Universidad Nacional Autónoma
de México
Facultad de Ingeniería

ANÁLISIS DE TÉCNICAS PARA REDUCCIÓN DE RUIDO AMBIENTAL EN VOZ

T E S I S
que presenta

José Alberto Pesado Santiago

a la
Facultad de Ingeniería de la
Universidad Nacional Autónoma de México
como requerimiento parcial para recibir el título de
Ingeniero en Computación

Febrero de 2005

Tutor: Dr. José Abel Herrera Camacho

AGRADECIMIENTOS

Agradezco principalmente a mis padres, quienes siempre me han brindado su apoyo y respaldo en mis proyectos personales y profesionales. Padres, gracias por confiar en mí.

A mi asesor Abel Herrera, por haber confiado en mi capacidad para realizar este proyecto y darme oportunidades y apoyos en mi formación profesional. Agradezco enormemente su apoyo para participar en el 11° Congreso Internacional Mexicano de Acústica, celebrado en la ciudad de Morelia Michoacán, México.

Quiero expresar mi gratitud a toda la gente que me ha dado muestras de amistad y compañerismo, especialmente a mi novia y amigos: Remedios, Juan (Nan), Adrián, Luis, Carlos (Eva), Samuel (Axe) y Augusto (Zet). A todos ustedes doy gracias por los innumerables momentos de reflexión que de alguna manera me han guiado hasta aquí. Sus comentarios me permitieron darle mejor forma a este trabajo.

Finalmente agradezco a mi universidad y sus profesores por mostrarme que la diversidad es un bien de gran valor, así como por darme un punto de referencia para seguir mi formación humana y profesional.

RESUMEN

El desarrollo de la voz como interfaz de comunicación entre sistemas de procesamiento natural y artificial ha generado la necesidad de tratar el problema de ruido acústico en señales de voz. Esta tesis hace un análisis de las técnicas de mayor relevancia para la reducción de ruido acústico ambiental de carácter aditivo, específicamente dirigidas a sistemas donde el escucha final es un oído humano. La *cancelación adaptable de interferencia* y *sustracción espectral* son estudiados como modelos de reducción de ruido, con énfasis en los requerimientos y calidad que cada modelo ofrece. Para cada modelo se presentan dos implementaciones distintas, por el *menor valor cuadrático medio* (LMS) y por LMS-Newton (LMS-N) como algoritmos de adaptación para un cancelador adaptable de interferencia, y sobre-sustracción con suelo espectral, de banda completa y por sub-bandas, como extensiones del modelo de sustracción espectral. Se hace un análisis comparativo de rendimiento basado en métricas subjetivas y objetivas de calidad de voz. Para ello se propuso un protocolo modificado de la *métrica de diagnóstico de aceptabilidad* (DAM), menos completo que el protocolo original, pero que permite una evaluación más práctica de la calidad subjetiva de la voz. Las pruebas de calidad objetivas fueron evaluadas mediante la métrica de *distorsión modificada del espectro Bark* (MBSD), la cual resultó ser de mayor confiabilidad por su mayor grado de correspondencia con las pruebas subjetivas. Los resultados obtenidos muestran que a pesar de que la sustracción espectral tiene menor rendimiento en la calidad percibida, tal calidad es comparable con la cancelación adaptable de interferencia. La sustracción espectral cobra mérito por tal situación debido a su número reducido de requerimientos en información y complejidad computacional.

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

NOMENCLATURA

Variables aleatorias

x	Muestra de un espacio aleatorio caracterizado por X
X	Variable aleatoria
$X(t)$	Proceso aleatorio
$f_X(x)$	Función de probabilidad de la variable aleatoria X
$F_X(x)$	Función de probabilidad acumulada de la variable aleatoria X
$E[X^n]$	n-ésimo momento de la variable aleatoria X
μ_X	Esperanza, valor esperado, o media de la variable aleatoria X
σ_X^2	Varianza, o segundo momento central de la variable aleatoria X
Cov_{XY}	Covarianza de las variables aleatorias X y Y
ρ_{XY}	Coeficiente de correlación de las variables aleatorias X y Y
$\mu_{X(t)}(t)$	Función media del proceso aleatorio $X(t)$
$r_{X(t)}(t_1,t_2)$	Función de autocorrelación del proceso aleatorio $X(t)$
$r_{X(t)}(\Delta)$	Función de autocorrelación del proceso aleatorio estacionario en sentido amplio $X(t)$

Señales

$x(t)$	Señal continua en el dominio del tiempo
$X(\omega)$	Señal continua en el dominio de la frecuencia
$x[t]$	Señal discreta en el dominio del tiempo
$X[k]$	Señal discreta en el dominio de la frecuencia
$x_w[t]$	Señal discreta en el dominio del tiempo, segmentada con la ventana w
$X_W[k]$	Señal discreta en el dominio de la frecuencia, segmentada con la ventana w

Reducción de Ruido

$s[t]$	Señal de voz libre de distorsiones acústicas
$\hat{s}[t]$	Estimación de la señal s
$n[t]$	Señal de ruido acústico aditivo
$sn[t]$	Señal de voz con distorsiones acústicas de carácter aditivo

Cancelación adaptable de interferencia

MSE, ξ	Error cuadrático medio
∇	Función gradiente
μ	Factor de convergencia LMS
R	Matriz de autocorrelación de n
P	Vector de correlación cruzada de sn y n
Q	Matriz de vectores de valores propios
Λ	Matriz de valores propios
λ	Valor propio
N_t	Vector de muestras secuenciales de n en el instante t

W_t	Vector de parámetros del filtro adaptable en el tiempo t
W^*	Vector de parámetros del filtro óptimo
$\varepsilon[t]$	Señal de error del filtro
L	Longitud y número de parámetros del filtro

Sustracción espectral

$w[t]$	Ventana de segmentación, para análisis en frecuencia de tiempo corto
$S_W[k]$	Espectro de s en el segmento w
$\|S_W[k]\|$	Magnitud del espectro S_W
$\angle S_W[k]$	Fase del espectro S_W
α	Factor de sobre-sustracción
β	Parámetro de suelo espectral

Capítulo 1

INTRODUCCIÓN

1.1 Importancia del tratamiento del ruido en señales de voz

Por naturaleza el ser humano tiene la capacidad de comunicarse con su entorno a través de diversas formas, cada una de las cuales llega a ser de mayor relevancia según el grado de su uso. Tal es el caso de la voz, capacidad natural humana que en años recientes ha sido adoptada en muchas aplicaciones de telecomunicaciones [29,37] y de interacción hombre-máquina [13] como método para transmitir información.

Como generalmente ocurre, la aplicación de modelos teóricos de laboratorio en situaciones reales implica el tratamiento de condiciones y efectos no previstos o difíciles de controlar, ya sea por su desconocimiento o por el nivel de abstracción del modelo inicial. El ruido, cualquiera que sea su tipo, representa un factor importante que limita los niveles de rendimiento de muchos sistemas de ingeniería, sobre todo de aquéllos basados en voz cuando de ruido acústico ambiental se trata. En varias investigaciones [33,23] se ha demostrado que aún con los más novedosos avances en las técnicas para la adquisición, transmisión o análisis de señales de voz, este tipo de sistemas son susceptibles a disminuir sus niveles de rendimiento cuando son probados en ambientes ruidosos.

El ruido acústico ambiental se encuentra presente en la gran mayoría de las aplicaciones *front-end*[1] del mundo real, lo cual nos obliga a buscar medios a través de los cuales los sistemas basados en voz puedan mantener un nivel de rendimiento cercano al que logran en condiciones libres de ruido.

1.2 Contexto del problema

Un ambiente acústicamente ruidoso puede producir una gran variedad de efectos nocivos en la voz, reflejados principalmente durante el procesamiento de la misma. La reducción de tales efectos es una tarea que puede realizarse en varias etapas; durante la etapa de adquisición suelen utilizarse micrófonos direccionales que responden únicamente a ondas sonoras recibidas dentro de un rango aproximado de 180°, en dirección del eje principal del micrófono, tal como se muestra en la Figura 1.1 [2]. Aún en esta circunstancia, la voz se encuentra en competencia con otras señales indeseables. Esta situación es resuelta en muchas de las aplicaciones usando el micrófono cerca de la boca; solución impráctica que no siempre es posible. Una opción es la modificación interna de los sistemas de procesamiento de voz ya existentes, de tal manera que sean lo menos susceptibles al ruido para que puedan mantener un nivel mínimo de rendimiento. Sin embargo, esta opción involucra mayores costos y

[1] Entiéndase *front-end* como sinónimo de interfaz de usuario final, término comúnmente utilizado para referirse a la parte de un sistema que se encuentra en contacto directo con el usuario.

complejidad en el diseño de los sistemas de procesamiento. Otra opción más aceptada es la utilización de un sistema de pre-procesamiento que permita aislar el problema sin alteración alguna del sistema de procesamiento, dedicado únicamente a reducir los efectos producidos por el ruido. El trabajo elaborado en esta tesis está dirigido al análisis de técnicas de procesamiento con tal objetivo.

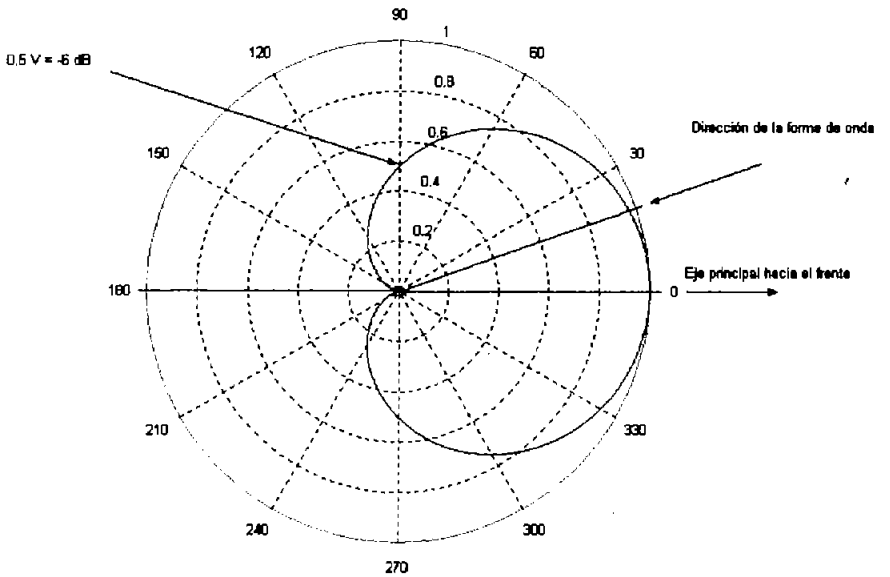


Figura 1.1 Patrón direccional de un micrófono cancelador de ruido.

1.3 Alcance

El problema de cancelación o reducción de ruido es un problema que difícilmente puede ser generalizado. El diseño de un sistema de procesamiento de tal naturaleza debe tomar en cuenta características específicas del ruido a tratar; la forma en la que la distorsión se lleva a cabo, el entorno de aplicación y los recursos disponibles. El análisis realizado en esta investigación se limita a los casos de reducción de distorsión debida a ruido ambiental de naturaleza acústica aditiva. Se consideraron las técnicas que particularmente están dirigidas a que un oído humano evalúe su rendimiento. Se hizo un análisis comparativo de la eficiencia de las técnicas seleccionadas, dirigido principalmente por la valoración de la cantidad de recursos utilizados y los resultados obtenidos de la evaluación objetiva y subjetiva de la eficacia de la técnica. Tal evaluación se realizó a partir de los resultados obtenidos del procesamiento digital de señales fuera de tiempo real. Debido a la falta de recursos, la implementación física de un sistema para la reducción de ruido se encontró fuera de los alcances de esta tesis.

1.4 Objetivos

Los objetivos que se persiguen con el análisis presentado son los de proveer una visión más amplia del desarrollo actual en el campo de la reducción de ruido acústico ambiental; identificar soluciones óptimas a problemas futuros, cuyas condiciones específicas correspondan o se aproximen a alguna de las técnicas analizadas; así como identificar limitaciones de las técnicas que sirvan como punto de partida de futuras investigaciones.

1.5 Estructura de la tesis

El primer paso para abordar el problema de reducción de ruido en señales de voz es identificar las características más relevantes de las señales involucradas. En el Capítulo 2 se presenta un descripción suficiente para el análisis de los aspectos principales de la voz y el ruido, además de una breve revisión de los modelos existentes para reducir ruido. Producción, adquisición, representación y percepción son los elementos de interés respecto de las señales de voz. Por su parte, el ruido acústico es tratado como un caso particular de la variedad de distorsiones debidas a ruido. La discusión se centra en el caso específico de ruido aditivo y cómo éste distorsiona la señal de voz. Se identifica una clasificación de las técnicas para reducción de ruido, basada en una extensa revisión de la literatura. De la variedad de técnicas presentadas se seleccionan la *cancelación adaptable de interferencia* y *sustracción espectral*, siendo éstas las de mayor relevancia por su extenso campo de aplicación.

Una vez que se tiene la posible solución al problema, es conveniente establecer la forma en la que se evaluará el nivel de éxito alcanzado con los resultados obtenidos. Si el propósito de la reducción de ruido es elevar la calidad de voz percibida, es evidente que su medición proporcione el nivel de eficacia de la solución. El Capítulo 3 resume las técnicas empleadas en la medición subjetiva y objetiva de la calidad de voz, considerando la inteligibilidad y la aceptabilidad como dimensiones de calidad. Se propone un protocolo de evaluación de calidad subjetiva para las señales de voz, basado principalmente en la métrica de diagnóstico de aceptabilidad. La *relación señal a ruido segmental* (SNR_{seg}), SNR_{seg} *por sub-bandas* ($SNR_{fw\text{-}seg}$), y la *distorsión modificada del espectro Bark* (MBSD) son seleccionadas como posibles métricas para el análisis posterior.

El análisis de las técnicas inicia en el Capítulo 4 con el modelo de cancelación adaptable de interferencia. Se analiza la estructura del modelo como caso particular de los sistemas adaptables de lazo cerrado, distinguiendo las señales involucradas y los módulos de procesamiento, un filtro transversal y un módulo de adaptación. La solución del caso estacionario ya se encuentra definida, sin embargo, esta solución presenta varias limitaciones para poder lograrse, por lo que se han tenido que generar varias aproximaciones de la solución. Estas aproximaciones toman forma de algoritmos de adaptación de los parámetros del filtro; el *menor valor cuadrático medio* (LMS) y LMS-Newton son los casos seleccionados para el análisis. El análisis de LMS y LMS-N parte de la distinción del paradigma que cada uno sigue en el proceso de adaptación. A este

respecto, se hace un estudio de sus comportamientos en velocidad y precisión de adaptación. De igual manera, se hace mención de aspectos importantes de la aproximación causal y de longitud finita del filtro generado mediante esta técnica.

En el Capítulo 5 se estudian las bases del modelo de sustracción espectral, con énfasis en los supuestos de cualidad estacionaria, la detección de actividad de voz y una segmentación no intrusiva. El incumplimiento de tales supuestos da lugar al análisis del *ruido residual*. En este sentido, se presentan algunas extensiones del modelo que pretenden reducir al mínimo los efectos perceptibles de tal ruido. Sobre-sustracción y suelo espectral, en banda completa y por sub-bandas son las principales extensiones analizadas a este respecto. De igual manera, se describe el funcionamiento del detector de actividad silencios / actividad de voz utilizado en las pruebas experimentales.

El análisis de los resultados obtenidos con cada una de las configuraciones de cancelación adaptable de interferencia y sustracción espectral se realiza en el Capítulo 6. Primero se analiza el comportamiento de la SNR_{seg}, $SNR_{fw\text{-}seg}$ y MBSD como métricas objetivas de calidad de voz. Tal análisis se realiza para seleccionar la métrica que mejor describe los resultados obtenidos con el protocolo modificado de la métrica de diagnóstico de aceptabilidad (PMDAM). En este sentido, la MBSD es seleccionada como referencia para el análisis siguiente. Se hace una comparación de la calidad MBSD lograda por la cancelación adaptable de interferencia y sustracción espectral para distintas configuraciones de distorsión, tipo y magnitud. Los tipos de ruido ambiental experimentados fueron: aproximadamente estacionario con distribución espectral uniforme, aproximadamente estacionario con distribución espectral no uniforme, y no estacionario con distribución espectral no uniforme. Los grados de distorsión que se utilizan en el análisis corresponden a –10, 0 y 10 dB de SNR. Se hace un análisis cualitativo de la complejidad computacional de cada técnica, análisis basado en el retardo que cada señal experimenta al ser procesada por dichas técnicas. Finalmente, se hace un estudio comparativo de sustracción espectral en sus variantes de sobre-sustracción y de análisis en sub-bandas.

Capítulo 2

RUIDO Y VOZ

2.1 Generalidades de procesos estocásticos

A lo largo de este trabajo se hace uso de términos asociados con señales aleatorias, así que conviene establecer el concepto de aquéllos que son utilizados en la terminología del procesamiento de señales. Este capítulo no pretende hacer una extensa revisión acerca de la teoría de señales aleatorias, únicamente homogenizar la terminología utilizada. Una explicación más extensa del contenido de este capítulo puede encontrarse en las referencias [8,18,26], o bien en cualquier libro dedicado al análisis de señales o variables aleatorias.

2.1.1 Variables aleatorias

Una *variable aleatoria* X de un experimento se define como una función de mapeo $X(s)$ entre una muestra s del espacio de posibles resultados S del experimento y el conjunto de los números reales R. Se parte del hecho que no es posible determinar con total certidumbre el resultado de dicho experimento. Si la variable aleatoria tiene un número infinito de posibles valores reales se puede hablar de una variable aleatoria continua, pero, si únicamente puede tomar un número finito de números reales se denomina variable aleatoria discreta.

La propiedad más importante para una variable aleatoria es su *función de probabilidad* $f_X(x)$, la cual se supone que es una medida de la posibilidad de que la variable tome cierto valor x, esto es: $f_X(x) = P[X = x]$. La función de probabilidad puede definirse tanto para variables aleatorias continuas como discretas, recibiendo el nombre de *función de densidad* y *función de masa*, respectivamente. La función de probabilidad $f_X(x)$ cumple con las siguientes propiedades

Variable aleatoria discreta	Variable aleatoria continua	
$0 \le f_X(x) \le 1$	$f_X(x) \ge 0$	(2.1)
$\sum_{\forall x} f_X(x) = 1$	$\int_{-\infty}^{+\infty} f_X(x)dx = 1$	(2.2)
$P[a \le X \le b] = \sum_{x_i \ge a}^{x_i \le b} f_X(x_i)$	$P[a \le X \le b] = \int_a^b f_X(x)dx$	(2.3)
$F_X(x) = \sum_{x_i \le x} f_X(x_i)$	$F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi$	(2.4)

Donde $F_X(x)$ es la *función de probabilidad acumulada* que determina la probabilidad que el valor tomado por la variable aleatoria sea menor o igual a x, $F_X(x) = P[X \leq x]$.

El *n-ésimo momento* de la variable aleatoria X se define como

$$E[X^n] = \sum_{\forall x_i} x_i^n f_X(x_i) \qquad \text{si } X \text{ es variable aleatoria discreta} \qquad (2.5)$$

$$E[X^n] = \int_{-\infty}^{+\infty} x^n f_X(x) dx \qquad \text{si } X \text{ es variable aleatoria continua} \qquad (2.6)$$

Los momentos más importantes para análisis de señales aleatorias son los dos primeros, cuando $n = 1$ y $n = 2$. El primero de ellos corresponde al valor esperado de la variable aleatoria, conocido también como *media*, $\mu_X = E[X]$. El segundo momento se conoce como *valor cuadrático medio*, $E[X^2]$.

Otra clase de momentos basados en la media de la variable aleatoria corresponde a los *momentos centrales*. El *n-ésimo momento central* de la variable aleatoria X se define como

$$E[(X - \mu_X)^n] = \sum_{\forall x_i} (x_i - \mu_X)^n f_X(x_i) \qquad \text{si X es v. aleatoria discreta} \qquad (2.7)$$

$$E[(X - \mu_X)^n] = \int_{-\infty}^{+\infty} (x - \mu_X)^n f_X(x) dx \qquad \text{si X es v. aleatoria continua} \qquad (2.8)$$

El primer momento central es cero. El segundo de ellos es el más útil en muchas de las aplicaciones, se conoce como *varianza* y se representa como σ_X^2. La raíz cuadrada de la varianza se conoce como desviación estándar de X, σ_X, que a su vez da una idea de las variaciones de X alrededor de su media, según su función de probabilidad. Puede observarse que cuando la variable aleatoria tiene media igual a cero, la varianza coincide con el valor cuadrático medio.

En ocasiones es necesario considerar más de una variable aleatoria en la descripción del comportamiento de un experimento o fenómeno. El análisis para tales situaciones es muy similar al caso de una única variable, sólo que ahora habrá que considerar las demás variables en todas las expresiones. Es entonces que por extensión se puede definir como *función de probabilidad conjunta* $f_{X,Y}(x,y)$ a la probabilidad de que las variables aleatorias X y Y tomen los valores específicos x y y respectivamente, para el caso de una función de dos variables aleatorias. Esto es: $f_{X,Y}(x,y) = P[X = x, Y = y]$. De igual forma, la función de probabilidad acumulada puede extenderse a un mayor número de dimensiones, en el caso bidimensional queda $F_{X,Y}(x,y) = P[X \leq x, Y \leq y]$.

Se define el concepto de *covarianza* entre las variables aleatorias X y Y como

$$Cov_{XY} = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y \tag{2.9}$$

En el que el primer término *E[XY]* comúnmente se conoce como *correlación*, especialmente cuando se tratan múltiples variables aleatorias organizadas en vectores, generando matrices en lugar de escalares. El uso de esta terminología puede llegar a causar confusión, ya que el valor del término *E[XY]* no es por sí solo representativo del nivel de correlación o dependencia lineal dado por el *coeficiente de correlación*.

Dos variables aleatorias X y Y son independientes si no existe alguna relación funcional entre los valores tomados por dichas variables, en tal caso se cumple que $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Una medida del grado de dependencia lineal entre dos variables aleatorias se obtiene a través del *coeficiente de correlación*, definido como

$$\rho = \rho_{XY} = \frac{Cov_{XY}}{\sigma_X \sigma_Y} \tag{2.10}$$

Limitado en el intervalo $-1 \leq \rho \leq 1$. Si $|\rho| \rightarrow 0$, se dice que X y Y son variables aleatorias no correlacionadas, por el contrario si $|\rho| \rightarrow 1$, X y Y se encuentran más correlacionadas mientras más se acerque a la unidad. Debe notarse que el coeficiente de correlación proporciona información acerca de la dependencia lineal de las variables, así que $\rho = 0$ no implica la independencia entre las variables.

2.1.2 Procesos aleatorios

Se denomina *proceso estocástico* o *aleatorio* al conjunto de variables aleatorias relacionadas con el mismo experimento, cada una asociada a un punto en el tiempo. Nos referiremos únicamente a procesos en tiempo discreto. Se denota como $X(t)$, y como $X(t_i)$ a la i-ésima variable aleatoria en la secuencia de tiempo. Dentro del análisis de señales, un proceso estocástico se conoce también como *señal aleatoria*. Como posición opuesta, puede hablarse de *proceso determinístico* cuando se habla de un conjunto de variables cuyos valores pueden ser determinados para cualquier instante de tiempo. Cada variable aleatoria de un proceso aleatorio posee su propia función de probabilidad, es así que un proceso aleatorio puede entenderse como una familia de funciones de probabilidad y describirse a través de su función de probabilidad conjunta

$$f_{X(t_1),X(t_2),\ldots,X(t_N)}(x_1, x_2, \ldots, x_N) = P[X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_N) = x_N] \tag{2.11}$$

para N puntos seleccionados del proceso aleatorio. Se entiende como *realización* al resultado de una experimentación de un proceso estocástico, formado por el conjunto de resultados obtenidos por las variables aleatorias del proceso. De esta manera, la forma de onda de una señal aleatoria corresponde a una realización de la misma.

Se dice que un proceso aleatorio es *estrictamente estacionario de orden* N si todas sus propiedades estadísticas son invariantes a cualquier traslación en el tiempo; esto es, el

proceso aleatorio $X(t)$ es estrictamente estacionario si y solo si, para cualquier valor de N, τ y cualquier conjunto de variables aleatorias tomadas en t_1, t_2, ..., t_N separadas arbitrariamente, se cumple que

$$f_{X(t_1),X(t_2),\ldots,X(t_N)}(x_1,x_2,\ldots,x_N) = f_{X(t_1+\tau),X(t_2+\tau),\ldots,X(t_N+\tau)}(x_1,x_2,\ldots,x_N) \tag{2.12}$$

Debe observarse que el desplazamiento τ es arbitrario y que por lo tanto un proceso o señal aleatoria no puede estar limitado en el tiempo. En muchas de las aplicaciones es suficiente considerar dos puntos del proceso en el cálculo de momentos de segundo orden, así que con una visión pragmática se define el concepto de proceso *débilmente estacionario*, *estacionario en sentido amplio* o simplemente *estacionario* para los efectos de esta tesis. El proceso $X(t)$ es estacionario en sentido amplio si su *función media* $\mu_{X(t)}(t)$ no depende de t, o bien, es constante, y su *función de autocorrelación* [1] $r_{X(t)}(x_1,x_2)$ puede calcularse como $r_{X(t)}(\Delta)$, donde Δ es la separación en el tiempo entre x_1 y x_2, esto es

$$\mu_{X(t)}(t) = E[X(t)] = \text{constante} \tag{2.13}$$

$$r_{X(t)}(t_1,t_2) = E[X(t_1)X(t_2)] = E[X(t)X(t-\Delta)] = r_{X(t)}(\Delta) \tag{2.14}$$

$$\Delta = t_2 - t_1$$

Se observa que para que se cumplan las condiciones anteriores el proceso debe ser estrictamente estacionario de segundo orden. Si lo anterior es cierto, es posible obtener de $r_{X(t)}(0)$ la *potencia media* (valor cuadrático medio) de la señal aleatoria. En el caso de que las variables aleatorias que componen la señal estacionaria sean independientes, es posible obtener $\mu_{X(t)}{}^2$ ya que

$$E[X(t)X(t-\Delta)] = E[X(t)]E[X(t-\Delta)] = \{E[X(t)]\}^2 \tag{2.15}$$

Sobre todo cuando $\Delta \to \infty$. La función de autocorrelación puede utilizarse como una prueba de la periodicidad de una señal estacionaria ya que

$$\lim_{\Delta\to\infty} r_{X(t)}(\Delta) = \begin{cases} \mu_{X(t)}^2 \text{ si } X(t) \text{ no es periódica} \\ r(\Delta - T) \text{ si } X(t) \text{ es periódica con periodo T} \end{cases} \tag{2.16}$$

Se deduce entonces que si $X(t)$ es periódica, la función de autocorrelación puede calcularse tomando únicamente un periodo completo de la señal. Otra importante función que describe el proceso aleatorio estacionario es la *función de autocovarianza*[2], misma que se define a partir de la función de autocorrelación como

---

[1] La función de autocorrelación es un caso particular de la función de correlación, misma que se define para dos procesos aleatorios cualesquiera, $E[X(t)Y(t+\Delta)]$.

[2] Al igual que la función de autocorrelación, la función de autocovarianza es un caso de la función de covarianza $E[(X(t)-\mu_{X(t)})(Y(t)-\mu_{Y(t)})]$, donde $X(t)$ y $Y(t)$ son el mismo proceso.

$$Cov_{X(t)}(t_1,t_2) = E\left[\left(X(t_1)-\mu_{X(t)}(t_1)\right)\left(X(t_2)-\mu_{X(t)}(t_2)\right)\right] = r_{X(t)}(t_1,t_2)-\mu_{X(t)}(t_1)\mu_{X(t)}(t_2) \qquad (2.17)$$

El caso $Cov_{X(t)}(0)$ proporciona la *potencia media de la componente AC* (varianza) de la señal aleatoria. Además, con media constante igual a cero puede verificarse que la función de autocovarianza equivale a la función de autocorrelación.

Se dice que un proceso aleatorio es *ergódico* si exhibe las mismas características estadísticas promedio sobre la dimensión temporal de una sola realización, así como en el espacio de diferentes realizaciones del proceso. Si lo anterior ocurre, dichas estadísticas no pueden ser función del tiempo así que el proceso debe ser estacionario, aún cuando el caso opuesto no es siempre cierto.

2.1.3 Autocorrelación y espectro de potencia

La expresión (2.14) es la definición estadística de la función de autocorrelación que para ser utilizada se requiere del conocimiento de la función de probabilidad $f_X(x)$. Otra forma de obtener la función de autocorrelación a partir de la señal $X(t)$ es

$$r_{X(t)}(\Delta) = \lim_{N\to\infty} \frac{1}{2N-1} \sum_{n=-N}^{N} X(n)X(n-\Delta) \qquad (2.18)$$

útil para señales de potencia, sin embargo, en procesamiento digital de voz normalmente se trabaja con señales de energía o longitud finita. Una opción es utilizar la sumatoria completa en lugar del promedio, quedando

$$r_{X(t)}(\Delta) = \sum_{n=-\infty}^{\infty} X(n)X(n-\Delta) \qquad (2.19)$$

de tal forma que si $X(t)$ es una señal aleatoria estacionaria, $X(\omega) = DFT\{ X(t) \}$, y $R_{X(t)}(\omega) = DFT\{ r_{X(t)}(\Delta) \}$, donde DFT es la Transformada Discreta de Fourier, entonces puede demostrarse [31] que

$$R_{X(t)}(\omega) = X(\omega)X^*(\omega) = |X(\omega)|^2 \qquad (2.20)$$

lo que significa que la DFT de la función de autocorrelación de una señal estacionaria, corresponde al espectro de potencia de la señal. Dado que este espectro provee información acerca de la distribución de la potencia de la señal sobre cada frecuencia, se le conoce también como Densidad Espectral de Potencia (PSD).

La estrecha relación que existe entre el espectro de una señal y su función de correlación sugiere una forma de identificar el carácter estacionario de la señal. Para que la ecuación (2.20) pueda satisfacerse, debe cumplirse que la función de autocorrelación pueda ser expresada a través de un único argumento, tal y como se muestra en la ecuación (2.14). Si la señal aleatoria es estacionaria deben cumplirse

(2.13) y (2.14) para cualquier origen temporal. De aquí se puede deducir que, como la función de correlación de una señal se encuentra totalmente relacionada con su espectro de potencia, se puede identificar que una señal es estacionaria si su espectro de potencia (y en general cualquier grado de magnitud) se mantiene constante a través del tiempo. Esta forma de identificar la cualidad estacionaria de una señal es más evidente cuando se trabaja con el espectro de la señal, así que será utilizada para los análisis de esta tesis.

En lo que respecta a aplicaciones de ingeniería, los conceptos revisados hasta ahora resultan ser muy útiles pero poco prácticos si se intentan aplicar directamente. Dichos conceptos generalmente se definen para señales aleatorias de longitud infinita, siendo que siempre se trabaja con señales limitadas de corta duración. Es por eso que es necesario utilizar estimadores que aproximen lo mejor posible las características de la señal como si fuera de larga duración. En la Tabla 2.1 se enumeran algunos estimadores para propiedades estadísticas utilizadas en este trabajo.

Función media	$\mu_{X(t)} \cong \frac{1}{N}\sum_{n=1}^{N} X(n)$
Función de correlación cruzada	$r_{X(t)Y(t)}(\Delta) \cong \frac{1}{N}\sum_{n=1}^{N} X(n)Y(n-\Delta)$
Función de autocorrelación	$r_{X(t)}(\Delta) \cong \frac{1}{N}\sum_{n=1}^{N} X(n)X(n-\Delta)$

Tabla 2.1 Estimadores estadísticos para procesos o señales aleatorias en tiempo corto.

El interés de revisar conceptos relacionados con señales aleatorias viene del hecho que en la mayoría de los desarrollos de tratamiento de voz y ruido acústico éstas señales se modelan como procesos o señales aleatorias.

2.2 Aspectos principales de la voz

La voz humana forma parte de un fenómeno más general conocido como sonido, un proceso físico que implica la propagación de perturbaciones debidas a cambios de presión sobre un medio elástico, normalmente el aire. Estas perturbaciones son generadas por cuerpos en vibración, que en el caso específico de la voz se realizan a través de los músculos y órganos que componen el tracto pulmonar, la laringe y el tracto vocal.

2.2.1 Digitalización

La voz en su papel de proceso de transmisión de información sólo puede ser observada y procesada como formas de onda de sonido, captadas a través de un transductor que proporcione una representación eléctrica del fenómeno. Una vez que se cuenta con la señal analógica, normalmente conviene realizar un proceso de digitalización de la señal

que facilite su procesamiento. Muestreo, cuantización y codificación son pasos esenciales en el proceso de conversión analógico/digital (A/D), que deben ser cuidadosamente realizados.

La voz humana de un adulto se encuentra ubicada en un ancho de banda con límite máximo promedio de 10 kHz, mientras que el intervalo supuesto de frecuencias audibles está entre 20 Hz y 20 kHz, aunque por deterioro natural del oído sólo se alcanzan 15 kHz en promedio. Utilizar dispositivos de procesamiento y transmisión que funcionen a esas frecuencias resultaría muy costoso debido a la calidad de recursos tecnológicos requeridos. Afortunadamente la información contenida en la voz no se encuentra uniformemente distribuida en todo su ancho de banda, por el contrario, se sabe que las componentes frecuenciales dominantes corresponden a los tres grupos de frecuencias conocidas como formantes, denotadas como F1, F2 y F3, de la más baja a la más alta, siendo esta última de valor más particular para cada hablante [12,6]. El ancho de banda ocupado por estas tres formantes va de 0.3 a 3.4 kHz en promedio, que de acuerdo con el teorema de muestreo coincide con el ancho de banda efectivo diseñado para una línea telefónica usando una tasa de muestreo de 8 kHz. El teorema de muestreo supone también que la señal se encuentra perfectamente limitada en banda, así que para evitar efectos de distorsión en las componentes de alta frecuencia por el efecto de *aliasing* o traslape espectral, es necesario aplicar un filtro paso bajas que restrinja el ancho de banda de la señal antes de ser muestreada.

El rango dinámico de la voz supera los 50 dB, así que para disminuir al mínimo la distorsión de la señal por efecto de la cuantización son necesarios 10 bits para el caso de cuantización lineal y 8 bits para una cuantización con compresión-expansión por Ley μ o Ley A [12,31,9]. En el caso específico de señales de voz, las componentes de alta frecuencia normalmente se encuentran por debajo de los niveles de energía promedio, por lo tanto, es posible incrementar la SNR de cuantización comprimiendo el rango dinámico de la señal a través del aplanamiento del espectro. Este procedimiento se realiza a través de la aplicación de un filtro paso altas antes de la conversión A/D, conocido como filtro de pre-énfasis.

Por último, se requiere de un proceso de codificación a través del cual se asignan identificadores numéricos a los niveles de voltaje resultado de la cuantización. En este proceso normalmente se hace uso de una codificación binaria.

La importancia de los requerimientos de conversión A/D antes establecidos impulsó a que la Unión Internacional de Telecomunicaciones (ITU), a través del Comité Consultivo Internacional de Telegrafía y Telefonía (CCITT), emitiera en 1972 la recomendación G.711 para la modulación de código de pulso (PCM) en señales de telefonía. En esta recomendación se especifica una tasa de muestreo de 8 kHz (64 kbits/s) y un cuantizador de 8 bits con Ley μ o Ley A como método de compresión instantánea [9]. A pesar de que en la actualidad existen varias recomendaciones que superan el rendimiento de la G.711, su simplicidad ha ofrecido un rendimiento adecuado para aplicaciones de voz por muchos años. Las señales de voz utilizadas en las demostraciones y pruebas experimentales de esta tesis siguen una tasa de muestreo de 8 kHz, 16 bits de cuantización y una codificación PCM. Esta codificación ofrece una

SNR de cuantización aproximada de 89 dB[1], suficiente para cubrir los 35 dB mínimos para una buena percepción [12,31,24].

2.2.2 Aspectos de producción

En el área de procesamiento de voz comúnmente se hacen clasificaciones de los tipos de sonidos que el sistema vocal humano puede producir. Una de las clasificaciones más utilizadas hace la distinción entre sonidos *sordos* y sonidos *sonoros*. Esta clasificación se basa en la forma en que los sonidos son producidos; específicamente, se habla de sonidos sonoros cuando su producción implica la vibración de las cuerdas vocales, y de sonidos sordos cuando éstos son producidos sólo a través del tracto pulmonar y el tracto vocal, sin el movimiento de vibración de las cuerdas vocales. La participación de las cuerdas vocales en la generación de sonido implica la interrupción aproximadamente periódica del paso de aire proveniente de los pulmones. Esto da lugar a que la forma de onda de los sonidos sonoros producidos por el tracto vocal tenga un comportamiento aproximadamente periódico. Lo anterior no ocurre para los sonidos sordos, cuyo comportamiento es parecido a ruido de banda ancha. Gran parte de los modelos de producción de voz utilizados para reducción de ruido y síntesis de voz se basan en esta clasificación.

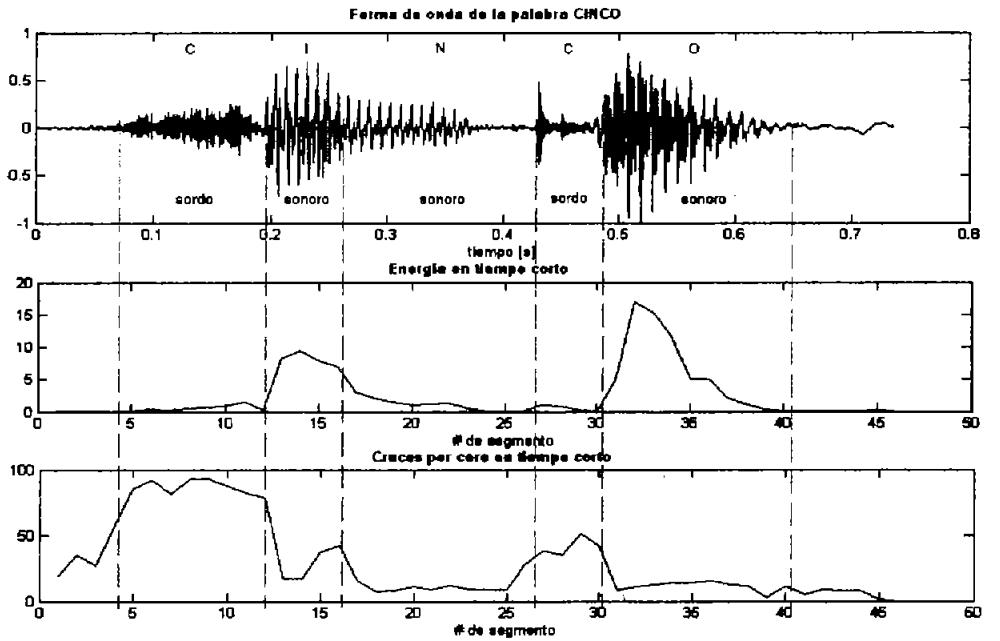


Figura 2.1 Forma de onda de sonidos sordos y sonidos sonoros de una señal de voz. Alta energía para sonidos sonoros, baja para sonidos sordos y mayor número de cruces por cero para sonidos sordos.

[1] Calculado a partir de la ecuación SNR = 6.02 B - 7.27, donde B es el número de bits de cuantización.

Otra clasificación de origen lingüístico con importancia para análisis de señales es la que distingue entre sonidos vocálicos y consonantes. Los sonidos vocálicos son todos sonidos de carácter sonoro, caracterizados por tener muy poca restricción al flujo de aire cuando éste pasa por el tracto vocal. Los sonidos consonantes pueden ser sonoros o sordos, producidos por distintas configuraciones orgánicas del tracto vocal para restringir el paso del aire, generando que sus formas de onda sean menos previsibles.

La falta de restricción en la circulación de aire de los sonidos vocálicos provoca que normalmente posean mayor energía que los sonidos consonantes, por otro lado, también hay evidencia que los sonidos sordos tienen componentes de mayor frecuencia, la cantidad de veces que la señal cruza por cero es mayor para los sonidos sordos que para los sonoros. La Figura 2.1 muestra la forma de onda de una señal de voz con sus respectivas gráficas de energía y de cruces por cero evaluadas en tiempo corto.

2.2.3 Aspectos acústicos

La naturaleza acústica del sistema humano de producción de voz tiene como consecuencia la presencia del efecto de resonancia. Este efecto se presenta como un conjunto de frecuencias resonantes para cada configuración física del sistema de producción, así que cada tipo de sonido tendrá su propio conjunto de frecuencias resonantes. Una forma de ver la localización de dichas frecuencias es a través de la magnitud del espectro de frecuencia de la señal, donde puede distinguirse que las componentes mas significativas de la señal se encuentran formadas por las frecuencias resonantes; es por eso que las frecuencias de resonancia se conocen también como frecuencias formantes.

En la Figura 2.2 se muestran los espectros de magnitud de un segmento de los sonidos */i/* y */c/*. Como puede notarse, la periodicidad del sonido sonoro */i/* se caracteriza por una estructura armónica con separación de F0 Hz, donde F0 es la frecuencia de oscilación de las cuerdas vocales, conocida como frecuencia fundamental o *pitch*[1]. Esta estructura no se presenta en el caso del sonido */c/*. Otra característica observable es la mejor definición de los grupos de frecuencias formantes F1, F2 y F3 para el caso del sonido sonoro */i/*.

[1] Dentro de la terminología de procesamiento de señales de voz comúnmente los términos pitch y frecuencia fundamental se usan indistintamente, sin embargo, en términos de psicoacústica el pitch corresponde a la frecuencia fundamental percibida por el oído humano.

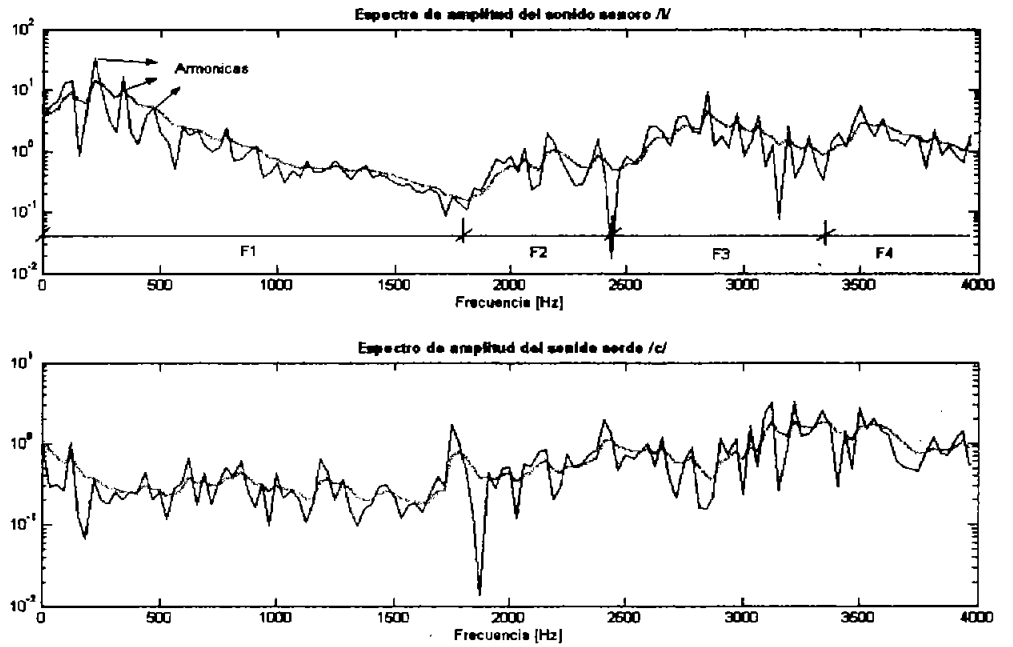
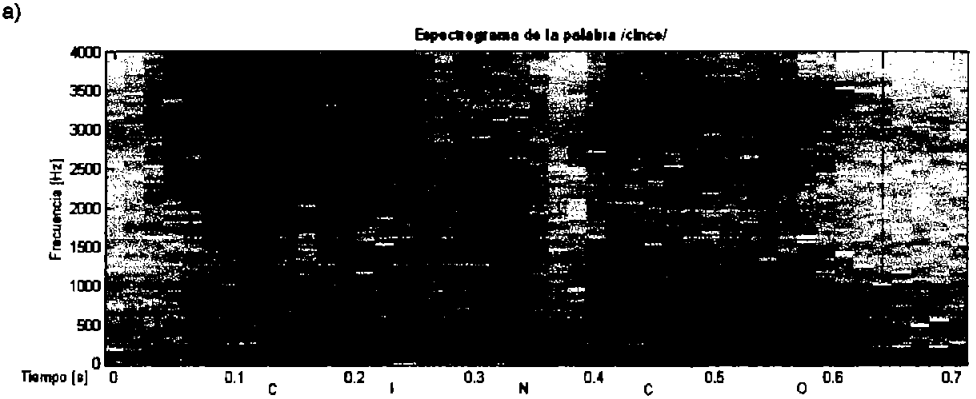


Figura 2.2 Magnitud del espectro de frecuencia del sonido sonoro */i/*, y del sonido sordo */c/*. En el primer caso se distingue una separación aproximadamente constante entre las componentes del espectro, también se distinguen los grupos de frecuencias formantes F1, F2 y F3. Esto no ocurre para los sonidos sordos.

El espectro de frecuencia de la voz se define a partir de la posición de cada uno de los músculos y órganos que componen el sistema de producción de voz. Dada la amplia variedad de sonidos que puede contener una señal de voz, queda implícito que su espectro de frecuencia no puede ser constante a lo largo del tiempo. Una representación útil de la voz como fenómeno variante en el tiempo es el *espectrograma*, una gráfica tridimensional que muestra la variación del espectro de magnitud a través del tiempo. La Figura 2.3 presenta el espectrograma de la palabra /cinco/.

a)



b)

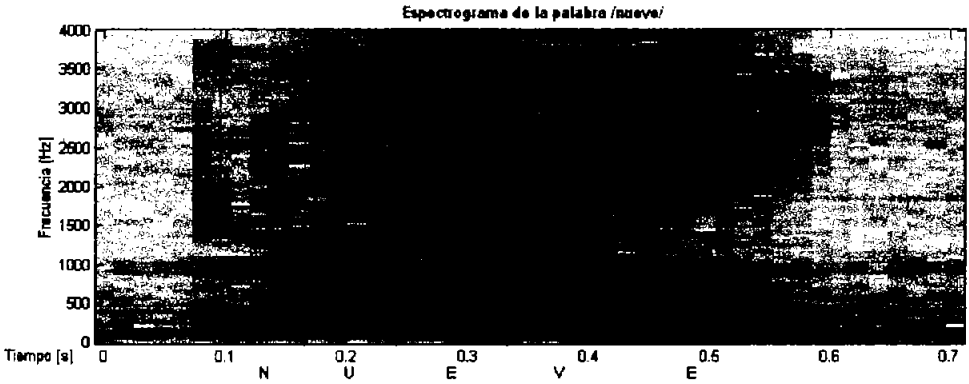


Figura 2.3 Espectrogramas de las palabras /cinco/ y /nueve/, para segmentos de 32 ms o 256 muestras, tonalidades oscuras representan componentes de mayor energía. a) Puede observarse cierto grado de cualidad estacionaria en los intervalos correspondientes a sonidos sonoros /i/, /n/, y /o/. b) En este caso permite ver con mayor claridad el comportamiento aproximadamente estacionario de las formantes del espectro de los sonidos sonoros.

Evidentemente la voz no es estacionaria debido a los cambios en su espectro de magnitud durante el tiempo, sin embargo, esta representación pone en evidencia la existencia de cierto grado de cualidad estacionaria en algunos segmentos, específicamente en los sonidos sonoros. Esta cualidad estacionaria aproximada normalmente se presenta en intervalos cortos de tiempo de 10 a 40 ms, equivalentes a 80 y 320 muestras respectivamente, para una frecuencia de muestreo de 8 kHz [31,1,12,21,11]. La cualidad aproximadamente estacionaria de la voz tiene mucha importancia ya que es base de un gran número de aplicaciones de procesamiento de voz, entre las que destacan las de reducción de ruido.

2.2.4 Aspectos perceptivos de la voz

En la actualidad se conoce poco acerca del proceso humano de percepción de la voz, que además del proceso fisiológico para percibir sonidos, consta de un complejo análisis simultáneo de información contenida en la señal de voz, encaminado hacia el entendimiento del mensaje recibido. Debe tomarse en cuenta que además de la información de carácter fonético, sintáctico y semántico del lenguaje hablado, también se recibe información ambiental que en este contexto se considera como ruido. Convenientemente para el problema de percepción de voz distorsionada por ruido acústico, ni la sintaxis ni la semántica fonéticas se ven afectadas directamente. Sin embargo, el contenido fonético de la voz es altamente susceptible a distorsiones acústicas que en general degradan la *calidad* percibida de la voz, medida conjunta de su grado de *aceptabilidad* e *inteligibilidad*. La aceptabilidad es una medida subjetiva que describe qué tan bien es percibida la señal en términos de naturalidad y esfuerzo del escucha. Por otro lado, la inteligibilidad es una medida menos subjetiva que indica qué tan fácil es extraer la información contenida en la señal de voz. El calificativo de tales métricas, y en general de la calidad de una señal de voz, depende de ciertos aspectos perceptivos conocidos.

La calidad percibida de una señal de voz depende fuertemente de una representación precisa del espectro de magnitud en tiempo corto, mientras que el espectro de fase resulta poco relevante. La localización de los grupos formantes es perceptivamente más importante que cualquier otro detalle en la forma del espectro. Tal localización puede variar entre personas, aunque generalmente F1 se encuentra entre 200 y 1200 Hz, y F2 entre 600 y 3500 Hz [12]. Se sabe además que la SNR debe ser por lo menos de 35 a 40 dB para que las distorsiones no sean percibidas por el oído, valores mayores representan cambios poco significativos en la percepción de la señal [24,8]. Por otro lado, la voz pierde toda inteligibilidad con SNR menor o igual a -10 dB [27,5,15].

En lo que a aceptabilidad se refiere, se sabe que las componentes de baja frecuencia (primer grupo formante) resultan ser de mayor importancia, contrario a lo que ocurre con la inteligibilidad, para la cual las altas frecuencias (segundo grupo formante) son perceptivamente más importantes. Tomando en cuenta lo que se ha mencionado hasta ahora, puede decirse que la aceptabilidad percibida depende en mayor grado de la definición de los sonidos sonoros que de los sonidos sordos, y que a su vez la inteligibilidad recibe mayor aportación de los sonidos sordos que de los sonidos sonoros, aún cuando estos últimos constituyan el mayor porcentaje de energía en la señal [20,23].

Otro aspecto perceptivo importante es la capacidad del sistema auditivo de enmascarar una señal con otra, fenómeno que ocurre cuando una señal audible llega a ser inaudible cuando otra señal de mayor intensidad ocurre simultáneamente. Modelos que intentan aprovechar esta característica se basan en la naturaleza selectiva del oído humano con respecto a las frecuencias percibidas, agrupadas en lo que se conoce como *bandas críticas*. En la Tabla 2.2 se enlistan 25 bandas críticas para el rango de frecuencias audibles, organizadas según la escala conocida como escala Bark[1]; puede observarse

[1] En memoria de Barkhausen, el científico que introdujo el *phon*.

que el oído humano tiene mayor capacidad discriminatoria para bajas frecuencias que para las altas.

Número de banda [Bark]	Límite inferior [Hz]	Centro [Hz]	Límite superior [Hz]
0	0	50	100
1	100	150	200
2	200	250	300
3	300	350	400
4	400	450	510
5	510	570	630
6	630	700	770
7	770	840	920
8	920	1000	1080
9	1080	1170	1270
10	1270	1370	1480
11	1480	1600	1720
12	1720	1850	2000
13	2000	2150	2320
14	2320	2500	2700
15	2700	2900	3150
16	3150	3400	3700
17	3700	4000	4400
18	4400	4800	5300
19	5300	5800	6400
20	6400	7000	7700
21	7700	8500	9500
22	9500	10500	12000
23	12000	13500	15500
24	15500	19500	

Tabla 2.2 Tabla de bandas críticas para el rango audible de 20 Hz a 20 kHz, como en [43,16].

Existen varias aproximaciones analíticas para expresar la relación entre frecuencias en Hz y Bark, entre las que se destacan

$$b = 13 angtan\left(\frac{0.76f}{1000}\right) + 3.5 angtan\left(\frac{f}{7500}\right)^2 \tag{2.21}$$

$$b = 7 angsenh\left(\frac{f}{650}\right) \tag{2.22}$$

Correcciones

$$b = \frac{26.81f}{1960+f} - 0.53 \qquad \begin{matrix} b < 2 & b' = b + 0.15(2-b) \\ b > 20.1 & b' = b + 0.22(b-20.1) \end{matrix} \tag{2.23}$$

donde f indica frecuencias en Hz y b frecuencias Bark, siendo (2.23) la más conveniente para el análisis de voz, debido a que su aproximación es la más cercana a la definición empírica de la escala; para la cual, según [14], el error cometido no supera 0.1 Bark.

Además de la existencia de bandas críticas, se debe considerar el hecho de que el oído humano presenta una variación en la intensidad percibida con respecto a la frecuencia

de la señal. Como puede observarse en la Figura 2.4, la relación que existe entre la intensidad[1] física del sonido (en dB) y la intensidad percibida (en *phons*) no es lineal [2].

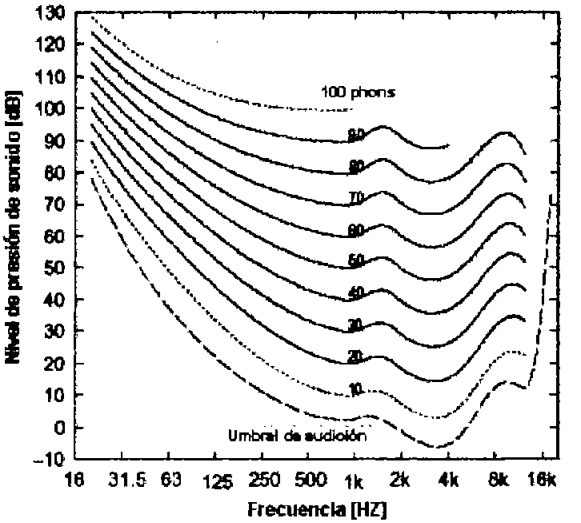


Figura 2.4 Curvas de equi-intensidad para tonos, según Suzuki en [35].

Las curvas de equi-intensidad[2], nombre que reciben las curvas mostradas en la Figura 2.4, sugieren que es más sencillo enmascarar componentes de alta frecuencia que de baja frecuencia, refiriéndose al nivel de intensidad requerido para enmascarar la señal. En lo que respecta a la percepción de la voz mezclada con ruido, el enmascaramiento de tonos por ruido de banda ancha resulta ser el caso de mayor relevancia. En este sentido, se sabe que el oído humano tiene mayor capacidad de percibir pequeñas variaciones en la amplitud de tonos (como es el caso del ruido de banda ancha) que tonos de nivel medio situados en la parte media de dicha banda. Para que lo contrario ocurriera se requeriría que la intensidad del tono mencionado fuera tan alta como para enmascarar la banda de ruido, 17 dB arriba de la intensidad del ruido si es blanco o 2 dB inferior si se trata de ruido de banda estrecha menor o igual a una banda crítica [24,43]. Este fenómeno es aprovechado para enmascarar tonos o ruido de banda estrecha cuya percepción es menos cómoda al oído que el caso de banda ancha.

[1] Según la gráfica mostrada, estrictamente debería hablarse de nivel de presión de sonido (SPL) y no de nivel de intensidad (IL). Aunque para condiciones atmosféricas normales de temperatura de 22 °C y presión de 0.751 mHg, puede considerarse que SPL = IL. SLP se define como SPL = 20 $\log_{10}$ (p/p_{ref}) e IL como IL = 10 $\log_{10}$ (I/I_{ref}).

[2] Las curvas de equi-intensidad también son conocidas como curvas de Fletcher-Munson por ser ellos los primeros en estudiar el fenómeno; sin embargo, se han realizado otros estudios que arrojan curvas con sutiles diferencias.

2.3 El ruido y sus características

Cuando se habla de ruido, tal término normalmente se asocia a algo negativo que implica efectos no deseados. En algunas ocasiones el termino ruido se asocia a procesos no determinísticos, y algunas otras simplemente a fenómenos no deseables que de alguna manera interfieren con la comunicación de una señal de interés. Algunos investigadores [32] sugieren que la mejor forma de definir al ruido es haciéndolo como todos aquellos datos o señales que no desean ser explicadas o tomadas en cuenta dentro del modelado de un sistema. El argumento para esta definición es que la asociación del concepto de ruido con procesos no determinísticos puede ser engañosa, pues el concepto de ruido es también usado para denotar a fenómenos determinísticos no explicados. Sin embargo, en lo que concierne a ruido acústico no existe tal determinación, así que las técnicas de reducción de ruido tienden considerar a éste como un proceso aleatorio que de alguna manera distorsiona la información contenida en la señal principal.

En otro sentido, el ruido en su carácter de señal y la diversidad de sus fuentes naturales justifican la extensa variedad de ruidos potenciales. Vaseghi, por ejemplo, clasifica al ruido en [36] según se muestra en la Tabla 2.3. Hay que notar que la existencia de varios tipos de ruido no implica que todos ellos estén involucrados en igual medida en la distorsión de alguna característica de una señal contaminada, dependerá de la aplicación y el tipo de caracterización que reciba la señal.

Ruido acústico	Proviene del movimiento, vibración o colisiones, siendo el más familiar debido ya que puede ser percibido por el oído humano. Generado por maquinaria en funcionamiento, voces de segundo plano, viento, lluvia, etc.
Ruido electromagnético	Presente en todas las frecuencias, en particular en las de radio. Todos los dispositivos eléctricos generan este tipo de ruido. Mientras mayor sea el nivel de voltaje o corriente generado, consumido o transmitido, mayor será el ruido inducido.
Ruido térmico	Generado por los movimientos aleatorios de partículas térmicamente energizadas. Mientras la temperatura de las partículas se incremente, su energía cinética aumentará y como consecuencia, la intensidad del ruido térmico también aumentará. Presente en circuitos eléctrico-electrónicos.
Ruido electrostático	Generado por la presencia de una diferencia de potencial eléctrico con o sin flujo eléctrico. La luz fluorescente es una fuente común de este tipo de ruido.
Distorsión de canal	Debida a las características no ideales de los canales de comunicación. Las distorsiones producidas son en magnitud y fase del espectro de la señal.
Ruido por procesamiento	Ruido que resulta de los procesos de acondicionamiento de la señal para poder ser almacenada, analizada o transmitida. Este tipo de ruido incluye distorsiones generadas por la amplificación, cuantización, codificación o compresión de la señal.

Tabla 2.3 Clasificación del ruido según su origen.

Como cualquier señal, el ruido lleva consigo información acerca de la fuente que lo produce, misma que de alguna manera se ve reflejada en las características de la señal

que lo representa tanto en tiempo como en frecuencia, tal como se muestra en la Tabla 2.4 [36].

Ruido blanco	Ruido teórico totalmente aleatorio que tiene un espectro de potencia uniformemente distribuido en todas las frecuencias.
Ruido blanco limitado en banda	Similar al ruido blanco puro, pero que usualmente se encuentra limitado por el espectro del dispositivo o señal de interés.
Ruido coloreado	Ruido que además de encontrarse limitado en una banda de frecuencia, su espectro de potencia no es plano, posiblemente ruido blanco modificado por el espectro del canal o dispositivo.
Ruido impulsivo	Consiste en pulsos aleatorios pero corta duración y amplitud aleatoria. Por ejemplo, en el contexto de señales de audio, se considera que los pulsos de corta duración de hasta 3 milisegundos (240 muestras a 8 kHz) pueden considerarse como ruido impulsivo.
Ruido transitorio	Consiste de pulsos de larga duración relativa, generalmente seguidos de oscilaciones de baja frecuencia. El pulso inicial se debe a una interferencia impulsiva interna o externa, mientras que las oscilaciones siguientes se deben a la resonancia del canal de comunicación excitado por el pulso inicial.

Tabla 2.4 Clasificación del ruido según sus características en tiempo y frecuencia.

2.4 Modificación del espectro de la voz en presencia de ruido

Ya se ha visto que una señal, como la voz, se encuentra sujeta a distorsiones por ruido de muy variada naturaleza. Considerando el caso de una señal de voz dentro de un ambiente de ruido acústico se pueden distinguir dos principales tipos de distorsión, aún hablando de la misma fuente física de ruido. Uno de éstos corresponde a la distorsión por ruido de carácter aditivo, generado por la adición de ondas sonoras provenientes de maquinaria en funcionamiento, fenómenos naturales o voces en segundo plano. El otro tipo de distorsión se debe a un proceso de filtrado no conocido, en ocasiones llamado ruido convolutivo, el cual ocurre cuando existen fuentes de reverberación o eco. En la Figura 2.5 se presenta un modelo utilizado para el tratamiento de ruido aditivo y convolutivo, en el cual $s[t]$ representa la voz limpia, $n[t]$ el ruido aditivo, $h[t]$ la función de transferencia del filtro desconocido y $sn[t]$ la voz degradada, todas ellas tratadas como señales en tiempo discreto [34]. Cada uno de estos tipos de distorsión acústica tiene características específicas, por lo que se hace evidente que las técnicas generadas para su tratamiento tengan enfoques distintos. Según Lim y Oppenheim en [20], existe la posibilidad de generalizar el asunto por medio de una transformación homomórfica, convirtiendo las distorsiones por ruido convolutivo en distorsiones por ruido aditivo.

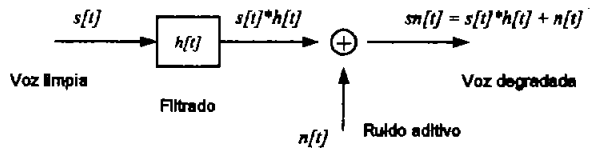


Figura 2.5 Modelo de distorsión por ruido acústico. *h[t]* representa la función de transferencia del canal de transmisión desde la fuente de voz hasta el micrófono.

En la Figura 2.6 c) y d) se muestran por separado los efectos causados por el ruido aditivo y por el ruido convolutivo respectivamente, tanto en la forma de onda como en el espectro de magnitud de la señal de voz. En a) se presenta la señal original sin distorsión y en b) una señal de ruido blanco aditivo. Debe notarse que el espectro resultante en c) equivale a la superposición de los espectros de voz y ruido. El efecto audible de esta operación resulta en el enmascaramiento de tonos de baja amplitud, principalmente de alta frecuencia. La explicación para el caso d) es más complicada debido a que la operación realizada por una fuente de eco sobre la señal es desconocida.

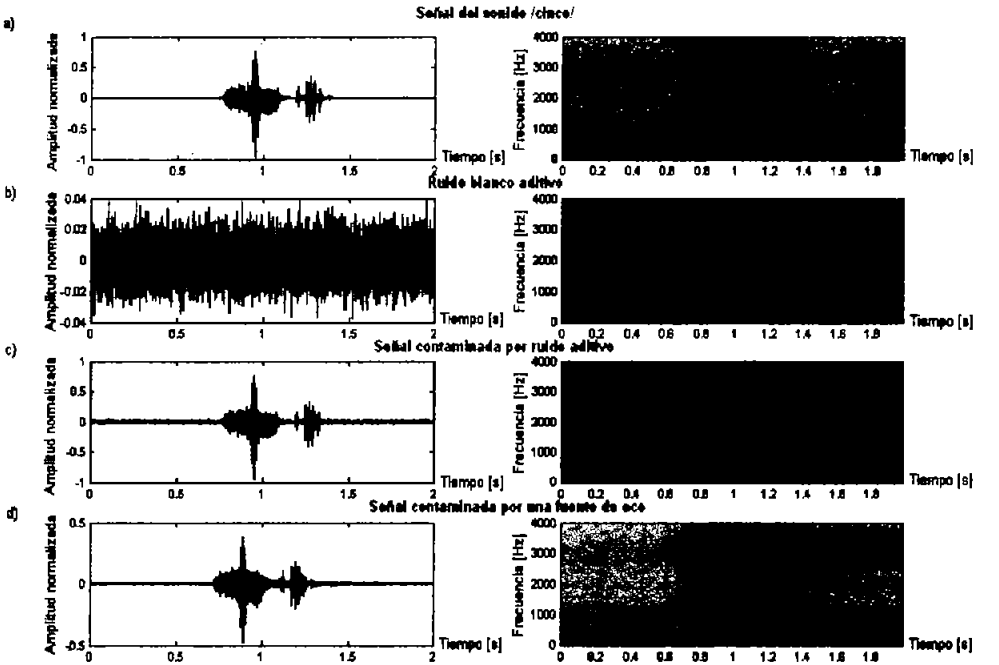


Figura 2.6 Efectos de distorsión acústica sobre una señal de voz en la forma de onda y en el espectro de magnitud.

Como hasta ahora se ha observado, el ruido tiene una gran variedad de representaciones que imposibilita un análisis y tratamiento generalizado para su reducción, es por esto que el problema comúnmente se divide en problemas menos complejos y más restringidos que se enfocan en la reducción de cierta clase de ruido. En esta tesis se analizan técnicas que se enfocan en la reducción de las distorsiones que sufre una señal de voz cuando se encuentra en un ambiente de ruido continuo no impulsivo, acústico, aditivo no estacionario y por lo tanto no ergódico. De esta forma, enfocándonos en las distorsiones por ruido aditivo, el modelo de distorsión por ruido acústico de la Figura 2.5 con un filtrado idealizado $|H[k]| = 1$, queda

$$sn[t] = s[t] + n[t] \tag{2.24}$$

Modelo que en la mayoría de las ocasiones es suficiente para obtener buenos resultados si no existen fuentes importantes de reverberación.

2.5 Modelos para la reducción de ruido en voz

A lo largo de la historia del procesamiento de voz se han generado por lo menos tres enfoques en el tratamiento del ruido [23,19,20], esquematizados en la Figura 2.7; por un lado, el que se inclina por la "restauración" de la forma de onda de la señal original, basándose principalmente en criterios matemáticos; el que sigue un modelo de producción de voz[1] para "reconstruir" la señal limpia a partir de la estimación de parámetros del modelo utilizado; y finalmente, el más reciente que opta por "realzar" la información contenida en la señal que la propia reducción del ruido, proceso realizado a través del uso de información relacionada con las características más importantes de la voz para la percepción humana. Cada uno de ellos dirigido a aplicaciones con ciertas condiciones.

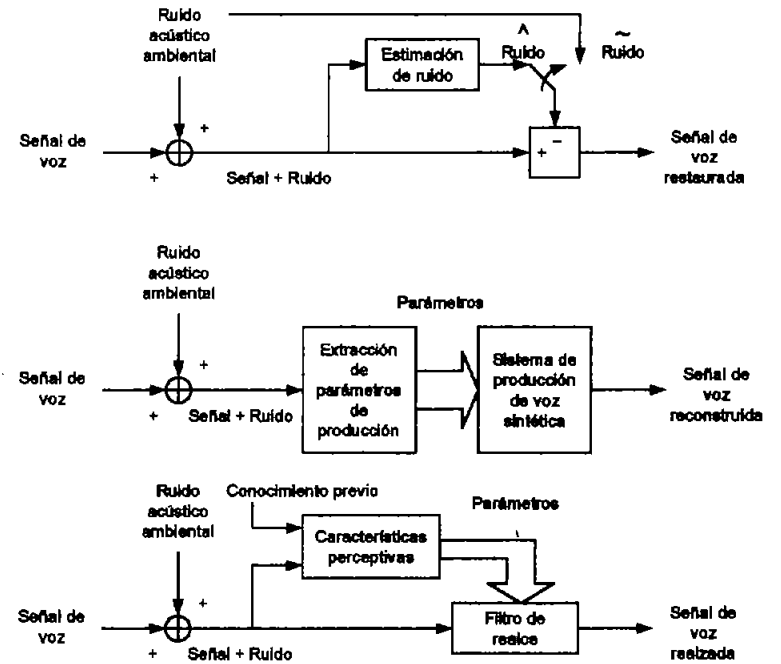


Figura 2.7 Enfoques para el tratamiento de ruido. a) Restauración a partir de la aplicación del proceso inverso de distorsión. b) Reconstrucción por medio de la extracción de parámetros característicos de la señal. c) Realce de los parámetros más importantes de la señal.

[1] Dado que la voz es una señal que depende directamente de la fuente que la produce, dichos modelos de producción también son utilizados para modelar la propia señal de voz en técnicas de realce.

El enfoque de restauración asume que la señal contaminada puede ser de alguna manera limpiada invirtiendo el proceso de degradación por medio del filtrado o sustracción de las componentes correspondientes al ruido. Por consiguiente, asume también que se posee o que es posible obtener información acerca de las características de la componente de ruido y de cómo se lleva a cabo el proceso de distorsión. Sin embargo, no siempre es posible lograr dichas condiciones, ya sea por el incremento en la cantidad de recursos necesarios o por la dificultad que representa la extracción de información directamente relacionada al ruido, cuando una señal degradada es la única señal disponible. A su vez, el enfoque de reconstrucción decide asumir que la parte importante de la señal degradada corresponde a la señal de voz, y que por lo tanto resulta más sencillo extraer información relacionada con ella. Una vez logrado esto, el objetivo es reproducir la señal de voz tal y como se produjo en el tracto vocal del emisor. Esto involucra la utilización de un modelo de producción de voz como el comúnmente utilizado [30,20] modelo LPC (Linear Prediction Coding) mostrado en la Figura 2.8. Este modelo aproxima el sistema generador de voz humano a través de un sistema lineal variante en el tiempo. Según el modelo, la fuente de excitación puede ser un tren de pulsos aproximadamente periódicos para los sonidos sonoros o ruido aleatorio para los sonidos sordos. La glotis, el tracto vocal y los efectos de los labios en la radiación del sonido son representados por un filtro digital variante en el tiempo, cuyos coeficientes corresponden a los coeficientes LPC de la voz. Estos coeficientes se relacionan directamente con las frecuencias formantes de la señal. Otros parámetros que deben ser estimados son la frecuencia fundamental de los sonidos sonoros, el parámetro de ganancia y la clasificación de sonidos sordos y sonoros.

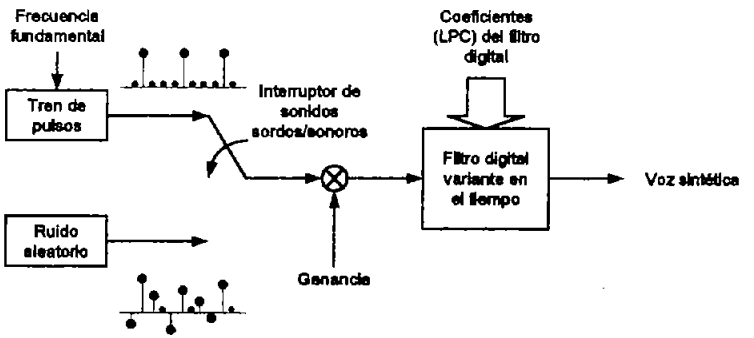


Figura 2.8 Modelo de producción de voz sintética basado en el modelo LPC.

El enfoque de realce intenta hacer uso eficiente del conocimiento de aspectos relacionados con la percepción y producción de la voz. Para este caso el objetivo es realzar las componentes de voz que se encuentran mezcladas con ruido, por lo que resulta necesario hacer uso de un modelo de voz que permita identificar dichas componentes. Dado que las características de una señal como la voz dependen directamente de la fuente que lo produce, dicho modelo de voz puede ser sustituido por el modelo del sistema de producción asociado. Como puede observarse, existe una estrecha relación entre este enfoque y el enfoque de reconstrucción; sin embargo, en el

enfoque de realce no se intenta de ningún modo sustituir la señal original, sino amplificar las componentes de voz con respecto al ruido. El enmascaramiento y filtrado paso banda son dos formas de realzar las componentes importantes para la percepción mencionadas en el apartado 2.2.4.

El enfoque de cada técnica puede estar totalmente definido por alguno de los enfoques mencionados, o bien, por alguna combinación de ellos, sobre todo cuando se trata de obtener provecho de los parámetros de producción o de los parámetros perceptivos de la voz. De igual forma, cabe destacar que algunas de las técnicas afiliadas al enfoque de restauración de la forma de onda pueden ser aplicables a problemas donde se involucren señales de otra naturaleza distinta a la voz, tal como se puede verificar en [40].

Además de los enfoques mencionados, el tratamiento del ruido en señales de voz puede ser caracterizado de varias maneras [11,23]: por las características del ruido; por el punto donde el ruido aparece dentro del sistema, en el emisor o en el receptor; por la forma en que éste interactúa con la señal de voz limpia, de forma aditiva o convolutiva; por el número de canales disponibles, y por la naturaleza del sistema demandante de una señal de voz sin distorsiones. Esta variedad de caracterizaciones ha generado una amplia variedad de modelos para reducción de ruido acústico, misma que se resume en la Tabla 2.5. Las técnicas que procesan la voz degradada para ser escuchada primordialmente por un oído humano son de interés específico en esta tesis, por lo que hay que notar que en la Tabla 2.5 no se enumeran técnicas específicas para cierto tipo de aplicaciones como compresión en sistemas de telefonía y reconocimiento automático de voz [1]. De la amplia variedad de técnicas para tratar el problema se han seleccionado para su análisis sólo las más representativas: *cancelación adaptable de interferencia* y *sustracción espectral*.

Basado en la periodicidad de la voz[1]			✓	✓				
Basado en la estimación de la amplitud espectral de las señales	✓	✓			✓			
Basado en modelos de percepción de la voz						✓		
Basado en modelos de producción de la voz							✓	
Basado en modelos estadísticos								✓
Técnicas	Filtrado de Wiener invariante en el tiempo	Filtrado adaptable o cancelación adap. de interferencia - LMS (Widrow)	Filtrado adaptable o cancelación adap. de interferencia -LMS (Sambur)	Filtrado dinámico selectivo - Comb Filtering (Shields y Frazier)	Sustracción espectral (Weiss y Boll)	Filtrado de pre-énfasis	Modelos paramétricos AR/Predicción lineal (Lim y Oppenheim)	Modelos basados en modelos ocultos de Markov
Enfoque								
Restauración	✓	✓	✓		✓			✓
Reconstrucción							✓	
Realce				✓	✓	✓		
Ruido								
Estacionario	✓							
Estacionario en tiempo corto		✓	✓	✓	✓		✓	✓
Previo conocimiento	✓					✓		
Voz								
Estacionaria	✓							
Estacionaria en tiempo corto	✓		✓	✓	✓		✓	✓
Ubicación								
Emisor	✓	✓	✓	✓	✓	✓	✓	✓
Receptor						✓		
Interferencia								
Aditiva	✓	✓	✓	✓	✓	✓	✓	✓
Convolutiva								
Número de canales								
1	✓		✓	✓	✓	✓	✓	✓
2		✓						
Naturaleza del receptor								
Humana	✓	✓	✓	✓	✓	✓	✓	✓
Máquina		✓	✓		✓		✓	✓

Tabla 2.5 Esquema general de técnicas para reducción de ruido acústico ambiental para voz.

[1] Basándose en la observación de las formas de onda de los sonidos sonoros.

Capítulo 3

MÉTRICAS DE CALIDAD DE VOZ

3.1 Inteligibilidad y aceptabilidad como dimensiones de calidad de voz

Como ya ha sido establecido con anterioridad, en esta tesis se analizaron técnicas para reducción de ruido acústico en señales de voz que serán escuchadas y evaluadas por un oído humano. Se hace evidente que las métricas que pretenden evaluar la calidad de voz estén dirigidas a predecir y remplazar el juicio humano sobre cómo la señal es percibida. Esto no implica que dichas técnicas no sean aplicables a sistemas artificiales de procesamiento que normalmente son menos tolerantes, esto dependerá tanto de la robustez del sistema con respecto al ruido como de la calidad de los resultados alcanzados por la técnica. Por ejemplo, aplicaciones de reconocimiento automático de voz requieren de una señal con una SNR superior a 20 dB, de otra forma el rendimiento del sistema se ve disminuido notablemente [22].

Como ya se mencionó en el apartado 2.2.4, la aceptabilidad e inteligibilidad son los dos elementos perceptivos principales en la determinación de la calidad de una señal de voz. La relación existente entre ambas dimensiones aún no se encuentra bien definida, aunque se sabe de cierto grado de independencia entre ellas. Es decir, una señal de voz puede ser altamente inteligible pero de poca aceptabilidad, como regularmente suena la voz sintética, o bien el caso menos común, ser de alta aceptabilidad pero poco inteligible. Este último caso representa la condición mínima para los sistemas donde la señal es evaluada por un oído humano, tal que la flexibilidad y eficacia del sistema humano de reconocimiento de voz disminuya la probabilidad de errores debidos al bajo nivel de inteligibilidad. Los sistemas de tal naturaleza, confiados en tal capacidad humana, utilizan técnicas para reducir ruido que eleven la calidad de voz en función de su grado de aceptabilidad más que de inteligibilidad.

3.2 Métricas subjetivas y objetivas de calidad de voz

La evaluación del grado de inteligibilidad y de aceptabilidad puede realizarse a través del juicio subjetivo emitido por un conjunto de oídos humanos, o bien, a través del cálculo directo y objetivo, a partir de la señal de voz, de una métrica fuertemente correlacionada con la percepción humana.

Por un lado, las métricas de carácter subjetivo son más representativas de la calidad de voz para un conjunto dado de sujetos de prueba, debido a que las mediciones son obtenidas directamente de los sujetos destino; sin embargo, su naturaleza subjetiva provoca que sean más susceptibles a variaciones en el conjunto de sujetos de prueba entre experimentos. La aplicación de métricas subjetivas requiere espacio y tiempo para calibrar el material de voz, como para entrenar a los sujetos de prueba, lo que evidentemente involucra mayor inversión de recursos que las métricas objetivas.

Por su parte, las métricas objetivas ofrecen una medida cuantitativa que puede ser directamente calculada de la señal (o señales) de voz, naturalmente el tiempo de evaluación es incomparable con el requerido por cualquier métrica subjetiva. Esta ventaja puede aprovecharse incluyendo un subsistema de medición automática de calidad en un sistema de procesamiento de voz, cuyos parámetros se optimen de acuerdo a la calidad de la señal resultante. El éxito de una métrica objetiva depende de su capacidad de predecir el resultado que arrojaría la experimentación con una métrica subjetiva. El problema con las métricas objetivas radica en la simplicidad de su manera de abordar el problema, pues no toman en cuenta la complejidad del proceso de percepción humana que incluye procesos de información de muy alto nivel como lenguaje, contexto y experiencia. Es por eso que no puede esperarse que las métricas objetivas mantengan el mismo rendimiento en sus predicciones para todo tipo de situaciones de prueba. Además, normalmente las métricas objetivas requieren de la utilización de una señal no distorsionada, como referencia en el cálculo de la distorsión.

En la actualidad no se ha logrado distinguir con precisión los elementos perceptivos de la voz que definen tanto a la inteligibilidad como a la aceptabilidad, razón por la cual no se puede asegurar independencia entre tales dimensiones. Una gran variedad de elementos perceptivos han sido propuestos, con los cuales se han diseñado técnicas para medir la calidad de voz. Tales técnicas se resumen en la Tabla 3.1 como métricas o pruebas, subjetivas y objetivas, para la estimación del grado de inteligibilidad, de aceptabilidad o calidad total de una señal de voz.

Métrica	**Dimensión de calidad**		
Métricas subjetivas	*Aceptabilidad*	*Inteligibilidad*	*Calidad total*
Prueba de articulación		✓	
Prueba modificada de rima (MRT)		✓	
Prueba de diagnóstico de rima (DRT)		✓	
Calificación promedio de opinión (MOS)	✓		
Método de grado apareado de aceptabilidad (PARM)	✓		
Prueba de grado de aceptación de calidad (QUART)	✓	✓	✓
Métrica de diagnóstico de aceptabilidad (DAM)	✓	✓	✓
Prueba de comunicabilidad	✓		
Métricas objetivas			
Índice de articulación (AI)		✓	
Índice de transmisión de voz (STI)		✓	
Relación señal a ruido (SNR)	✓		
Relación señal a ruido segmental (SNR_{seg})	✓		
Relación señal a ruido segmental por sub-bandas ($SNR_{fw\text{-}seg}$)	✓		
Métricas de distancia espectral	✓		
Métricas de distancia paramétrica basadas en LPC	✓		
Métrica de distancia espectral por ponderación de pendientes (WSSM)	✓		

Tabla 3.1 Métricas subjetivas y objetivas para evaluar la calidad de señales de voz.

Las pruebas de inteligibilidad son normalmente utilizadas en condiciones de mediana a alta distorsión, ya que para señales de baja distorsión la inteligibilidad se ve ligeramente afectada y las pruebas dejan de tener sentido. Por su parte, las pruebas de aceptabilidad tienen mejor resolución en señales de alta inteligibilidad.

Ya se ha mencionado que los resultados de una métrica subjetiva son más representativos de la calidad percibida, razón por la cual son considerados como patrones de comparación para valorar el rendimiento de las métricas objetivas. Un estudio realizado en [28] por Quackenbush, Barnwell y Clements muestra una comparativa entre diversas métricas objetivas y el patrón de calidad tomado de la métrica subjetiva conocida como *métrica de diagnóstico de aceptabilidad* (DAM). El análisis comparativo se realizó a partir de la regresión lineal entre datos objetivos y subjetivos. Para cada métrica objetiva se obtuvo un coeficiente de correlación, ρ_{est}, con los resultados de la DAM. Los resultados se resumen en la Tabla 3.2.

Métrica objetiva de calidad de voz	$\lvert\rho_{est}\rvert$
*Relación señal a ruido (SNR)	0.24
*Relación señal a ruido segmental (SNR_{seg})	0.77
*Relación señal a ruido por sub-bandas ($SNR_{fw\text{-}seg}$)	0.93
Métricas de distancia espectral	
Lineal	0.38
Logarítmica	0.60
No lineal	0.61
Métricas de distancia paramétrica basadas en LPC	
Lineal con los coeficientes del predictor	0.06
Logarítmica con los coeficientes del predictor	0.11
Lineal con los coeficientes PARCOR	0.46
Logarítmica con los coeficientes PARCOR	0.11
Razón lineal de área	0.24
Razón logarítmica de área	0.62
Itakura	0.59
Métrica de distancia espectral por ponderación de pendientes	0.74

Tabla 3.2 Comparativa del rendimiento de algunas métricas objetivas de calidad de voz, tomando como referencia la métrica de diagnóstico de aceptabilidad (DAM). *Correlaciones válidas únicamente para la evaluación de distorsiones de carácter aditivo en la forma de onda.

Las métricas listadas en la Tabla 3.2 pueden ser divididas en dos grupos, distinguibles por el dominio que toman para caracterizar la calidad de la señal. Por un lado, SNR y SNR_{seg} son métricas para el dominio del tiempo, el resto corresponde a métricas para el dominio de la frecuencia. Aún cuando estas últimas resultan ser más confiables por ser menos sensibles a diferencias de fase entre las señales de comparación, no están directamente ligadas a la percepción humana.

En los últimos diez años se ha observado un esfuerzo por desarrollar nuevas métricas objetivas basadas en modelos de percepción acústica, más que en modelos de producción de voz. Estas métricas, del dominio perceptivo, intentan aprovechar el conocimiento de los aspectos perceptivos de la voz mencionados en el apartado 2.2.4.

En la Tabla 3.3 se enumeran algunas de las métricas objetivas desarrolladas para el dominio perceptivo [42]. Hasta este momento, no existe un estudio comparativo completo del rendimiento de esta clase de métricas que sugiera la aplicación de alguna

de ellas, posiblemente por la diversidad de aplicaciones para las que fueron desarrolladas. De ellas, la más reconocida es la *métrica de calidad perceptiva de voz* (PSQM), una métrica diseñada para evaluar distorsiones debidas al proceso de codificación, cuyo desempeño ha sido reconocido por la ITU como recomendación P.861. A excepción de la *distorsión modificada del espectro Bark* (MBSD), el rendimiento del resto de las métricas que evalúan un campo más amplio de distorsiones depende de un extenso proceso entrenamiento. La MBSD ha sido estudiada en varias investigaciones en las cuales ha demostrado una alta correlación con la calidad de voz percibida, comparable con PSQM y para un rango más amplio de distorsiones. La extensión a la distorsión modificada del espectro Bark (EMBSD) supera los resultados de la MBSD, sin embargo, requiere de mayor capacidad de procesamiento. Por tales razones, se considera conveniente que en esta investigación se utilice la MBSD como métrica objetiva de calidad.

Métrica objetiva de dominio perceptivo	Año de publicación
Distorsión del Espectro Bark (BSD)	1992
Métrica de Calidad Perceptiva de Voz (PSQM)	1994
PSQM+	1997
Medición por Bloques Normalizantes (MNB)	1997
Sistema de Medición por Análisis Perceptivo (PAMS)	1998
Qvoice	1993
Determinación de Calidad Objetiva de Voz para Telecom. (TOSQA)	1997
Distorsión modificada del espectro Bark (MBSD)	1997
Extensión a la distorsión modificada del espectro Bark (EMBSD)	1999

Tabla 3.3 Métricas objetivas del dominio perceptivo.

3.3 Protocolo modificado de la métrica de diagnóstico de aceptabilidad

La DAM, descrita por William Voiers en 1977, forma parte de un conjunto de pruebas que realizan una evaluación indirecta de calidad de voz. La evaluación no se consigue a partir del juicio directo sobre la calidad total percibida, sino del juicio independiente de inteligibilidad y de aceptabilidad. La aceptabilidad, además de ser evaluada por preferencia directa del escucha, es evaluada a partir del juicio independiente de distorsiones específicas. Cada una de estas mediciones se realiza con base en escalas paramétricas y metamétricas de calidad, que finalmente generan una calificación representativa de la calidad total de voz. Esta métrica ha llegado a ser una de las más importantes por su índice de confiabilidad 0.96 y la amplia variedad de distorsiones que valora. Sin embargo, el análisis de esta investigación se interesó únicamente en la evaluación de distorsiones por ruido acústico aditivo, que interpretando los resultados en [28], forman un subconjunto de las escalas de la DAM. Además, el protocolo original de aplicación de la prueba está diseñado para evaluar hasta veinticuatro sistemas de codificación o comunicación de voz. Por tales razones, en este trabajo se optó por diseñar un protocolo modificado de la DAM (PMDAM) para evaluar las técnicas presentadas, siguiendo el concepto de parametrización y escalas de la DAM. Las escalas utilizadas en este protocolo fueron diseñadas como en los métodos directos de evaluación (*calificación promedio de opinión*, MOS).

En la Tabla 3.4 se presenta el conjunto de escalas propuestas para el PMDAM; seis escalas paramétricas, tres para distorsiones de los sonidos de ambiente y tres para distorsiones únicamente de la señal de voz; dos escalas que conjuntan las anteriores, una para la calidad de la voz y otra para la del ambiente (TSQ y TSB); tres escalas metamétricas, una para el juicio directo de inteligibilidad (I), otra para el de aceptabilidad (A) y una de aceptabilidad compuesta por las calidades totales del ambiente y la voz (AC); y finalmente una escala isométrica total compuesta por las escalas metamétricas (CT).

Descriptor del sonido	Nombre	Interpretación
	Escalas paramétricas	
Señal de voz		
Tenue, distante	SH	Filtro paso altas
Áspero, crujiente	SD	Centro o picos recortados
Amortiguado, suavizado	SL	Filtro paso bajas
Calidad de voz	TSQ	f(SH,SD,SL)
Ambiente		
Seseante	BN	Ruido blanco
Chirriante	BF	Errores de banda estrecha, ruido musical
Retumbante	BR	Ruido de muy baja frecuencia
Calidad del ambiente	TBQ	f(BN,BB,BF,BR)
	Escalas metamétricas	
Aceptabilidad general	A	Naturalidad, escuchabilidad, fatiga
Aceptabilidad compuesta	AC	f(TSQ,TBQ)
Inteligibilidad general	I	Entendimiento, identificación de palabras
	Escala isométricas	
Calidad total	CT	f(A, AC, I)

Tabla 3.4 Escalas de calidad del PMDAM.

La Tabla 3.5 especifica la puntuación asignada a la calidad descrita por el sujeto en distintas escalas.

Calificaciones para escalas paramétricas

Puntos	*Descripción*
5	Imperceptible
4	Esfuerzo requerido para ser percibido
3	Perceptible sin llegar a ser molesto
2	Perceptivamente molesto pero aceptado si es necesario
1	Inaceptablemente molesto

Calificaciones para la escala de aceptabilidad general

Puntos	*Descripción*
5	Agradable
4	Naturalidad dudosa pero aceptable sin fatiga alguna
3	Ligeramente fatigante por falta de naturalidad poco aceptable
2	Nada natural, fatigante, y difícilmente aceptable
1	Totalmente inaceptable

Calificaciones para la escala de inteligibilidad general

Puntos	*Descripción*
5	Entendible con completa claridad
4	Muy poca dificultad para entender
3	Dificultad de entendimiento pero poco confundible
2	Contenido difícilmente entendible y altamente confundible
1	Imposible identificar fonemas del lenguaje

Tabla 3.5 Calificaciones para las escalas de calidad del PMDAM.

3.3.1 Procedimiento de evaluación

El procedimiento de evaluación consiste en recolectar la opinión de un conjunto de sujetos en relación con las características del material de prueba. El material de prueba corresponde a las señales de audio generadas por los sistemas de procesamiento basados en las técnicas analizadas.

Suponiendo que la opinión de los sujetos tiene una distribución de probabilidad normal (o de Gauss) y que el error máximo permitido es 0.5 del valor CT con una probabilidad de 0.9, entonces se requiere, en el peor de los casos, de un conjunto muestra de hasta aproximadamente 10 evaluaciones por cada señal. Este tamaño de muestra fue calculado a partir de la máxima desviación estándar de CT, misma que fue registrada para el caso de mayor distorsión. Como debe notarse, el número de pruebas necesarias es alto, sobre todo si se desean evaluar varias técnicas con distintas configuraciones y en niveles distintos de distorsión. Para agilizar este proceso se propuso el uso de la tecnología Web en la recolección de datos de opinión. En la Figura 3.1 se muestra el aspecto que tiene el sitio Web construido para este efecto. En este sitio, el usuario realiza la evaluación de cuatro sistemas de reducción de ruido con base en las ocho escalas directas de la Tabla 3.4.

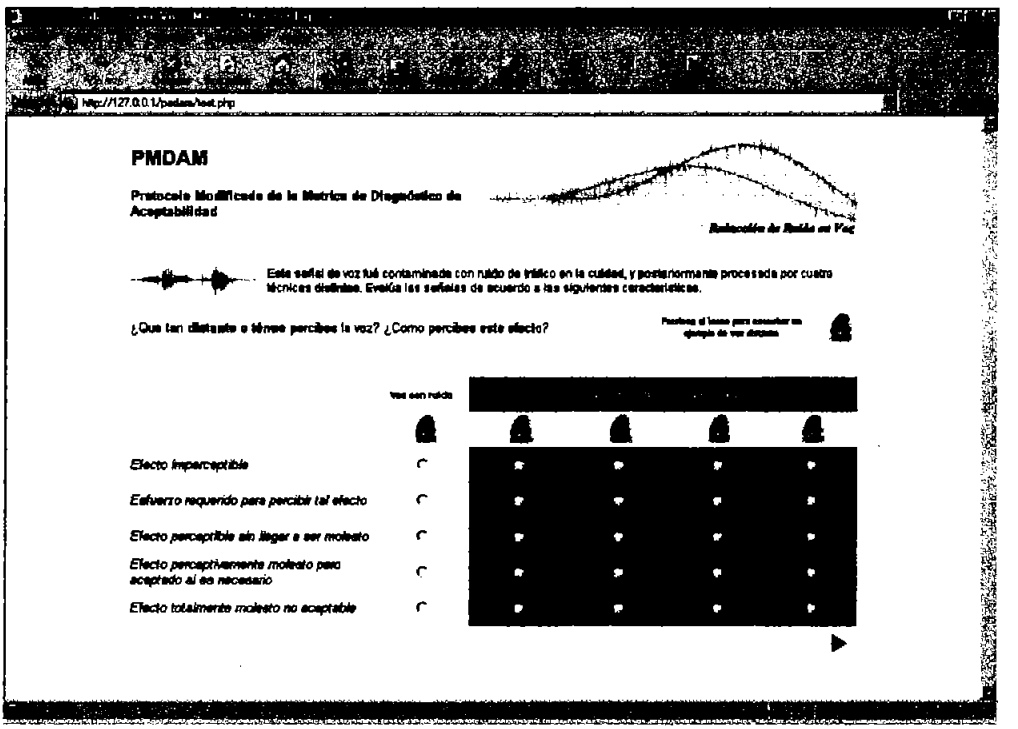


Figura 3.1 Evaluación PMDAM por medio de un sitio Web.

3.3.2 Cálculo de calificaciones

Sea p_{ij} el puntaje asignado por el sujeto j a la i-ésima escala paramétrica, y r_j el coeficiente de correlación entre la calificación del sujeto j y las calificaciones promedio para la escala i; entonces, la calificación obtenida de N sujetos de prueba en la escala paramétrica i, se calculó como

$$S_i = \frac{\sum_{j=1}^{N} r_j p_{ij}}{\sum_{j=1}^{N} r_j} \tag{3.1}$$

donde S_i, con i desde 1 y hasta 6, representan SH, SD, SL, BN, BF y BR.

De las seis escalas paramétricas se obtienen dos escalas más generales que describen la calidad de la voz y del ambiente. TSQ y TBQ fueron obtenidas como el promedio ponderado de las calificaciones en cada escala j, tal que

$$TSQ = w_1 SH + w_2 SD + w_3 SL \tag{3.2}$$

$$TBQ = w_4 BN + w_5 BF + w_6 BR \tag{3.3}$$

Los coeficientes de ponderación w_i fueron aproximados a partir de un análisis comparativo del nivel de percepción de cada escala, nivel de porcentaje que fue determinado por el juicio directo de un grupo reducido de sujetos de prueba. El valor de cada coeficiente fue ajustado de tal manera que la calidad total fuera lo más representativa de la opinión de los sujetos. Los valores obtenidos para tales coeficientes fueron: $w_1 = 0.3$, $w_2 = 0.4$, $w_3 = 0.3$, $w_4 = 0.45$, $w_5 = 0.45$, y $w_6 = 0.1$.

La aceptabilidad compuesta paramétricamente, AC, se obtuvo como el promedio ponderado de TSQ y TBQ, tal que

$$AC = 0.4TSQ + 0.6TBQ \tag{3.4}$$

La calidad total estimada se obtuvo a partir de AC y de las escalas metamétricas de aceptabilidad A y de inteligibilidad I; calificaciones obtenidas directamente del sujeto. Promediando los términos se obtuvo

$$CT = \frac{AC + A + I}{3} \tag{3.5}$$

Con valores en el intervalo de 1 a 5, según los valores establecidos en la Tabla 3.5. La interpretación práctica de este dato se encuentra en la Tabla 3.6.

Calificativo de calidad total de voz	
Intervalo	*Descripción*
4-5	Buena
3-4	Suficiente
2-3	Deficiente
1-2	Mala

Tabla 3.6 Interpretación práctica de la escala PMDAM

La importancia de la DAM y de esta variante es la independencia de la preferencia de los sujetos hacia un tipo específico de distorsión cuando éstos realizan su juicio sobre la calidad total de la señal (como en los métodos directos). Además, la granularidad del método permite identificar con mayor claridad la dimensión en que la voz requiere de mayor atención.

3.4 Evaluación objetiva

En esta investigación, la evaluación objetiva se realizó a partir de las métricas basadas en la *relación señal a ruido* (SNR), SNR *segmental* (SNR_{seg}) y SNR *segmental por sub-bandas* ($SNR_{fw\text{-}seg}$), y de la *distorsión modificada del espectro Bark* (MBSD). Puede verse en la Tabla 3.2 que aunque las métricas basadas en la SNR poseen los más altos niveles de correlación con las métricas subjetivas, tal correlación puede asegurarse únicamente cuando se evalúan distorsiones de la forma de onda. En general, las métricas SNR realizan un análisis numérico en función de características de la forma de onda, más que de características de importancia perceptiva de la señal. Las técnicas de reducción de ruido de los capítulos siguientes siguen principalmente el enfoque de restauración de la forma de onda, tal que las métricas SNR son completamente aplicables. Sin embargo, algunas de las variantes presentadas hacen uso de elementos perceptivos asociados con el enfoque de realce, esto obligó a hacer uso de métricas con representatividad de la calidad percibida y no de la forma de onda. Para estos casos se propuso el uso de dos métricas, MBSD y una variante de la SNR_{seg} con tendencias perceptivas conocida como $SNR_{fw\text{-}seg}$.

La SNR es una métrica basada en datos estadísticos de largo plazo, por lo tanto no puede ser aplicada directamente en la evaluación de señales cuyas características importantes se encuentran definidas en tiempos cortos. SNR_{seg} es una variante que proporciona una mejor estimación de calidad, debido a que el cálculo de la SNR se realiza en segmentos de tiempo corto con estadísticas aproximadamente constantes. $SNR_{fw\text{-}seg}$, variante de SNR_{seg} mejor correlacionada con la percepción, realiza el cálculo de la SNR fraccionando el espectro en bandas críticas de percepción del oído humano. Ambas métricas caracterizan muy bien la distorsión de la forma de onda, siempre y cuando las señales de prueba y de referencia se encuentren sincronizadas. MBSD, por su parte, hace una estimación de la calidad perceptiva calculando la distancia euclidiana perceptible entre los espectros de las señales de prueba y de referencia. Este cálculo toma en cuenta tres importantes aspectos de la percepción: la variabilidad de la intensidad percibida con la frecuencia, el enmascaramiento entre sonidos y un análisis de bandas críticas del oído humano.

3.4.1 Relación señal a ruido segmental

La definición de la SNR_{seg} parte de la definición de la SNR. Si se establece que para el caso de ruido acústico aditivo la SNR es el cociente de la energía de la señal limpia E_s y la energía del ruido persistente en la señal estimada $E_{s-\hat{s}}$, ambas con longitud T, tal que

$$SNR = 10\log\frac{E_s}{E_{s-\hat{s}}} = 10\log\frac{\sum_{t=0}^{T} s^2[t]}{\sum_{t=0}^{T}(s[t]-\hat{s}[t])^2} \tag{3.6}$$

donde $s[t]$ es la señal de voz limpia de referencia y $\hat{s}[t]$ es la estimación de la señal limpia, entonces, se define la SNR_{seg} como el promedio de las SNR's obtenidas en cada uno de los M segmentos de la señal, cada uno con longitud N de 15 a 30 ms, obteniéndose

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{E_s(m)}{E_{s-\hat{s}}(m)} = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1} s^2[n]}{\sum_{n=Nm}^{Nm+N-1}(s[n]-\hat{s}[n])^2} \tag{3.7}$$

El subíndice m indica el número de segmento sobre el cual se calcula cada SNR. Nótese que se realiza el promedio de valores en escala logarítmica y no en escala lineal [15,28]. Realizar el promedio de esta manera se justifica con una SNR_{seg} que considera igualmente importantes las SNR's de sonidos de baja y de alta amplitud.

3.4.2 Relación señal a ruido segmental por sub-bandas

De igual forma que SNR_{seg} divide el cálculo en regiones temporales, $SNR_{fw\text{-}seg}$ calcula la SNR de un segmento dividiéndolo en regiones espectrales. Normalmente se busca que las regiones espectrales coincidan con las bandas críticas mostradas en la Tabla 2.2. Así, la $SNR_{fw\text{-}seg}$ puede ser calculada promediando las SNR's de los M segmentos, las que a su vez equivalen al valor SNR promedio de J sub-bandas espectrales, es decir

$$SNR_{fw-seg} = \frac{1}{M}\sum_{m=0}^{M-1}\left(\frac{\sum_{j=1}^{J} w_j 10\log\left[\frac{E_s(m,j)}{E_{s-s}(m,j)}\right]}{\sum_{j=1}^{J} w_j}\right)$$

$$\frac{E_s(m,j)}{E_{s-s}(m,j)} = \frac{\sum_{\forall k,j} S^2[k]}{\sum_{\forall k,j}\left(S[k]-\hat{S}[k]\right)^2} \qquad (3.8)$$

Los términos $E_s(m,j)$ y $E_{s-s}(m,j)$ representan la energía contenida en la sub-banda j del segmento m de la señal limpia y el ruido persistente, respectivamente. Nótese que $S[k]$ corresponde a la representación en frecuencia de $s[t]$. El término w_j representa el factor de ponderación asignado a una banda durante el cálculo, este refinamiento permite ponderar la medida en cada sub-banda de acuerdo a su importancia perceptiva, *C-Message* (Laboratorios Bell) y *ponderación sofométrica* (CCITT) son funciones utilizadas en telefonía para asignar el valor de tales parámetros. En los experimentos que se presentan en este trabajo, se consideraron bandas que contribuyen equitativamente a la calidad percibida, así que $w_j = 1/J$, y (3.8) queda

$$SNR_{fw-seg} = \frac{1}{M}\sum_{m=0}^{M-1}\left(\frac{1}{J}\sum_{j=1}^{J} 10\log\left[\frac{E_s(m,j)}{E_{s-s}(m,j)}\right]\right)$$

$$\frac{E_s(m,j)}{E_{s-s}(m,j)} = \frac{\sum_{\forall k,j} S^2[k]}{\sum_{\forall k,j}\left(S[k]-\hat{S}[k]\right)^2} \qquad (3.9)$$

Para las evaluaciones realizadas con SNR_{seg} y $SNR_{fw\text{-}seg}$, basándose en lo mencionado en el apartado 2.2.4, se consideraron únicamente valores SNR contenidos en el intervalo de decibeles $-10 < SNR < 35$. De esta forma, la evaluación total es menos susceptible a sesgos por valores SNR muy altos o muy bajos poco representativos. La longitud del segmento utilizado en las pruebas realizadas es de 128 muestras equivalentes a 16 ms, traslapados temporalmente al 50% y espectros promediados con los factores $w_{s-1} = 0.25$, $w_s = 0.5$, $w_{s+1} = 0.25$.

3.4.3 Distorsión modificada del espectro Bark

La *distorsión modificada del espectro Bark* (MBSD) es una versión modificada de la métrica BSD propuesta en 1992 por Wang, Skey y Gersho. La característica sobresaliente de la MBSD sobre la *distorsión del espectro Bark* (BSD) es la incorporación del concepto de enmascaramiento entre sonidos y la utilización de la distancia euclidiana entre espectros, que según [42] es la más apropiada.

MBSD calcula la distorsión de una señal de prueba con respecto a otra de referencia, como la distancia euclidiana entre sus espectros de dominio en bandas críticas. Se hace uso del umbral de enmascaramiento de ruido para excluir del cálculo las componentes del espectro que no representan distorsiones perceptivas al oído. La señales se dividen en segmentos de 256 muestras traslapadas temporalmente al 50% y espectros promediados con los factores $w_{s-1} = 0.25$, $w_s = 0.5$, $w_{s+1} = 0.25$. El cálculo de la distorsión se realiza en tres pasos:

1) Cálculo del espectro perceptivo de la señal de referencia y la señal distorsionada

El procedimiento de cálculo del espectro perceptivo inicia aplicando un filtro de énfasis que aproxima el efecto de no linealidad de la sensitividad con la frecuencia, según la Figura 2.4. Este filtro se encuentra definido, como en [7], con la expresión

$$E[\omega] = \frac{\left(\omega^2 + 56.8\times10^6\right)\omega^4}{\left(\omega^2 + 6.3\times10^6\right)^2\left(\omega^2 + 0.38\times10^9\right)} \qquad (3.10)$$
$$\omega = 2\pi f$$

Enseguida se realiza un análisis en bandas críticas del espectro de energía de la señal de referencia. Con este análisis se obtiene un espectro con dominio en la escala Bark, tal que

$$B[b] = \sum_{k=fl}^{fu} P[k] \qquad (3.11)$$

donde $B[b]$ corresponde a la energía en la banda crítica b, $P[k]$ el espectro de energía en Hz y los términos fl y fu a los límites inferior y superior de la banda crítica b. Los límites de cada banda crítica pueden ser calculados a partir de (2.23). Las señales analizadas en este trabajo cubren un espectro de 18 bandas críticas, $0 \leq b \leq 17$.

Después del análisis anterior, es necesario aplicar una función de dispersión que estima los efectos de enmascaramiento en cada banda crítica del espectro. El espectro disperso $C[b]$ se obtiene convolucionando el espectro $B[b]$ con la función de dispersión $S[b]$, es decir,

$$C[b] = S[b] * B[b] \qquad (3.12)$$

La función de dispersión $S_{dB}[b]$ definida en decibeles, se expresa como en [4], por

$$S_{dB}[b] = 15.81 + 7.5(b + 0.474) - 17.5\sqrt{1 + (b + 0.474)^2} \qquad (3.13)$$

Esta función se encuentra definida para el dominio de bandas críticas $-25 \leq b \leq 25$. Las tres operaciones anteriores pueden ser resumidas en la Figura 3.2, donde se muestran las funciones de dispersión para cada banda crítica con énfasis de equi-intensidad.

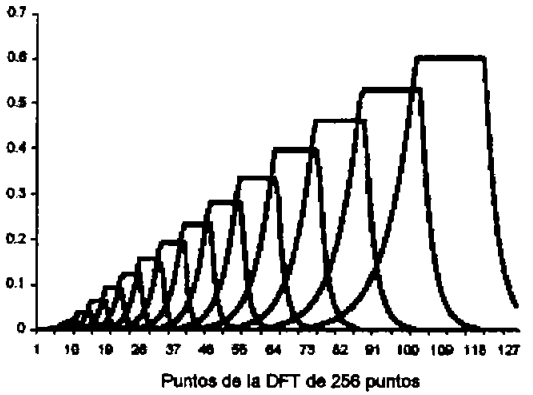


Figura 3.2 Banco de filtros aplicados al espectro de una señal para obtener un espectro perceptivo con dominio Bark, según [7].

2) Cálculo del umbral de enmascaramiento de ruido

El umbral de enmascaramiento, denotado como $Th_{dB}[b]$, se obtiene del espectro perceptivo de la señal de referencia. Este umbral es obtenido a partir de un desplazamiento u *offset* relativo al espectro de energía $C_{db}[b]$, offset cuya magnitud depende de la naturaleza tipo tono o tipo ruido de la señal de voz, tal que

$$Th_{dB}[b] = C_{dB}[b] - O_{dB}[b] \tag{3.14}$$

donde la magnitud en decibeles del offset, $O_{dB}[b]$, se determina a partir del coeficiente de tonalidad α como

$$O_{dB}[b] = \alpha(14.5+b) + (1-\alpha)5.5 \tag{3.15}$$

Con $\alpha = 1$ para voz totalmente de naturaleza tonal, y $\alpha = 0$ para voz totalmente tipo ruido. Este parámetro puede ser calculado para cada segmento, según [42], a partir de

$$\alpha = \min\left(\frac{10\log\frac{Gm}{Am}}{-60}, 1\right) \tag{3.16}$$

Los términos Gm y Am corresponden a las medias geométrica y aritmética del espectro de energía $P[k]$. De (3.14) se obtiene la magnitud lineal del umbral de enmascaramiento como

$$Th[b] = 10^{\log C[b] - \frac{O_{dB}[b]}{10}} \tag{3.17}$$

3) Cálculo de la métrica

Finalmente, el valor MBSD es obtenido como el promedio de la suma de diferencias perceptibles del espectro $C[b]$, en todos los segmentos temporales de la señal. Es decir,

$$MBSD = \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{b=0}^{17} I[m,b]D[m,b]\right] \tag{3.18}$$

donde el término $I[m,b]$ es un indicador binario de la perceptibilidad de la distorsión, $D[m,b]$, para el segmento m y la banda crítica b. El término $D[m,b]$ se define como la diferencia entre el espectro de la señal de referencia, $C_s[b]$, y el de la señal degradada, $C_{se}[b]$. El indicador $I[m,b]$ se obtiene de la siguiente manera

$$I[m,b] = \begin{cases} 1 & \text{si} \quad D[m,b] \geq Th[m,b] \\ 0 & \text{si} \quad D[m,b] < Th[m,b] \end{cases} \tag{3.19}$$

Capítulo 4

CANCELACIÓN ADAPTABLE DE INTERFERENCIA

4.1 Filtros adaptables

Un logro importante de la ingeniería es la automatización de procesos a través del diseño de sistemas capaces de ajustarse automáticamente a las condiciones cambiantes del entorno. Un ejemplo importante de mucho éxito en el desarrollo de sistemas de procesamiento de señales son los filtros adaptables, cuyas primeras aplicaciones en reducción de ruido fueron realizadas por Howells y Applebaum en 1960 para la General Electric Company. Los sistemas con filtros adaptables son sistemas que tienen una estructura no lineal variante en el tiempo, caracterizados por buscar continuamente el nivel óptimo de rendimiento según un criterio establecido. Una de las razones de éxito de los filtros adaptables es el notable incremento en el rendimiento con respecto a los sistemas no adaptables cuando las señales de entrada son variantes en el tiempo y la estructura del sistema no es lineal. Los filtros adaptables ofrecen la capacidad de ser utilizados en distintas configuraciones según la aplicación, entre las que destacan: predicción de señales, identificación o modelado de sistemas, filtrado inverso o deconvolución y cancelación de interferencia [40].

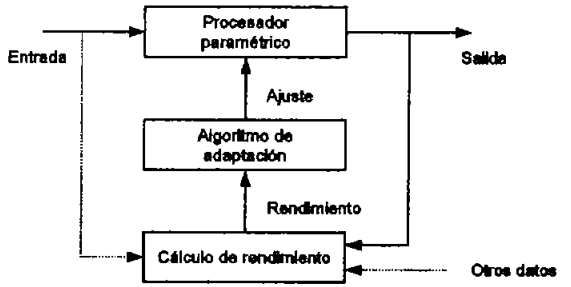


Figura 4.1 Modelo conceptual de un sistema adaptable de lazo cerrado.

Conceptualmente, un sistema adaptable de lazo cerrado puede presentarse como en la Figura 4.1, donde puede apreciarse un módulo principal de procesamiento dirigido por parámetros adaptables, un módulo encargado de adaptar dichos parámetros en función del rendimiento obtenido y un módulo destinado a calcular el rendimiento del sistema.

4.2 Cancelador adaptable de interferencia

La configuración utilizada en el tratamiento de ruido es la que se conoce como *cancelador adaptable de interferencia* (ANC). Como puede observarse en la Figura 4.2, la configuración para la cancelación de interferencia aplicada a voz y ruido acústico considera una fuente de voz, una fuente de ruido acústico y de manera implícita el

modelo de distorsión (2.24). El canal principal del ANC corresponde a sn, la señal de voz, s, contaminada aditivamente por una versión n' del ruido generado por la fuente de ruido. El canal de referencia, que en el esquema de la Figura 4.1 puede corresponder a "otros datos", es otra versión n de la señal emitida por la fuente de ruido. Se asume que las señales n' y n se encuentran correlacionadas de alguna forma, pero no correlacionadas con s, todas ellas con media igual a cero tal que

$$E[n'[t]n[t-\Delta]] = r_{n'[t]n[t]}[\Delta] \quad \text{para toda } \Delta \tag{4.1}$$

$$E[s[t]n'[t-\Delta]] = 0 \quad \text{para toda } \Delta \tag{4.2}$$

La razón de hablar en términos de versiones para señales de ruido radica en la necesidad de adquirir una señal de referencia que describa únicamente a la fuente de ruido. Este requerimiento implica una colocación física de los micrófonos tal que la señal de voz no interfiera en el canal de referencia, separados por una barrera acústica o simplemente alejados entre sí. Así es muy probable que la componente de ruido n' en el canal principal sea una versión convolucionada por h (normalmente el aire) de la señal n o viceversa, es decir

$$n'[t] = h[t]*n[t] \tag{4.3}$$

En la estructura interna del ANC se distinguen los módulos de procesamiento y adaptación de la Figura 4.2, como un solo *módulo de filtrado adaptable*.

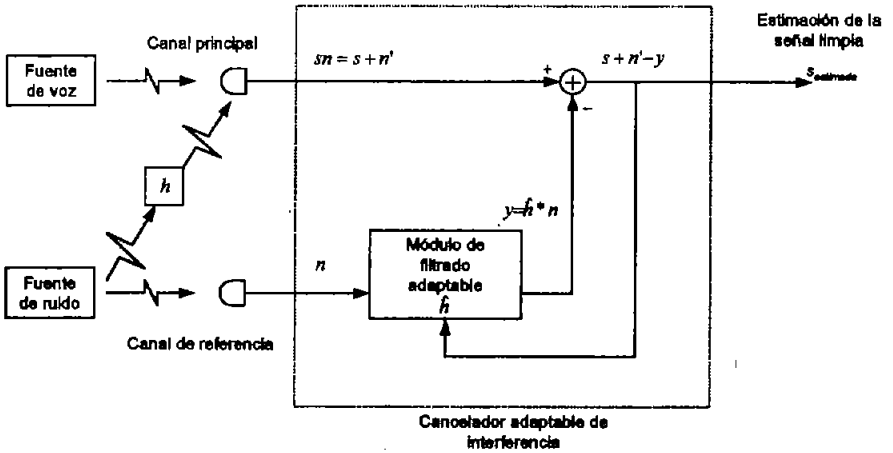


Figura 4.2 Configuración de un cancelador adaptable de interferencia.

El módulo de filtrado adaptable contiene un filtro por medio del cual se intenta estimar los efectos del canal h, para producir una salida y lo más parecida posible a n. Parte del cálculo de rendimiento del filtro se realiza sustrayendo y del canal principal para obtener una señal de error, ε, tal que

$$\begin{aligned}\varepsilon[t] &= sn[t] - y[t] \\ &= s[t] + n'[t] - y[t] \\ &= s[t] + \left(h[t] * n[t]\right) - \left(\hat{h}[t] * n[t]\right) \\ &= s[t] + \left(h[t] - \hat{h}[t]\right) * n[t]\end{aligned} \tag{4.4}$$

Esta señal es enviada al módulo de filtrado adaptable como primer indicador de su rendimiento. Este módulo, a través de un algoritmo de adaptación, modifica los parámetros del filtro interno de tal forma que la función criterio sea progresivamente minimizada. La función criterio seleccionada para regular el comportamiento de los filtros adaptables es el error cuadrático medio (MSE) ξ, que se define como

$$\xi = E\left[\varepsilon^2[t]\right] \tag{4.5}$$

Que equivale a la potencia de salida del sistema. Según se ve en el desarrollo (4.4), el mayor grado de minimización se logra cuando $\hat{h} = h$, lo que da lugar a que ε sea la mejor estimación de la señal limpia.

4.3 Estructura interna del módulo de filtrado adaptable

El módulo de filtrado adaptable del cancelador de interferencia de la Figura 4.2 puede descomponerse en dos principales elementos: un filtro adaptable o unidad de procesamiento y un módulo de adaptación encargado de modificar los parámetros del filtro.

4.3.1 Filtro Transversal Adaptable

La estructura base de procesamiento del filtro adaptable se conoce como *combinador lineal adaptable* [40]. De forma general, esta estructura se compone de un vector de entradas X de longitud L, un correspondiente vector de parámetros ajustables W, un sumador y una salida y. Como se observa en la Figura 4.3, la estructura es un *combinador lineal* porque asigna a su salida la combinación lineal de sus elementos de entrada de acuerdo al valor instantáneo del vector de parámetros. El adjetivo *adaptable* proviene de la disponibilidad de ajuste de tales parámetros.

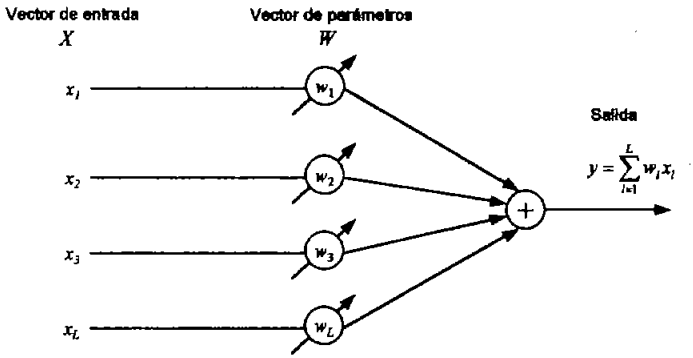


Figura 4.3 Estructura general de un combinador lineal adaptable.

Un filtro digital de una sola entrada que sigue el concepto del combinador lineal se conoce como *filtro transversal*. De esta manera, la estructura del filtro transversal adaptable tiene como única entrada un vector columna N_t de L muestras secuenciales de la señal de ruido en el canal de referencia, tal que

$$N_t = [n[t] \quad n[t-1] \quad n[t-2] \quad \cdots \quad n[t-L+1]]^T \tag{4.6}$$

Se asume que la distancia temporal entre cada muestra es constante y que equivale al periodo de muestreo de la señal según el teorema de muestreo. Se define al vector columna de la misma longitud

$$W_t = [w_{1t} \quad w_{2t} \quad n_{3t} \quad \cdots \quad n_{Lt}]^T \tag{4.7}$$

como vector de parámetros o de ponderación del filtro lineal, nótese que tales parámetros son variantes en el tiempo. Para tal situación, la respuesta del filtro y es el producto interno de N_t^T y W_t tal que

$$y[t] = \sum_{l=1}^{L} w_{lt} n[t-l+1] = W_t^T N_t = N_t^T W_t \tag{4.8}$$

Según (4.4) la señal de error ε puede escribirse en función de N_t y W_t como

$$\varepsilon[t] = sn[t] - N_t^T W_t \tag{4.9}$$

Calculando ε^2 se obtiene el valor cuadrático instantáneo de la señal de error del sistema

$$\varepsilon^2[t] = sn^2[t] + W_t^T N_t N_t^T W_t - 2sn[t] N_t^T W_t \tag{4.10}$$

4.3.2 Filtro óptimo de Wiener y su función criterio

El modelo del ANC parte directamente de un análisis presentado alrededor de 1940 por Norbert Wiener, cuyo problema era básicamente el caso estacionario del problema de filtrado que aquí se presenta. El problema de filtrado de Wiener consiste en diseñar un filtro de parámetros fijos que a partir de la señal estacionaria *n[t]*, genere una señal *y[t]* lo más parecida a una señal estacionaria deseada *sn[t]* tal que la diferencia entre ellas, referida como señal de error, observe su menor valor cuadrático medio posible. Para tal situación, la función criterio se obtiene como la esperanza de la versión estacionaria de (4.10), con W_t^T constante, tal que

$$\xi = E\left[\varepsilon^2[t]\right] = E\left[sn^2[t]\right] + W^T RW - 2P^T W \tag{4.11}$$

Donde R es la matriz de autocorrelación de n y P es el vector correlación cruzada de sn y n, que se definen como

$$\begin{aligned} R &= E\left[N_t N_t^T\right] \\ &= E\begin{bmatrix} n^2[t] & n[t]n[t-1] & n[t]n[t-2] & \cdots & n[t]n[t-L+1] \\ n[t-1]n[t] & n^2[t-1] & n[t-1]n[t-2] & \cdots & n[t-1]n[t-L+1] \\ n[t-2]n[t] & n[t-2]n[t-1] & n^2[t-2] & \cdots & n[t-2]n[t-L+1] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n[t-L+1]n[t] & n[t-L+1]n[t-1] & n[t-L+1]n[t-2] & \cdots & n^2[t-L+1] \end{bmatrix} \end{aligned} \tag{4.12}$$

$$\begin{aligned} P &= E\left[sn[t]N_t\right] \\ &= E\left[sn[t]n[t] \quad sn[t]n[t-1] \quad sn[t]n[t-2] \quad \cdots \quad sn[t]n[t-L+1]\right]^T \end{aligned} \tag{4.13}$$

Como puede observarse en la Figura 4.4, la función criterio ξ es una función cuadrática del vector de parámetros W, que describe un hiperparaboloide en el hiperplano de dimensión L (paraboloide para el caso de $L = 2$) cóncavo hacia valores positivos de ξ. El vector de parámetros W^* que define el punto mínimo de dicha superficie corresponde a la solución óptima de (4.11) y se obtiene igualando a cero el vector gradiente de ξ

$$\begin{aligned} \nabla(\xi) &= \left[\frac{\partial \xi}{\partial w_1} \quad \frac{\partial \xi}{\partial w_2} \quad \frac{\partial \xi}{\partial w_3} \quad \cdots \quad \frac{\partial \xi}{\partial w_L}\right]^T \\ &= 2RW - 2P \end{aligned} \tag{4.14}$$

si R es invertible,

$$\begin{aligned} 2RW^* - 2P &= 0 \\ RW^* &= P \\ W^* &= R^{-1}P \end{aligned} \tag{4.15}$$

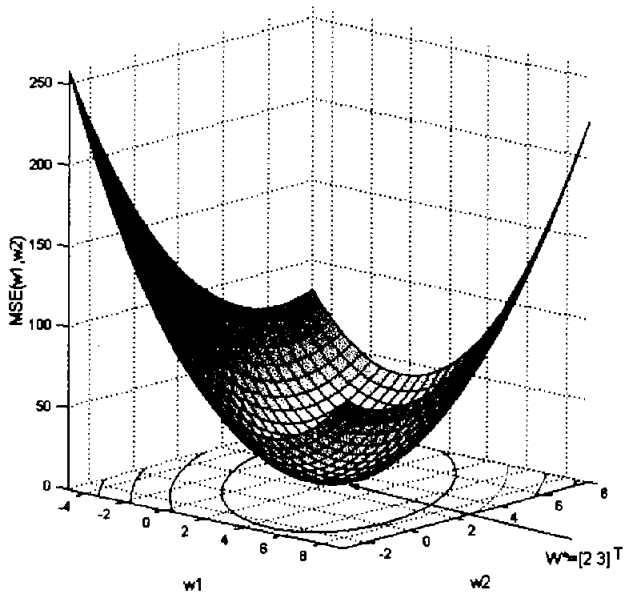


Figura 4.4 Ejemplo de la función criterio ξ para el caso estacionario donde $L = 2$, $W^* = [2\ 3]^T$.

Donde $RW^* = P$ es la versión realizable en forma matricial de la ecuación de Wiener-Hopf. El vector W^* obtenido en (4.15) se conoce como la solución restringida de Wiener, solución que define el filtro óptimo en términos de valores cuadráticos mínimos [40,38]. Sin embargo, en el contexto de voz y ruido acústico, los parámetros W^* no pueden ser directamente calculados por las razones siguientes:

a) Las señales de voz y de ruido no presentan un comportamiento estrictamente estacionario, así que R y P son variantes en el tiempo.

b) Las estadísticas R y P generalmente no son conocidas y requieren de una considerable cantidad de recursos y tiempo para ser estimadas.

c) La solución (4.15) requiere que la matriz R sea invertible, operación que requiere de más recursos.

Debido a tales circunstancias se ha optado por incluir un *módulo de adaptación* al modelo de filtrado óptimo de Wiener, con la finalidad de buscar continuamente la solución óptima para las circunstancias que se presenten. En la Figura 4.5 se muestra la relación que existe entre el módulo de adaptación y el filtro transversal adaptable dentro del módulo de filtrado en el ANC. Nótese que el subíndice t señala la dependencia del tiempo para los elementos de N y W.

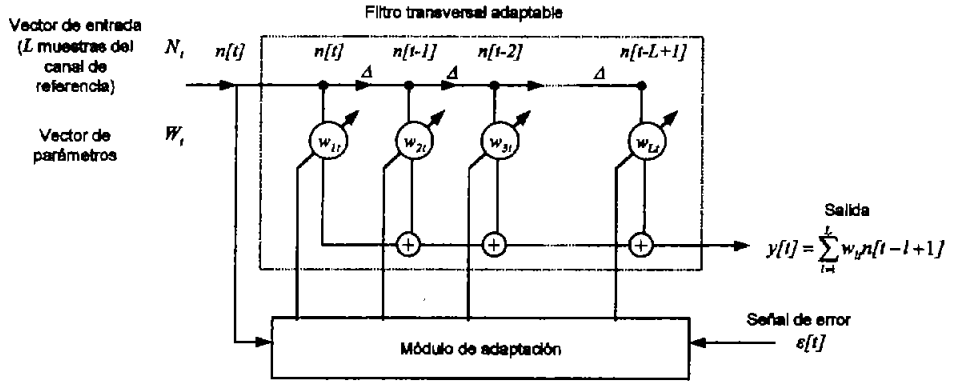


Figura 4.5 Estructura del módulo de filtrado adaptable. Se distingue un módulo de adaptación alimentado por la señal de error, la señal de referencia y posiblemente algún otro dato de conocimiento previo. En la parte superior se observa un filtro transversal adaptable con parámetros w_{lt} dependientes del módulo de adaptación.

Es importante notar que para un instante de tiempo t, el módulo de filtrado adaptable tiene una estructura lineal guiada por la estructura lineal del filtro. Sin embargo, más adelante se verá que los parámetros del filtro también son modificados constantemente por el módulo de adaptación en función no lineal de la entrada (valor cuadrático medio) y otros datos. De esta forma, la salida del filtro y del sistema será una función no lineal de la entrada, alcanzando una condición estable sólo si la entrada n llega a ser estacionaria y se alcance el filtro con parámetros óptimos.

4.3.3 Algoritmo de adaptación

El módulo de adaptación, ilustrado en la Figura 4.5, contiene la implementación de un algoritmo de aprendizaje o adaptación que, basado en información disponible y guiado por la minimización de ξ, actualiza los parámetros W_t del filtro adaptable convergiendo hacia la solución óptima W^*. Podría asegurarse que la esencia del cancelador adaptable de interferencia se concentra en este módulo, eficiencia de cuyo algoritmo depende la calidad de los resultados obtenidos, razón por la cual la mayoría de las propuestas para mejorar la técnica de cancelación adaptable se centran en el diseño del algoritmo de adaptación. Las variantes de cancelación adaptable analizadas en lo que resta del capítulo serán con respecto a este módulo.

Las soluciones propuestas hasta ahora consisten en algoritmos iterativos que, a través de la información que continuamente se recibe de n, obtienen la versión de W_t más cercana al óptimo W^*. Minimizar la cantidad de datos y operaciones utilizadas, y maximizar la calidad del resultado son objetivos perseguidos por tales técnicas. En esta tesis se analizaron los casos de dos técnicas conocidas como *técnicas de descenso*, reciben este nombre por utilizar estimaciones del gradiente de la función ξ para indicar

la dirección en la que se encuentra su valor mínimo ξ_{min}. Estas técnicas son *meno valor cuadrático medio* (LMS) y LMS-Newton (LMS-N). Tales técnicas se relacionan fuertemente en la manera de estimar el gradiente de ξ, sin embargo, cada una sigue un paradigma distinto para llegar a W^*. LMS por su parte se origina directamente de la técnica de *pasos de mayor descenso* (Steepest Descent), que tiene como objetivo cambiar el vector de parámetros W_t en dirección del gradiente negativo de ξ. Por otro lado, LMS-N parte de la técnica de Newton que busca cambiar el vector de parámetros W_t en dirección directa del punto mínimo de la función ξ. Estas técnicas resultan ser equivalentes únicamente en el caso en el que las curvas de nivel de $\xi(W)$ describen circunferencias perfectas a cualquier nivel arbitrario de ξ donde la función se encuentre definida [40].

Ambas técnicas, pasos de mayor descenso y Newton, han tenido importancia fundamental para el desarrollo de las técnicas que aquí se analizan, sin embargo, dependen fuertemente del conocimiento exacto tanto de $\nabla(\xi)$ como de R^{-1}, lo cual no permite su aplicación directa en la práctica.

4.4 Observaciones sobre el cancelador adaptable de interferencia

4.4.1 Garantía de máxima distorsión

La aparente sencillez de la estructura de la Figura 4.2 puede llegar a sugerir una idea engañosa. ¿Cuál es la necesidad de un filtro adaptable si se dispone de una señal de ruido n que puede ser restada directamente de la señal de voz contaminada sn'? A pesar de que tal procedimiento puede resultar exitoso en el caso especial de $|H[k]| = 1$, es mas probable que sea fuente de mayores distorsiones en la señal de voz ya que no se asegura la correspondencia entre el nivel de ruido presente en el canal principal y el que se resta con el canal de referencia. Por otro lado, el filtro adaptable tiene la ventaja de garantizar un nivel máximo de distorsión si por lo menos se satisface (4.2), este nivel de distorsión corresponde al que se presenta en la señal sin procesar cuando (4.1) no se cumple. Como ya se ha dicho, el objetivo es minimizar la potencia o el valor cuadrático medio de la señal de error resultante, tal que de (4.4) se obtiene

$$\begin{aligned} E[\varepsilon^2] &= E[(s+n'-y)^2] \\ &= E[s^2]+E[n'^2]+E[y^2]+2E[sn']-2E[sy]-2E[n'y] \end{aligned} \tag{4.16}$$

Partiendo de que s no se correlaciona con el ruido, si n' y n se correlacionan se cumple que $E[sn'] = 0$ y $E[sy] = 0$. Minimizando (4.16) para tales condiciones

$$E_{min}[\varepsilon^2] = E[s^2]+E_{min}[(n'-y)^2] \tag{4.17}$$

Así que minimizar la potencia total de salida equivale a generar la mejor estimación de s, en términos de valores cuadráticos mínimos. En cambio, si n' y n no se correlacionan, $E[sn'] = 0$, $E[sy] = 0$ y $E[n'y] = 0$. Para este caso, al minimizar (4.16) se obtiene

$$E_{\min}\left[\varepsilon^2\right] = E\left[s^2 + n'^2\right] + E_{\min}\left[y^2\right] \tag{4.18}$$

Minimizar (4.18) significa minimizar la potencia de salida del filtro, lo cual se logra a partir de una respuesta nula del filtro, $E[y^2] = 0$. La potencia obtenida para tal caso equivale a la potencia de la señal más la potencia del ruido inicial.

El peor caso ocurre cuando no puede lograrse que se cumpla (4.2), es decir, que en el canal de referencia exista cierto nivel de información relacionada con la señal de voz. Estas componentes causarán cierto grado de cancelación de la señal de voz dependiendo de su grado de intervención en el canal de referencia. Esto es

$$E_{\min}\left[\varepsilon^2\right] = E\left[s^2\right] + E_{\min}\left[(n'-y)^2\right] - 2E[sy] \tag{4.19}$$

Donde el término $2E[sy]$ corresponde a las componentes s filtradas en el canal de referencia.

4.4.2 Aproximación causal y de longitud finita del filtro de Wiener

La solución obtenida en (4.15) es una aproximación de la solución de Wiener en su forma no restringida, que idealmente asume la capacidad de construcción de un filtro no causal[1] y de longitud infinita, donde la solución óptima se obtiene de la versión no restringida de la ecuación de Wiener-Hopf definida como

$$\sum_{l=-\infty}^{\infty} w_l^* r_{n[l]n[l]}[\Delta - l] = r_{sn[l]n[l]}[\Delta] \tag{4.20}$$

En general, en aplicaciones de tiempo real no es posible conocer valores futuros de una señal, mucho menos si se trata de voz. De igual manera es imposible construir físicamente un filtro con un número infinito de parámetros. En este sentido, al utilizar la aproximación causal y de longitud finita (4.15) es necesario tomar ciertas consideraciones para alcanzar un rendimiento que tienda al ideal.

La longitud del filtro debe ser lo suficientemente grande para cubrir la respuesta al impulso del canal desconocido h. Evidentemente, mientras mayor sea la longitud del filtro su respuesta en frecuencia será mas cercana al filtro ideal, pero al mismo tiempo el costo computacional es mayor. En este sentido, cualquier filtro con un gran número de parámetros puede resolver el problema, sin embargo, es necesario considerar que a mayor resolución en frecuencia corresponde menor resolución en el tiempo. Para el

[1] No causal en el sentido que el conjunto de elementos muestra de la señal de entrada al filtro transversal, contiene elementos anteriores y elementos futuros al tiempo presente.

análisis de voz ambas características son importantes, por lo que la longitud del filtro debe diseñarse tal que ambas resoluciones sean menos afectadas. En las pruebas que se realizaron para esta investigación se observó una longitud apropiada de 64 parámetros, que además de la resolución tiempo-frecuencia, considera un tiempo razonable de procesamiento.

Suponiendo que la señal de ruido llega al mismo tiempo al canal principal y al de referencia, es necesario insertar un bloque de atraso de la señal del canal principal, de tal manera que la respuesta al impulso se encuentre centrada en la línea de atraso del filtro [38]. Esto se logra con un atraso igual a la mitad de la longitud total de la línea de atraso, es decir, la mitad de unidades de atraso que conforman la longitud del filtro. Por ejemplo, un filtro de longitud L = 64 (64 parámetros) requiere de una unidad de atraso de 32 muestras en el canal principal.

4.4.3 Solución sin canal de referencia

Característica esencial en la configuración que hasta ahora se ha presentado es la dependencia de un canal auxiliar que provea información de referencia. Una opción para disminuir el costo incurrido por esta dependencia fue propuesta por Marvin R. Sambur en 1977. Sambur propuso un sistema para cancelación de ruido en voz basado conceptualmente en el filtrado adaptable, en el que la señal de referencia se genera con base en la periodicidad de los sonidos sonoros. Sin embargo, resultados obtenidos en [20] revelan niveles bajos de rendimiento para este tipo de técnica, sobre todo en la que respecta a la inteligibilidad de la señal resultante. Este resultado se debe principalmente a que el método esta fuertemente basado en la supuesta periodicidad de la voz, suposición que no siempre se satisface.

4.5 Adaptación LMS

El algoritmo de adaptación por medio del menor valor cuadrático medio (LMS) se origina directamente de la técnica de pasos de mayor descenso (Steepest Descent). Este método es aplicado al problema de filtrado de Wiener cuando las condiciones del sistema varían lentamente, tal que sea posible rastrear la solución óptima W^* a partir del conocimiento aproximado de la función criterio. La mejor aproximación del algoritmo de pasos de mayor descenso se logra llevando el vector de parámetros W_t en dirección del gradiente negativo de la superficie cuadrática ξ, dirección que eventualmente apuntará al mínimo ξ_{min} si la superficie se mantiene estacionaria. Este procedimiento se puede ser expresado como

$$W_{t+1} = W_t + \mu(-\nabla(\xi,t)) \tag{4.21}$$

El algoritmo de pasos de mayor descenso requiere del conocimiento exacto del gradiente de la superficie de error para cada punto t, sin embargo, como se puede observar en (4.14), el cálculo del gradiente depende de las estadísticas no conocidas R

y P. Para solventar esta situación, LMS hace una estimación del gradiente a partir del valor instantáneo de ε^2 (y no de su valor medio), derivando el valor cuadrático de (4.9) respecto de W_t, el estimador queda

$$\hat{\nabla}(\xi,t) = -2\varepsilon[t]N_t \tag{4.22}$$

Nótese que ahora el estimador del gradiente depende explícitamente de la señal de error. Al considerar el estimador (4.22) se consigue reducir el costo computacional de calcular el gradiente exacto en cada paso, sin embargo, es claro que ahora el proceso de adaptación puede resultar en una trayectoria no óptima hacia W^* debido a la estimación imperfecta del gradiente. La adaptación del vector de parámetros W_t hacia el óptimo W^* se lleva acabo por medio del proceso iterativo descrito con las expresiones

$$\begin{cases} W_{t+1} = W_t + 2\mu\varepsilon[t]N_t \\ \varepsilon[t] = sn[t] - N_t^T W_t \end{cases} \tag{4.23}$$

El término μ, conocido como factor de convergencia, controla la magnitud de cambio del vector de parámetros, de su valor depende la convergencia del algoritmo y de la velocidad con la que ésta se lleve a cabo. El intervalo de μ que asegura la convergencia de (4.23) depende directamente del *valor propio* máximo de R, denotado como λ_{max}. Un intervalo más restringido, pero más sencillo de calcular, es el que depende de la potencia de n y de la longitud del filtro. Ambos intervalos se especifican en (4.24).

$$\begin{aligned} 0 < \mu < \frac{1}{\lambda_{max}} \\ 0 < \mu < \frac{1}{L \cdot E\left[n^2[t]\right]} \end{aligned} \tag{4.24}$$

4.6 Adaptación LMS-Newton

Así como LMS se deriva de pasos de mayor descenso, LMS-Newton (LMS-N) se deriva de la técnica de Newton. Este método busca cambiar el vector de parámetros W_t siempre en dirección del punto mínimo de la función ξ y no en dirección negativa de su gradiente. Para ello, la técnica de Newton deriva la solución óptima combinando (4.14) y (4.15), de donde se obtiene

$$\begin{aligned} W^* &= R^{-1}P \\ W^* &= R^{-1}\left(RW - \frac{1}{2}\nabla(\xi)\right) \\ W^* &= W - \frac{1}{2}R^{-1}\nabla(\xi) \end{aligned} \tag{4.25}$$

Si encontrar el vector de parámetros óptimo es el objetivo, evidentemente adaptar en dirección del óptimo implica obtener el mejor rendimiento en exactitud y velocidad de convergencia que pueda alcanzarse; en un solo paso según la solución (4.25), si se satisface completamente la cualidad estacionaria de ξ. Sin embargo, este procedimiento se vuelve ideal desde que asume el conocimiento de dos cantidades que normalmente son desconocidas y están en cambio constante, R y $\nabla(\xi)$. LMS-N reduce el nivel de idealización convirtiendo la solución (4.25) en una solución iterativa como en LMS, donde se usa el estimador (4.22) y se introduce un factor de convergencia μ con un intervalo de convergencia $0 < \mu < 1$. Debe mencionarse que se sigue asumiendo la cualidad estacionaria de n de tal manera que R sea representativa del punto de análisis y que R^{-1} pueda ser calculada. La solución LMS-N se expresa como

$$\begin{cases} W_{t+1} = W_t + 2\mu R^{-1}\varepsilon[t]N_t \\ \varepsilon[t] = sn[t] - N_t^T W_t \end{cases} \tag{4.26}$$

Con el fin de establecer condiciones de comparación entre LMS y LMS-N, se introduce un escalamiento del intervalo de μ tal que para ambos algoritmos se tenga el mismo intervalo de convergencia. Este escalamiento introduce un factor más a (4.26) tal que

$$\begin{cases} W_{t+1} = W_t + 2\mu\lambda_{av} R^{-1}\varepsilon[t]N_t \\ \varepsilon[t] = sn[t] - N_t^T W_t \end{cases} \tag{4.27}$$

Donde λ_{av} es el promedio de los valores propios de R. Como puede intuirse, LMS-N será superior a LMS en el grado que n sea estacionaria. Ambos algoritmos tendrán el mismo comportamiento mientras la dispersión de los valores propios de R sea mínima[1]; en este caso R tiende a ser una matriz diagonal y el producto $\lambda_{av}R^{-1}$ se aproxima a la matriz identidad, tal que (4.27) llega a ser idéntica a (4.23). Partiendo de una primera estimación de R^{-1} a partir de un entrenamiento previo, la matriz puede ser actualizada a partir del *lema de inversión de matrices*, como se desarrolla en [10], con la expresión

$$R_t^{-1} = \frac{1}{1-a}\left(R_{t-1}^{-1} - \frac{R_{t-1}^{-1} N_t N_t^T R_{t-1}^{-1}}{\frac{1-a}{a} + N_t^T R_{t-1}^{-1} N_t} \right) \tag{4.28}$$

El término a se conoce como parámetro de convergencia de R, con valores positivos cercanos pero distintos a 0. Con base en lo que se propone en [40] y según las pruebas

[1] El factor de dispersión se calcula como el cociente del máximo valor propio entre el mínimo, siendo la unidad el factor de dispersión mínimo. El factor de dispersión indica el grado de excentricidad de las curvas de nivel de ξ, el cual describe la distribución del espectro de la señal, factores pequeños de dispersión indican una distribución uniforme de las componentes en todo el espectro.

realizadas, se consideró conveniente que a sea calculada con base en la longitud del filtro, tal que

$$a = 1 - 2^{\left(\frac{1}{3L}\right)} \tag{4.29}$$

Siendo a = 0.0036, cuando L = 64.

4.7 Observaciones sobre el rendimiento de LMS y LMS-N

4.7.1 Condiciones iniciales de adaptación

El factor de convergencia utilizado en los algoritmos LMS y LMS-N es determinante para el comportamiento de los mismos. Valores cercanos al límite superior del intervalo (4.24) proveen rápida convergencia hacia la solución óptima, sin embargo, esta condición provoca que el MSE se mantenga oscilando sobre el mínimo sin alcanzarlo, debido a que la magnitud de los cambios realizados a W_i no permite mayor resolución en la solución. Por otro lado, valores pequeños en μ permiten mayor resolución y calidad de la solución, aunque generan una respuesta lenta, posiblemente insuficiente a las condiciones del sistema. La importancia del factor μ ha generado varias propuestas de algoritmos para establecer dinámicamente su valor en función de las condiciones del sistema, como ejemplos están [10,25]. En los experimentos de este trabajo se consideró la recomendación de Widrow en [40] sobre un factor de convergencia fijo, tal que

$$\mu = \frac{1}{10L \cdot E\left[n^2[i]\right]} \tag{4.29}$$

De esta forma se pretende alcanzar la máxima rapidez de convergencia posible, considerando que el error cometido en la estimación de las estadísticas de n puede ocasionar que el algoritmo no converja si se eligen valores cercanos al limite superior de (4.24).

Según (4.29), la determinación de μ requiere de una estimación de la potencia media de la señal de ruido. De igual forma, el algoritmo LMS-N requiere de una estimación inicial de la matriz R. Tales estimadores son calculados por medio de un entrenamiento previo. En [39] se sugiere un entrenamiento con un conjunto de muestras de tamaño igual a diez veces la longitud del filtro. Considerando L = 64, el entrenamiento realizado para las pruebas de esta técnica consta del análisis de 640 muestras de la señal de ruido.

A menos que se tenga idea de los valores aproximados del vector de parámetros óptimo, ambos algoritmos de adaptación inician con $W_0 = \bar{o}$. Se utilizó el valor fijo de

0.0036 para el parámetro de convergencia a, con el cual se obtuvieron los mejores resultados.

4.7.2 Velocidad y precisión en la adaptación

Ya se ha mencionado anteriormente que la función criterio ξ, expresión (4.11), corresponde a una función cuadrática de W, entonces, es posible representarla con la ecuación del hiperparaboloide

$$\xi = \xi_{min} + (W - W^*)^T R(W - W^*) \tag{4.30}$$

El término ξ_{min} corresponde al menor MSE alcanzado cuando $W = W^*$. Aplicando la transformación geométrica de traslación $V = W - W^*$ de tal forma que ξ quede expresada en un sistema de coordenadas con centro en W^*, (4.30) llega a ser

$$\xi = \xi_{min} + V^T R V \tag{4.31}$$

Y si además se aplica la rotación $V' = Q^T V$, para expresar ξ en el sistema de coordenadas principal (ejes principales de ξ), (4.31) queda

$$\xi = \xi_{min} + V'^T \Lambda V' \tag{4.32}$$

Donde Q y Λ corresponden a las matrices de vectores de valores propios y de valores propios de R, respectivamente. Como se verá enseguida, expresar en términos de las coordenadas V y V' permite analizar el comportamiento de W sin importar la posición de ξ con respecto al origen del sistema de coordenadas naturales.

Una manera de describir el comportamiento de los algoritmos LMS y LMS-N es observando el recorrido que sigue W_t hacia el vector óptimo W^*. Para ello, los algoritmos pueden ser expresados con las progresiones geométricas

LMS: $$V_t' = (I - 2\mu\Lambda)^t V_0' \tag{4.33}$$

LMS-N: $$V_t = (I - 2\mu\lambda_{av})^t V_0 \tag{4.34}$$

Las cuales proporcionan explícitamente el valor del vector de parámetros para la iteración t partiendo del valor inicial $V_0 = W_0 - W^*$. Estas progresiones son obtenidas en [40] por inducción matemática partiendo de (4.23) y (4.27). En (4.33) y (4.34) pueden observarse dos situaciones distintas: el algoritmo LMS proporciona una razón geométrica para cada componente del vector de parámetros V_t', en cambio, LMS-N muestra una única razón geométrica para todas las componentes V_t; es decir,

LMS: $$r = 1 - 2\mu\lambda_n \qquad 1 \geq n \geq L \tag{4.35}$$

LMS-N: $r = 1 - 2\mu\lambda_{av}$ (4.36)

El efecto transitorio producido en el MSE obtenido puede observarse graficando ξ a lo largo de t. Sustituyendo (4.33) y (4.34) en (4.32) y (4.31), respectivamente, se obtienen las expresiones para lo que se conoce como *curvas de aprendizaje* del proceso adaptable, esto es,

LMS: $\xi_t = \xi_{min} + V_0'^T (I - 2\mu\Lambda)^{2k} \Lambda V_0'$ (4.37)

LMS-N: $\xi_t = \xi_{min} + (I - 2\mu\lambda_{av})^{2k} V_0^T R V_0$ (4.38)

El proceso de adaptación normalmente inicia en un punto W_0 no óptimo y con un MSE excesivo. Este proceso esta guiado por dos objetivos principales: reducir lo mas pronto posible el MSE hacia el mínimo y mantener el MSE final ξ_{fin} cerca del mínimo ξ_{min} con la mayor precisión posible. Una manera de ver como LMS y LMS-N desarrollan el primer objetivo es calculando la constante de tiempo τ_{mse} de las aproximaciones exponenciales para las progresiones geométricas (4.37) y (4.38). La constante de tiempo puede interpretarse como el tiempo o número de iteraciones necesarias para alcanzar el 63.2% del valor final. La aproximaciones exponenciales de las progresiones de los vectores de parámetros y de la curva de aprendizaje pueden escribirse como

$$r = \exp\left(-\frac{1}{\tau}\right) \approx 1 - \frac{1}{\tau} \qquad \text{para valores } \tau > 6 \tag{4.39}$$

$$r_{mse} = r^2 = \exp\left(-\frac{2}{\tau}\right) = \exp\left(-\frac{1}{\tau_{mse}}\right) \tag{4.40}$$

Con las razones geométricas (4.35) y (4.36) sustituidas en las expresiones anteriores, se obtiene una aproximación de τ_{mse}, como

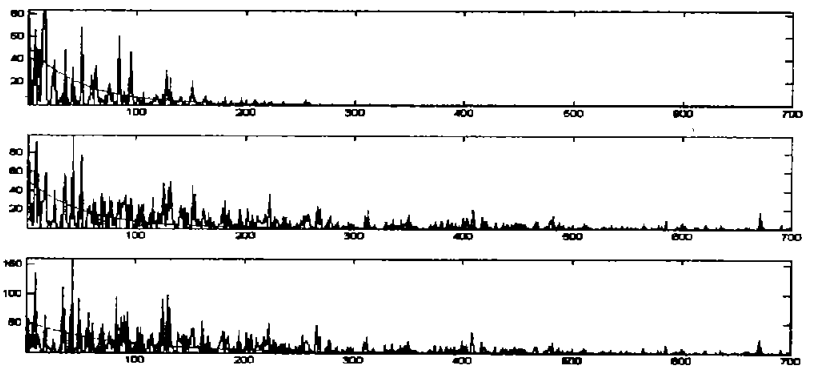
LMS: $(\tau_{mse})_n = \frac{1}{4\mu\lambda_n} \qquad 1 \geq n \geq L$ (4.41)

LMS-N: $\tau_{mse} = \frac{1}{4\mu\lambda_{av}}$ (4.42)

La curva de aprendizaje LMS corresponde a la suma de n progresiones exponenciales, cada una con su propia constante de tiempo según (4.41). Las L constantes de tiempo serán tan diferentes como tan alta sea la dispersión de los valores propios λ, dependiendo directamente de la distribución en frecuencia del ruido. Por otro lado, LMS-N posee una curva de aprendizaje aproximada a una exponencial con la constante de tiempo (4.42), la cual puede ser mayor o menor que la del *modo* más veloz o del más lento de la curva correspondiente a LMS. Donde *modo* se refiere a la curva de aprendizaje asociada a un parámetro w_l. Considerando que la convergencia del algoritmo se alcanza a cuatro veces la constante de tiempo, la curva de aprendizaje LMS resultante puede ser más veloz o más lenta que la correspondiente LMS-N, lo cual

depende de la dispersión de los valores propios y de las condiciones iniciales tomadas. En la Figura 4.6 a) y b) se muestran tres curvas de aprendizaje para LMS y LMS-N, respectivamente. Estas curvas fueron generadas partiendo de tres condiciones iniciales distintas, sobre el caso general en el que los valores propios de R se encuentran dispersos. Las condiciones iniciales corresponden a dos puntos localizados en ejes principales y otro en un punto intermedio, todos sobre la misma curva de nivel de ξ. Como puede observarse en las figuras, la velocidad de adaptación de LMS puede llegar a ser mayor o menor que la de LMS-N. Este comportamiento puede ser mejor entendido con la Figura 4.7 a) y b), donde se muestra el recorrido que sigue W_t sobre el plano w_1w_2, para los casos presentados en la Figura 4.6. El hecho de que LMS-N adapte el vector de parámetros en dirección del mínimo hace que su trayectoria sea más consistente que con LMS, cuya trayectoria siempre en dirección del gradiente negativo, ortogonal a la curva de nivel, no siempre apunta directamente al mínimo.

a)



b)

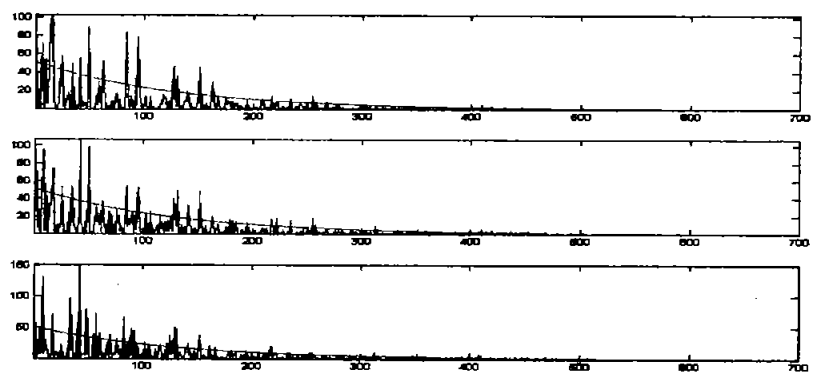


Figura 4.6 Curvas de aprendizaje para procesos de adaptación LMS (a) y LMS-N (b), con μ = 0.01, L = 20, $\xi_{inicial}$ = 50, para una señal de voz contaminada con ruido ambiental.

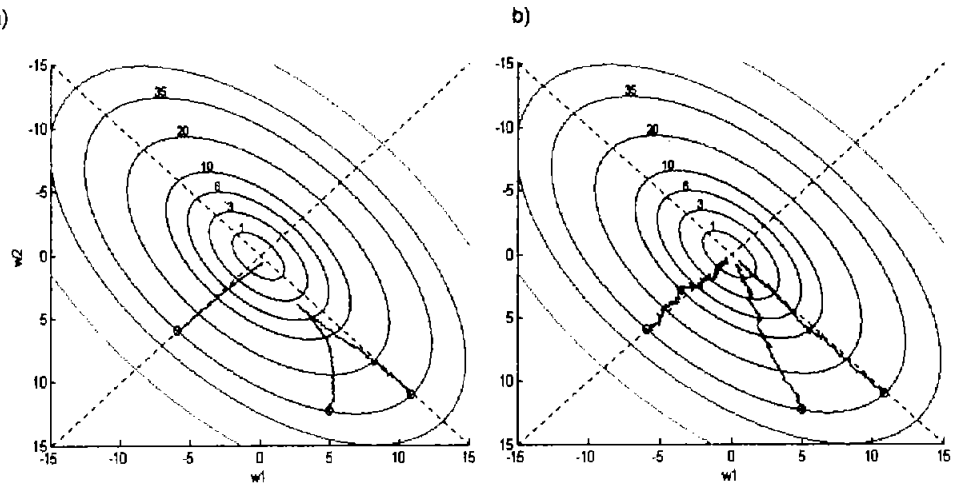


Figura 4.7 Comportamiento de los procesos de adaptación LMS(a) y LMS-N(b) para el caso general de λ_n dispersos, con μ = 0.01, L = 20, $\xi_{inicial}$ = 35.

Una manera común de analizar la precisión alcanzada con LMS y LMS-N es a través de lo que se conoce como *desajuste*. Ésta es una medida adimensional que describe la diferencia entre la aproximación adaptable y el filtro óptimo. El desajuste se encuentra definido como el cociente del excedente MSE entre el mínimo MSE cuando el proceso de adaptación ha llegado a un estado estable, tal que

$$M = \frac{MSE_{excedente}}{MSE_{min}} = \mu \cdot tr[R] \tag{4.43}$$

Siendo esta última una expresión válida para ambos algoritmos, su derivación puede encontrarse en [40]. La forma de esta expresión justifica lo que se menciona en el apartado 4.7.1, se alcanzará mayor precisión, desajuste menor, mientras más pequeño sea el tamaño de cada paso en el proceso de adaptación, tamaño que se encuentra determinado por el factor de convergencia μ.

Capítulo 5

SUSTRACCIÓN ESPECTRAL

5.1 Conceptos básicos de sustracción espectral

El análisis en frecuencia ha sido una herramienta matemática fundamental en el desarrollo histórico del procesamiento de señales. Sustracción espectral es una técnica que pretende hacer uso efectivo de las ventajas que ofrece una análisis de ésta naturaleza. En 1975, Weiss, Aschkenasy y Parsons publicaron el desarrollo de un sistema conocido como INTEL (INTelligibility Enhancement by Liftering), cuyo objetivo era proveer de mayor inteligibilidad a una señal de voz distorsionada por ruido ambiental. Originalmente, este sistema tuvo una visión en el dominio del tiempo basada en la estimación de la función de correlación en tiempo corto, que finalmente representa la versión generalizada de sustracción espectral. Sin embargo, fue en 1979 cuando Boll explicó el concepto base de esta técnica de manera comprensiva en el dominio de la frecuencia, resultando ser una de las técnicas más sencillas de entender y de implementar.

La técnica de *sustracción espectral* también hace uso del modelo de distorsión por ruido aditivo (2.24), sólo que esta vez llevando el modelo al espacio de la frecuencia a través de la transformada discreta de Fourier,

$$SN(\omega) = S(\omega) + N(\omega) \tag{5.1}$$

donde $SN(\omega)$, $S(\omega)$ y $N(\omega)$ corresponden a las transformadas discretas de Fourier de las señales discretas estacionarias $sn[t]$, $s[t]$ y $n[t]$, respectivamente. Sin embargo, este modelo no es completamente útil en casos prácticos de análisis por dos razones principales: 1) La transformada discreta de Fourier de una secuencia $x[t]$ se define como una sumatoria infinita de productos de $x[t]$ por una función exponencial compleja, sin embargo, en casos prácticos únicamente se cuenta con una secuencia finita de muestras de una señal digital. Además, la representación espectral de la voz tiene sentido únicamente en intervalos de comportamiento estacionario los cuales son relativamente cortos. 2) La variable argumento ω es una variable continua, por lo que no es posible calcular dicha función a través de un sistema de procesamiento digital.

Por las razones anteriores, lo conveniente es seleccionar una porción de la señal con dinámica lo más estacionaria posible, cuya descripción en el dominio de la frecuencia sea lo más representativa posible. Este procedimiento, conocido como *segmentación* o *ventaneo*, se realiza a partir de la multiplicación de una ventana $w[t]$ de longitud N por la señal de análisis, sea

$$x_w[t] = x[t]w[t] \tag{5.2}$$

donde *w[t]* se encuentra valuada en cero para todos los puntos fuera de la ventana. Ver referencias [8,12] para una mayor profundización en este proceso. Para este caso, el análisis en frecuencia se realiza a partir de la *transformada discreta de Fourier de N puntos* (DFTN), la cual permite tener una representación digital del espectro en frecuencia de una señal de longitud N, que corresponde efectivamente al fragmento de señal seleccionado. Luego entonces, el modelo de distorsión a utilizar queda

$$SN_W[k] = S_W[k] + N_W[k] \tag{5.3}$$

Conociendo las propiedades lineales de la DFTN, puede deducirse que el espectro de la señal de voz puede obtenerse a partir del modelo de sustracción espectral

$$S_W[k] = SN_W[k] - N_W[k] \tag{5.4}$$

Sin embargo, en el caso de sustracción espectral se parte del hecho que únicamente contamos con un canal de información, mismo que corresponde a la señal de voz contaminada por ruido. Esta limitante sugiere el cálculo de un estimador del espectro del ruido $N_W[k]$. En el caso específico de la voz, se considera que perceptivamente el espectro de fase en tiempo corto tiene muy poca relevancia en comparación con el espectro de magnitud [20], lo cual brinda cierta flexibilidad en el modelo de sustracción espectral. Así, el modelo de sustracción espectral se reduce a

$$\left|\hat{S}_W[k]\right| = \left|SN_W[k]\right| - \left|\hat{N}_W[k]\right| \tag{5.5}$$

Según Boll [5], el estimador del espectro de magnitud del ruido requerido por este modelo, puede ser calculado como la esperanza o media aritmética del espectro de $SN_W[k]$ durante intervalos donde la voz no se encuentre presente, esto es

$$\left|\hat{N}_W[k]\right| = E\left\{\left|N_W[k]\right|\right\} = \mu_{N_W}[k] = \mu_{SN_W}[k]\Big|_{\text{espacios sin presencia de voz}} \tag{5.6}$$

Una vez que se logra tener una estimación del espectro de magnitud de la señal de voz, finalmente puede procederse a dar una estimación de la señal de voz en el espacio del tiempo. Siguiendo con la noción de la poca importancia que representa la precisión en la estimación de la fase del espectro de la señal de voz, en el dominio del tiempo la estimación de la señal de voz limpia queda

$$\hat{s}_w[t] = DFTN^{-1}\left\{\left|\hat{S}_W[k]\right| \cdot \angle SN_W[k]\right\} \tag{5.7}$$

Una medida de rendimiento del modelo (5.5) se representa por el *error espectral* $|E_W[k]|$, que a partir de la magnitud de (5.4) queda como

$$\left|E_W[k]\right| = \left|\hat{S}_W[k]\right| - \left|S_W[k]\right| = \left|N_W[k]\right| - \mu_{N_W}[k] \tag{5.8}$$

Naturalmente, el objetivo es reducir lo más posible a cero el error cometido durante la sustracción espectral. En la Figura 5.1 se ilustra esquemáticamente el funcionamiento del modelo de sustracción espectral.

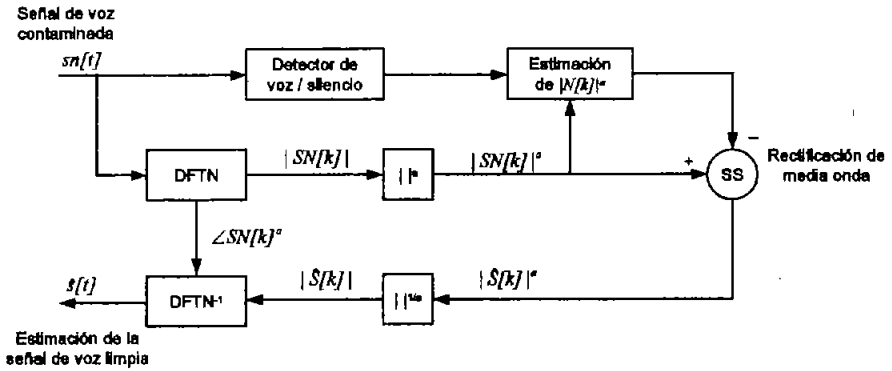


Figura 5.1 Modelo de sustracción espectral.

5.2 Observaciones sobre el rendimiento de sustracción espectral

5.2.1 Supuestos

El modelo de sustracción espectral (5.5) asume ciertas condiciones para alcanzar el mejor rendimiento. Primero, se considera que la voz y el ruido son señales no correlacionadas entre sí, ambas con comportamiento estacionario. Se espera entonces que los mejores resultados sean evaluados durante el cumplimiento de dichas condiciones, siendo la cualidad estacionaria la más difícil de alcanzar. Por lo pronto se sabe que la voz es aproximadamente estacionaria en intervalos de sonido sonoro. A pesar de que el ruido ambiental generalmente no mantiene un comportamiento estacionario esta técnica asume cierto grado de cualidad estacionaria, por lo menos lo suficiente para considerar que su espectro de magnitud se mantiene constante desde unos instantes previos a la presencia de voz y durante el tiempo que ésta se mantenga.

Se asume la capacidad de determinar un estimador del espectro de magnitud del ruido con un alto nivel de fidelidad con cada segmento procesado. Cualquier incompatibilidad entre el modelo y las condiciones reales puede verse reflejada en el aumento del error espectral (5.8), por ejemplo, si el carácter aleatorio del ruido aumenta el espectro real del ruido se ve más alejado de su estimador.

Además, se asume que el proceso de segmentación previo a la aplicación de sustracción espectral no representa fuente de distorsiones para la señal. Para que esto se logré es necesario considerar el traslape temporal de los segmentos de análisis.

5.2.2 Ruido residual

El error espectral puede presentarse como una inconsistencia en el modelo, o bien como *ruido residual*. Si $|E_W[k]| < 0$, el espectro real de magnitud del ruido es menor que su estimador, por lo que se genera una inconsistencia en (5.5) al resultar un espectro de magnitud negativa como estimador de la señal limpia. Por otra parte, $|E_W[k]| > 0$ indica que el espectro real de magnitud del ruido es mayor que su estimador, así que la estimación de la señal limpia mantendrá componentes de ruido que se encuentren sobre su media, o cualquiera que sea su estimador. Componentes que desde el punto de vista perceptivo se conocen como ruido residual, percibido principalmente durante la ausencia de actividad de voz o en regiones donde la voz no puede enmascararlo. Esto se debe a que el ruido, por su naturaleza aleatoria, puede presentar picos espectrales que no son eliminados cuando son restados por una versión suavizada (5.6) de tal espectro. En la Figura 5.2 b) se ilustra el espectro obtenido al aplicar el modelo (5.5) sobre una señal contaminada a). De esos picos espectrales residuales, los más estrechos y de mayor amplitud pueden distinguirse auditivamente como tonos musicales variantes, conocido como *ruido musical.*

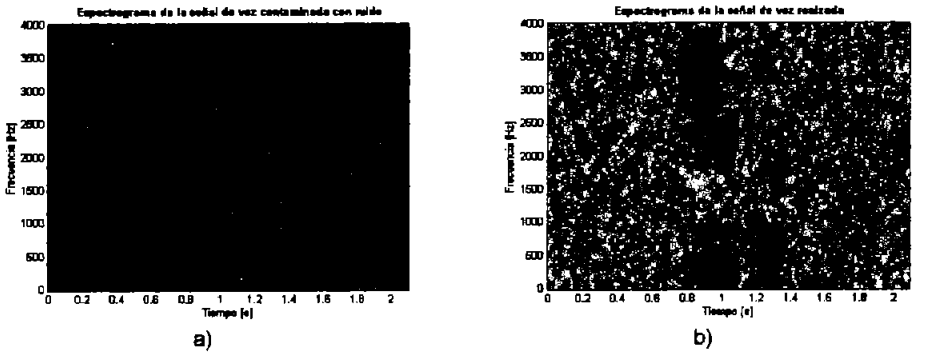


Figura 5.2 Ruido residual después de la sustracción espectral. a) Señal contaminada con ruido blanco a 5 dB. b) Señal estimada con el modelo de sustracción espectral.

5.2.3 Detector de voz

Dado que el ruido generalmente no es estacionario, se espera que su espectro en tiempo corto no sea fielmente aplicable a intervalos de tiempo siguientes. Es por eso que se requiere un detector de voz que permita identificar espacios de silencio (en lo que a voz se refiere) y poder actualizar cualquier cambio en el estimador del espectro de magnitud del ruido, previo a la siguiente actividad de voz. El problema de detección de voz en una señal ruidosa no es trivial, además de las complicaciones generadas por las características variantes de la voz se deben resolver aquellas relacionadas con el ruido. Primero, el detector de voz debe tomar en cuenta las características de la voz revisadas en el Capitulo 2: niveles de amplitud variantes en el tiempo, alta energía en

sonidos sonoros y baja en sonidos sordos, comportamiento estacionario únicamente en sonidos sonoros y en tiempos muy cortos, así como la existencia de características espectrales específicas de la voz de cada persona. Segundo, cuando se trabaja en ambientes reales, el detector de voz debe considerar que el ruido puede tener cambios en su espectro de magnitud, tanto en los niveles de amplitud como en la ubicación de sus componentes en frecuencia.

5.2.4 Segmentación

El proceso de segmentación tiene mucha importancia en la calidad de los resultados que se obtienen, resultados que están sujetos a la magnitud y forma en que el proceso distorsiona la información. A pesar de que la segmentación distorsiona la señal, éste es un proceso necesario para poder llevar a cabo el análisis en frecuencia de señales de longitud finita, que estrictamente hablando es imposible tratar [8,31]. Para que una técnica de procesamiento en el dominio de la frecuencia como sustracción espectral produzca los resultados esperados, el proceso de segmentación debe ser tal que si una señal definida en tiempo es segmentada, llevada a la frecuencia a través de una transformación y procesada por un sistema de función identidad, su correspondiente temporal sea lo más idéntica posible a la original, tal y como se presenta en la Figura 5.3.

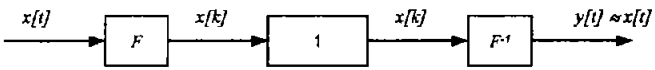


Figura 5.3 Proceso ideal de segmentación o ventaneo.

El tipo de ventana, su longitud y forma de aplicación son aspectos que influyen en gran medida en los resultados del proceso de segmentación. Cuando la forma de onda de una señal se multiplica por una ventana no rectangular, la longitud del intervalo de análisis efectivo es aproximadamente 40% menor que la longitud de la ventana, debido a que ambos extremos de la ventana son atenuados. Esta reducción en la longitud efectiva de análisis temporal resulta en una reducción del mismo grado en la resolución en frecuencia [12]. De esta forma, el proceso de segmentación de la Figura 5.4 a) provoca que la información contenida entre segmentos no tenga la mejor representación espectral posible, siendo más susceptible a perderse durante el procesamiento. Por otro lado, si el análisis se lleva a cabo con base en segmentos temporalmente traslapados es posible aumentar la longitud efectiva de análisis reducida por la aplicación de ventanas no rectangulares. Después de procesar cada uno de los segmentos, la señal sintética resultante se obtiene de la concatenación de cada uno de los segmentos y la suma de las componentes temporales traslapadas, Figura 5.4 b). La cantidad de traslape entre segmentos de análisis depende de la longitud efectiva del tipo de ventana utilizada, tomando en cuenta que la segmentación resultante del traslape y suma de ventanas debe aproximarse a una ventana rectangular unitaria para evitar mayores distorsiones en la forma de onda. Para el caso de ventanas Hanning y Hamming es común utilizar segmentos traslapados al 50%.

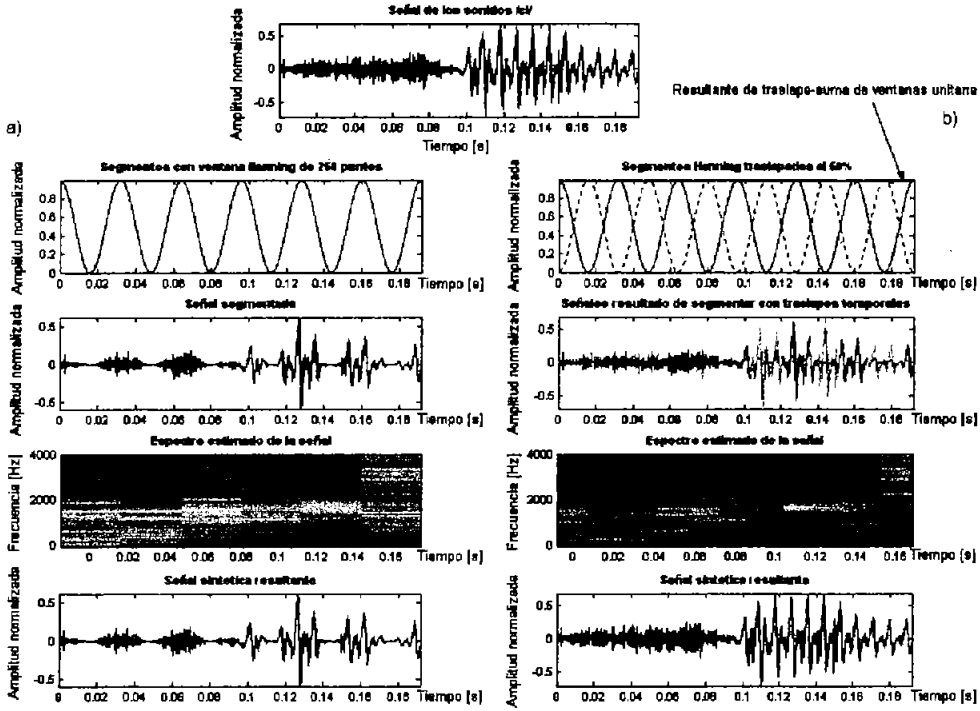


Figura 5.4 Proceso de segmentación de la señal de voz previa y posterior al procesamiento unitario. a) Segmentación de la señal por medio de la aplicación de ventanas adyacentes. b) Segmentación por medio de la aplicación de ventanas traslapadas temporalmente. Se observan mejores resultados para el segundo caso, cuya representación espectral se encuentra mejor definida por los cambios espectrales menos abruptos y más naturales.

Resultados obtenidos en [5] indican que para lograr una resolución espectral aceptable se requiere de una ventana lo suficientemente corta para ajustar en el intervalo temporal estacionario de la voz (hasta 40 ms) y lo suficientemente larga para cubrir dos veces el periodo fundamental más largo esperado (alrededor de 8 ms, periodo fundamental correspondiente a la voz masculina) [31,12]. Además, puede lograrse un menor tiempo en el cálculo del espectro si se utiliza el algoritmo de la *transformada rápida de Fourier* (FFT), en el cual se espera que la longitud de la ventana sea potencia de 2. Esto sugiere la elección de ventanas de 128 o 256 muestras, equivalentes a 16 y 32 ms respectivamente.

5.3 Extensiones del modelo de sustracción espectral

Una vez planteado el modelo y las condiciones requeridas para un buen desempeño, el objetivo es reducir al mínimo el nivel de ruido residual sobre la señal resultante, por lo que en la literatura se sugieren ciertos cambios al modelo inicial.

5.3.1 Rectificación de media onda

Una manera de eliminar la posibilidad de un espectro de magnitud negativo en (5.5) es aplicar una rectificación de media onda al espectro de magnitud estimado, lo que indica que cualquier componente que resulte negativa será sustituida por valor nulo. De esta forma el recorrido del error espectral (5.8) se reduce a cantidades positivas, tal que la versión rectificada del estimador de la señal limpia en (5.5) queda

$$|\hat{S}_W[k]| = \begin{cases} |SN_W[k]| - \mu_{N_W}[k] & si \quad |SN_W[k]| > \mu_{N_W}[k] \\ 0 & si \quad |SN_W[k]| \leq \mu_{N_W}[k] \end{cases} \tag{5.9}$$

La desventaja de rectificar el espectro de esta forma es la incorrecta eliminación de componentes espectrales de voz de baja amplitud, que ocurre cuando la magnitud del espectro de la señal de voz contaminada es menor que el espectro promedio del ruido. Por otro lado, este modelo no intenta por ningún modo reducir los efectos audibles del ruido residual y el ruido musical.

5.3.2 Sobre-sustracción y suelo espectral

Berouti, Schwartz y Makhoul presentan en [3] una opción para disminuir la percepción auditiva del ruido musical basándose en el enmascaramiento de señales. En este modelo se pretende disminuir el rango dinámico de los picos espectrales asociados al ruido musical, de tal forma que sean menos perceptibles al oído humano. El costo es el incremento del nivel de ruido residual de banda ancha, que se sabe es menos incómodo al oído que el de banda estrecha. El modelo consiste en realizar la sustracción de una *sobre-estimación* del espectro de magnitud del ruido y restringir los niveles de magnitud espectral resultantes a un mínimo configurable conocido como *suelo espectral*. La rectificación de media onda realizada en (5.9), utilizando el modelo propuesto queda

$$|\hat{S}_W[k]| = \begin{cases} |SN_W[k]| - \alpha\mu_{N_W}[k] & si \quad |SN_W[k]| - \alpha\mu_{N_W}[k] > \beta\mu_{N_W}[k] \\ \beta\mu_{N_W}[k] & si \quad |SN_W[k]| - \alpha\mu_{N_W}[k] \leq \beta\mu_{N_W}[k] \end{cases} \tag{5.10}$$

donde α es el factor de sobre-sustracción y β es el parámetro que define el suelo espectral. Valores de $\alpha > 1$ reducen la magnitud de los picos espectrales residuales y en general el ruido de banda ancha. Es claro que al utilizar una sobre-estimación del ruido además de disminuir el ruido residual también se distorsionan componentes propias de la voz, con un mayor impacto negativo cuando la SNR es baja. Cuando seguir incrementando α implica un alto grado de distorsión en la voz, lo mejor es negociar cierto nivel aceptable de ruido residual de banda ancha a cambio de la disminución perceptiva de los picos espectrales. Este procedimiento se realiza con

valores de $0 < \beta << 1$, lo que permite enmascarar los picos espectrales remanentes con un ligero nivel de ruido de banda ancha.

Si la SNR de entrada es suficientemente alta para que la voz enmascare al ruido, la necesidad de la sobre-sustracción disminuye, siendo más necesaria conforme el nivel de ruido residual reste la SNR. Este comportamiento deja la posibilidad de definir el factor de sobre-sustracción en función de la SNR, como en [3,17] donde se obtiene

$$\alpha(SNR_W) = \begin{cases} \frac{5}{4}\alpha_0 - \frac{1}{4} & si \quad SNR_W < -5 \\ \alpha_0 - SNR_W \frac{\alpha_0 - 1}{20} & si \quad -5 \le SNR_W \le 20 \\ 1 & si \quad SNR_W > 20 \end{cases} \tag{5.11}$$

siendo SNR_W la relación señal a ruido de entrada para el segmento w y α_0 el factor de sobre-sustracción para $SNR_W = 0$ dB, con un rango adecuado de 3 a 6 unidades.

A su vez, el parámetro β ofrece mejores resultados cuando se mantiene constante debido a que establece cierto nivel de ruido natural al oído. Se establece de 0.02 a 0.06 para $SNR_W < 0$ dB y de 0.005 a 0.02 para $SNR_W \geq 0$ dB.

El término SNR_W se puede obtener a partir del espectro estimado de la señal limpia y de la estimación de la magnitud del ruido de entrada como

$$SNR_W = 10\log\left(\frac{\sum_{\forall k}|\hat{S}_W[k]|^2}{\sum_{\forall k}\mu^2_{N_W}[k]}\right) \tag{5.12}$$

donde cada punto k corresponde a cada punto sobre todo el ancho de banda espectral.

5.3.3 Sub-bandas espectrales

El espectro en frecuencia de una señal de ruido generalmente no se encuentra definido con la misma intensidad para todo el ancho de banda, razón por la cual las señales de voz normalmente no se encuentran contaminadas uniformemente en todo su espectro, Figura 5.5. Entonces, es conveniente adaptar el modelo de sobre-sustracción de tal manera que durante el procesamiento se considere la cantidad de ruido involucrado en cada región espectral. Para tal efecto, el espectro de la voz es dividido en cierto número de regiones o *sub-bandas* espectrales. En esta investigación se consideró apropiado

realizar la distribución en sub-bandas según la distribución de bandas críticas tratadas en el apartado 2.2.4. El modelo en sub-bandas queda representado de la siguiente manera

$$|\hat{S}_{W,j}[k]| = \begin{cases} |SN_{W,j}[k]| - \alpha_j \mu_{N_W,j}[k] & si \quad |SN_{W,j}[k]| - \alpha\mu_{N_W,j}[k] > \beta\mu_{N_W,j}[k] \\ \beta\mu_{N_W,j}[k] & si \quad |SN_{W,j}[k]| - \alpha\mu_{N_W,j}[k] \leq \beta\mu_{N_W,j}[k] \end{cases} \tag{5.13}$$

Donde el subíndice j indica el número de las J sub-bandas sobre la cual se está operando. Como se observa, se realizan J sustracciones por bandas independientes cuyo factor de sobre-sustracción α_j debe ser acorde al monto de ruido en dicha sub-banda. El valor de α_j puede ser calculado de igual forma que en (5.11), únicamente que ahora debe considerarse el cálculo de la relación señal a ruido por segmentos w y por sub-banda j, $SNR_{W,j}$, tal que

$$SNR_{W,j} = 10\log\left(\frac{\sum_{\forall k,j}|\hat{S}_W[k]|^2}{\sum_{\forall k,j}\mu^2_{N_W}[k]}\right) \tag{5.14}$$

El modelo de sub-bandas antes presentado es una versión modificada de la que se presenta en [17], en el cual se considera un factor de sobre-sustracción adicional por sub-banda que aquí no se considera necesario.

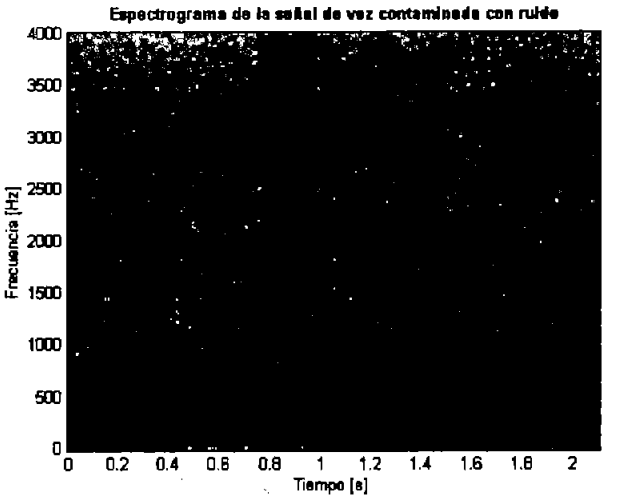


Figura 5.5 Espectrograma de una señal de voz contaminada con ruido ambiental de espectro con distribución no uniforme.

5.3.4 Promediado espectral

Un artificio útil para disminuir aún más el ruido residual corresponde al promediado espectral. Usar el promediado espectral para obtener una versión suavizada de $SN_{\overline{m}}[k]$ permite disminuir la varianza de las componentes espectrales del ruido, esto a su vez propicia que las estimaciones realizadas con (5.6) sean mas precisas y que el monto de ruido residual sea menor. Sin embargo, esta operación también puede cancelar componentes tipo ruido asociadas a la voz, lo cual se vería reflejado directamente en la percepción de los sonidos sordos. En [5] se sugiere que el promediado espectral no cubra más de tres segmentos traslapados al 50%, cada uno con 256 puntos.

5.3.5 Detector de actividad de voz

Cuando se trabaja en condiciones libres o de bajo ruido es común que el detector de voz esté basado en la energía y el número de cruces por cero en tiempo corto de la señal de voz. Para tales condiciones, este tipo de detectores resultan ser muy eficientes ya que el inicio y fin de palabra pueden ser determinados con un alto grado de confiabilidad y con un costo de cómputo muy bajo. Sin embargo, la condición libre de ruido pocas veces se logra fuera del laboratorio de diseño y cuando este tipo de detectores son trasladados a entornos con ruido su rendimiento se ve disminuido notablemente. La disminución del rendimiento se debe principalmente a que la energía y el número de cruces por cero son modificados por la adición de la señal de ruido, lo cual genera que los umbrales de detección se vuelvan inestables.

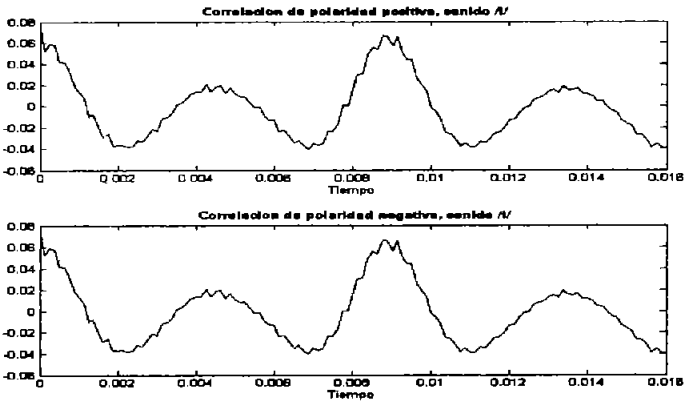
En esta tesis el problema de detección de voz se resolvió con una versión simplificada del método propuesto por Wu y Li en [41]. Este método se basa en dos métricas, la similitud de las funciones de correlación de los polos positivo y negativo de la señal y la energía en tiempo corto. Para obtener las funciones de correlación de polaridad positiva y negativa, la señal en tiempo corto[1] $sn_l[t]$ puede ser expresada como la suma de sus señales positiva y negativa, esto es

$$sn_l[t] = sn_l^+[t] + sn_l^-[t] \tag{5.15}$$

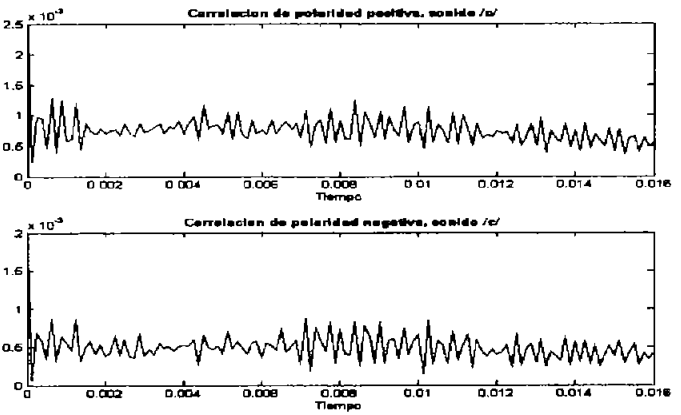
Como puede observarse en la Figura 5.6, sus funciones de correlación en el polo positivo r_l^+ son muy parecidas a las correspondientes del polo negativo r_l^- tanto en sonidos sonoros como sordos. En el caso de señales de ruido esta situación regularmente no ocurre, lo cual sugiere utilizar tal característica en la detección de la presencia de voz.

[1] Nótese que se utiliza el subíndice *l* en vez de *w* para indicar que la segmentación se realiza con ventanas rectangulares.

a)



b)



c)

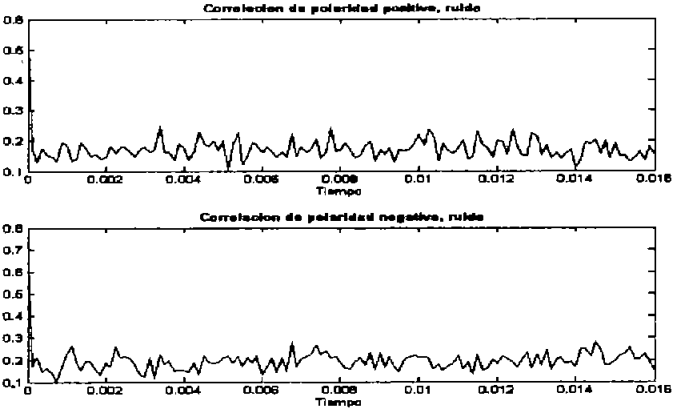


Figura 5.6 Similitud característica entre las funciones de correlación polar para sonidos sonoros (a) y sonidos sordos (b), característica no observable en señales de ruido (c).

La detección de actividad de voz se realiza calculando el grado de similitud entre r_l^+ y r_l^- por cada segmento l. Esta métrica se obtiene a partir del coeficiente de correlación (2.10), considerando cada función de correlación como una variable aleatoria, esto es

$$\rho_l = \frac{Cov_{r_l^+ r_l^-}}{\sigma_{r_l^+}\sigma_{r_l^-}} \tag{5.16}$$

Donde el término de covarianza se obtiene a partir del estimador

$$Cov_{r_l^+ r_l^-} = \frac{1}{N}\sum_{j=0}^{N-1}\left(r_l^+[j] - \mu_l^+\right)\left(r_l^-[j] - \mu_l^-\right) \tag{5.17}$$

De igual forma se requiere el cálculo de la energía E_l por cada segmento l, tal que para cada segmento de 16 o 32 ms se tiene el parámetro

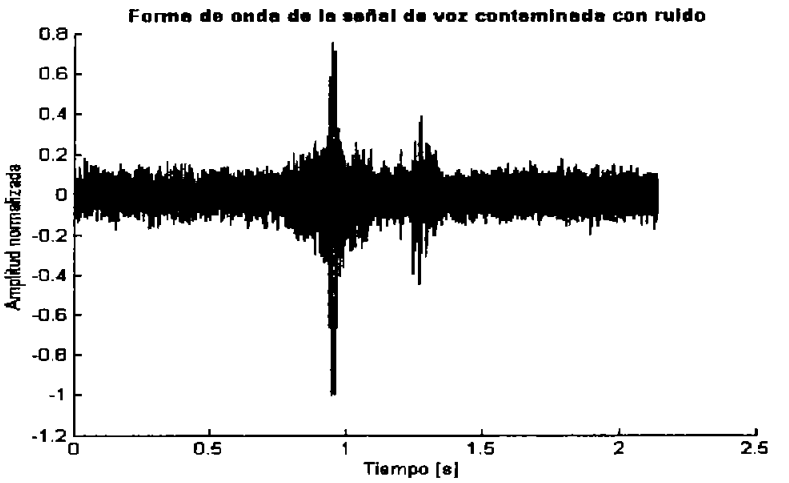
$$SE_l = \rho_l \cdot \log E_l \tag{5.18}$$

La versión adaptada para esta investigación considera que si el parámetro SE_l es superior que el umbral TH, entonces el segmento l corresponde a un segmento de voz, de lo contrario, el segmento corresponde a una señal de ruido. El umbral TH utilizado en las pruebas es

$$TH = \mu_{SE_l} + \sigma_{SE_l} \tag{5.19}$$

En la Figura 5.7 se ilustra un ejemplo del funcionamiento de la detección de silencios o actividad de voz con este método.

a)



b)

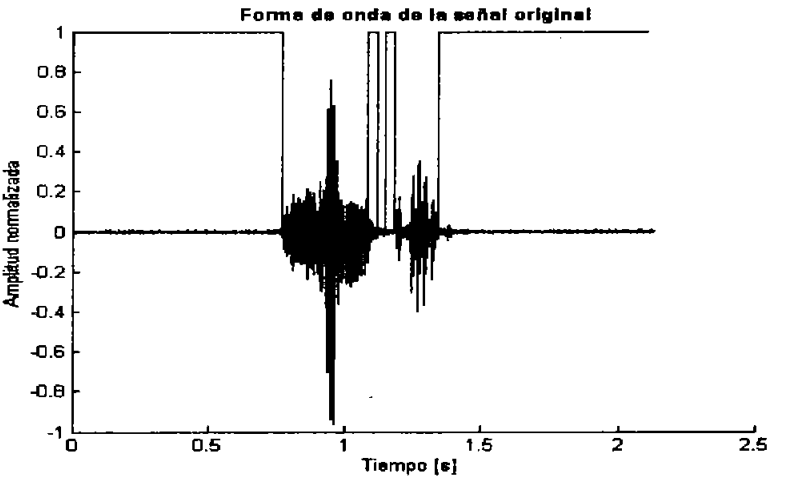


Figura 5.7 Detección de silencio o actividad de voz en la señal de voz contaminada. a) Señal de voz contaminada con ruido y b) Comparación de los espacios de silencio detectados, de valor unitario, con la señal original.

Capítulo 6

EVALUACIÓN EXPERIMENTAL

6.1 Métricas de calidad

Como ya se ha visto, medir el rendimiento de las técnicas para reducción de ruido a partir de la evaluación subjetiva es un procedimiento muy costoso, así que normalmente se recurre a alguna de las métricas objetivas revisadas en el Capítulo 3. Sin embargo, es necesario reconocer que no todas las métricas objetivas descritas en dicho capítulo son lo suficientemente representativas de la calidad subjetiva. Así que con el objetivo de asegurar la utilización de la métrica más adecuada se realizó el siguiente análisis.

Se generó un conjunto de 48 señales de prueba, resultado del procesamiento de una señal con tres niveles de distorsión, –10, 0 y 10 dB de SNR, por tres fuentes de ruido ambientales y una sintética (ruido blanco). Las técnicas utilizadas en la generación de esta muestra fueron la cancelación adaptable de interferencia con LMS y LMS-N, y sustracción espectral con sobre-sustracción y suelo espectral, por análisis de banda completa y en sub-bandas. De las 280 evaluaciones recolectadas se generó la estadística ilustrada en la Figura 6.1.

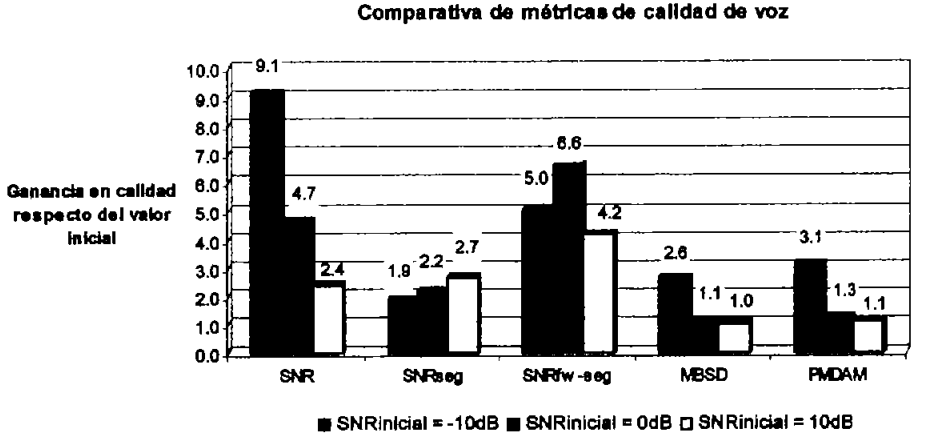


Figura 6.1 Comportamiento estadístico de las métricas de calidad objetiva respecto de la evaluación subjetiva con PMDAM.

Puede intuirse que el oído humano percibe mayor incremento en la calidad de voz cuando el monto de ruido reducido es mayor e incremento casi nulo cuando la reducción es mínima. Tal efecto puede verificarse en la métrica PMDAM, según la cual la calidad resultante no va más allá de tres o cuatro veces la calidad inicial. Como se observa, únicamente MBSD y SNR siguieron tal comportamiento aunque esta última se

observa más susceptible a distorsiones de baja amplitud. Tal susceptibilidad provoca que se reporten incrementos relativamente altos en la calidad. Como también puede apreciarse, ninguna de las variantes SNR_{seg} y $SNR_{fw\text{-}seg}$ demostraron correlación con las pruebas objetivas realizadas. De esta forma se justificó la utilización de MBSD como métrica de evaluación de calidad y rendimiento en los experimentos que siguen.

6.2 Técnicas de reducción de ruido

Para el análisis comparativo del rendimiento de las técnicas se generó un conjunto de 12 casos de distorsión para 4 señales de voz distintas. Los niveles de distorsión utilizados fueron –10, 0 y 10 dB de SNR. Para analizar el comportamiento de las técnicas se seleccionaron tres tipos de distorsión (distribuidos en cuatro señales de ruido): ruido aproximadamente estacionario con distribución espectral uniforme (factor de dispersión igual a 1), ruido aproximadamente estacionario con distribución espectral no uniforme (factor de dispersión mayor a 1) y ruido no estacionario con distribución espectral no uniforme (factor de dispersión mayor a 1), todos ellos tomados de fuentes ambientales reales. La reducción de ruido por medio de cancelación adaptable de interferencia se experimentó con los algoritmos LMS y LMS-N. El canal entre la fuente de ruido y el punto de contaminación de la señal de voz se modeló como $n' = 0.9\, n \pm 0.1\, |n|\, wn$, donde n' es ruido, n es su señal de referencia y wn es ruido blanco de potencia unitaria. Con sustracción espectral la experimentación fué más variada dada la cantidad de parámetros ajustables: se realizaron pruebas con sustracción de espectros de magnitud y de potencia, con y sin sobre-sustracción, para el caso de sobre-sustracción se experimentó con tres valores distintos para α_0 (2,3 y 4), con análisis en banda completa y por sub-bandas. Se evaluaron 16 configuraciones distintas para reducción de ruido, por lo que en total se generaron 768 estimaciones de señales limpias, cuyas métricas MBSD sirvieron de base para los análisis siguientes.

6.2.1 Cancelación adaptable de interferencia vs. sustracción espectral

En las Figuras 6.2, 6.3 y 6.4 se muestra comparativamente el rendimiento de LMS, LMS-N, sustracción espectral de magnitud (SS-Mag) y sustracción espectral de potencia (SS-Pot) para diferentes tipos de distorsión. Dado que la MBSD es una métrica de distancia, debe interpretarse que valores menores representan menor grado de distorsión residual. La Figura 6.2 corresponde al caso contaminación por ruido de estadísticas aproximadamente estacionarias y espectro con distribución uniforme. El espectro de este tipo de ruido puede distinguirse por mantener un espectro de distribución casi plana en todo el ancho de banda a lo largo del tiempo. El ruido blanco se incluye en esta categoría. En la Figura 6.3 se ilustra el comportamiento de las técnicas para el caso en el que el espectro del ruido tiene estadísticas aproximadamente estacionarias pero no tiene una distribución uniforme, es decir, el factor de dispersión de los valores propios de su matriz de correlación es mayor que la unidad. Y finalmente, la Figura 6.4 presenta el caso del ruido no estacionario y con espectro de distribución no uniforme, distinguible por el cambio constante de la ubicación y magnitud de las componentes principales.

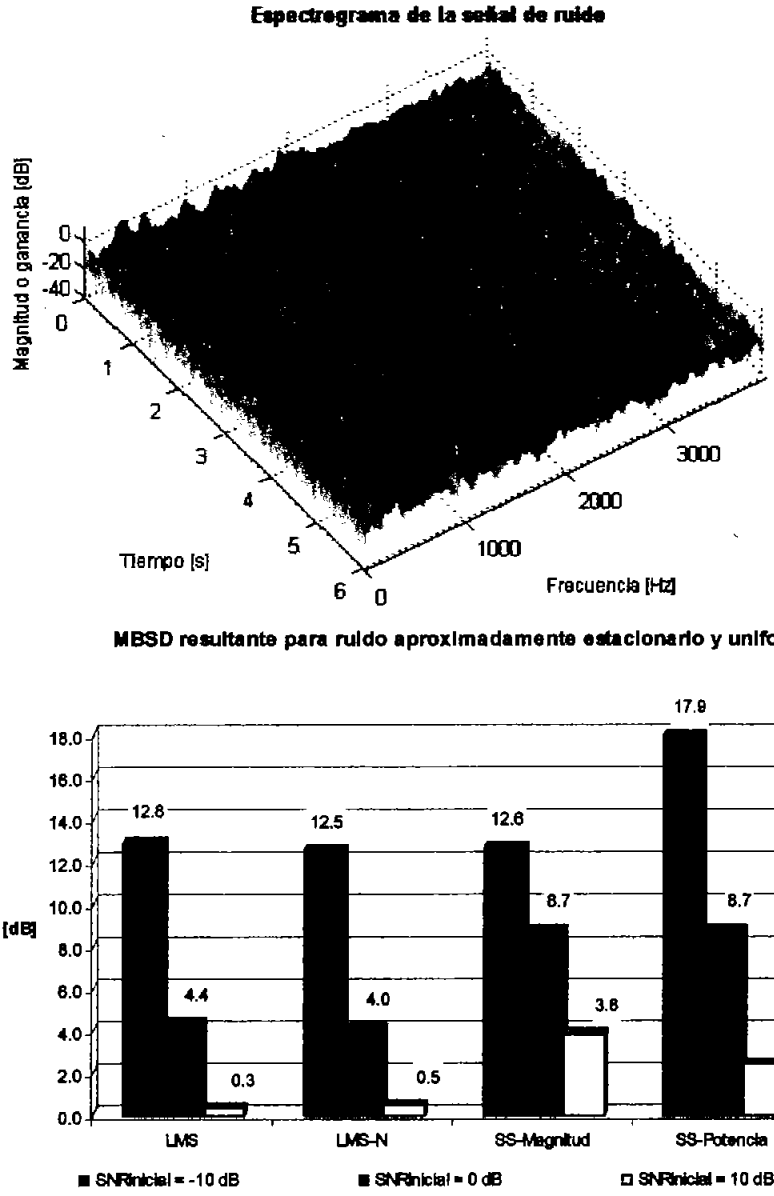
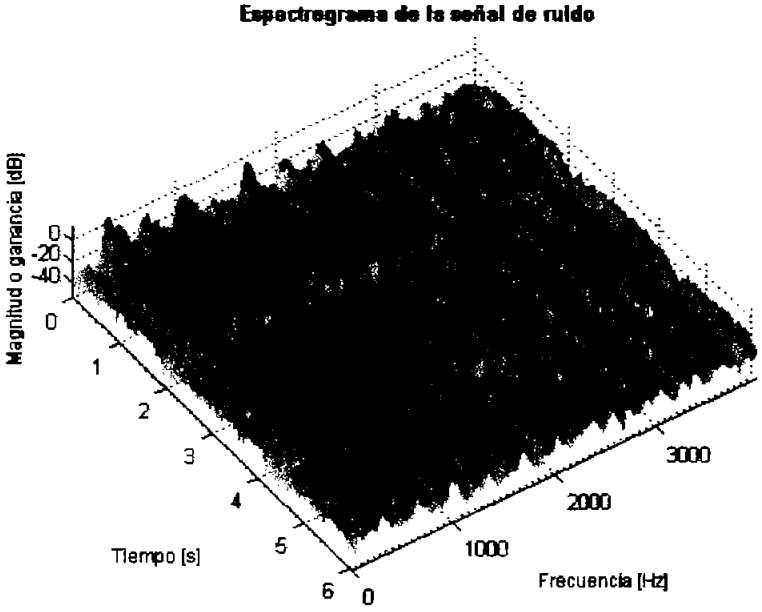


Figura 6.2 Comportamiento MBSD promedio de la cancelación adaptable de interferencia y de sustracción espectral para ruido aproximadamente estacionario con distribución espectral uniforme.



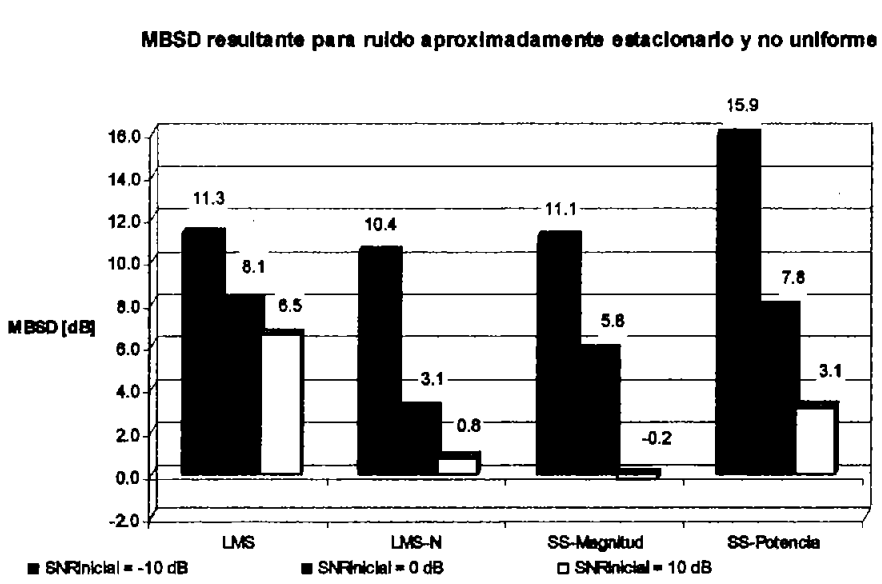
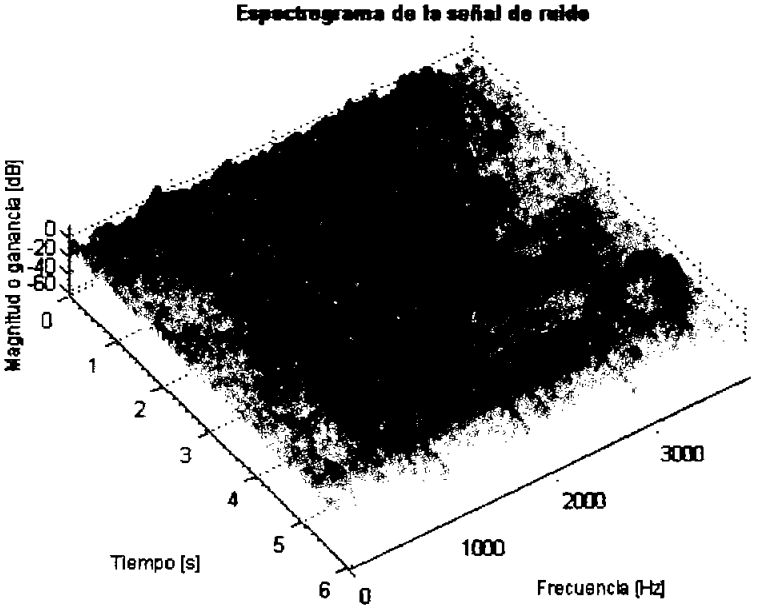


Figura 6.3 Comportamiento MBSD promedio de la cancelación adaptable de interferencia y de sustracción espectral para ruido aproximadamente estacionario con distribución espectral no uniforme.



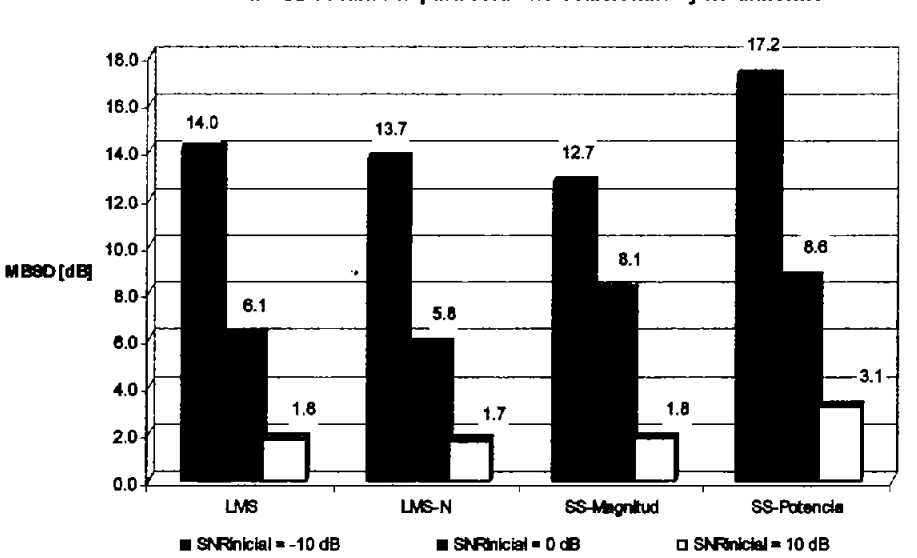


Figura 6.4 Comportamiento MBSD promedio de la cancelación adaptable de interferencia y de sustracción espectral para ruido no estacionario y con distribución espectral no uniforme.

En general, se observó que LMS y LMS-N tienen un comportamiento muy parecido (LMS-N ligeramente mejor) cuando el espectro es estacionario y se encuentra uniformemente distribuido. Sin embargo, cuando el ruido no fue uniforme LMS disminuyó notablemente su rendimiento, contrario a lo que ocurrió con LMS-N, cuyo comportamiento fue casi el mismo. Este fenómeno ya se esperaba a partir de lo que se observa en la Figura 4.7. Para tal situación, LMS ofreció constantes de tiempo de adaptación muy grandes en alguno o algunos de los parámetros del filtro, provocando que la salida del filtro no consiguiera seguir fielmente las condiciones de las señales de entrada. Por otro lado, cuando la cualidad estacionaria no se cumplió, LMS-N perdió ventaja sobre LMS debido a que las estimaciones de R se volvieron más imprecisas, caso en el que LMS y LMS-N reflejó su menor rendimiento.

En lo respecta a sustracción espectral, evaluaciones subjetivas y con MBSD reportaron mayor rendimiento con SS-Mag que con SS-Pot. Distinto a lo que ocurrió con LMS, SS-Mag tendió a elevar su rendimiento cuando se aplicó a ambientes con ruido de espectro no uniforme. La razón es que en tal situación la sustracción se realizó entre espectros mejor definidos, afectando así en menor grado a las componentes de voz que originalmente no fueron distorsionadas. Se observó que cuando el ruido tuvo un espectro uniforme, SS-Mag no pudo ofrecer menor distancia MBSD promedio que 3.8 dB aún cuando la distorsión original fue apenas de 10 dB de SNR. En este sentido parece ser que el contraste del espectro del ruido importó más que su cualidad estacionaria.

A grandes rasgos, la cancelación adaptable de interferencia con LMS-N proveyó estimaciones de mayor calidad que con sustracción espectral. Las excepciones observables en las figuras, como en $SNR_{Inicial}$ = -10 dB de la Figura 6.4, se debieron principalmente a la falta de correlación entre la señal de ruido contaminante y la referencia del modelo de cancelación adaptable de interferencia de la Figura 4.2.

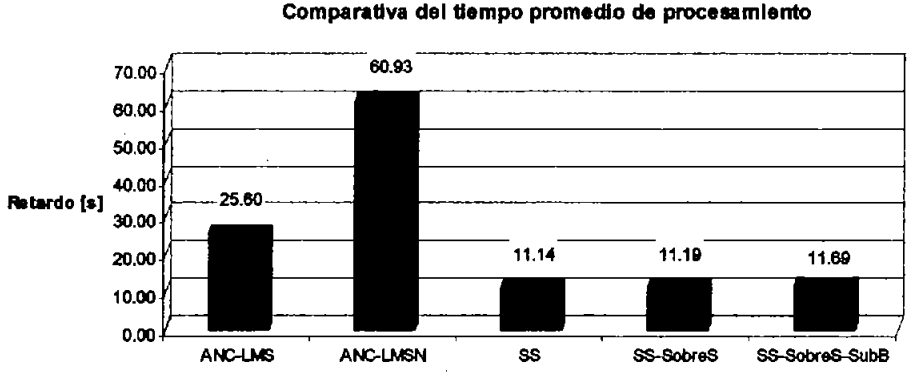


Figura 6.5 Comparativa del tiempo de procesamiento requerido por las técnicas de reducción de ruido, expresado como retardo de la señal procesada.

Otro aspecto comparable entre las técnicas es la complejidad computacional del procesamiento. Esta complejidad se encuentra asociada al número de operaciones requeridas del procesador digital, cantidad que puede reflejarse en el retardo que una señal experimenta al ser procesada por el sistema. En la Figura 6.5 se ilustra el tiempo promedio que cada técnica consumió en el procesamiento de 1 segundo de la señal procesada. Debe notarse que la intención de la Figura 6.5 es mostrar de manera cualitativa la complejidad computacional de cada técnica, ya que el tiempo de procesamiento además de depender del número de operaciones, depende de la capacidad de cómputo del procesador. De la figura antes mencionada, puede deducirse que sustracción espectral, en su forma más simple, fue más eficiente en lo que respecta a respuesta de tiempo real. En este sentido, el caso de menor eficiencia correspondió a LMS-N, debido a la cantidad de operaciones realizadas continuamente en el cálculo de los L parámetros del filtro, así como las requeridas para la estimación de R^{-1}.

6.2.2 Variantes de sustracción espectral

Según lo que se observa que la Figura 6.6, el grado de sobre-sustracción requerido fue directamente proporcional al grado de distorsión de la señal. Por ejemplo, el caso de señales de –10 dB de $SNR_{inicial}$ tuvo mejor respuesta con sobre-sustracción de $\alpha_0 = 3$, $\alpha_0 = 2$ para 0 dB, y sobre-sustracción nula para 10 dB. Pudo verse que aplicar sobre-sustracción sobre señales de voz con baja distorsión podría ocasionar el deterioro de la misma. Tal deterioro se debe a la cancelación de componentes asociadas a la propia voz, de las cuales las más susceptibles corresponden a sonidos sordos.

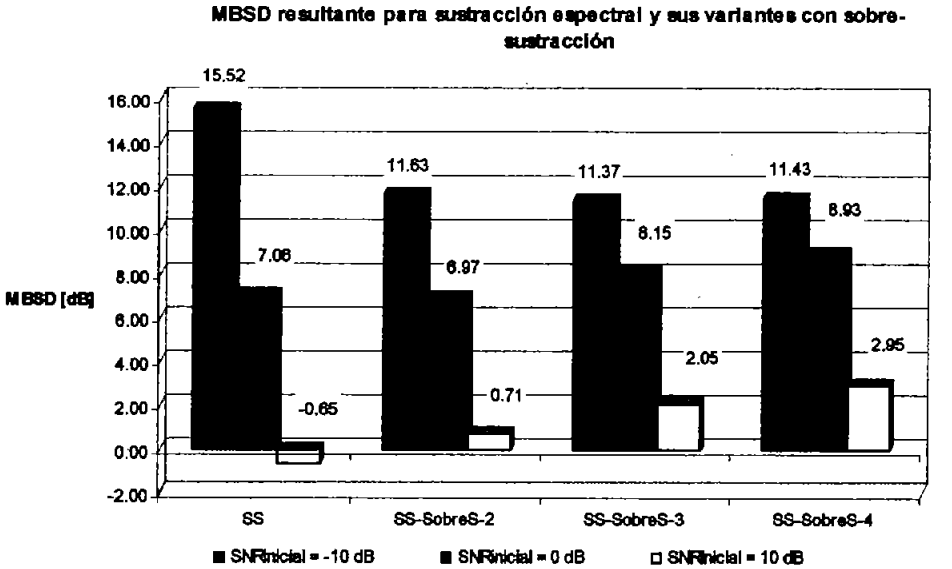


Figura 6.6 Comportamiento MBSD promedio de sustracción espectral y sus variantes con sobre-sustracción y suelo espectral para $\alpha_0 = 2$, 3 y 4.

Por otro lado, la experimentación con el modelo de sustracción espectral por sub-bandas ha dejado ver que la métrica MBSD aún requiere de refinamiento y que no puede sustituir completamente al juicio humano. En varias pruebas, la aplicación del modelo de sub-bandas no representó grandes mejoras en la calidad de la voz percibida, sin embargo, se percibió que las señales obtenidas por este medio fueron de mayor naturalidad que las obtenidas por análisis de banda completa. Este efecto se debe principalmente a que las componentes de baja amplitud (sonidos sordos) son muy susceptibles a ser canceladas cuando se opera con estimaciones del espectro completo. Cuando la sustracción se realiza por sub-bandas, estas componentes obtienen mayor representatividad en la banda de pertenencia, lo que hace que sea menos probable que sean canceladas. La desventaja es que tales componentes de baja amplitud pueden corresponder a sonidos sordos de voz, o bien, a ruido. Esto hace que, sobre todo en altos niveles de distorsión, existan componentes residuales de ruido confundidas con voz. La Figura 6.7 muestra los resultados obtenidos con los modelos de análisis de banda completa y por sub-bandas, aunque tales resultados no reflejan la preferencia del oído por mayor definición de sonidos sordos pese al aumento de ruido residual.

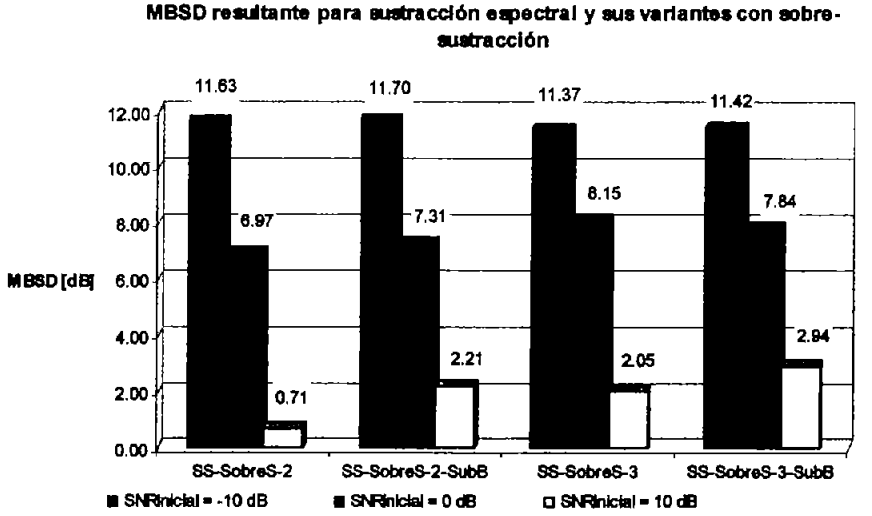
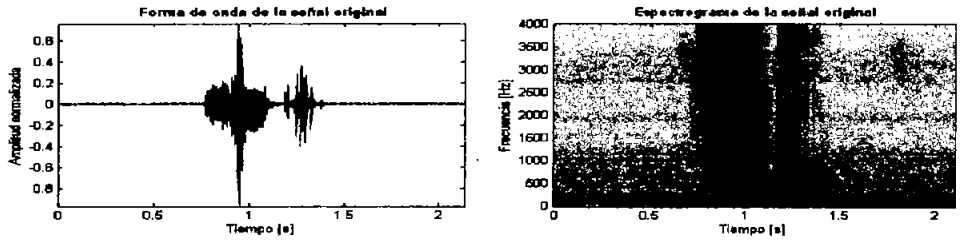


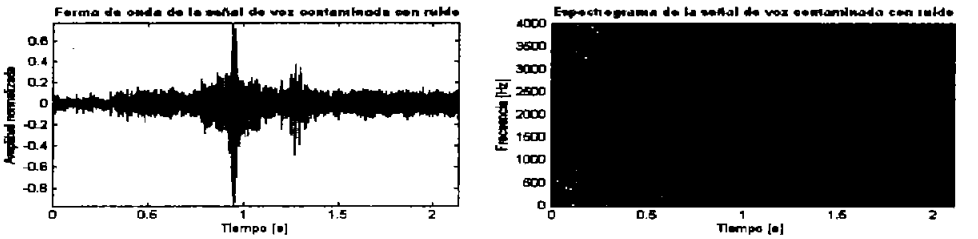
Figura 6.7 Comportamiento MBSD promedio de sustracción espectral por banda completa y por sub-bandas.

Un ejemplo del comportamiento de las técnicas analizadas puede observarse en las figuras 6.8 y 6.9. Esta figura corresponde a las estimaciones de la forma de onda y el espectro de magnitud de la señal /cinco/ contaminada a 0 dB de SNR con ruido ambiental generado por aplausos dentro de un auditorio.

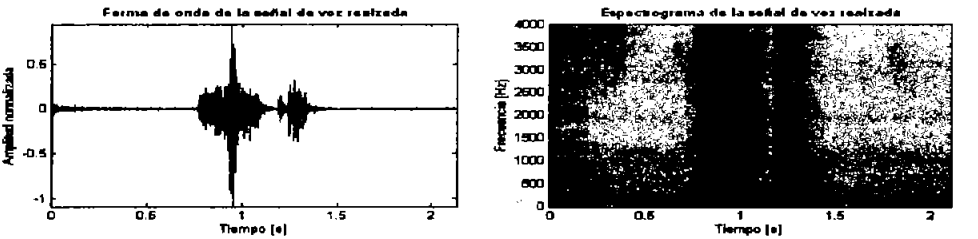
Señal original



Señal contaminada - MBSD = 12.5780



LMS - MBSD = 4.4577



LMS-N - MBSD = 0.6846

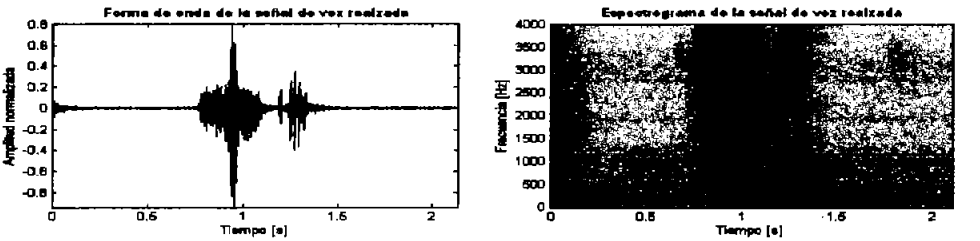
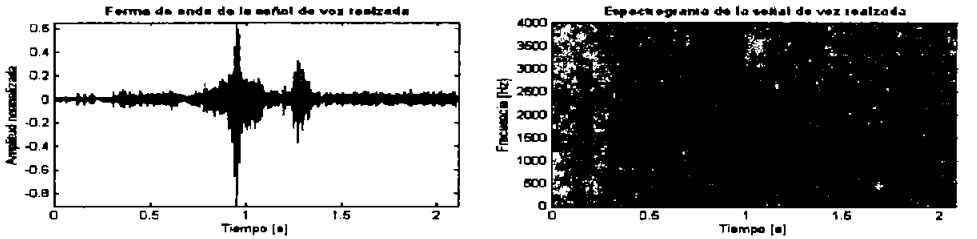
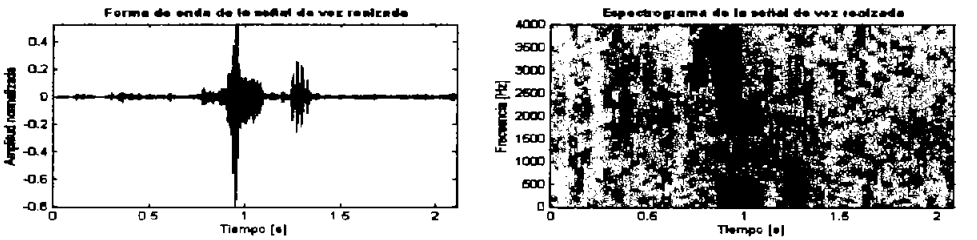


Figura 6.8 Forma de onda y espectrograma de la señal de voz /cinco/ contaminada con ruido ambiental a 0 dB de SNR, con estimaciones de señal limpia por LMS y LMS-N.

SS en magnitud - MBSD = 6.8206



SS en magnitud con sobre-sustracción 2 - MBSD = 7.0548



SS en magnitud con sobre-sustracción 2 por sub-bandas - MBSD = 7.4980

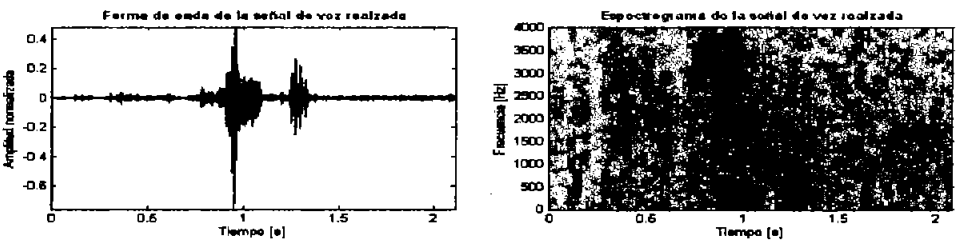


Figura 6.9 Forma de onda y espectrograma de la señal de voz /cinco/ contaminada con ruido ambiental a 0 dB de SNR, con estimaciones de señal limpia por SS en magnitud, SS en magnitud con sobre-sustracción y SS en magnitud con sobre-sustracción por sub-bandas.

Capítulo 7

CONCLUSIONES

Se han analizado dos técnicas para reducción de ruido acústico ambiental en señales de voz, cancelación adaptable de interferencia y sustracción espectral. El rendimiento de tales técnicas fue evaluado a partir del juicio directo de oídos humanos y por medio de métricas objetivas de calidad. Se encontró que de las métricas objetivas utilizadas, MBSD es la más representativa y por lo tanto la más adecuada en la determinación automática de la calidad de señales de voz. En general, ambas técnicas de reducción de ruido proporcionaron mejoras notables en la calidad de voz en comparación con la señal en estado de distorsión, aunque el grado de calidad alcanzada depende de la técnica y la variante utilizada.

La técnica de cancelación adaptable de interferencia basada en el algoritmo LMS-N proporcionó el nivel más alto de calidad en la mayoría de los casos, aunque tal superioridad fue más notable únicamente en los casos de ruido estacionario. Por otro lado, en lo que respecta a complejidad computacional (tiempo de procesamiento) la cancelación adaptable de interferencia resultó ser menos eficiente, siendo en este sentido LMS el algoritmo de mayor aceptación. Además, la cancelación adaptable de interferencia tiene como desventaja el requerimiento de un canal de referencia, el cual puede ser limitante para su aplicación en sistemas reales. A pesar de la efectividad del error cuadrático medio como medida de distorsión en el procesamiento de señales, en lo que corresponde a voz no es particularmente efectivo debido a su naturaleza estadística más que perceptiva. Un ejemplo de lo anterior se observó en la sensibilidad que el cancelador adaptable de interferencia presentó cuando ocurre una diferencia en la fase de la señal de ruido de referencia y la señal de voz distorsionada.

En lo que respecta a la cantidad de recursos requeridos, la técnica de sustracción espectral representó la mejor opción. Esta técnica prescinde de un canal de referencia para poder conocer la señal contaminante, en su lugar necesita del apoyo de algún método para detectar silencio o actividad de voz sobre la única señal disponible. Evidentemente este requerimiento representa una desventaja en la aplicación de la técnica, sin embargo, un método más o menos eficiente en la detección de silencio no representa una gran carga computacional para la técnica, tanto que sustracción espectral presentó el menor tiempo de procesamiento durante las pruebas realizadas. Con relación a la calidad de voz lograda con esta técnica, puede decirse que en el caso general de ruido no estacionario y no uniforme los resultados fueron inferiores pero comparables a los obtenidos con la cancelación adaptable de interferencia. Una ventaja de sustracción espectral sobre la cancelación adaptable de interferencia es la integración de información de carácter perceptivo (enmascaramiento de sonidos) en el modelo de reducción.

Se observó que ambas técnicas dependen fuertemente en la naturaleza estacionaria del ruido contaminante, así que puede esperarse que tengan mejor comportamiento en este tipo de ambientes y no así en su aplicación a señales contaminadas con ruido de carácter impulsivo.

Recientemente se han propuesto extensiones a las técnicas básicas aquí presentadas, extensiones que se fundamentan principalmente en los aspectos perceptivos de la voz. Pueden mencionarse como ejemplos, el *análisis perceptivo lineal predictivo* (PLP) y *sustracción espectral con ponderación perceptiva*, técnicas que además de otros aspectos perceptivos toman en cuenta la variación de la sensitividad del oído con la frecuencia del sonido, tal y como en MBSD. Se espera que el desarrollo de este tipo de técnicas mejore aún más la calidad percibida de señales de voz contaminadas con ruido ambiental.

REFERENCIAS BIBLIOGRÁFICAS

[1] Acero, A., "Acoustical and Environmental Robustness in Automatic Speech Recognition", Tesis de Doctorado en Ingeniería Eléctrica, Univ. Carnegie Mellon, E.U.A., 1990.

[2] Beranek, L., *Acoustics*, McGraw-Hill, E.U.A., 1954.

[3] Berouti, M., Schwartz, R. y Makhoul, J., "Enhancement of Speech Corrupted by Acoustic Noise", *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79*, Abril de 1979, vol. 4, pp 208-211.

[4] Bhatnagar, M., "A Modified Spectral Subtraction Method Combined With Perceptual Weighting For Speech Enhancement", Tesis de Maestría en Ciencias de Ingeniería Eléctrica, Universidad de Texas en Dallas, E.U.A., 2002.

[5] Boll, S.F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Abril de 1979, vol. 27, no. 2, pp. 113-120.

[6] Brokish, Ch.W. y Lewis, M., "A-Law and mu-Law Compandings Implementations Using the TMS320C54x", *Digital Signal Processing Solutions*, Texas Instruments Inc., Diciembre de 1997.

[7] Costache, G., Railneau, A., Pavel, R., Perisoara, L., y Ionita, M., "Improved Speech Recognition In Noisy Environments Using Spectral Noise Reduction And Perceptual Linear Prediction", *GSPx & International Signal Processing Conference*, E.U.A., Marzo – Abril de 2003.

[8] Deller, J.R., Proakis, J.G. y Hansen, J.H., *Discrete-Time Processing of Speech Signals*. Prentice Hall, E.U.A., 1987.

[9] Decina, M. y Modena, G., "CCITT Standards on Digital Speech Processing", *IEEE Journal on Selected Areas in Communications*, Febrero de 1988, vol. 6, no. 2, pp. 227-234.

[10] Diniz, P.S.R., Campos, M.L.R., y Antoniou, A., "Analysis of LMS-Newton Adaptive Filtering Algorithms with Variable Convergence Factor", *IEEE Transactions on Signal Processing*, E.U.A., Marzo de 1995, vol. 43, no. 3, pp. 617-627

[11] Ephraim, Y., "Statistical-Model-Based Speech Enhancement Systems", *Proc. IEEE*, Octubre de 1992, vol. 80, no. 10, pp. 1526-1555.

[12] Furui, S., *Digital Speech Processing, Syntesis, and Recognition 2nd Ed.*, Marcel Dekker, Inc., E.U.A., 2001.

[13] Furui, S., Iwano, K., Hori, Ch., Shinozaki, T., Saito, Y. y Tamura, S., "Ubiquitous Speech Processing", *IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, Mayo de 2001, vol. 1, pp. 13-16.

[14] Graham, P.C., "Structured Variation in British English Liquids: The Role of Resonance", Tesis de Doctorado en Ciencia Lingüística, Universidad de York, Reino Unido, 2002.

[15] Hansen, J.H.L. y Pellom, B, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms", *Inter. Conf. on Spoken Language Processing 98*, Sydney Australia, Diciembre de 1998, vol. 7, pp. 2819-2822.

[16] Johnston, J.D., "Transform Coding of Audio Signals Using Perceptual Noise Criteria", *IEEE Journal on Selected Areas in Communications*, Febrero de 1988, vol. 6, no. 2, pp. 314-323.

[17] Kamath, S.D., "A Multi-Band Spectral Subtraction Method for Speech Enhancement", Tesis de Maestría en Ciencias de Ingeniería Eléctrica, Universidad de Texas en Dallas, E.U.A., 2001.
[18] Komo, J.J., *Random Signal Analysis in Engineering Systems*. Academic Press Inc., E.U.A., 1987.
[19] Lim, J.S., *Speech Enhancement*, Prentice Hall, E.U.A., 1983.
[20] Lim, J.S. y Oppenheim, A.V., "Enhancement and Bandwidth Compression of Noisy Speech", *Proc. IEEE*, Diciembre de 1979, vol. 67, no. 12, pp. 1586-1604.
[21] Milner, B.P. y Vaseghi, S.V. "Comparison of some noise-compensation methods for speech recognition in adverse environments", *IEE Proc. - Vision, Image and Signal Processing*, Octubre de 1994, vol. 141, no. 5, pp. 280-288.
[22] Milner, B.P., Lewis, A.V., y Vaseghi, S.V., "Speech Enhancement", Capítulo en *Speech Technology for Telecommunications*, Westall, F.A., Johnston, R.D., y Lewis, A.V., Chapman & Hall Eds., Reino Unido, 1998, pp. 376-405.
[23] Niederjohn, R.J. y Heinen, J.A., "Understanding speech corrupted by noise", *Proc. IEEE International Conference on Industrial Technology*, Shanghai, China, Diciembre de 1996, pp. 1-5.
[24] O'Saughnessy, D., *Speech Communications, Human and Machine*, IEEE Press, E.U.A., 2000.
[25] Pazaitis, D.I., y Constantinides, A.G., "A Kurtosis-Driven Variable Step-Size LMS Algorithm", *ICASSP-96, Conference Proceedings*, E.U.A., Mayo de 1996, vol. 3, pp. 1846-1849.
[26] Picinbono, B., *Random Signals and Systems*, Prentice Hall, E.U.A., 1993.
[27] Pols, L.C.W., "Flexible Human Speech Recognition", *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding*, Piscataway, NJ, 273-283.
[28] Quackenbush, S.R., Barnwell, T.P, y Clements, M.A., *Objective Measures of Speech Quality*, Prentice Hall, E.U.A., 1988.
[29] Rabiner, L., "Applications of Voice Processing to Telecommunications", *Proc. IEEE*, Febrero de 1994, vol. 82, no. 2, pp. 199-228.
[30] Rabiner, L. y Juang B. H., *Fundamentals of Speech Recognition*, Prentice Hall, E.U.A., 1993.
[31] Saito, S. y Nakata, K., *Fundamentals of Speech Signal Processing*. Academic Press Inc., E.U.A., 1981.
[32] Scales, J. y Snieder R., "What is noise", *Geophysics*, Julio-Agosto de 1998, vol. 63, no. 4, pp. 1122--1124.
[33] Stern, R.M., Acero, A., Liu, F.-H., y Ohshima, Y., "Signal Processing for Robust Speech Recognition", Capítulo en *Speech Recognition*, Lee C.-H. y Soong F., Kluwer Academic Publishers, Boston, 1996, pp. 351-378.
[34] Stern, R.M. Raj, B. y Moreno, P.J., "Compensation for Environmental Degradation in Automatic Speech Recognition", *Proc. of the ESCA Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-au-Mousson, Francia, Abril de 1997, pp. 33-42.
[35] Suzuki, Y., "Precise and Full-range Determination of Two-dimensional Equal Loudness Contours", Publicación de *New Energy and Industrial Technology Development Organization (NEDO)-ISD*, Marzo de 2003.
[36] Vaseghi S.V., *Advanced Digital Signal Processing and Noise Reduction 2nd Ed.*, John Wiley & Sons Ed., Reino Unido, 1996.

[37] Westall, F.A., "Review of Speech Technologies for Telecommunications", *Electronics & Communication Engineering Journal*, Octubre de 1997, pp. 197-207.
[38] Widrow, B., Glover, J.R., McCool, J.M., Kaunitz, J., Williams, Ch.S., Hearn, R.H., Zeidler, J.R., Dong, E., y Goodlin, R.C., "Adaptive Noise Cancelling: Principles and Applications", *Proc. IEEE*, Diciembre de 1975, vol. 63, no. 12, pp. 1692-1716.
[39] Widrow, B. y Kamenetsky, M., "On the Statistical Efficiency of the LMS Family of Adaptive Algorithms", *Proc. IEEE Joint Conf. Neural Networks*, Portland, Julio de 2003, vol. 4, pp. 2872-2880.
[40] Widrow, B. y Stearns, S.D., *Adaptive Signal Processing*, Prentice Hall, E.U.A., 1985.
[41] Wu, Y. y Li, Y., "Robust Speech/Non-Speech Detection in Adverse Conditions Using the Fuzzy Polarity Correlation Method", *2000 IEEE International Conference on, Systems, Man, and Cybernetics*, E.U.A., Octubre de 2000, vol. 4, pp. 2935-2939.
[42] Yang, W., "Enhanced Modified Bark Spectral Distortion (EMBSD): An Objective Speech Quality Measure Based On Audible Distortion And Cognition Model", Tesis de Doctorado en Filosofía, Universidad de Temple, E.U.A., 1999.
[43] Zwicker, E. y Fastl, H., *Psychoacoustics: Facts and Models*, Springer-Verlag Berlin Heidelberg Ed., Alemania, 1999.