ejemplo, si para producir un sonido entran en contacto los dos labios, se crearán sonidos bilabiales como es el caso de las realizaciones de los fonemas /p/, /b/ y /m/.

- ✓ ***Modo de articulación.*** Es la postura que adoptan los órganos que producen los sonidos. Por ejemplo, si los órganos cierran total y momentáneamente la salida del aire, los sonidos serán plosivos. Ese es el caso de los sonidos /p/, /t/ y /k/.

Dentro de los fonemas consonánticos, existen dos clases distintas de sonidos, los denominados "*pulmonares*" y los "*no pulmonares"*. En el caso de las consonantes pulmonares, se utilizan los pulmones para impulsar el aire hacia el exterior. En el caso de las no pulmonares, el aire es impulsado desde la glotis, o son simples chasquidos. En la mayoría de los idiomas, solo se utilizan las consonantes pulmonares, siendo el japonés y el húngaro ejemplos de ello.

Los fonemas vocálicos

Cuando articulamos los sonidos vocálicos, el aire no encuentra obstáculos en su salida desde los pulmones al exterior. Para clasificar estos fonemas, tendremos en cuenta los siguientes factores:

- ✓ *La* ***localización (punto de articulación).*** Se refiere a la parte de la boca donde se articulan. Pueden ser anteriores (/e/, /i/), medio o central (/a/) o posteriores (/o/, /u/).

- ✓ *La* ***abertura (modo de articulación).*** Se refiere a la abertura de la boca al pronunciarlos. Pueden ser de abertura máxima o abierto (/a/), de abertura media o semiabiertos (/e/, /o/) y de abertura mínima o cerrados (i, u).

3.3.2. Agrupación de los fonemas en clases amplias fonéticas

Como hemos comentado anteriormente, para poder realizar una comparación entre distintos idiomas a nivel fonético, es necesario que definamos las *clases amplias fonéticas.* Esta agrupación de fonemas se basa en la clasificación de los mismos según el modo de articulación, en el caso de los sonidos consonánticos. Los sonidos vocálicos, dada su gran diversidad, se han agrupado en una sola clase fonética.

Atendiendo a esta forma de agrupar los fonemas, se definen las siguientes clases fonéticas [10]:

Consonantes Oclusivas

El flujo de aire es retenido firmemente por los órganos del habla, hasta que este es liberado, generando de esta forma el sonido deseado. Ejemplos de estos fonemas en el castellano son [p] [t] [k]

Consonantes fricativas

El aire ha de atravesar una estrecha obertura formada por los órganos del habla, generando el sonido gracias a la fricción que se produce en el tracto vocal. Ejemplos de estos sonidos pueden ser [f] o [s].

Consonantes africadas

Es una combinación de una consonante oclusiva, seguida de una consonante fricativa. Para ello, el aire es retenido por los órganos vocales, para ser liberado posteriormente de forma paulatina. En castellano, un sonido africado se produce con el fonema [tʃ], utilizado en *"**ch**ubasquero"*.

Consonantes Aproximantes

En este caso, el sonido es producido por la aproximación de los órganos vocales sin llegar a cerrarse (como en el caso de las oclusivas) y sin la existencia de fricción aérea (como en las consonantes fricativas). Se encuentran muy relacionadas con los sonidos vocálicos, sin llegar a ser considerados como tal. Un ejemplo de estos fonemas es [j] utilizado en *"famil**i**a"* o *"chirimo**y**a"*.

Consonantes Aproximantes Laterales

Estas consonantes son un subgrupo de las aproximantes. Se consideran aproximantes laterales aquellos sonidos que son formados por la aproximación de la lengua y el paladar superior o los dientes. Un fonema aproximante lateral es [l].

Consonantes nasales

Este tipo de sonido es generado cuando el flujo de aire, incapaz de atravesar los orificios orales, es desviado hacia la cavidad nasal, generando ese sonido tan característico. Ejemplo de consonantes nasales son [m] o [n].

Consonantes vibrantes

Son sonidos generados gracias cuando uno de los órganos vocales golpea de forma rápida y repetitiva sobre el otro, mientras el flujo de aire atraviesa la cavidad. Un ejemplo de estos es el fonema [r].

3.3.3. Correspondencia IPA-SAMPA

El reconocedor húngaro que vamos a utilizar [11] para segmentar los audios dispone de un diccionario de fonemas, que engloba todos aquellos sonidos que éste va a ser capaz de reconocer. Dichos fonemas están expresados en el alfabeto SAMPA [12] (*Speech Assessment Methods Phonetic Alphabet*), el más utilizado en procesamiento fonético por ser totalmente legible por un ordenador.

Grupo Fonético	Fonema
Vocal	A: E e: i i: O o o: u u: y y: :2 _2
Oclusiva	b b: d d_ d_: g k k: p t t: t1 t1:
Fricativa	f h h1 S S: s s: v x Z z z:
Africada	dz tS tS_ ts ts_
Nasal	F J J: m m: N n n:
Aproximante	j j:
Aproximante lateral	l l:

Tabla 3. Diccionario de fonemas reconocibles por el reconocedor húngaro **Phnrec.**

La tabla de fonemas reconocibles equivalente al IPA la podemos obtener hallando la correlación de cada uno de los fonemas en los dos estándares SAMPA que se utilizan en la actualidad (SAMPA y X-SAMPA). En este caso, se ha utilizado el traductor del Laboratorio de Fonética Experimental "*Arturo Genre*" de la Universidad de Turín conjuntamente con el IPA de la edición del año 2005 [13]

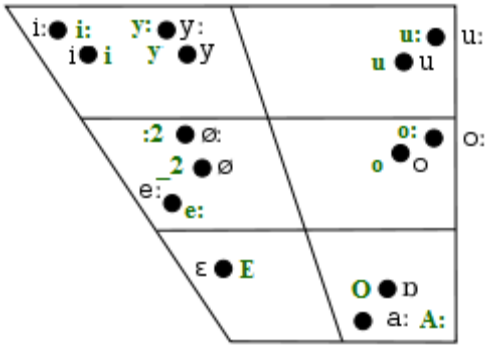

Figura 6. Diagrama de correspondencia de sonidos vocálicos entre el diccionario de Phnrec e IPA. En verde los fonemas de Phnrec.

Oclusivas	
SAMPA	**IPA**
b b:	[b] [b:]
d d_ d_:	[d] [d̪] [d̪:]
g	[g]
k k:	[k] [k:]
P	[p]
t t: t1 t1:	[t] [t:] [t̪] [t̪:]

Fricativas	
SAMPA	**IPA**
f	[f]
h h1	[h] [ɦ]
S S: s s:	[ʃ] [ʃ:] [s] [s:]
v	[v]
x	[x]
Z z z:	[ʒ] [z] [z:]

Nasales	
SAMPA	**IPA**
F	ɱ
J J:	ɲ ɲ:
m m:	m m:
N	ŋ
n n:	n n:

Africadas	
SAMPA	**IPA**
dz	dz
tS tS_	[tʃ] [t͡ʃ]
ts ts_	[ts] [t͡s]

Aproximantes lateral	
SAMPA	**IPA**
l l:	l l:

Aproximantes	
SAMPA	**IPA**
j j:	j j:

Tabla 4. Tablas de correspondencia de sonidos consonánticos entre el diccionario de Phnrec e IPA.

3.4. Aplicación del condicionamiento fonético

Una vez que tenemos definido como vamos a agrupar los distintos fonemas en lo que hemos denominado "clases amplias fonéticas", estamos en disposición de poder estudiar cómo afecta la distorsión generada por el filtro de Wiener sobre estos grupos de fonemas. Conocer qué grupos se ven más afectados por la distorsión generada en la etapa de filtrado, es una ventaja a la hora de contrarrestar estos efectos sobre los audios que pretendemos tratar, que es el objetivo que estamos persiguiendo.

3.4.1. Cálculo de AFD a nivel de clase amplia fonética

Partiendo de la ecuación de cálculo del AFD, y teniendo en cuenta las consideraciones prácticas para su aplicación indicadas en el punto 3.3.2., la obtención de los valores del AFD a nivel fonético se resume en implementar dicho algoritmo y aplicarlo sobre el conjunto de locuciones de CENSREC-2 que presentan simultaneidad,

es decir, locuciones grabadas en el mismo instante, pero obtenidas con distintos micrófonos.

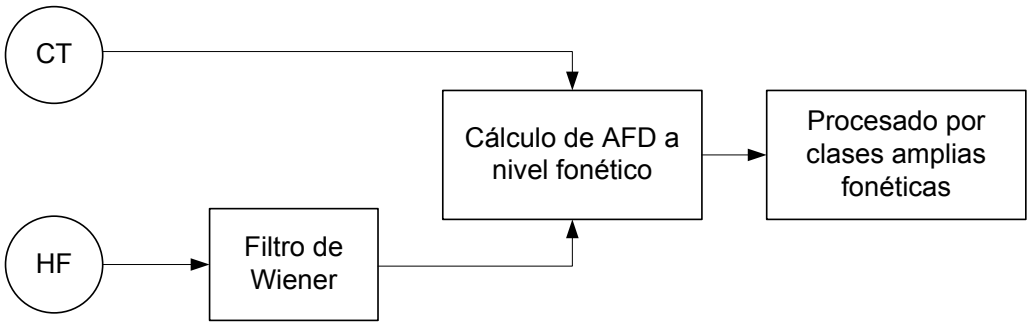

Figura 7. Obtención del AFD de las distintas clases amplias fonéticas a partir de los conjutnos de locuciones CT y HF.

Para la implementación del algoritmo y su aplicación, se ha desarrollado un conjunto de scripts en Matlab que permite el fácil procesado de los resultados que se obtengan.

Grupo fonético	AFD (Media)	Desviación estándar
Vocal	1.115	0.265
Oclusiva	12.691	19.052
Fricativa	1.968	1.007
Africada	4.482	0.684
Nasal	1.339	0.244
Aproximante	1.106	0.152
Aproximante Lateral	1.437	0.223

Tabla 5. Valor medio obtenido del AFD y su correspondiente desviación estándar para cada clase fonética.

Con los resultados de la tabla 5 obtenidos tras el cálculo de los valores del AFD, se confirma que las distintas clases fonéticas que se han definido presentan un comportamiento distinto frente al filtrado de Wiener, tal y como supusimos al principio. A partir de aquí adaptar estos resultados al modelo de filtrado de Wiener subóptimo definido en el estado del arte es muy sencillo, dado que en su definición ya se incluye un factor de ponderación de la cantidad de ruido que se pretende eliminar.

$$h_{sub} = h_1 - \alpha R_y^{-1} r_{vv}$$

En el caso del filtro de Wiener definido en el dominio temporal, se introduce el factor α, que pondera la cantidad de ruido a eliminar. Para el caso que nos ocupa, la relación entre α y el AFD, es inversamente proporcional, esto es, cuanto mayor sea el AFD de un fragmento de la grabación, menos deberá ser la cantidad de ruido a eliminar, y por tanto, menor deberá ser el valor de α.

3.5. Entorno Experimental

3.5.1. Implementación del filtro de Wiener utilizada

Como ya se ha dicho anteriormente, para llevar a cabo la reducción de ruido se ha utilizado una implementación del filtro de Wiener. Dado que el objetivo del proyecto no es el desarrollo de software propio, sino la mejora de los ya existentes, hemos optado por elegir una implementación que trabaja con el filtro de Wiener en su versión definida en frecuencia.

Dicha implementación forma parte de *"Qualcomm-ICSI-OGI front-end feature extraction"* desarrollado en 2002 y propuesto para la evaluación WI008 [14]. En este se hace uso de un filtrado de Wiener previo al procesado de la señal para eliminar el ruido aditivo presente en la misma. Dicho filtro está definido en el dominio de la frecuencia y tiene la forma:

$$|H_r(k,n)|^2 = \max\left(\frac{|X(\omega_i,t)|^2 - \alpha|\hat{N}(\omega_i,t)|^2}{|X(\omega_i,t)|^2}, \beta\right)^{\gamma} \quad (73)$$

Esta definición del filtro coincide con la que ya se vio anteriormente (71), que se corresponde con la forma generalizada del filtro de Wiener paramétrico. En esta forma de definir el filtro existen tres parámetros. El primero de ellos, "α" es el factor de sobreestimación de ruido, y se utiliza para corregir la cantidad de ruido a eliminar. El valor de α depende del valor de la SNR local de la ventana donde nos encontremos, eliminando más ruido para valores de SNR bajos y viceversa. Este factor de sobreestimación se obtiene a partir de:

$$\alpha = \frac{-1.875}{20} PosteriorSNR_n + 3.125 \quad (74)$$

donde

$$PosteriorSNR_n = 10 * log10\left(\frac{FrameEnergy_n}{NoiseEnergy_n}\right) \quad (75)$$

Este factor de sobreestimación de ruido, tal y como está definido anteriormente, está comprendido entre los valores [1.25, 3.125] dado que la SNR máxima que se considera es de 20dB y la mínima de 0dB, como se puede ver en la figura 8. El segundo parámetro que aparece en la definición del filtro,"β", sirve para definir un valor mínimo en la función de transferencia del filtro, evitando así valores negativos o demasiado bajos. El último parámetro del filtro,"γ", se utiliza para controlar el comportamiento del filtro, es decir, para $\gamma = 1$ el filtro aplica substracción espectral, para $\gamma = 2$ se aplica filtrado de Wiener, pudiendo usarse otros valores, aparte de los ya mencionados.

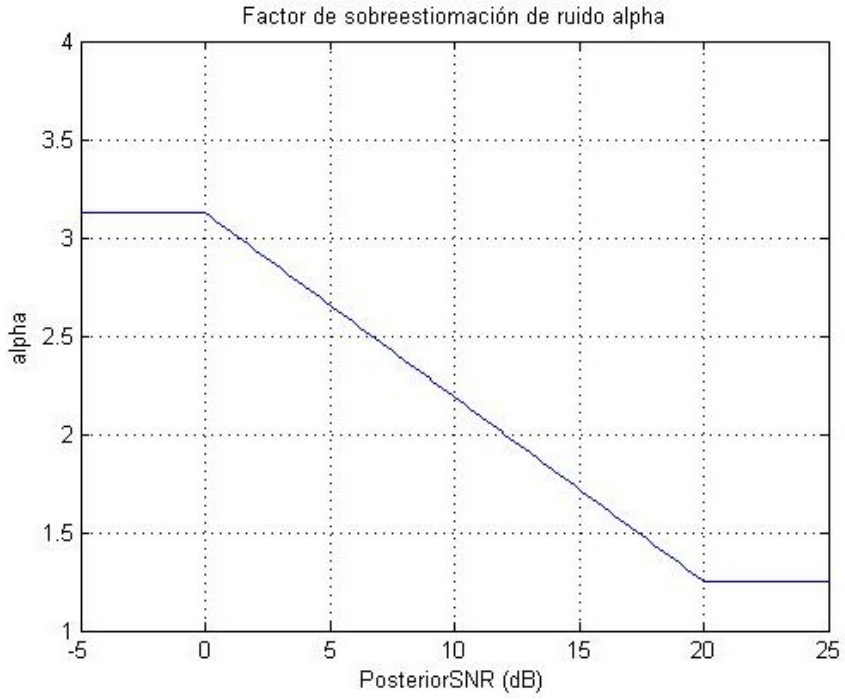

Figura 8. Valor del factor de sobreestimación de ruido en función de la SNR calculada para QIO.

Valores recomendados de los parámetros

El grupo de desarrollo del filtro pudo comprobar a través de los experimentos realizados con la base de datos Aurora Speech-Dat Car, que los valores que se muestran a continuación son los más recomendados para obtener los mejores resultados, con la opción del filtro propuesto.

- $\alpha = [1.25,3.125]$
- $\beta = 0.01$
- $\gamma = 2$

3.5.2. Base de datos sonora utilizada

Para poder evaluar la técnica de reducción de ruido planteada en este PFC, es necesaria la realización de diversos experimentos utilizando para ello grabaciones de voz. Para verificar el correcto funcionamiento y aplicación de la técnica, dichas grabaciones han de cumplir con una serie de requisitos. En nuestro caso, necesitamos un entorno acústico agresivo, en condiciones de ruido variable, donde la reducción de ruido sea una herramienta necesaria que ayude a la comprensión de los audios o bien por parte de un oído humano o bien por parte de un sistema reconocedor de voz. En esta misma línea de trabajo se han desarrollado diversos corpus orientados al procesamiento de voz en condiciones acústicas ruidosas, incluyendo en ellos los mecanismos de evaluación, que permitan identificar las mejoras realizadas.

La base de datos que hemos utilizado y que cumple con las condiciones planteadas es CENSREC-2. Dicho corpus ha sido desarrollado por el *IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group* y está pensado para el reconocimiento de voz de dígitos en condiciones de conducción real, utilizando el japonés. Consta de un total de 17.651 locuciones, grabadas por 104 personas, 52 mujeres y 52 hombres [15].

Disposición de micrófonos

Las locuciones grabadas para CENSREC-2 fueron tomadas usando dos tipos de micrófonos distintos, uno de ellos de habla cercana (*close talking*, *CT*) y el otro de manos libres (*hands free, HF*). El micrófono HF fue colocado en la zona del techo

correspondiente al conductor, mientras que el micrófono CT se dispuso lo más cerca posible a la boca del conductor.

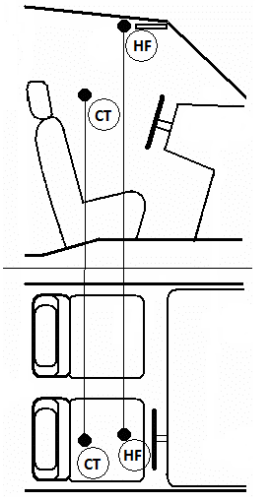

Figura 9. Ubicación de los micrófonos en el interior del vehículo.

El modelo y marca de los micrófonos utilizados fue en ambos casos la misma, para uniformizar los resultados de la obtención de los datos. En este caso se utilizaron micrófonos Sony ECM77B, uno de ellos (CT), montado sobre unos auriculares para el conductor del coche.

Vocabulario utilizado

El vocabulario de CENSREC-2, como base de datos de dígitos, consta de once modelos de dígitos distintos, correspondientes a los diez primeros números, incluyendo dos pronunciaciones para el cero. Además, fueron definidos un silencio ('sil') y una pausa corta ('sp'). La secuencia de dígitos utilizada en las grabaciones es la misma que la utilizada en la AURORA-2J (versión en japonés de AURORA-2).

Número	1	2	3	4	5	6	7	8	9	0
Pronunciación	Ichi	Ri	San	Yon	Go	Roku	Nana	Hachi	Kyu	Zero

Tabla 6. Lista de dígitos y pronunciación utilizados en CENSREC-2.

Condiciones de grabación

Se utilizaron once condiciones de grabación distintas, como resultado de combinar tres velocidades de movimiento del vehículo (*idling*, *low-speed* y *high-speed*),

representando cada uno de los entornos en los que un coche se puede desplazar, y cuatro tipos de condiciones acústicas en el interior del coche (normal, aire acondicionado encendido, reproductor de CD's encendido y ventanas abiertas).

Velocidad de Coche	Condiciones interiores
Parado (ralentí)	Normal, Aire acondicionado, Reproductor CD, Ventana abierta
Baja Velocidad	Normal, Aire acondicionado, Reproductor CD, Ventana abierta
Alta Velocidad	Normal, Aire acondicionado

Tabla 7. Combinación de velocidades y condiciones acústicas en el vehículo.

Protocolo de evaluación de resultados

Para poder evaluar el procesamiento aplicado sobre los audios de la base de datos, los desarrolladores de la base de datos han definido un protocolo de evaluación, basado en el reconocimiento de voz de las locuciones, que previamente han sido etiquetadas con la transcripción de su contenido.

Dicha evaluación consta de cuatro condiciones acústicas, que se forman a partir de las combinaciones entre velocidad y condiciones internas del vehículo para las distintas fases del reconocimiento, que comprenden un entrenamiento del reconocedor (*train*) y la prueba de reconocimiento (*test*). Las características de cada condición de *train* y *test* son las siguientes:

- ✓ **Condición 1:** Las grabaciones usadas para el entrenamiento y el test se tomaron con el mismo micrófono en las mismas condiciones acústicas.
- ✓ **Condición 2:** Las grabaciones usadas para el entrenamiento y test se tomaron utilizando el mismo micrófono en distintas condiciones acústicas.
- ✓ **Condición 3:** Las grabaciones usadas para el entrenamiento y test se tomaron utilizando distintos micrófonos en mismas condiciones acústicas.
- ✓ **Condición 4:** Las grabaciones usadas para el entrenamiento y test se tomaron utilizando distintos micrófonos en distintas condiciones acústicas.

Condición Micrófono	Cond. 1 CT	Cond. 1 HF	Cond. 2 CT	Cond. 2 HF	Cond. 3 CT	Cond. 3 HF	Cond. 4 CT	Cond. 4 HF
Parado (ralentí)	-	✓	-	✓	✓	-	✓	-
Baja velocidad	-	✓	-	-	✓	-	-	-
Alta velocidad	-	✓	-	-	✓	-	-	-

Tabla 8. Datos entrenamiento para cada condición de evaluación.

Condición	Cond. 1	Cond. 2	Cond. 3	Cond. 4
Parado (ralentí)	✓	-	-	-
Baja velocidad	✓	✓	✓	✓
Alta velocidad	✓	✓	✓	✓

Tabla 9. Datos test para cada condición de evaluación.

Software utilizado

El protocolo de evaluación de resultados que acabamos de explicar se aplica haciendo uso de un reconocedor de voz. En el caso que nos ocupa, el reconocedor utilizado, y recomendado por los creadores de la base de datos es HTK (*Hidden Markov Model Toolkit*) [16], un software de investigación en reconocimiento de voz de uso libre y que goza de gran reputación en el procesamiento de señales de voz. Este grupo de herramientas puede reconocer el contenido de una locución, basando su criterio de decisión en un conjunto de parámetros extraídos del audio, a partir de los cuales, es capaz de realizar una transcripción del audio analizado.

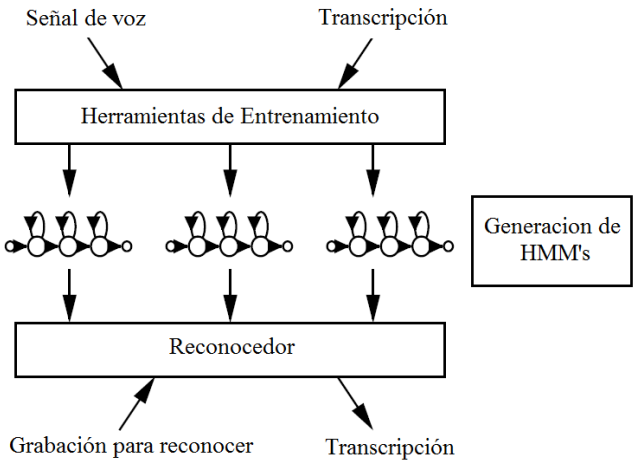

Figura 10. Funcionamiento general de HTK

Para poder llevar a cabo este reconocimiento, es necesario que el paquete de herramientas haya sido entrenado previamente con un conjunto de muestras de audio que representen todos los fragmentos reconocibles de una locución, es decir, que sea capaz de completar todo el diccionario de elementos que van a ser reconocidos posteriormente. Esta fase es conocida como "*Training*", y es la encargada de crear los modelos (*HMM's* o *Hidden Markov Models*) que van a ser la base para el posterior reconocimiento. En la fase de entrenamiento, son necesarios dos elementos indispensables, las muestras de audio de ejemplo y la transcripción del contenido de estas muestras. A través de la combinación de ambas, se realizarán los HMM's que relacionan una serie de parámetros de la voz con un elemento reconocible (en nuestro caso, números).

Los ficheros de configuración para realizar el reconocimiento están disponibles junto con las locuciones de CENSREC-2, pudiendo ser modificados en función de los parámetros que se quieran establecer para el entrenamiento y test. En el caso de utilizar la configuración recomendada, el reconocimiento es llevado a cabo con las siguientes características:

- ✓ El reconocimiento de voz se lleva a cabo haciendo uso de los HMM's (*Hidden Markov Model*) generados en la fase de entrenamiento. Son modelos estadísticos basados en procesos de Markov.
- ✓ Cada modelo HMM de cada dígito está compuesto de 18 estados con 16 distribuciones de salida, el silencio 'sil' tiene cinco estados con tres salidas, y la pausa 'sp' tiene tres estados con una salida.
- ✓ Para la extracción de características se han utilizado 12 MFCC's (*Mel-frequency cepstral coefficients*), coeficientes basados en el modelo de percepción auditiva humana, y coeficientes log-energia. Con ventana de Hamming de 20ms, con 50% de superposición. También se utilizan los coeficientes de velocidad (Delta) y aceleración (Delta-Delta) de dichos coeficientes.

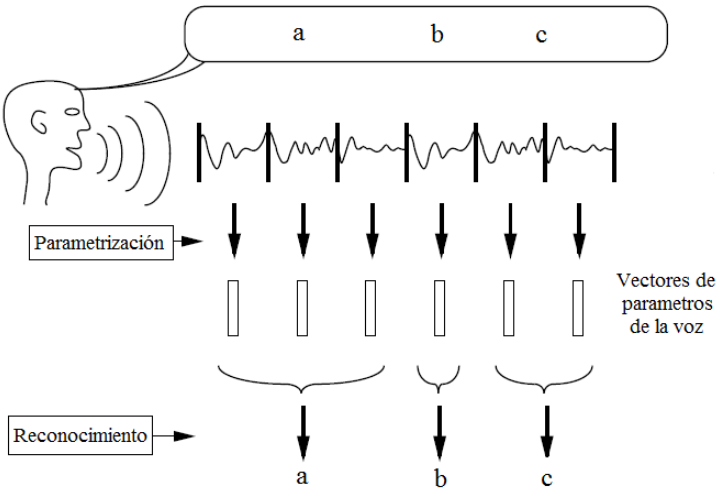

Figura 11. Esquema básico de reconocimiento de HTK

Para poder verificar que la configuración utilizada es correcta, junto con la base de datos se incluyen los resultados de referencia del proceso de reconocimiento.

CENSREC-2 Resultados de referencia (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,58	74,49	61,46	48,87	66,35

Tabla 10. Resultados de referencia proporcionados por CENSREC-2.

En la Tabla 10 están representados los resultados de referencia del reconocimiento de la base de datos original. En ella, los datos que se reflejan indican el porcentaje de palabras reconocidas con éxito en cada una de las condiciones definidas en el protocolo de evaluación.

Para comprobar la calibración de nuestro sistema de reconocimiento y la configuración, hemos realizado la misma prueba con las muestras de sonido originales, y comparar así nuestros resultados con los aportados por la base de datos.

CENSREC-2 Resultados obtenidos (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,64	74,57	61,53	49,05	66,45

Tabla 11. Resultados obtenidos en el reconocimiento de las muestras originales de la base de datos.

Como se puede observar en la Tabla 11, los resultados obtenidos en nuestro caso son muy similares a los de referencia, por lo que podemos considerar que la configuración del reconocimiento es correcta e igual a la recomendada.

3.5.3. Reconocedor de voz empleado como VAD

Para poder probar la efectividad de las modificaciones y mejoras planteadas anteriormente sobre el sistema inicial, es necesario el uso de un reconocedor de voz que sustituya al detector de actividad de voz. Con este propósito hemos hecho uso de un reconocedor fonético desarrollado por el "Speech Processing Group", perteneciente a la Universidad tecnológica de Brno, en la Republica Checa. Dicho reconocedor está pensado para trabajar con varios idiomas, como son el checo, húngaro, ruso e inglés [11].

Este reconocedor representa un gran potencial, ya que según los mismos autores y desarrolladores, está siendo utilizado en multitud de aplicaciones, tales como:

- Reconocimiento de idioma.
- Reconocimiento de voz de amplio vocabulario.
- Búsqueda de palabras clave.
- Detección de la actividad de voz.

En el área que nos interesa, la detección de actividad de voz, el reconocedor ya ha sido utilizado para la fase de pre procesado en las evaluaciones del NIST, obteniendo buenos resultados.

Para nuestra implementación, el idioma escogido para el reconocimiento es el húngaro, dado que es el que presenta un diccionario fonético más amplio, lo que nos

facilita la tarea de hacer el sistema independiente al idioma, es decir, cuantos más fonemas seamos capaces de reconocer, mayor capacidad de adaptación a un idioma distinto presenta. En el caso del húngaro, su diccionario dispone de 56 fonemas distintos. En la tabla 12 se encuentran todos los fonemas del diccionario, clasificados en función de su modo de articulación y expresados en formato SAMPA (*Speech Assessment Methods Phonetic Alphabet*), definiendo de esta manera las clases amplias fonéticas que vamos a utilizar de aquí en adelante.

Grupo Fonético	Fonema
Vocal	A: E e: i i: O o o: u u: y y: :2 _2
Oclusiva	b b: d d_ d_: g k k: p t t: t1 t1:
Fricativa	f h h1 S S: s s: v x Z z z:
Africada	dz tS tS_ ts ts_
Nasal	F J J: m m: N n n:
Aproximante	j j:
Lateral	l l:

Tabla 12. Conjunto de fonemas del diccionario del reconocedor para el húngaro. Los fonemas están presentados en formato SAMPA, para uso con computadores.

Motivos de la selección

Para el húngaro, el reconocedor ha sido entrenado previamente con la base de datos SpeechDat húngara, formada por locuciones obtenidas a través de la red telefónica, lo que favorece su portabilidad, dado que no es necesario realizar este entrenamiento y su integración en nuestro sistema es bastante sencillo. Además, el hecho de que sea un reconocedor fonético, nos permite utilizar la información de los fonemas reconocidos para condicionar la reducción de ruido, tal y como se explicó anteriormente.

Su facilidad en el manejo, la sencilla integración en nuestro sistema, y la aportación de la información del reconocimiento para su uso en la reducción de ruido, hacen que este reconocedor sea ideal, y nos permite la posibilidad de evaluar la independencia del sistema frente al idioma, poniendo en liza a dos idiomas tan dispares, como son el húngaro y el japonés. Cabe destacar que el código fuente del reconocedor se encuentra disponible en la página web del grupo de desarrollo de la Universidad Tecnológica de Brno, y que además, este tiene licencia de software libre para su uso académico y de investigación.

Capítulo 4

Pruebas y Resultados

En este capítulo vamos a presentar los distintos experimentos desarrollados a lo largo de este PFC. El objetivo es mostrar los resultados obtenidos en las distintas condiciones de filtrado, para evaluar así si los cambios introducidos mejoran el comportamiento del método de filtrado, o al contrario, lo empeoran.

En general, los resultados que van a ser analizados tras la realización de la reducción de ruido, son las evaluaciones SNR y la evaluación final por medio del reconocedor de voz HTK, con el cual obtendremos los porcentajes de las locuciones que han sido reconocidas con éxito. Todos estos valores son evaluados antes y después del filtrado, lo que nos será de gran utilidad para poder comparar ambos resultados.

4.1. Pruebas Iniciales

Este primer experimento tiene como objetivo principal probar la integración de los principales componentes del sistema de reducción de ruido que hemos planteado, y comprobar así su funcionalidad. Asimismo, los resultados que arroje nos servirán de referencia para poder establecer los posibles puntos de mejora.

En esta primera prueba, se ha utilizado la implementación del filtro de Wiener QIO, junto con el VAD que viene incluido en el conjunto de herramientas del filtro, con la misma configuración que recomiendan los desarrolladores. Asimismo, se han considerado todas las locuciones presentes en la base de datos CENSREC-2 (tanto locuciones de *test* como de *train*), puesto que el objetivo no es comparar los resultados de la reducción de ruido en distintas condiciones acústicas.

Experimento I	
Tipo de filtro	*Wiener subóptimo QIO*
VAD utilizado	*QIO*
Conjunto de locuciones	*CENSREC-2 completo*

Tabla 13. Resumen del experimento I

4.1.1. Evaluación de la SNR

La evaluación de la SNR (*Signal to Noise Ratio*) nos permite calcular la cantidad de ruido que ha sido eliminado en la locución a tratar. En este caso, el valor obtenido

por si solo carece de valor experimental si no lo comparamos con otro, por lo que siempre vendrá acompañado de los valores de SNR de referencia, que no son otros que la SNR de las locuciones originales, es decir, sin filtrar.

En la siguiente gráfica están representados los distintos valores de SNR para cada una de las locuciones. En el eje de ordenadas están indicados los valores de SNR de la locución antes de pasar por la etapa de reducción de ruido (SNR_{in}), mientras que el eje de abscisas representa los valores de SNR de la misma locución después de haber sido aplicado el filtrado de Wiener (SNR_{out}). Sobre la misma gráfica se ha representado la recta $f(x) = x$ que delimita que locuciones presentan un mayor valor de SNR_{out} con respecto a SNR_{in} o viceversa.

De esta forma, se puede comprobar el resultado global del filtrado, observando que locuciones caen por encima o por debajo de la recta $f(x) = x$, y por tanto, que locuciones han logrado mejorar o empeorar su SNR.

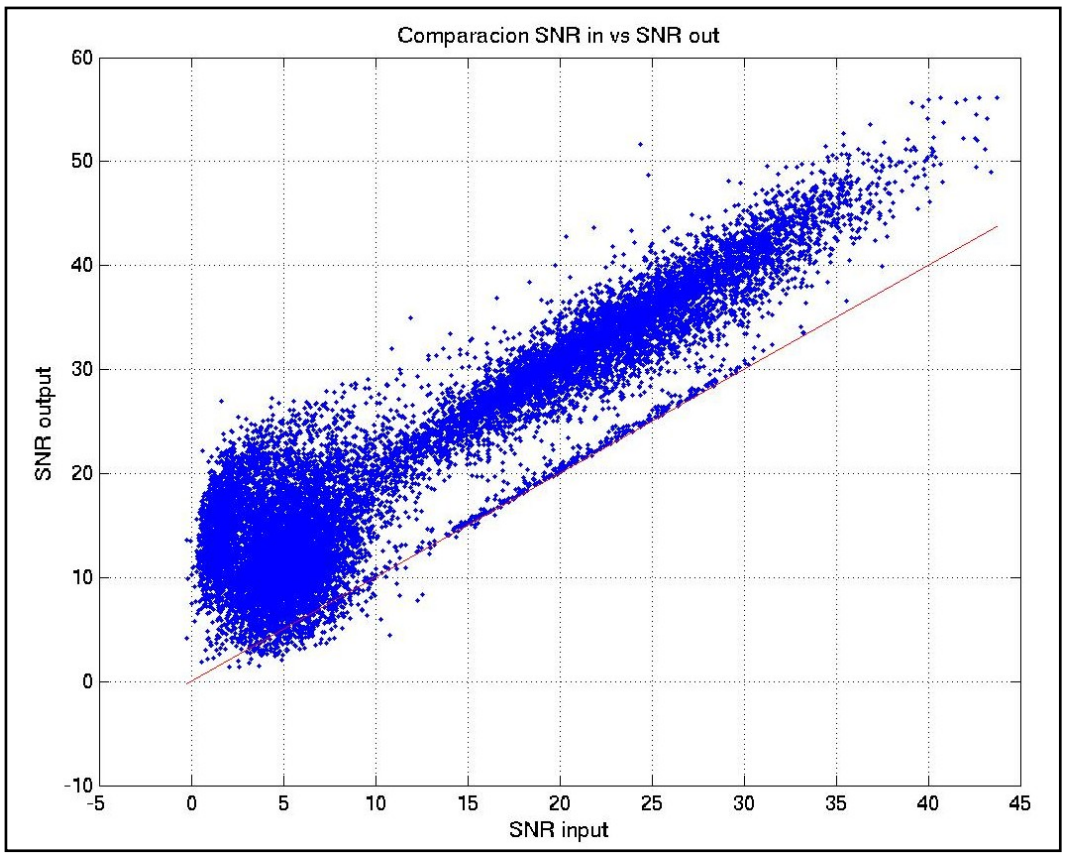

Figura 12. Comparación de $\boldsymbol{SNR_{in}}$ y $\boldsymbol{SNR_{out}}$ del experimento I.

Tal y como vemos en el gráfico superior, la gran mayoría de las grabaciones presentan mayor SNR tras la reducción de ruido aplicada. Este hecho era de esperar, puesto que la definición del filtro de Wiener subóptimo indica que este siempre es capaz de mejorar la SNR de la señal sobre la que se aplique.

El siguiente histograma ilustra este aspecto. En él, se representa el nivel diferencial de SNR (SNR_{out}– SNR_{in}) para todas las muestras filtradas. Una vez más, se comprueba que la mayoría de las locuciones han mejorado su nivel de SNR tras el filtrado, siendo apenas un grupo reducido de ellas las que han empeorado, y las que lo han hecho, apenas han perdido unos decibelios.

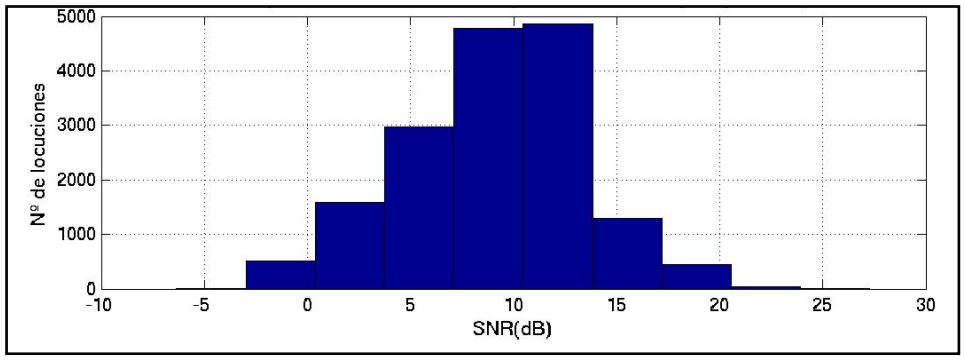

Figura 13. Histograma de SNR diferencial entre $\boldsymbol{SNR_{in}}$y $\boldsymbol{SNR_{out}}$ del experimento I

En la siguiente tabla se reflejan a modo de resumen, los datos estadísticos de la evaluación SNR realizada.

	Valor Medio	σ
SNR_{in}	12.9545 dB	10.0024 dB
SNR_{out}	21.9893 dB	11.4116 dB
$SNR_{out} - SNR_{in}$	9.0348 dB	4.3076 dB

Tabla 13. Parámetros estadísticos de la evaluación SNR

4.1.2. Evaluación con HTK

Como ya se ha visto en el apartado de *entorno experimental*, HTK es un potente conjunto de herramientas pensado para el ASR (*Automatic Speech Recognition*) el cual

es capaz de, a partir de una serie de muestras de ejemplo para el entrenamiento del sistema, reconocer el contenido de una locución dada. Además de eso, las herramientas de análisis de los resultados obtenidos tras el reconocimiento nos permite calcular que porcentaje de éxito (y por tanto también de error) ha cometido el reconocedor con las locuciones. En nuestro caso, todas las locuciones de la base datos CENSREC-2 estan transcritas para facilitar esta tarea de análisis. Así es posible calcular cómo de bueno ha sido el reconocimiento, y por tanto, cómo de bueno ha sido el filtrado.

Al contrario que con la evaluación de la SNR, la evaluación HTK de las muestras filtradas se centra sobre el contenido de las mismas, es decir, podemos medir la calidad de las locuciones y su inteligibilidad, y comprobar si la reducción de ruido ha logrado mejorar la calidad de los audios, o por el contrario ha distorsionado la señal de voz contenida en el mismo.

Para establecer el criterio de "mejora" de la calidad de las locuciones filtradas, vamos a comparar directamente el porcentaje de aciertos del reconocedor antes y después del filtrado. En la siguiente tabla, podemos comprobar cuales han sido los resultados del reconocimiento para el experimento que nos ocupa. En primer lugar se muestran los resultados de reconocimiento básicos del conjunto de locuciones de la base de datos, a modo de referencia, indicando el porcentaje de locuciones reconocidas con éxito. En segundo lugar se representan los resultados arrojados por el experimento, y posteriormente, el porcentaje de mejora para cada una de las condiciones de evalución, definidas con anterioridad.

Datos CENSREC-2 (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,58	74,49	61,46	48,87	66,35
Resultados Experimento I (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,64	75,26	60,02	49,23	66,29
Mejora				
Condición 1	Condición 2	Condición 3	Condición 4	Media
0,31 %	3,02 %	-3,74 %	0,70 %	-0,19 %

Tabla 14. Resultados (porcentaje de aciertos (*Word accuracy*)) de reconocimiento con HTK del experimento I.

Para este caso concreto, los resultados del reconocimiento nos muestran como la calidad/inteligibilidad de las locuciones ha mejorado en algunos casos, pero por lo general, los resultados obtenidos no siempre mejoran, si tenemos en cuenta los valores de SNR mostrados anteriormente, y que las muestras reconocidas han sido filtradas previamente.

4.2. Sustitución del VAD

El segundo experimento que vamos a realizar tiene un cambio sustancial con respecto al anterior. En este caso, queremos comprobar la funcionalidad de un VAD basado en energía, y por ello hemos sustituido el propio detector de voz del paquete de herramientas de Qio, por otro mucho más sencillo.

El detector de actividad de voz que proponemos en este caso clasifica los fragmentos de la locución en función de su nivel de energía. El funcionamiento de este VAD se puede resumir en los siguientes pasos:

1. Se enventana la señal para poder trabajar con pequeñas porciones de la misma.
2. Se calculan los niveles de energía de cada ventana.
3. Se obtienen los niveles máximos y mínimos de toda la secuencia de audio y en función de estos, el margen dinámico resultante.
4. Se clasifican las ventanas como *voz/no voz* en función de un determinado umbral dependiente del margen dinámico obtenido anteriormente.
5. Se eliminan los silencios demasiado cortos y los picos de ruido por vecindad.

De esta forma, en función de los niveles de energía de la señal, se discrimina entre fragmentos de voz y fragmentos de no voz.

Experimento II	
Tipo de filtro	*Wiener subóptimo QIO*
VAD utilizado	*Basado en energía*
Conjunto de locuciones	*CENSREC-2 completo*

Tabla 15. Tabla resumen del experimento II

4.2.1. Evaluación de la SNR

Al igual que en el experimento anterior, se ha realizado en análisis de los niveles de SNR antes y después del filtrado, para poder comprobar si la condición de filtrado subóptimo se sigue cumpliendo. En la siguiente figura, están representados los valores de SNR_{in} frente a los valores de SNR_{out}, junto con la recta de referencia $f(x) = x$.

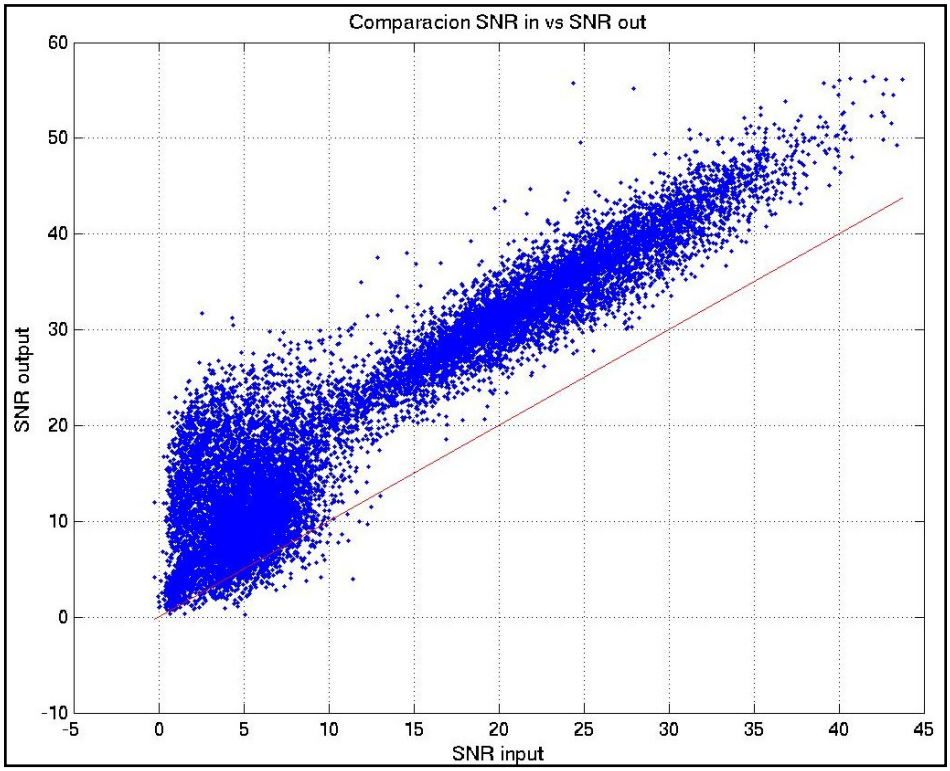

Figura 14. Comparación de $\boldsymbol{SNR_{in}}$ y $\boldsymbol{SNR_{out}}$ del experimento II

Como en el caso anterior, el histograma de los valores de la SNR diferencial nos da una idea del comportamiento del filtro en cuanto a SNR. En este caso, los valores se encuentran mucho más localizados en torno a los 10 dB de subida de la SNR, hecho que se puede comprobar en la grafica anterior de forma visual.

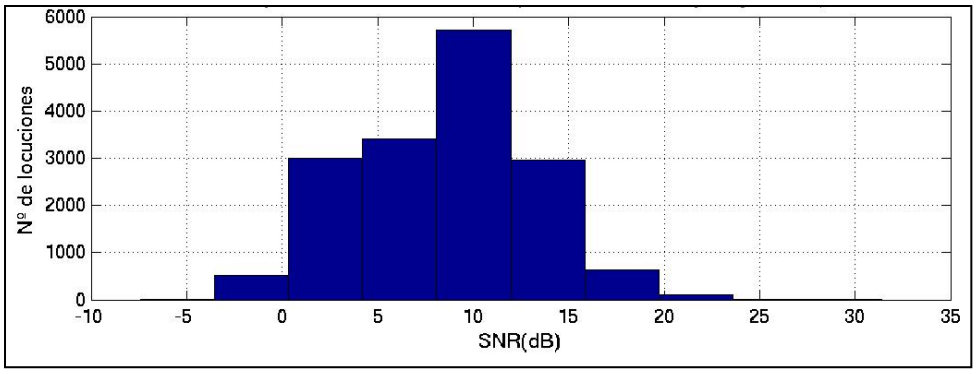

Figura 15. Histograma de SNR diferencial entre $\boldsymbol{SNR_{in}}$ y $\boldsymbol{SNR_{out}}$ del experimento II

En la siguiente tabla se reflejan a modo de resumen, los datos estadísticos de la evaluación SNR realizada.

	Valor Medio	σ
SNR_{in}	12.9545 dB	10.0024 dB
SNR_{out}	21.6104 dB	12.5637 dB
$SNR_{out} - SNR_{in}$	8.6545 dB	4.6154 dB

Tabla 16. Parámetros estadísticos de la evaluación **SNR**

4.2.2. Evaluación con HTK

En la siguiente tabla se encuentran reflejados los resultados del reconocimiento a través de HTK para este experimento.

Datos CENSREC-2 (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,58	74,49	61,46	48,87	66,35
Resultados Experimento I (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
85,42	78,14	57,70	45,78	66,76
Mejora				
Condición 1	Condición 2	Condición 3	Condición 4	Media
24,92 %	14,31 %	-9,76 %	-6,04 %	1,22 %

Tabla 17. Resultados de reconocimiento con HTK del experimento II.

El efecto del cambio de detector de actividad de voz es notable en este caso. Para las condiciones 1 y 2, la mejora de la calidad del audio es notable, mientras que para las condiciones 3 y 4 los resultados son bastante pobres. Esta diferencia tan acusada en este aspecto está directamente relacionada con el VAD basado en energía. Para las condiciones 1 y 2 los micrófonos utilizados en las fases de entrenamiento y test fueron el mismo, es decir, el micrófono de manos libres (*HF*). En estas dos condiciones, el funcionamiento del VAD ha resultado muy satisfactorio, al conseguir mejorar la calidad del filtrado. Este hecho no se da en las condiciones 3 y 4, donde la tasa de reconocimiento correcto ha bajado con respecto al original.

Hay que destacar también la influencia de las condiciones acústicas sobre los resultados obtenidos. Para los casos 1 y 3, las condiciones acústicas fueron las mismas, al contrario que para los casos 2 y 4. La sensible variación de la tasa de reconocimiento en los pares de condiciones 1-2 y 3-4 tiene que ver, por tanto, con la diferencia de las condiciones acústicas en las fases de entrenamiento y test.

4.3. Filtro de Wiener ETSI standard v1.1.3

Con objeto de poder tener una referencia, y a modo de comparación con los experimentos realizados hasta ahora, la siguiente prueba trata de comprobar el funcionamiento de otra implementación de Wiener. En este caso se trata de la implementación de Wiener utilizada en la fase de reducción de ruido del ETSI ES 202 050 V1.1.3 front-end pensado para reconocimiento de voz.

Experimento III	
Tipo de filtro	*Wiener ETSI standard v1.1.3*
VAD utilizado	*VADNest (basado en energía)*
Conjunto de locuciones	*CENSREC-2 completo*

Tabla 18. Tabla resumen del experimento III

Esta implementación del filtro de Wiener tiene la particularidad de estar desarrollada en dos etapas, o lo que es lo mismo, la reducción de ruido se realiza a través de dos filtros de Wiener concatenados.

Este modo de aplicar la reducción de ruido supone una novedad con lo visto hasta ahora, puesto que combina los dos modos de filtros vistos, el filtro óptimo y el filtro subóptimo. En el siguiente diagrama, podemos ver el esquema básico de funcionamiento propuesto, con las dos etapas de filtrado.

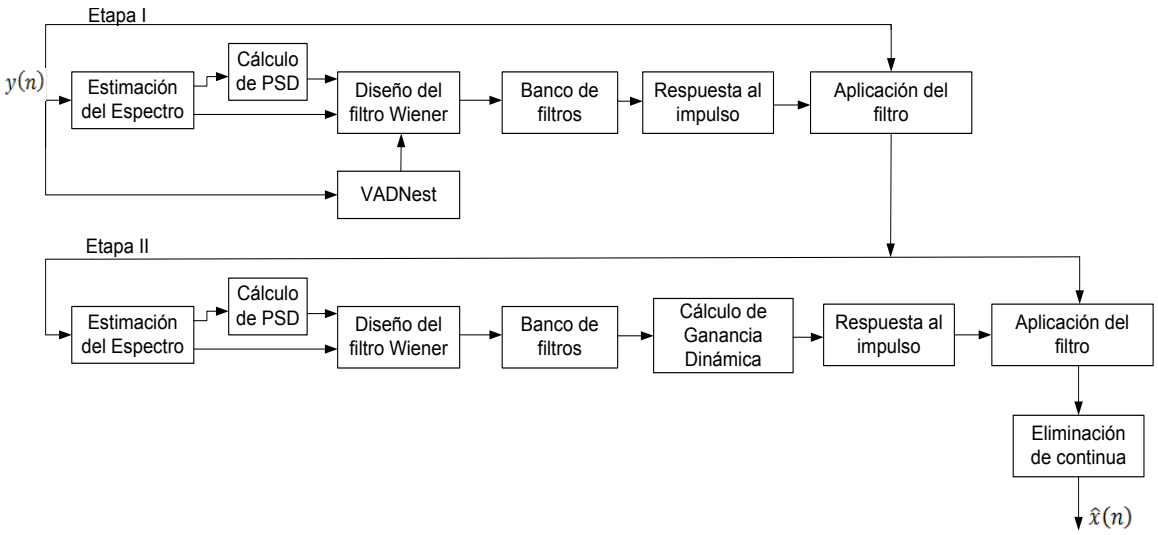

Figura 16. Diagrama de bloques del doble filtro de Wiener propuesto en el ETSI standard v1.1.3

La primera etapa de filtrado, consiste sencillamente en un filtro de Wiener óptimo, con un VAD sencillo, basado en energía. El objetivo esta primera etapa no es más que la realización de un prefiltrado de la señal, rebajando el nivel de ruido medio presente en la señal. La segunda etapa es algo más compleja. Se trata de un filtro de Wiener subóptimo, mediante el cual, la reducción de ruido se aplica de forma dinámica en función de los niveles de SNR de la señal en los fragmentos a filtrar.

En esta prueba no se ha realizado un análisis de la SNR de los resultados tras el filtrado, puesto que el objetivo es la comparación de los resultados a nivel de evaluación HTK, que es realmente el indicador válido de la calidad del filtrado.

4.3.1. Evaluación con HTK

En la siguiente tabla se encuentran reflejados los resultados del reconocimiento a través de HTK para este experimento.

Datos CENSREC-2 (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,58	74,49	61,46	48,87	66,35
Resultados Experimento I (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
84,06	70,50	62,18	56,44	68,30
Mejora				
Condición 1	Condición 2	Condición 3	Condición 4	Media
17,92 %	-15,64 %	1,87 %	14,81 %	5,78 %

Tabla 19. Resultados de reconocimiento con HTK del experimento III.

A primera vista, los resultados de la evaluación HTK son aceptables, si tenemos en cuenta que en la mayoría de las condiciones evaluadas la tasa de aciertos ha sido superior a la estándar. Sin embargo, los resultados presentan poca uniformidad con respecto a los anteriores, dado que la Condición 2, obtiene unos resultados realmente pobres, mientras que en el caso del filtro Qio, es en esa condición, donde se han obtenido los mejores resultados hasta el momento.

A pesar de este inconveniente, los resultados se pueden considerar aceptables, dado que en media, la tasa de reconocimiento con éxito es superior a la estándar. No hay que olvidar que esta implementación fue planteada y diseñada como paso previo a un sistema ASR, por lo que este funcionamiento entra dentro de lo esperado. No obstante, con los resultados obtenidos hasta ahora en los experimentos previos, esta solución no plantea una mejora frente a lo que ya hemos visto, por lo que no será tenida en cuenta para las mejoras posteriores.

4.4. Sustitución del VAD por un reconocedor

Tras comprobar la importancia del VAD y su influencia sobre los resultados finales del proceso de reducción de ruido, el siguiente paso, tal y como se propuso anteriormente, consiste en sustituir el VAD por un reconocedor, en este caso, un reconocedor fonético.

Experimento IV	
Tipo de filtro	*Wiener subóptimo QIO*
VAD utilizado	*Reconocedor "phnrec"*
Conjunto de locuciones	*CENSREC-2 completo*

Tabla 20. Tabla resumen del experimento IV

El uso del reconocedor fonético aporta ciertas ventajas frente a los VAD convencionales. En estos, los errores de tipo MSC (*Mid Speech Clipping*) son muy comunes, dado que los VAD's basado en energía, tienden a discriminar los fragmentos de voz muy cortos. Puede darse el caso de palabras de muy corta duración que sean clasificadas como no-voz, en lugar de cómo voz. Un reconocedor fonético, independientemente de la duración de una determinada palabra, va a identificar dicha palabra, y clasificar como voz en la gran mayoría de casos. Además, ayuda a reducir el número de fragmentos de silencio adyacentes a fragmentos de voz que son clasificados como voz (error de tipo "*over*"), dado que el reconocedor se ajusta estrictamente al contenido de la grabación.

4.4.1. Evaluación de la SNR

Como en los experimentos anteriores, se ha realizado en análisis de los niveles de SNR antes y después del filtrado, para poder comprobar si la condición de filtrado subóptimo se sigue cumpliendo. En la siguiente figura, están representados los valores de SNR_{in} frente a los valores de SNR_{out}, junto con la recta de referencia $f(x) = x$.

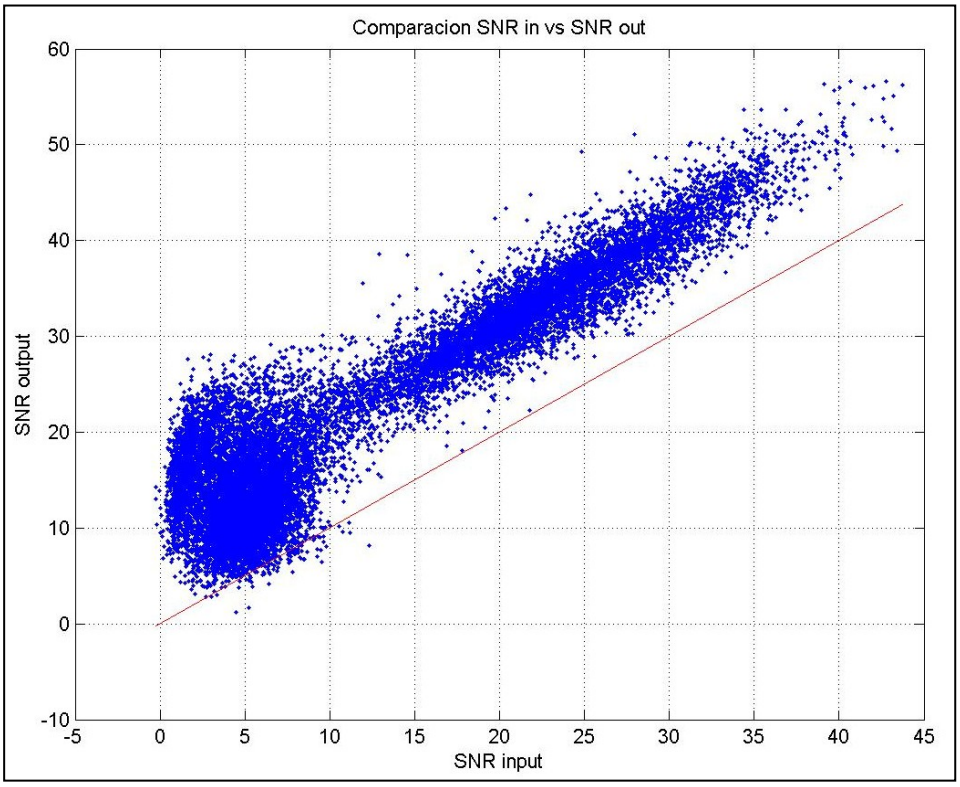

Figura 17. Comparación de $\boldsymbol{SNR_{in}}$ y $\boldsymbol{SNR_{out}}$ del experimento IV

En la figura anterior se puede comprobar que el comportamiento a nivel de SNR es muy similar al de los experimentos anteriores. En este sentido, no podemos esperar grandes novedades, la condición de filtro de Wiener subóptimo se sigue cumpliendo casi al cien por cien.

En la siguiente tabla se reflejan a modo de resumen, los datos estadísticos de la evaluación SNR realizada.

	Valor Medio	σ
SNR_{in}	12.9545 dB	10.0024 dB
SNR_{out}	23.0002 dB	11.4237 dB
$SNR_{out} - SNR_{in}$	8.6545 dB	4.6154 dB

Tabla 21. Parámetros estadísticos de la evaluación SNR

De estos resultados podemos concluir que el cambio del VAD, ha sido satisfactorio a nivel SNR. No solo ha conseguido contener los niveles de $\mathrm{SNR_{out}}$ en los valores esperados, sino que además, ha logrado mejorar los valores que se obtuvieron con el VAD del ATVS. En la siguiente tabla podemos ver las diferencias en ambos casos, y como el reconocedor utilizado como VAD mejora levemente los resultados. Es necesario tener en cuenta que el VAD del experimento II ya logró mejorar los resultados con respecto al caso incial, por lo que tenemos que valorar la mejora introducida en este experimento de forma muy positiva.

	Tasa de mejora	$\mathrm{SNR_{in}}$	$\mathrm{SNR_{out}}$	$\mathrm{SNR_{out}} - \mathrm{SNR_{in}}$
Experimento II	97,59 %	13,04dB	21.61dB	8,79dB
Experimento IV	99,52 %	13.03dB	23.00dB	10.13dB
Diferencia	1,93 %	---	2.61dB	1.34dB

Tabla 22. Comparativa de los experimentos II y IV. Las cifras de $\mathbf{SNR_{in}}$ y de $\mathbf{SNR_{out}}$ están referidos a sus respectivos valores medios.

4.4.2. Evaluación con HTK

En la siguiente tabla se encuentran reflejados los resultados del reconocimiento a través de HTK para este experimento.

Datos CENSREC-2 (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,58	74,49	61,46	48,87	66,35
Resultados Experimento I (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
87,44	75,69	64,22	53,41	70,19
Mejora				
Condición 1	Condición 2	Condición 3	Condición 4	Media
35,32 %	4,70 %	7,16 %	8,88 %	11,41 %

Tabla 23. Resultados de reconocimiento con HTK del experimento IV.

A tenor de los resultados obtenidos en la evaluación SNR, presentados en el punto anterior, no es de extrañar que para el caso de la evaluación HTK , estos sean similares. De nuevo, la influencia del VAD sobre la prueba de reconocimiento queda patente, dado que la tasa de reconocimiento con éxito vuelve a subir con respecto al experimento anterior. En este caso, se ha logrado mejorar dicha tasa en todas las condiciones propuestas, incluso en las condiciones 3 y 4, donde anteriormente los resultados siempre habían sido negativos.

Estos resultados, junto con los obtenidos en la evaluación SNR, dejan patente la clara mejora que representa sustituir un VAD basado en energía por un reconocedor fonético, lo que demuestra la importancia de fase de segmentación de la señal, y la extracción correcta de los segmentos que contienen voz.

4.5. Filtro de Wiener subóptimo con dependencia fonética

Gracias a la introducción del reconocedor fonético en la fase de clasificación de los fragmentos de la locución, podemos hacer uso de los datos que a partir de este obtenemos, para mejorar las etapas siguientes. El condicionamiento fonético va a permitir que ajustemos el filtrado al contenido de la locución, y por tanto, evitemos añadir distorsión en exceso sobre la señal que se está tratando.

Experimento V	
Tipo de filtro	*Wiener subóptimo QIO con dependencia fonética*
VAD utilizado	*Reconocedor "phnrec"*
Conjunto de locuciones	*CENSREC-2 completo*

Tabla 24. Tabla resumen del experimento V

4.5.1. Aplicación del condicionamiento fonético

Para poder hacer uso del condicionamiento fonético, tenemos que tener en cuenta los valores de AFD obtenidos para cada clase fonética, y la implementación del filtro de Wiener que estemos utilizando.

Para el caso que nos ocupa, se ha utilizado un modelo de filtro de Wiener paramétrico, visto en el estado del arte, en el cual, podemos definir la cantidad de ruido a eliminar ajustando el parámetro α.

$$|H_r(k,n)|^2 = \max\left(\frac{|X(\omega_i,t)|^2 - \alpha\left|\widehat{N}(\omega_i,t)\right|^2}{|X(\omega_i,t)|^2}, \beta\right)^{\gamma}$$

En la implementación de Qio, el parámetro α está comprendido en un rango de valores recomendado (entre 1.125 y 3.125), y es dependiente de la SNR calculada para la muestra que se esté midiendo.

$$\alpha = \frac{-1.875}{20} PosteriorSNR_n + 3.125$$

De esta manera, cuando se detecta una muestra con un valor alto de SNR, se le aplica una reducción de ruido más agresiva, al contrario que cuando el valor de SNR medido es bajo. Lo que se pretende hacer con el condicionamiento fonético es un comportamiento parecido, con la diferencia que el factor que va a determinar la agresividad de la reducción de ruido es el factor AFD previamente calculado.

El mecanismo de funcionamiento del filtro en este caso, es muy distinto a los vistos previamente. En la fase segmentación de voz, el reconocedor fonético no solo se encarga de clasificar como voz o no-voz, sino que además etiqueta cada fonema encontrado y reconocido, para la aplicación del factor α en la fase de reducción de ruido.

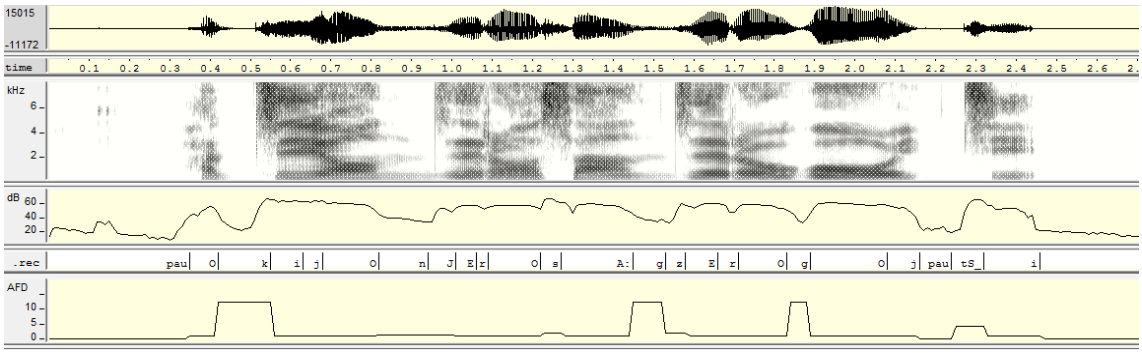

Figura 18. Forma de onda, espectrograma, energía, transcripción fonética y valor de AFD instantáneo para una grabación filtrada.

Posteriormente, con la información disponible de los fonemas que están presentes en el audio, se clasifican los en función del grupo fonético asociado, se calcula el nivel estimado de ruido presente en el audio en función de los silencios detectados, se aplica el factor α de sobreestimación de ruido, y finalmente, se realiza la reducción de ruido en función de lo estimado anteriormente.

4.5.2. Evaluación con HTK

En la siguiente tabla se encuentran reflejados los resultados del reconocimiento a través de HTK para este experimento.

Datos CENSREC-2 (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
80,58	74,49	61,46	48,87	66,35
Resultados Experimento I (%)				
Condición 1	Condición 2	Condición 3	Condición 4	Media
87,61	78,14	65,68	55,90	71,83
Mejora				
Condición 1	Condición 2	Condición 3	Condición 4	Media
36,20 %	14,31 %	10,95 %	13,75 %	16,29 %

Tabla 25. Resultados de reconocimiento con HTK del experimento V.

En línea con los resultados que se han ido obteniendo hasta este momento, el condicionamiento fonético aplicado en la reducción de ruido logra reducir la distorsión generada por el filtro de Wiener, logrando cumplir el compromiso entre reducción de ruido y distorsión. Hay que destacar el gran crecimiento de la tasa de reconocimiento con éxito en las condiciones de pruebas 2 y 4, en las cuales, hasta ahora, los resultados han sido muy modestos, debido a que en ambos casos, las condiciones acústicas de la fase de entrenamiento del reconocedor, y la de test, son distintas, mientras que en los casos 1 y 3, las condiciones acústicas son exactamente iguales.

4.6. Discusión de resultados

A continuación, vamos a comparar los resultados obtenidos para cada uno de las propuestas de mejora planteadas en los experimentos realizados. En la siguientes tabla se reflejas un resumen de la evaluación HTK de las pruebas anteriores para cada una de las condiciones de evaluación. Los datos representados hacen referencia a los porcentajes de mejora sobre la tasa de reconocimiento con éxito de cada una de las pruebas.

Condición de test 1		
	Tasa de reconocimiento con éxito	Mejora
Experimento I	80.64 %	0.31%
Experimento II	85.42 %	24.92 %
Experimento III	84.06 %	17.92 %
Experimento IV	87.44 %	35.32 %
Experimento V	87.61 %	36.20 %

Tabla 26. Resumen de los resultados de la condición de test 1 de la evaluación HTK.

Una de las características principales de la **Condición de test 1** es que tanto el entrenamiento del reconocedor, como el test, se realizaron con audios obtenidos con el mismo micrófono y en las mismas condiciones acústicas. Esto se ve reflejado en la alta tasa de reconocimiento con éxito de partida. En esta situación, el marco de mejora teórico es muy estrecho, pero a la vez, es una de las condiciones más homogéneas de todas, por lo que será más fácil obtener grandes resultados. Este hecho se ve reflejado en la evolución de los resultados obtenidos de la evaluación.

El porcentaje de mejora es el más alto de todos los casos, y donde, tras aplicar el filtrado de Wiener con condicionamiento fonético, es donde mejor se pueden apreciar sus efectos. Cabe destacar que en ninguna de las pruebas realizadas, para esta condición, los resultados han sido negativos, lo que también nos da una idea de la homogeneidad de esta prueba.

Condición de test 2		
	Tasa de reconocimiento con éxito	Mejora
Experimento I	75.26 %	3.02 %
Experimento II	78.14 %	14.31 %
Experimento III	70.50 %	-15.64 %
Experimento IV	75.69 %	4.70 %
Experimento V	78.14 %	14.31 %

Tabla 27. Resumen de los resultados de la condición de test 2 de la evaluación HTK**.**

En la **Condición de test 2**, se utilizaron grabaciones obtenidas en distintos entornos acústicos para las fases de entrenamiento y test. Comparando los resultados con la Condición 1, es fácil comprobar este dato. El punto de partida nos da una tasa de reconocimiento con éxito más baja, y la mejora máxima obtenida no llega a ser ni la mitad de buena que la obtenida en la condición 1. No obstante, la evolución de dicho dato es clara, y tanto en el experimento II como en el experimento V los datos obtenidos son muy positivos.

Condición de test 3		
	Tasa de reconocimiento con éxito	Mejora
Experimento I	60.02 %	-3.74 %
Experimento II	57.70 %	-9.76 %
Experimento III	62.18 %	1.87 %
Experimento IV	64.22 %	7.16 %
Experimento V	65.68 %	10.95 %

Tabla 28. Resumen de los resultados de la condición de test 3 de la evaluación HTK.

La **Condición de test 3** supone un cambio sustancial con respecto a lo visto hasta ahora. En este caso el micrófono utilizado para las condiciones de entrenamiento y test

fue distinto. Nuevamente, la tasa de reconocimiento con éxito vuelve a bajar con respecto a la condición anterior, y además es la primera condición en la que el experimento II, que hasta ahora estaba dando buenos resultados, no ha logrado mejorar la tasa de reconocimiento con éxito. Nuevamente se hace patente que el uso del condicionamiento fonético logra mejorar los resultados, incluso en las condiciones más adversas.

Condición de test 4		
	Tasa de reconocimiento con éxito	Mejora
Experimento I	49.23 %	0.70 %
Experimento II	45.78 %	-6.04 %
Experimento III	56.44 %	14.81 %
Experimento IV	53.41 %	8.88 %
Experimento V	55.90 %	13.75 %

Tabla 29. Resumen de los resultados de la condición de test 4 de la evaluación HTK.

En el último caso, la **Condición de test 4** supone un reto de cara a la reducción de ruido. En esta ocasión, para las fases de entrenamiento y test se utilizaron grabaciones obtenidas con distinto micrófono y en distintas condiciones acústicas. En este sentido, la prueba resulta muy poco homogénea, pero unos resultados positivos, nos puede dar una idea de la robustez del sistema de reducción de ruido y de las mejoras empleadas.

En este caso, el experimento V vuelve a presentar los mejores resultados con respecto al resto de experimentos, consolidándose como el más regular, y el más robusto a la vez. Nuevamente, el experimento II vuelve a fallar, como pasaba en la condición anterior, debido a la poca robustez del empleo de un VAD basado en energía.

Capítulo 5

Conclusiones y trabajo futuro

5.1. Conclusiones

A través de todos las pruebas realizadas y reflejadas en este proyecto, hemos podido ver como la aplicación de las mejoras sucesivas del sistema de filtrado de Wiener ha logrado mejorar los resultados de los que se partía inicialmente. En este sentido, podemos considerarnos satisfechos, puesto que el objetivo principal del proyecto ha sido cubierto con creces.

Gracias a los resultados del experimento V, se ha demostrado como el condicionamiento fonético puede ser decisivo. No solo se ha comprobado su completa funcionalidad, sino que además, se ha demostrado como los distintos grupos fonéticos que hemos clasificado presentan un comportamiento muy distinto frente a la distorsión, y como el modo de generación de los distintos fonemas en el tracto vocal, hace que determinados "sonidos" presenten mayor robustez frente al ruido.

También se ha comprobado como a nivel de las clases amplias fonéticas aquí expuestas, dos idiomas tan distintos como son el húngaro y el japonés, comparten una serie de características comunes, que los hace compatibles a dicho nivel. El rápido desarrollo de los reconocedores de voz ha permitido poder analizar estas características del lenguaje hablado, y poder utilizarlas en nuestro provecho.

Podemos concluir, por tanto, que las mejoras expuestas en cuanto a condicionamiento fonético y el uso de reconocedores de voz como VAD, funcionan y son aplicables a los sistemas de reducción de ruido más utilizados de hoy en día.

El resultado del trabajo aquí expuesto ha sido publicado en el ICPR 2010 (International Conference Pattern Recognition) [17], uno de los congresos internacionales de investigación que goza del máximo prestigio, mostrando los resultados de esta nueva técnica de reducción de ruido vistos en este proyecto.

5.2. Trabajo futuro

A pesar de los buenos resultados obtenidos, es necesario resaltar los puntos que son sensibles a mejorar, para poder utilizar esta herramienta de forma mucho más efectiva. Para que este novedoso método de reducción de ruido sea útil y aplicable, es condición

necesaria su implementación en un sistema de tiempo real. El desarrollo de esta herramienta en tiempo real supondría su uso en una gran cantidad de aplicaciones, como puede ser, por ejemplo, la telefonía o los sistemas de comunicación embarcados.

También, es necesario resaltar que las pruebas realizadas, se han hecho con idiomas distintos. Esto ha servido para probar la robustez de las propuestas de mejora, pero es necesario comprobar que resultados se obtendrían si, tanto las grabaciones sobre las que trabaja, como el reconocedor utilizado como VAD fueran el mismo. Es de esperar que los resultados en ese caso fueran incluso mejores que los vistos en este proyecto, por lo que sería necesario analizar que límites presenta el condicionamiento fonético.

De las conclusiones arrojadas sobre este proyecto, se abren nuevas vías de investigación. Hay que destacar la relaciones a nivel fonético (realmente, a nivel de clases amplias fonéticas) entre idiomas tan dispares. Del estudio del porqué de este comportamiento y su análisis se pueden obtener nuevos resultados que pueden ser utilizados nuevamente en los sistemas de procesamiento automático del habla, que pueden suponer nuevas mejoras.

Bibliografía

[1] **J. Chen, J. Benesty, Y. Huang and E.J. Diethorn.** Fundamentals of Noise Reduction. *Springer Handbook.* s.l. : Springer, 2008.

[2] **Iser, B., Minker, W. and Schmidt, G.** *Bandwith extensions of speech signals.* s.l. : Springer, 2008.

[3] **ITU-T.** *Recommendation P.800: Methods for subjective determination of transmission quality.* 1996.

[4] **Fukunaga, K.** *Introduction to Statistial Pattern Recognition.* San Diego : s.n., 1990.

[5. **Bullington, K. and Fraser, J. M.** Engineering aspects of TASI. *The Bell System Technical Journal.* 1959, pp. 353-364.

[6] **ITU.** *A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications.* 1996.

[7] **ETSI.** *Voice activity detector (VAD) for adaptative mult-rate (AMR) speech traffic channels.* 1999.

[8] **ETSI**. *Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advance front-end feature extraction algorithm; compression algorithms.* 2002.

[9] **Association, International Phonetic.** *Handbook of the International Phonetic Association.* s.l. : Cambridge, 1999.

[10] **Malmkjær, Kirsten.** *The Linguistics Encyclopedia.* London : s.n., 2004.

[11] **Schwarz, P.** *Phoneme Recognition based on Long Temporal Context, PhD Thesis.* s.l. : Brno University of Tecnology, 2009.

[12] **Wells, J. C.** *SAMPA computer readable phonetic alphabet.* s.l. : Mouton de Gruyter, 1997.

[13] **Romano, A., Interlandi, G. and Mairano, P.** *Multimedia IPA chart. [Online] Laboratorio di Fonetica Sperimentale "Arturo Genre" di Torino*. http://www.lfsag.unito.it/ipa/.

[14] **Adami, A., et al.** *Qualcomm-ICSI-OGI features for ASR.* 2002.

[15] **Nakamura, S., Fujimoto, M. and Takeda, K.** *CENSREC2: Corpus and Evaluation Environments for In Car Continuous Digit Speech Recognition.* 2006. p. paper 1726.

[16] **Young, S., et al.** *The HTK Book.* s.l. : Cambridge Universisty Engineering Department, 2002.

[17] **Gonzalez-Caravaca, Guillermo, Toledano, Doroteo Torre and Puertas, Maria.** *Phone-Conditioned Suboptimal Wiener Filtering.* in Proc. IEEE International Conference on Pattern Recognition (ICPR) 2010. ISSN: 1051-4651, DOI:10.1109/ICPR.2010.1088, pp. 4480-4483.

Anexo A

Lema. Con los valores de λ_i $(i = 1, 2, \dots, L)$ tal que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_L \geq 0$ que fueron definidos en (34) y con $\mu > 0$ tenemos que

$$\left[\sum_{i=1}^{L} \frac{\lambda_i^3}{(\lambda_i+\mu)^2} q_i^2\right] \sum_{i=1}^{L} q_i^2 \geq \left[\sum_{i=1}^{L} \frac{\lambda_i^2}{(\lambda_i+\mu)^2} q_i^2\right] \sum_{i=1}^{L} \lambda_i q_i^2 \qquad [a]$$

donde q_i puede ser cualquier número real.

Demostración. Esta inecuación puede ser probada a través del método de inducción.

- *Paso inicial*

  Si consideramos *L = 2*

$$\left(\sum_{i=1}^{2} \frac{\lambda_i^3}{(\lambda_i+\mu)^2} q_i^2\right) \sum_{i=1}^{2} q_i^2 = \frac{\lambda_1^3}{(\lambda_1+\mu)^2} q_1^4 + \frac{\lambda_2^3}{(\lambda_2+\mu)^2} q_2^4 + \left(\frac{\lambda_1^3}{(\lambda_1+\mu)^2} + \frac{\lambda_2^3}{(\lambda_2+\mu)^2}\right) q_1^2 q_2^2 \qquad [b]$$

Teniendo en cuenta que $\lambda_1 \geq \lambda_2 \geq 0$, es fácil comprobar que

$$\frac{\lambda_1^3}{(\lambda_1+\mu)^2} + \frac{\lambda_2^3}{(\lambda_2+\mu)^2} \geq \frac{\lambda_1^2 \lambda_2}{(\lambda_1+\mu)^2} + \frac{\lambda_1 \lambda_2^2}{(\lambda_2+\mu)^2} \qquad [c]$$

donde ambos lados de la ecuación son iguales cuando $\lambda_1 = \lambda_2$. Por tanto tenemos que

$$\left(\sum_{i=1}^{2}\frac{\lambda_i^3}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{2}q_i^2$$
$$\geq \frac{\lambda_1^3}{(\lambda_1+\mu)^2}q_1^4+\frac{\lambda_2^3}{(\lambda_2+\mu)^2}q_2^4+\left(\frac{\lambda_1^2\lambda_2}{(\lambda_1+\mu)^2}+\frac{\lambda_1\lambda_2^2}{(\lambda_2+\mu)^2}\right)q_1^2q_2^2$$
$$=\left(\sum_{i=1}^{2}\frac{\lambda_i^2}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{2}\lambda_i q_i^2$$

[d]

Por lo que la propiedad es cierta para $L=2$, y la igualdad se mantiene cuando $\lambda_1=\lambda_2$ o cuando al menos q_1 o q_2 es igual a 0.

- *Paso inductivo*

En este caso, asumimos que la propiedad es cierta cuando $L=m$

$$\left(\sum_{i=1}^{m}\frac{\lambda_i^3}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m}q_i^2\geq\left(\sum_{i=1}^{m}\frac{\lambda_i^2}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m}\lambda_i q_i^2 \qquad [e]$$

Para continuar, debemos demostrar que la propiedad sigue siendo cierta para $L=m+1$

$$\left(\sum_{i=1}^{m+1}\frac{\lambda_i^3}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m+1}q_i^2=\left(\sum_{i=1}^{m}\frac{\lambda_i^3}{(\lambda_i+\mu)^2}q_i^2+\frac{\lambda_{m+1}^3}{(\lambda_{m+1}+\mu)^2}q_{m+1}^2\right)$$
$$\times\left(\sum_{i=1}^{m}q_i^2+q_{m+1}^2\right)$$
$$=\left(\sum_{i=1}^{m}\frac{\lambda_i^3}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m}q_i^2+\frac{\lambda_{m+1}^3}{(\lambda_{m+1}+\mu)^2}q_{m+1}^4$$
$$+\sum_{i=1}^{m}\left(\frac{\lambda_i^3}{(\lambda_i+\mu)^2}\frac{\lambda_{m+1}^3}{(\lambda_{m+1}+\mu)^2}\right)q_i^2q_{m+1}^2$$

[f]

Utilizando la hipótesis de inducción, y teniendo en cuenta el hecho de que

$$\frac{\lambda_i^3}{(\lambda_i+\mu)^2}+\frac{\lambda_{m+1}^3}{(\lambda_{m+1}+\mu)^2}\geq\frac{\lambda_i^2\lambda_{m+1}}{(\lambda_i+\mu)^2}+\frac{\lambda_i\lambda_{m+1}^2}{(\lambda_{m+1}+\mu)^2} \qquad [g]$$

Podemos obtener

$$\left(\sum_{i=1}^{m+1}\frac{\lambda_i^3}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m+1}q_i^2$$
$$\geq \left(\sum_{i=1}^{m}\frac{\lambda_i^2}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m}\lambda_i q_i^2 \; + \; \frac{\lambda_{m+1}^3}{(\lambda_{m+1}+\mu)^2}q_{m+1}^4$$
$$+ \sum_{i=1}^{m}\left(\frac{\lambda_i^2\lambda_{m+1}}{(\lambda_i+\mu)^2}+\frac{\lambda_i\lambda_{m+1}^2}{(\lambda_{m+1}+\mu)^2}\right)q_i^2 q_{m+1}^2$$
$$= \left(\sum_{i=1}^{m+1}\frac{\lambda_i^2}{(\lambda_i+\mu)^2}q_i^2\right)\sum_{i=1}^{m+1}\lambda_i q_i^2$$

[h]

Donde como en el caso anterior, se mantiene la igualdad cuando todos los λ_i correspondientes a los q_i distinto de cero son iguales.

Con esto se da por concluida la demostración.

Anexo B

Presupuesto

1) Ejecución Material

▪ Compra de ordenador personal (Software incluido)	2.200 €
▪ Alquiler de impresora laser durante 6 meses	280 €
▪ Material de oficina	150 €
▪ Total de ejecución material	2.630 €

2) Gastos generales

▪ sobre Ejecución Material	420 €

3) Beneficio Industrial

▪ sobre Ejecución Material	157 €

4) Honorarios Proyecto

▪ 1500 horas a 18 € / hora	27.000 €

5) Material fungible

▪ Gastos de impresión	90 €
▪ Encuadernación	200 €

6) Subtotal del presupuesto

▪ Subtotal Presupuesto	30.497 €

7) I.V.A. aplicable

▪ 18% Subtotal Presupuesto	5.489,46 €

8) Total presupuesto

- Total Presupuesto 35.986,46 €

Madrid, JULIO 2011

El Ingeniero Jefe de Proyecto

Fdo.: Guillermo González Caravaca

Ingeniero Superior de Telecomunicación

Anexo C

Publicaciones

Título: Phone-Conditioned Suboptimal Wiener Filtering
Autores: Guillermo González Caravaca, Doroteo Torre Toledano
Conferencia: International Conference Pattern Recognition (ICPR). Agosto 2010, Estambul.

2010 International Conference on Pattern Recognition

Phone-Conditioned Suboptimal Wiener Filtering

Guillermo Gonzalez-Caravaca, Doroteo Torre Toledano, Maria Puertas
ATVS, Escuela Politecnica Superior, UAM
{guillermo.gonzalez, doroteo.torre, maria.puertas}@uam.es

Abstract

A novel way of managing the compromise between noise reduction and speech distortion in Wiener filters is presented. It is based on adjusting the amount of noise reduced, and therefore the speech distortion introduced, on a phone-by-phone basis. We show empirically that optimal Wiener filters produce different amounts of speech distortion for different phones. Therefore we propose a phone-conditioned suboptimal Wiener filter that uses different amounts of noise reduction for each phone, based on a previous estimation of the amount of distortion introduced. Speech recognition results have shown that phone-conditioning suboptimal Wiener filtering can provide almost a 5% additional relative improvement in word accuracy over comparable optimal Wiener filtering.

1. Introduction

One of the most successful and widely used techniques in noise reduction in speech recordings is Wiener filtering [1]. This filter is very effective in reducing noise in speech recordings, but this is achieved at the cost of a distortion of the speech signal [2]. The Wiener filter is optimal in the sense that it is obtained by minimizing the Minimum Squared Error (MSE) between the clean signal and the noise reduced signal, but this criterion does not necessarily produce always the best results according to the quality of the filtered speech due to the presence of speech distortion. For this reason, in some cases sub-optimal Wiener filters less effective in noise reduction but also less aggressive with speech are preferred [1, 2]. In some implementations, it applies factor conditioning to achieve suboptimal filtering [3]. This section presents a very brief review of both types of Wiener filters to present the main hypothesis of this article: that it is possible to improve Wiener filtering of speech (for some applications such as speech recognition) by adjusting the amount of noise removed and the speech distortion introduced on phone-by-phone basis in what we have called *phone-conditioned sub-optimal Wiener filtering*.

1.1. Optimal Wiener filter

Considering that we have a noisy observed speech signal corrupted by additive noise

$$y(n) = x(n) + v(n)\,, \tag{1}$$

where $v(n)$ is the additive noise and $x(n)$ is the clean speech signal, the Wiener filter estimates the speech signal by processing the corrupted signal with a finite impulse response (FIR) filter with impulse response determined by a vector $\boldsymbol{h}$

$$\hat{x}(n) = \boldsymbol{h}^T \boldsymbol{y}(n). \tag{2}$$

The Wiener filter coefficients, $\boldsymbol{h}$, are determined by minimizing the MSE cost function

$$J_x(h) \triangleq E[e_x^2(n)], \tag{3}$$

where $e_x(n)$ is the error signal between the clean speech sample and its estimate from the corrupted signal at time n. It is well known [1] that the minimization of (3) yields the Wiener–Hopf equations

$$R_y h_0 = r_{yx}\,, \tag{4}$$

where R_y is the autocorrelation matrix of the observed signal $y(n)$ and r_{yx} is the cross-correlation vector between the noisy and clean speech signals. The former can be directly estimated from $y(n)$, while r_{yx} can be expanded into

$$r_{yx} = E[y(n)y(n)] - E[v(n)v(n)] = r_{yy} - r_{vv}\,. \tag{5}$$

Using (5) and solving (4) we obtain impulse response of the optimal Wiener filter

$$_oh = R_y^{-1} r_{yx} = R_y^{-1} r_{yy} - R_y^{-1} r_{vv} = h_1 - R_y^{-1} r_{vv}\,, \tag{6}$$

Where h_1 is the impulse response of the identity filter,

$$h_1 = [1 \;\; 0 \;\; \ldots \;\; 0]^T\,. \tag{7}$$

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1088

1.2. Suboptimal Wiener filter

When we apply the optimal Wiener filter the noise in the signal is reduced but this is achieved at the cost of a *distortion of the speech signal* that can reduce its quality and be as undesirable as the noise or even more. Fortunately, it is relatively simple to control the compromise between noise reduction and speech distortion since the optimal Wiener filter in (6) can be considered as composed of two filters in parallel, an identity filter (h_1) that replicates the input signal and a filter ($-R_y^{-1}r_{vv}$) that removes an estimate of the noise. By weighting the effect of the second filter we have a *suboptimal Wiener filter*

$$h_{sub} = h_1 - \alpha R_y^{-1} r_{vv} \qquad (8)$$

where α is a weighting factor. It can be demonstrated [1] that if α satisfies $0 < \alpha < 1$, we achieve noise reduction with less distortion over the signal than in the case of an optimal Wiener filter.

1.3. Phone-conditioned suboptimal Wiener filter

The main hypothesis of this article is that different phones are distorted in different degrees by the Wiener filter (possibly due to the fact that they have different spectral energy distributions) and that it is possible to improve the results of the sub-optimal Wiener filter by using a *phone-conditioned suboptimal Wiener filter* using a different weighting factor for different phones.

The rest of the article tries to verify these hypotheses. Section 2 describes the systems used in the experiments. Section 3 describes the databases and experimental setup. Section 4 verifies the hypothesis that different phonemes are distorted in different degrees by the optimal Wiener filter and Section 5 shows that phone-conditioned sub-optimal Wiener filtering can be better than Wiener filtering or sub-optimal Wiener filtering for the particular application of speech recognition.

2. System description

Several Wiener filters have been developed for this work. Our baseline system is an optimal Wiener filter using a standard energy-based VAD, as shown in Figure 1.a.

The second system is also an optimal Wiener filter but using a phone recognizer as VAD, as shown in Figure 1.b. Using a phoneme recognizer as VAD provides more accurate and robust speech/non-speech segmentation [5]. Given that noise reduction achieved

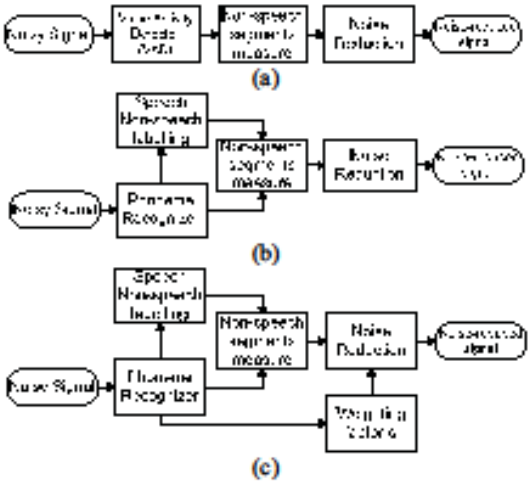

Fig.1. Systems compared in this article: baseline optimal Wiener filter with energy-based VAD (a), optimal Wiener filter with phone recognizer as VAD (b) and phone-conditioned sub-optimal Wiener filter (c).

by Wiener filtering depends on the accuracy of the VAD [4], it is reasonable to expect that using a phone recognizer as VAD increases performance of Wiener filtering, as shown in [4]. For this system we use the Hungarian phoneme recognizer developed at Brno University of Technology [6].

Finally, our *phone-conditioned sub-optimal Wiener filter* (Figure 1.c) is based on exploiting the information provided by the phone recognizer used in our second system to adjust the compromise between noise reduction and speech distortion on a phone-by-phone basis by using a different weighting factor (α) for each phone. One drawback of using phone conditioning is that it can make Wiener filtering language dependent, which is clearly undesirable. In order to have a high degree of language independency we use broad phonetic classes present in most languages. Thus we grouped the Hungarian phones used by the recognizer into the following classes: vowels, occlusives, fricatives, affricates, nasals, approximants and laterals.

3. Experimental setup

The database used for the experiments presented in this paper is CENSREC-2 [7], a database for the evaluation of in-car speech recognition in Japanese. The vocabulary of the database consists of 11 digit isolated words. The selection of this database for the experiments is based on two factors: it is a database developed for research in noise robust speech

Table 1. Train (○) and Test (△) combinations for the four conditions of CENSREC-2 testing protocol.

Condition	Cond. 1		Cond. 2		Cond. 3		Cond. 4	
Microphone	CT	HF	CT	HF	CT	HF	CT	HF
Idling	○	△	○		○		○	
Low speed	○	△	△		○	△		△
High speed	○	△	△		○	△		△

recognition and it contains a language (Japanese) very different from the Hungarian language used in the phonetic recognizer used for phone conditioning.

Speech data were recorded under different environmental conditions using combinations of three kinds of vehicle speeds and four kinds of in-car environments with two different microphones, one of them, a close-talking microphone (CT) and the other a hands-free microphone (HF). The database defines a baseline speech recognition experiment and the protocol for performing tests [7], including four different conditions explained in Table 1. We have followed this protocol to assess whether phone-conditioning in sub-optimal Wiener filtering provides improvements over optimal Wiener filtering.

4. Analysis of speech distortion for different phonetic classes

In order to verify the hypothesis that Wiener filtering distorts different phonemes in different degrees we used a measure to quantify the speech distortion due to a noise-reduction algorithm, the *attenuation frequency distortion* (AFD) [2]

$$\phi_{sd} \triangleq \frac{E\left[|X(i\omega)|^2 - |\hat{X}_{nr}(i\omega)|^2\right]}{E\left[|X(i\omega)|^2\right]} = \frac{P_x(\omega) - P_{\hat{x}_{nr}}(\omega)}{P_x(\omega)} \quad (9)$$

where $X(i\omega)$ and $P_x(\omega)$ are the Fourier spectrum and power spectral density of the clean speech $x(n)$, and $\hat{X}_{nr}(i\omega)$ and $P_{\hat{x}_{nr}}(\omega)$ are, respectively, the Fourier spectrum and power spectral density of the noise-reduced signal, $\hat{x}_{nr}(n)$.

We computed the AFD on a phone by phone basis using Hamming windows and comparing the audios recorded with the close-talking microphone (taken as clean speech) and the noisy audios captured by the hands-free microphone after Wiener filtering with our second system (corresponding to Fig 1.b). This allowed us to represent an AFD plot along with the phonetic transcription obtained by the Hungarian recognizer, the spectrogram and the waveform (Fig. 2), which clearly shows that the AFD is very different for different phonemes. Average and standard deviation of the AFD were measured for each phonetic class on the training part of the CENSREC-2 corpus (Table 2). Table 2 confirms that different phonetic classes experiment different amounts of distortion and is also the basis for setting α for each phone class. In particular we made α inversely proportional to the average AFD for each phone class according to the equation

$$\alpha = -0161 \times AFD + 3.03, \quad (10)$$

in which alpha varies in the range $1.25 < \alpha < 3.125$ as in the Wiener implementation used for this paper [3].

Table 2. Average values for AFD and corresponding standard deviations for each phonetic class.

Phoneme Group	AFD (Average)	Standard Deviation
Vowels	1.115	0.265
Occlusive	12.691	19.052
Fricative	1.968	1.007
Affricate	4.482	0.684
Nasal	1.339	0.244
Aproximant	1.106	0.152
Lateral	1.437	0.223

5. Speech recognition results

To measure noise filtering results, we performed speech recognition experiments using CENSREC-2 database and the HTK toolkit. Results are presented as the speech recognition word accuracy in percentage.

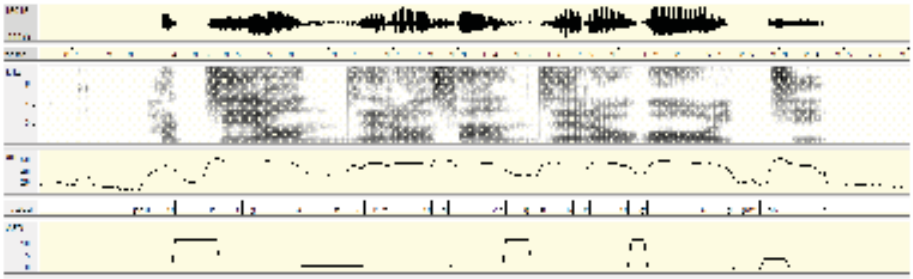

Fig. 2. Waveform, spectrogram, energy, phoneme transcription and AFD plot. Note that this is for a noise-reduced audio and the AFD measure denotes the ratio of likeness between the original and filtered audio defined as in (8).

Table 3. Word accuracy using optimal Wiener filter using a conventional energy-based VAD.

Baseline Results (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
80.58	74.49	61.46	48.87	66.35
Word Accuracy (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
84.42	75.19	57.86	37.28	63.69
Relative Improvement (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
19.77	2.74	-9.34	-22.67	-7.91

All tables compare results obtained from the original noisy speech and the noise reduced speech.

Table 3 shows the results of noise reduction with the optimal Wiener filter using a energy-based VAD (system in Figure 1.a). Table 4 shows the results with the optimal Wiener filter using the phonetic recognizer as a VAD (system in Figure 1.b) and finally Table 5 shows results for the suboptimal Wiener filter with noise weighting factor by phone class (system in Figure 1.c). Results show that, as previously shown in the literature [4], the effect of using the phonetic recognizer as VAD is very important, achieving in this experiment an additional 10% of average relative improvement. Comparing Table 4 and Table 5 it is also clear that using phone-conditioning to adjust the amount of noise reduction for each phone class is also clearly beneficial, allowing for improvements on all conditions of CENSREC-2 and achieving almost an additional 5% average relative improvement over results in Table 4.

6. Conclusions and future work

We have showed empirically that the distortion introduced by an optimal Wiener filter for different phone classes is very different and that it is possible to apply phone conditioning to adjust the amount of noise removed (and hence the amount of speech distortion introduced) by a suboptimal Wiener filter for different phone classes. This is the central idea of the proposed *phone-conditioned suboptimal Wiener filter*. We have shown that this filter is able to produce better results (in terms of speech recognition performance in noisy environments) than optimal Wiener filters.

Although our results are still somewhat limited, we consider that the possibility of applying phone conditioning to adapt the compromise between noise reduction and speech distortion deserves further study and consideration. As future work it would be desirable to check our conclusions with other databases, languages and acoustic conditions.

Table 4. Word accuracy using optimal Wiener filter with the recognizer as a VAD.

Baseline Results (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
80,58	74,49	61,46	48,87	66,35
Word Accuracy (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
85.75	69.10	56.81	45.34	64.25
Relative Improvement (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
26.62	-21.13	-12.07	-6.9	-6.24

Table 5. Word accuracy using suboptimal Wiener filter with the recognizer as a VAD and phoneme conditioning.

Baseline Results (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
80.58	74.49	61.46	48.87	66.35
Word Accuracy (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
87.61	78.14	65.68	55.90	71.83
Relative Improvement (%)				
Cond. 1	Cond. 2	Cond. 3	Cond. 4	Average
36.20	14.31	10.95	13.75	16.29

References

[1] J. Chen, J. Benesty, Y. Huang and E.J. Diethorn, Chapter 43, "Fundamentals of Noise Reduction" of J. Benesty, M. M. Shondi and Y. Huang (eds.) Springer Handbook of Speech Processing, Springer, 2008.

[2] J. Chen, J.Benesty, Y. Huang and S. Doclo, "New insights into the noise reduction Wiener filter", IEEE Trans. on Audio, Speech and Language Process., vol. 14, no. 4, pp. 1218-1234, Jul. 2006.

[3] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan and S. Sivadas, "Qualcomm-ICSI-OGI Features for ASR", ICSLP 2002.

[4] A. de la Torre, J. Ramírez, M. C. Benítez, J.C. Segura, L. García and A.J. Rubio, "Noise robust model-based Voice Activity Detection", Proc. Interspeech, 2006, pp. 1954-1957.

[5] J. Žibert, N. Pavesić and F. Mihelič, "Speech/Non-speech segmentation based on phoneme recognition features", EURASIP Journal on Applied Signal Processing, Volume 2006, Article ID 90495, Pages 1–13, February 2006.

[6] P. Schwarz, "Phoneme Recognition based on Long Temporal Context", PhD Thesis, Brno University of Technology, 2009.

[7] S. Nakamura, M. Fujimoto, and K. Takeda, "CENSREC2: Corpus and Evaluation Environments for In Car Continuous Digit Speech Recognition", Proc. Interspeech, pp.2330-2333, 2006.

Anexo D

liego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un sistema de reducción de ruido en grabaciones de audio. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

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

Condiciones generales

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

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

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

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

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

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

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

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

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

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos

precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

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

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

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

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

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

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

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

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

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

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

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

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

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

Condiciones particulares

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

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

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

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

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

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

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

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

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

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

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

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

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