UNIVERSITAT POLITECNICA DE CATALUNYA
DEPARTAMENT DE TEORIA DEL SENYAL I COMUNICACIONS

Tesi Doctoral

TECNICAS DE SPEECH ENHANCEMENT CONSIDERANDO ESTADISTICAS DE ORDEN SUPERIOR

AUTOR : Josep Mª Salavedra Molí

DIRECTOR : Enrique Masgrau Gómez

TUTORA : Asunción Moreno Bilbao

Barcelona, Juny 1995

UNIVERSITAT POLITECNICA DE CATALUNYA
DEPARTAMENT DE TEORIA DEL SENYAL I COMUNICACIONS

Tesi Doctoral

TECNICAS DE SPEECH ENHANCEMENT CONSIDERANDO ESTADISTICAS DE ORDEN SUPERIOR

AUTOR : Josep Mª Salavedra Molí

DIRECTOR : Enrique Masgrau Gómez

TUTORA : Asunción Moreno Bilbao

Barcelona, Juny 1995

UNIVERSITAT POLITÈCNICA DE CATALUNYA
ADMINISTRACIÓ D'ASSUMPTES ACADÈMICS

Aquesta Tesi ha estat enregistrada a la pàgina 77 amb el número 705

Barcelona, 5-10-95

L'ENCARREGAT DEL REGISTRE,

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Indice.

CAPITULO I

Introducción al Speech Enhancement.

I.1. Introducción.

El procesado digital de señal es un área en constante evolución que juega un papel esencial en la revolución industrial referida a alta tecnología, que actualmente tiene lugar. Desde la década de los años sesenta está representando un importantísimo campo donde se han realizado contribuciones muy significativas a partir de combinar ideas y metodologías procedentes de Teoría de Sistemas, Estadística, Análisis Numérico, Informática y Tecnología VLSI (Very Large Scale Integrated Circuits). El objetivo final perseguido por el procesado digital de señal consiste en procesar un conjunto finito de datos, obtenidos mediante un sensor o varios sensores, y extraer la información significativa que se encuentra oculta en estos datos. Normalmente, este fin se alcanza cuando se combina el desarrollo de formulaciones matemáticas con sus implementaciones algorítmicas y sus aplicaciones a un conjunto de datos reales. Estas técnicas de procesado de señales digitales llevan asociados unos conceptos de calidad como pueden ser la calidad de las estimaciones, la complejidad de cálculo, el coste de implementación o montaje o los efectos de la aritmética finita.

Durante las últimas décadas se han dedicado muchos esfuerzos dirigidos hacia la mejora de los algoritmos de procesado de la señal de voz. Por una parte los avances en los codificadores de voz han permitido una reducción de la velocidad de transmisión desde los 64Kbps del estandard PCM hasta velocidades pertenecientes al margen de velocidades comprendidas entre 6 y 13Kbps, donde una variedad de distintas estrategias permiten obtener

un nivel de calidad suficientemente bueno para su uso en telefonía comercial. Así, sistemas actuales, como por ejemplo el sistema digital Paneuropeo celular GSM de Telefonía Móvil, permiten la disponibilidad de un mayor número de canales dada una velocidad radioeléctrica prefijada para la transmisión de la información.

Esta mayor compresión de la información ha generado un aumento notable del coste computacional debido a los algoritmos de procesado de voz. Pero el gran desarrollo experimentado a nivel herramientas Hardware ha permitido su implementación en tiempo real, cuando no hace mucho tiempo hubiera sido impensable. Especial mención merece la reducción del tiempo de cálculo y el incremento del espacio de memoria disponible, y su menor consumo, experimentado por los DSP's. Este avance tecnológico también ha permitido tomar en consideración la posible implementación de técnicas de codificación de voz en Banda Ancha, para velocidades de transmisión medias y, asimismo, técnicas de reconocimiento de voz con tasas de error muy pequeñas.

Desgraciadamente, la mayoría de técnicas de procesado de señal de voz han sido diseñadas en ambientes pertenecientes al entorno laboratorio, donde ningún tipo de señales indeseadas se mezclan interfieren la señal deseada de voz. Sin embargo, a nivel práctico, la mayoría de técnicas que ofrecen unas excelentes prestaciones en ambientes silenciosos, experimentan una gran degradación en sus prestaciones al considerar entornos reales de funcionamiento, especialmente en los ambientes más ruidosos: oficina, calle urbana, coche, avión, ... Así, un factor que suele afectar a los codificadores de voz a baja velocidad viene dado por el ruido de fondo. Esto se debe a que están diseñados específicamente para voz y no se adaptan a la presencia de otro tipo de señales. Esta degradación es completamente indeseable en aplicaciones tales como la telefonía móvil, donde es bastante probable la presencia de ruido de fondo en forma de otras voces, ruido de la calle o ruido de motor debido al propio vehículo.

Se aprecian, pues, grandes divergencias en el procesado de la voz (codificación o reconocimiento, por ejemplo) entre su comportamiento ante condiciones reales o según las condiciones ideales de laboratorio. Por este motivo, actualmente se destinan muchos recursos a la investigación de técnicas de tratamiento de la señal de voz en ambientes ruidosos, dando origen a las denominadas técnicas robustas de procesado de voz. Especial dedicación recibe el área de Reconocimiento de Voz puesto que el reconocedor está más limitado respecto al caso de un codificador-decodificador donde el receptor sea el oído humano, pues éste se muestra más robusto frente a la inteligibilidad del mensaje. Así, se entiende por Speech Enhancement el procesado de señales de voz destinadas a ser escuchadas, pudiéndose realizar una primera

clasificación dentro de las técnicas de Speech Enhancement según la aplicación a que se destinan:

a) procesado de señales de voz destinadas directamente al oido humano.

b) procesado de señales de voz como preparación previa a otros sistemas de procesado (codificación, reconocimiento, ...), anteriores al oido humano.

El objetivo fundamental que persiguen las técnicas de Speech Enhancement consiste en mejorar uno o varios aspectos perceptuales de la voz: aumentar su calidad global, incrementar su inteligibilidad, reducir el grado de fatiga del oyente, etc. Según cada aplicación particular, el sistema de Speech Enhancement puede orientarse solamente hacia uno de estos objetivos o a varios de ellos. Por ejemplo, un sistema de comunicación de señal de voz puede introducir un eco de baja amplitud y bastante retardo o una perturbación aditiva de banda estrecha de tal manera que estas degradaciones no representan, por si mismas, una reducción de la inteligibilidad, segun los propósitos de uso de este canal. Sin embargo, estos efectos son generalmente desagradables y, entonces, una mejora de la calidad puede ser deseable, aunque sea a expensas de una cierta pérdida de inteligibilidad. En algunos contextos donde las condiciones sean muy adversas, como puede ser el caso de una comunicación entre la cabina de un avión y la torre de control de tráfico aéreo, donde la comunicación en sí ya representa un problema considerable, la mejora de la inteligibilidad es normalmente el único aspecto a considerar, aunque ello implique un sacrificio de calidad. No obstante, en la mayoría de aplicaciones el objetivo consiste en transformar una señal de voz de calidad media en una de mayor calidad.

Llegados a este punto puede ser interesante la distinción entre dos técnicas de procesado de voz muy parecidas, que a menudo se confunden: Speech Restoration y Speech Enhancement. Bajo la denominación de Speech Restoration se incluyen todas las técnicas que persiguen como objetivo principal el procesado de la señal de voz degradada para que se parezca tanto como sea posible a la señal de voz "ideal" original, donde la voz "ideal" dependerá de cada entorno de aplicación y, a menudo, no será conocida. En cambio, las técnicas de Speech Enhancement presentan como objetivo básico el hecho de que la voz procesada suene mejor que la no procesada. Como ejemplo ilustrativo, la señal de voz original no puede ser restaurada (Speech Restoration) pues ya coincide con la "ideal", pero si puede ser realzada (Speech Enhancement) mediante la aplicación de un filtrado paso alto para que suene más clara y fresca, percibiéndose normalmente como una mejora de calidad. Segun lo comentado anteriormente, cuando la señal de voz esté degradada, su posible restauración conduce, a menudo, hacia la aplicación de técnicas de Speech Enhancement.

Las técnicas de Speech Enhancement pueden aplicarse a una gran variedad de contextos. Los principales, y más comunes, a considerar en la presente tesis doctoral, son aquellos donde un ruido interferente, dependiente de cada tipo de situación (ruidos de oficina, ruidos propios de la calle, ruidos del motor de vehículos en la cabina de éstos), se superpone a la señal de voz originándose una pérdida de inteligibilidad y/o calidad de la voz. La señal de fondo que se superpone también puede corresponder a señal de voz procedente de otro locutor en aquellos entornos donde concurren varios locutores. Otras posibles aplicaciones pueden ser:

- las correcciones debidas a la reverberación de salas;

- corrección de distorsiones debidas a dificultades patológicas del locutor, o debidas a un intento de hablar demasiado rápido o quizás debido a que el locutor es un buzo respirando una mezcla sintética de oxigeno (mezcla de Helio y Oxigeno) a cierta profundidad (Hyperbaric Speech);

- realce de la voz original para personas con deficiencias auditivas (Impaired Hearing);

- la reducción del ruido presente en la voz para mejorar la operatividad de los Vocoders.

Naturalmente, la estrategia a elegir está estrechamente ligada al contexto de aplicación considerado, y puede variar considerablemente de un contexto a otro, según la disponibilidad de otras señales o alguna información adicional [Lim-83]. Así, en el caso de la cabina de un avión, se puede realzar la voz mediante la colocación de otro micrófono que capte sólamente el ruido de fondo y use sus características para ajustar y actualizar el sistema de Speech Enhancement. En este supuesto el sistema de realce de la voz se denomina Sistema Multimicrófono. Nótese, sin embargo, que en la torre de control únicamente se dispone de la señal de voz degradada y, por consiguiente, se puede aplicar sólamente alguna de las técnicas conocidas como técnicas de Micrófono Simple, ya que únicamente se dispone de la señal de voz ruidosa.

Las prestaciones de los sistemas actuales de compresión de banda de la señal de voz experimentan una rápida degradación en presencia de ruido aditivo u otro tipo de distorsiones y, por esta razón, existe un considerable interés orientado hacia el desarrollo de sistemas de compresión de voz más robustos. Para afrontar esta situación existen en la actualidad dos metodologías básicas frecuentemente consideradas: técnicas que incorporan una etapa de preprocesado, previa al sistema de compresión de voz, y técnicas que incorporan información relativa a la degradación en el modelo de voz usado por el sistema de compresión de la voz.

Cualquiera de ambas estrategias puede ser la más adecuada, según las condiciones particulares de cada aplicación, sin que se pueda razonar a priori una elección general.

Para la primera estrategia, el sistema de compresión de voz supone que recibe una señal libre de degradación y se desarrolla un preprocesador previo que realce la señal de voz degradada y la prepare para el procesado posterior correspondiente al sistema de compresión de banda de la voz. La efectividad de esta etapa de preprocesado se valora comparando la salida del sistema de compresión de voz cuando se ha considerado la etapa de preprocesado previa y la salida de éste en ausencia de ésta. Puede ocurrir que algún oyente valore, según una medida específica, la salida de la etapa de preprocesado como de calidad inferior a la señal degradada presente a su entrada y, al mismo tiempo, prefiera la salida del sistema de compresión de banda correspondiente a la presencia de preprocesador previo por encima de la salida sin etapa de preprocesado. En este caso, la etapa de preprocesado se considera claramente efectiva en su función de realzar la señal de voz en vistas a prepararla para su posterior compresión en banda.

La segunda estrategia de compresión de banda de una señal de voz degradada consiste en incorporar información acerca de la degradación en el modelo usado para la señal de voz. La señal de voz representa una subclase dentro de las señales de audio y, en consecuencia, existen modelos bastante razonables que permiten una buena caracterización y clasificación de dicha señal de voz. Al tratar de modelar la señal de voz de una forma muy específica se adquiere más potencialidad en relación a suprimir el ruido de fondo presente en esta señal de voz. Pero, por otra parte, cuanto mayor sea el número de suposiciones impuestas, acerca de la señal de voz, mayor sensibilidad presentará el sistema de Speech Enhancement ante posibles imprecisiones o desviaciones en relación a dichas suposiciones. De este modo, la incorporación de restricciones e información relativa a la señal de voz es algo habitual en los distintos sistemas existentes de realce de voz. Existen multitud de técnicas, en la bibliografía, basadas en la forma de generarse la voz y en sus características básicas.

De una forma similar, estos sistemas pueden tratar de incorporar información detallada acerca del ruido de fondo. Por ejemplo, el tipo de procesado adecuado cuando el ruido de fondo es otro locutor es distinto en relación al necesario cuando se ataca el supuesto ruido aleatorio de banda ancha. Así, los sistemas de Speech Enhancement también operan de forma distinta según las suposiciones extraidas a partir del ruido de fondo a tratar. Tal como ocurría con la señal de voz, cuanto más el sistema de realce trate de capitalizar las características específicas del ruido, sus prestaciones resultan ser más sensibles ante desviaciones de éstas.

Otra consideración importante en un sistema de Speech Enhancement viene dada por el hecho que los criterios de realce de la voz deben estar en concordancia con la evaluación final que usualmente debe emitir un receptor humano. En distintos contextos, los criterios de evaluación pueden ser bastante diferentes segun si se considera como más importante la calidad, la inteligibilidad u otros atributos de la voz. De esta forma el sistema de realce debe, inevitablemente, tener en cuenta los aspectos propios de la percepción humana, puesto que representa el juez final evaluador de las prestaciones de este procesado. Mientras algunos sistemas están profundamente motivados por consideraciones perceptuales relacionadas con el oido humano, ótros depositan más su confianza en base a criterios matemáticos. En este último caso, el criterio matemático debe ser consecuente con la percepción humana. Aunque no se conoce un criterio matemático óptimo, se puede hablar sobre la existencia de algunos criterios de error que ofrecen mejores prestaciones en base a los aspectos de la percepción humana.

I.2. Estado del Arte.

Las primeras técnicas de Speech Enhancement aparecierón a finales de los años sesenta y principios de la década de los setenta. La mayoría de estas técnicas se caracterizan por su simplicidad y permiten solventar situaciones de realce de la voz que no sean demasiado desfavorables. Así, una de las estrategias más simples para realzar la voz consiste en el empleo de un filtrado paso bajo o paso banda para atenuar el ruido que cae fuera de la banda perceptualmente importante de la voz. Evidentemente, desde sus inicios, la mayoría de Sistemas de Speech Enhancement se han apoyado en aspectos de la producción y la percepción de la voz, pudiéndose afirmar que todos estos sistemas de realce de la voz dependen, en distinta medida, de los aspectos de producción y/o percepción de la señal de voz.

El enmascaramiento y el filtrado paso banda representan dos maneras bastante simples de como sacar provecho de los aspectos perceptuales del sistema auditivo. Otra técnica, que contempla profundamente los aspectos perceptuales, fue propuesta por Thomas y Niederjohn [Thom1-68b] para aplicaciones donde se dispone de la señal de voz original. Este método consiste en realizar un preprocesado, previo a la degradación de la señal de voz mediante ruido. Este método realiza un filtrado paso alto para reducir o eliminar el primer formante seguido de un recorte infinito. La idea general detrás de este método es: para una SNR fija, el recorte infinito aumenta, en relación a las vocales, la amplitud de los sucesos de baja amplitud y perceptualmente importantes y esto provoca que las consonantes sean menos sensibles en relación al ruido de enmascaramiento. Además, para el caso de los sonidos vocálicos, este filtrado incrementa el valor relativo de los formantes superiores y, de esta manera, adquieren una menor sensibilidad frente a la degradación. Un claro inconveniente de esta técnica viene dado por la implícita disponibilidad de la señal de voz original, lo cual no representa una aplicación realista segun lo comentado anteriormente. Esta idea relacionada con la inmunización de la señal de voz original, como un paso previo a la degradación del ruido, fue bastante típica entre los primeros autores, desde la década de los cincuenta hasta los inicios de los setenta. Estos sistemas suelen distinguirse por lograr mejoras de inteligibilidad bastante importantes.

En cambio, aquellos sistemas que sólo pueden operar sobre la señal de voz ruidosa suelen obtener mejoras de calidad a cambio de una cierta pérdida de inteligibilidad [Lim-79], [Boll-79]. Un ejemplo que rompe esta regla corresponde al sistema propuesto por Thomas y Ravindran en 1974 [Thom1-74], cuando aplicaron su idea anterior [Thom1-68b] directamente

sobre señal de voz ruidosa. Aunque la calidad puede degradarse debido al filtrado paso alto y al recorte, se obtuvo una notoria mejora en la inteligibilidad al realzar la voz bajo la suposición de ruido aleatorio de banda ancha. La explicación dada por Lim y Oppenheim sobre este hecho [Lim-79] suponía que el filtrado paso alto reduce la máscara de ruido presente en los formantes superiores, perceptualmente importantes, a cambio de penalizar las componentes de baja frecuencia, menos importantes a nivel perceptual.

Otro sistema que considera las características de la percepción humana de una forma amplia fue propuesto por Drucker en [Druc-68]. Basándose en pruebas perceptuales realizadas, concluyó que la confusión entre sonidos plosivos y fricativos era la causa primaria de la perdida de inteligibilidad apreciada cuando la voz se degrada mediante ruido de banda ancha. Este hecho se debe, en parte, a la pérdida de las pausas existentes inmediatamente antes del inicio de los sonidos plosivos. Entonces, la técnica propuesta por Drucker consiste en un filtrado paso alto sobre uno de los sonidos fricativos, el sonido /s/, y en la inserción de pausas cortas antes de cada sonido plosivo. Se presupone que las localizaciones de estos sonidos pueden ser determinadas con total exactitud. Sus trabajos mostraban una significante mejora en la inteligibilidad cuando localizaba estos sonidos de forma manual.

A un nivel perceptual el módulo espectral por trama tiene una importancia primordial mientras la fase espectral es poco importante. Basándose en esta propiedad surgieron varias clases de sistemas de Speech Enhancement cuya motivación común es intentar, por diversos caminos, estimar este módulo espectral para cada trama, sin prestar atención a la fase, y utilizarlo para recuperar o reconstruir la señal de voz original. Dentro de este grupo de metodologías pueden diferenciarse dos grandes familias: las técnicas de sustracción espectral y las técnicas por filtrado óptimo.

El método de Sustracción Espectral presenta la ventaja de una superior simplicidad en relación al Método de Filtrado óptimo de Wiener. Por esta razón fue considerado, en principio, por varios investigadores. Lim en [Lim-78c] evalúa el método de Sustracción Espectral generalizado (ver Apartado II.3.3.2.) para distintos valores del parámetro α=2, 1, 1/2, 1/4 y ruido AWGN. Posteriormente Boll en [Boll-79] lo evalúa para α=1 cuando la degradación es ruido de helicóptero. Ambos autores evaluaban las prestaciones de estas técnicas mediante el Test DRT (Diagnostic Rhyme Test). Boll aplicó, también, esta técnica de Sustracción Espectral como preprocesador de un sistema Vocoder a una velocidad de 2.4 Kbps, obteniendo resultados de calidad e inteligibilidad bastante superiores en relación al supuesto de no considerar el preprocesado. En 1979 Lim y Oppenheim publicaron un tutorial [Lim-79] muy completo conteniendo la mayor parte de técnicas existentes en esta época: examinaron los distintos sistemas de Speech Enhancement existentes y llegaron a la

conclusión que la técnica de Sustracción Espectral era la más efectiva. Este estudio, juntamente con una recopilación de artículos relacionados con el realce de la voz [Lim-83] realizada por ambos autores y editada en forma de libro, es considerada por la mayor parte de autores como la primera contribución pionera en estos temas relacionados con la teoría y aplicaciones del Speech Enhancement.

En esta época, finales de los setenta e inicios de los ochenta, la metodología de Sustracción Espectral y sus variantes vivió una época muy productiva con una multitud de trabajos publicados. Merecen destacarse aquellas variantes que consideraban un alisado o promediado del módulo espectral [Boll-79], aquéllas que controlaban la cantidad de ruido sustraído [Lim-79], [Bero-79] o aquéllas que consideraban distintos grados de alinealidades durante la estimación del módulo espectral [Lim-79], [Weis-74]. El mayor inconveniente de estos métodos de Sustracción Espectral viene dado por el ruido residual que aparece tras procesar la señal de voz ruidosa degradada por ruido de banda ancha. Este ruido residual se conoce también como ruido musical y se compone de varias señales de banda estrecha (tonos) cuyas amplitudes y frecuencias varían con el tiempo, produciendo una sensación muy desagradable. Evidentemente esta gran dedicación recibida por los métodos de Sustracción Espectral se debió principalmente a su simplicidad. Especial mención merece el trabajo realizado por McAulay y Malpass [Aula-80] quiénes formularon la Sustracción Espectral como un problema de estimación segun el criterio de Máxima Verosimilitud.

Esta nueva técnica propuesta por McAulay y Malpass obtiene prestaciones superiores en relación a las otras técnicas de Sustracción Espectral. Su sistema obedece a una idea parecida a la formulada previamente por Drucker en [Druc-68], donde clasificaba la voz en cinco clases distintas (fricativas, nasales, vocales, pausas y deslizantes). McAulay y Malpass mejoraron las prestaciones del método de Sustracción Espectral aplicando este mismo concepto aunque de una forma distinta: consideraron dos estados para la voz, silencio y actividad de voz, y desarrollaron un estimador compuesto por la adición ponderada de diversos estimadores individuales para cada estado. Como la mejor estimación del silencio es cero, este estimador propuesto por McAulay y Malpass equivale al producto de los estimadores correspondientes a la Sustracción Espectral y a la probabilidad a posteriori de que la voz esté presente dada la señal de ruido.

Combinando las ideas básicas contenidas en los trabajos de Drucker, McAulay y Malpass y el modelado AR propuesto por Lim y Oppenheim en [Lim-78a], Ephraim propuso en 1992 [Ephr-92b] una técnica de realce de la voz basada en modelos estadísticos. Ephraim propone modelar la voz mediante modelos ocultos de Markov. Este modelado considera la señal de voz como la composición de un conjunto de subfuentes estadísticamente

independientes, donde cada una representa una clase particular (estado) de sonidos similares a un nivel estadístico. Se considera cada subfuente como un proceso AR y las transiciones entre subfuentes se tratan de forma Markoviana. De este modo la transición hacia un nuevo estado viene determinada por el último estado y así el sistema dispone de memoria, ya que ve la señal de voz como un proceso correlativo. Ephraim ha evaluado su sistema con ruido AWGN para niveles de SNR superiores a los 5dB y sus resultados son bastante superiores en relación al método de Sustracción Espectral. Al aplicar su método a los Sistemas de Reconocimiento de voz sólo ha obtenido mejoras para valores de SNR superiores a los 10dB. Recientemente se ha considerado alguna versión modificada de la estrategia de Sustracción Espectral tal como la Sustracción Espectral Alisada (SSS) [Arsl-95] para limitar el ruido musical residual propio de estas técnicas.

Otros sistemas de realce de la voz se centran en aspectos relativos a la producción de la señal de voz, de formas muy diversas. Así en [Lim-79] se presentan varios sistemas que aprovechan la información relacionada con la periodicidad de la voz durante los sonidos sonoros. Para este tipo de sonidos, el espectro de la voz presenta una estructura armónica (Fig.II.4) que sugiere la posibilidad de aplicar un filtrado en peine. Una de sus variantes, debida a Parsons [Pars-76] intenta extraer las componentes del espectro de la voz sólamente en estas frecuencias armónicas. Es decir, el conocimiento de la estructura armónica correspondiente a los sonidos sonoros nos permite eliminar, en principio, el ruido existente en las bandas frecuenciales entre armónicos. Este realce de la voz mediante un filtro en peine también puede verse como un promediado de sucesivos periodos para cancelar parcialmente el ruido. Posteriormente Sambur en [Samb-78] propuso una técnica que intenta sacar provecho de la naturaleza cuasiperiódica de la voz, basándose en los principios de cancelación adaptativa de ruido.

En el modelo de producción de la voz, representado en la Fig.II.7, se genera la señal de voz excitando un sistema lineal cuasiestacionario mediante un tren de pulsos (sonidos sonoros) o mediante ruido (sonidos sordos). Otra posible estrategia, basada en dicho modelo, consiste en tratar de estimar los parámetros del modelo en lugar de la propia voz y, entonces, usar éstos para sintetizarla. Es decir, se realza la voz mediante un sistema de análisis-síntesis de voz. Una aplicación novedosa para este tipo de técnicas se debió a Miller quién en su tesis doctoral (1973) suprimió los acompañamientos musicales presentes en viejas grabaciones de Enrico Caruso [Lim-79]. Para estimar la respuesta impulsional del filtro lineal del modelo de la Fig.II.7 utilizó la deconvolución homomórfica. Suzuki en [Suzu-76] y [Suzu-77] desarrolló una técnica similar para reducir el ruido: se considera la función de correlación de la voz degradada, como una estimación de la respuesta impulsional deseada. Esta técnica se conoce

popularmente como SPAC (SPlicing of Auto-Correlation function) y una de sus variantes como SPOC (SPlicing of crOss-Correlation function).

Con posterioridad aparecieron bastantes sistemas cuya filosofía de trabajo se basa en tratar de modelar el tracto vocal de una forma más detallada. La función de transferencia del tracto vocal se caracteriza por un conjunto de resonancias (formantes), importantes a nivel perceptual. Esto sugiere la posibilidad de representar la respuesta impulsional del tracto vocal mediante un modelo ARMA cuyos polos ofrecen una razonable representación de estos formantes. Los modelos todo polos (AR) tuvieron un notable auge para los sistemas que combinan análisis-síntesis de voz , aplicados sobre voz limpia de ruido. Esto animó a su aplicación para el caso de voz ruidosa y los primeros resultados fueron obtenidos por Lim y Oppenheim [Lim-78a] y Done y Rushforth en [Done-79]. Rápidamente aparecierón extensiones al caso de considerar un modelado mediante polos y ceros (ARMA) como el debido a Musicus y Lim en [Musi-79].

Un inconveniente de esta técnica consiste en que la varianza del estimador de parámetros AR no puede reducirse tanto como sería deseable, debido a que la trama de observaciones ruidosas no suele ser lo suficientemente larga. Sistemas posteriores han evitado este problema considerando la trama de señal ruidosa conjuntamente con otros vectores adyacentes de esta señal de voz durante la estimación del modelo [Hansl-87], [Hansl-91]. Paliwal en [Pali-86], [Pali-87a] propone otra variante de la técnica de estimación AR debida a Lim y Oppenheim [Lim-78a]: inicialmente estima el modelado AR, variante con el tiempo, a partir de la señal de voz ruidosa y luego lo usa para diseñar el estimador que debe originar la señal de voz limpia de ruido. Para ello considera un Filtrado de Kalman y obtiene mejores resultados al compararlo con el Filtrado de Wiener. Hansen y Clements en [Hansl-91] también proponen una versión modificada del Filtrado de Wiener: obtiene mejores prestaciones al imponer restricciones sobre los formantes de la señal de voz, aunque la complejidad de cálculo se dispara a más del doble. Previamente en [Hansl-87] los citados autores evaluaron esta técnica iterativa de Wiener para ruido coloreado.

I.3. Objetivos.

El objetivo principal de la presente Tesis Doctoral consiste en contribuir al estudio de la Técnicas de Speech Enhancement o realce de la señal de voz, especialmente en su aplicación a entornos reales donde el nivel de ruido sea medio-alto.

Tal como se ha expuesto en la introducción, la mayoría de Sistemas de Reconocimiento o Codificación de la voz han sido diseñados a partir de la señal de voz en condiciones de laboratorio, es decir, en un ambiente caracterizado por la ausencia de ruido. Sin embargo, las personas que usan estos sistemas no están encerradas en un habitáculo insonorizado, sino que estos entornos reales de aplicación se caracterizan por la presencia de un ruido de fondo, el cuál puede llegar a ser bastante intenso y molesto en algunas aplicaciones determinadas, tales como la cabina de un avión o la propia Telefonía Móvil. Bajo estas condiciones reales de uso, los sistemas de Tratamiento de la Voz degradan rápidamente sus prestaciones ante la presencia de ruido superpuesto a la señal de voz.

Existen multitud de sistemas diseñados para condiciones libres de ruido y cuyas prestaciones son muy buenas, que al aplicarlos en ambientes ruidosos su utilización resulta poco aceptable, debido a la pérdida de calidad experimentada. Para resolver este problema existe una primera posibilidad consistente en rediseñar, de nuevo, estos sistemas bajo las premisas de la presencia de un ruido de fondo. El mayor inconveniente de esta solución es que los sistemas anteriores pierden validez. Por otra parte, una segunda estrategia permite aprovechar estos sistemas, diseñados en condiciones de laboratorio, simplemente añadiendo una etapa de preprocesado consistente en un sistema de realce de la voz que intente eliminar el ruido presente en la señal de voz. La presente Tesis Doctoral se ubica bajo la consideración de esta segunda posibilidad. Así, los algoritmos presentados se deben interpretar como una etapa de preprocesado que intenta eliminar el ruido presente en la voz, sin causar degradaciones significativas en la señal de voz.

En la literatura puede encontrarseuna gran variedad de trabajos pertenecientes al uso de reconocedores robustos de voz bajo condiciones reales de ruido. Así, por ejemplo, en [Hana-94] se presenta el comportamiento del sistema SPHINX [Lee1-90] ante distintos tipos de ruido presentes en el entorno de un automóvil (motor, viento, radio, calefacción, ruido exterior, ...). Para evitar la degradación del sistema reconocedor se trata de compensar el ruido actuando sobre los coeficientes cepstrales para que el sistema reconocedor pierda cierta sensibilidad al ruido. Los resultados presentados por Hanai y Stern muestran buenas reducciones en la tasa de error de reconocimiento a excepción del ruido correspondiente a

otro locutor procedente de la radio. Para este caso se propone un sistema de cancelación adaptativa de ruido, aprovechando que en esta situación el ruido interferente corresponde a una señal eléctrica disponible dentro del coche. Otros autores como Trompf, Eckhard y Mekhaiel [Trom-94] proponen una reducción de la componente de ruido presente en el dominio del vector de referencias mediante la combinación de técnicas de Sustracción Espectral y redes neuronales.

El procesado de voz en ambientes ruidosos ha recibido especial atención en los últimos años, pero todavía sigue siendo un campo bastante subdesarrollado dentro del área del Procesado de la señal. Por esta razón, la presente Tesis apareció como una primera aproximación a esta problemática, para en pasos posteriores intentar integrar el Sistema de Speech Enhancement dentro del Sistema de Tratamiento de la voz. El carácter innovador de esta temática se pone de manifiesto en el reducido grupo de investigadores dedicados a este tema en comparación a otros campos como el Reconocimiento o la Codificación de la voz, aunque algunas aplicaciones como el reconocimiento robusto de la voz han experimentado una notoria expansión en el pasado más inmediato [Juan-91], [Wilp-90]. El planteamiento general relativo al desarrollo de los ordenadores durante el futuro próximo prevé la comunicación hombre-máquina mediante mensajes orales, es decir, un sistema de reconocimiento de voz en ambientes ruidosos debe ser el dispositivo privilegiado como entrada del mundo exterior. Actualmente, no existen soluciones satisfactorias cuando el nivel de ruido es significativo.

La mayoría de estrategias de reconocimiento robusto de la voz pretenden eliminar parte del efecto debido al ruido actuando sobre los parámetros de representación de la voz [More2-95], [Sand-95], [Nade-94], [Open-94], [Pali-94], [Ster-94], considerando medidas de distancia menos sensibles a la presencia del ruido [Carl2-94], o bién aprovechar ciertas propiedades espectrales de la voz [Haya-94], [Haya-95]. Según el entorno de aplicación donde se sitúa el sistema de reconocimiento merecen destacarse los entornos correspondientes al ruido de coche [Haue-95], [Yang-95] y los ruidos ocasionados por los canales de comunicación (preferentemente linea telefónica) [Reyn-95], [Rahi-94].

Otra estrategia para combatir el problema de la presencia de ruido consiste en colocar uno o más micrófonos en el entorno del locutor para disponer de señales de referencia relacionadas con el ruido presente en la voz disponible en el micrófono principal. Esta estrategia Multimicrófono da lugar al campo de la Cancelación Activa de ruido [Elli-93], [Elli-94] y suele ser habitual cuando se afronta la presencia del ruido en aquellos entornos donde se tiene la presencia del locutor. Así, por ejemplo, numerosas empresas dedicadas a la fabricación de automóviles realizan grandes esfuerzos para eliminar el ruido presente en la

cabina del automóvil. En [Fels-90] se desarrolla un detallado estudio acerca de las prestaciones de estos sistemas segun la ubicación de distintos micrófonos en distintos lugares de la cabina del coche, realizados por la marca Audi-Volkswagen.

Sin embargo, nuestro objetivo consiste en dedicarnos sólamente a las técnicas de Micrófono Simple, es decir, aquellas aplicaciones donde se recibe voz ruidosa y no se dispone de información adicional acerca de las características del ruido presente en el entorno del locutor. Esta situación suele ser habitual en las aplicaciones reales y la única señal disponible es la señal de voz ruidosa y, a partir de ésta, se debe intentar recobrar la señal de voz original. Dentro de esta problemática no se ha considerado el ruido de fondo originado por la presencia de otros locutores. El presente estudio sólamente considera el enfoque hacia la resolución de la degradación originada según el contexto de ruido interferente, donde el ruido presente en el entorno del locutor degrada la señal de voz original, mientras las otras aplicaciones mencionadas (Reberveración de Salas, Impaired Hearing, ...) no se han considerado.

Durante la etapa inicial de esta Tesis se consideraron las técnicas de Sustracción Espectral para solucionar este problema bajo las condiciones arriba mencionadas. Esta etapa fue muy enriquecedora para asimilar toda la problemática relacionada con el Speech Enhancement. A pesar de que las técnicas de Sustracción Espectral destacan por su simplicidad, no nos condujeron a prestaciones satisfactorias ante la presencia de un nivel de ruido importante.

Con posterioridad se consideró la estrategia del Filtrado iterativo de Wiener de una forma adaptativa. Esta técnica básica se debe a Lim y Oppenheim [Lim-79] aunque algunos autores han considerado versiones modificadas [Hansl-91] con posterioridad. Esta estrategia se caracteriza por una mayor complejidad pero también permite afrontar niveles de ruido superiores. Los resultados obtenidos mediante esta técnica básica son bastante buenos hasta que la relación señal a ruido global cae por debajo del margen comprendido entre los 12dB y los 15dB. El objetivo de la presente Tesis consiste en obtener técnicas de Speech Enhancement que, partiendo de esta técnica básica, permitieran afrontar niveles superiores de ruido, donde un sistema de Speech Enhancement tiene más razón de existir. De este modo, en el presente trabajo se han considerado niveles de ruido bastante altos, hasta una SNR de 0dB e incluso, en algún caso, se han considerado valores negativos de la SNR global.

El punto débil de esta técnica de Filtrado de Wiener es que precisa conocer el espectro de la voz original durante la etapa de diseño del filtro. Según lo expuesto previamente, en nuestras condiciones de trabajo no se dispone de la señal de voz original. Bajo este supuesto, se impone la estimación de la voz original a partir de la señal de voz ruidosa. Todos los

autores anteriores, que han considerado esta técnica, realizan dicha estimación segun distintas estrategias (Sustracción Espectral, modelado paramétrico AR, ...) pero con un factor en común: el uso de las estadísticas de segundo orden, es decir, el procesado de la señal de voz ruidosa mediante la función autocorrelación. El principal inconveniente de las estadísticas de segundo orden viene dado por su elevada sensibilidad ante la presencia del ruido.

Por esta razón, un objetivo primordial de este trabajo viene dado por la aplicación de la Estadísticas de Orden Superior (HOS) a la señal de voz. Las características propias de estas estadísticas de orden superior, expuestas en el Capítulo III, hacían prever la obtención de un mayor desacoplo voz-ruido que se debe traducir en una convergencia más rápida de la técnica iterativa de Wiener y, en consecuencia, se pensó alcanzar unos niveles superiores de calidad e inteligibilidad. La combinación de las HOS con técnicas de Speech Enhancement aporta una gran originalidad al trabajo presentado, ya que existen numerosos desarrollos analíticos correspondientes a dichas estadísticas pero sus campos de aplicación real son todavía muy limitados. De este modo este trabajo representa una de las primeras aplicaciones reales de las HOS sobre señal de voz.

Un inconveniente importante de las HOS viene dado por su superior varianza y sesgo durante las estimaciones de los cumulantes (ver Apartado III.2.). Para evitarlo se debe disponer de conjuntos de datos lo suficientemente extensos. Sin embargo, en algunas aplicaciones la no estacionariedad limita fuertemente el número de muestras disponibles para realizar la estimación de los cumulantes de orden superior. La señal de voz es un claro ejemplo y quizás debido a este aspecto no han aparecido muchas aplicaciones de las HOS aplicadas a señal de voz. La mayor parte de aplicaciones reales de las HOS están encaminadas hacia la ecualización ciega de canales reales. En esta aplicación concreta la mayoría de autores han llegado a la conclusión de que se precisa, como mínimo, la disposición de unas 1000 muestras de señal para ecualizar razonablemente el canal [Boss-95], [Gaet-95]. Esta conclusión descarta en principio la posible aplicación de las HOS a los sistemas actuales de telefonía móvil digital como el sistema GSM [Gsm-90] e incluso para los futuros sistemas de telefonía móvil UMTS [Chia-92], previstos para inicios del siglo XXI.

Una primera incógnita a resolver consiste en saber si la voz es suficientemente estacionaria para obtener estas estimaciones de los cumulantes de un modo lo suficientemente fiable. Por esta razón, en el presente trabajo se ha considerado la aplicación de las HOS sobre señal de voz sintética. De este modo se ha pretendido estudiar el comportamiento de los cumulantes de tercer y cuarto orden en su aplicación sobre la señal de voz.

Así, la presente Tesis Doctoral pretende aportar varias posibles estimaciones del espectro de potencia de la señal de voz limpia, que proporcionan originalidad al trabajo presentado, y que permiten obtener sustanciales mejoras, tanto en calidad como inteligibilidad: estimación mediante un modelado AR a partir de estadísticas de orden superior o la estimación efectuada en el dominio de la autocorrelación causal.

I.4. Estructura de la Tesis.

Tras la Introducción inicial referente a las técnicas de Speech Enhancement correspondiente a este Capítulo I, en el Capítulo II se han expuesto las principales metodologías o grupos de técnicas que permiten realzar la señal de voz ruidosa a partir de la disposición de la señal de voz ruidosa. Estas técnicas han sido clasificadas segun las distintas familias existentes, originadas básicamente por los distintos aspectos perceptuales o los aspectos de producción de la voz. Estas técnicas expuestas en el Capítulo II forman la base de partida de todos los algoritmos que se presentan posteriormente en los últimos capítulos, especialmente la estimación paramétrica AR y el Filtrado Iterativo de Wiener.

En el Capitulo III se presentan las Estadísticas de Orden Superior (HOS). La primera parte contiene las definiciones básicas y sus propiedades para el dominio temporal y el frecuencial. En este apartado se definen los conceptos de cumulante, biespectro, slice de cumulantes entre otros. Posteriormente se considera la estimación convencional para los cumulantes así como la estimación paramétrica AR a partir de los cumulantes. Obviamente, este último caso se trata con mayor profundidad y se presentan algunos posibles métodos que conducen a dicha estimación.

En el Capítulo IV se presentan los algoritmos básicos de Speech Enhancement, en los que el presente estudio se fundamenta. Para describir el entorno de simulación, inicialmente se describe la parte común a estos algoritmos, así como las medidas de evaluación objetiva consideradas. Seguidamente se evalúa el comportamiento del algoritmo clásico de segundo orden (AR2) debido a Lim y Oppenheim. Este estudio nos sirve de base de comparación para los distintos algoritmos que se presentan a continuación: los algoritmos de tercer (AR3) y cuarto orden (AR4). Seguidamente se expone un estudio sobre la convergencia de estos tres algoritmos y la posible distorsión ocasionada, así como un estudio realizado sobre señal sintética para evaluar los efectos de la cuasiestacionariedad de la voz en dichos algoritmos. También se exponen algunas versiones modificadas, algoritmos AR3H, AR34, originadas a partir de la consideración de los efectos distorsionadores del algoritmo iterativo de Wiener. Finalmente se presenta el algoritmo realimentado de Wiener (ARre).

En el Capítulo V se consideran versiones modificados de los algoritmos presentados en el capítulo anterior para tratar de acelerar la convergencia de dichos algoritmos sin que ello ocasione un incremento apreciable de la distorsión. De esta manera se ha considerado el algoritmo de Wiener generalizado bajo la consideración de los cumulantes de tercer y cuarto orden, acompañados de sus respectivos estudios de convergencia. Seguidamente se presenta

alguna técnica de promediado intertrama de coeficientes para el caso de segundo orden (AR2_IF) y cumulantes de tercer orden (AR3_IF). Los estudios de estos algoritmos se han efectuado en tres ambientes distintos segun el nivel de ruido considerado.

En el Capítulo VI se ha considerado un algoritmo de segundo orden actuando en el dominio de la función autocorrelación causal (OSA_AR2). Con este algoritmo se persiguen las mismas características comentadas en el capítulo anterior para algoritmos de orden superior. En este caso también se ha considerado la ponderación intertrama (OSA_AR2_IF), especialmente en aquellos ambientes más ruidosos. Finalmente en el Capítulo VII se presentan las principales conclusiones y futuras lineas de trabajo originadas por el presente trabajo.

CAPITULO II

Principales Metodologías de Speech Enhancement.

II.1. Técnicas de Micrófono Simple y Multimicrófono.

Una posible clasificación de los sistemas de Speech Enhancement viene determinada por el número de micrófonos (o señales a procesar) disponibles, clasificándose en: sistemas de micrófono simple y sistemas multimicrófono. Los sistemas de micrófono simple disponen de una única señal a procesar: la voz degradada por un ruido o una interferencia. Este tipo de sistemas de Speech Enhancement serán el objeto de la presente tesis. Su aplicación viene condicionada por la estacionariedad, o cuasi estacionariedad, de la interferencia y, además, requieren una señal degradada cuya relación señal a ruido (SNR) sea positiva en la mayoría de frecuencias que componen el espectro de la voz.

Cuando la interferencia a cancelar sea muy potente o no estacionaria, se impone la necesidad de los sistemas multimicrófono, para poder eliminarla con éxito. Para este tipo de sistemas se requiere un cierto conocimiento del lugar donde está situada la fuente deseada, con respecto a la situación de los distintos micrófonos. Dos técnicas pertenecientes a esta estructura multimicrófono son la Cancelación Activa de ruido [Elli-93], usualmente aplicada a situaciones de ruidos no estacionarios, y las técnicas de Conformación de Haz (Speech

Beamforming) que aprovechan la información espacial para su aplicación a problemas relacionados con las distorsiones multicamino, las reverberaciones o aplicaciones de separación ciega de fuentes [Comp-92]. Una aplicación de esta estrategia al reconocimiento robusto del habla se debe a Sullivan y Stern [Sull-93]: la señal de voz recibida por los distintos micrófonos se procesa por subbandas para simular en cierta manera el comportamiento del oído humano.

El objetivo de la presente tesis doctoral se dirige hacia la resolución del problema de realce de una señal de voz que llega degradada al receptor. Para ello no se dispone de ninguna otra información adicional, tal como ocurre con los sistemas multimicrófono. Sólo se dispone de la salida de un micrófono que contiene la señal de voz ruidosa y por ello nos encontramos ante una de las situaciones más adversas en Speech Enhancement, ya que no se dispone de una señal de referencia del ruido y, además, la señal de voz original no puede ser procesada de forma previa a la aparición del ruido. Las estadísticas de la señal de voz original y del ruido no se conocen y tampoco se dispone de una medida de distorsión plenamente satisfactoria. En consecuencia, todas las soluciones presentadas en los próximos apartados, son subóptimas.

Una buena selección de todas las técnicas de micrófono simple desarrolladas hasta 1982 pueden encontrarse en [Lim-83], [Lim-79]. El deseo de mejorar la robustez que presentaban los vocoders, cuyas prestaciones se degradan rápidamente en presencia de ruido, fue la razón principal que motivó la aparición de estas primeras técnicas de Speech Enhancement. Según lo descrito en el Apartado I.2, referente al Estado del Arte de las técnicas de micrófono simple, se concluye la existencia de dos posibles estrategias para tratar de eliminar el ruido degradador de la señal de voz original:

a) técnicas de Sustracción Espectral

b) técnicas basadas en el filtrado de Wiener.

Ambas familias de técnicas se encuentran descritas en detalle en el Apartado II.3. Otras posibles técnicas como el filtrado en peine (Comb Filter) no se han considerado debido a la importante reducción de inteligibilidad que comportan [Lim-86].

En el presente trabajo se considera el siguiente modelo, donde la señal de voz original $\mathbf{s(n)}$ se ha degradado de forma aditiva mediante una interferencia $\mathbf{r(n)}$ resultando la señal de voz degradada $\mathbf{x(n)}$:

$$x(n) = s(n) + r(n) \qquad \text{(II.1)}$$

bajo la suposición de incorrelación entre la señal de voz original y el ruido. Este modelo permite cubrir, de forma razonable, un amplio margen de situaciones reales donde la interferencia suele ser ruido y la voz degradada se ha obtenido en un ambiente ruidoso. Un aspecto importante que no se considera en este modelo es el efecto Lombard. El efecto Lombard refleja el hecho de que el locutor aumenta la intensidad del habla ante la presencia de ambientes ruidosos [Junq-92]: varía la posición de los formantes de la voz (el primer formante se desplaza hacia frecuencias mayores y su anchura disminuye) y aumenta la duración de los sonidos vocálicos, entre otros efectos. En aplicaciones de reconocimiento de voz ruidosa resulta de crucial importancia la consideración de este efecto Lombard [Mak-92], [Junq-94], pero en las aplicaciones de Speech Enhancement basadas en el Filtrado de Wiener las repercusiones son bastante menores.

La idea básica contenida en los algoritmos de este trabajo hace referencia a la estacionariedad de la señal de voz durante intérvalos de tiempo de corta duración (menores a 32ms). Asimismo, este procesado se realiza frecuentemente en el dominio frecuencial porque intuitivamente nuestro conocimiento es mejor que en el dominio temporal. Además, la característica fundamental viene dada por el módulo espectral, debido a que el oido humano es muy poco sensible frente a las variaciones experimentadas por la fase de una determinada señal. La exposición anterior nos conduce a enventanar la señal y procesarla por tramas, cuya longitud no supere el margen de estacionariedad de la voz y con un solapamiento entre tramas del 50% de su longitud. Aunque la reconstrucción sea virtualmente plana se tiene siempre el efecto de enventanado en el cálculo de la Transformada Discreta de Fourier de cada trama.

Las principales aplicaciones donde se desee la resolución del problema anterior pueden ser :

a) mejorar aspectos perceptuales de la voz ruidosa;

b) inmunizar codificadores de voz frente al ruido presente a la entrada;

c) mejorar las prestaciones de los reconocedores de voz en presencia de ruido.

Este trabajo considera básicamente la primera de dichas aplicaciones. En estas aplicaciones se intenta mejorar fundamentalmente dos características de la voz ruidosa: la calidad y la inteligibilidad. La calidad es una medida subjetiva de cuán placentera suena la voz realzada a los oyentes. Una mejor calidad de la voz comporta, además de un mayor placer, una menor fatiga y una mayor comprensión en los oyentes. La inteligibilidad es, en cambio, una medida objetiva sobre cuánta información puede extraerse de la voz recibida, independientemente de su calidad. Así, los sistemas de Speech Enhancement conducen frecuentemente a una mejora en la calidad de la voz ruidosa a cambio de una cierta reducción de la inteligibilidad.

Obviamente, esta pérdida de inteligibilidad debe ser lo menor posible. Los criterios para evaluar ambas características pueden ser matemáticos (medidas de relación señal a ruido o distancias espectrales) o criterios perceptuales del oyente. Aunque ámbos deben ser consistentes, no se conoce un criterio óptimo calificable como el mejor.

II.2. Técnicas basadas en la producción de la voz.

A continuación se expone una breve descripción sobre algunos aspectos relacionados con la producción de la voz, mientras los aspectos característicos de la percepción humana se tratan en el Apartado II.3. Ambas, la producción y la percepción de la voz, juegan un papel muy importante en los sistemas de realce de la voz. Por esta razón en el Apartado II.2.1. se presenta un breve resumen sobre las principales características de la señal de voz. A partir de la forma como se genera la señal de voz se obtiene un modelo analítico, representado en la Fig.II.7, utilizado en multitud de técnicas que pretenden sintetizar la señal de voz. Seguidamente en el Apartado II.2.2. se presentan las principales técnicas de realce de la voz que hacen uso de estos aspectos de producción de la voz. Finalmente, en el Apartado II.2.3 se discute la estimación del modelado paramétrico AR de la voz.

II.2.1. El Modelado de la señal de Voz.

El conocimiento de las principales características de la voz, así como del mecanismo generador que la define, constituye sin duda el paso previo obligado que nos ha de proporcionar una mejor comprensión de las técnicas a aplicar. El modelado del mecanismo de producción del habla aparece como el punto de partida de las diferentes técnicas de Speech Enhancement en ambientes ruidosos. Gran parte de las primeras técnicas de Speech Enhancement se basaban en la forma de generación de la señal de voz e intentaban sacar provecho de sus aspectos principales. Estas primeras técnicas eran bastante intuitivas y simples, pero sus prestaciones eran limitadas y los niveles de ruido considerados solían ser poco elevados (Apartado II.2.2.). A continuación se presenta un breve resumen acerca de la producción de la señal de voz y sus características fundamentales.

II.2.1.1. La Señal de Voz.

La producción del habla puede verse desde un punto de vista fisiológico o bien desde una perspectiva analítica, más adecuada para el posterior procesado de la señal de voz. Desde el punto de vista fisiológico podemos decir que el responsable de la generación de la voz es el aparato fonador humano, el cual está constituido por tres partes: las cavidades infraglóticas, formadas por los pulmones y la tráquea; la cavidad glótica, donde se encuentran la laringe y las cuerdas vocales; y las cavidades supraglóticas, formadas por la faringe, las cavidades nasal y bucal, y los labios. A efectos de un modelado del sistema se puede simplificar básicamente a una fuente de excitación y dos cavidades resonantes: la cavidad bucal y la cavidad nasal, dado que son éstas las cavidades que definen el tracto vocal.

El locutor produce la señal de voz a partir de variaciones temporales de la presión del aire, que sale expulsado rítmicamente de los pulmones y es modificado a su paso por las cuerdas vocales cuando éste pasa a través de ellas. A partir de la variación de la forma de las cavidades resonantes se modelan los diferentes fonemas, que se unen formando secuencias de sonidos u ondas acústicas. Estos sonidos propios del habla, con unas particularidades y características determinadas, reciben el nombre de alófonos. La cavidad bucal es la que nos proporciona la forma e intensidad de los armónicos de estos sonidos, y por tanto, el timbre de la voz.

En función de la actuación de las cuerdas vocales en el proceso de construcción de la voz podemos dividir los sonidos en sonoros y sordos. Los sonidos sonoros se deben a la vibración lateral de las cuerdas vocales por el paso del aire procedente de los pulmones. Ello genera una forma de onda periódica o cuasiperiódica muy rica en armónicos, denominada

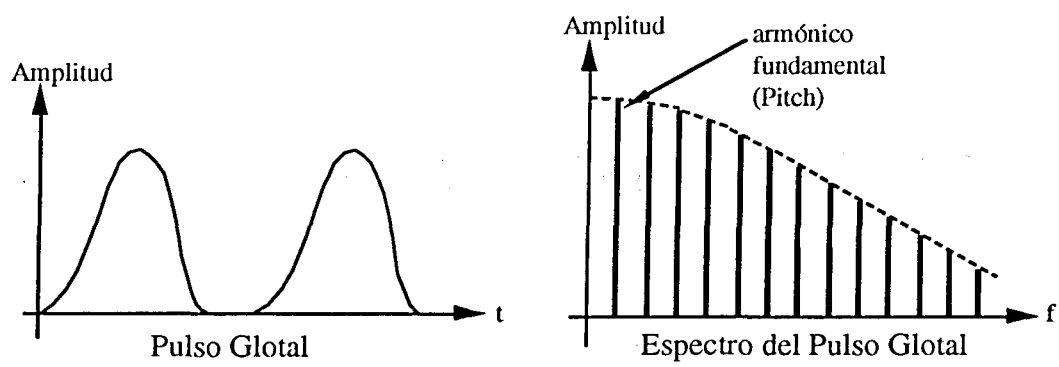


Figura II.1 : Pulso Glotal y su espectro de rayas periódicas.

Pulso Glotal, que actua como señal de excitación del sistema formado por las cavidades resonantes. En la Fig.II.1 puede observarse el espectro de la señal pulsante excitadora. Su periodicidad realmente se restringe a los regímenes estacionarios de producción de sonidos sonoros, por lo que en realidad dichas rayas no son tan puras, sinó que presentan una cierta anchura espectral (ver Figura II.4).

El periodo de repetición de dicha señal, es decir, el armónico fundamental del Pulso Glotal, depende de la frecuencia de vibración de las cuerdas vocales, y se conoce como el pitch del locutor. Esta frecuencia nos determina el tono de la señal que se encuentra en el margen de 80 a 200Hz para los hombres, y entre 150 y 350Hz para las mujeres. Los sonidos sonoros, generados a partir de esa excitación cuasiperiódica, dan lugar a las vocales y a todas las consonantes sonoras, lo que corresponde aproximadamente a unos dos tercios de los sonidos del habla.

Los sonidos sordos resultan del paso rápido del aire expulsado por los pulmones a través de las cuerdas vocales sin provocar su vibración. Este tipo de excitación puede ser modelada por ruido aleatorio de banda ancha, y da lugar a las consonantes sordas de la voz, que pueden ser fricativas o plosivas, y a los transitorios (por ejemplo, paradas y enlace de diptongos).

En la representación temporal de una secuencia de señal de voz se puede apreciar cómo la energía y duración de los sonidos sonoros es mayor, especialmente cuando están acentuados. Vemos en la Fig.II.2 que cada golpe de energía de la señal representa una sílaba o un diptongo. Las consonantes, aunque tienen poca energía (sobretodo las sordas), son de una importancia crucial para la inteligibilidad del mensaje.

En la Fig.II.3.a podemos ver la representación temporal de un sonido sonoro. Se aprecia perfectamente su periodicidad, aproximadamente de 33 muestras, que a una frecuencia de

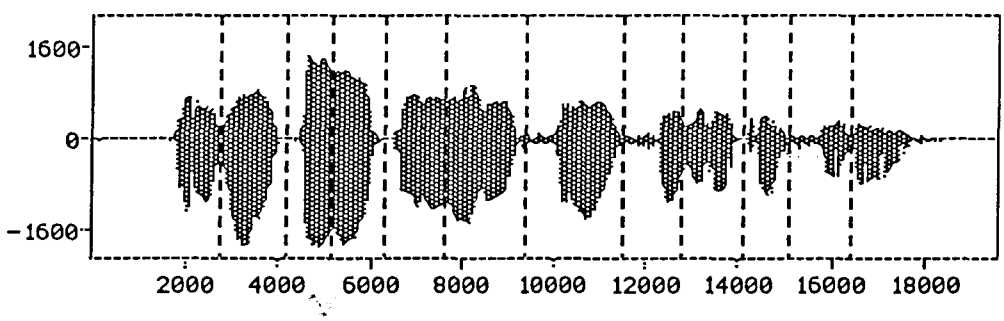


Figura II.2 : Aspecto de la señal en el dominio temporal, 'El-gol-pe-de-ti-món-fue-so-bre-co-ge-dor' *correspondiente al fichero de señal ASUN1.*

muestreo de 8KHz, como es el caso, equivale a un pitch de 242Hz (4.1mseg). En el caso de un sonido sordo, Fig.II.3.b, la señal no presenta la periodicidad anterior y es de baja energía.

Una vez analizadas las dos posibles excitaciones del sistema, cabe ahora examinar cómo se ven modificadas por las distintas cavidades resonantes. El paso de dichas señales excitadoras por la laringe y por las cavidades bucal y nasal, es decir, el tracto vocal, actúa a modo de filtros paso-banda en la selección de los armónicos, dando lugar a un espectro de la señal resultante modulado por la aparición de unas componentes espectrales dominantes. Dichas componentes corresponden a las frecuencias de resonancia del tracto vocal, es decir, las frecuencias de resonancia de la laringe y las cavidades nasal y bucal, y reciben el nombre de formantes. Su número y posición dependen de la articulación de las cuerdas vocales y del tracto vocal en el instante de fonación, variando de un fonema al siguiente y de una persona a otra.

Los principales elementos que intervienen en el tracto vocal son la lengua, los labios, los dientes, el paladar y la mandíbula. Este conjunto de elementos nos permiten modificar la energía dentro del espectro del habla, amplificando unas frecuencias y atenuando otras, de forma que se producen los diferentes fonemas. La cavidad bucal, por ejemplo, puede verse

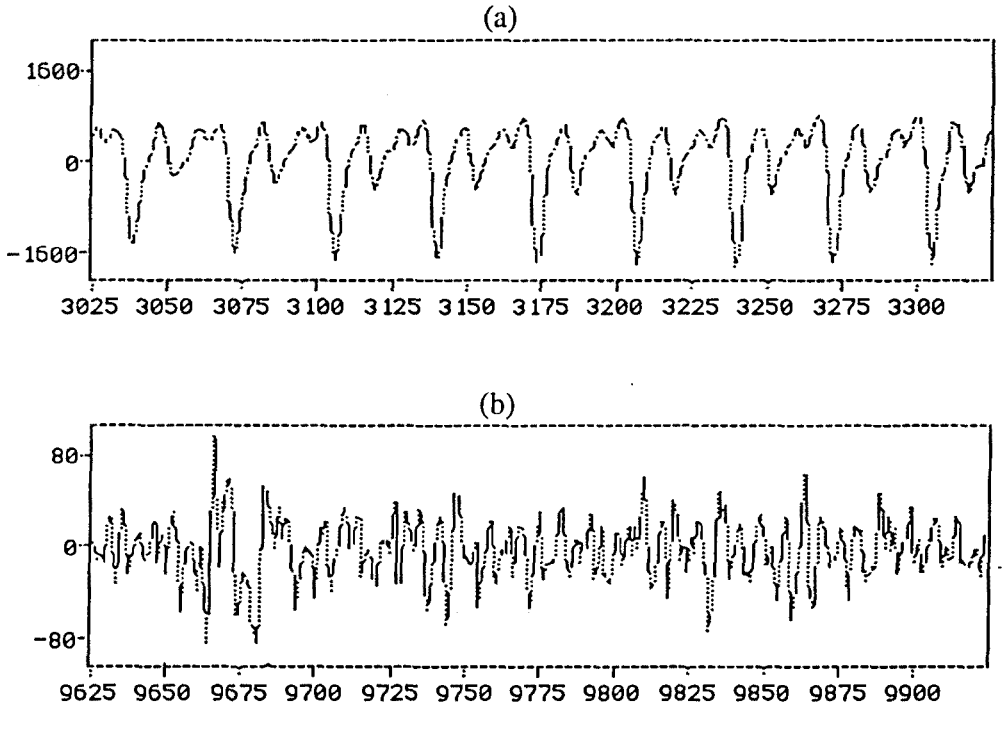


Figura II.3 : a) Intérvalo de 37mseg de un sonido sonoro en el dominio temporal, donde se aprecia su cuasiperiodicidad (periodo aproximado de 4,1mseg equivalente a un pitch de 242Hz); b) representación de un sonido sordo (aperiódico y de baja energía).

pluralizada por la acción de la lengua, pues la variación de su posición da lugar a sonidos totalmente distintos, aún manteniendo el resto del tracto igualmente gesticulado. Acortar la zona de la cavidad supone un aumento de la frecuencia de resonancia.

El último paso para la producción del habla es la radiación al exterior de la señal resultante de todo el proceso. Esta radiación se produce mayoritariamente a través de la boca y supone un filtrado paso-alto de unos 6dB/octava de ganancia, que viene a compensar en cierto modo la atenuación producida en las cavidades resonantes.

Los sonidos sonoros, especialmente las vocales, tienen de dos a cinco formantes. En la Fig.II.4 se ha representado el espectro y la envolvente de un sonido sonoro, donde puede observarse la concentración de energía en aquellas frecuencias discretas múltiplos del pitch, así como la presencia de tres formantes moduladores de la señal.

Normalmente, para el caso de las vocales, sus formantes se encuentran equiespaciados pero su posición y forma varía en función de los fonemas adyacentes. Este fenómeno se denomina coarticulación, y es especialmente influyente en la caracterización de los formantes de las consonantes [Atal-83], [Dix-94]. Se observa también que a partir de los 4KHz la energía del espectro es muy baja, por lo que podremos limitar el ancho de banda a este valor sin perder apenas calidad, aunque si se pierde naturalidad.

Para el caso de sonidos sordos el sistema funciona de forma similar. Aparecen unas frecuencias de resonancia o formantes, pero debido a la menor energía de la señal excitadora y a su aleatoriedad resultan poco marcados, más irregulares y sin periodicidad apreciable, tal

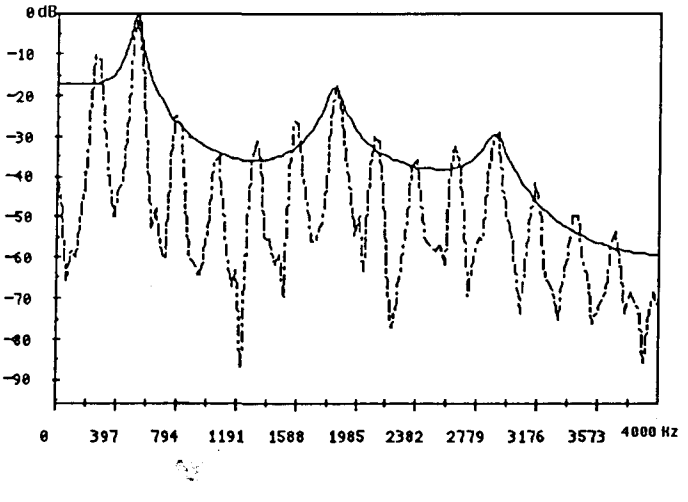


Figura II.4 : Espectros DFT (raya discontinua) y LPC (raya continua) correspondientes a una trama sonora de voz real.

como se puede apreciar en la Fig.II.5. Nótese que el margen de la escala logarítmica en dicha figura es bastante menor.

Existe un tercer tipo de sonidos, que son los nasales, cuyo tracto y radiación difieren del resto de sonidos sonoros debido al acoplo de la cavidad nasal a la bucal. Como consecuencia de esto, su tracto se caracterizará por una función de transferencia con ceros y polos, al contrario de lo que sucede con el resto de sonidos cuyos formantes se modelan solamente a partir de un cierto número de polos en el denominador. No obstante, para los cálculos en tiempo real, la localización de los ceros resulta un factor ciertamente limitador, por lo que prescindiremos de ellos subsanando su falta con el uso de mayor número de polos.

II.2.1.2. Estacionariedad de la señal de Voz.

Podemos considerar la señal de voz como una señal aleatoria que está representada por un conjunto de muestras temporales. Para obtener un mejor modelado, se considera la señal de voz como un proceso estocástico. De este modo se puede caracterizar mediante una serie de funciones de densidad de probabilidad.

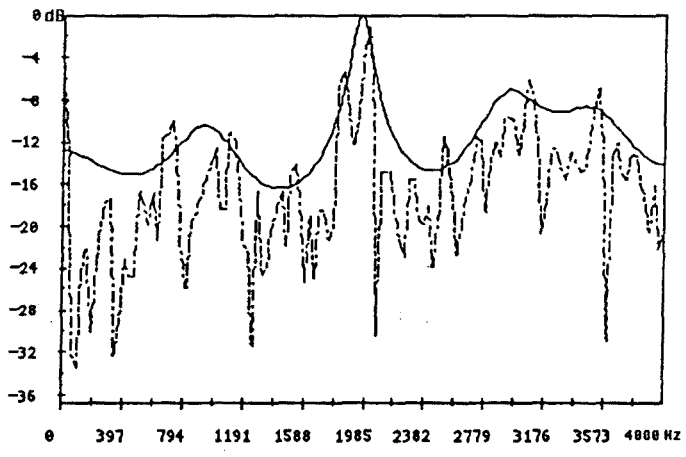


Figura II.5 : Espectros DFT y LPC de una trama sorda, donde no se aprecia ninguna periodicidad ni la presencia de unos formantes claramente marcados.

La señal de voz no es estacionaria, es decir, varía en el tiempo según la variación de las diferentes cavidades resonantes, según la variación del tracto vocal. Sin embargo, al tratarse de un sistema mecánico con inercia, el tracto no puede variar bruscamente: las transiciones entre diferentes fonemas y entre los diferentes niveles de amplitud de la voz se producirán de forma gradual. Por esa razón, se podrá dividir la señal de voz en segmentos o tramas con propiedades constantes, es decir, tramas de duración tal que el tracto vocal se mantenga invariante, para poder analizar y procesar la señal de voz como una señal estacionaria.

Podemos considerar que la señal de voz es cuasiestacionaria durante intervalos de señal de 25 a 35 mseg., en los cuales no hay apenas variación. Si consideramos que estamos trabajando con una frecuencia de muestreo de 8KHz, ello se traduce en tramas de 200 a 280 muestras de duración. Puesto que el valor intermedio más cercano a esos valores que sea potencia de 2 es el 256, lo hemos tomado como valor estándar de longitud de nuestras tramas, de cara a facilitar los futuros cálculos de FFT sobre las distintas señales. De esta manera podemos aislar tramas correspondientes a un sonido sonoro, cuasiperiódico de periodo igual al pitch, y tramas sordas aperiódicas. En cualquier caso, el procesado de la señal por tramas de corta duración se realiza con un solapamiento del 50% entre tramas adyacentes, con lo que aseguramos en mayor medida la estacionariedad de las mismas y limitamos aún más la posible variación del tracto de una trama a la siguiente.

Debido a que la señal de voz, en intervalos suficientemente cortos de tiempo, se comporta como un proceso aleatorio estacionario, podemos decir también que es ergódico, y por tanto podremos calcular estas estimaciones a partir del promediado de las muestras temporales del proceso.

II.2.1.3. Modelado del Tracto Vocal

Desde un punto de vista más analítico, el tracto vocal puede analizarse como un sistema de parámetros distribuidos consistente en un tubo acústico unidimensional de sección discontinua, considerando la señal de voz como una onda plana. Este modelado discreto del tracto, conocido como Modelo Multitubos o de Kelly, se caracteriza por que cada una de las cavidades que quedan determinadas por las diferentes secciones del tubo, llevan asociadas una frecuencia de resonancia f_r que depende de la longitud de la cavidad. Una longitud mayor supone un aumento del retardo de propagación de la energía, y en consecuencia una f_r menor.

Esto repercute directamente en la posición de los formantes de cada fonema. La excitación sigue siendo el Pulso Glotal, que asciende desde la glotis a través del tubo acústico definiendo el pitch característico del locutor.

Considerando las ecuaciones de presión y velocidad volumétrica del aire dentro de cada una de las cavidades de sección constante, su solución queda representada por una combinación lineal de la onda que se propaga de la glotis a los labios (onda progresiva) y de la onda que viaja en sentido contrario (onda regresiva). Aplicando las condiciones de continuidad a la presión y la velocidad volumétrica entre secciones y las condiciones de contorno en los extremos (glotis y labios), se obtienen los coeficientes de reflexión entre cada una de las áreas fronterizas. Una vez calculados esos coeficientes, la propagación del sonido por el tracto vocal queda totalmente descrita. Podremos obtener así la señal de voz de salida a partir de una entrada apropiada.

Se puede comprobar, a partir del modelo multitubos del tracto vocal, que la relación entre la entrada y la salida viene dada por la siguiente función de transferencia todo-polos:

$$H(z) = \frac{g}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \qquad \text{(II.2)}$$

donde **g**, término ajustable de ganancia, y los parámetros $\mathbf{a_k}$, dependen directamente de los coeficientes de reflexión y definen, consecuentemente, el tubo generador. Se trata, por tanto, de un modelado lineal de la producción de la voz, basado en la simulación o caracterización de los formantes a través de los polos de dicha función de transferencia. Los polos se asocian por pares conjugados y están relacionados con los formantes.

La cavidad nasal presenta unas frecuencias de resonancia menores debido a ser una cavidad mayor. Cuando esta cavidad, para sonidos nasales, se acopla a la cavidad bucal,

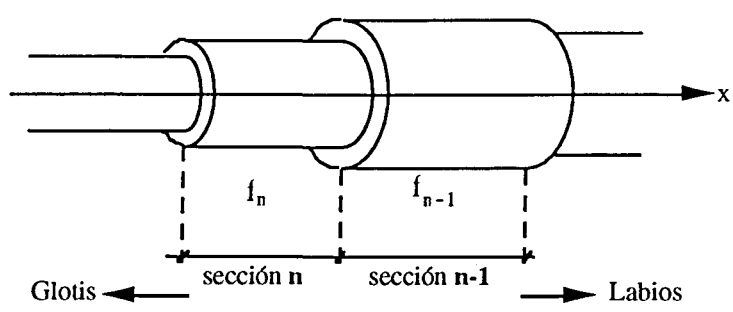


Figura II.6 : Modelo Multitubos del Tracto Vocal.

aparecen algunos ceros en H(z). Pero tal como se ha mencionado en el apartado anterior, añadiremos más polos al modelado con el objetivo de no tener que incluir esos ceros.

Este modelado varía de un fonema a otro y por lo tanto debe ir actualizándose trama a trama para conservar su validez. La variación de la función de transferencia responde directamente a la variación de las cavidades resonantes del tracto vocal y por tanto de la posición, número y ancho de banda de los formantes.

II.2.1.4. Modelo General de Producción de la Voz.

En función de lo que hemos visto en los anteriores apartados, podemos agrupar todo el mecanismo fisiológico de producción del habla en un solo modelo lineal simplificado que contempla los dos tipos de excitación posibles y el modelado del tracto vocal según ya se ha comentado, es decir, mediante una función de transferencia todo-polos. En la Fig.II.7 se ha representado el diagrama de bloques de todo el proceso de producción de la voz, a la salida del cual se obtiene la señal de voz.

El modelo distingue entre sonidos sonoros y sonidos sordos. Para simular la excitación sonora, el Pulso Glotal, se utiliza un tren de pulsos cuyo periodo se corresponde con el pitch deseado. La excitación de los sonidos sordos tiene un espectro plano y se puede modelar por

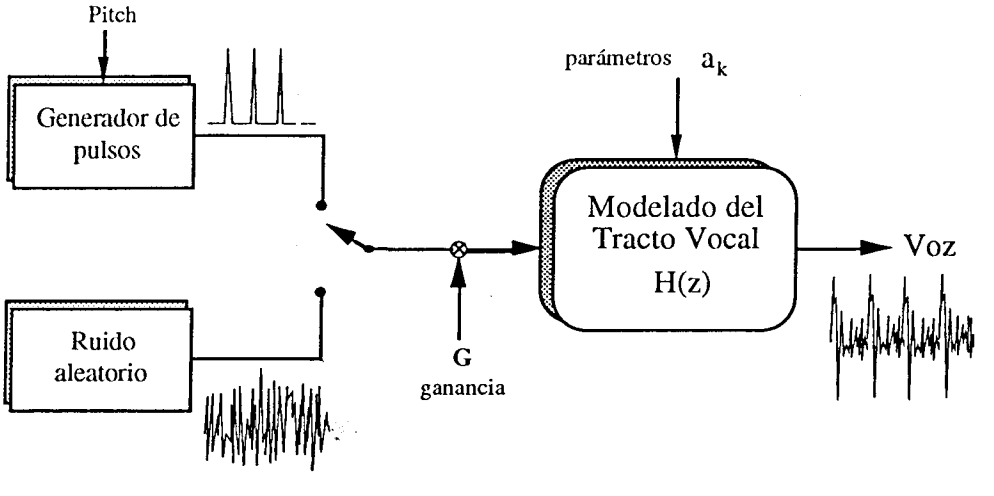


Figura II.7 : Diagrama de bloques del modelo de generación del habla.

un generador de ruido aleatorio con distribución gaussiana. Un conmutador accionado por el conocimiento del tipo de fonema a generar, sonido sordo o sonoro, es el encargado de seleccionar una u otra fuente como entrada al sistema modulante.

El bloque modulador del tracto vocal viene caracterizado por la función de transferencia H(z), caracterizada a su vez por los coeficientes $\mathbf{a_k}$ y el factor **G** de ganancia:

$$G \cdot H(z) = \frac{G}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \qquad \text{(II.3)}$$

Su forma cambia de una trama a otra, a medida que cambia también el tracto vocal.

La mayor parte de técnicas de realce de la voz, tales como la Sustracción Espectral o el Filtrado de Wiener, se fundamentan en la interpretación de la señal de voz como un proceso estocástico. Realmente, esta caracterización de la voz resulta mucho más apropiada para los sonidos sordos, cuando el filtro del tracto vocal se excita mediante ruido espectralmente plano. El tracto vocal varía en su forma a medida que cambian los sonidos generados y, en consecuencia, se debe hablar de una función de transferencia H(z) variante con el tiempo para el sistema lineal de la Fig.II.7. Sin embargo, debido a las restricciones mecánicas y fisiológicas presentes en el movimiento del tracto vocal y las restricciones articulatorias (lengua y labios) parece razonable considerar este sistema como un sistema lineal que varía lentamente, es decir, dentro de una trama de voz (32 mseg) se puede realizar la aproximación de estacionariedad. De esta manera, los atributos anteriormente citados para la voz, envolvente espectral caracterizada por un conjunto de resonancias y la estructura armónica asociada con los sonidos sonoros (Fig.II.4), sólamente tiene validez cuando se consideran intérvalos de tiempo relativamente cortos (tramas). Un desarrollo más detallado acerca de los modelos de generación de la voz puede encontrarse en [Rabi-78].

II.2.2. Técnicas basadas en la Periodicidad de la Voz Sonora.

En este apartado la discusión se dirige hacia aquellas técnicas de Speech Enhancement que centran su enfoque en las características de periodicidad, típicas de las tramas sonoras de la señal de voz. En estas tramas sonoras se aprecia un periodo que se corresponde con la frecuencia fundamental o pitch propio de cada locutor. Limitándose en los principios básicos

de los aspectos característicos de la producción de la voz han aparecido una gran variedad de posibles estrategias. Primeramente se expone la técnica del Filtrado en Peine, caracterizado por dejar pasar los armónicos de la señal de voz y rechazar las componentes frecuenciales entre armónicos. En segundo lugar se considera la extracción de los armónicos de la voz a partir de un espectro de alta resolución, obtenido a partir de la señal de voz ruidosa. Finalmente se presentan algunas técnicas de cancelación adaptativa de ruido que reducen el ruido de fondo mediante la estimación de una señal de referencia obtenida a partir de la periodicidad de las tramas sonoras.

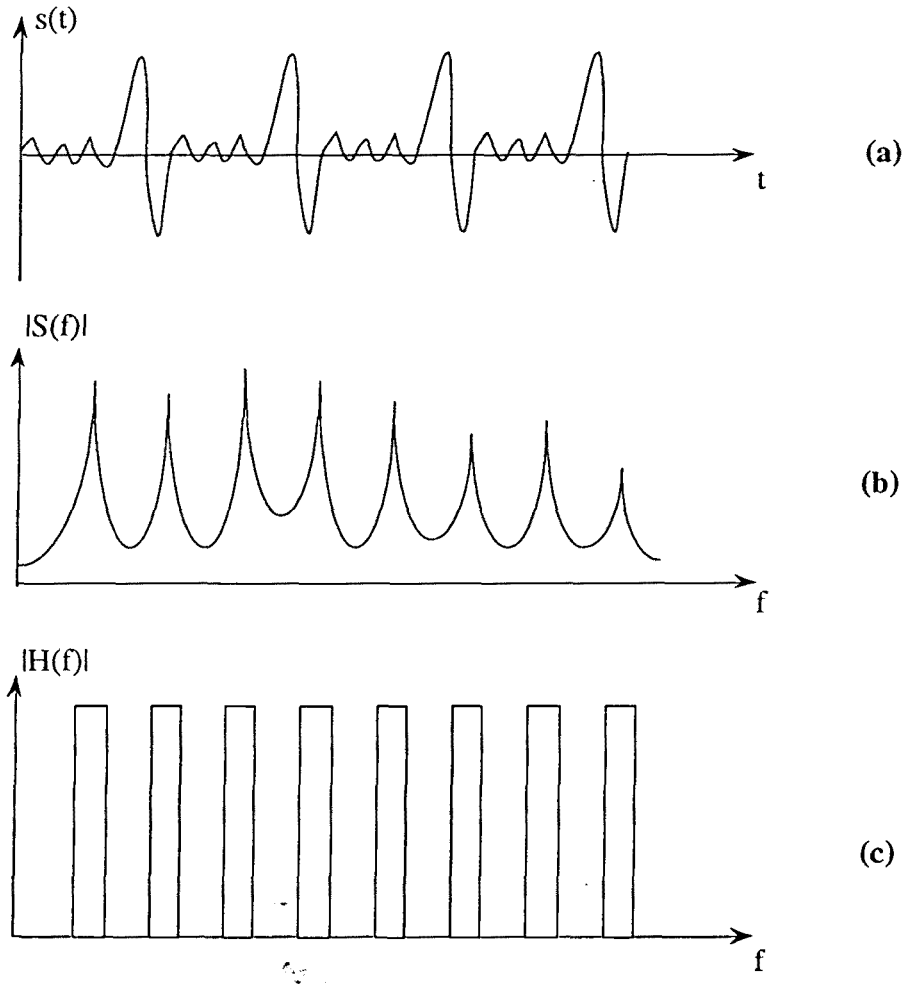


Figura II.8 : a) Señal temporal s(t); b) Módulo del espectro de s(t); c) Respuesta frecuencial del filtro en peine ideal.

II.2.2.1. El Filtrado en Peine adaptativo.

La periodicidad observable en la representación temporal de los sonidos sonoros se manifiesta, asimismo, en el dominio frecuencial con la aparición de armónicos cuya frecuencia fundamental se corresponde con el periodo de la representación temporal, tal como se muestra en la Fig.II.8. En la Fig.II.8.a se muestra un segmento temporal correspondiente a un sonido sonoro. En la Fig.II.8.b y Fig.II.4 se puede observar como la energía de la señal periódica se concentra en unas bandas específicas, correspondientes a dichos armónicos, mientras que las señales interferentes suelen tener la energía distribuida a lo largo de toda la banda del espectro. En estas condiciones, un filtro en peine (Fig.II.8.c) permite reducir el ruido mientras mantiene la señal. Nótese que para ello se debe disponer de información fiel acerca de la frecuencia fundamental de la señal de voz.

Aunque la voz sonora es periódica sólo aproximadamente, el concepto de Filtro en Peine puede aplicarse, sin muchos problemas, para reducir el ruido de fondo presente en la señal de voz ruidosa. Una de las primeras técnicas de realce de la voz mediante Filtrado en Peine se debe a Shields. En la Fig.II.9.a se muestra una respuesta impulsional típica, propuesta por Shields en su Tesis Doctoral [Lim-79], para definir el Filtro en peine. El valor de **T** en dicha figura se refiere al periodo de pitch. Diferentes valores de **T** se consideran al procesar las distintas tramas de voz sonora para adaptar globalmente este filtro a la naturaleza no estacionaria de la señal de voz.

Sin embargo, Frazier y sus colaboradores en [Fraz-76] observaron que, incluso en el caso más favorable donde se dispone de la frecuencia fundamental de forma exacta, el Filtrado adaptativo en Peine de Shields distorsiona considerablemente la señal de voz debido

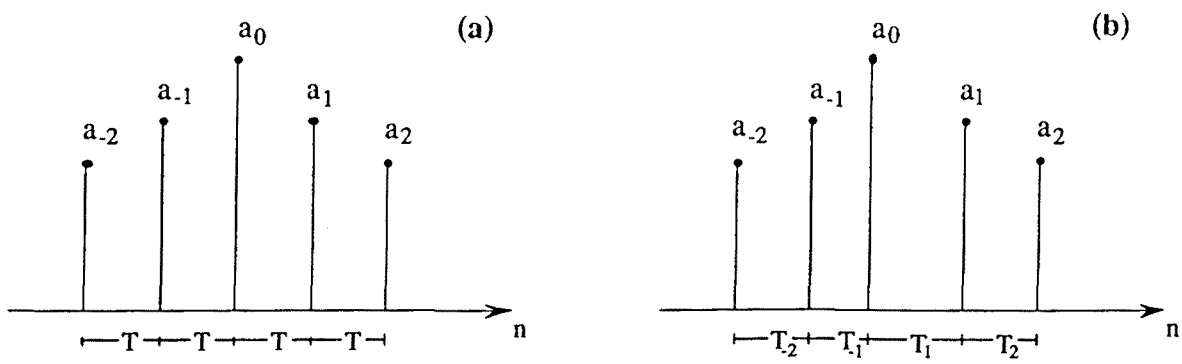


Figura II.9 : Respuestas impulsionales típicas para un Filtro adaptativo en Peine debidas a: a) Shields; b) Frazier y sus colaboradores.

a la naturaleza cambiante de la voz, incluso dentro del intérvalo local (trama) donde se supone estacionariedad. Por esta razón, Frazier y sus colaboradores sugirieron un filtro que se autoadapta de forma global y de forma local en relación a la naturaleza cambiante de la señal de voz. En la Fig.II.9.b se muestra una respuesta impulsional típica de las consideradas por Frazier. En este filtro el espaciado T_i se adapta localmente a las variaciones del periodo de pitch de la voz sonora. En la Fig.II.10 se presenta un algoritmo típico correspondiente al realce de la voz mediante la técnica de Filtrado Adaptativo en Peine.

II.2.2.2. Método de Selección Armónica.

El método anterior precisa de información fiel acerca del pitch y, en consecuencia, precisa de otro sistema para estimar el pitch. Para el supuesto donde el ruido de fondo sea otro locutor, Parsons en [Pars-76] desarrolla un sistema parecido al filtrado en peine pero obtiene la información del pitch como parte integral del mismo sistema. Se enventana cada trama de

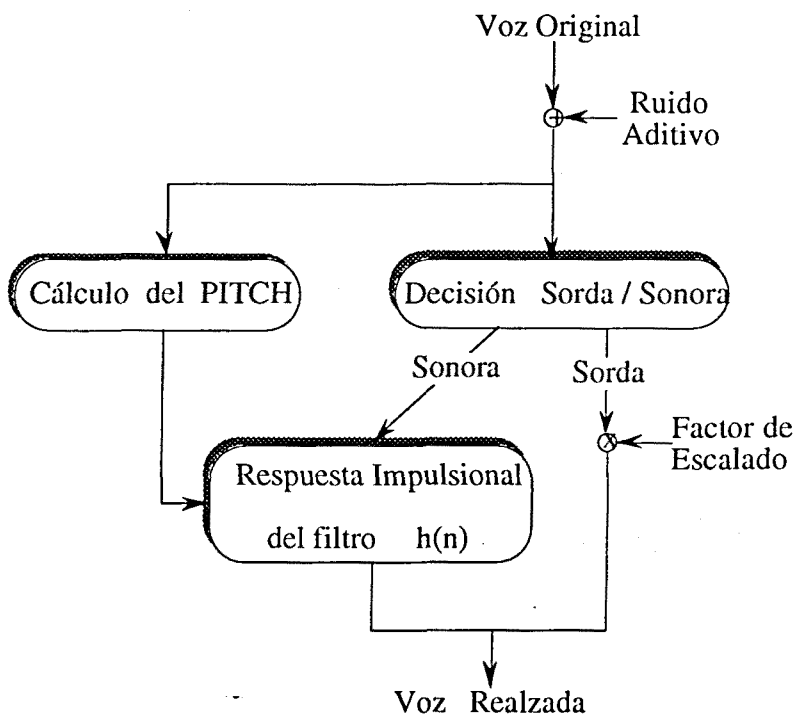


Figura II.10 : Esquema general del Filtrado en Peine Adaptativo.

voz sonora y se obtiene su espectro de alta resolución, donde aparece la periodicidad de la voz en forma de picos. Parsons desarrolló una técnica para distinguir entre los picos correspondientes al locutor interferente y los correspondientes al locutor principal. A partir de la parte seleccionada como perteneciente al locutor principal se genera su voz. Así, la esencia del sistema de Parsons consiste en una localización y selección de armónicos de un locutor a partir del espectro de alta resolución obtenido sobre la señal de voz ruidosa. También puede interpretarse en el dominio frecuencial como un extractor de pitch y un filtro adaptativo.

II.2.2.3. Técnicas de Cancelación adaptativa de Ruido.

Bajo la consideración del modelo aditivo voz-ruido expresado en (II.1), estas técnicas consideran la disponibilidad de la señal de voz ruidosa **x(n)** y una señal de referencia **r'(n)**, la cual está incorrelada con la señal de voz original **s(n)** pero correlada con respecto al ruido **r(n)**. Un diagrama de bloques correspondiente a esta técnica se ha representado en la Fig.II.11. Al filtrar adecuadamente **r'(n)** de una forma adaptativa se estima la componente ruidosa **r(n)** y se puede sustraer de la señal de voz ruidosa **x(n)**, resultando la señal de voz realzada. Nótese como esta técnica precisa de una señal de referencia exterior y, en consecuencia, estaríamos considerando una técnica multimicrófono. Sin embargo, Sambur en [Samb-78] desarrolla un sistema, basado en los principios de cancelación de ruido adaptativa, que genera asimismo la señal de referencia a partir de la periodicidad de la señal de voz.

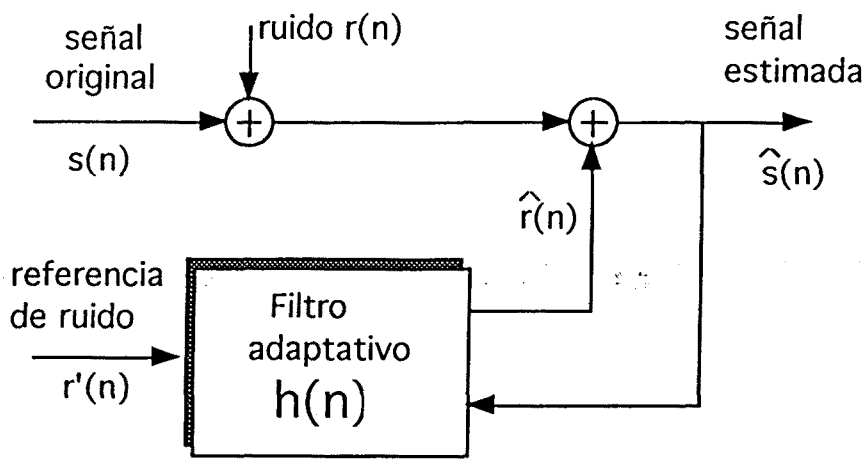


Figura II.11 : Esquema general de las técnicas de cancelación adaptativa de ruido.

II.2.3. Técnicas basadas en un modelado de la voz.

Los sistemas de compresión de banda de la señal de voz estan jugando un rol creciente y muy importante dentro del entorno de los sistemas de comunicación de voz, debido a la creciente importancia de los canales de comunicaciones digitales junto a la necesidad de encriptar la voz y al creciente énfasis en las redes integradas de voz y datos. La base conceptual de los sistemas de compresión de voz de banda estrecha procede de un modelo para la señal de voz. Este modelo se fundamenta en nuestro conocimiento físico y fisiológico acerca de la producción de la señal de voz. Debido a la dependencia en este modelo de la voz, no parece descabellado esperar una degradación en las prestaciones del sistema de compresión de voz, en términos de calidad, inteligibilidad, etc, cuando se introducen degradaciones tales como el ruido aditivo.

En el modelo de producción de la voz, representado en la Fig.II.7, se genera la señal de voz excitando un sistema lineal cuasiestacionario mediante un tren de pulsos (sonidos sonoros) o mediante ruido (sonidos sordos). Otra posible estrategia, basada en dicho modelo, consiste en tratar de estimar los parámetros del modelo en lugar de la propia voz y usar éstos para sintetizarla. Es decir, se realza la voz mediante un sistema de análisis-síntesis de voz. Una aplicación novedosa para este tipo de técnicas se debió a Miller quién en su tesis doctoral (1973) suprimió los acompañamientos musicales presentes en viejas grabaciones de Enrico Caruso. Para estimar la respuesta impulsional del filtro lineal del modelo de la Fig.II.7 utilizó la deconvolución homomórfica. Suzuki en [Suzu-76] y [Suzu-77] desarrolló una técnica similar para reducir el ruido: se considera la función de correlación, obtenida a partir de la voz degradada, como una estimación de la respuesta impulsional deseada. Esta técnica se conoce popularmente como SPAC (SPlicing of Auto-Correlation function) y una de sus variantes como SPOC (SPlicing of crOss-Correlation function).

Con posterioridad aparecieron bastantes sistemas cuya filosofía de trabajo se basa en tratar de modelar el tracto vocal de una forma más detallada. Tal como hemos visto anteriormente, la función de transferencia del tracto vocal se caracteriza por un conjunto de resonancias (formantes), importantes a nivel perceptual. Esto sugiere la posibilidad de representar la respuesta impulsional del tracto vocal mediante un modelo ARMA cuyos polos ofrecen una razonable representación de estos formantes. Los modelos todo polos (AR) tuvieron un notable auge para los sistemas que combinan análisis-síntesis de voz , aplicados sobre voz limpia de ruido. Esto animó a su aplicación para el caso de voz ruidosa y los primeros resultados fueron obtenidos por Lim y Oppenheim [Lim-78a] y Done y Rushforth

en [Done-79]. Rápidamente aparecieron extensiones al caso de considerar un modelado mediante polos y ceros (ARMA) como el debido a Musicus y Lim en [Musi-79].

A continuación se realiza un estudio detallado de una posible estimación de la señal de voz original $\mathbf{P_s}$ a partir de la señal ruidosa mediante un modelado autorregresivo de la señal de voz. En principio no se considera la presencia de ruido, para considerarlo posteriormente y observar las limitaciones y distorsiones introducidas.

II.2.3.1 Estimación de la señal de voz en ambientes silenciosos.

Desde un punto de vista fisiológico, el aire impulsado rítmicamente por los pulmones atraviesa el tracto vocal resultando una onda radiada por la boca. Los pulmones crean una onda pulsante de excitación hacia las cuerdas vocales, las cuales con su vibración modulan el pulso excitante y luego se vuelve a modificar en las distintas cavidades que atraviesa (laringe, nariz, boca) hasta su radiación final. Esta excitación se denomina Pulso Glotal y toma una forma de onda de pulso de aire pseudoperiódico, modulado por el cierre de las cuerdas vocales, causando la apariencia periódica de la voz durante los sonidos sonoros (unos dos tercios de los sonidos producidos durante el habla). Los sonidos sordos resultan del paso rápido de aire a través de las cuerdas vocales sin provocar su vibración. De esta manera, la señal de voz puede modelarse como la respuesta a un sistema lineal cuasiestacionario (Fig.II.7), donde se pueden distinguir dos tipos de excitaciones:

1) para sintetizar los sonidos sonoros se considera una excitación periódica (tren de deltas cuyo periodo fundamental es el pitch)

2) para sintetizar los sonidos sordos se considera una excitación ruidosa (ruido Gaussiano de media nula y varianza unidad)

El tracto vocal resulta, pues, modelado por el siguiente filtro lineal todo polos, también denominado modelo autorregresivo (AR) de orden **p** de la señal de voz :

$$V(z) = \frac{g}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \qquad \text{(II.4)}$$

El problema que se plantea consiste en encontrar los parámetros a_k que permiten modelar el tracto vocal a partir de un cierto número de observaciones de la señal de voz :

$$s(n) = \sum_{k=1}^{P} a_k \, . \, s(n-k) + w(n) + e(n) = \sum_{k=1}^{P} a_k \, . \, s(n-k) + g \, . \, u(n) \qquad (II.5)$$

donde **w(n)** representa la excitación del sistema y **e(n)** el error de modelado cometido. Para el caso de sonidos sordos se puede considerar un factor de ganancia **g** para combinar las dos señales anteriores y considerar **u(n)** como la excitación. Aunque esta relación no se verifica en el caso de sonidos sonoros, la forma de su excitación (tren de deltas) provoca que este factor **g** no tenga apenas influencia en vistas al proceso de estimación, resultando el mismo modelo para ambos tipos de sonidos, obteniéndose la siguiente expresión para su espectro de potencia :

$$P_s(w) = \frac{g^2}{\left| 1 - \sum_{k=1}^{P} a_k \, . \, e^{-jwk} \right|^2} \qquad (II.6)$$

En (II.5) se observa claramente que **s(n)** depende de **2p+1** parámetros: los **p** coeficientes a_k, el factor de ganancia **g** y los **p** valores anteriores de **s(0)** como condiciones iniciales (notado como $\underline{s_i}$). Para su estimación se supone que todos estos parámetros son aleatorios y se les puede asociar una función densidad de probabilidad Gaussiana a priori. A continuación se puede ejecutar la estimación en base al criterio de Máximo a Posteriori (MAP). En principio sólo nos interesa el cálculo de los **p** coeficientes a_k que maximizan la función densidad de probabilidad de $\underline{a}$, supuesto conocido el vector de observaciones $\underline{s_o}$) de la señal **s(n)** :

$$\max \; P(\underline{a} \; / \; \underline{s_o}) \;\; ,, \;\; \underline{s_o}^T = (\, s(N-1), \, s(N-2), \, .. \, , \, s(0) \,) \;\; ,, \;\; \underline{a}^T = (\, a_1, \, a_2 \, , \, .. \, , \, a_p \,) \qquad (II.7)$$

Aunque sólamente nos interesa el cálculo de estos coeficientes a_k del modelo AR de la voz, se puede plantear su cálculo estimando los **2p+1** parámetros mediante la maximización de la función $P(\underline{a}, g, \underline{s_i} \; / \; \underline{s_o})$ suponiendo conocido el vector de observaciones $\underline{s_o}$. Según la forma de tratar la estimación de **g** y $\underline{s_i}$ se pueden distinguir cuatro estrategias distintas que conducen a la estimación de los coeficientes a_k :

Opción 1 : Estimar conjuntamente todos los parámetros $\underline{a}$, **g**, $\underline{s_i}$ suponiendo que no se conoce ninguna información a priori, resultando $P(\underline{a}, g, \underline{s_i})$ constante y el criterio de estimación MAP se corresponde con el de Máxima Verosimilitud (ML). Esta estimación de $\underline{a}$ se corresponde con el Método de Covarianza, conocido en Predicción Lineal :

$$\sum_{n=p}^{N-1} \left[s(n) - \underline{a}^T . s(n-1 : n-p) \right] . s(n-i) = 0 \quad ,, \quad i = 1, \ldots , p \qquad \text{(II.8)}$$

donde

$$s(n-1 : n-p)^T = (s(n-1), s(n-2), \ldots , s(n-p))$$

Opción 2 : Estimar conjuntamente $\underline{a}$, y g suponiendo las condiciones iniciales $\underline{s_i}$ conocidas. Entonces se debe maximizar $P(\underline{a}, g / \underline{s_o} ; \underline{s_i})$ donde $P(\underline{a}, g / \underline{s_i})$ es constante porque $\underline{s_i}$ está totalmente determinada, llegándose al siguiente conjunto lineal de ecuaciones que permiten el cálculo del vector de coeficientes $\underline{a}$ deseado :

$$\sum_{n=0}^{N-1} \left[s(n) - \underline{a}^T . s(n-1 : n-p) \right] . s(n-i) = 0 \quad ,, \quad i = 1, \ldots , p \qquad \text{(II.9)}$$

donde se puede apreciar que la única diferencia respecto a la opción anterior consiste en variar el índice inferior del sumatorio ya que ahora disponemos de **N+p** observaciones de **s(n)**. Corresponde, pues, al mismo Método de Covarianza anterior pero como se ha impuesto la condición de conocer $\underline{s_i}$, esta estimación de $\underline{a}$ se suele denominar Estimación por Máxima Verosimilitud Condicional (CML).

Una alternativa posible consiste en suponer respuesta nula antes de la trama de señal considerada y asignar a $\underline{s_i}$ las primeras **p** observaciones de la presente trama y añadir **p** ceros al final de las **N** observaciones. De este modo resulta la maximización de $P(\underline{a}, g / s(N+p-1: 0))$ respecto $\underline{a}$ y **g**, que conduce al siguiente conjunto de ecuaciones, correspondiente al Método de Correlación conocido en Predicción Lineal :

$$\sum_{n=0}^{N+p-1} \left[s(n) - \underline{a}^T . s(n-1 : n-p) \right] . s(n-i) = 0 \quad ,, \quad i = 1, \ldots , p \qquad \text{(II.10)}$$

donde

$$s(N+p-1 : N) = s(-1 : -p) = 0$$

Una posible y conocida ventaja de este Método de Correlación, respecto al Método de Covarianza, se debe a que origina una matriz Toeplitz y se garantiza la estabilidad de la solución $\underline{a}$ así obtenida.

Opción 3 : Estimar conjuntamente $\underline{a}$ y $\underline{s_i}$ suponiendo g conocida, resultando la maximización de $P(\underline{a}, \underline{s_i} / \underline{s_o}; g)$ que conduce a un conjunto de ecuaciones lineales más complejo, que puede interpretarse como la **opción 1** cuando la función de covarianza de los coeficientes $\underline{a}$ sea arbitrariamente grande.

Opción 4 : Estimar solo $\underline{a}$. Se debe maximizar $P(\underline{a} / \underline{s_o})$ y resulta un conjunto de ecuaciones no lineales, excepto en el supuesto de que g y $\underline{s_i}$ sean ámbas conocidas, en cuyo caso da origen a un conjunto lineal de ecuaciones.

II.2.3.2. Estimación de la señal de voz en ambientes ruidosos.

Todos estos métodos permiten calcular los coeficientes a_k que modelan el tracto vocal con bastante fiabilidad en ausencia de ruido. Sin embargo sus prestaciones se degradan rápidamente ante la presencia de ruido, al considerar la señal de voz ruidosa **x(n)** en lugar de la voz original limpia **s(n)**. Si en los métodos arriba mencionados consideramos la influencia del ruido en las observaciones de la señal **s(n)** se obtiene, para las cuatro variantes anteriores, un conjunto de ecuaciones no lineales para hallar los coeficientes a_k. No obstante, se puede obtener un sistema lineal de ecuaciones al aplicar la consideración de soluciones subóptimas. Para este supuesto se realizan dos suposiciones :

1) las señales de voz **s(n)** y ruido **r(n)** estan incorreladas

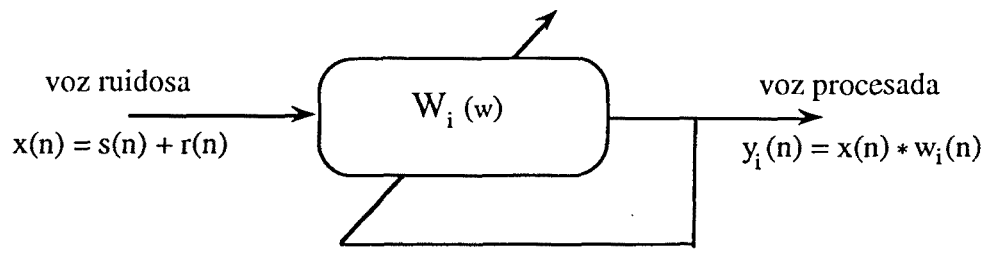
2) el ruido **r(n)** se modela como un Proceso Gaussiano blanco de media nula.

Considerando, por ejemplo, la **opción 4** anterior, se debe maximizar $P(\underline{a} / \underline{x_o})$ suponiendo g y $\underline{s_i}$ conocidas. Para obtener un algoritmo de cálculo más tratable operacionalmente, evitando el sistema no lineal de ecuaciones, se considera una solución subóptima donde $\underline{s_i}$ no tiene mucha importancia. En realidad $\underline{s_i}$ se puede tomar artificialmente nula o bien definirla a partir de las observaciones actuales ruidosas, aunque para el supuesto usual de **N>>p** se ha visto que su influencia final en el resultado es insignificante a un nivel práctico [Lim-78a]. Tampoco se dispone de $\underline{s_o}$ pero se puede obtener una estimación MAP de $\underline{s_o}$ a partir de las observaciones $\underline{x_o}$ de la señal ruidosa disponible, y para unos coeficientes $\underline{a}$

dados. Esto conduce a la consideración de un algoritmo iterativo, donde se parte de una estimación inicial de los coeficientes $\underline{a}^0$ y se estima $\underline{s}_o$ maximizando $P(\underline{s}_o / \underline{a}^0, \underline{x}_o ; g, \underline{s}_i)$ y se obtiene una primera estimación $\underline{s}_o^1$. A partir de esta estimación se obtiene una primera estimación $\underline{a}^1$ de $\underline{a}$. Este proceso se puede repetir iterativamente para obtener una estimación final subóptima $\underline{a}^\infty$ de los coeficientes $\underline{a}$.

Este algoritmo iterativo converge a un máximo local de la función densidad de Probabilidad conjunta $P(\underline{a}, \underline{s}_o / \underline{x}_o ; g, \underline{s}_i)$. Si la estimación inicial $\underline{a}^0$ es tal que conduce al máximo global, o bien la función anterior es unimodal, este método se corresponde con el criterio de estimación conjunta MAP de los parámetros $\underline{a}$ y $\underline{s}_o$. De esta forma el problema de la resolución del sistema no lineal se ha reducido a la resolución sucesiva de varios conjuntos de ecuaciones lineales, que conduce a un Filtro de Wiener no causal tal como se muestra en la Fig.II.12 . Estudios realizados han demostrado que este algoritmo iterativo produce un efecto de estrechamiento del ancho de banda y desplazamiento de los formantes de la voz, tanto para el caso de ruido blanco Gaussiano [Lim-78a] como en el supuesto de considerar ruido coloreado [Hansl-87]. Nótese que la primera estimación del filtro de Wiener se realiza a partir de la señal de voz ruidosa y, en consecuencia, se cumple $\mathbf{y_0(n)=x(n)}$. Este efecto se conoce con el nombre de "picado espectral" y adquiere un carácter más notorio a medida que aumenta el número de iteraciones procesadas por cada trama de voz. Este efecto distorsión se describe detalladamente en el Apartado IV.5.2.

Otra posible aproximación consiste en maximizar $P(\underline{a} / \underline{s}_o ; g, \underline{s}_i)$ tomando los valores de $\underline{s}_o$ estimados en la iteración anterior. Esto conduce al siguiente filtro :



$$W_i(w) = \frac{P_{y_{i-1}}(w)}{P_{y_{i-1}}(w) + P_r(w)} \qquad P_{y_{i-1}}(w) = \frac{g^2}{\left| 1 - \sum_{k=1}^{p} a_k \cdot e^{-jwk} \right|^2}$$

Figura II.12 : Esquema general del algoritmo iterativo de Wiener para la iteración i-ésima.

$$H(w) = \frac{P_s(w) \cdot P_r(w)}{P_s(w) + P_r(w)} \qquad \text{(II.11)}$$

para la estimación final de $\underline{s}_o$. Sin embargo, a diferencia del Filtro de Wiener anterior, la convergencia teórica de este filtro no puede obtenerse, aunque experimentalmente se ha visto que converge en la mayoría de aplicaciones. Presenta la ventaja de producir un menor efecto de picado alrededor de las frecuencias asociadas con los formantes de la voz, cuando varias iteraciones del algoritmo han sido procesadas.

II.3. Técnicas basadas en Aspectos Perceptuales de la Voz.

Los aspectos perceptuales de la voz son considerablemente más complejos en relación a los aspectos de producción de la voz y, en general, su conocimiento no es tan bueno. Sin embargo, existe un determinado conjunto de aspectos de la percepción de la voz, ampliamente aceptados, que juegan un papel primordial en los sistemas de Speech Enhancement. Por ejemplo, se conoce que las consonantes son muy importantes para la inteligibilidad de la voz, a pesar de representar un porcentaje relativamente pequeño de su energía global. También se conoce extensamente que el espectro por trama (short-time) tiene una importancia capital para el estudio de la percepción de la señal de voz y que, en este espectro por trama, los formantes tienen una mayor importancia en relación a otros detalles de la envolvente espectral. El primer formante se localiza, típicamente, en el margen frecuencial comprendido entre 250 y 800Hz y suele ser menos importante que el segundo formante, desde el punto de vista perceptual [Thoml-68a], [Agra-75]. Segun estas características perceptuales, se puede aplicar un cierto filtrado paso alto que puede afectar al primer formante y, al mismo tiempo, no implique una degradación significativa en la inteligibilidad. De forma similar, un filtrado paso bajo, cuya frecuencia de corte este por encima de los 4kHz, puede afectar la naturalidad y la calidad de la voz pero no degrada notoriamente su inteligibilidad.

También se considera de vital importancia, el hecho de disponer de una buena representación para el módulo espectral, mientras que la información de fase se considera muy poco importante. La capacidad de enmascarar una señal con otra señal representa otro aspecto perceptual del sistema auditivo humano, bastante importante en Speech Enhancement. Así, por ejemplo, el ruido de banda estrecha y muchos tipos de ruidos artificiales o degradaciones, como la originada por un vocoder, suenan de una forma más desagradable que un ruido de banda ancha. Consecuentemente, un sistema de Speech Enhancement puede considerar la posibilidad de introducir un cierto nivel de ruido de banda ancha para enmascarar estos ruidos de banda estrecha o artificiales.

II.3.1. La Voz y el Oído.

Desde un punto de vista lingüístico, sabemos que el habla puede verse como un conjunto de 5.000 a 10.000 palabras de uso común, o como un conjunto de 2.000 a 3.000 sílabas, o simplemente como 40 o 50 fonemas distintos enlazados en multitud de combinaciones. Existen otros tipos de divisiones fonéticas utilizadas en otros campos del Procesado de la Señal de Voz, como son el Reconocimiento o la Síntesis, pero en nuestro entorno de trabajo la división básica verdaderamente práctica es sin duda el fonema.

El conocimiento de cómo el aparato de fonación humano genera cada uno de los fonemas de la voz, es indudablemente de vital importancia para el planteamiento de sistemas de procesado que pretendan ser altamente eficientes. No obstante, no hay que olvidar que todo sistema de comunicaciones está compuesto, por lo menos, de un emisor de la información y de un receptor del mensaje emitido. Por lo tanto, habrá que analizar también el comportamiento del aparato receptor humano del habla, es decir, el oído. Sin ánimo de extendernos demasiado con la descripción fisiológica del sistema auditivo humano y de los elementos que lo componen, nos limitaremos al análisis de su comportamiento frente a los distintos tipos de sonidos presentes en la voz.

Se observa que el oído actúa subjetivamente, tanto en frecuencia como en nivel de presión acústica. La sensación percibida no tiene porqué estar directamente relacionada con la tonalidad o sonoridad real de un determinado sonido. Esto lleva a definir las curvas isofónicas o de Robinson-Dadson, que representan la sensación auditiva creada por distintos tonos de diferentes frecuencias y diversos niveles de presión acústica. Cada curva produce la misma sensación de percepción, y lleva asociado un nivel de sonoridad que coincide con el valor numérico del nivel de presión de un tono de 1KHz que tenga la misma sonoridad que la señal, y su unidad se denomina Phon.

De las curvas isofónicas de la Fig.II.13 podemos observar los siguientes puntos:

- Para niveles bajos de señal el oído se comporta de forma no lineal, debemos aumentar mucho los niveles para obtener la misma sonoridad, tanto a bajas como a altas frecuencias.

- El umbral de percepción es de 20mPa a 1KHz, aunque este valor varía en función de la frecuencia, pudiendo ser superior o inferior.

- A medida que aumenta el nivel, la tendencia del comportamiento del oído es igualar la curva de percepción hacia la linealidad (curva más plana).

Esta caracterización de los sistemas tanto de fonación como de audición del habla nos da una idea de las zonas más delicadas y las más superfluas en el procesado de la voz. Es importante tener en cuenta la naturaleza sonora o sorda del sonido, su pitch y el filtro modelador del tracto vocal H(z), siendo las demás características de generación menos importantes. Respecto al oído, se puede afirmar que no es sensible a la fase de los sonidos sonoros, pero sí a la de los sonidos sordos y los transitorios, así como a la distorsión que pueda resultar de un determinado sistema de procesado.

La potencia media asociable a la voz es de 10mW, estando concentrada en su mayor parte en el margen de 500 a 2.000 Hz. La inteligibilidad de la voz tiene mayor peso en la zona de 1 a 4KHz. Esta concentración de la energía por debajo de los 4KHz permite limitar el ancho de banda en aquellas aplicaciones que no requieran elevada calidad; por encima de los 4KHz la energía del espectro es muy pequeña y tan sólo contribuye a mejorar la naturalidad.

Por otra parte, la gran redundancia de información que contiene la señal de voz permite comprender el mensaje con tan sólo parte de él, debido a la lenta variación del espectro y su gran periodicidad. Esto permite una descripción con relativamente pocos parámetros.

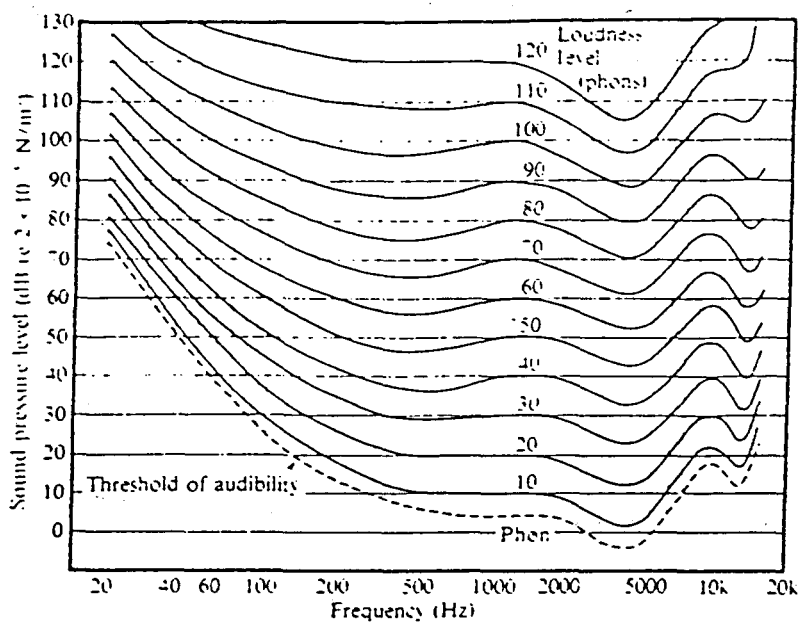


Figura II.13 : Curvas isofónicas.

Los algoritmos de realce y codificación de voz concentran su esfuerzo en no modificar los picos del espectro de amplitud; se basan en explotar la mayor sensibilidad del sistema auditivo a la presencia de energía, dejando más de lado la fase o la energía en frecuencias de menor amplitud. Las partes sonoras, zonas de gran amplitud y concentración de la energía a bajas frecuencias, son más importantes que los fonemas sordos para garantizar la calidad del sistema. La mayoría de algoritmos tienden a mejorar estas partes periódicas correspondientes a los sonidos sonoros.

Algunos tests de audición demuestran que los sonidos sordos y débiles son menos importantes que los sonoros. No obstante la tarea más difícil reside en resolver la eliminación de ruido en las zonas de menor redundancia, las zonas de poco nivel energético y las zonas de transición.

II.3.2. Técnicas Perceptuales.

En apartados anteriores del presente trabajo se ha visto como muchas técnicas de Speech Enhancement tratan de asimilar atributos relacionados con la percepción humana o con la producción de la voz. Este apartado se dedica a aquellos métodos que tratan de incorporar información acerca de la percepción humana. Las primeras estrategias pertenecientes a este grupo consistían en un filtrado paso alto [Thom1-68b], [Thom1-74], aprovechando que la degradación relativa al primer formante es poco perceptible, o bien, colocaban pausas cortas justo antes de los sonidos plosivos [Druc-68]. Sin embargo, estas primeras técnicas son bastante artesanas y presentan serias limitaciones de trabajo, como la necesidad de disponer de la voz original o la localización exacta de los sonidos sordos (fricativos y plosivos).

De una forma más general, cuando se conoce la densidad espectral de potencia del ruido, se puede considerar el uso de un Filtrado de Wiener basado en el espectro de potencia a largo plazo de la voz. Estos sistemas que usan un Filtrado de Wiener invariante con el tiempo hacen uso, sólamente, de las características de la voz promediadas durante varias tramas, tales como el hecho de que el módulo del espectro promediado decae con una pendiente aproximada de 6dB/octava. Aunque en algunos casos esta técnica obtiene buenas prestaciones, como por ejemplo ante la presencia de ruido de fondo de banda estrecha, en

general resulta bastante limitada debido a la no estacionariedad de la voz. Incluso en el supuesto que la voz fuera un proceso estacionario, el criterio de error a minimizar, utilizado para definir el Filtro de Wiener, el criterio MSE o error cuadrático medio mínimo, no está muy correlado con la percepción de la voz y consecuentemente pierde efectividad. Esto se pone en evidencia, por ejemplo, durante el uso de técnicas de enmascaramiento para realzar la voz: al añadir ruido de banda ancha para tratar de enmascarar otras degradaciones más molestas se está, en realidad, incrementando el error cuadrático medio. Otro ejemplo que sugiere que este criterio MSE no contempla correctamente los atributos perceptualmente importantes viene dado por la distorsión que sufre la señal de voz cuando se filtra paso alto: si la respuesta impulsional del filtro paso alto es relativamente corta, no se aprecian diferencias auditivas, pero en cambio, al comparar la señal de voz original y la filtrada puede aparecer un error cuadrático medio bastante importante. Se puede concluir que la medida MSE es sensible a la fase espectral mientras que la percepción humana presenta una tendencia a no serlo.

El enmascaramiento y el filtrado paso banda representan dos maneras bastante simples de cómo sacar provecho de los aspectos perceptuales del sistema auditivo. Otra técnica, que contempla profundamente los aspectos perceptuales, fue propuesta por Thomas y Niederjohn [Thom1-68b] para aplicaciones donde se dispone de la señal de voz original. Este método consiste en realizar un preprocesado sobre la voz original, previo a la degradación de la señal de voz mediante ruido: realiza un filtrado paso alto para reducir o eliminar el primer formante seguido de un recorte ideal. La idea general que hay detrás de este método es que para una SNR fija, el recorte infinito aumenta, en relación a las vocales, la amplitud de los sucesos de baja amplitud y perceptualmente importantes y esto provoca que las consonantes sean menos sensibles en relación al ruido de enmascaramiento. Además, para el caso de los sonidos vocálicos, este filtrado incrementa el valor relativo de los formantes superiores y, de esta manera, adquieren una menor sensibilidad frente a la degradación. Un claro inconveniente de esta técnica viene dado por la implícita disponibilidad de la señal de voz original, lo cual no representa una aplicación realista segun lo comentado anteriormente. Posteriormente, Thomas y Ravindran [Thom1-74] aplicaron esta técnica directamente sobre señal de voz ruidosa. Aunque la calidad puede degradarse debido al filtrado paso alto y al recorte, se obtuvo una notoria mejora en la inteligibilidad al realzar la voz bajo la suposición de ruido aleatorio de banda ancha. La explicación dada por Lim y Oppenheim sobre este hecho [Lim-79] suponía que el filtrado paso alto reduce la máscara de ruido presente en los formantes superiores, perceptualmente importantes, a cambio de penalizar las componentes de baja frecuencia, menos importantes a nivel perceptual.

Otro sistema basado en las características de la percepción humana de una forma amplia fue propuesto por Drucker en [Druc-68]. Basándose en pruebas perceptuales realizadas,

concluyó que la confusión entre sonidos plosivos y fricativos era la causa primaria de la pérdida de inteligibilidad apreciada cuando la voz se degrada mediante ruido de banda ancha. Este hecho se debe, en parte, a la pérdida de las pausas existentes inmediatamente antes del inicio de los sonidos plosivos. La técnica propuesta por Drucker consiste en un filtrado paso alto sobre uno de los sonidos fricativos, el sonido /s/, y en la inserción de pausas cortas antes de cada sonido plosivo. Se presupone que la localización de estos sonidos puede ser determinada con total exactitud. Sus trabajos mostraban una significante mejora en la inteligibilidad cuando se localizaban estos sonidos de forma manual.

Posteriormente aparecieron una serie de métodos bastante más potentes. Tal como se ha comentado durante la discusión de los atributos perceptuales, el módulo espectral por trama tiene una importancia primordial mientras la fase espectral es poco importante. Basándose en esta propiedad surgieron varias clases de sistemas de Speech Enhancement cuya motivación común es intentar, por diversos caminos, estimar este módulo espectral para cada trama, sin prestar atención a la fase, y utilizarlo para recuperar o reconstruir la señal de voz original. La estimación del módulo espectral, a partir de la señal de voz ruidosa, resulta bastante más sencillo en relación al supuesto de tener que estimar el módulo y la fase. Además, según se comentó previamente, la inteligibilidad y la calidad dependen principalmente del módulo del espectro. De esta forma, las técnicas descritas en este apartado se centran en realzar sólamente el módulo del espectro mientras que la fase se toma directamente a partir de la señal de voz ruidosa disponible. Dentro de este grupo de metodologías pueden diferenciarse dos grandes familias: las técnicas de sustracción espectral y las técnicas por filtrado óptimo. La importancia actual y el desarrollo experimentado por estas técnicas durante las dos últimas décadas merecen la dedicación detallada en sendos apartados del presente capítulo. Además, este trabajo se fundamenta en una de estas dos técnicas: el filtrado óptimo de Wiener. Aunque superficialmente estas dos metodologías parecen bastante diferentes, en dichos apartados se muestra como ámbas estan estrechamente relacionadas y bajo ciertas condiciones se pueden identificar.

En la familia de técnicas de Sustracción Espectral, el módulo espectral se estima en el dominio frecuencial a partir del espectro de la voz degradada y se toma la fase de la señal de voz ruidosa y aplicando transformada inversa de Fourier se obtiene la señal en el dominio temporal. Para el segundo grupo de técnicas, a partir de la voz ruidosa se diseña un filtro que después se aplica a esta voz degradada. Como este procedimiento tiene fase nula, sólamente se está realzando el módulo espectral y, en consecuencia, la fase de la señal filtrada o realzada coincide con la de la señal ruidosa. En comparación a las técnicas expuestas en el apartado anterior, nótese que no se efectúa ninguna distinción entre sonidos sordos y sonoros, ni se explota la periodicidad presente en las tramas sonoras de la voz. Ambos grupos se interpretan,

a continuación, mediante la consideración de procesos estocásticos. Aunque esta representación se justifica mejor para los sonidos sordos, se ha visto empíricamente que también puede aplicarse a los sonidos sonoros. En los apartados siguientes se desarrollan detalladamente estas dos familias de técnicas.

II.3.3. La Sustracción Espectral.

La idea original correspondiente a esta metodología de Sustracción Espectral se debe a Weiss y data del año 1974 [Weis-74]. Aunque seguidamente aparecieron estudios sobre su comportamiento y el de diversas de sus variantes [Lim-78c], [Boll-79], [Bero-79]. Las técnicas de Sustracción Espectral constituyen la forma más sencilla de resolución del problema anteriormente especificado. Un estudio comparativo publicado por Lim y Oppenheim en 1979 [Lim-79] comparaba distintas técnicas de Speech Enhancement, conocidas hasta entonces, y llegaba a la conclusión que los sistemas basados en Sustracción Espectral eran los más efectivos cuando se actuaba sobre ruido aditivo incorrelado. Tal como ya ha sido mencionado, la fase de la señal de voz a mejorar no constituye un factor importante para el oido humano, y por este motivo se suele asignar la fase de la señal de voz ruidosa a la señal de voz realzada, presente a la salida del sistema. El problema se reduce a la simple estimación del módulo del espectro correspondiente a la señal de voz mejorada. Una vez se dispone del módulo y la fase de la señal a estimar, en el dominio frecuencial, aplicando transformada inversa de Fourier se obtiene la señal de voz realzada que se estaba buscando.

La técnica de Sustracción Espectral de Potencia es la más simple y por esta razón fue la primera en aparecer [Boll-79]. A continuación aparecieron diversas variantes simples que permiten obtener mejoras de calidad bastante significativas. Alguna nueva versión modificada ha sido objeto de estudio del presente trabajo pero, básicamente, se consideran estas técnicas de Sustracción Espectral para tener una base de comparación respecto a las nuevas configuraciones correspondientes a Filtrado de Wiener, presentadas en el siguiente apartado.

II.3.3.1. Sustracción Espectral de Potencia.

Bajo la suposición de ruido aditivo y estacionario, las señales de voz y ruido pueden modelarse como procesos estocásticos estacionarios e incorrelados segun el modelo representado en (II.1). Si los procesos estocásticos fueran realmente estacionarios se cumpliría la igualdad considerando sus funciones densidad espectral de potencia :

$$\Phi_x(w) = \Phi_s(w) + \Phi_r(w) \qquad \text{(II.12)}$$

pero al ser estacionarios solamente durante intérvalos cortos de tiempo la igualdad se cumple como aproximación, donde a partir de la DFT se obtiene la relación entre sus espectros de potencia :

$$P_x(w) \cong P_s(w) + P_r(w) \qquad \text{(II.13)}$$

donde ya se incluye el efecto enventanado. El espectro de potencia de la señal de voz no degradada se estima como la diferencia entre el espectro de potencia de la señal de voz ruidosa y la densidad espectral de potencia del ruido :

$$\hat{P}_s(w) = P_x(w) - E\{ P_r(w) \} \qquad \text{(II.14)}$$

ya que se puede aproximar el espectro de potencia del ruido por una estimación de su densidad espectral. La expresión (II.14) se aplica para cada coeficiente de la DFT, suponiendo que todos los coeficientes frecuenciales son independientes unos de los otros. De forma estricta, esto sólo es cierto cuando se trabaja con procesos estacionarios y ventanas de análisis infinitamente largas. Como ello no se cumple, se están considerando dos aproximaciones básicas cuando se considera esta técnica de Sustracción Espectral de Potencia:

1) independencia entre coeficientes frecuenciales

2) sustitución del espectro de potencia ruido enventanado por una estimación de su densidad espectral.

Se puede demostrar que, bajo las aproximaciones anteriores y para un conjunto de señales Gaussianas, el método de Sustracción Espectral equivale a una estimación de Máxima Verosimilitud y, al mismo tiempo, a una estimación cuadrática mínima para los distintos coeficientes espectrales :

$$\min \quad E\left\{ \left| P_s(w) - \hat{P}_s(w) \right|^2 \right\} \qquad \text{(II.15)}$$

Finalmente se obtiene el espectro estimado de la voz no ruidosa tomando el módulo de (II.14) y como fase la correspondiente a la señal de voz ruidosa :.

$$S(w) = \left| \hat{P}_s(w) \right|^{1/2} . e^{j\phi x} \qquad \text{(II.16)}$$

Aplicando posteriormente DFT inversa se obtiene la señal de voz deseada **s(n)** . Sin embargo la expresión (II.14) no puede ser aplicada directamente, ya que pueden aparecer valores negativos en el espectro de potencia y por definición se espera que sea una función positiva. En la literatura se ha resuelto el problema mediante la asignación de un valor positivo o un valor nulo a estas frecuencias. En el caso de un valor no nulo se suele asignar un valor umbral fijo para todas o aplicando un cambio de signo al valor negativo. Considerando este problema obtenemos una nueva expresión para (II.14):

$$\hat{P}_s(w) = \max \left[P_x(w) - E\{ P_r(w) \} , P_0 \right] \quad ,, \quad 0 \le P_0 \qquad \text{(II.17)}$$

donde $\mathbf{P_0}$ suele tomar el valor del paso de cuantificación del convertidor A/D y en algunos casos se ha considerado dependiente de la frecuencia.

Otra dificultad que presenta este método viene dada por la estimación del espectro de potencia del ruido. Esta estimación se halla promediando los espectros de potencia correspondientes a varias tramas consecutivas donde se presupone que no hay actividad de voz. Al principio de cada conversación se precisan algunas tramas de silencio para obtener la

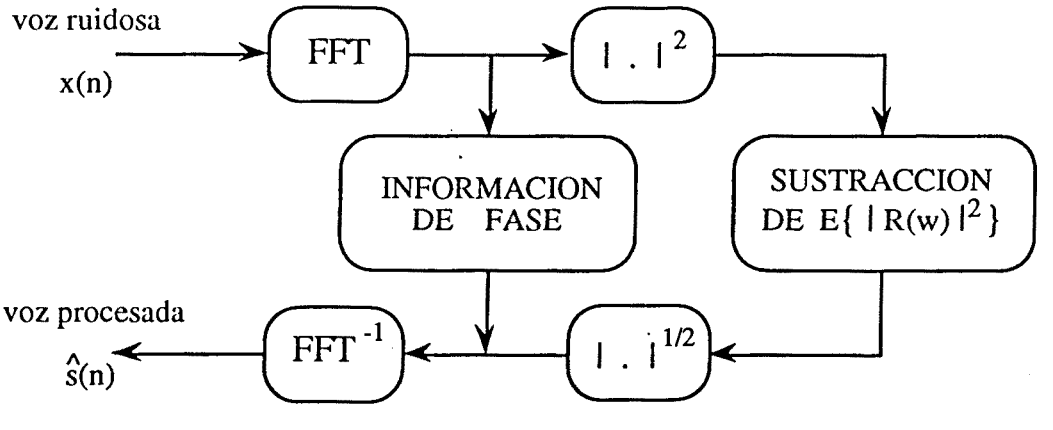


Figura II.14 : *Esquema general de la técnica de Sustracción Espectral de Potencia.*

primera estimación del ruido. Después, como el ruido no suele ser totalmente estacionario, esta estimación debe ser actualizada cada cierto intervalo de tiempo, dependiente de cada aplicación concreta. Se precisa, pues, un detector de actividad de voz-silencio cuyas prestaciones sean buenas incluso para niveles elevados de ruido.

En general, este método permite mejorar la calidad y la inteligibilidad cuando se enfrenta a ruidos de banda muy estrecha, pero deteriora bastante la inteligibilidad cuando el ruido presente es de banda ancha. Además introduce algunas distorsiones bastante molestas, siendo la más importante el "ruido musical", que aparece en la señal de voz estimada debido al efecto de recorte inherente al sistema: los espectros de entrada parecidos a la estimación del ruido promedio presentan, después de aplicar la Sustracción Espectral, un espectro con un cierto número de pequeños picos espectrales distribuidos aleatoriamente con un vacío entre ellos. Esta secuencia de espectros origina en el dominio temporal una continua conmutación de tonos de bajo nivel.

II.3.3.2. La Sustracción Espectral Generalizada.

El método anterior se puede generalizar mediante la introducción de un parámetro α en (II.17) :

$$\hat{P}_s(w) = \left| \max \left[P_x^{\alpha}(w) - \beta.E\{ P_r(w) \}^{\alpha} , P_0^{\alpha} \right] \right|^{1/\alpha} \quad ,, \quad 0 \leq P_0 \qquad (II.18)$$

El factor de escalado ß permite la consideración de subestimaciones (ß<1) o sobreestimaciones (ß>1) del espectro de potencia del ruido, mientras el parámetro α permite numerosas variantes [Lim-79], incluyendo el método anterior (α=1) de Sustracción Espectral de Potencia [Boll-79] y la Sustracción simple del módulo espectral (α=0.5). En [Lim-79] pueden encontrarse los resultados obtenidos con varias generalizaciones y se aprecia un comportamiento ligeramente superior para el caso de Sustracción Espectral de Potencia respecto al de sustracción por módulo espectral. En [Bero-79] se concluye que una sobreestimación del ruido es deseable y, además, se obtienen mejores resultados cuando se efectua un recorte $\mathbf{P_0 > 0}$ respecto al caso de rectificación de media onda $\mathbf{P_0 = 0}$. Considerando lo anterior, se puede alcanzar un compromiso entre una reducción del ruido

musical y un incremento del ruido blanco que resulta menos molesto. Variantes más complejas han sido estudiadas en [Aula-80] y [Ephr-84] .

II.3.3.3. La Sustracción Espectral no lineal.

Considerando que el oido humano presenta un comportamiento logarítmico en el dominio frecuencial, parece más lógico minimizar el error de estimación en el dominio logarítmico espectral, o incluso mejor en el dominio de la intensidad de los sonidos. Desde este punto de vista perceptual no sorprende pensar que, al minimizar el error de estimación en el apartado anterior, aparezcan efectos indeseados como el ruido musical. Sin embargo, cuando la adición de espectros de potencia deja de tener validez, los desarrollos teóricos se complican y se pierde la simplicidad que representaba el método anterior de Sustracción Espectral. El objetivo del presente método consiste en minimizar :

$$\left| L_s(w) - \hat{L}_s(w) \right|^2 \qquad \text{(II.19)}$$

donde $\mathbf{L_s(w)}$ representa el espectro logarítmico de la señal de voz **s(n)**. Un primer intento de enfrentarse a este problema puede encontrarse en [Port-84], donde la distancia espectral logarítmica es uno de los criterios considerados y se compara a una gran diversidad de posibles criterios. Pero ahora nos encontramos ante el hecho que los espectros logarítmicos de

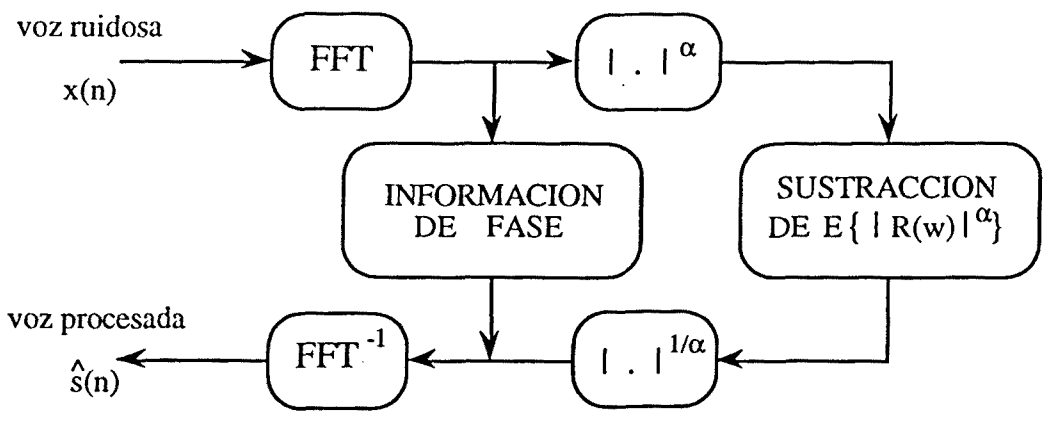


Figura II.15 : Esquema general de la técnica de Sustracción Espectral Generalizada.

la voz y el ruido no son aditivos y, por consiguiente, la estimación bajo el criterio de error cuadrático medio mínimo implica un modelado conjunto de las estadísticas de la voz y del ruido, o bien, las de voz limpia y voz ruidosa. Además, el modelado de ruido debería tener en cuenta el efecto de enventanado que produce un error en la estimación, y también su carácter no estacionario. Para la voz se han propuesto varios modelos (Gaussiana o uniforme logarítmica) mientras se utiliza una distribución Gaussiana para el ruido.

Sin embargo, todos estos modelos comportan desarrollos muy complejos y, además, aparecen importantes contradicciones en las suposiciones de partida. Así, soluciones más empíricas para modelar estas dos distribuciones y/o la implementación de la Sustracción Espectral no lineal han resultado ser una alternativa válida. Además, esto comporta la ventaja de que automáticamente quedan consideradas las relaciones de dependencia entre índices frecuenciales (bins) y los efectos debidos al espectro de ruido enventanado. Ejemplos recientes pueden encontrarse en [Erel-90], para su aplicación al reconocimiento de voz.

Aunque la Sustracción Espectral no lineal se considera una técnica claramente superior a la técnica anterior, especialmente para SNR bajas, su aceptación se ha visto enormemente limitada debido a la complejidad que comporta. Los resultados obtenidos con esta técnica concuerdan con los descubrimientos empíricos de la Sustracción Espectral Generalizada donde se consideraba una sobreestimación del ruido y se situaba un umbral $\mathbf{P_0}$ relativamente positivo.

II.3.3.4. Otras variantes.

Otras posibles soluciones al considerar la metodología de Sustracción Espectral han sido propuestas en la literatura. Especial mención merece la propuesta por McAulay y Malpass [Aula-80], donde se considera un criterio de Máxima Verosimilitud en la estimación por sustracción resultando:

$$\left|\hat{S}(w)\right| = \frac{1}{2} \cdot |X(w)| + \frac{1}{2} \cdot \left[|X(w)|^2 - E\left\{|R(w)|^2\right\} \right]^{1/2} \qquad \text{(II.20)}$$

Otra posible solución se debe a Drucker [Ephr-92b] donde la señal de voz se clasifica en cinco categorías distintas para luego procesarla por caminos distintos segun cada categoría.

En [Ephr-92b] se combinan varios modelos para obtener un modelado de la voz como un conjunto de subprocesos estadísticamente independientes donde cada uno representa una clase particular de sonidos similares estadísticamente. Se asigna cada subproceso a un estado y se consideran modelos ocultos de Markov para modelar la señal de voz.

II.3.4. El Filtrado de Wiener.

Las técnicas de Filtrado de Wiener se basan en la estimación de un filtro óptimo **W(w)** a partir de la señal de voz ruidosa **X(w)** para luego filtrarla y obtener a su salida la estimación de la señal limpia **S(w)**. En principio este filtrado puede realizarse tanto en el dominio temporal como en el frecuencial, aunque en el presente estudio se considera sólamente el dominio frecuencial:

$$\hat{S}(w) = W(w) \cdot X(w) \qquad \text{(II.21)}$$

En principio se estima un filtro óptimo a partir de la voz ruidosa y luego este filtro actúa como una ponderación frecuencial sobre el módulo del espectro de la señal ruidosa. De forma análoga al caso anterior, se supone que la señal de voz ha sido segmentada en tramas y dentro de cada trama la señal ha sido enventanada. Tal como se ha visto en [Weis-74], las técnicas de sustracción espectral también admiten una expresión similar. La diferencia esencial entre estos dos grupos de técnicas reside en la forma de estimar el filtro ponderador frecuencial **H(w)**. Aquí se presenta la estimación de **H(w)** según los principios de la estimación de un filtro óptimo en relación a la aplicación de un criterio de error a minimizar. El estimador lineal de la voz original **s(n)**, que minimiza el error cuadrático medio (MSE), se obtiene al filtrar la voz ruidosa **x(n)** mediante el filtro de Wiener (II.21). No obstante, hacia el final de este subapartado veremos como esta técnica puede conducir a ciertas ponderaciones frecuenciales idénticas a las de algunos casos particulares de sustracción espectral. Tal como se especificó previamente en (II.1), se supone que la voz original **s(n)** y el ruido **r(n)** son procesos aleatorios incorrelados. Si suponemos que ambos procesos son estacionarios y sus respectivas funciones densidad espectral de potencia viene dadas por $\Phi_s(w)$ y $\Phi_x(w)$, se podría obtener la estimación del filtro de Wiener a partir de la expresión siguiente:

$$W(w) = \frac{\Phi_s(w)}{\Phi_x(w)} \qquad \text{(II.22)}$$

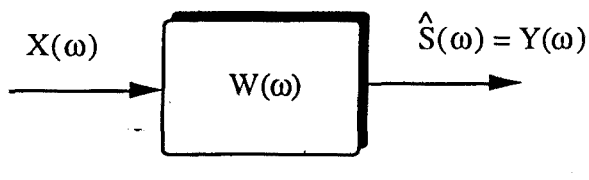


Figura II.16 : Esquema general de la técnica de Filtrado de Wiener.

Es decir, el filtro de Wiener adquiere esta expresión (II.22) cuando $\hat{s}(n)$ es la mejor estimación de $s(n)$ según el criterio MSE. A primera vista se vislumbran dos problemas:

1) este filtro óptimo de Wiener normalmente no es causal y, además, la señal de voz no cumple las premisas de estacionariedad,

2) las funciones densidad espectral de potencia, correspondientes a ambos procesos, deberían ser conocidas con antelación, antes del diseño de dicho filtro.

Como estos procesos no son estacionarios, especialmente la voz, se debe actualizar la expresión del filtro de Wiener para cada trama de ambas señales. De este modo se trata cada trama enventanada por separado y se consideran estimaciones de sus densidades espectrales de energía respectivas $\mathbf{P_s(w)}$ y $\mathbf{P_r(w)}$. Entonces la expresión (II.22) correspondiente al diseño del filtro de Wiener no causal se puede aproximar por la siguiente expresión

$$W(w) = \frac{P_s(w)}{P_s(w) + P_r(w)} = \frac{E\left\{ |S(w)|^2 \right\}}{E\left\{ |S(w)|^2 \right\} + E\left\{ |R(w)|^2 \right\}} \qquad \text{(II.23)}$$

donde el espectro de potencia del ruido $\mathbf{P_r(w)}$ se puede estimar a partir de la estadística del ruido. Normalmente no se suele conocer y se realiza un promediado de $\mathbf{|R(w)|^2}$ a lo largo de varias tramas consecutivas, durante intérvalos de silencio, aplicándose la suposición de estacionariedad a lo largo de todo este intérvalo. En nuestro caso, normalmente se considera la estimación del ruido mediante un alisado del periodograma para intérvalos compuestos de unas 6 a 8 tramas de longitud 256 muestras (Frecuencia de Muestreo F_m=8KHz).

El mayor problema de la presente técnica viene motivado por la estimación espectral de la señal de voz original y, por esta razón, han aparecido diversas estrategias. La primera en aparecer fue la de sustraer espectralmente la estimación del ruido $\mathbf{E\{ |R(w)|^2 \}}$ a partir de la estimación de la voz ruidosa $\mathbf{E\{ |X(w)|^2 \}}$, o bien, aplicando las técnicas de Sustracción Espectral, especificadas anteriormente. En este caso la expresión (II.23) del filtro de Wiener se puede expresar como:

$$W(w) = \frac{P_x(w) - P_r(w)}{P_x(w)} = \frac{E\left\{ |X(w)|^2 \right\} - E\left\{ |R(w)|^2 \right\}}{E\left\{ |X(w)|^2 \right\}} \qquad \text{(II.24)}$$

Los primeros estudios basados en esta técnica consideraban un promedio de $\mathbf{|X(w)|}$, a lo largo de algunas tramas, que posteriormente derivó hacia un alisado. A esta estimación se le sustraía la estimación del ruido $\mathbf{E\{ |R(w)|^2 \}}$, realizada durante los intérvalos cuya actividad de voz es nula. Posteriormente se aproximó $\mathbf{E\{ |S(w)|^2 \}}$ por $|\hat{S}(w)|^2$ o un alisado suyo,

obteniéndose $|\hat{S}(w)|^2$ a partir de alguna variante de las técnicas de sustracción espectral descritas previamente. Nótese que este filtro de Wiener no tiene término de fase y, por consiguiente, se está asignando implícitamente la fase de la señal de voz ruidosa a la estimación de la señal de voz limpia, justamente la misma suposición elegida para el caso de las técnicas de Sustracción Espectral. Por esta razón, algunos autores [Comp-92], [Lim-79] han interpretado el Filtrado de Wiener como un cierto tipo de Sustracción Espectral, con algunas salvedades.

El desarrollo efectuado en este apartado se sustenta en el criterio de error cuadrático medio mínimo, resultando el Filtro de Wiener como el mejor filtro lineal que conduce a un menor error cuadrático medio en la estimación de la señal de voz limpia. En principio, esta medida de distorsión parece bastante pobre, e incluso inferior, de alguna forma, al caso del criterio de Sustracción Espectral de Potencia, pero veremos que las versiones modificadas de Filtrado de Wiener, que se presentarán, conducen a unas prestaciones muy superiores. Los límites de estacionariedad de la señal de voz son muy similares a la longitud de trama considerada y, por esta razón, el filtro de Wiener debe estimarse trama a trama. La estacionariedad de la señal de ruido se puede considerar mucho mayor, en la mayoría de aplicaciones reales, y por este motivo la suposición de adaptar la estimación del ruido aprovechando los periodos carentes de actividad de voz parece ser una medida plenamente satisfactoria, para utilizar dicha estimación a lo largo de la siguiente ráfaga de tramas de voz.

II.3.4.1. El Filtrado de Wiener Generalizado.

La técnica de Filtrado de Wiener puede generalizarse mediante la adición de dos parámetros a la expresión general obtenida en (II.23) :

$$W(w) = \left| \frac{P_s(w)}{P_s(w) + \beta.P_r(w)} \right|^{\alpha} = \left| \frac{E\{ |S(w)|^2 \}}{E\{ |S(w)|^2 \} + \beta.E\{ |R(w)|^2 \}} \right|^{\alpha} \qquad \text{(II.25)}$$

donde los parámetros α y ß permiten hablar de un Filtrado de Wiener parametrizado. De esta manera se pueden obtener filtros con características muy variadas mediante la asignación de distintos valores a ambos parámetros. Por ejemplo, si α=ß=1, la expresión (II.25) se corresponde con la expresión general del filtro de Wiener (II.23). Si ß=1 y α=0.5 la expresión

(II.25) corresponde a la de un filtrado por espectro de potencia, que se caracteriza porque la señal realzada tiene el mismo espectro de potencia $\mathbf{P_s(w)}$, usado en la expresión del filtro (II.25).

Se puede demostrar que ciertos valores de los parámetros conducen al caso de Sustracción Espectral (α=0.5) o al caso de Sustracción Espectral bajo la consideración del criterio de Máxima Verosimilitud (α=1, ß=0.25). Si se estima la densidad espectral de la voz original $P_s(w)$ como:

$$\hat{P}_s(w) = \left| \hat{S}(w) \right|^2 \qquad \text{(II.26)}$$

entonces el módulo espectral de la expresión (II.21) verifica:

$$\left|\hat{S}(w)\right| = \left| \frac{\left|\hat{S}(w)\right|^2}{\left|\hat{S}(w)\right|^2 + \beta . E\left\{ |R(w)|^2 \right\}} \right|^{\alpha} . |X(w)| \qquad \text{(II.27)}$$

resultando una relación en forma implícita para $|\hat{S}(w)|$, referido por algunos autores como Filtrado de Wiener Implícito.

Si se considera un valor α=0.5 esta relación (II.27) ofrece dos soluciones para la variable $|\hat{S}(w)|$:

$$\left|\hat{S}(w)\right| = 0 \qquad \text{(II.28.a)}$$

$$\left|\hat{S}(w)\right| = \left| |X(w)|^2 - \beta . E\left\{ |R(w)|^2 \right\} \right|^{1/2} \qquad \text{(II.28.b)}$$

De este modo una solución consecuente para $|\hat{S}(w)|$ en (II.27) sería tomar (II.28.b) siempre que origine valores positivos y la solución nula (II.28.a) para el resto de valores frecuenciales. Obsérvese que esta particularización del Filtrado de Wiener se corresponde con el método de Sustracción Espectral descrito por (II.18) tomando α=2.

De forma similar, para α=1 la ecuación (II.27) conduce a la solución:

$$\left|\hat{S}(w)\right| = 0.5 |X(w)| + 0.5 \left| |X(w)|^2 - 4 . \beta . E\left\{ |R(w)|^2 \right\} \right|^{1/2} \qquad \text{(II.29)}$$

En este caso la solución suministrada por técnicas de Filtrado de Wiener, particularizada para ß=0.25, se corresponde con el caso de aplicar el criterio de Máxima Verosimilitud a las técnicas de Sustracción Espectral, representado por la expresión (II.20).

II.3.4.2. El Filtrado de Wiener Iterativo.

Puede considerarse como otra generalización para conseguir una mejor estimación del espectro de potencia $\mathbf{P_s(w)}$ en (II.23). Al principio se obtiene una primera estimación de P_s para el diseño del filtro de Wiener a partir de la señal de voz ruidosa. Aplicando el filtrado se obtiene una señal de voz más limpia, a la salida del filtrado de Wiener. Parece lógico lograr una mejora significativa si se itera repitiendo el proceso pero considerando para, el diseño del filtro de Wiener, la señal saliente del filtro en la iteración anterior en lugar de la voz entrante, mucho más ruidosa :

$$W_i(w) = \frac{P_{y_{i-1}}(w)}{P_{y_{i-1}}(w) + P_r(w)} \qquad \text{(II.30)}$$

y aplicando el filtrado de la forma siguiente :

$$Y_i(w) = W_i(w) \cdot X(w) \qquad \text{(II.31)}$$

donde para la primera iteración (i=1) se verifica $\mathbf{P_x(w)= P_{y_0}(w)}$, y se corresponde al caso del Filtrado general de Wiener; mientras la salida obtenida durante la ultima iteración del algoritmo se toma para la estimación final de la señal de voz original de la presente trama. Este algoritmo de Wiener iterativo se debe a Lim-Oppenheim [Lim-79]. En principio parece lógico repetir el proceso durante unas cuantas iteraciones, pues permite obtener un mejor diseño del filtro, iteración a iteración, pues se estima a partir de una señal cada vez menos ruidosa a su salida. Sin embargo, el principal inconveniente viene dado, además del aumento de la carga computacional, por la distorsión introducida en cada iteración, que puede conducir a una significativa distorsión final y, por supuesto, a una pérdida de inteligibilidad, cuando varias iteraciones han sido procesadas. Esto produce que, en algunas frecuencias, el filtro estimado no converja hacia el filtro óptimo deseado y, en consecuencia, la estimación del

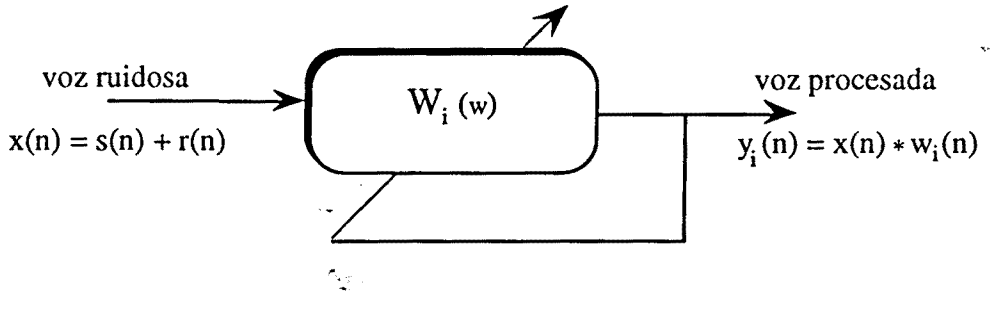


Figura II.17 : Esquema general del algoritmo iterativo de Wiener

módulo espectral $|Y_i(w)|$ no converge para un número infinito de iteraciones hacia la señal deseada $|S(w)|$.

CAPITULO III

Las Estadísticas de Orden Superior (HOS).

Durante las dos últimas décadas, las técnicas de estimación espectral han resultado de vital importancia en la aparición de nuevos sistemas de comunicaciones, sonar, radar, procesado de voz e imagen, biomedicina, oceanografía, mecánica de fluidos o geofísica. Estas técnicas hacen uso, sólamente, de la información estadística de segundo orden, o sea, suponen que las señales o procesos a tratar son inherentemente Gaussianas. Sin embargo, la mayoría de señales pertenecientes a entornos reales de aplicación no son Gaussianas. Así, resulta relativamente sencillo justificar las serias dificultades que estas técnicas de estimación espectral presentan cuando se consideran entornos reales. Las señales deterministas o estocásticas tienen, intrínsecamente, mucha más información que la contenida en su función de autocorrelación o en su correspondiente espectro. Esta información adicional se puede obtener a partir de los espectros de orden superior, obtenidos a partir de las estadísticas de orden superior.

La reciente explosión en el uso de las Estadísticas de Orden Superior (HOS) se debe, principalmente, a los trabajos relativos a múltiples posibles aplicaciones realizados por Nikias [Niki-87c] y a Lii y Rosenblatt [Lii-82] cuyo estudio demostró que era posible obtener la Función de Transferencia de un sistema de fase no mínima a partir únicamente de su salida, sin precisar la información de la señal a su entrada, cuando se usan las estadísticas de orden superior en lugar del método clásico basado en la función autocorrelación. De esta manera,

actualmente se dispone de multitud de algoritmos de procesado de señal basados en los espectros de orden superior, tanto para su uso comercial como militar. La aparición de componentes hardware de alta velocidad y bajo coste, asociado con una mayor disponibilidad de computadores más potentes, conduce a extraer más información presente en las señales que la extraída en un pasado próximo mediante las estadísticas de segundo orden y, por consiguiente, se pueden tomar mejores soluciones.

Los entornos de aplicación de estas técnicas basadas en estadísticas de orden superior pueden agruparse en tres categorías principales:

a) tratamiento de fenómenos físicos: aplicaciones marinas u oceánicas (resistencia de los barcos frente a las olas, anomalías de temperatura entre la superficie y el fondo marino, ondas gravitatorias de los bancos de peces, ruido, interacción de ondas), terrestres (oscilaciones libres), atmosféricas (presión, turbulencias), interplanetarias (brillos), viento (corrientes, turbulencias), plasmas (interacción de ondas, fenómenos no lineales), electromagnetismo (señales de baja frecuencia) y cristalografía (estructuras).

b) tratamiento de señales unidimensionales: diagnóstico (de asperezas de una superficie, defectos de maquinaria, sistemas mecánicos ruidosos), análisis de vibraciones (reconocimiento de modelos, medidas, detección de golpes), voz (detección de pitch, decisión sorda-sonora, supresión de ruido), ruido (cancelación, bioeléctricos, ruido de engranajes), identificación de sistemas (canales de fase no mínima, entrada-salida), detección, enganche de fase, señales de FM, deconvolución sísmica, supresión de efecto doppler, ecualización ciega, series temporales en economía, potencialidad cerebral y ritmos biológicos, recuperación de armónicos, sistemas no lineales (acoplamiento de fase, Volterra), procesado en arrays, sonar y radar.

c) tratamiento de señales bidimensionales o tridimensionales: imagen (modelado, reconstrucción, restauración, codificación, estimación de movimiento, análisis de secuencias), texturas (discriminación, validación de modelos), tomografía, astronomía y filtrado inverso de imágenes ultrasónicas.

III.1. Introducción.

La mayor parte de Métodos clásicos de estimación espectral, mencionados en el Capítulo II, se caracterizan porque durante la estimación espectral se procesa una señal cuyo espectro se interpreta como una superposición de componentes frecuenciales armónicas, estadísticamente incorreladas, y se obtiene su distribución de potencia a lo largo de sus componentes frecuenciales, perdiéndose las relaciones de fase existentes entre estas componentes. La información contenida en el módulo espectral de una señal se corresponde, esencialmente, con la información presente en su secuencia autocorrelación, y resulta ser suficiente para la caracterización estadística completa de una señal con distribución Gaussiana, cuya media sea conocida. Sin embargo, hay situaciones reales donde es necesario extraer información referente a la desviación que una determinada señal o proceso presenta respecto a la Gaussianidad o, incluso, disponer de la información contenida en su fase espectral y, entonces, el método clásico de autocorrelación resulta ser ciego en lo referente a las características anteriores [Niki-87c], [Mend-91].

En cambio, esta información se encuentra presente en los espectros de orden superior (también conocidos como poliespectros), definidos a partir de las estadísticas de orden superior (cumulantes) de una señal. Casos particulares de estos espectros de orden superior son el espectro de tercer orden (Biespectro), que por definición se corresponde con la Transformada de Fourier de las estadísticas de tercer orden, y el espectro de cuarto orden (Triespectro), definido como la Transformada de Fourier de los cumulantes de cuarto orden de una señal estacionaria. Nótese que el espectro de potencia clásico puede verse como el espectro de segundo orden dentro del entorno de las estadísticas de orden superior (HOS). Así, en la Fig.III.1 se ha representado la clasificación de los espectros de orden superior correspondientes a una señal discreta dada.

Las estadísticas y espectros de orden superior pueden expresarse en función de los momentos o de los cumulantes. Mientras que los momentos, y sus correspondientes espectros, resultan muy provechosos para el análisis de señales deterministas (periódicas o transitorios), los cumulantes y sus espectros son de gran importancia durante el análisis de procesos estocásticos [Niki-93a]. El uso de los espectros de orden superior parece bastante adecuado en determinados campos del procesado de la señal, especialmente, en técnicas para:

a) suprimir ruido aditivo Gaussiano, blanco o coloreado, cuyo espectro de potencia sea desconocido,

b) identificar sistemas de fase no mínima o reconstruir señales de fase no mínima,

c) extraer información relativa a las desviaciones respecto a las características de un proceso Gaussiano,

d) detectar y caracterizar propiedades no lineales de algunas señales determinadas o identificar sistemas no lineales [Niki-87c].

La primera de estas posibles aplicaciones se refiere a la supresión o reducción de ruido Gaussiano que ha degradado una señal no Gaussiana. El contenido de la presente Tesis Doctoral hace referencia a esta aplicación de las HOS, bajo la suposición de considerar una distribución no Gaussiana para la señal de voz. Las HOS pueden aplicarse en este contexto porque una de sus propiedades fundamentales es el tener idénticamente nulos todos los cumulantes, y sus espectros, de orden superior a dos sólamente para el caso de señales con distribución Gaussiana. De esta manera, cuando se recibe una señal no Gaussiana degradada con ruido aditivo Gaussiano, se puede observar esta señal sin degradación al situarnos en el

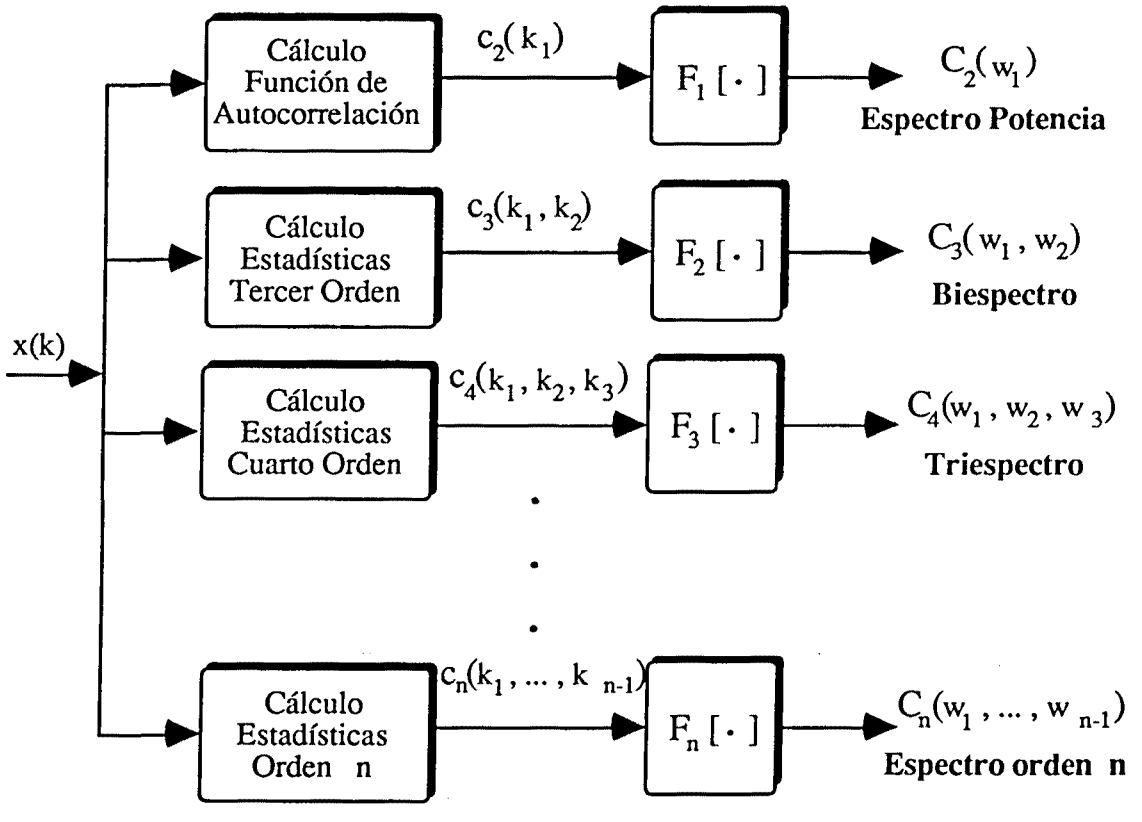


***Figura.III.1** : Esquema de clasificación de los Espectros de Orden Superior para una señal discreta **x(k)**, donde **F_n[.]** representa la Transformada de Fourier de dimensión **n**.*

dominio de los cumulantes de orden superior, y obtener unas prestaciones muy superiores en relación al dominio clásico de los momentos de segundo orden o dominio de la función autocorrelación.

La segunda aplicación se fundamenta en la propiedad que todos los poliespectros, de cumulantes o de momentos, conservan la información de fase de la señal tratada. Durante el modelado de series temporales, en procesado de la señal, suelen usarse las estadísticas de segundo orden porque resultan de aplicar el criterio de optimización de Mínimos Cuadráticos, que conduce a la estimación de Máxima Verosimilitud de los parámetros de Procesos Gaussianos y, además, conduce a sistemas de ecuaciones lineales donde aparece la función autocorrelación. Cuando el proceso no sea Gaussiano el criterio de Mínimos Cuadráticos ya no conduce a la solución de Máxima Verosimilitud. Si consideramos un proceso **x(k)** real, estacionaria y aleatoria de media nula, su función autocorrelación $\mathbf{c_2(k_1)}$ da una medida de lo correlada que está esta secuencia respecto a desplazamientos de ella misma:

$$c_2(k_1) = E\{ x(k) \cdot x(k+k_1) \} \qquad \text{(III.1)}$$

Nótese que esta función autocorrelación presenta simetría par respecto al origen:

$$c_2(-k_1) = c_2(k_1) \qquad \text{(III.2)}$$

y, en consecuencia, su Transformada de Fourier también es una función real y par en el dominio frecuencial, es decir, resulta una función de fase nula y la información contenida en la fase del proceso **x(k)** se ha perdido al considerar $\mathbf{c_2(k_1)}$. Así, la información contenida en la fase desaparece en el dominio de la autocorrelación, pudiéndose recuperar sólamente para el caso de procesos de fase mínima. De esta manera, el uso de espectros de orden superior se impone para la reconstrucción de procesos de fase no mínima, o para la identificación de sistemas de fase no mínima, puesto que los poliespectros contienen las informaciones correspondientes al módulo y a la fase no mínima. Así, por ejemplo, dos procesos distintos, uno de fase mínima y ótro espectralmente equivalente de fase no mínima, pueden presentar la misma función de autocorrelación y diferentes estadísticas de tercer orden, es decir, el mismo espectro de potencia y distintos biespectros. Nótese, que en el caso de procesos Gaussianos de fase no mínima, ningún método permite recuperar la información de fase.

La mayor parte de señales o procesos reales no presentan una distribución Gaussiana y, por este motivo, presentan unos espectros de orden superior no nulos. Interpretando el esquema de la Fig.III.1, un proceso no Gaussiano puede descomponerse entre sus funciones espectrales de orden superior y cada una de éstas puede contener distinta información acerca de este proceso. Esto resulta muy útil en aplicaciones de clasificación de procesos donde

distintas características de clasificación pueden extraerse a partir de los distintos dominios espectrales de orden superior.

Finalmente, el uso de los espectros de orden superior parece bastante lógico cuando se pretenda analizar alguna no linealidad de un sistema, operando con una entrada aleatoria. Durante los últimos años se han estudiado, de forma bastante extensa, las relaciones entre determinadas señales aleatorias estacionarias y su paso a través de determinados sistemas lineales. Sin embargo, la mayoría de estas propiedades se han establecido considerando criterios basados en el espectro de potencia o en la función de autocorrelación y, además, la mayoría de estas relaciones dejan de cumplirse cuando se trabaja con sistemas no lineales. Cada tipo de alinealidad se debe investigar como un caso especial y los Poliespectros pueden jugar un papel clave en vistas a detectar y caracterizar cada tipo de alinealidad de un sistema, a partir de la señal a la salida del sistema analizado. Bajo las premisas anteriores se han desarrollado distintos métodos para la detección y caracterización de alinealidades en series temporales mediante el uso de espectros de orden superior [Niki-93a].

III.2. Definiciones Temporales y Frecuenciales.

En este apartado se presentan las principales definiciones correspondientes a las estadísticas de orden superior y sus correspondientes espectros de orden superior. En principio no se distingue entre señales deterministas y estocásticas porque todas las definiciones que se presentan tienen validez para ambos tipos de señales, bajo la suposición de estacionariedad. Aunque en este trabajo las señales deteministas no son de interés, en [Niki-93b] pueden encontrarse las expresiones finales de los espectros para los casos de señales deterministas de energía finita y de potencia media finita periódicas. En el primer subapartado se discuten estas definiciones desde el dominio temporal, dejando la discusión correspondiente al dominio frecuencial para el segundo subapartado. En ambos casos se parte de la definición correspondiente a un orden **n** genérico para, seguidamente, situarse en los casos de segundo, tercer y cuarto orden, donde se sedimenta el trabajo realizado en el presente estudio.

III.2.1. Momentos y Cumulantes.

Sea **x(k)** un proceso discreto, real y estacionario cuyos momentos existen hasta un orden **n**, entonces se define su momento de orden **n** como:

$$m_n(k_1, k_2, \dots, k_{n-1}) \int E\{ x(k) \cdot x(k+k_1) \cdot \dots \cdot x(k+k_{n-1}) \} \qquad \text{(III.3)}$$

donde

$$k, k_1, k_2, \dots, k_{n-1} = 0, \pm 1, \pm 2, \pm 3, \dots$$

y **E{.}** representa el operador Esperanza estadística. Puede observarse que este momento de orden **n** depende, sólamente, de los desplazamientos temporales $\mathbf{k_1, k_2, \dots, k_{n-1}}$ debido a la estacionariedad del proceso **x(k)**. Cuando el proceso considerado no sea estacionario, también depende de la posición temporal $\mathbf{m_n(k, k_1, k_2, \dots, k_{n-1})}$. Se aprecia, también, que el momento de segundo orden $\mathbf{m_2(k_1)}$ se corresponde claramente con la función autocorrelación clásica, mientras que los momentos de tercer y cuarto orden vienen representados, respectivamente, por $\mathbf{m_3(k_1, k_2)}$ y $\mathbf{m_4(k_1, k_2, k_3)}$.

Para el caso de un proceso **x(k)** aleatorio, estacionario y no Gaussiano los cumulantes de tercer y cuarto orden pueden expresarse en función de los momentos como sigue (n=3,4):

$$c_n(k_1, k_2, \dots, k_{n-1}) = m_n(k_1, k_2, \dots, k_{n-1}) - m^G_n(k_1, k_2, \dots, k_{n-1}) \qquad \text{(III.4)}$$

donde

$$m^G_n(k_1, k_2, \dots, k_{n-1}) \equiv E\{ g(k) \cdot g(k+k_1) \cdot \dots \cdot g(k+k_{n-1}) \} \qquad \text{(III.5)}$$

representa el momento de orden **n** de un proceso Gaussiano **g(k)** equivalente a **x(k)**, de tal manera, que ambas presenten la misma media y la misma secuencia de autocorrelación, es decir, con idénticos momentos de primer y segundo orden. Evidentemente, esta definición concuerda con una propiedad de los cumulantes, mencionada en el apartado anterior, respecto a la nulidad de los cumulantes de tercer y cuarto orden, $\mathbf{c_n(k_1, k_2, \dots, k_{n-1}) = 0}$, cuando el proceso **x(n)** presenta una distribución Gaussiana, puesto que se verifica:

$$m_n(k_1, k_2, \dots, k_{n-1}) = m^G_n(k_1, k_2, \dots, k_{n-1}) \qquad \text{(III.6)}$$

Nótese, que la propiedad anterior se verifica para cualquier cumulante cuyo orden sea superior a dos, aunque la expresión (III.4) solo es válida para el caso de tercer y cuarto orden. Obsérvese, también, que los cumulantes de orden **n** están dando una medida sobre cuánto se desvía un proceso cualquiera respecto de la Gaussianidad.

A continuación se presentan las relaciones básicas entre cumulantes y momentos, resultantes de combinar las expresiones (III.3) y (III.4), particularizando para los cuatro primeros órdenes y que se utilizan en capítulos posteriores del presente trabajo de investigación:

1) cumulantes de primer orden:

$$c_1 = m_1 = E\{ x(k) \} = \text{media} \qquad \text{(III.7)}$$

2) cumulantes de segundo orden:

$$c_2(k_1) = m_2(k_1) - [m_1]^2 = m_2(-k_1) - [m_1]^2 = c_2(-k_1) \qquad \text{(III.8)}$$

donde $\mathbf{m_2(k_1)}$ se corresponde con la secuencia de autocorrelación y, asimismo, los cumulantes de segundo orden $\mathbf{c_2(k_1)}$ con la secuencia de covarianza.

3) cumulantes de tercer orden:

$$c_3(k_1, k_2) = m_3(k_1, k_2) - m_1[m_2(k_1) + m_2(k_2) - m_2(k_1 - k_2)] + 2[m_1]^3 \quad \text{(III.9)}$$

4) cumulantes de cuarto orden:

$$\begin{aligned} c_4(k_1, k_2, k_3) &= m_4(k_1, k_2, k_3) - m_2(k_1) \cdot m_2(k_3-k_2) \\ &- m_2(k_2) \cdot m_2(k_3-k_1) - m_2(k_3) \cdot m_2(k_2-k_1) - 6 \cdot [m_1]^4 \\ &- m_1[m_3(k_2-k_1, k_3-k_1) + m_3(k_2, k_3) + m_3(k_1, k_3) + m_3(k_1, k_2)] \\ &+ [m_1]^2 \cdot [m_2(k_1) + m_2(k_2) + m_2(k_3) + m_2(k_3-k_1) + m_2(k_3-k_2) + m_2(k_2-k_1)] \end{aligned} \quad \text{(III.10)}$$

Si el proceso **x(k)** presenta media nula, $\mathbf{m_1=0}$, entonces las expresiones anteriores se simplifican y los cumulantes de segundo y tercer orden se corresponden idénticamente con los momentos de segundo y tercer orden respectivamente:

$$c_2(k_1) = m_2(k_1) \quad \text{(III.11)}$$

$$c_3(k_1, k_2) = m_3(k_1, k_2) \quad \text{(III.12)}$$

sin embargo, para generar los cumulantes de cuarto orden se precisa del conocimiento de los momentos de cuarto y segundo orden:

$$\begin{aligned} c_4(k_1, k_2, k_3) &= m_4(k_1, k_2, k_3) - m_2(k_1) \cdot m_2(k_3-k_2) \\ &- m_2(k_2) \cdot m_2(k_3-k_1) - m_2(k_3) \cdot m_2(k_2-k_1) \end{aligned} \quad \text{(III.13)}$$

Cuando nos situamos en el origen, es decir, considerando desplazamientos temporales nulos, $\mathbf{k_1=k_2=k_3=0}$, se presentan los conceptos de varianza, skewness y kurtosis correspondientes, respectivamente, a los dominios de segundo, tercer y cuarto orden:

varianza : $$\gamma_2 = E\{ x^2(k) \} = c_2(0) \quad \text{(III.14)}$$

skewness : $$\gamma_3 = E\{ x^3(k) \} = c_3(0, 0) \quad \text{(III.15)}$$

kurtosis : $$\gamma_4 = E\{ x^4(k) \} - 3 [\gamma_2]^2 = c_4 (0, 0, 0) \qquad \text{(III.16)}$$

y para los casos de las estadísticas de tercer y cuarto orden también se utilizan los conceptos de skewness y kurtosis normalizadas definidos, respectivamente, como $\gamma_3/[\gamma_2]^{3/2}$ y $\gamma_4/[\gamma_2]^2$. Nótese, que en el dominio de los cumulantes de orden superior, un valor nulo en el origen no implica que se anulen los cumulantes para cualquier punto del plano multidimensional. Así, por ejemplo, un valor nulo de la skewness no implica que los cumulantes de tercer orden sean idénticamente cero.

Aunque los cumulantes de cuarto orden implican un incremento considerable de la complejidad de cálculo, resultan especialmente necesarios cuando los cumulantes de tercer orden se anulan para el caso de procesos distribuidos simétricamente, tales como los procesos uniformes, procesos de Laplace, procesos Gaussianos y los procesos de Bernoulli-Gaussianos. Los cumulantes de tercer orden no se anulan para los procesos cuya función densidad de probabilidad no sea simétrica, como por ejemplo los procesos exponenciales o los de Rayleigh, pero pueden tomar valores extremadamente pequeños en comparación a los valores que presentan sus cumulantes de cuarto orden y, entonces, también parece lógico usar éstos últimos.

III.2.1.1. Propiedades de los Cumulantes.

Los cumulantes pueden usarse como un operador, de la misma forma que tratamos el operador Esperanza Estadística. Las principales propiedades de los cumulantes, que sostienen esta afirmación, son las siguientes (su demostración puede hallarse en [Mend-91]):

1) Los cumulantes de señales o procesos escalados, no siendo estos factores de escala aleatorios, se corresponde con el producto de todos estos factores por los cumulantes del proceso sin escalar:

$$\text{cum}\left\{\lambda_0 x(k), \ldots, \lambda_{n-1} x(k-k_{n-1})\right\} = \left(\prod_{i=0}^{n-1} \lambda_i\right) \cdot \text{cum}\left\{x(k), \ldots, x(k-k_{n-1})\right\} \qquad \text{(III.17.a)}$$

donde λ_i son constantes y

$$\text{cum}\left\{x(k), \ldots, x(k-k_{n-1})\right\} = c_n(k_1, \ldots, k_{n-1}) \qquad \text{(III.17.b)}$$

2) Los cumulantes son simétricos respecto la posición de sus argumentos ($\mathbf{k_0=0}$):

$$\text{cum}\left\{x(k-k_0), \ldots, x(k-k_{n-1})\right\} = \text{cum}\left\{x(k-k_{i_0}), \ldots, x(k-k_{i_{n-1}})\right\} \qquad \text{(III.18)}$$

donde $(\mathbf{i_0, i_1, \ldots, i_{n-1}})$ es una permutación de $\mathbf{(0, 1, \ldots, n-1)}$. Esto significa que se pueden intercambiar los argumentos de los cumulantes sin modificar su valor. De este modo los cumulantes de cuarto orden verifican:

$$c_4(k_1, k_2, k_3) = c_4(k_3, k_1, k_2) = c_4(k_2, k_3, k_1) \qquad \text{(III.19)}$$

3) Los cumulantes son aditivos respecto a sus argumentos, es decir, los cumulantes de una suma de argumentos se corresponde con la suma de cumulantes:

$$\text{cum}\left\{x(k)+y(k), x(k-k_1), \ldots, x(k-k_{n-1})\right\} = \qquad \text{(III.20)}$$
$$\text{cum}\left\{x(k), x(k-k_1), \ldots, x(k-k_{n-1})\right\} + \text{cum}\left\{y(k), x(k-k_1), \ldots, x(k-k_{n-1})\right\}$$

De ahí viene el nombre "cumulant" en inglés.

4) Los cumulantes son transparentes respecto la adición de constantes. Siendo ∂ una constante, entonces se verifica:

$$\text{cum}\left\{\partial + x(k), x(k-k_1), \ldots, x(k-k_{n-1})\right\} = \text{cum}\left\{x(k), x(k-k_1), \ldots, x(k-k_{n-1})\right\} \qquad \text{(III.21)}$$

5) Si dos procesos $\mathbf{x(k)}$ e $\mathbf{y(k)}$ son independientes, los cumulantes del proceso suma toma el valor de la suma de los cumulantes de cada proceso por separado:

$$\text{cum}\left\{x(k)+y(k),\ ..\ ,\ x(k-k_{n-1})+y(k-k_{n-1})\right\} = \text{cum}\left\{x(k),\ ..\ ,\ x(k-k_{n-1})\right\} + \text{cum}\left\{y(k),\ ..\ ,\ y(k-k_{n-1})\right\} \quad \text{(III.22)}$$

Nótese que si los procesos **x(k),y(k)** no fueran independientes, según la propiedad **3)**, aparecerían **2n** términos en el lado derecho de esta última expresión.

6) Si un subconjunto de **r** argumentos (**r≤n**) son independientes del resto entonces se verifica:

$$\text{cum}\left\{x(k),\ y(k-k_1),\ ..\ ,\ z(k-k_{n-1})\right\} = 0 \quad \text{(III.23)}$$

7) Los cumulantes de orden **n** presentan **n!** regiones de simetría [Gian-90a]:

$$c_n(k_1, k_2, \ldots, k_{n-1}) = c_n(k_2, k_1, \ldots, k_{n-1}) = \ldots = c_n(k_{n-1}, k_{n-2}, \ldots, k_1) =$$
$$c_n(-k_1, k_2-k_1, \ldots, k_{n-1}-k_1) = \ldots = c_n(k_{n-1}-k_1, k_{n-2}-k_1, \ldots, -k_1) = \quad \text{(III.24)}$$
$$\vdots$$
$$c_n(k_1-k_{n-1}, k_2-k_{n-1}, \ldots, -k_{n-1}) = \ldots = c_n(-k_{n-1}, k_{n-2}-k_{n-1}, \ldots, k_1-k_{n-1})$$

Pero, contrariamente a lo que sucede en el dominio de la función autocorrelación, no se verifica generalmente la simetría par para **n≥3** :

$$c_n(k_1, k_2, \ldots, k_{n-1}) \neq c_n(-k_1, -k_2, \ldots, -k_{n-1}) \quad \text{(III.25)}$$

Nótese que para el caso de segundo orden (III.24) se reduce a $\mathbf{c_2(k_1)=c_2(-k_1)}$ y se cumple la simetría par.

Sea **v(k)** un proceso Gaussiano independiente de **x(k)** (blanco o coloreado), que ha degradado el proceso **x(k)** originando el proceso **y(k)=x(k)+v(k)**, entonces para **n≥3** se verifica:

$$c_n^y (k_1, k_2, \dots, k_{n-1}) = c_n (k_1, k_2, \dots, k_{n-1}) \qquad \text{(III.26)}$$

mientras que en el caso de segundo orden clásico se cumple:

$$c_n^y (k_1) = c_n (k_1) + c_n^v (k_1) \qquad \text{(III.27)}$$

Esta última característica muestra la mayor robustez de las estadísticas de orden superior frente a la función autocorrelación clásica, incluso cuando este ruido sea coloreado. En consecuencia, los cumulantes pueden obtener información de procesos no Gaussianos sin afectarles la presencia de ruidos Gaussianos y, por ello, están viendo unas relaciones señal a ruido efectivas superiores.

III.2.2. Espectros de Orden Superior.

Los espectros de orden superior se obtienen al aplicar la Transformada de Fourier multidimensional $\mathbf{F_n[.]}$ sobre las estadísticas de orden superior. Para un orden **n** genérico se define el espectro de momentos como:

$$M_n (w_1, w_2, \dots, w_{n-1}) = F_n [\, m_n (k_1, k_2, \dots, k_{n-1}) \,] \qquad \text{(III.28)}$$

y, análogamente, se define el espectro de cumulantes:

$$C_n (w_1, w_2, \dots, w_{n-1}) = F_n [\, c_n (k_1, k_2, \dots, k_{n-1}) \,] \qquad \text{(III.29)}$$

Nótese que el espectro de cumulantes de orden **n** es también periódico con periodo 2π:

$$C_n (w_1, w_2, \dots, w_{n-1}) = C_n (w_1 + 2\pi, w_2 + 2\pi, \dots, w_{n-1} + 2\pi) \qquad \text{(III.30)}$$

Esta representación (III.29) de los cumulantes en el dominio frecuencial se debe a Kolmogorov y, asimismo, el término "Poliespectro" se debe a Brillinger y el término "espectro de orden superior" fue acuñado por Brillinger y Akaike [Niki-93a]. Al trabajar con procesos estocásticos, como pueden ser la señal de voz o el ruido, el espectro de cumulantes presenta una serie de ventajas respecto al espectro de momentos:

a) Para procesos Gaussianos todos los cumulantes de orden superior a dos se anulan y, por esta razón, el espectro de cumulantes puede medir la no Gaussianidad de un proceso concreto.

b) Los cumulantes dan una medida bastante conveniente de la extensión de las relaciones estadísticas que las series temporales presentaban en el caso de segundo orden.

c) Para el caso de ruido blanco de media no nula, sólamente su función covariancia se corresponde con la función impulso y, por consiguiente, presenta un espectro plano. Sus cumulantes de orden superior presentan la forma de una función impulso multidimensional y los poliespectros de este ruido son multidimensionalmente planos.

d) Los cumulantes de dos procesos aleatorios estadísticamente independientes se corresponden con la suma de los cumulantes de cada proceso individual, mientras que esta propiedad no se cumple para el caso de momentos de orden superior, como se citó en la expresión (III.21). Esta propiedad permite trabajar con los cumulantes, de una manera sencilla, usándolos como un operador.

Llegados a este punto, en adelante, se consideran sólamente los espectros de cumulantes y, similarmente al desarrollo correspondiente al dominio temporal, se particulariza para los casos del Espectro de Potencia (o Densidad Espectral de Potencia), el Biespectro y el Triespectro:

1) Espectro de Potencia:

$$C_2(w_1) = \sum_{k_1=-\infty}^{+\infty} c_2(k_1) \cdot e^{-j(w_1 k_1)} \qquad \text{(III.31)}$$

donde $|w_1| \leq \pi$ y $\mathbf{c_2(k_1)}$ representa la secuencia de covariancia del proceso $\mathbf{x(k)}$. Esta expresión se conoce, también, como Teorema de Wiener-Khintchine. A partir de la expresión (III.2) se deduce con facilidad que $\mathbf{C_2(w_1)}$ es una función real, par y no negativa.

2) Biespectro:

$$C_3(w_1, w_2) = \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} c_3(k_1, k_2) \cdot e^{-j(w_1 k_1 + w_2 k_2)} \qquad \text{(III.32)}$$

$$|w_1| \leq \pi \ , \ |w_2| \leq \pi \ , \ |w_1 + w_2| \leq \pi$$

donde $c_3(k_1, k_2)$ representa la secuencia de cumulantes de tercer orden de $\mathbf{x(k)}$. Al combinar la expresión (III.9) con las propiedades propias de los momentos, se deducen unas relaciones de simetría muy importantes para estos cumulantes de tercer orden (III.26):

$$c_3(k_1, k_2) = c_3(k_2, k_1) = c_3(-k_2, k_1 - k_2) = c_3(k_2 - k_1, -k_1)$$
$$= c_3(k_1 - k_2, -k_2) = c_3(-k_1, k_2 - k_1) \qquad \text{(III.33)}$$

a partir de estas cinco ecuaciones aparece una división del plano $\mathbf{k_1k_2}$ en seis regiones donde se repite esta función y, en consecuencia, al conocer los cumulantes de tercer orden en cualquiera de estas seis regiones, representadas en la Fig.III.2.a, se puede reconstruir la secuencia completa correspondiente a los cumulantes de tercer orden. Nótese que cada una de estas regiones contiene a su frontera. Así, por ejemplo, el sector **I** es una región infinita caracterizada por $\mathbf{0 < k_2 \leq k_1}$ (sector de 45º perteneciente al primer cuadrante). Para procesos no estacionarios estas seis regiones de simetría desaparecen. A partir de estas relaciones (III.30) y de la definición del espectro de cumulantes de tercer orden, se obtienen las siguientes relaciones en el dominio frecuencial bidimensional:

$$C_3(w_1, w_2) = C_3(w_2, w_1) = C_3^*(-w_2, -w_1) = C_3(-w_1 - w_2, w_2)$$
$$= C_3(w_1, -w_1 - w_2) = C_3(-w_1 - w_2, w_1) = C_3(w_2, -w_1 - w_2) \qquad \text{(III.34)}$$

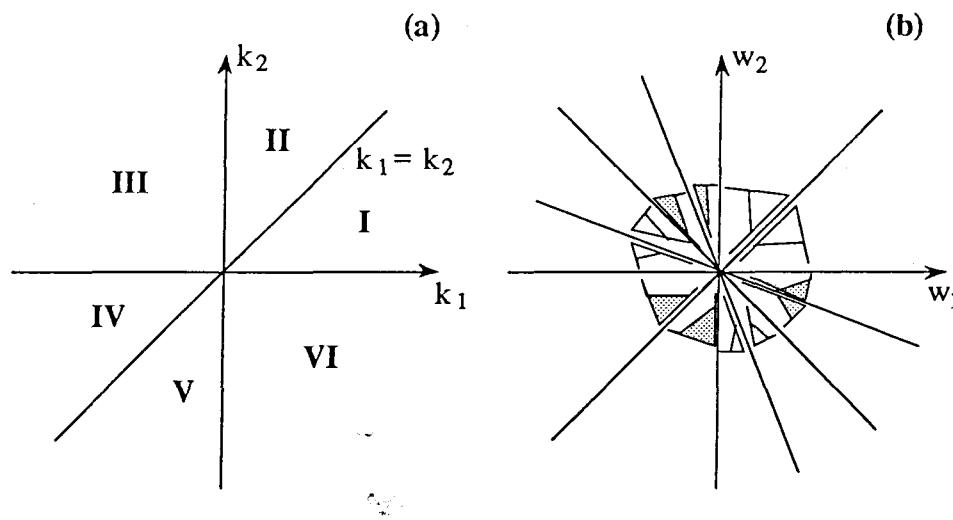


Figura III.2 : Regiones de simetría para: a) cumulantes de tercer orden; b) Biespectro

En la Fig.III.2.b se representan las 12 regiones de simetría del Biespectro cuando se consideran procesos estocásticos reales y, análogamente al dominio temporal, el conocimiento del Biespectro en la región triangular $w_2 \geq 0$, $w_1 \geq w_2$, $w_1 + w_2 \leq \pi$ es suficiente para una total reconstrucción del Biespectro. Nótese que, en el dominio frecuencial las regiones de simetría presentan un área finita y en ellas, en general, el Biespectro toma valores complejos y, consecuentemente, no se destruye la información de fase. El Biespectro es una función doblemente periódica con periodo 2π , $C_3(w_1, w_2) = C_3(w_1 + 2\pi, w_2 + 2\pi)$.

3) Triespectro:

$$C_4(w_1, w_2, w_3) = \sum_{k_1=-\infty}^{+\infty} \sum_{k_2=-\infty}^{+\infty} \sum_{k_3=-\infty}^{+\infty} c_4(k_1, k_2, k_3) \cdot e^{-j(w_1 k_1 + w_2 k_2 + w_3 k_3)} \qquad \text{(III.35)}$$

$$|w_1| \leq \pi \,, \quad |w_2| \leq \pi \,, \quad |w_3| \leq \pi \,, \quad |w_1 + w_2 + w_3| \leq \pi$$

donde $c_4(k_1, k_2, k_3)$ representa la secuencia de cumulantes de cuarto orden, definida en (III.10). Al combinar la definición del Triespectro y la de los cumulantes de cuarto orden se pueden deducir 96 regiones de simetría [Pflu-92], cuando se evalúan procesos reales.

A partir de los espectros de cumulantes de orden superior, en el dominio frecuencial, se pueden recuperar las expresiones de sus respectivos cumulantes, en el dominio temporal, aplicando la Transformada de Fourier Inversa de orden **n**:

$$c_n(k_1, .., k_{n-1}) = \frac{1}{(2\pi)^{n-1}} \int_{-\pi}^{+\pi} .. \int_{-\pi}^{+\pi} C_n(k_1, .., k_{n-1}) . e^{j(w_1 k_1 + .. + w_{n-1} k_{n-1})} \, dw_1 .. \, dw_{n-1} \qquad \text{(III.36)}$$

Asimismo, los valores de variancia, skewness y kurtosis pueden hallarse en el dominio frecuencial mediante las siguientes expresiones:

$$\gamma_2 = \frac{1}{2\pi} \int_{-\pi}^{+\pi} C_2(w_1) \, dw_1 \qquad \text{(III.37)}$$

$$\gamma_3 = \frac{1}{(2\pi)^2} \int_{-\pi}^{+\pi} \int_{-\pi}^{+\pi} C_3(w_1, w_2) \, dw_1 \, dw_2 \qquad \text{(III.38)}$$

$$\gamma_4 = \frac{1}{(2\pi)^3} \int_{-\pi}^{+\pi} \int_{-\pi}^{+\pi} \int_{-\pi}^{+\pi} C_4(w_1, w_2, w_3) \, dw_1 \, dw_2 \, dw_3 \qquad \text{(III.39)}$$

III.2.2.1. Procesos no Gaussianos Lineales.

En el dominio frecuencial se utilizan, en algunas aplicaciones, los espectros de orden superior **n** normalizados por el Espectro de Potencia (**n=2**), dando lugar a las denominadas funciones de Coherencia de orden **n**. Así, la función de Bicoherencia, o Coherencia de tercer orden, se define como:

$$P_3(w_1, w_2) = \frac{C_3(w_1, w_2)}{\sqrt{C_2(w_1) \cdot C_2(w_2) \cdot C_2(w_1 + w_2)}} \qquad \text{(III.40)}$$

y, de forma similar, la función de Tricoherencia, o Coherencia de cuarto orden, viene dada por:

$$P_4(w_1, w_2, w_3) = \frac{C_4(w_1, w_2, w_3)}{\sqrt{C_2(w_1) \cdot C_2(w_2) \cdot C_2(w_3) \cdot C_2(w_1 + w_2 + w_3)}} \qquad \text{(III.41)}$$

Estas funciones de coherencia se aplican, principalmente, para la detección y caracterización de alinealidades de ciertas series temporales y para distinguir entre procesos lineales y no lineales. De esta manera, se dice que un proceso no Gaussiano es lineal de orden **n** cuando el módulo de su función Coherencia de orden **n**, $|P_n(w_1, \ldots, w_{n-1})|$, sea constante para cualquier frecuencia, considerándose no lineales los restantes procesos [Ragh-85].

Un proceso no Gaussiano de media nula, $E\{v(k)\}=0$, se denomina Blanco de orden **n** si su secuencia de cumulantes de orden **n** presenta la expresión siguiente:

$$c_n^v(k_1, \ldots, k_{n-1}) = \gamma_n^v \cdot \delta(k_1, \ldots, k_{n-1}) \qquad \text{(III.42)}$$

donde $\delta(k_1, \ldots, k_{n-1})$ es la función delta de Krönecker. Entonces, aplicando la Transformada de Fourier multidimensional de orden **n-1** se obtiene su Poliespectro de orden **n**:

$$C_n^v(w_1, \ldots, w_{n-1}) = \gamma_n^v \qquad \text{(III.43)}$$

que resulta ser totalmente plano para todo el dominio frecuencial y para cualquier orden **n** considerado. Algunos casos especiales de Poliespectros correspondientes a Procesos Blancos se refieren a su Espectro de Potencia, su Biespectro y su Triespectro cuyos valores constantes se corresponden, respectivamente, con la variancia, la skewness y la kurtosis:

$$C_2^v(w_1) = \gamma_2^v \qquad \text{(III.44)}$$

$$C_3^v(w_1, w_2) = \gamma_3^v \qquad \text{(III.45)}$$

$$C_4^v(w_1, w_2, w_3) = \gamma_4^v \qquad \text{(III.46)}$$

donde γ_2^v , γ_3^v y γ_4^v representan la variancia, la skewness y la kurtosis, respectivamente, de **v(k)**.

El Poliespectro de orden **n** de un Proceso no Gaussiano Lineal **x(k)** admite siempre la siguiente expresión:

$$C_n(w_1, \ldots, w_{n-1}) = \gamma_n^v \cdot H(w_1) \ldots H(w_{n-1}) \cdot H^*(w_1 + \ldots + w_{n-1}) \qquad \text{(III.47)}$$

donde γ_n^v es una constante característica de un proceso Ruido Blanco no Gaussiano **v(k)** que excita un Sistema **H(w)** Lineal e Invariante (LTI), que a su salida origina el Proceso **x(k)** no Gaussiano Lineal, tal como se muestra en la Fig.III.3 .

Para los casos particulares anteriormente citados, se obtienen las siguientes expresiones correspondientes a sus respectivos Poliespectros:

$$C_2(w_1) = \gamma_2^v \cdot |H(w_1)|^2 \qquad \text{(III.48)}$$

$$C_3(w_1, w_2) = \gamma_3^v \cdot H(w_1) \cdot H(w_2) \cdot H^*(w_1 + w_2) \qquad \text{(III.49)}$$

$$C_4(w_1, w_2, w_3) = \gamma_4^v \cdot H(w_1) \cdot H(w_2) \cdot H(w_3) \cdot H^*(w_1 + w_2 + w_3) \qquad \text{(III.50)}$$

Existe, también, la posibilidad de que un proceso no Gaussiano sea Lineal para un(os) cierto(s) orden(es) y no Lineal para los restantes, es decir, puede presentarse el caso que un Proceso no Gaussiano sea Lineal en el dominio de su Biespectro pero no Lineal los dominios correspondientes al Triespectro y Poliespectros de órdenes superiores. Nótese, que la expresión (III.48) es válida siempre, para Procesos Lineales y no Lineales, mientras que los Poliespectros sólamente verifican (III.47) cuando el proceso considerado sea lineal, o sea, cuando presente el módulo plano de la función Coherencia. Así, para el caso de Procesos

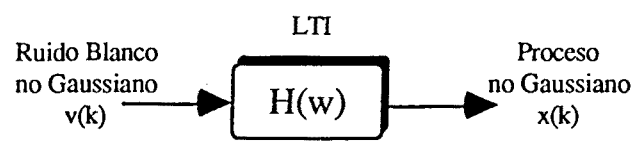


Figura III.3. : Generación de un Proceso no Gaussiano Lineal.

Lineales, se puede obtener el Espectro de Potencia a partir de su Biespectro o su Triespectro combinando las igualdades (III.48), (III.49) y (III.50), conocido **H(0)≠0** :

$$C_3(w_1,\ 0\ ) = \frac{\gamma_3^v}{\gamma_2^v} \cdot H(0) \cdot C_2(w_1) \qquad \text{(III.51)}$$

$$C_4(w_1,\ 0\ ,\ 0\ ) = \frac{\gamma_4^v}{\gamma_3^v} \cdot H(0) \cdot C_3(w_1,\ 0\ ) \qquad \text{(III.52)}$$

y, análogamente, se puede obtener el Biespectro a partir del conocimiento de su Triespectro:

$$C_4(w_1,\ w_2,\ 0\ ) = \frac{\gamma_4^v}{\gamma_3^v} \cdot H(0) \cdot C_3(w_1,\ w_2) \qquad \text{(III.53)}$$

III.3. Los Slices o Cumulantes unidimensionales.

La mayor complejidad de cálculo asociada a los cumulantes conduce, en ciertas aplicaciones, a considerar sus valores en una determinada dirección unidimensional, dentro del plano de dimensión **n-1**. Así, para el caso de cumulantes de orden **n** superior a dos, normalmente se evalúan éstos en una sola dimensión y se colocan los restantes **n-2** grados de libertad a un valor fijo, dando origen al concepto de slice unidimensional $r_{i,j}(k_s)$, donde k_s representa la variable unidimensional, **i** la situación de la primera dimensión no nula y **j** el número de dimensiones no nulas. Se puede hablar de distintos tipos de slices unidimensionales según la forma en que se evalúa cada slice. Así, se denomina diagonal al slice que verifica $k_1 = k_2 = \ldots = k_{n-1} = k$ y, análogamente, se puede considerar radial, vertical u horizontal. Evidéntemente, los slices resultan de gran utilidad en la mayor parte de aplicaciones de procesado de señal, puesto que evaluar simultáneamente varios grados de libertad de los cumulantes puede resultar bastante complejo.

Si nos situamos en el dominio de los cumulantes de tercer orden, se suelen definir los dos slices unidimensionales resultantes de trazar dos lineas de 90º (slice vertical) y 45º (slice diagonal):

$$r_{2,1}(k_s) \equiv \text{cum}\{ x(k) \cdot x(k) \cdot x(k+k_s) \} = c_3(0, k_s) \qquad \text{(III.54.a)}$$

$$r_{1,2}(k_s) \equiv \text{cum}\{ x(k) \cdot x(k+k_s) \cdot x(k+k_s) \} = c_3(k_s, k_s) \qquad \text{(III.54.b)}$$

$$r_{1,2}(k_s) \equiv r_{1,2}(-k_s) \qquad \text{(III.54.c)}$$

que conducen a la definición de dos funciones que presentan, respectivamente, simetría par e impar:

$$s_{2,1}(k_s) \equiv 0.5 \cdot [\, r_{2,1}(k_s) + r_{1,2}(k_s) \,] \qquad \text{(III.55.a)}$$

$$q_{2,1}(k_s) \equiv 0.5 \cdot [\, r_{2,1}(k_s) - r_{1,2}(k_s) \,] \qquad \text{(III.55.b)}$$

Si definimos el espectro unidimensional como:

$$R_{2,1}(w) \equiv \sum_{k_s=-\infty}^{+\infty} r_{2,1}(k_s) \cdot e^{-j(w k_s)} \qquad \text{(III.56)}$$

combinando las expresiones anteriores resulta:

$$R_{2,1}(w) = \sum_{k_s=-\infty}^{+\infty} \left\{ s_{2,1}(k_s) \cdot \cos(w k_s) - j\, q_{2,1}(k_s) \cdot \operatorname{sen}(w k_s) \right\} \quad \text{(III.57)}$$

Se observa que la contribución a la skewness procede sólamente de la parte real de este espectro, pues $s_{2,1}(0)=\gamma_3$ y $q_{2,1}(0)=0$, y se obtiene la siguiente expresión que relaciona este espectro unidimensional con el Biespectro:

$$R_{2,1}(w) = \frac{1}{2\pi} \cdot \int_{-\pi}^{+\pi} C_3(w, \sigma)\, d\sigma \quad \text{(III.58)}$$

que es la integración del Biespectro a lo largo de una de sus dimensiones.

III.4. Estimación de los Cumulantes y sus Poliespectros.

En una situación real de aplicación no se dispone de un conjunto infinito de valores de un proceso determinado sino, sólamente, de un conjunto finito de **N** valores o muestras del proceso. Además, algunos procesos solo presentan estacionariedad durante intérvalos relativamente cortos de tiempo, resultando necesaria la obtención de los cumulantes de una forma adaptativa a lo largo del tiempo, reduciéndose, aún más, el conjunto de valores disponibles del proceso **x(k)**.

En estas condiciones se impone, pues, la necesidad de una estimación de los cumulantes, o sus poliespectros asociados, a partir de un conjunto finito de observaciones de este determinado proceso. En este apartado, la discusión se centra en el caso de la estimación de tercer orden, pudiéndose deducir el caso de cuarto orden por simple extensión a partir del caso de tercer orden, usado de forma mayoritaria en el presente estudio. En principio, existen dos estrategias básicas para estimar estos poliespectros:

1) métodos convencionales o tipo Fourier,

2) métodos paramétricos fundamentados en modelos ARMA, AR o MA.

III.4.1. Métodos de Estimación Convencionales.

Los métodos convencionales para estimar los cumulantes de tercer orden, o el Biespectro, asociados a cada proceso particular pueden clasificarse en dos grupos principales:

1.a) técnicas indirectas que hacen uso de aproximaciones relacionadas directamente con las definiciones presentadas en (III.4) y (III.29),

1.b) técnicas directas que aproximan según una definición equivalente para los cumulantes o el Biespectro.

El comportamiento obtenido por ambas clases de técnicas es bastante similar y, por esta razón, aquí se presenta sólamente la primera de éllas. A continuación se detalla la

metodología correspondiente a las técnicas indirectas. Suponiendo que se dispone de un conjunto de **N** valores $\{x(1), x(2), \ldots, x(N)\}$ del proceso $x(k)$. Para estimar su Biespectro se deben considerar los siguientes pasos:

1) segmentar el conjunto de **N** valores en **K** registros con **M** valores cada uno, **N=K.M,**

2) sustraer la media m_1 de cada registro, si la hubiera,

3) tomando $\{x^i(k), k=0,1, \ldots, M-1\}$ como el conjunto de valores correspondiente al segmento i-ésimo, entonces, se obtiene la secuencia de cumulantes o momentos de tercer orden:

$$r^i(k_1, k_2) = \frac{1}{M} \cdot \sum_{k=s_1}^{s_2} x^i(k) \cdot x^i(k + k_1) \cdot x^i(k + k_2) \qquad \text{(III.59)}$$

donde

$i = 1, 2, \ldots, K$

$s_1 = \text{máx}\ (0, -k_1, -k_2)$

$s_2 = \text{mín}\ (M-1, M-1-k_1, M-1-k_2)$

4) promediar los valores obtenidos por cada uno de los **K** segmentos:

$$\hat{c}_3(k_1, k_2) = \frac{1}{K} \cdot \sum_{i=1}^{K} r^i(k_1, k_2) \qquad \text{(III.60)}$$

5) generar la estimación del Biespectro:

$$\hat{C}_3(w_1, w_2) = \sum_{k_1=-L}^{L} \sum_{k_2=-L}^{L} \hat{c}_3(k_1, k_2) \cdot w(k_1, k_2) \cdot e^{-j(w_1 k_1 + w_2 k_2)} \qquad \text{(III.61)}$$

donde $L < M-1$ y $w(k_1, k_2)$ representa una ventana temporal en el espacio bidimensional. La complejidad propia de esta aproximación puede simplificarse al aplicar las propiedades de simetría de los cumulantes (III.33) durante la obtención de $r^i(k_1, k_2)$ en (III.59) y las correspondientes simetrías propias del Biespectro (III.34) en (III.61).

Tal como sucede en la estimación convencional del Espectro de Potencia, se pueden considerar ventanas adecuadas para la obtención de mejores estimaciones. Estas ventanas bidimensionales han sido estudiadas por Sasaki, Sato y Yamashita [Niki-87c]. Se han estipulado algunas condiciones a satisfacer por éstas:

1) propiedades de simetría de los cumulantes de tercer orden (III.33):

$$w(k_1, k_2) = w(k_2, k_1) = w(k_1-k_2, -k_2) = w(-k_1, k_2-k_1) \qquad \text{(III.62.a)}$$

2) fuera de la región donde se define $\hat{c}_3(k_1, k_2)$ se debe verificar:

$$w(k_1, k_2) = 0 \qquad \text{(III.62.b)}$$

3) condición de normalización:

$$w(0, 0) = 1 \qquad \text{(III.62.c)}$$

4) para todo par frecuencial (w_1, w_2) debe verificarse:

$$W(w_1, w_2) \geq 0 \qquad \text{(III.62.d)}$$

Una familia de funciones que satisface las condiciones anteriores es la representada por la expresión:

$$w(k_1, k_2) = d(k_1) \cdot d(k_2) \cdot d(k_2-k_1) \qquad \text{(III.63)}$$

donde

$$d(k) = d(-k) \qquad \text{(III.64.a)}$$

$$d(k) = 0 \ , \quad \text{si} \quad k > L \qquad \text{(III.64.b)}$$

$$d(0) = 1 \qquad \text{(III.64.c)}$$

$$D(w) \geq 0 \ , \quad \text{para todo } w \qquad \text{(III.64.d)}$$

Estas ecuaciones permiten construir ventanas bidimensionales a partir de las ventanas unidimensionales standard. Sin embargo, no todas las ventanas convencionales, usadas para estimar el Espectro de Potencia, satisfacen la condición (III.64.d). Así, por ejemplo, la ventana de Hanning presenta lóbulos secundarios negativos en el dominio frecuencial. Algunos ejemplos de ventanas que cumplen las restricciones arriba mencionadas son:

1) ventana Optima (en el sentido se mínimo sesgo en la estimación del Biespectro)

$$d_o(k) = \begin{cases} \frac{1}{\pi} \cdot \left| \operatorname{sen}\frac{\pi k}{L} \right| + \left(1 - \frac{|k|}{L}\right) \cdot \cos\frac{\pi k}{L} & , \quad \text{si } |k| \le L \\ 0 & , \quad \text{si } |k| > L \end{cases} \qquad \text{(III.65)}$$

2) ventana de Parzen

$$d_p(k) = \begin{cases} 1 - 6.\left(\frac{|k|}{L}\right)^2 + 6.\left(\frac{|k|}{L}\right)^3 & , \quad \text{si } |k| \le \frac{L}{2} \\ 2.\left(1 - \frac{|k|}{L}\right)^3 & , \quad \text{si } \frac{L}{2} \le |k| \le L \\ 0 & , \quad \text{si } |k| > L \end{cases} \qquad \text{(III.66)}$$

3) ventana Uniforme en frecuencia

$$W_u(w_1, w_2) = \begin{cases} \frac{4\ \pi}{3\ \Omega o} & , \quad |w| < \Omega o = \frac{a_o}{L} \\ 0 & , \quad |w| > \Omega o = \frac{a_o}{L} \end{cases} \qquad \text{(III.59)}$$

donde $|w|=\text{máx}(|w_1|, |w_2|, |w_1+w_2|)$ y a_o representa un parámetro de valor constante. Esta ventana presenta una distribución uniforme a lo largo de la región hexagonal frecuencial representada en la Fig.III.2.b y no pertenece a la clase de funciones separables representada por (III.63), aunque si pertenece a la clase general originada por (III.62).

Algunos estudios [Niki-87c] han comparado estas tres ventanas en términos de la calidad obtenida en la estimación del Biespectro. En principio la ventana Uniforme origina un sesgo en la estimación cuatro veces superior al que se obtiene bajo consideración de la ventana Optima, para valores similares de sus varianzas. Al comparar las ventanas de Parzen y Optima se obtienen resultados similares: para un mismo valor de **L** la ventana Optima presenta un sesgo inferior en un 18% respecto a la ventana de Parzen, pero esta última presenta una variancia inferior en un 26%.

No obstante, estos estimadores convencionales se caracterizan generalmente por su alta variancia en las estimaciones que generan y, en consecuencia, se precisa promediar un número elevado de segmentos **K** para conseguir una estimación alisada. Pero, esto conlleva un incremento de la complejidad de cálculo, el sesgo de la estimación aumenta y, además, empeora la capacidad de discernir entre picos biespectrales próximos. Obviamente, estos métodos convencionales de estimación ofrecen unas prestaciones bastante pobres para procesos como la señal de voz, cuando la hipótesis de estacionariedad es válida únicamente durante intérvalos de duración inferior a los 25 ó 30 mseg. En resumen, estas metodologías convencionales presentan la ventaja de su sencillez y su facilidad de implementación, ya que posibilitan el uso de algoritmos eficientes de Transformada Rápida de Fourier (FFT). Pero sus limitaciones en la variancia estadística, el tiempo de cálculo requerido y las disponibilidades de espacio de memoria imponen serias limitaciones a su implementación práctica.

III.4.2. Estimadores Paramétricos.

Tal como se ha comentado anteriormente, los métodos convencionales aplicados a la estimación de un proceso no Gaussiano presentan las ventajas de facilidad de implementación y fidelidad de la estimación cuando se dispone de grandes registros de datos correspondientes al proceso a estimar [Ragh-85]. Sin embargo, su capacidad para discernir componentes armónicas en el dominio biespectral es limitada debido al "Principio de Incertidumbre" de la Transformada de Fourier. Este problema cobra especial importancia en aplicaciones como la detección de acoplamientos de fase cuadráticos cuando los pares frecuenciales estan próximos [Ragh-86]. Además, para el caso de procesos paramétricos estos métodos convencionales obtienen una fidelidad biespectral bastante pobre [Ragh-85], [Ragh-86]. Nótese que la voz puede modelarse con bastante fidelidad por un modelado paramétrico Autoregresivo (AR).

Una de las técnicas más populares y eficientes para la interpretación de series temporales consiste en la generación de modelos paramétricos lineales, impulsados por ruido blanco, a partir de las muestras disponibles del proceso físico en consideración. Sin embargo, en la literatura correspondiente a las series temporales dominan las técnicas fundamentadas en la suposición de Gaussianidad y/o relacionadas con el Espectro de Potencia (modelado segun la función de autocorrelación). Estas técnicas presentan dos limitaciones principales:

1) si el proceso no es Gaussiano, estas técnicas basadas en el Espectro de Potencia no pueden exprimir toda la información que se oculta en los datos disponibles del proceso,

2) los procesos Gaussianos de fase no mínima se identifican como procesos de fase mínima, cuando se usan estos modelos de segundo orden.

Por otra parte, existen tres razones básicas para buscar un modelado paramétrico impulsado por ruido blanco no Gaussiano:

1) recuperar fielmente la información contenida en la fase,

2) aumentar la capacidad de resolución del estimador para discernir entre picos cercanos en el dominio biespectral,

3) incrementar la fidelidad biespectral cuando los procesos no Gaussianos considerados sean paramétricos o puedan aproximarse fidedignamente por modelos paramétricos.

Se considera un proceso **x(k)** ARMA (Autoregresive Moving-Average), real, y de órdenes (p,q) descrito por:

$$\sum_{i=0}^{p} a_i \, . \, x(k-i) \;=\; \sum_{j=0}^{q} b_j \, . \, w(k-j) \qquad \text{(III.68)}$$

donde **w(k)** representa un proceso independiente e idénticamente distribuido de media nula, **E{w(k)}=0**, y con momentos de segundo y tercer orden dados por

$$E\{w(k),\, w(k+k_1)\} = Q.\, \delta(k_1) \qquad \text{(III.69.a)}$$

$$E\{w(k),\, w(k+k_1),\, w(k+k_2)\} = ß.\, \delta(k_1, k_2) \qquad \text{(III.69.b)}$$

y **x(k')** es independiente respecto **w(k)** para todo **k'<k** . Nótese que **w(k)** y **x(k)** no son procesos Gaussianos. Como el proceso **w(k)** es estacionario de tercer orden y se supone que el modelo ARMA es estable, entonces, el proceso generado **x(k)** también presenta estacionariedad de tercer orden. Su Espectro de Potencia viene dado por:

$$C_2(w_1) = P(w_1) = Q \, . \, |\, H(z)\, |^2 \qquad \text{para} \qquad z = e^{\,jw_1} \qquad \text{(III.70.a)}$$

donde

$$H(z) = \frac{N(z)}{D(z)} \qquad \text{(III.70.b)}$$

$$N(z) = \sum_{j=0}^{q} b_j \cdot z^{-j} \qquad \text{(III.70.c)}$$

$$D(z) = \sum_{i=0}^{p} a_i \cdot z^{-i} \quad , \quad a_0 = 1 \qquad \text{(III.70.d)}$$

y su Biespectro se puede representar como:

$$C_3(w_1, w_2) = B(w_1, w_2) = ß \cdot H(w_1) \cdot H(w_2) \cdot H^*(w_1+w_2) \qquad \text{(III.71.a)}$$

donde

$$H(w) = H(z)\Big|_{z=e^{jw}} = |H(w)| \cdot e^{j\,h(w)} \qquad \text{(III.71.b)}$$

Si **w(k)** es un proceso Gaussiano y, por lo tanto **x(k)** también, entonces cualquier raíz real z_r de **N(z)** o **D(z)** puede reemplazarse por su inversa $(1/z_r)$, y cualquier par de raíces complejas conjugadas z_c por su par conjugado inverso $(1/z_c)^*$, sin originar modificaciones ni en su Espectro de Potencia ni en su Función de Distribución de Probabilidad. Así, por ejemplo, bajo la suposición de tener todas las raices reales y distintas aparecen 2^{p+q} posibles combinaciones de raices que conducen a la misma secuencia de autocorrelación $C_2(w_1)$, es decir, existen 2^{p+q} posibles maneras de elegir la fase **h(w)** sin modificar su módulo **|H(w)|**.

Durante los últimos años se ha obtenido un amplio abanico de resultados referentes a la estimación de los coeficientes a_i y b_j segun el criterio de Mínimos Cuadráticos que, en el entorno Gaussiano, son asintóticamente equivalentes al criterio de Máxima Verosimilitud. En el supuesto no Gaussiano, estas técnicas aportan soluciones de Mínimos Cuadráticos pero no de Máxima Verosimilitud. En ambos casos, sin embargo, los coeficientes estimados corresponden a raices situadas en el interior del círculo unidad (modelos ARMA de fase mínima) [Niki-87c]. En resumen, para estimar correctamente la Función de Transferencia de un proceso ARMA(p, q) debe considerarse:

1) Si el proceso **w(k)** es Gaussiano y **H(w)** es de fase mínima, entonces, los métodos basados en la autocorrelación, usando el criterio de Mínimos Cuadráticos, identifican correctamente el módulo **|H(w)|** y su fase **h(w)** [Makh-75].

2) Si **w(k)** es Gaussiano y **H(w)** no es de fase mínima, entonces, ningún método puede recuperar la fase auténtica **h(w)** [Benv-80].

3) Si el proceso **w(k)** no es Gaussiano y **H(w)** no es de fase mínima, entonces, los métodos de segundo orden pueden obtener correctamente el módulo **|H(w)|** pero no su fase **h(w)** [Benv-80].

4) Para el supuesto anterior se pueden obtener el módulo y fase correctos a partir del conocimiento de la distribución no Gaussiana de **w(k)**. Para ello se aplica el criterio de Máxima Verosimilitud durante la estimación de los coeficientes a_i y b_j, aunque conduce a la resolución de sistemas no lineales de ecuaciones [Benv-80].

5) Considerando este último supuesto, se pueden obtener el modulo y la fase verdaderos sin la necesidad del conocimiento de la distribución de **w(k)**. En este caso los parámetros a_i y b_j, o directamente su módulo y su fase, se deben estimar en el dominio espectral de orden superior [Niki-93a], [Gian-90a].

Los métodos espectrales de orden superior son más atractivos porque no requieren del conocimiento de la distribución no Gaussiana de **w(k)**, habitualmente desconocida en las aplicaciones reales o bien su estimación presenta serias dificultades, como en el caso de la propagación multicamino.

A continuación se discute el problema práctico de disponer de un conjunto finito de datos **{x(1), ... , x(N)}**, suponiendo que el proceso disponible **x(k)** se corresponde con un proceso AR de orden **p**. Entonces, este proceso se puede representar a partir de (III.68) fijando **q=0** y $a_0=b_0=1$.

$$x(k) + \sum_{i=1}^{p} a_i \cdot x(k-i) = w(k) \qquad \text{(III.72)}$$

resultando una secuencia de momentos de tercer orden

$$c_3(k_1, k_2) = E\{x(k), x(k+k_1), x(k+k_2)\} \qquad \text{(III.73)}$$

que satisface la siguiente recursión de tercer orden [Ragh-85]

$$c_3(-k_1, -k_2) + \sum_{i=1}^{p} a_i \cdot c_3(i-k_1, i-k_2) = \beta \cdot \delta(k_1, k_2) \quad , \quad k_1, k_2 \geq 0 \qquad \text{(III.74)}$$

donde $\delta(k_1, k_2)$ es la función impulso unidad bidimensional. A partir de (III.74), tomando la recta $k_1=k_2$ (slice diagonal) aparecen **2p+1** valores de los cumulantes de tercer orden que satisfacen la ecuación [Ragh-85], [Ragh-86]:

$$\underline{\underline{R}}_c \cdot \underline{a} = \underline{\beta} \tag{III.75.a}$$

donde

$$\underline{\underline{R}}_c = \begin{bmatrix} c_3(0,0) & c_3(1,1) & \cdots & c_3(p,p) \\ c_3(-1,-1) & c_3(0,0) & \cdots & c_3(p-1,p-1) \\ \vdots & \vdots & \ddots & \vdots \\ c_3(-p,-p) & c_3(-p+1,-p+1) & \cdots & c_3(0,0) \end{bmatrix} \tag{III.75.b}$$

$$\underline{a} = [1, a_1, a_2, \ldots, a_p]^T \tag{III.75.c}$$

$$\underline{\beta} = [\beta, 0, 0, \ldots, 0]^T \tag{III.75.d}$$

Esta matriz $\underline{\underline{R}}_c$ es Toeplitz pero, en general, no es simétrica. Además si se escribe el vector de parámetros de la forma $\underline{a}=[a_p, a_{p-1}, \ldots, a_1, 1]^T$, entonces $\underline{\underline{R}}_c$ es una matriz de Hankel. Otra posible representación de (III.74) consiste en permitir valores de (k_1, k_2) pertenecientes al primer sector de la Fig.III.2.a, siendo ahora una región triangular porque se dispone de un conjunto finito de datos pertenecientes al proceso $x(k)$:

$$C_3(-k_1, -k_2) + \sum_{i=1}^{p} a_i \cdot C_3(i-k_1, i-k_2) = \beta \cdot \delta(k_1, k_2) \tag{III.76}$$

$$k_1 = 0, 1, \ldots, L_1$$

$$k_2 = \begin{cases} 0, 1, \ldots, k_1 & \text{si } k_1 < L_1 \\ 0, 1, \ldots, L_2 & \text{si } k_1 = L_1 \end{cases}$$

donde (k_1, k_2) se eligen de manera tal que se verifique:

$$L_2 \leq L_1$$

$$p = 1 + L_2 + \frac{(L_1-1).(L_1+2)}{2} \tag{III.77}$$

La matriz correspondiente a las ecuaciones anteriores no se corresponde con una matriz de Toeplitz, pero representa un caso más general que (III.75) porque obtiene la información a partir de un área finita de los cumulantes y no a lo largo de una linea recta $k_1=k_2$. La expresión representada en (III.75) obtiene el modelado AR a partir de $2p+1$ valores de la secuencia de cumulantes o momentos de tercer orden, pertenecientes a la recta $k_1=k_2$ de la Fig.III.2.a : uno correspondiente al origen, $k_1=k_2=0$, y p puntos a cada lado del origen. Este modelo se ajusta al proceso en el sentido de ajuste perfecto entre la secuencia de momentos de tercer orden de la salida del filtro generador AR y las muestras disponibles del proceso en

estos instantes correspondientes, es decir, si se encuentra un modelo AR de orden **p**, impulsado por ruido blanco no Gaussiano, cuya secuencia de momentos de tercer orden a su salida se iguale a las muestras dadas del proceso para los puntos $k_1=k_2=0, \pm1, .. ,\pm p$, entonces, sus parámetros **a** verificarían (III.75) como condición necesaria. Si se considera la expresión (III.76) se precisan más de **2p+1** valores para ajustar un modelo AR segun el criterio anterior. Si las muestras **x(k)** disponibles han sido generadas a partir de valores reales de la secuencia de momentos de tercer orden correspondientes a un proceso de orden **p** que satisface todas las condiciones anteriormente especificadas, entonces, los parámetros **a** obtenidos en ambos casos son los mismos. En cualquier otra situación las dos soluciones anteriores pueden ser distintas.

El problema de ajustar un Biespectro dado con el de un proceso AR puede interpretarse, también, desde el punto de vista de la Teoría de Predicción. Si se considera un predictor lineal que obtiene una aproximación del valor actual **x(k)** a partir de **p** valores anteriores:

$$\hat{x}(k) = - \sum_{i=1}^{p} a_i \, . \, x(k-i) \qquad \text{(III.78)}$$

entonces el error de predicción viene dado por:

$$e(k) = x(k) - \hat{x}(k) = x(k) + \sum_{i=1}^{p} a_i \, . \, x(k-i) \qquad \text{(III.79)}$$

cuyo Biespectro $E(w_1, w_2)$ puede relacionarse con el Biespectro $C_2(w_1, w_2)$ del proceso **x(k)** mediante la expresión siguiente:

$$E(w_1, w_2) = A(w_1) \, . \, A(w_2) \, . \, A^*(w_1+w_2) \, . \, C_2(w_1, w_2) \qquad \text{(III.80.a)}$$

donde

$$A(w) = \left[\; 1 + \sum_{i=1}^{p} a_i \, . \, e^{-j\,w.i} \; \right] \qquad \text{(III.80.b)}$$

Si los coeficientes del predictor lineal (III.78) son tales que se cumple:

$$E(w_1, w_2) = ß \qquad \text{(III.81)}$$

siendo **ß** una constante real, entonces se concluye que el Biespectro del proceso **x(k)** viene dado, exactamente, por el proceso AR de parámetros $\{a_i, i=1, 2, .. , p\}$, impulsado por ruido blanco no Gaussiano cuya skewness tome valor **ß**:

$$C_2(w_1,w_2) = \frac{ß}{A(w_1) \cdot A(w_2) \cdot A^*(w_1+w_2)} \qquad \text{(III.82)}$$

De esta manera se puede afirmar que el parecido del Biespectro $\mathbf{E(w_1, w_2)}$ del error de predicción respecto a un Biespectro totalmente plano proporciona una medida del parecido de los Biespectros correspondientes al proceso **x(k)** disponible y a su modelo AR asociado [Ragh-86].

A continuación se presentan algunos métodos para estimar el Biespectro de un proceso **x(k)** mediante un modelado AR. Estos dos métodos se diferencian en la forma de estimar los cumulantes de tercer orden, a partir del conjunto finito de **N** muestras del proceso **x(k)**, para resolver el sistema de ecuaciones de tercer orden (III.74). En todos estos métodos se consideran unas condiciones comunes de trabajo: un proceso **v(k)**, uniformemente distribuido, independiente y no Gaussiano, excita un modelo **h(k)** paramétrico AR y causal, cuya fase puede no ser mínima, y a su salida se obtiene un proceso **x(k)** no Gaussiano, del cual se dispone un conjunto finito de **N** muestras. Normalmente el ruido **v(k)** y su distribución se suponen desconocidos.

III.4.2.1. Método recursivo de tercer orden (TOR) :

La matriz $\underline{\underline{R}}_c$ que aparece en la expresión (III.75) es Toeplitz pero, en general, no es simétrica. Si el filtro AR es estable, luego, esta ecuación (III.75) existe. Una condición suficiente, aunque no necesaria, para la estabilidad de **H(z)** impone que $\underline{\underline{R}}_c$ sea una matriz Toeplitz, simétrica y definida positiva [Makh-75]. De este modo se puede asegurar la estabilidad de las representaciones AR, sólamente, para aquellos procesos cuya matriz de cumulantes de tercer orden verifique las tres condiciones previamente citadas.

Sin embargo, en el caso que nos ocupa se dispone de un conjunto limitado de datos. Así, se consideran las **p+1** ecuaciones compuestas por los slices diagonales (III.75) y la aproximación de los cumulantes de tercer orden mediante $\hat{\mathbf{c}}_3(\mathbf{k_1}, \mathbf{k_2})$, hallados segun (III.59) y (III.60), se considera el sistema representado por:

$$\hat{\underline{\underline{R}}}_c \cdot \hat{\underline{a}} = \hat{\underline{\beta}} \qquad \text{(III.83.a)}$$

donde

$$\hat{\underline{\underline{R}}}_c = \begin{bmatrix} \hat{c}_3(0,0) & \hat{c}_3(1,1) & \cdots & \hat{c}_3(p,p) \\ \hat{c}_3(-1,-1) & \hat{c}_3(0,0) & \cdots & \hat{c}_3(p-1,p-1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{c}_3(-p,-p) & \hat{c}_3(-p+1,-p+1) & \cdots & \hat{c}_3(0,0) \end{bmatrix} \qquad \text{(III.83.b)}$$

$$\hat{\underline{a}} \int [1, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]^T \qquad \text{(III.83.c)}$$

$$\hat{\underline{\beta}} = [\hat{\beta}, 0, 0, \ldots, 0]^T \qquad \text{(III.83.d)}$$

siendo $\hat{\underline{a}}$ una estimación de los parámetros AR y la estimación del momento de tercer orden correspondiente al ruido blanco se representa por $\hat{\underline{\beta}}$. También se pueden aplicar las ecuaciones (III.76) para la estimación $\hat{\underline{a}}$ de los parámetros AR. Debe remarcarse que la aplicación de este método presupone la ergodicidad del proceso **x(k)**. Puede demostrarse que si el proceso **x(k)** cumple (III.68), entonces, este método (TOR) facilita estimaciones consistentes para los parámetros AR [Ragh-85].

III.4.2.2. Método de los momentos promedio de tercer orden (CTOR).

Si se forma la función de tercer orden

$$\hat{q}^k(k_1, i) = x(k-i) \cdot x^2(k-k_1) \quad , \quad i, k_1 = 1, .. , p \qquad \text{(III.84)}$$

aplicando el operador Esperanza se verifica:

$$E\left\{ \hat{q}^k(k_1, i) \right\} = c_3(i-k_1, i-k_1) \qquad \text{(III.85)}$$

Si en lugar de disponer de la muestras del proceso **x(k)** se dispusiera de muestras pertenecientes a su secuencia de cumulantes de tercer orden $\mathbf{c_3(k_1, k_2)}$, entonces, el proceso AR de orden **p** se ajustaría al proceso dado mediante la resolución de las ecuaciones (III.74):

$$E\left\{ \hat{q}^k(k_1, 0) + \sum_{i=1}^{p} \hat{a}_i \cdot \hat{q}^k(k_1, i) \right\} = 0 \quad , \quad k_1 = 1, .. , p \qquad \text{(III.86)}$$

Si la expresión interior al operador se nota como $\hat{e}_3(k, k_1)$:

$$E\left\{ \hat{e}_3(k, k_1) \right\} = 0 \quad , \quad \begin{matrix} k_1 = 1, 2, .. , p \\ k = p+1, p+2, .. , N \end{matrix} \qquad \text{(III.87)}$$

donde $\hat{e}_3(k, k_1)$ se refiere al proceso error de predicción de tercer orden. A partir de las **N** muestras disponibles del proceso **x(k)** se pueden obtener **N-p** valores de $\hat{e}_3(k, k_1)$ por cada valor de k_1. Para el caso que nos ocupa el operador Esperanza se traduce en un cálculo de la media de la secuencia error, y se llega a un sistema de **p** ecuaciones lineales, cuya resolución conduce a los parámetros $\underline{a}$ estimados:

$$\frac{1}{N-p} \cdot \sum_{k=p+1}^{N} \hat{e}_3(k, k_1) = 0 \quad , \quad k_1 = 1, .. , p \qquad \text{(III.88)}$$

que admite la siguiente representación matricial:

$$\underline{\underline{\hat{Q}}} \cdot \underline{\hat{a}} = \underline{\hat{b}} \qquad \text{(III.89.a)}$$

donde

$$\underline{\underline{\hat{Q}}} = \begin{bmatrix} \hat{q}_{11} & \cdots & \hat{q}_{1p} \\ \vdots & \ddots & \vdots \\ \hat{q}_{p1} & \cdots & \hat{q}_{pp} \end{bmatrix} \qquad \text{(III.89.b)}$$

$$\underline{\hat{a}} = [\hat{a}_1, \hat{a}_2, \ldots , \hat{a}_p]^T \qquad \text{(III.89.c)}$$

$$\underline{\hat{b}} = [\, \hat{q}_{10}, \ldots , \hat{q}_{p0}]^T \qquad \text{(III.89.d)}$$

$$\hat{q}_{ij} \equiv \sum_{k=p+1}^{N} \hat{q}^k(i,j) \qquad \text{(III.89.e)}$$

Los resultados obtenidos mediante este algoritmo son significativos si se verifica, como condición necesaria, la ergodicidad de tercer orden para el proceso **x(k)**. En [Ragh-86] se demuestra que estos métodos TOR y CTOM son asintóticamente equivalentes para un mismo

orden **p** prefijado. Las estimaciones de parámetros AR obtenidos mediante este método CTOM son consistentes mientras el proceso **x(k)** satisfaga la ecuación (III.72), elegida como suposición inicial.

III.4.2.3. Método AR Optimizado (OARM).

Este algoritmo se debe a An, Kim y Powers [An-88] y consiste en una extensión de la metodología TOR hacia un sistema sobredeterminado de ecuaciones. Se considera la ecuación recursiva de tercer orden (III.74) y se toman **p+1** posibles valores para:

$$k_1, k_2 = 0, 1, .., p \tag{III.90}$$

resultando la siguiente formulación matricial:

$$\underline{\underline{r}} \cdot \underline{a} = \underline{b} \tag{III.91.a}$$

donde

$$\underline{\underline{r}} = \begin{bmatrix} c_3(0,0) & c_3(1,1) & \cdots & c_3(p,p) \\ c_3(0,-1) & c_3(1,0) & \cdots & c_3(p,p-1) \\ \vdots & \vdots & \ddots & \vdots \\ c_3(0,-p) & c_3(1,1-p) & \cdots & c_3(p,0) \\ c_3(-1,0) & c_3(0,1) & \cdots & c_3(p-1,p) \\ \vdots & \vdots & \ddots & \vdots \\ c_3(-1,-p) & c_3(0,1-p) & \cdots & c_3(p-1,0) \\ \vdots & \vdots & \ddots & \vdots \\ c_3(-p,0) & c_3(1-p,1) & \cdots & c_3(0,p) \\ \vdots & \vdots & \ddots & \vdots \\ c_3(-p,-p) & c_3(1-p,1-p) & \cdots & c_3(0,0) \end{bmatrix} \tag{III.91.b}$$

$$\underline{a} = [1, a_1, a_2, \ldots, a_p]^T \tag{III.91.c}$$

$$\underline{b} = [\beta, 0, 0, \ldots, 0]^T \tag{III.91.d}$$

Nótese que $\underline{\underline{r}}$ es una matriz de $(p+1)^2$ filas y $p+1$ columnas que contiene todos los cumulantes de tercer orden. Al tratarse de un sistema sobredeterminado, la solución se obtiene aplicando el criterio de Mínimo Error Cuadrático Medio:

$$\underline{\hat{a}} = (\underline{\underline{r}}^T \underline{\underline{r}})^{-1} \underline{\underline{r}}^T . \underline{b} \qquad \text{(III.92)}$$

Mediante la resolución de la ecuación (III.92) se obtienen los parámetros AR. Este algoritmo origina una buena estimación especialmente para el caso de disponer de un conjunto pequeño de datos $\{x(1), x(2), .. , x(N)\}$ y/o para entornos altamente ruidosos. Bajo estas condiciones de trabajo, los dos métodos anteriores, TOR y CTOM, ofrecen pobres prestaciones segun se demuestra en [An-88].

III.4.2.4. Método de las ecuaciones de Yule-Walker de orden superior.

Giannakis demostró que esta metodología ofrece siempre una solución y ésta es única [Gian-90a]. En principio, el desarrollo analítico correspondiente a este algoritmo se deduce para cualquier orden **n** de las estadísticas de orden superior consideradas, no restringiéndose al caso más simple de tercer orden. Se consideran los cumulantes de orden $n \geq 3$ con **n-3** grados de libertad fijados a cero y, por simplicidad, se hace uso de la siguiente nomenclatura:

$$d_n(k_1, k_2) \int c_n(k_1, k_2, 0, \ldots, 0) \qquad \text{(III.93)}$$

Considerando (III.47), en el dominio temporal, se verifica la fórmula de Brillinger-Rosenblatt [Bril-67], demostrada también en [Gian-89a], particularizada para esta situación (III.93):

$$d_n(k_1, k_2) = \gamma_n^v . \sum_{k=0}^{\infty} h^{n-2}(k) . h(k+k_1) . h(k+k_2) \qquad \text{(III.94)}$$

Por otra parte, el filtro AR satisface la recursión:

$$h(k_1, k_2) = - \sum_{i=1}^{p} a_i . h(k+k_1-i) + b(k+k_1) \qquad \text{(III.95)}$$

y sustituyendo en (III.94):

$$d_n(k_1, k_2) + \sum_{i=1}^{p} a_i.d_n(k_1 - i, k_2) = \gamma_n^v . \sum_{k=0}^{\infty} h^{n-2}(k).b(k+k_1).h(k+k_2) \quad \text{(III.96)}$$

Como el modelo considerado es AR, entonces, el término $\mathbf{b(k+k_1)}$ se anula siempre para $\mathbf{k_1>0}$ y se obtienen las ecuaciones de Yule-Walker en el dominio de los cumulantes de orden superior:

$$\sum_{i=1}^{p} a_i.d_n(k_1 - i, k_2) = - d_n(k_1, k_2) \quad , \quad \begin{matrix} k_1 > 0 \\ k_2 \geq 0 \end{matrix} \quad \text{(III.97)}$$

Fijando $\mathbf{k_2}$ y tomando $\mathbf{k_1=1, .. , p}$ aparece un sistema lineal de **p** ecuaciones:

$$\underline{\underline{Q}}_{k_2} \cdot \underline{a} = \underline{b}_{k_2} \quad \text{(III.98.a)}$$

donde,

$$\underline{\underline{Q}}_{k_2} = \begin{bmatrix} d_n(0, k_2) & d_n(-1, k_2) & \cdots & d_n(1-p, k_2) \\ d_n(1, k_2) & d_n(0, k_2) & \cdots & d_n(2-p, k_2) \\ \vdots & \vdots & \ddots & \vdots \\ d_n(p-1, k_2) & d_n(p-2, k_2) & \cdots & d_n(0, k_2) \end{bmatrix} \quad \text{(III.98.b)}$$

$$\underline{a} = [a_1, a_2, \ldots, a_p]^T \quad \text{(III.98.c)}$$

$$\underline{b}_{k_2} = [d(1, k_2), d(2, k_2), \ldots, d(p, k_2)]^T \quad \text{(III.98.d)}$$

En [Gian-89a] se demuestra que esta matriz presenta rango completo **p** y origina una solución única cuando se toman **p+1** slices correspondientes a los valores $\mathbf{k_2=-p, 1-p, .. , -1, 0}$. En principio, parece que un slice de la secuencia de cumulantes de orden **n** podría ser suficiente puesto que origina **p** ecuaciones para las **p** incógnitas $\underline{a}$. En [Gian-89a] se hacían conjeturas acerca de la posibilidad de encontrar, para todos los casos, algun slice que fuera de rango completo **p**, pero en [Swam-89c] se descarta tal posibilidad. Así, se puede afirmar que ningún slice proporciona una matriz **pxp** de Hankel que sea de rango completo **p**. Por esta razón se debe considerar el siguiente sistema:

$$\underline{\underline{Q}} . \underline{a} = \underline{b} \quad \text{(III.99.a)}$$

donde,

$$\underline{\underline{Q}} = \begin{bmatrix} d_n(0,0) & d_n(-1,0) & \cdots & d_n(1-p,0) \\ \vdots & \vdots & & \vdots \\ d_n(0,-p) & d_n(-1,-p) & & d_n(1-p,-p) \\ d_n(1,0) & d_n(0,0) & & d_n(2-p,0) \\ \vdots & \vdots & \cdots & \vdots \\ d_n(1,-p) & d_n(0,-p) & & d_n(2-p,-p) \\ \vdots & \vdots & \ddots & \vdots \\ d_n(p-1,0) & d_n(p-2,0) & & d_n(0,0) \\ \vdots & \vdots & \cdots & \vdots \\ d_n(p-1,-p) & d_n(p-2,-p) & & d_n(0,-p) \end{bmatrix} \qquad \text{(III.99.b)}$$

$$\underline{a} = [a_1, a_2, \ldots, a_p]^T \qquad \text{(III.99.c)}$$

$$\underline{b}_{k_2} = [d_n(1,0), .., d_n(1,-p), d_n(2,0), .., d_n(2,-p), \ldots, d_n(p,0), .., d_n(p,-p)]^T \qquad \text{(III.99.d)}$$

Obsérvese que esta matriz $\underline{\underline{Q}}$ está formada por **p(p+1)** filas y **p** columnas y, aplicando las propiedades de simetría de los cumulantes de orden **n** (III.24), precisa calcular los siguientes valores de los cumulantes:

$$d_n(0..2p\text{-}1, 0..p) \quad \& \quad d_n(2p,p) \qquad \text{(III.100)}$$

En los métodos presentados anteriormente se ha asumido, implícitamente o explícitamente, la existencia de slices unidimensionales con rango completo **p**. Esta suposición puede ser apropiada para un modelado donde la respuesta impulsional desconocida se aproxima por un modelo AR, pero no para la identificación de un proceso donde **H(z)** se suponga que equivale a un proceso AR. En [Swam-89c] se demuestra mediante algunos ejemplos que un slice unidimensional no suele originar una matriz $\underline{\underline{Q}}$ de rango completo **p**. Posteriormente se demuestra analíticamente que no se puede asegurar la existencia de un slice de rango completo **p**, pues un slice de rango completo debe satisfacer unas restricciones muy fuertes respecto, por ejemplo, la posición de los polos del sistema AR y, en consecuencia, es bastante probable que origine un rango menor a **p**, que incluso puede ser cero. Sin embargo, tomando **p+1** slices en las ecuaciones normales (III.97), con $\mathbf{k_2}$**=-p, .. , 0** , se obtiene siempre una matriz de rango completo y, en consecuencia, se puede asegurar la consistencia de su estimación AR resultante [Swam-89c]. En [Tugn-89], también se presenta un método para obtener información referente a los polos a partir de un número finito de **p+1** slices, convenientemente elegidos, y originar un modelo causal AR(p) únicamente definido.

III.4.2.5. Método de la combinación lineal de slices ponderados (w-slice).

Esta metodología representa la respuesta impulsional del filtro AR a partir de una combinación lineal de slices de cumulantes de cualquier orden, aceptándose la posibilidad de combinar cumulantes de distinto orden. Esta combinación lineal ponderada de cumulantes de distinto orden pretende mejorar la calidad de la estimación AR, disminuyendo su varianza y su coste de cálculo con respecto a otras metodologías existentes. Se puede decir que esta técnica combina las estadísticas de orden superior con los algoritmos adaptativos.

Tal como se ha visto anteriormente, la respuesta impulsional **h(k)** de un sistema AR lineal e invariante satisface la recursión:

$$\sum_{i=0}^{p} a_i \,.\, h(k-i) \;=\; \delta(k) \qquad , \qquad (a_0 = 1) \tag{III.101}$$

Esta respuesta impulsional se puede relacionar con los cumulantes de orden n mediante la fórmula de Brillinger-Rosenblatt [Bril-67]:

$$c_n(k_1, k_2, \ldots, k_{n-1}) = \gamma_n^v \,.\, \sum_{k=-\infty}^{+\infty} \prod_{j=0}^{n-1} h(k+k_j) \qquad , \qquad \begin{matrix} k_0 = 0 \\ n \geq 2 \end{matrix} \tag{III.102}$$

Combinando linealmente (III.102) con los parámetros AR y haciendo uso de (III.101), bajo la suposición inicial **h(0)=1**:

$$\sum_{i=0}^{p} a_i \,.\, c_n(-i, k_2, \ldots, k_{n-1}) \;=\; \gamma_n^v \,.\, \prod_{j=2}^{n-1} h(k_j) \tag{III.103}$$

y al repetir la misma operación **n-1** veces para las **n-1** variables $\mathbf{k_1, .., k_{n-1}}$ se obtiene [Tugn-91]:

$$\sum_{l_1=0}^{p} \sum_{l_2=0}^{p} \ldots \sum_{l_{n-1}=0}^{p} a_{l_1} \,.\, a_{l_2} \ldots a_{l_{n-1}} \,.\, c_n(-l_1, -l_2, \ldots, -l_{n-1}) \;=\; \gamma_n^v \tag{III.104}$$

En el método anterior [Gian-90a], se vio como la respuesta impulsional **h(k)** se tomaba a partir de (III.103) fijando $\mathbf{k_3=k_4=...=k_{n-1}=0}$:

$$\sum_{l=0}^{p} a_l \,.\, c_n(-l, k, 0, \ldots, 0) \;=\; \gamma_n^v \,.\, h(k) \tag{III.105}$$

Sin embargo, esta no es la única combinación lineal de cumulantes que puede conducir a **h(k)**. A partir de (III.103) se deduce fácilmente que cualquier combinación lineal del tipo:

$$w(k) = \sum_{l_1=0}^{p} \sum_{l_2} \ldots \sum_{l_{n-2}} a_{l_1} \cdot w_{l_2} \ldots w_{l_{n-2}} \cdot c_n(-l_1, l_2, \ldots, l_{n-2}, k) \quad \text{(III.106)}$$

reconstruye la respuesta impulsional como:

$$w(k) = \mu \cdot h(k) \quad \text{(III.107.a)}$$

donde

$$\mu = \gamma_n \cdot \prod_{i=2}^{n-2} \sum_{l_i=0}^{p} w_{l_i} \cdot h(l_i) \neq 0 \quad \text{(III.107.b)}$$

Cualquier combinación lineal de slices de orden **n≥2** satisface las relaciones (III.105), (III.106) y (III.107) anteriores. A continuación se considera la siguiente combinación ponderada de slices:

$$c_w(k) = w_2.c_2(k) + \sum_{j=-M}^{N} w_3(j).c_3(k, j) + \sum_{j=-M}^{N} \sum_{l=-M}^{j} w_4(j, l).c_4(k, j, l) + \ldots \quad \text{(III.108)}$$

y mediante una elección adecuada de los coeficientes de ponderación **w**, $c_w(k)$ conduce a la respuesta impulsional **h(k)** sin la necesidad del conocimiento previo de los coeficientes **a**. Así, en [Vida-94], se demuestra que para $c_w(k)$ causal se verifica:

$$c_w(k) = c_w(0) \cdot h(k) \quad \text{(III.109)}$$

imponiendo las condiciones siguientes:

$$c_w(k) = 0 \quad , \quad k < 0 \quad \text{(III.110.a)}$$

$$c_w(0) \neq 0 \quad \text{(III.110.b)}$$

que permiten obtener los coeficientes de ponderación w_2, $w_3(j)$, $w_4(j,l)$, ... En principio, las condiciones (III.110) originan un conjunto infinito de ecuaciones lineales. Este problema se resuelve limitando el sistema a un número **P** de ecuaciones, siendo **P** una cota superior de **p**:

$$c_w(k) = 0 \quad , \quad -P \leq k < 0 \quad \text{(III.111)}$$

entonces se satisface [Vida-94]:

$$c_w(k) = c_w(0) \,.\, h(k) \qquad , \qquad k \geq -P \qquad \text{(III.112)}$$

Obsérvese que este método permite el uso de los cumulantes de segundo orden en la generación del slice ponderado $\mathbf{c_w(k)}$, siendo especialmente indicado cuando el proceso $\mathbf{x(k)}$ observado provenga de un entorno no ruidoso. Además, la calidad de la respuesta impulsional estimada está estrechamente relacionada con la fidelidad de la estimación de los cumulantes de tercer y cuarto orden. En [Gian-92] se ha estudiado la convergencia de las estimaciones de cumulantes hacia sus respectivos valores auténticos, habiéndose obtenido tres condiciones a satisfacer:

1) el sistema lineal $\mathbf{h(k)}$ debe ser exponencialmente estable,

2) el proceso aleatorio $\mathbf{v(k)}$ entrante debe ser estacionario,

3) los cumulantes de tercer orden deben ser absolutamente sumables hasta orden 6 y, asimismo, para el caso de considerar cumulantes de cuarto orden, entonces, éstos deben ser absolutamente sumables hasta orden 8.

La obtención de los coeficientes de ponderación (III.111) puede representarse en notación matricial:

$$\underline{\underline{S}}_a \,.\, \underline{w} = \underline{1} \qquad \text{(III.113.a)}$$

donde $\underline{\underline{S}}_a$ representa la matriz w-slice anticausal:

$$\underline{\underline{S}}_a = \begin{bmatrix} c_2(-P) & c_3(-P, j) & \cdots & c_4(-P, j, l) & \cdots & c_5(-P, j, l, m) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ c_2(0) & c_3(0, j) & \cdots & c_4(0, j, l) & \cdots & c_5(0, j, l, m) & \cdots \end{bmatrix} \qquad \text{(III.113.b)}$$

donde $\mathbf{j,l,m}$ toman valores entre $\mathbf{-M}$ y $\mathbf{N}$; $\underline{w}$ es el vector de ponderación:

$$\underline{w} = [w_2, w_3(j), \ldots, w_4(j, l), \ldots, w_5(j, l, m), \ldots]^T \qquad \text{III.113.c)}$$

y $\underline{1}$ es la respuesta impulsional anticausal:

$$\underline{1} = [0, 0, \ldots, 1]^T \qquad \text{(III.113.d)}$$

En principio, esta ecuación (III.113) tiene solución pero puede no ser única. Sin embargo, considerando la matriz pseudo-inversa de Moore-Penrose $\underline{\underline{S}}_a^{\#}$, se asegura la elección de la solución de norma mínima dentro del subespacio de soluciones exactas. Entonces, si se conoce la cota superior **P** del orden **p** del modelado AR se debe verificar:

$$N \geq 0 \quad , \quad M \geq P \tag{III.114}$$

Si se usa la Descomposición en Valores Singulares para el cálculo de $\underline{\underline{S}}_a^{\#}$, entonces, se puede decir que la solución está correctamente condicionada, es decir, errores en la estimación de los componentes de esta matriz se propagan débilmente hacia los parámetros AR a estimar..

Sintetizando lo deducido anteriormente se puede concretar el algoritmo w-slice en tres etapas que permiten hallar los coeficientes $\underline{a}$ del modelo AR deseado:

Paso 1: cálculo de los coeficientes de ponderación de norma mínima que conduce a la elaboración del slice ponderado que verifica $\mathbf{c_w(0)=1}$:

$$\underline{w}_n = \underline{\underline{S}}_a^{\#} \cdot \underline{1} \tag{III.115}$$

Paso 2: estimar la parte causal de la respuesta impulsional:

$$\underline{\hat{h}} = \underline{\underline{S}}_c \cdot \underline{w}_n \tag{III.116.a}$$

donde la matriz $\underline{\underline{S}}_c$ corresponde a una compañera causal de $\underline{\underline{S}}_a$:

$$\underline{\underline{S}}_c = \begin{bmatrix} c_2(P) & c_3(P, j) & \cdots & c_4(P, j, l) & \cdots & c_5(P, j, l, m) & \cdots \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ c_2(0) & c_3(0, j) & \cdots & c_4(0, j, l) & \cdots & c_5(0, j, l, m) & \cdots \end{bmatrix} \tag{III.116.b}$$

y la respuesta impulsional estimada viene representada por:

$$\underline{\hat{h}} = \left[\hat{h}(P), \ldots, \hat{h}(1), 1\right]^T \tag{III.116.c}$$

Paso 3: resolver (III.101) mediante los valores estimados para $\underline{\hat{h}}$:

$$\begin{bmatrix} \hat{h}(0) & 0 & \cdots & 0 \\ \hat{h}(1) & \hat{h}(0) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{h}(P-1) & \hat{h}(P-2) & \cdots & \hat{h}(0) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = - \begin{bmatrix} \hat{h}(1) \\ \hat{h}(2) \\ \vdots \\ \hat{h}(P) \end{bmatrix} \tag{III.117}$$

Para este algoritmo pueden deducirse algunas observaciones. En principio, la ecuación (III.113) parece necesitar un número elevado de términos cuando se consideren órdenes $n \geq 3$ para los slices, puesto que los índices **j, l, ...** pueden tomar valores entre **-M** y **N**. Sin embargo, tomando slices unidimensionales, por ejemplo fijando $l_2=l_3=..=l_{n-2}=0$ en (III.106) y (III.107), se garantiza la obtención de estimaciones AR no sesgadas. Bajo estas condiciones el número necesario de valores de cumulantes es idéntico al método anterior de las ecuaciones de Yule-Walker de orden superior.

Los valores de los parámetros **M** y **N** pueden eligirse independientemente para cada orden **p**, como parámetros de diseño. Evidentemente, si se eligen valores **M, N** que impliquen la utilización de un mayor número de valores de cumulantes se puede esperar una progresiva reducción de la varianza. Aunque esta propiedad también dependerá de la varianza de los nuevos valores correspondientes a los cumulantes adicionales añadidos. En [Pora-89] puede encontrarse un estudio sobre la influencia de la calidad de la estimación de los cumulantes en la estimación posterior de los coeficientes AR. Además, en (III.112) debe elegirse un valor suficientemente grande para **P** si se desea evitar el efecto acumulativo de varianza en los parámetros AR, siendo necesaria una resolución de esta sobredeterminación mediante el criterio de Mínimos Cuadráticos.

CAPITULO IV

El Algoritmo Iterativo de Wiener.

IV.1 Introducción.

Los algoritmos considerados en este capítulo se caracterizan por eliminar el ruido presente en la señal de voz mediante un filtrado iterativo de Wiener, de manera que a la salida de dicho filtro se obtenga una señal de voz realzada **y(n)** lo más parecida posible a la señal de voz original **s(n)**. Tal como se vio en el Capítulo II, este filtro de Wiener, óptimo y no causal, se diseña a partir de las densidades espectrales correspondientes a las señales de voz y ruido, segun la expresión (II.22). Un primer problema viene dado por la no estacionariedad de los procesos voz y ruido. La no estacionariedad de ambos procesos estocásticos obliga a actualizar continuamente sus estimaciones espectrales, segun se expone en (II.23). No obstante, debe reseñarse que en la literatura alguna vez se ha tomado en consideración el uso de un filtro de Wiener invariante, donde el filtro no se actualiza trama a trama cuando la voz pierde la hipótesis de estacionariedad. Este filtro invariante se basa en un espectro de potencia estimado para el ruido y para la señal de voz se obtienen sus características promediando un extenso intérvalo de tiempo. Alguna de estas características a largo plazo viene representada, por ejemplo, por el hecho que el espectro promediado de voz decae con una pendiente de 6dB/octava al aumentar la frecuencia.

Evidentemente, el filtro de Wiener actualizado con el transcurrir de las tramas de voz conduce a unas prestaciones de calidad e inteligibilidad muy superiores. Mientras que el ruido puede considerarse estacionario durante intérvalos relativamente largos, la señal de voz puede verse como un proceso cuasiestacionario, donde las premisas de estacionariedad pierden validez cuando los intérvalos considerados superan los 30-35 mseg. Por otra parte, la estimación espectral de voz es más fiel, menor sesgo y menor varianza, cuando se considera un conjunto de muestras lo más extenso posible. Por esta razón se ha realizado el procesado mediante tramas de duración 32 mseg y en cada trama debe obtenerse una estimación espectral de la voz. Como las características espectrales de la señal de voz varían, trama a trama, se debe ejecutar la estimación espectral de la voz a partir de la trama actual disponible, y así poder filtrar esta trama de voz ruidosa. Obsérvese que la estimación del espectro de potencia de la voz debería ser conocido a priori.

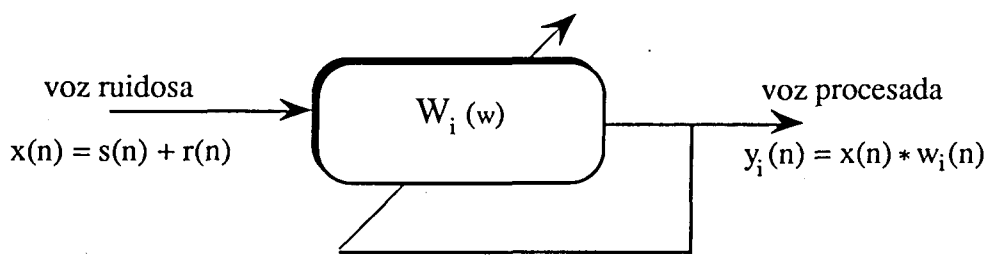
Otro problema de mayor magnitud aparece cuando se dispone sólamente de la señal de voz ruidosa. A primera vista, la información espectral correspondiente al ruido puede ser extraída durante los intérvalos de silencio, caracterizados por su nula actividad de voz, y sacar provecho de su mayor estacionariedad. En cambio, las tramas de voz siempre aparecen contaminadas por ruido de fondo. Aunque alguna trama puntual de voz tuviera ruido nulo, esta buena estimación solo sirve para esta trama y no sirve de ayuda para las estimaciones correspondientes a tramas posteriores. Anteriormente se han comentado dos posibles estrategias a tener en cuenta durante la etapa de diseño del filtro de Wiener. Al principio, los primeros sistemas consideraban la expresión (II.24), bastante más lógica si tenemos en cuenta que sólamente se dispone de la señal de voz ruidosa y a lo sumo del ruido durante los intérvalos de silencio. Otra posibilidad considerada en éstos inicios, consiste en estimar directamente el espectro de potencia de la voz original a partir, usualmente, de alguna técnica de Sustracción Espectral. Sin embargo, todos estos métodos muestran bastante sensibilidad al ruido y, en consecuencia, el ruido residual es demasiado significativo después de aplicar el sistema de realce de la voz. En relación a estos primeros sistemas se puede concluir que tienen a su favor una cierta simplicidad de cálculo pero, en cambio, cuando el nivel de ruido es importante sus prestaciones no son buenas.

En consecuencia, este trabajo se fundamenta sobre una técnica básica más compleja pero, a su vez, bastante mejor en sus prestaciones. Se diseña el filtro de Wiener a partir de la expresión (II.23) y, por tanto, el principal problema reside en como obtener una buena estimación espectral de la voz original. Para ello se considera un modelado paramétrico AR de la voz original. Nótese que las prestaciones obtenidas mediante esta técnica de Filtrado de Wiener dependen, en gran medida, de la bondad de esta estimación de $P_s(w)$. En el apartado III.4.2. del capítulo anterior, se ha comentado la superioridad de las técnicas de estimación

paramétrica por encima de los estimadores convencionales, que hacen uso de la Transformada de Fourier o algoritmos rápidos FFT. Además de su menor capacidad para discernir entre componentes armónicos próximos, estas técnicas convencionales obtienen una fidelidad espectral o biespectral bastante pobre cuando se trabaja con procesos paramétricos. Recuérdase que la señal de voz puede interpretarse como un proceso paramétrico AR cuyos polos estan relacionados con las frecuencias de resonancia de la cavidad vocal.

En el Capítulo II se ha discutido como la estimación de los coeficientes $\underline{a}$ comporta la resolución de un sistema lineal de ecuaciones. El problema aparece cuando la señal de voz disponible está contaminada con ruido, pues, el sistema de ecuaciones pasa a ser no lineal y su resolución se complica bastante. Para evitarlo se considera un método que proporciona una solución subóptima, a cambio de tratar sistemas lineales de ecuaciones: el Filtrado Iterativo de Wiener. Obsérvese que esta solución subóptima conduce a una estimación paramétrica aproximada de la señal de voz y, como se discute posteriormente en presencia de ruido, esta aproximación es la causante de la aparición de una cierta distorsión que aumenta con el número de iteraciones ejecutadas. Esta distorsión se traduce básicamente en un desplazamiento y estrechamiento de los formantes de la señal de voz. Este efecto también se conoce como efecto de picado espectral y se estudia ampliamente en el apartado IV.5, donde se discute como el filtro estimado no converge hacia el óptimo para ciertas frecuencias.

En el presente capítulo se analiza el comportamiento de este algoritmo iterativo de Wiener cuando se usan estadísticas de orden superior durante el proceso de estimación paramétrica de la señal de voz. Primeramente, en el Apartado IV.2, se evalua el comportamiento del algoritmo iterativo de Wiener clásico, que se caracteriza por realizar la estimación de los coeficientes $\underline{a}$ del modelo AR de la voz mediante el empleo de las



$$W_i(w) = \frac{P_{y_{i-1}}(w)}{P_{y_{i-1}}(w) + P_r(w)} \qquad P_{y_{i-1}}(w) = \frac{g^2}{\left| 1 - \sum_{k=1}^{P} a_k \cdot e^{-jwk} \right|^2}$$

Figura IV.1 : *Esquema del algoritmo iterativo de Wiener con estimación paramétrica AR.*

estadísticas clásicas de segundo orden, basadas en la función autocorrelación. Así, este algoritmo iterativo de Wiener, que estima un modelo AR para la voz mediante estadísticas de segundo orden, se denomina en lo sucesivo como algoritmo AR2. De forma análoga los algoritmos cuya estimación paramétrica AR de la voz se realiza mediante el uso de estadísticas de tercer y cuarto orden, se han denominado respectivamente algoritmos AR3 y AR4. Segun las características propias de las estadísticas de orden superior, presentadas en el capítulo precedente, se espera un mayor desacoplo voz-ruido que permita una estimación paramétrica AR menos sensible. Estos algoritmos de orden superior se estudian en los Apartados IV.3 y IV.4 respectivamente, donde se comparan sus prestaciones a las propias del algoritmo clásico AR2, bastante conocido en estos entornos de aplicación cuando el ruido aditivo es Gaussiano blanco [Lim-79] o ruido Gaussiano coloreado [Hansl-87]. Un estudio comparativo de estos algoritmos AR2, AR3 y AR4 para distintos tipos de ruido ha sido discutido en [Sala-93a] y [Sala-93b].

Como síntesis a lo expuesto, se considera un filtrado iterativo de Wiener originario de Lim y Oppenheim [Lim-79]: la estimación del filtro óptimo de Wiener se estima iterativamente a partir de las sucesivas salidas obtenidas por los sucesivos filtros, tal como se muestra en la Fig.IV.1. En la primera iteración del algoritmo se obtiene la estimación del filtro de Wiener a partir de la señal de voz ruidosa disponible. En lugar de aceptar su salida, $\mathbf{Y_1(w)=W_1(w).X(w)}$, se puede volver a estimar el filtro por segunda vez $\mathbf{W_2(w)}$ a partir de la nueva señal de voz disponible $\mathbf{y_1(n)}$, menos ruidosa en relación a la disponible en la iteración anterior $\mathbf{y_0(n)=x(n)}$. Con este segundo diseño del filtro se espera obtener una segunda señal $\mathbf{y_2(n)}$ cuyo nivel de ruido debe ser inferior al presente en $\mathbf{y_1(n)}$. De esta manera se puede ir iterando sucesivamente. Parece lógico obtener un diseño menos sensible al ruido para dicho filtro cuando se procesan varias iteraciones y, en consecuencia, la señal de voz realzada a la salida debería presentar menos ruido a medida que aumenta el número de iteraciones. Sin embargo, un mayor número de iteraciones conlleva una mayor distorsión ocasionada, tal como se muestra en el apartado IV.5, y se debe alcanzar una solución de compromiso entre ambos efectos. Nótese que los sucesivos diseños del filtro de Wiener siempre se aplican sobre la señal de voz ruidosa $\mathbf{x(n)}$.

En este trabajo, se obtiene la estimación espectral de la voz a partir de un modelado paramétrico AR y, entonces, la estimación de $\mathbf{P_{y_i}(w)}$ se traslada al problema de la estimación de los parámetros $\underline{\mathbf{a}}$ de este modelo de la voz. Un esquema más detallado del diagrama de bloques asociado con este método iterativo de Wiener se muestra en la Fig.IV.2.

Además de lo expuesto anteriormente, se debe tener presente el incremento de coste de cálculo implícito en este algoritmo iterativo. Este factor resulta muy importante si se pretende

obtener un sistema competitivo para aplicaciones en tiempo real. En consecuencia, parte de nuestro esfuerzo se encamina, también, hacia la reducción del número de iteraciones necesarias para la obtención de la señal de voz realzada definitiva.

Llegados a este punto, vamos a exponer con cierto detalle los distintos bloques que componen nuestro algoritmo iterativo de Wiener. Este algoritmo iterativo representa la columna vertebral sobre la que se sustentan todas las variantes descritas en apartados posteriores. Su diagrama de bloques general se ha representado en la Fig.IV.3, donde la secuencia propia del algoritmo se ha representado en linea continua mientras en linea discontinua se ha representado la disponibilidad de una cierta señal, presente a la salida de una etapa de procesado y que se usa como entrada para otro bloque de procesado.

La evaluación de las prestaciones ofrecidas por los distintos algoritmos se efectúa a partir de la disponibilidad de una frase de voz original correspondiente a un determinado locutor. Se dispone de varios locutores masculinos y femeninos cuya voz ha sido muestreada a $\mathbf{F_m}$=**8kHz**. La señal de voz ruidosa se genera añadiendo al fichero de voz original un nivel de ruido, en base a una SNR global prefijada. Así, la mayor parte de medidas espectrales y temporales resultan de la comparación entre el fichero de voz realzada, suministrado por el algoritmo considerado, y el fichero original de voz. Al inicio de cada fichero de voz existe siempre un intérvalo de silencio superior a $\mathbf{M_r}$ tramas, puesto que la densidad espectral correspondiente al ruido $\mathbf{P_r(w)}$ se estima siempre durante estas $\mathbf{M_r}$ tramas iniciales. Este algoritmo procesa la señal de voz ruidosa trama a trama y, para cada trama, procesa un número de iteraciones prefijado al inicio.

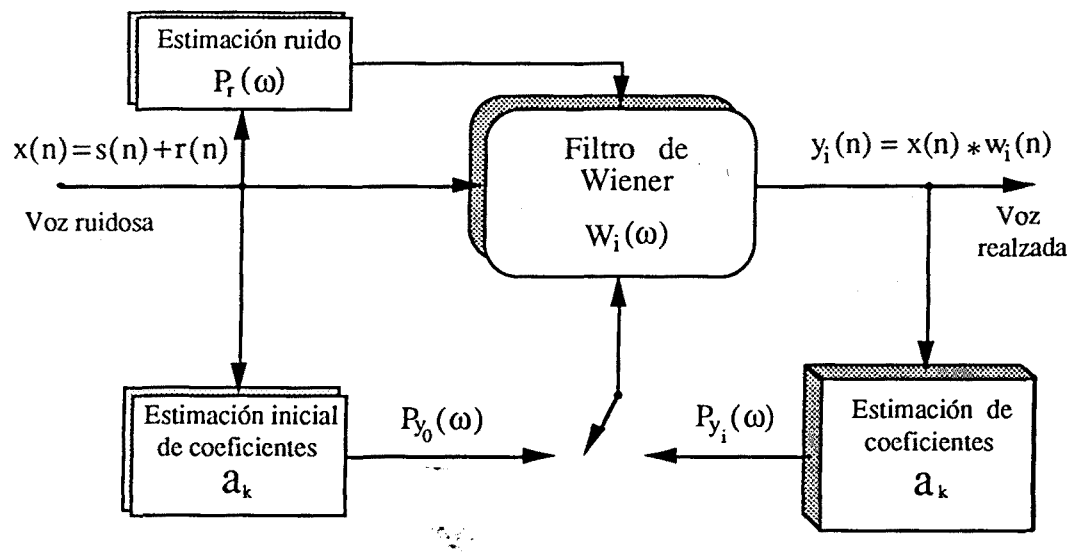


Figura IV.2 : Esquema básico del algoritmo iterativo de Wiener considerando estimación paramétrica todo polos (AR) durante el diseño del filtro.

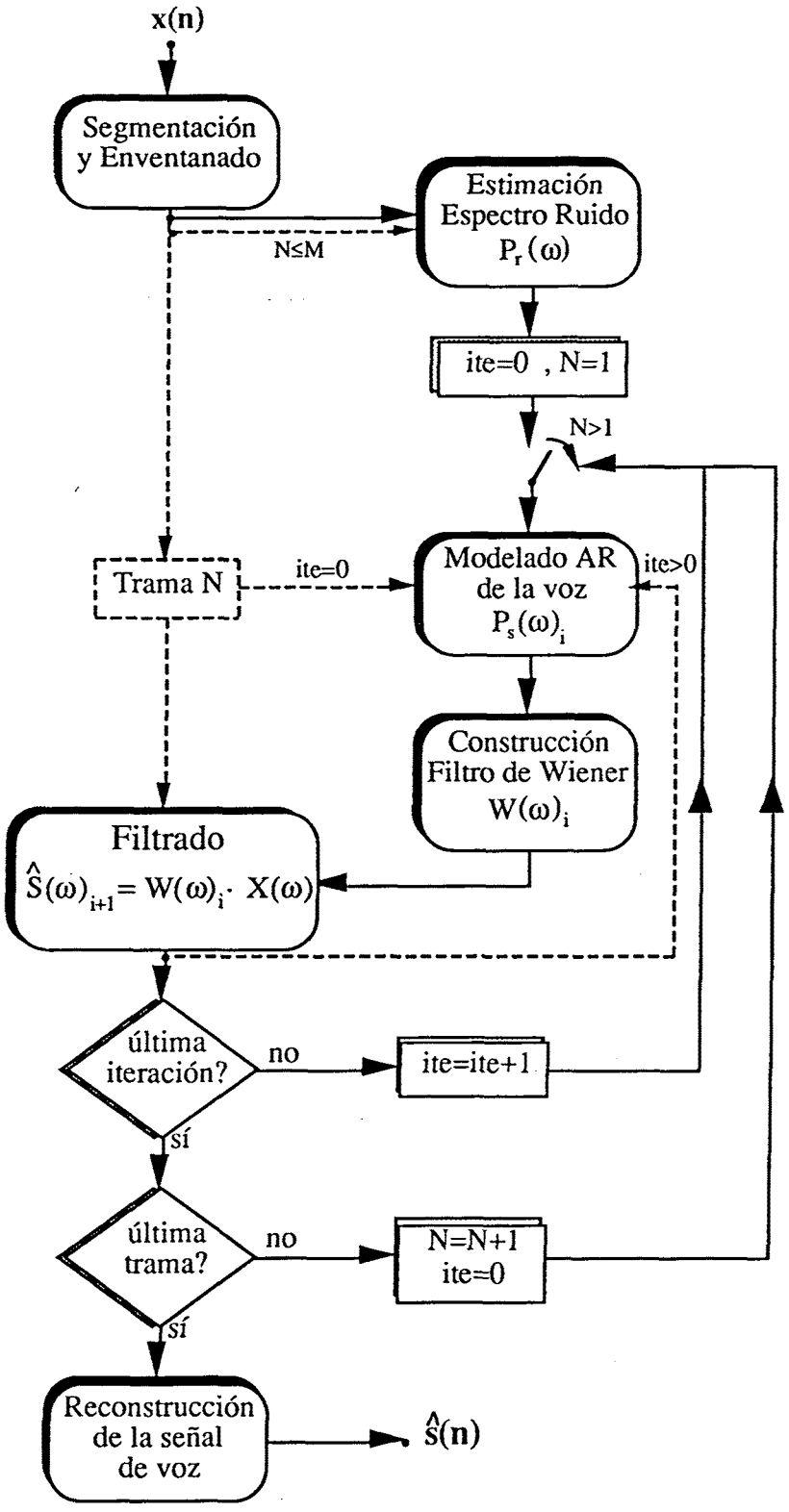


***Figura IV.3** : Diagrama de bloques detallado del algoritmo iterativo de Wiener (línea continua: flujo del algoritmo, línea discontinua: flujo de la señal).*

IV.1.1. Segmentación y Enventanado. Reconstrucción

Vimos en el Capítulo II cómo la señal de voz no cumple la propiedad de estacionariedad para intérvalos de tiempo relativamente largos: la variación continuada del tracto vocal y su excitación al generar los distintos fonemas que componen la voz imposibilita totalmente un procesado directo de toda la señal. Aparece, entonces, la necesidad de segmentarla en bloques o tramas de corta duración, de manera que pueda considerarse estacionaria dentro de ese intervalo relativamente corto. Se ha tomado una longitud de trama de **N=256** muestras, valor múltiplo de 2 que facilita los cálculos de FFT. A una frecuencia de muestreo típica de 8KHz, que es la utilizada en las señales de que disponemos, equivale a 32mseg. de duración, intérvalo de tiempo situado en el límite de la estacionariedad de la voz.

Una vez segmentada la señal procedemos a enventanar cada una de las tramas obtenidas con una ventana adecuada, de modo que suavice los extremos y evite así las transiciones bruscas entre tramas. Para que no existan problemas en la reconstrucción de la señal, dicha ventana debe cumplir la siguiente condición:

$$\sum_{m=0}^{M-1} v\,(k - m.D) \;=\; 1 \qquad \text{(IV.1)}$$

donde **v(k)** es la ventana, **D** el desplazamiento entre tramas y **M** el número de tramas que forman el fichero de voz considerado. Se ha elegido la ventana de Hanning definida por la siguiente expresión:

$$v_{Hanning}(k) \;=\; \frac{1}{2} \cdot \left[1 \;-\; \cos\left(2\pi . \frac{k}{N-1} \right) \right] \qquad ,, \qquad k=0..N-1 \qquad \text{(IV.2)}$$

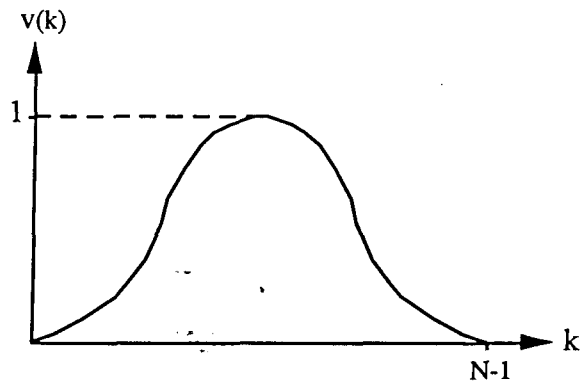


Figura IV.4 : Ventana de Hanning.

Se ha considerado un solapamiento del 50 por ciento entre tramas consecutivas, es decir, se ha utilizado un desplazamiento, entre ventanas, de media longitud de trama **D=N/2.** Bajo estas condiciones, se comprueba fácilmente el cumplimiento de la condición (IV.1). En la Fig.IV.5 se muestra gráficamente el proceso de segmentación y enventanado, así como el cumplimiento de dicha condición durante el proceso de reconstrucción.

Como conclusión, la señal ruidosa de entrada se corta en distintas tramas de longitud **N=256** y, una vez enventanada mediante una ventana de Hanning, se envía al algoritmo para que la procese. Así, la señal a procesar puede verse como el producto temporal de dichas señales:

$$x_v(n) = x(n) \cdot v_{Hanning}(n) \qquad \text{(IV.3)}$$

que en el dominio frecuencial tiene como expresión equivalente la convolución de ámbas:

$$X_v(w) = X(w) * V_{Hanning}(W) \qquad \text{(IV.4)}$$

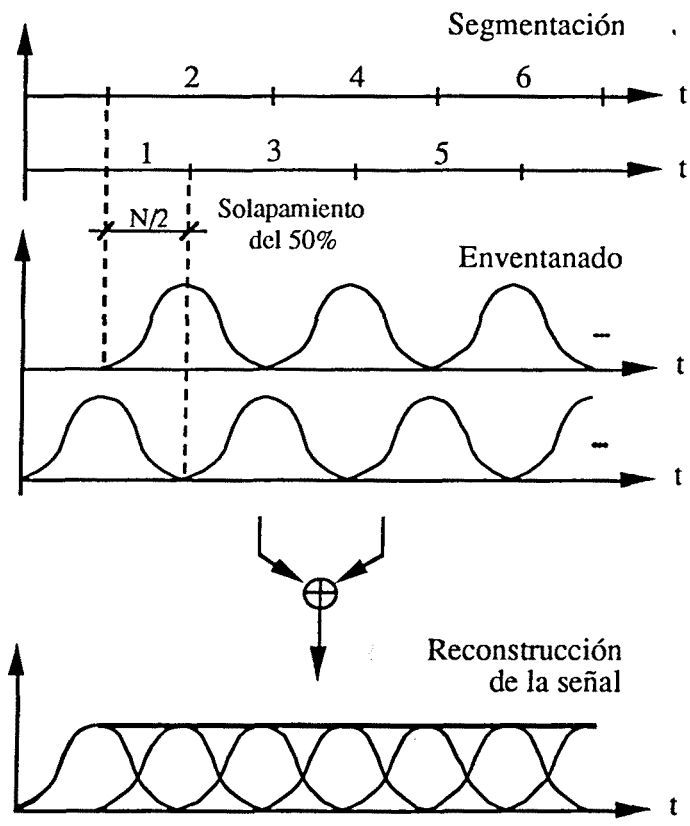


Figura IV.5 : Procesos de Segmentación y Enventanado,y su posterior Reconstrucción.

Nos interesa, pues, una ventana que modifique lo menos posible el espectro de la señal original, es decir, una ventana cuyo espectro tenga un lóbulo principal lo más estrecho posible (alta resolución) y cuyo nivel de lóbulos secundarios sea también lo más bajo posible (bajo rizado).

La reconstrucción de la señal, una vez procesadas todas las iteraciones para cada una de las tramas que la componen, se lleva a término segun la representación de la Fig.IV.6. Siguiendo el proceso inverso de la segmentación, en cada segmento de 256 muestras de la señal reconstruida, contribuirán de forma adecuada las muestras correspondientes a 3 tramas de señal segmentada: las últimas 128 muestras de la trama solapada inmediatamente anterior, las 256 muestras de la trama actual y las 128 primeras muestras de la trama solapada inmediatamente posterior.

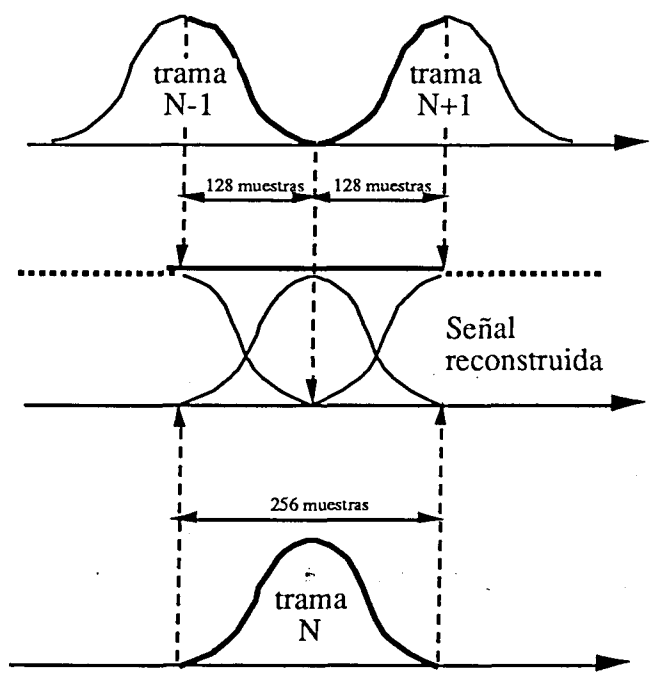


Figura IV.6 : Proceso de Reconstrucción de la señal realzada.

IV.1.2. Estimación del Espectro del Ruido

En el Capítulo I, cuando presentamos el filtro de Wiener, se expuso la necesidad de obtener una buena estimación del espectro del ruido presente en la señal de voz que conduzca a una implementación de dicho filtro lo más óptima posible. En el supuesto de disponer de una señal de referencia de ruido (Técnicas Multimicrófono), totalmente incorrelada con la señal de voz, podemos obtener una estimación de su espectro realmente precisa, gracias a la posibilidad de ir actualizándola trama a trama. Sin embargo, en el caso que nos ocupa, disponemos de una única señal: la señal compuesta de voz y ruido **x(n)**. La ventaja fundamental al tratar de estimar el ruido reside en la existencia de intérvalos de silencio, localizados entre las distintas ráfagas con actividad de voz.

$$\begin{array}{ll} x(n) = s(n) + r(n) & \text{períodos de actividad} \\ x(n) = r(n) & \text{períodos de silencio} \end{array} \qquad \text{(IV.5)}$$

Como la estacionariedad de la señal de ruido es mucho mayor que la de la señal de voz, la hipotesis de estacionariedad a lo largo de la siguiente ráfaga de tramas con actividad vocal puede ser, en general, bastante realista. Al aparecer un nuevo intérvalo carente de actividad de voz, la estimación del espectro de ruido podrá ser adecuadamente actualizada. Esto nos plantea la necesidad de implementar un detector de zonas de actividad/silencio si queremos que el sistema funcione sistemáticamente en tiempo real.

En condiciones de laboratorio, como es nuestro caso, disponemos de frases de corta duración, de manera que utilizamos la zona inicial del fichero para realizar la estimación del ruido y lo consideramos estacionario a lo largo de toda la frase. Esto no supone ninguna limitación de cara a una implementación real del algoritmo, pues, se precisan de 5 a 8 tramas de ruido al principio de cada ráfaga de voz.

Para la estimación de la densidad espectral de potencia del ruido se ha utilizado el algoritmo de Welch. Este algoritmo realiza un promediado de los espectros de potencia, obtenidos para cada una de las $\mathbf{M_r}$ tramas iniciales, mediante una estimación por el Periodograma. Así, tras la segmentación y enventanado de la trama m-ésima, se obtiene su Transformada de Fourier Discreta $\mathbf{R^m(k)}$. Entonces se obtiene el Periodograma para dicha trama:

$$\frac{\left|R^m(k)\right|^2}{N} \qquad \text{(IV.6)}$$

siendo N la longitud de la trama. Del promediado de los periodogramas de las M_r tramas de ruido disponibles, resulta la estimación del espectro que estamos buscando:

$$E\left\{ |R(k)|^2 \right\} = P_r(k) \equiv \frac{1}{M_r} \cdot \sum_{m=1}^{M_r} \frac{|R^m(k)|^2}{N} \qquad \text{(IV.7)}$$

El proceso seguido por el algoritmo de Welch se muestra en la Fig.IV.7.

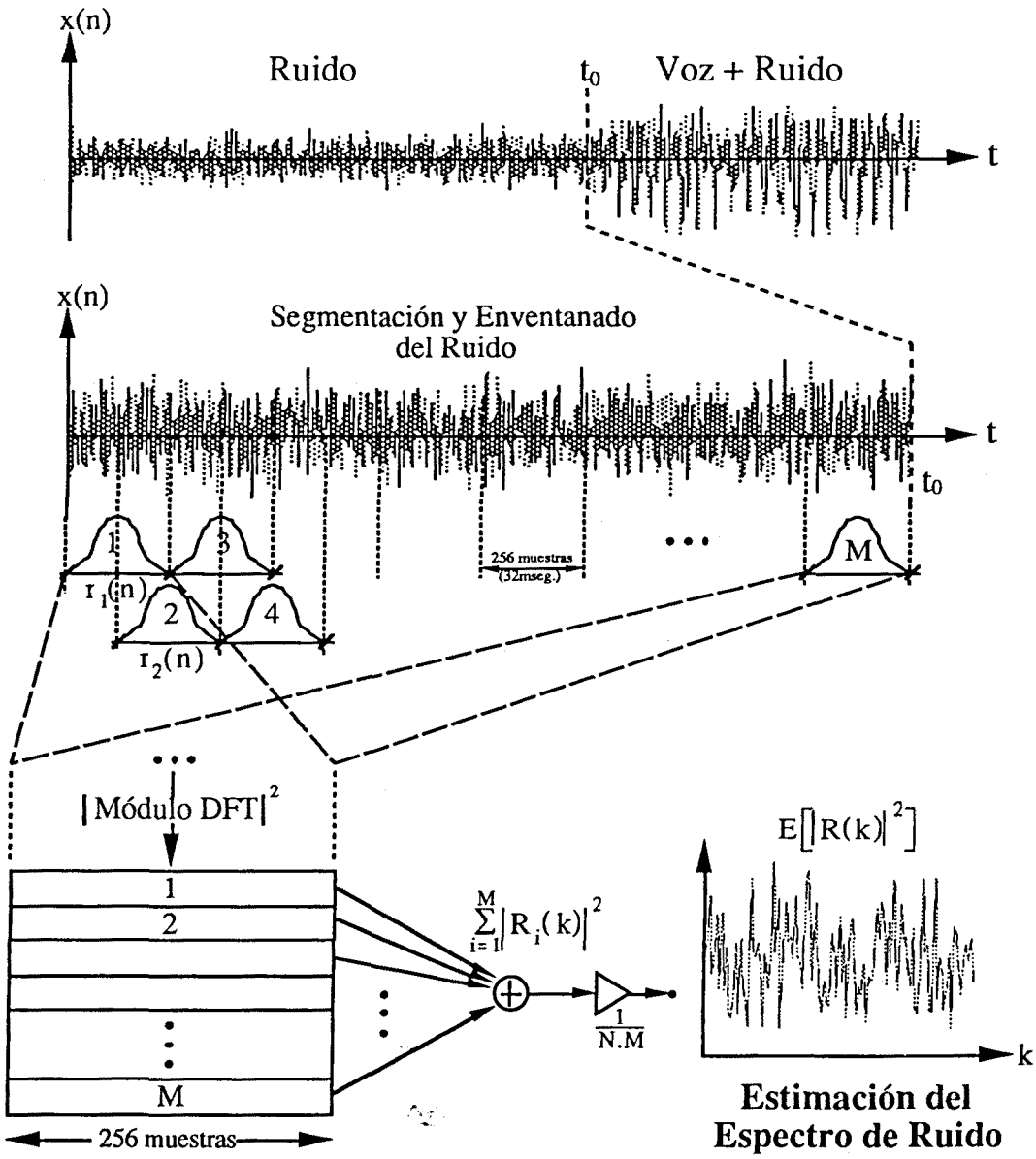


Figura IV.7 : Estimación de la densidad espectral del Ruido mediante el algoritmo de Welch.

IV.1.3. El Modelado AR de la Voz.

Una vez estimado el espectro de potencia del ruido, el algoritmo penetra dentro de un bucle de procesado que recorre, trama a trama, la totalidad de la señal. Para cada una de las tramas, una vez enventanadas, se llevarán a cabo las iteraciones de filtrado requeridas, independientemente del resto de la señal de voz. El primer paso de este proceso de filtrado es la estimación del espectro de la voz $P_S(w)$ a partir del modelo autorregresivo expuesto los capítulos anteriores. Recordemos que dicho modelo respondía a una función de transferencia de la forma:

$$V(z) = \frac{g}{1 - \sum_{k=1}^{p} a_k \cdot z^{-k}} \qquad \text{(IV.8)}$$

donde **p** es el orden del predictor lineal y **g** un factor de ganancia. Hemos obtenido un valor **p=10** como un buen compromiso para representar los formantes espectrales sin una carga excesiva de cálculo. Entonces, se resuelve la estimación de $P_{y_{i-1}}(w)$ a partir de este modelo:

$$P_{y_{i-1}}(w) = \frac{g^2}{\left| 1 - \sum_{k=1}^{p} a_k \cdot e^{-jwk} \right|^2} \qquad \text{(IV.9)}$$

El cálculo de los coeficientes a_k, como veremos en posteriores apartados, se puede llevar a cabo utilizando estadísticas de orden dos (correlaciones) o de orden superior (cumulantes). Por otro lado, el factor **g** de ganancia del modelo AR se calcula haciendo un balance de energías de la trama en consideración. En efecto, si:

$$x(n) = s(n) + r(n) \qquad \text{(IV.10)}$$

entonces se debe verificar la siguiente relación para sus energías:

$$E_x = E_s + E_r \qquad \text{(IV.11)}$$

bajo la hipótesis de incorrelación entre **s(n)** y **r(n)**. Como disponemos de la señal ruidosa, **x(n)**, y de la estimación de la señal de ruido que hemos realizado en el apartado anterior, podemos hallar así el valor de **g** que ajuste el modelo. Si el valor obtenido es negativo, $g \leq 0$, decidiremos que la trama en que nos encontramos contiene solamente ruido y le aplicaremos, consecuentemente, una atenuación arbitraria a elegir (generalmente 40dB), para evitar posteriormente el filtrado sobre dicha trama (ahorramos así tiempo de cálculo).

IV.1.3. Diseño del Filtro de Wiener.

Tal como se ha mencionado previamente, el filtro de Wiener no causal queda descrito por la siguiente expresión:

$$W(w) = \frac{P_s(w)}{P_s(w) + P_r(w)} \qquad \text{(IV.12)}$$

donde $P_s(w)$ y $P_r(w)$ son, respectivamente, las densidades espectrales de la señal de voz y del ruido, obtenidas por estimación en los bloques descritos anteriormente. Puesto que la condición de diseño del filtro viene impuesta sobre su módulo, y no sobre su fase, se aplica un desplazamiento de fase lineal, en frecuencia, de la forma:

$$\phi_w(w) = e^{-jw\frac{N-1}{2}} \qquad \text{(IV.13)}$$

Esta fase lineal en el dominio frecuencial significa un desplazamiento en el dominio temporal de **(N-1)/2** muestras en la respuesta del filtro, lo cual consigue convertir el filtro no causal en causal. Sin embargo dicho retardo deberá tenerse en cuenta tanto en el filtrado como en la evaluación temporal de la señal de salida respecto de la de entrada, por la presencia de un retardo adicional de media muestra.

Con el módulo y la fase que hemos descrito construimos el filtro de Wiener. Aplicando Transformada Inversa de Fourier se obtiene su respuesta impulsional **w(n)**. Cualquier presencia no causal en esta respuesta temporal del filtro, o la existencia de residuos complejos, por motivos de falta de precisión en los coeficientes, son eliminados tomándose sólamente la parte real de la respuesta temporal del filtro. En cualquier caso, para conseguir mayor precisión en los cálculos del filtrado posterior, se ha utilizado una longitud de trama de 512 muestras en la obtención de la respuesta frecuencial del filtro.

IV.1.4. El Filtrado de Wiener.

Una vez obtenido el filtro, el procedimiento de filtrado es simple. Si $\mathbf{x^m(n)}$ es la trama m-ésima de la señal de voz ruidosa **x(n)**, entonces para filtrarla hay que realizar la operación:

$$y_i^m(n) = x^m(n) * w_i^m(n) \qquad \text{(IV.14)}$$

siendo $\mathbf{w_i^m(n)}$ la respuesta impulsional del filtro de Wiener en la iteración i-ésima del algoritmo e $\mathbf{y_i^m(n)}$ la señal filtrada resultante de esa iteración, es decir, la estimación de la señal de voz limpia.

Sin embargo resulta más fácil realizar la operación en el dominio frecuencial, donde la operación de convolución se convierte en un simple producto de espectros:

$$Y_i^m(k) = X^m(k) \cdot W_i^m(k) \qquad \text{(IV.15)}$$

Hay que tener en cuenta que se está realizando una convolución circular, por lo que deberemos evitar los errores de superposición que puedan surgir. Como dijimos en el anterior apartado, doblamos la longitud del filtro para evitar posibles errores en el filtrado. Por la misma razón utilizamos como señal de entrada al filtro una secuencia compuesta por la trama actual y la anterior; ello contribuye a la obtención de una respuesta del filtro más próxima a la que obtendríamos sin enventanar.

Recordemos que se trata de un algoritmo iterativo: para una determinada trama de señal, la estimación del espectro de potencia de la señal de entrada del filtro de la iteración i-ésima del algoritmo, se realiza a partir de la señal de salida resultante de la iteración anterior, aunque la señal de entrada al filtro siempre se corresponde con la señal ruidosa inicial.

$$\hat{s}_{i-1}^m(n) = y_{i-1}^m(n) \xrightarrow{\text{estimación AR}} P_{y_{i-1}}^m(k) \qquad \text{(IV.16)}$$

IV.1.5. Medidas objetivas de Evaluación del Sistema.

A continuación exponemos brevemente el cálculo de cada uno de los tipos de medidas utilizadas sobre la señal de voz, así como la interpretación física o intuitiva que de ellas hacemos.

IV.1.5.1. La Relación Señal a Ruido global (SNRglobal).

La SNR_{global} es una relación entre la potencia total de la señal procesada y la potencia de la señal de ruido. Viene expresada por:

$$SNR_{global} = 10.\log\left\{\frac{\sum x^2(n)}{\sum e^2(n)}\right\} \qquad \text{(IV.17)}$$

donde:

$$e(n) = s(n) - x(n) \qquad \text{(IV.18)}$$

siendo **s(n)** la señal de referencia (señal limpia) y **x(n)** la señal a medir (señal ruidosa). Sin embargo, al realizar el promediado sobre intervalos largos de señal, las zonas de mayor energía contribuirán con mayor peso al valor definitivo de la medida. Por esta razón, la SNR_{global} se tendrá solamente en consideración de manera aproximada, para tener una idea más bien general de los niveles medios de ruido que presenta una determinada señal.

IV.1.5.2. La relación Señal a Ruido Segmentada (SNR_{Seg}).

Para la trama i-ésima de la señal, de longitud **N** muestras, se define su SNR como:

$$SNR_i = 10.\log\left\{\frac{\sum_{n=1}^{N} x_i^2(n)}{\sum_{n=1}^{N} e^2(n)}\right\} \qquad \text{(IV.19)}$$

Se define la SNR_{seg} de toda la señal de voz, como el promediado de todas las SNR_i parciales. Se puede escribir entonces de la siguiente manera:

$$SNR_{seg} = \frac{1}{M} \cdot \sum_{i=1}^{M} SNR_i = \frac{1}{M} \cdot \sum_{i=1}^{M} 10.\log\left\{\frac{\sum_{n=1}^{N} x_i^2(n)}{\sum_{n=1}^{N} e^2(n)}\right\} \qquad \text{(IV.20)}$$

Se intenta de esta forma equilibrar los distintos pesos de influencia de las zonas de alta y baja energía de la señal. Hay por ello dos puntos a tener en cuenta para la lectura de los resultados:

- Si la SNR_{seg} es mayor que la SNR_{global}, la distorsión es mayor para los sonidos de mayor energía que para los débiles.
- Por contra, si la SNR_{seg} es menor que la SNR_{global}, la distorsión será más alta para los sonidos de poca energía.

Tanto la SNR_{global} como la SNR_{seg} tienen un problema básico que convierte su uso en delicado. El cálculo del ruido **e(n)** como muestra la ecuación (IV.18) puede llevar a serias complicaciones cuando el retardo o la fase de ambas señales no sean exactamente iguales. El retardo de media muestra introducido por el filtro de Wiener en el algoritmo de mejora no permite una buena evaluación de las medidas temporales, por lo que se ha retardado la señal de referencia también en media muestra para que ambas señales puedan ser comparadas. Este problema es despreciable en el dominio frecuencial.

IV.1.5.3. Distancia de Itakura o de Máxima Verosimilitud (ITAKU).

La distancia de Máxima Verosimilitud Logarítmica, LLR (Log Likelihood Ratio), es una variación de la distancia de Itakura, que definiremos a continuación.

El error de predicción **e(n)** de una señal **x(n)**, expresado por medio de una combinación lineal de sus **p** muestras anteriores, viene dado por:

$$e(n) = \sum_{i=0}^{p} a_i \cdot x(n-i) \qquad \text{con} \quad a_0 = 1 \tag{IV.21}$$

de forma que el error cuadrático medio vendrá dado por:

$$\alpha = \sum_{-\infty}^{+\infty} e^2(n) \tag{IV.22}$$

El filtro que minimiza dicho error puede obtenerse siguiendo los criterios habituales del método de las correlaciones, y queda definido por la siguiente expresión:

$$A(z) = \sum_{i=0}^{p} a_i \cdot z^{-i} \qquad \text{con} \quad a_i = 1 \tag{IV.23}$$

Si la misma secuencia **x(n)** se pasa a través de otro filtro, que minimiza la energía del error α' correspondiente a otra secuencia de datos **x'(n)**, entonces el error δ que estaremos cometiendo será mayor que el mínimo, α:

$$A'(z) = \sum_{i=0}^{p} a'_i \cdot z^{-i} \qquad \text{con} \quad a'_i = 1 \tag{IV.24}$$

$$\delta = \sum_{n=-\infty}^{+\infty} \left[\sum_{i=0}^{p} a'_i \cdot x(n-i) \right]^2 \geq \alpha \tag{IV.25}$$

Ambos errores únicamente serán iguales, $\alpha=\delta$, cuando **A(z)=A'(z)**. La relación δ/α , de alguna forma, refleja la diferencia o distancia existente entre los espectros de ambas estimaciones, **x(n)** y **x'(n)**. Dicha relación será siempre mayor que la unidad, excepto en el caso en que **A(z)=A'(z)**, donde toma ese valor.

De forma análoga podemos definir también la relación δ'/α'. Tanto un caso como el otro son las denominadas relaciones de máxima verosimilitud, y sus logaritmos son lo que hemos llamado distancia LLR. La distancia de Itakura en el dominio espectral queda entonces definida por la siguiente expresión:

$$d_{Itak} = Ln(\delta/\alpha) = Ln\left[\frac{1}{2\pi}\int_{-\pi}^{+\pi}\left|\frac{A'(e^{j\epsilon})}{A(e^{j\epsilon})}\right|^2 d\theta\right] \qquad \text{(IV.26)}$$

IV.1.5.4. La Distancia Cosh.

Para evitar problemas de simetría, la distancia Cosh intenta hacer un promediado de las distancias δ/α y δ'/α', de forma que se define a partir del parámetro Ω, cuya expresión es la siguiente:

$$\Omega = \frac{\frac{\delta}{\alpha} + \frac{\delta'}{\alpha'}}{2} - 1 \qquad \text{(IV.27)}$$

Si definimos:

$$\cosh(\omega) - 1 = \Omega \qquad \text{(IV.28)}$$

entonces la distancia Cosh, ω, queda definida a partir de la expresión:

$$\omega = Ln\left\{1 + \Omega + [\Omega.(2 + \Omega)]^{1/2}\right\} \qquad \text{(IV.29)}$$

IV.1.5.5. La Distancia Cepstrum.

Supongamos que tenemos dos modelos espectrales, $\sigma^2/A(z)$ y $\sigma^2/A'(z)$. El error entre estos dos modelos se define como sigue:

$$V(\theta) = \mathrm{Ln}\left[\frac{\sigma^2}{\left|A(e^{j\theta})\right|^2}\right] - \mathrm{Ln}\left[\frac{\sigma^2}{\left|A'(e^{j\theta})\right|^2}\right] \qquad \text{(IV.30)}$$

Una posible forma de medir la distancia entre ambos modelos es elegir la norma L_p^p como:

$$L_p^p = (d_p)^p = \frac{1}{2\pi} \cdot \int_{-\pi}^{+\pi} |V(\theta)|^p d\theta \qquad \text{(IV.31)}$$

donde para **p=1** se define la Medida Espectral Logarítmica de Media Absoluta, y para **p=2** la Medida Espectral Logarítmica RMS. El problema de esta medida es de tipo computacional, puesto que requiere el uso de FFT's y logaritmos para tener una estimación de (IV.31) como un sumatorio. La distancia Cepstrum evita estos costosos cálculos.

Si desarrollamos **A(z)** en serie de Taylor se obtiene:

$$\mathrm{Ln}[A(z)] = -\sum_{k=1}^{\infty} c_k \cdot z^{-k} \qquad \text{(IV.32)}$$

donde los c_k se denominan coeficientes cepstrales. Teniendo en cuenta que el logaritmo de una magnitud al cuadrado es la parte real del logaritmo, podemos escribir la siguiente igualdad:

$$\mathrm{Ln}\left[\frac{\sigma^2}{A(e^{j\theta})}\right] = \sum_{-\infty}^{+\infty} c_k \cdot e^{-j\theta_k} \qquad \text{(IV.33)}$$

con $c_0 = \mathrm{Ln}(\sigma)$ y $c_{-k}=c_k$. Igualmente podemos escribirla para el otro modelo:

$$\mathrm{Ln}\left[\frac{\sigma^2}{A'(e^{j\theta})}\right] = \sum_{-\infty}^{+\infty} c'_k \cdot e^{-j\theta_k} \qquad \text{(IV.34)}$$

Aplicando entonces la relación de Parseval llegamos a la siguiente expresión:

$$d_2^2 = \sum_{-\infty}^{+\infty} (c_k - c'_k)^2 = (c_0 - c'_0)^2 + 2.\sum_{k=1}^{+\infty} (c_k - c'_k)^2 \qquad \text{(IV.35)}$$

Como la serie infinita tampoco es práctica, truncamos el sumatorio en un término **M** arbitrariamente grande. Tendremos entonces que la distancia Cepstral quedará definida por la expresión aproximada que escribimos a continuación:

$$d_2^2 = \sum_{-\infty}^{+\infty}(c_k - c'_k)^2 \approx (c_0 - c'_0)^2 + 2.\sum_{k=1}^{M}(c_k - c'_k)^2 \qquad \text{(IV.36)}$$

Para que la medida sea válida habrá que utilizar valores de M mayores o iguales que **p**, el orden del modelo (**M≥p**). Los coeficientes $\mathbf{c_0}$ y $\mathbf{c'_0}$ representan la energía de las señales. Para tener una idea de la exactitud de la aproximación del sumatorio, podemos decir que para el caso concreto de **M=p=10**, la correlación entre la serie truncada y la exacta es del orden de 0.98, valor que consideramos suficientemente cercano a la unidad (caso ideal).

IV.1.5.6. Parámetros de Evaluación de las Medidas.

Se han utilizado los siguientes factores para el cálculo de las diversas medidas que acabamos de exponer:

a) SNR_{global}: Para evaluarla se usa toda la señal, sin suprimir los intervalos de silencio.

b) SNR_{seg}: Se utiliza una longitud de trama de **N=128** muestras, que a una frecuencia de muestreo de 8KHz equivalen a 16mseg.

c) Distancia de Itakura: Para esta distancia se ha elegido un modelo de orden 10. No se han suprimido tampoco los silencios, ni se ha realizado preénfasis de la señal. Se ha utilizado una ventana de Hamming.

d) Distancia Cosh: Los parámetros de evaluación son idénticos a los utilizados para el cálculo de la distancia de Itakura.

e) Distancia Cepstrum: Utiliza también un modelo de orden 10 para el cálculo. Los coeficientes $\mathbf{c_0}$ y $\mathbf{c'_0}$ (energía de las señales) no se tienen en cuenta. Idéntica a Itakura para el resto de parámetros.

IV.2. Evaluación del Algoritmo Clásico de Segundo Orden (AR2).

A lo largo del apartado anterior, se ha presentado el algoritmo iterativo de Wiener. Este algoritmo se debe a Lim y Oppenheim [Lim-79] en lo referente a su estructura general, dada por las ecuaciones (II.30) y (II.31). El punto crucial de este algoritmo consiste en obtener una buena estimación de la densidad espectral $P_S(w)$ correspondiente a la señal de voz original. Dicha estimación debe realizarse a partir de la señal de voz ruidosa. Lim y Oppenheim aplicaron técnicas de Sustracción Espectral, descritas en el Apartado II.3.3., para obtener dicha estimación. Ante la presencia de bajos niveles de ruido de fondo, este sistema suele obtener buenas prestaciones. Pero, cuando el ruido alcanza valores medios de SNR este tipo de técnicas no permiten eliminarlo. En el Apartado II.2.3. se expuso una segunda estrategia consistente en estimar la señal de voz original S(w) a partir de un modelado paramétrico. El problema de esta técnica es que trata de aproximar el espectro DFT mediante su envolvente o espectro LPC y, en consecuencia, la calidad final no es muy buena. Además, la estimación de los parámetros AR de segundo orden se realiza a partir de la función autocorrelación y ésta es bastante sensible al ruido cuando éste presenta niveles notorios.

En este trabajo se considera una estrategia basada en la combinación de ambos métodos: se utiliza el Filtrado Iterativo de Wiener estimando la densidad espectral de la voz original $P_S(w)$ a partir de un modelo paramétrico AR. Se evaluan las prestaciones de este algoritmo iterativo de Wiener basado en un modelado AR, estimado mediante estadísticas de segundo orden. Este algoritmo lo vamos a denominar algoritmo AR2. Inicialmente, se testea este algoritmo AR2 frente a una fuente perturbadora consistente en ruido aditivo Gaussiano blanco (AWGN). A menos que se indique lo contrario, los resultados expuestos corresponden a un fichero de voz facilitado por la Asociación europea ESCA. Evidentemente, los resultados varían para cada fichero de voz, aunque el comportamiento de sus prestaciones suele ser similar para todos los ficheros testeados.

En la Fig.IV.8 se presentan los valores de distancia Cepstrum obtenidos mediante este algoritmo AR2. Se han considerado distintos niveles de ruido para observar la capacidad de dicho algoritmo para afrontar distintos entornos ruidosos. Parece lógico a priori, esperar que el algoritmo AR2 pierda su capacidad de eliminar el ruido existente cuando el nivel de dicho ruido supere un cierto umbral y, en consecuencia, el presente estudio pretende obtener el margen de niveles de ruido donde el algoritmo es útil. En el lado opuesto, se prevé un buen comportamiento al aplicarlo en ambientes poco ruidosos. En este caso, la mayor parte de algoritmos conducen a una buena supresión de ruido y el factor diferenciador entre éllos

puede venir dado por el nivel de distorsión que se ha ocasionado o por el tiempo de proceso necesario. Para simular el margen de distintos niveles de ruido, abarcando desde los entornos más ruidosos hasta los más silenciosos, se ha degradado un mismo fichero de señal de voz con distintos niveles de ruido desde una SNR global (SNR_G) de 0dB hasta unos 24dB. Así, en los siguientes gráficos se muestra la evolución de las distintas medidas, temporales y espectrales, en función de la SNR a la entrada del sistema.

La traducción de estos valores de SNR global a valores de distancia Cepstrum se ha representado en la linea superior de la Fig.IV.8. En cada una de las gráficas podemos ver cinco curvas; la primera, definida como *Sin procesar*, nos sirve de referencia, pues nos informa sobre la distancia existente entre la señal limpia original y la ruidosa de partida. El resto de curvas nos indica el valor alcanzado en las sucesivas iteraciones de filtrado.

Después del primer filtrado de la señal ruidosa, se observa una reducción bastante uniforme, entre 1.2dB y 1.6dB en los valores de distancia Cepstrum, a lo largo de los distintos niveles de ruido. Se aprecia ligeramente una mayor reducción de ruido en los ambientes menos ruidosos. En la segunda iteración el comportamiento es similar, aunque el nivel de ruido suprimido es ligeramente inferior, entre 1dB y 1.2dB, y la reducción es un poco superior para niveles altos de ruido. En la tercera iteración, se obtienen mejoras de 0.6 a 0.8dB para niveles de ruido medios y altos (SNR_G<17dB) mientras que para ambientes poco ruidosos esta mejora se reduce hasta unos 0.2dB para SNR_G= 24dB. Durante la cuarta

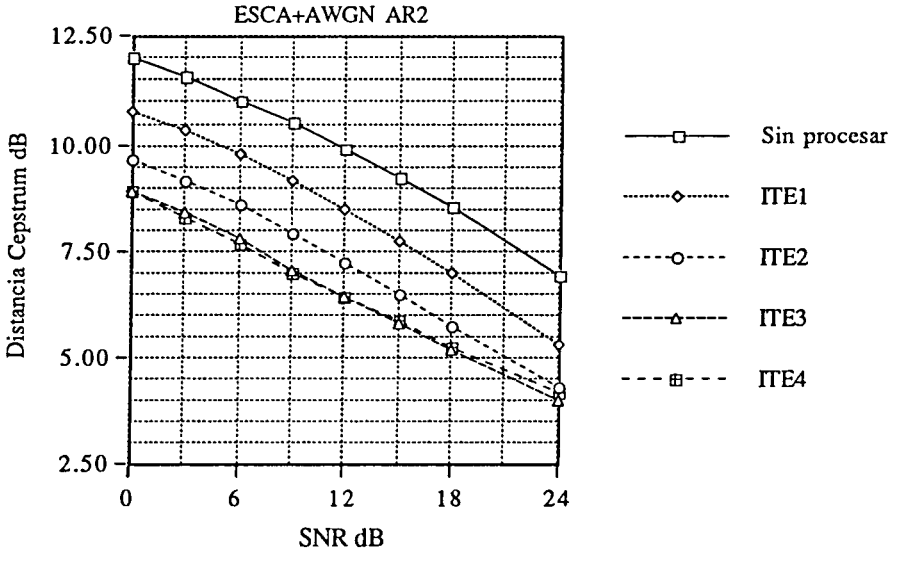


Figura IV.8 : Distancia Cepstrum obtenida por el algoritmo AR2 ante distintos niveles de ruido (señal ESCA+AWGN).

iteración, el algoritmo AR2 llega al equilibrio para valores medio-altos de SNR a la entrada. Esta zona de equilibrio se caracteriza porque la mayor parte de ruido ha sido suprimido durante las iteraciones precedentes y, de este modo, en la iteración actual se equilibran los efectos de reducción de ruido y distorsión ocasionada. Si bien para el margen de 0 a 9dB las mejoras son insignificantes, para SNR's mayores la distancia espectral empieza a empeorar, con la correspondiente carga de distorsión ocasionada en la señal. Llegados a este punto, una iteración más en ése margen induce al sistema a acusar los efectos nocivos de un sobrefiltrado propios del filtrado de Wiener con modelos AR (Capítulo IV.5).

En la Fig.IV.9 se ha representado como se reduce el nivel de ruido con el devenir de las iteraciones procesadas. Se han considerado tres casos representativos de los distintos ambientes reales con que podemos encontrarnos: entornos silenciosos (SNR_G=18dB), entornos con un nivel intermedio de ruido (SNR_G=9dB) y entornos altamente ruidosos (SNR_G=0dB). Esta distinción de los niveles de ruido, concretada en estos tres casos particulares, aparecerá otras veces a lo largo de los próximos capítulos. En el caso de este algoritmo AR2 no se aprecian diferencias significativas de comportamiento al variar el nivel de ruido presente en la voz. Durante las tres primeras iteraciones el efecto reducción de ruido predomina claramente mientras en la cuarta iteración la distorsión ocasionada enmascara claramente el nivel de ruido suprimido. Tras tres iteraciones el ruido eliminado se puede concretar cuantitativamente en una reducción de 3dB en la medida de distancia Cepstrum.

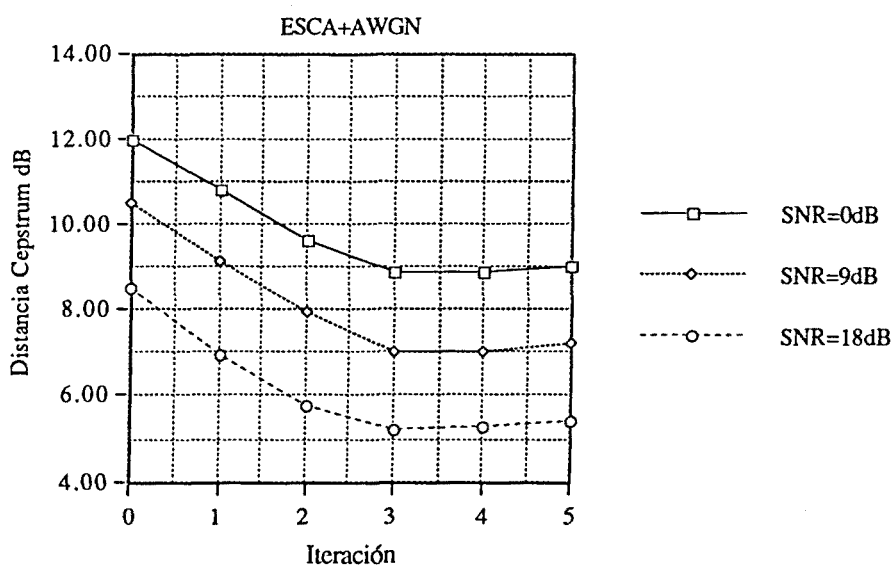


Figura IV.9 : Reducción de ruido en función del número de iteraciones procesadas y del nivel de ruido (señal ESCA+AWGN).

Esta mejora es ligeramente superior a medida que aumenta la SNR_G de la señal ruidosa x(n). Sin embargo, parte de esta distorsión ocasionada por las primeras iteraciones, visible en el dominio de la distancia Cepstrum, no es perceptible para el oido humano. De este modo las pruebas de audición demuestran una mejor calidad al procesar 4 iteraciones en relación a la señal estimada por la tercera iteración. Esta similar reducción de ruido para estos tres niveles de ruido demuestra la relativa incapacidad asociada con el algoritmo AR2 para atacar niveles de ruido altos o intermedios. Por otro lado, nótese la efectividad del algoritmo iterativo de Wiener en relación al Filtrado de Wiener clásico (primera iteración): las reducciones de ruido tras 3 ó 4 iteraciones son superiores al doble de la reducción obtenida tras la primera iteración.

En resumen, el método de correlaciones consigue una mejora gradual, iteración a iteración, hasta saturarse y empezar a empeorar. Tal saturación será más temprana cuanto mayor sea la SNR de la señal de entrada. Así, para relaciones señal a ruido medias y bajas (0-9dB), necesitamos 4 iteraciones para llegar al mínimo de distancia Cepstrum, mientras que para SNR's mayores serán suficientes 3. A partir de la iteración óptima, donde se alcanza el equilibrio entre estos efectos opuestos, el continuar procesando más iteraciones no implica un grave deterioro de la calidad, es decir, las medidas de distancia Cepstrum no aumentan

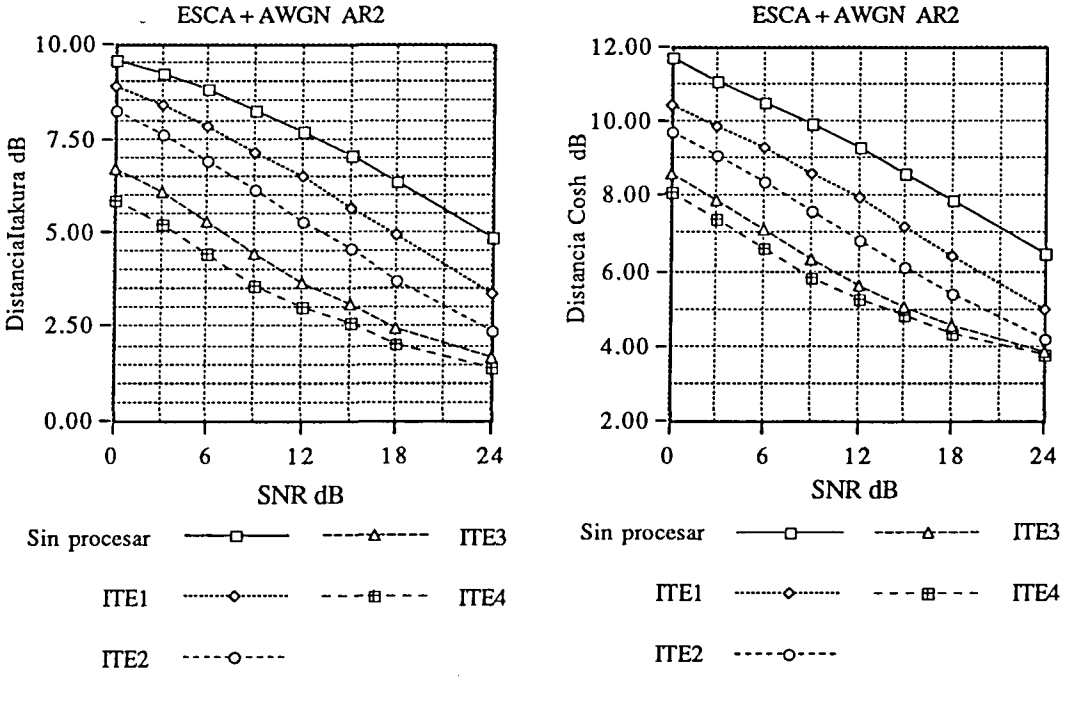


Figura IV.10 : Reducción de ruido, medida en términos de distancias Itakura y Cosh, en función del nivel de ruido (método AR2 sobre la señal ESCA+AWGN).

notoriamente. Por tanto, a pesar de sus limitaciones para discernir voz de ruido, el método de las correlaciones demuestra su bondad, no sólo en tiempo y sencillez, sinó también el la distorsión relativamente leve que introduce al utilizar un número de iteraciones más allá del óptimo.

Analizando el resto de distancias espectrales, Figura IV.11, podemos apreciar un comportamiento similar, si bien su saturación se produce después de un número superior de iteraciones. Es curioso observar cómo un incremento importante de mejora se produce en la tercera iteración (más que en la segunda), cuando la distancia Cepstrum ha empezado ya a saturarse. Recordemos que las distancias Cosh e Itakura nos informan, respectivamente, de la posición y forma de los formantes y menosprecian un poco lo que ocurre en los valles espectrales. Así, una posible explicación, confirmada por los tests de audición, es que el primer y segundo diseño del filtro eliminan sobretodo el ruido de fondo, para a partir de ahí afinar en la estimación de los formantes. Tal como veremos en el Capítulo IV.5, la distorsión producida por este algoritmo AR2 se concentra en los valles espectrales y mantiene intacta la zona correspondiente a los formantes. Si subimos en SNR el proceso de mejora está más repartido y el ajuste de los picos espectrales es más constante, ya que debe eliminarse un velo de ruido menor.

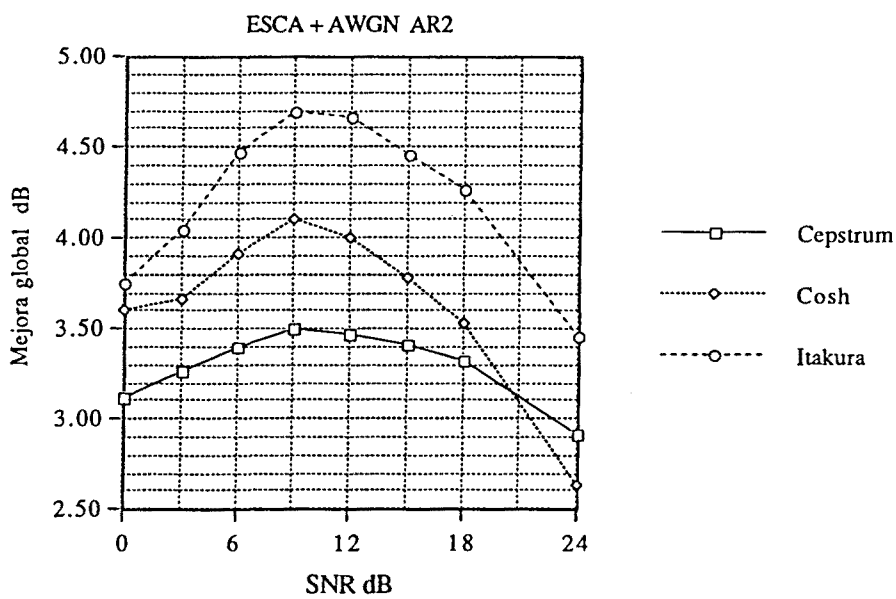


Figura IV.11 : Evolución de las mejoras máximas (iteración óptima) obtenidas por las distancias espectrales en función del nivel de ruido .

En la Fig.IV.11 se han representado las mejoras acumuladas tras procesar su número óptimo de iteraciones para cada una de las tres medidas espectrales consideradas y para cada nivel de ruido evaluado. Debe remarcarse que para un determinado nivel de ruido (SNR_G fijo), la iteración óptima correspondiente a una determinada medida de distancia espectral no tiene porque coincidir con la iteración óptima alcanzada por las otras dos medidas. El nivel de supresión de ruido se cuantifica de forma distinta segun cada medida específica de distancia. En concreto, la distancia de Itakura disminuye entre 3.7 y 4.7dB para el margen de 0 a 18dB de SNR de entrada, frente a los 3.5 - 4.1dB de la distancia Cosh y los 3.1 - 3.5dB de la Cepstrum. Las dos primeras, sobretodo la Itakura, centran su análisis en los formantes, así que mejoran en cada iteración sin verse afectadas por el efecto de picado que castiga y degrada los valles. La distancia Cepstrum, que pondera todo el espectro de forma más homogénea, sí se ve afectada y se satura un poco antes. En consecuencia, su mejora es menor que la obtenida mediante las otras distancias.

Parece lógico esperar una mejora más importante a medida que disminuye la SNR_G puesto que hay un superior nivel de ruido a eliminar. Este comportamiento se verifica durante el margen de valores de SNR_G comprendidos entre 9dB y 24dB. Para niveles de ruido superiores (SNR_G<9dB), las mejoras alcanzadas empeoran considerablemente a medida que aumenta el nivel de ruido. Este hecho muestra claramente la impotencia de este algoritmo AR2 para afrontar estos niveles de ruido, debido a su gran sensibilidad a la presencia de éste. El resultado final se traduce en una pobre supresión de ruido y una distorsión considerable, incluso en la posición y forma de los formantes.

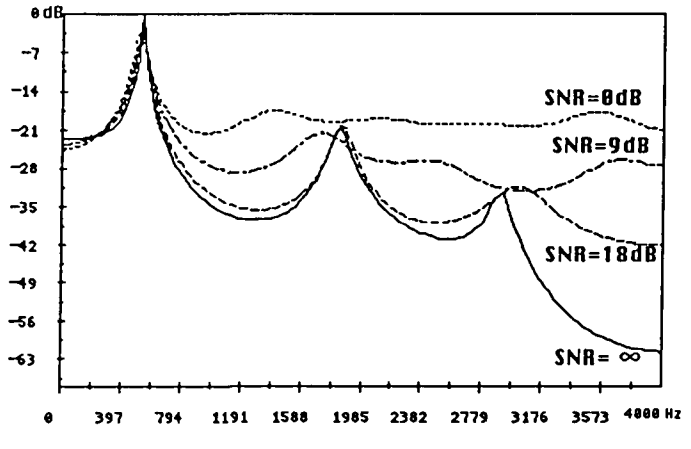


Figura IV.12 : Envolvente LPC, del modelado AR correspondiente a una /e/ real, obtenido por el algoritmo AR2 ante la presencia de distintos niveles de ruido.

Esta sensibilidad al ruido puede apreciarse en la Fig.IV.12, donde se ha representado el espectro LPC para un sonido /e/, extraido de un fichero de voz real (512 muestras). En linea continua se ha representado el espectro LPC en ausencia de ruido (SNR_G=∞dB). Los tres espectros restantes corresponden a los tres niveles representativos de ruido presentados anteriormente. Cuando el nivel de ruido es bajo (SNR_G=18dB) los dos primeros formantes se reproducen con bastante fidelidad, pero el tercer formante aparece ligeramente desplazado. El peor comportamiento se aprecia en la zona de alta frecuencia donde el algoritmo AR2 prácticamente no alcanza a suprimir el bajo nivel de ruido existente, debido a su vez a la baja energía que presenta la señal de voz original en esta zona. Ante la presencia de niveles de ruido superiores pierde, incluso, la información correspondiente a los formantes. Solamente logra rescatar la información de baja frecuencia alrededor del primer formante debido, en gran parte, a la alta energía espectral de la señal en esta zona frecuencial.

Analizando las medidas temporales correspondientes a la relación señal a ruido global y segmentada, SNR_G y SNR_S, se obtiene la evolución de extracción de ruido en el dominio temporal, obtenida por el algoritmo AR2 iteración a iteración (Fig.IV.13). Como ya deducíamos de las medidas espectrales, se aprecia una sustancial reducción del ruido inicial en la primera iteración (7.3dB de SNR_G para una SNR_G de entrada de 0dB), complementada con un incremento de 1.6 a 0.6dB más en la segunda. Ir más allá carece de sentido pues toda la mejora se relata en las distancias espectrales, mientras la relación señal a ruido se satura. La

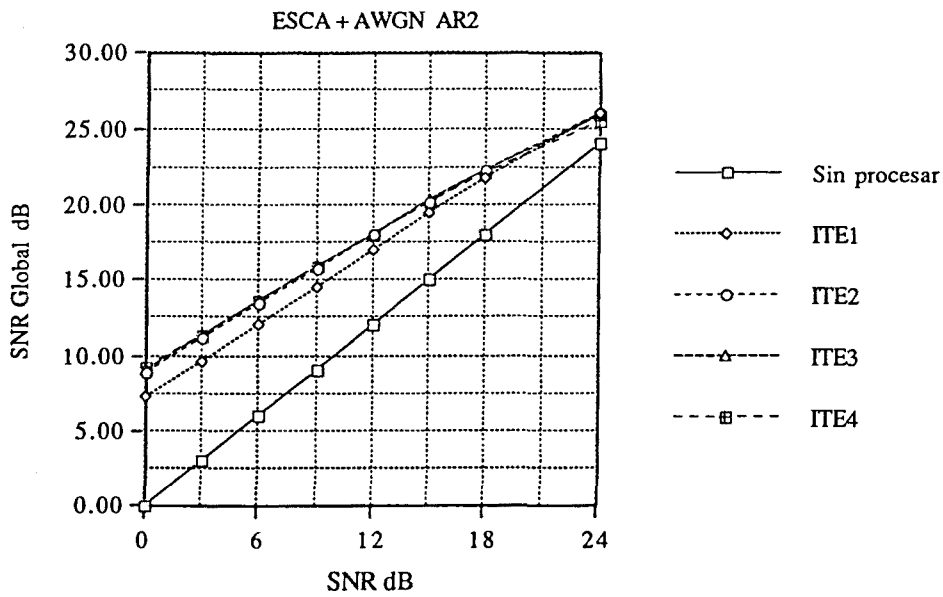


***Figura IV.13** : Relación señal a ruido global (SNR_G) obtenida con el método AR2 ante distintos niveles de ruido y bajo la consideración de distintas iteraciones.*

SNR_S mantiene el mismo comportamiento, aunque con valores inferiores dada la ponderación que ésta realiza entre las zonas de alta y baja energía de la señal.

Las pruebas de audición corroboran todo lo comentado anteriormente. Para SNR_G=0dB la mayor parte de la eliminación del ruido se produce en la primera iteración. En la segunda se continúa eliminando ruido, pero aparecen ya algunos espurios (ruido musical) bastante molestos, a la vez que empieza a notarse un aumento del dominio de los graves sobre los agudos. La inteligibilidad ha mejorado en cada iteración, pero a partir de aquí el efecto de la distorsión empieza a hacer mella en la señal. Aunque en la tercera iteración se aprecia cierta reducción del nivel de los espurios, a partir de ésta no se aprecia ninguna mejora significativa. y la señal empieza lentamente a degradarse.

Para una SNR_G de entrada de 9dB el proceso se condensa en menos pasos. Gracias a ello, y puesto que ahora la señal está menos enmascarada, se consigue mayor mejora con menos distorsión. Tras la tercera iteración podemos escuchar una señal bastante pura, aunque con una cierta presencia de espúreos o ruido musical residual. Finalmente, para SNR_G=18dB una sola estimación del filtro basta para eliminar el ruido existente y, puesto que los espúreos son prácticamente inexistentes, parece casi innecesario realizar una segunda iteración.

Podemos concluir, por tanto, que el método de las correlaciones será especialmente válido en situaciones donde el nivel del ruido que interfiera la señal sea bajo. En estos casos, además, el hecho de ir más allá del filtro aconsejable, aunque no mejore la señal, no introduce una distorsión excesiva. Sin embargo, para ambientes bastante ruidosos (SNR_G entre 0dB y 9dB), la dificultad del algoritmo de segundo orden para discriminar el ruido de la voz lo hacen lento e inseguro, y tanto la calidad como la inteligibilidad de la señal resultante se ven seriamente afectadas.

Los métodos alternativos propuestos en este trabajo pretenden precisamente cubrir ese campo de acción consiguiendo mejores prestaciones. Será por ese motivo por lo que a partir de ahora, en la mayoría de las ocasiones, centraremos nuestro estudio en los casos más adversos: relaciones señal a ruido intermedias o bajas. No obstante, también se evalua el comportamiento de estas nuevas estrategias en ambientes de trabajo poco ruidosos.

A continuación se presentan los resultados numéricos correspondientes a las cinco primeras iteraciones del algoritmo AR2 para estos tres niveles de ruido representativos, concretamente los casos 0dB, 9dB y 18dB correspondientes a SNR_G de entrada de la señal original de voz contenida en el fichero ESCA:

0dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	0.022	0.765	9.575	11.665	12.020
1 iter.	7.354	4.392	8.855	10.429	10.810
2 iter.	8.899	6.029	8.267	9.726	9.657
3 iter.	9.090	6.334	6.725	8.575	8.910
4 iter.	9.132	6.424	5.824	8.074	8.902
5 iter.	9.065	6.422	5.734	8.140	9.037

Tabla IV.1 : Evaluación del algoritmo AR2 en un entorno muy ruidoso (SNR_G=0dB).

9dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	9.021	8.073	8.276	9.923	10.510
1 iter.	14.464	10.126	7.154	8.611	9.172
2 iter.	15.742	11.471	6.127	7.575	7.934
3 iter.	15.912	11.680	4.422	6.302	7.054
4 iter.	15.862	11.694	3.578	5.814	7.007
5 iter.	14.338	11.323	3.529	5.816	7.197

Tabla IV.2 : Evaluación del algoritmo AR2 en un entorno ruidoso (SNR_G=9dB).

18dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	18.019	13.408	6.328	7.893	8.518
1 iter.	21.762	16.738	4.903	6.430	6.960
2 iter.	22.347	17.469	3.751	5.423	5.751
3 iter.	22.289	17.524	2.487	4.586	5.208
4 iter.	22.039	17.331	2.069	4.362	5.262
5 iter.	21.619	16.975	2.060	4.380	5.385

Tabla IV.3 : Evaluación del algoritmo AR2 en un entorno poco ruidoso (SNR_G=18dB).

IV.3. Evaluación del Algoritmo de Cumulantes de Tercer Orden (AR3).

En el apartado anterior se ha concluido que el algoritmo AR2 obtiene buenas prestaciones sólamente ante niveles bajos de ruido. A continuación se presenta el algoritmo AR3. La única diferencia, en relación al caso del algoritmo AR2, reside en la estimación de los parámetros **a** del modelado AR de la voz, puesto que dicha estimación se realiza en el dominio de los cumulantes de tercer orden. El cálculo de cumulantes se aplica a la señal de voz ruidosa disponible $\mathbf{x(n)}$ o a las sucesivas salidas del Filtro de Wiener $\mathbf{y_i(n)}$. Entonces, se hallan los coeficientes AR resolviendo las ecuaciones de Yule-Walker de tercer orden descritas en (III.96) para $\mathbf{n=3}$.

Vamos a ver ahora los avances que nos puede aportar la introducción, en el algoritmo de filtrado, de las estadísticas de orden superior (HOS). Recordemos que la principal característica de las HOS, que nos impulsaba a su utilización durante la estimación de los coeficientes AR, era su capacidad para discernir entre procesos gaussianos, o incluso no gaussianos con p.d.f. simétrica para orden tres, y procesos no gaussianos con p.d.f. asimétrica. Puesto que el ruido presente en muchos entornos reales puede modelarse como un proceso gaussiano y dado que la voz presenta una característica no gaussiana, sobretodo en sus tramas sonoras, las HOS nos proporcionan un cierto desacoplo voz-ruido verdaderamente útil para realizar una estimación de los coeficientes a_k más transparente al ruido. Hay que decir, no obstante, que ni la voz es totalmente no gaussiana ni el ruido real totalmente gaussiano Analizamos posteriormente lo que sucede con otra clase de ruidos (motor diesel, ruido de reactor...).

Análogamente al caso anterior, examinamos las posibilidades del algoritmo AR3 cuando se contamina la señal de voz original con AWGN, en el margen de 0 a 24dB de nivel de SNR_G. Veremos la evolución de las distintas distancias espectrales, así como de las medidas temporales.

El análisis en términos de de distancia Cepstrum, Figura IV.14, nos confirma plenamente el aumento de velocidad de convergencia del algoritmo iterativo asociado con los cumulantes de tercer orden. Tanto es así que en una sola iteración absorbemos e incluso sobrepasamos la mejora que consigue AR2 tras dos iteraciones: una reducción de 1.9dB a 3.3dB tras la primera iteración. Se pueden apreciar, también, los defectos propios de su alta agresividad frente al ruido. Mientras que para el margen $SNR_G \leq 9$dB, el sistema continúa mejorando apreciablemente (aún hay ruido que eliminar), para las SNR_G superiores el efecto es el contrario y se empieza a degradar la señal.

El punto de inflexión se podría situar en torno a SNR_G=12dB, donde las medidas de distancia Cepstrum correspondientes a las cuatro primeras iteraciones son prácticamente idénticas: en la primera iteración el efecto supresión de ruido domina claramente y obtiene una reducción de distancia Cepstrum de unos 3.3dB. Además, este valor representa la máxima reducción posible durante el procesado de la primera iteración y se puede interpretar como el nivel óptimo de ruido (SNR_G=12dB) para el cual este algoritmo AR3 resulta más efectivo: ante niveles inferiores de ruido (SNR_G>12dB) suprime el ruido durante una única iteración sin necesidad de emplearse a fondo; para niveles superiores de ruido (SNR_G≤9dB) sólamente logra eliminar parcialmente el ruido existente y, en consecuencia, muestra una menor efectividad a medida que decrece el valor de SNR_G. Sin embargo, fijémonos que en el peor caso, (SNR_G=0dB), alcanza una reducción de 2.1dB mientras que el algoritmo AR2 precisa 2 iteraciones para lograr reducciones similares.

Se debe constatar, también, que el uso de las estadísticas de tercer orden dotan de tal agresividad al método de Filtrado de Wiener que hacen innecesaria la posible consideración de su versión iterativa. Así, para el margen de niveles de ruido inferiores al correspondiente a SNR_G=15dB, se puede utilizar la técnica clásica de Filtrado de Wiener con estimación AR de tercer orden. Ello conlleva un ahorro considerable en el tiempo de cálculo debido a la ejecución de una única iteración por trama de voz. Si se continua iterando, a partir de la iteración óptima, aparece una apreciable distorsión, especialmente para los niveles de ruido

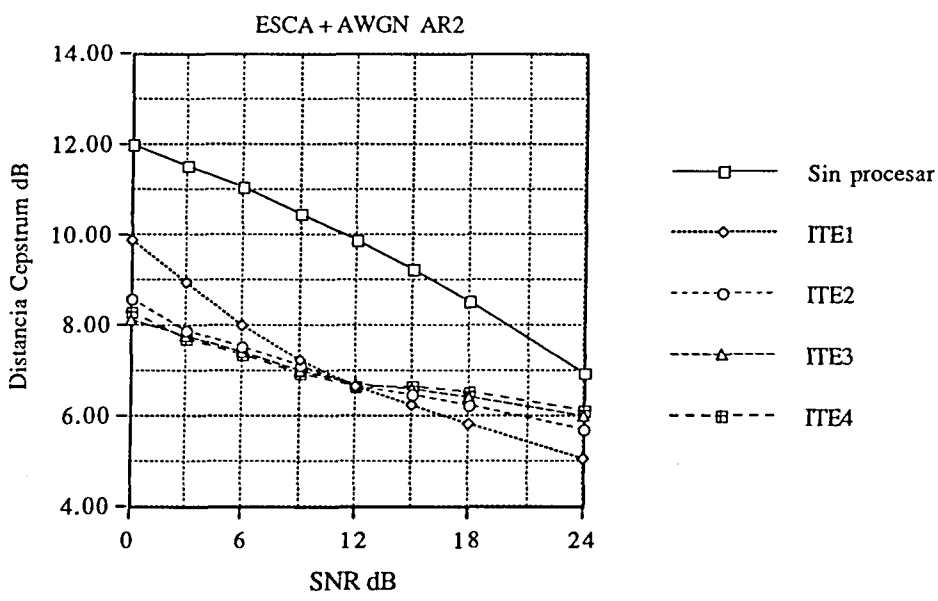


Figura IV.14 : Distancia Cepstrum obtenida por el algoritmo AR3 ante distintos niveles de ruido y segun el número de iteraciones procesadas.

más bajos.

Para niveles de ruido medio-altos debemos considerar el algoritmo iterativo de Wiener y procesar algunas iteraciones para reducir el ruido existente. Después de procesar unas tres iteraciones se llega al equilibrio, cuando el nivel de ruido suprimido es comparable al nivel de distorsión ocasionada. Los valores obtenidos superan apreciablemente los suministrados por el algoritmo AR2 y, además, se alcanzan tras procesar un menor número de iteraciones por trama de señal de voz.

Con las distancias Cosh e Itakura (Fig.IV.15), que centran su atención sobretodo en la posición y forma de los formantes de la voz, sucede exactamente lo mismo, si bien el umbral de inflexión se sitúa un poco más arriba, entre los 15dB y los 18dB de SNR_G de entrada. Nótese como la distorsión ocasionada por el hecho de sobrepasar la iteración óptima, para valores de SNR_G grandes, es poco notoria en términos de distancia Itakura o distancia Cosh.

En cualquier caso, si tenemos en cuenta lo anterior y analizamos las máximas mejoras introducidas por el método AR3 en todo el margen de SNR (Fig.IV.16), queda perfectamente claro que el filtrado mediante el uso de los cumulantes de tercer orden nos servirá, sobretodo, para casos con nivel medio-alto de ruido interferente, mientras que para SNR mayores se

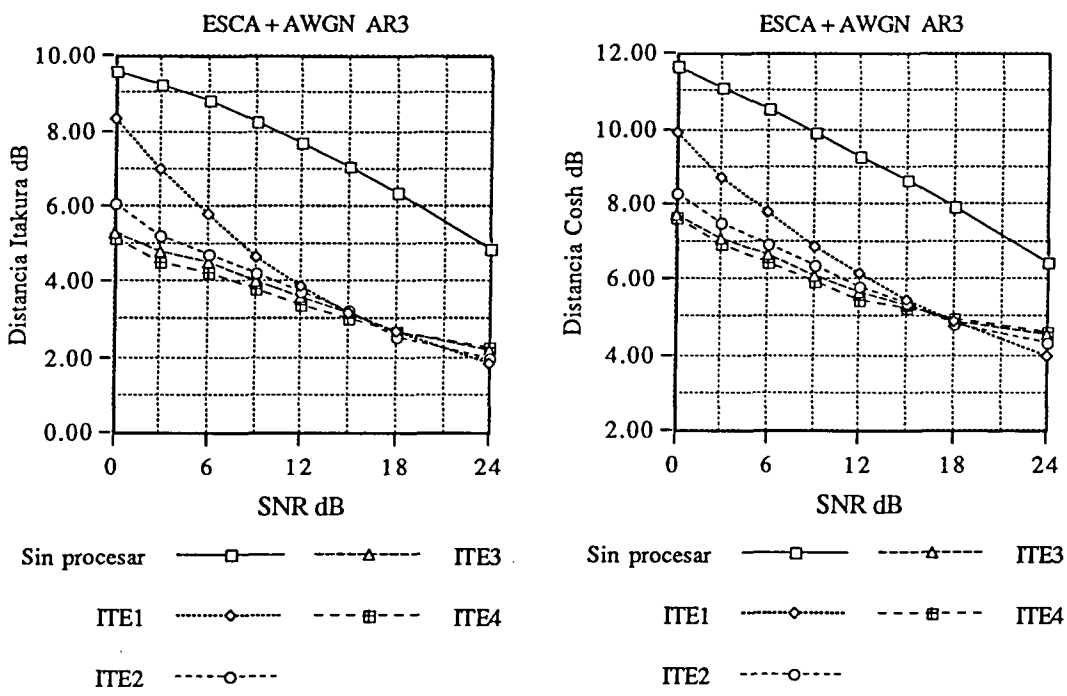


Figura IV.15 : Reducción de ruido, medida en términos de distancia Cosh e Itakura, en función del nivel de ruido para el algoritmo AR3.

obtienen buenos resultados para ambos algoritmos, AR2 y AR3. Si comparamos la gráfica IV.16 con su equivalente para el algoritmo AR2 (Fig.IV.11), se observa como el algoritmo AR3 logra peores reducciones en el margen de SNR_G altos. No obstante, este algoritmo AR3 alcanza la iteración óptima tras procesar únicamente la primera iteración y, en cambio, el algoritmo AR2 precisa como mínimo tres iteraciones para llegar a la óptima. Se puede interpretar que la menor sensibilidad al ruido conduce a la supresión de la mayor parte de ruido durante la primera iteración, y al procesar la segunda iteración, su mayor agresividad provoca un efecto distorsión que enmascara totalmente la supresión del ruido sobrante de la primera iteración. Es decir, se intuye que el algoritmo AR3 ocasiona una distorsión espectral mucho más significativa.

Evaluando en términos de distancia Cepstrum (Fig.IV.14), es fácil apreciar cómo en el margen inferior de SNR_G de entrada, tras una primera iteración muy agresiva, el sistema logra reducir entre 1.3 y 0.5dB más en la segunda iteración. En un tercer filtrado la mejora es ya bastante insignificante, solamente entre 0.1 y 0.4dB, saturándose a partir de ahí de modo que iterar más es perjudicial para la señal. Así, este algoritmo AR3 presenta una gran capacidad para afrontar niveles altos de ruido. En la Fig.IV.16, prácticamente se verifica la lógica relativa a alcanzar una mayor reducción de ruido cuando existe un nivel de ruido más elevado. Es decir, las mejoras obtenidas verifican una curva monótona decreciente al aumentar la SNR_G. Sólamente empieza a saturar para el margen $SNR_G \leq 3dB$. Recuérdase que

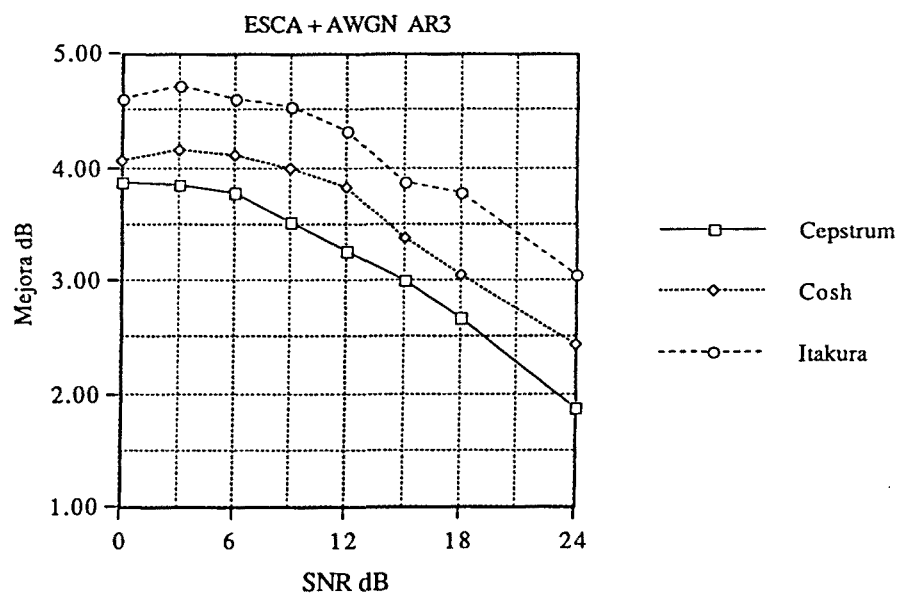


Figura IV.16 : Evolución de las mejoras máximas (iteración óptima) obtenidas por las distancias espectrales en función del nivel de ruido.

el algoritmo AR2 satura a partir de valores de SNR_G entre 9dB y 12dB, y se muestra ineficaz ante niveles altos de ruido (Fig.IV.11).

Las distancias Cosh e Itakura tienen el mismo comportamiento: la máxima reducción se concentra en las dos primeras iteraciones, obteniendo mejoras de hasta 3.5dB y 4.0dB respectivamente. Durante las dos iteraciones siguientes, si $SNR_G \leq 12dB$, la mejora es mínima (0.1 - 0.3dB), y para SNR_G superiores a 15dB una sola iteración es suficiente para alcanzar el mínimo.

Esta menor sensibilidad al ruido puede interpretarse como una evolución condensada de varias iteraciones AR2 en una sola iteración AR3. Así, aún en los casos más adversos, bastan 3 iteraciones para obtener mínimas distancias espectrales, inferiores además a las que obteníamos con AR2 procesando un mayor número de iteraciones. Sin embargo, como ya hemos mencionado, relaciones señal a ruido por encima de 12dB acusan la agresividad del método tras una sola estimación del filtro, pasando seguidamente a percibir la carga de distorsión que el uso de las HOS conlleva.

En la Fig.IV.17 se ha representado el espectro LPC correspondiente al sonido sonoro /a/ generado sintéticamente segun el modelo de la Fig.I.1. La estimación de sus parámetros AR se ha realizado a partir de una longitud de trama de 1024 muestras, aplicando las estadísticas de segundo orden (algoritmo de Levinson-Durbin) y las estadísticas de tercer orden (ecuaciones de Yule-Walker de tercer orden). En trazo continuo se han representado los espectros LPC correspondientes a ambas estimaciones AR cuando no se considera la presencia de ruido (condiciones ideales de laboratorio). Obsérvese como no coinciden la estimación del modelo AR en el dominio de la función autocorrelación **(a)** con la realizada en el dominio de los cumulantes de tercer orden **(b)**.

Se ha procedido a degradar esta señal sintética mediante los tres niveles de ruido AWGN utilizados con anterioridad. En el dominio de la función autocorrelación se aprecia una cierta incapacidad para eliminar el ruido de alta frecuencia cuando el nivel de ruido es bajo $SNR_G=18dB$. Así, prácticamente se pierde el cuarto formante, enmascarado por el ruido, aunque el tercer formante todavía se distingue. Al situarnos en un entorno con un nivel intermedio de ruido, la mitad superior del espectro LPC no puede ser recuperado fielmente, y los valles entre los dos primeros formantes presentan algo de ruido. Al considerar un nivel de ruido superior ($SNR_G=0dB$), sólamente se recupera el margen frecuencial correspondiente al primer formante, mientras que el resto del espectro resulta enmascarado por la abundante presencia de ruido.

Si nos situamos en el dominio de los cumulantes de tercer orden **(b)**, se aprecia una parte considerablemente menor de ruido existente durante el proceso de estimación del modelo AR. Esto es especialmente cierto para un nivel medio-bajo de ruido, aunque en el caso SNR_G=9dB se pierde la información correspondiente a los dos formantes superiores. Para SNR_G=0dB, la parte correspondiente a los dos primeros formantes se degrada en menor medida al compararlo con el caso de segundo orden. Además, el nivel de ruido que se observa en la mitad superior del espectro es considerablemente menor. Así, la menor sensibilidad al ruido, propia de los cumulantes de tercer orden, origina una menor presencia de ruido en el espectro LPC, especialmente en sus valles espectrales.

En referencia a las medidas temporales hay que decir que tanto la SNR_G como la SNR_S también saturan rápidamente su mejora tras tan sólo una o dos iteraciones del algoritmo. Es

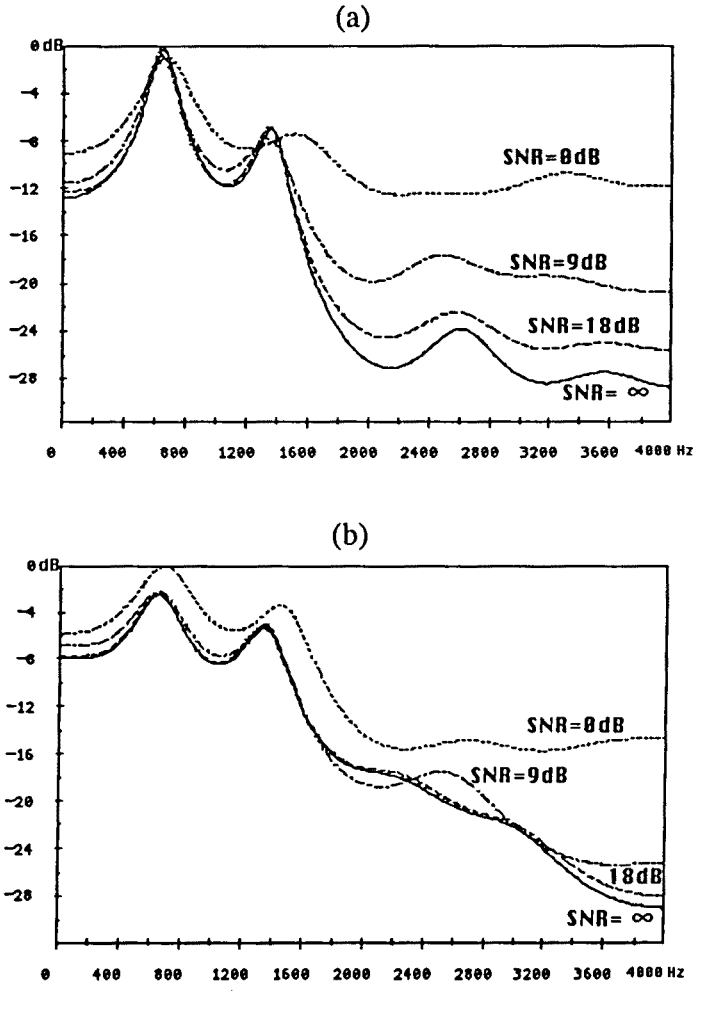


Figura IV.17 : Espectros LPC correspondientes al modelado AR de una vocal /a/ sintética mediante: (a) Método de correlaciones (2º orden); (b) Método de cumulantes de tercer orden.

curioso observar cómo en este tipo de medidas no logramos superar en ningún caso los resultados obtenidos con el método de correlaciones, que extrae el ruido de una forma más pausada pero segura.

En las pruebas de audición queda perfectamente comprobado el buen comportamiento del algoritmo en ambientes adversos. Así, para SNR=0dB, en una sola pasada se reduce el ruido a una fracción remanente muy pequeña, en comparación con AR2, que se elimina en un segundo paso. En este momento aparece el ruido musical de fondo para las tramas de baja energía, igual que sucedía con el método de las correlaciones. Un filtrado adicional nos proporciona, a pesar de estar a 0dB, una reducción del ruido musical a no más de 3 ó 4 espurios cuyo nivel tiende a desaparecer (recordemos que con AR2 el ruido de fondo se eliminaba progresivamente, pero no los espurios musicales que permanecían en un nivel no muy elevado, pero sí bastante molesto). Hemos ganado en calidad e inteligibilidad en las dos o tres primeras iteraciones. Sin embargo, la velocidad de operación se empieza a notar al apreciar una voz más apagada, una sensación subjetiva que responde a la distorsión de la banda alta del espectro inducida por el dominio de los dos primeros formantes. En la cuarta iteración tenemos una señal sin apenas ningún tipo de ruido, pero con una fuerte pérdida de calidad y resulta una voz bastante más ronca.

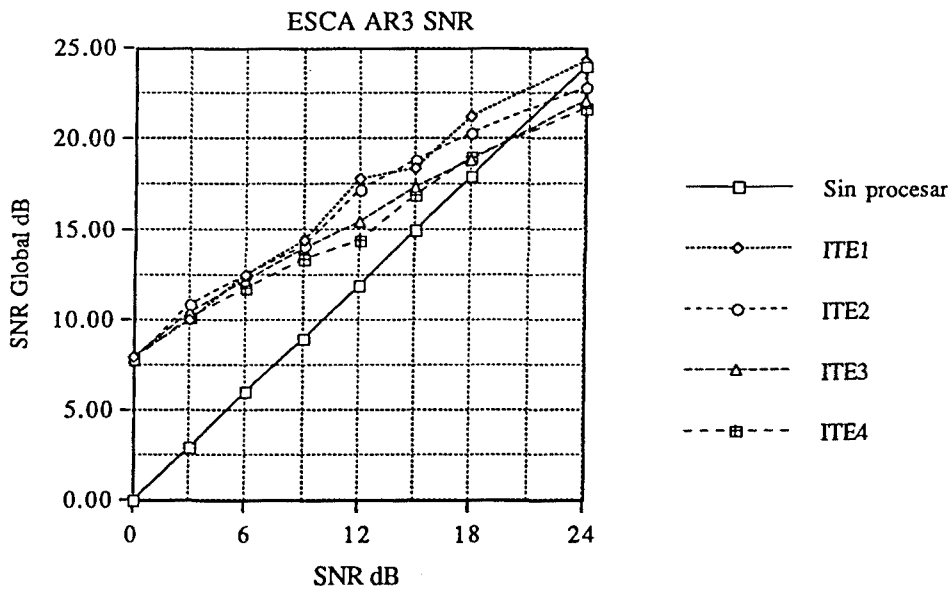


Figura IV.18 : Relación señal a ruido global (SNR_G) obtenida con el método AR3 ante distintos niveles de ruido y bajo la consideración de distintas iteraciones.

Se observa cómo para SNR_G=9dB el proceso se repite; primero eliminamos casi todo el ruido de fondo e introducimos tonos espurios, que la segunda iteración elimina o logra rebajar de nivel. En la tercera pasada obtenemos una señal bastante buena, pero a partir de aquí la calidad e inteligibilidad han llegado al óptimo y tienden a degradarse. Finalmente, para SNR_G=18dB, tal como queda patente en las medidas espectrales, dos iteraciones son más que suficientes y el continuar filtrando aporta principalmente distorsión en la señal.

A continuación se presentan los resultados numéricos correspondientes a las distintas medidas para los casos SNR_G=0dB, 9dB y 18dB, que consideramos representativos de los casos de mucho, medio y poco nivel de ruido interferente.

0dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	0.022	0.765	9.575	11.665	12.020
1 iter.	7.922	4.855	8.288	9.871	9.911
2 iter.	7.894	5.429	6.028	8.284	8.578
3 iter.	7.965	5.741	5.308	7.683	8.154
4 iter.	7.842	5.920	5.113	7.632	8.289
5 iter.	7.815	6.074	4.964	7.603	8.293

Tabla IV.4 : Evaluación del algoritmo AR3 en un entorno muy ruidoso (SNR_G=0dB).

9dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	9.021	8.073	8.276	9.223	10.510
1 iter.	14.497	10.770	4.636	6.847	7.230
2 iter.	14.168	10.898	4.261	6.391	7.150
3 iter.	14.066	10706	3.990	6.068	7.012
4 iter.	13.464	10.627	3.766	5.924	6.980
5 iter.	13.040	10.653	3.735	5.959	7.002

Tabla IV.5 : Evaluación del algoritmo AR3 en un entorno ruidoso (SNR_G=9dB).

18dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	18.019	13.408	6.328	7.893	8.518
1 iter.	21.177	16.784	2.682	4.895	5.845
2 iter.	20.258	16.115	2.549	4.835	6.216
3 iter.	18.779	15.400	2.654	4.925	6.422
4 iter.	19.037	15.372	2.670	4.961	7.556
5 iter.	18.861	15.202	2.762	5.031	6.523

Tabla IV.6 : Evaluación del algoritmo AR3 en un entorno poco ruidoso (SNR_G=18dB).

IV.4. Evaluación del Algoritmo de Cumulantes de Cuarto Orden (AR4).

En este apartado se presenta el algoritmo iterativo de Wiener cuando se efectúa la estimación AR en el dominio de los cumulantes de cuarto orden. Tras el cálculo de los cumulantes de cuarto orden, a partir de la señal disponible correspondiente a cada trama de voz ruidosa, se estiman los coeficientes AR mediante la resolución de las ecuaciones de Yule-Walker de cuarto orden descritas en (III.96) para **n=4**. Nótese como uno de los grados de libertad se fija a cero para los cumulantes de cuarto orden. La consideración de esta circunstancia reduce el coste de cálculo asociado con dichos cumulantes de cuarto orden. No obstante el coste de cálculo se incrementa bastante en relación al caso de los cumulantes de tercer orden y, en consecuencia, cada iteración del algoritmo AR4 resulta mucho más costosa.

Las propiedades de interés de estos cumulantes de cuarto orden hacen referencia a su valor nulo cuando se aplican sobre procesos Gaussianos. A diferencia de los cumulantes de tercer orden, no se anulan ante procesos no Gaussianos cuya función densidad de probabilidad sea simétrica. Esta última característica hace vislumbrar una menor distorsión asociada con su uso. Así, en principio, se pensó que su uso podía comportar una rápida convergencia del algoritmo, típica del uso de estadísticas de orden superior, combinada con una menor distorsión en relación al caso del algoritmo AR3. Es decir, el análisis realizado en el dominio de los cumulantes de cuarto orden preserva las componentes simétricas de la señal de voz y sigue manteniendo un cierto desacoplo voz-ruido.

A continuación se analiza qué nos puede aportar este método. Sabemos que ni el ruido es totalmente Gaussiano o no Gaussiano con p.d.f. simétrica, ni la voz totalmente no gaussiana con p.d.f. no simétrica. Ello incide directamente en la capacidad del método AR3 de diferenciar entre una cosa y otra, afectando directamente también sobre la carga de distorsión que produzca. Tanto es así que para determinados locutores los resultados obtenidos con AR3 son ciertamente inverosímiles, en el sentido de que ni su comportamiento ni sus posibilidades están a la altura de los conseguidos con otros. Se observa que los cumulantes de tercer orden obtenidos en estos casos (voz con fuerte componente simétrica) toman valores muy bajos, síntoma de que se está eliminando una parte importante de la señal. Cuanto más simétrica sea la p.d.f. de la señal que se genera, más confundirá el método de tercer orden a voz y a ruido, más distorsión y peores resultados se obtendrán. En esta situación parece lógico el uso de los cumulantes de cuarto orden, aunque comporten un mayor coste operacional.

Tal como hicimos con los algoritmos AR2 y AR3, vamos a estudiar los resultados obtenidos al procesar la señal ESCA degradada aditivamente por AGWN en un margen de SNR_G de entrada comprendido entre 0 y 24dB. En la Fig.IV.19 se puede observar la evolución de la distancia Cepstrum durante las primeras cuatro iteraciones del algoritmo AR4. Aunque en líneas generales presenta un comportamiento similar al de AR3, en la zona de 0 a 3dB se aprecian diferencias evidentes.

Mientras para el método de tercer orden 3 iteraciones son suficientes para alcanzar su mínimo con una mejora total de 3.9dB, los cumulantes de cuarto orden continúan mejorando todavía hasta la sexta iteración para una mejora acumulada de 3.7dB. Queda entonces patente la naturaleza de este nuevo método ante señales de baja calidad: con unos resultados numéricamente similares, aunque algo inferiores a los obtenidos por AR3, se trata de un procedimiento más lento, menos agresivo, y por ello posiblemente menos distorsionante.

Ello tiene su efecto positivo en la zona de SNR's medias y altas. Tal como sucedía en el caso del algoritmo AR3, a partir de SNR_G=14dB es prácticamente suficiente una iteración de filtrado para llegar al mínimo, aunque ahora éste es ligeramente menor (de 0.1dB para SNR_G=15dB, a 0.8dB para SNR_G=24dB). Nótese como su mayor lentitud traslada el punto de inflexión hacia menores niveles de ruido (SNR_G=14dB). En este punto las cuatro primeras iteraciones conducen a valores idénticos de distancia Cepstrum, puesto que durante la primera

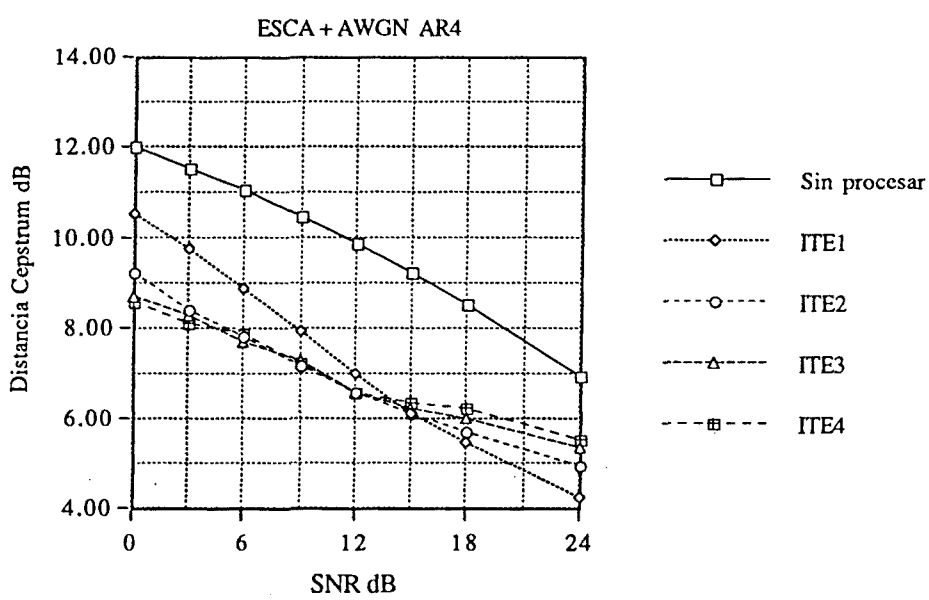


Figura IV.19 : Distancia Cepstrum proporcionada por el algoritmo AR4 ante distintos niveles de ruido y segun el número de iteraciones procesadas

iteración se elimina la mayor parte de ruido, mientras en las tres siguientes se equiparan los efectos supresión de ruido y distorsión ocasionada.

En la zona correspondiente a un bajo nivel de ruido ($SNR_G \geq 14dB$) se puede considerar el algoritmo clásico de Wiener, ya que no se precisa el refinamiento originado por una segunda iteración. Además, los valores obtenidos por el algoritmo AR4 son algo mayores, en comparación al caso AR3, debido principalmente a la menor distorsión ocasionada cuando actúa ante niveles de ruido donde no precisa emplearse a fondo: a partir de $SNR_G=20dB$ el algoritmo AR4 conduce a valores de distancia Cepstrum inferiores en más de 0.5dB en relación a los valores obtenidos mediante el algoritmo AR3. Como conclusión se puede afirmar que por encima de este punto de inflexión ($SNR_G=14dB$), el algoritmo AR4 conduce a reducciones de ruido con una menor distorsión y, en consecuencia, supera claramente al algoritmo AR3. En esta zona mantiene la agresividad asociada con las HOS y sólamente precisa operar la primera iteración. Para niveles superiores de ruido ($SNR_G<14dB$) su mayor lentitud provoca la ejecución de un mayor número de iteraciones y, además, conduce a peores valores mínimos en relación a AR3, ya que alcanza sus valores mínimos con un número superior de iteraciones y, entonces, se le añade la distorsión propia del algoritmo iterativo de Wiener (efecto de picado espectral en la región de los formantes), bastante perceptible a partir de la segunda iteración.

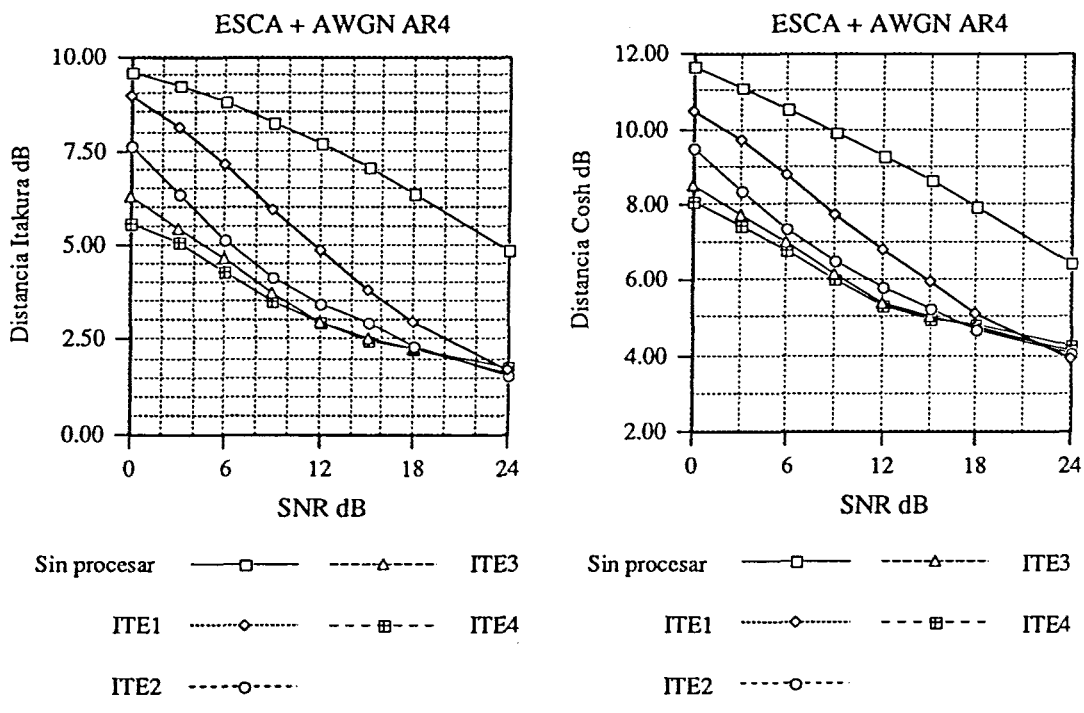


***Figura IV.20** : Reducción de ruido, medida en términos de distancia Itakura y Cosh en función del nivel de ruido para el algoritmo AR4.*

Para las distancias Cosh e Itakura (Fig.IV.20) se aprecia un comportamiento parecido al de AR3, aunque acusando la menor agresividad propia de este método: se obtienen peores resultados para SNR<12dB y valores ligeramente mejores para SNR_G superiores. El menor nivel de distorsión se aprecia en unos menores valores de distancia cuando el nivel de ruido es bajo y, además, al sobrepasar la iteración óptima la distorsión añadida no es tan notoria como en el caso AR3. Para niveles altos de ruido también se aprecia una mayor lentitud.

Lo comentado anteriormente puede observarse en la Fig.IV.21, donde las máximas mejoras globales para cada distancia han sido representadas. La condición de método intermedio entre AR2 y AR3 queda de manifiesto en su evolución. Teniendo nuevamente las curvas una forma parecida a las obtenidas con AR3, las medidas en el margen 0-9dB se sitúan por debajo de éstas, pero por encima de las que obteníamos con AR2. A partir de los 12dB los papeles de AR2 y AR3 se invierten, quedando nuevamente AR4 en una situación intermedia (también en cuanto al número de iteraciones necesarias). En esta zona el algoritmo AR4 ofrece mejores prestaciones que el algoritmo AR3. El algoritmo AR2 puede originar mejores valores de distancia Cepstrum, pero a costa de procesar un número de iteraciones superior. La distancia Cepstrum presenta un comportamiento similar al observado para el caso AR3: mayor supresión de ruido al decrecer SNR_G. Sin embargo, las distancias Itakura y Cosh saturan sus mejoras a partir de SNR_G=9dB.

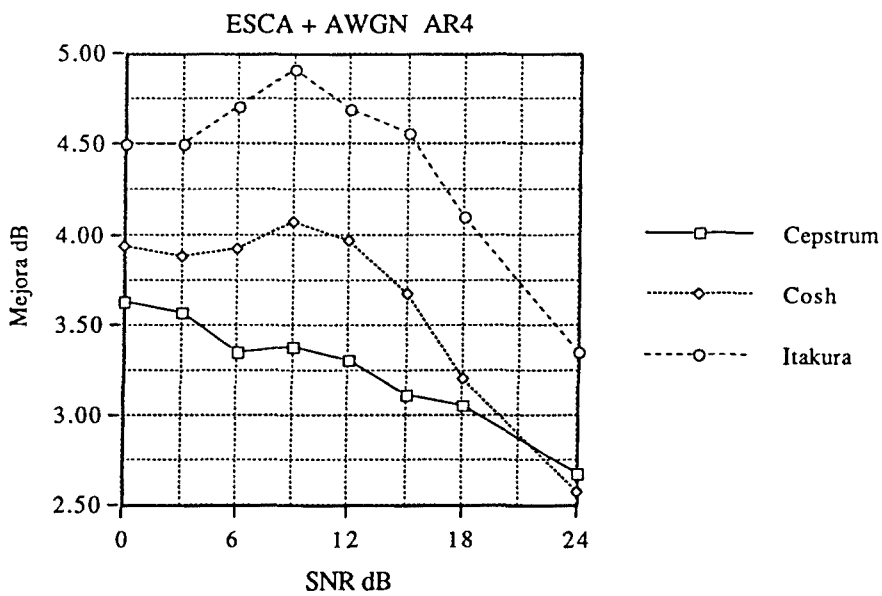


Figura IV.21 : Evolución de las mejoras máximas (iteración óptima) obtenidas por las distancias espectrales en función del nivel de ruido.

En la Fig.IV.22 se ha representado el espectro LPC correspondiente al sonido sintético sonoro /a/ para el supuesto de ausencia de ruido y para cuando se considera la presencia de tres niveles distintos de ruido, de forma similar a la Fig.IV.17 donde se muestran el comportamiento de los métodos de segundo y tercer orden. Para niveles bajos de ruido (SNR_G=18dB), el espectro LPC se ajusta de forma bastante precisa al espectro LPC de la señal original y presenta un comportamiento similar al caso de las estadísticas de tercer orden, siendo ambas técnicas bastante menos sensibles al ruido en relación al algoritmo basado en las estadísticas de segundo orden., especialmente en la zona de alta frecuencia. Incluso se puede apreciar ligeramente un mejor comportamiento de los cumulantes de cuarto orden en los alrededores del tercer formante al compararlo con el caso de los cumulantes de orden 3.

En cambio, para niveles intermedios de ruido los cumulantes de cuarto orden presentan una mayor sensibilidad al ruido, aunque bastante menor a la estimación obtenida a partir de la función autocorrelación. En entornos muy ruidosos presenta un comportamiento claramente inferior al método de cumulantes de tercer orden y tiende al comportamiento, bastante sensible al ruido, propio de las estadísticas de segundo orden, especialmente en la zona de alta frecuencia.

Respecto a las medidas temporales (Fig.IV.23) hay que decir que si nos fijamos únicamente en la máxima cota alcanzada en cada caso, AR4 es seguramente el que peor resulta. Recordemos, sin embargo, que AR3 degradaba rápidamente ese valor a causa de la distorsión que introducía. Los cumulantes de cuarto orden, por contra, mantienen, si no mejoran, un nivel de SNR de salida bastante estable a lo largo de las sucesivas iteraciones del algoritmo.

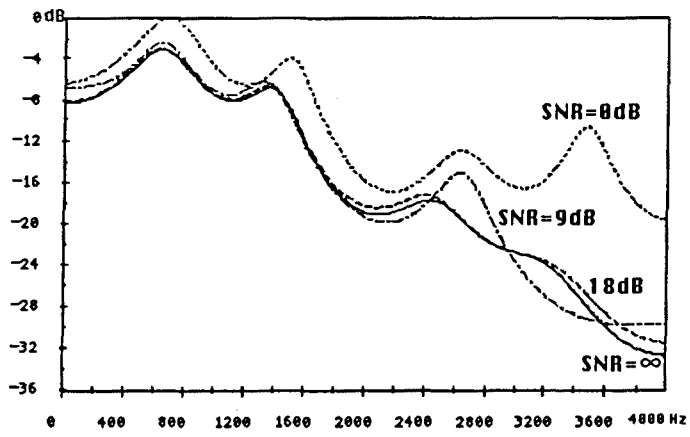


Figura IV.22 : Espectro LPC correspondiente al modelado AR obtenido a partir de los cumulantes de cuarto orden para una vocal /a/ generada sintéticamente.

Si escuchamos la señal a la salida del filtro a 0dB, tras una iteración obtenemos una reducción del ruido parecida a AR3; sin embargo, el ruido musical que aparece es más persistente. Tras una cuarta iteración nos queda algún espurio, pero la inteligibilidad es ligeramente superior a AR3. Para SNR=9 y 18dB el proceso se condensa , dando lugar a una señal de calidad muy similar a la del algoritmo de tercer orden.

La mejora no se refleja en una reducción drástica del ruido, quizás existan más espurios tras la última iteración con AR4 que con AR3; sinó en una disminución progresiva del ruido, al estilo del método de correlaciones, acompañada de un filtrado suave que logra, al tiempo que elimina la perturbación, no distorsionar la voz.

Con AR4 no aparecen los efectos derivados de un sobrefiltrado en la banda superior del espectro, tales como nos deparaba AR3 con una reducción del nivel de los formantes superiores, una voz más ronca. Tampoco, aún llegando a la cuarta iteración, a 0dB, se pierde la identificación del locutor, la distorsión sobre el primer y segundo formante parece menor que AR3.

En conclusión, si comparamos con AR3, queda patente la influencia de las componentes simétricas de la señal, sean de voz o de ruido, a la hora de discernir entre ambas cosas. Con los cumulantes de tercer orden eran todas consideradas como ruido, y en

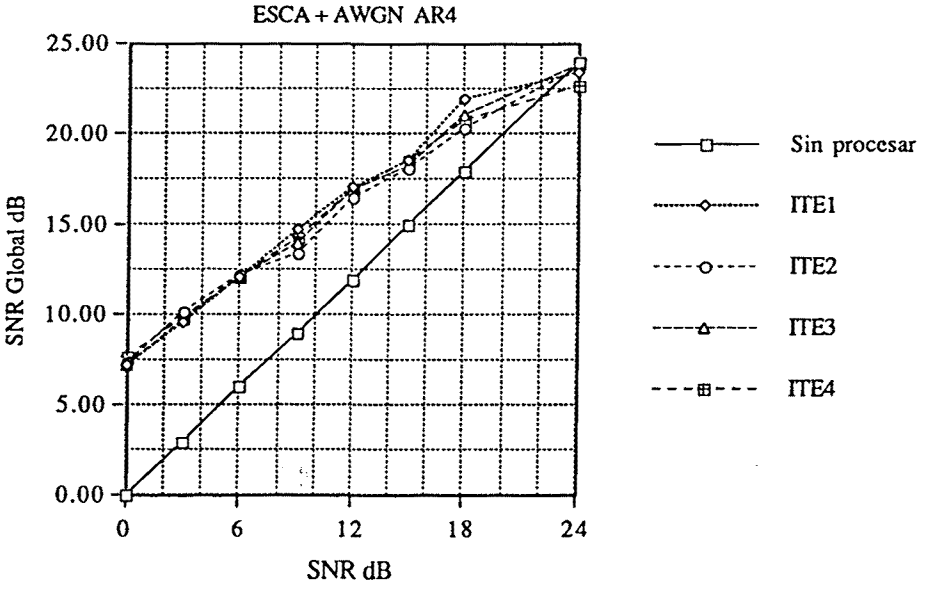


Figura IV.23 : Valores de SNR_G obtenidos por el algoritmo AR4 ante distintos niveles de ruido y bajo la consideración de distintas iteraciones.

consecuencia resultaban eliminadas. Se conseguía así mayor agresividad, que resultaba provechosa en los casos más adversos, pero también pagando con mayor distorsión, que se imponía como dominante en señales de mayor calidad.

Con AR4 las componentes simétricas son preservadas, y la diferenciación entre voz y ruido se sustenta únicamente en la gaussianidad o no gaussianidad de la señal. Resulta un método menos agresivo pero menos distorsionante, que continúa siendo mejor que AR2 para SNR's bajas. Se trata, en definitiva, de una buena alternativa para aquellos casos en que la señal de voz tenga una fuerte componente simétrica, y el procedimiento de tercer orden sea excesivamente destructivo.

No hay que olvidar, por otro lado, el incremento de carga computacional que AR4 supone, pues además de tardar el doble que AR3 para el cálculo de sus cumulantes, necesita mayor número de iteraciones. Es este un punto que nos puede llevar incluso a descartar el método si queremos que el sistema funcione en tiempo real.

En las siguientes tablas vemos los resultados numéricos de las medidas sobre ESCA para los casos representativos de calidad baja, media y alta de la señal, es decir, para SNR_G=0dB, 9dB y 18dB.

0dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	0.022	0.765	9.575	11.665	12.020
1 iter.	7.313	4.301	8.959	10.489	10.556
2 iter.	7.447	4.864	7.637	9.490	9.249
3 iter.	7.234	4.980	6.301	8.491	8.722
4 iter.	7.735	5.455	5.572	8.039	8.602
5 iter.	7.892	5.517	5.064	7.723	8.389

Tabla IV.7 : Evaluación del algoritmo AR4 en un entorno muy ruidoso (SNR_G=0dB).

9dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	9.021	8.073	8.276	9.923	10.510
1 iter.	14.730	10.376	5.916	7.710	7.951
2 iter.	13.464	10.515	4.171	6.519	7.189
3 iter.	13.987	10.764	3.754	6.167	7.285
4 iter.	14.278	10.801	3.538	6.004	7.223
5 iter.	14.492	10.880	3.368	5.846	7.125

***Tabla IV.8** : Evaluación del algoritmo AR4 en un entorno ruidoso (SNR_G=9dB).*

18dB	SNR_{Global}	$SNR_{Seg.}$	ITAKURA	COSH	CEPSTRUM
Original	18.019	13.408	6.328	7.893	8.518
1 iter.	21.912	17.040	2.292	5.081	5.461
2 iter.	20.361	16.645	2.299	4.684	5.708
3 iter.	21.034	16.538	2.215	4.731	6.009
4 iter.	20.699	16.306	2.265	4.783	6.224
5 iter.	20.508	16.234	2.368	4.786	6.195

***Tabla IV.9** : Evaluación del algoritmo AR4 en un entorno poco ruidoso (SNR_G=18dB).*