Evaluación de técnicas clásicas de reducción de ruido en señales de voz (*)

D. R. Tomassi[1], L. Aronson[2], C. E. Martínez[1], D. H. Milone[3], M. E. Torres[1], H. L. Rufiner[1]

[1]*Facultad de Ingeniería, Universidad Nacional de Entre Ríos, CC47, Suc. 3 (CP 3100), Paraná, Entre Ríos, Argentina*
[2] *Departamento de Implante Coclear, Fundación Arauz*
[3] *Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral*

Resumen

En el presente trabajo se evalúa el desempeño de un conjunto de técnicas clásicas de reducción de ruido en habla en el marco del idioma Español Rioplatense. Se consideran las técnicas de sustracción espectral, filtrado de Wiener y las reglas de Ephraim-Malah. El desempeño de estos algoritmos se evalúa en términos de la inteligibilidad y la calidad de las señales de voz obtenidas luego de su aplicación. La inteligibilidad se mide por medio del porcentaje de palabras reconocidas correctamente por sujetos normoyentes, y se identifican las confusiones más frecuentes ocurridas con cada algoritmo mediante matrices de confusión de consonantes. La calidad del habla obtenida con cada técnica se evalúa en forma subjetiva de acuerdo a la calificación de los oyentes con respecto a un grupo de factores intervinientes en su percepción. También se efectúa la evaluación de calidad mediante un grupo de medidas objetivas seleccionadas al efecto. Se presenta el desempeño relativo de cada algoritmo considerado y se discute la correlación entre las calificaciones obtenidas con ambos tipos de métodos.

Palabras claves: • inteligibilidad • calidad del habla • reducción de ruido • prótesis auditivas •

Introducción

Numerosos algoritmos de reducción de ruido en señales de voz han sido propuestos durante los últimos años, acompañando la expansión de las tecnologías digitales para el procesamiento de señales y el desarrollo de las comunicaciones móviles. Las técnicas clásicas se han centrado sobre los enfoques de sustracción espectral [1-4], filtrado óptimo y adaptativo [3-6], y el procesamiento basado en modelos estadísticos [7-9]. Debido a que estos sistemas capturan la señal ruidosa con un único micrófono, esta tarea de limpieza requiere además la estimación de las características del ruido [10].

Cualquiera sea el método escogido para estimar el ruido y la técnica elegida para suprimirlo, el objetivo de este procesamiento es mejorar la calidad del habla y preservar, o incluso incrementar, su inteligibilidad [11]. La evaluación del éxito de los distintos algoritmos en esta tarea no es un problema sencillo. A diferencia de otros campos donde una medida computacional simple (como el error cuadrático medio) puede ser suficiente, la inteligibilidad y la calidad del habla son fenómenos complejos de percepción cuya descripción en términos puramente matemáticos no resulta eficaz. Esto ha conducido a la aceptación general de que las pruebas subjetivas son el mejor instrumento para evaluar en última instancia el desempeño de las distintas tecnologías de procesamiento de voz [12, 13]. Debido a los elevados costos que implican estas pruebas, el desempeño de los nuevos algoritmos propuestos suele evaluarse sólo a partir de un conjunto de medidas analíticas, prescindiendo de la valoración de los oyentes. Numerosos métodos de este tipo han sido propuestos en la bibliografía, cada uno con sus ventajas y limitaciones [3, 14]. La abundancia de medidas objetivas ha conducido al mismo tiempo al uso arbitrario de una diversidad de ellas. Este hecho, sumado a que en general no se ha utilizado un conjunto común de señales para la evaluación, ha dificultado la existencia de una base de referencia verdaderamente comparativa del desempeño de las distintas técnicas de reducción de ruido disponibles.

Por otra parte, la eficacia de los sistemas de procesamiento de voz depende de las características del lenguaje en el cual se aplican [15]. No obstante, la mayoría de las técnicas propuestas son evaluadas en el marco del idioma Inglés, extendiéndose muy pocas veces esta evaluación a múltiples idiomas. Hasta donde conocen los autores, no se dispone actualmente de una base de referencia de algoritmos de reducción de ruido construida

(*) Este trabajo ha sido soportado mediante el PICT 11-12700 "Técnicas no convencionales aplicadas a la reducción de ruido en audífonos digitales", Agencia Nacional de Promoción Científica y Tecnológica y Universidad Nacional de Entre Ríos.

sobre el idioma Español. Más aún, muchas medidas objetivas de evaluación encontradas en la bibliografía tampoco están validadas para distintos idiomas. Por otro lado, estas técnicas de reducción de ruido pueden ser incluidas en prótesis auditivas a fin de mejorar la calidad e inteligibilidad del habla ofrecida por estos dispositivos [16]. En estos casos, el uso de medidas objetivas de evaluación presenta un problema aún mayor, ya que usualmente sus resultados sólo han sido correlacionados con las calificaciones de sujetos normoyentes.

Advirtiendo esta situación, el presente trabajo aborda la evaluación de un conjunto de estrategias clásicas de reducción de ruido en habla, basada en un conjunto común de registros de voz del Español Rioplatense. El trabajo representa además un primer paso hacia la construcción de una base de referencia de técnicas de reducción de ruido para audífonos digitales en este idioma. Para ello, el material de un corpus específicamente diseñado, contaminado con distintos tipos e intensidades de ruido aditivo, fue procesado por un conjunto de algoritmos clásicos de reducción de ruido. La inteligibilidad y calidad del habla obtenida se evaluó por medio de métodos objetivos y subjetivos, incluyendo por el momento únicamente sujetos normoyentes. Las pruebas se orientaron no sólo a establecer el desempeño relativo de las estrategias escogidas, sino también a estudiar la eficacia de un conjunto de medidas objetivas para la predicción de los mismos y a recoger información sobre las aptitudes de cada técnica ante señales con características particulares. Esta información permite identificar las debilidades de las distintas técnicas y dirigir la búsqueda de mejores soluciones sobre la base de un sistema de referencia.

Materiales y métodos

Técnicas de reducción de ruido

La mayoría de las técnicas de reducción de ruido disponibles realizan la limpieza del habla contaminada luego de efectuar alguna transformación sobre la señal temporal. En este sentido, las técnicas clásicas han recurrido usualmente a representaciones en términos de la *Transformada Discreta de Fourier* (DFT, del Inglés *Discrete Fourier Transform*), aunque también se ha propuesto el uso de la *Transformada de Karhunen-Loeve*, la *Transformda Discreta en Cosenos* (DCT, del Inglés *Discrete Cosine Transform*) y la *Transformada Paquetes de Onditas* [17]. Debido a las restricciones impuestas por el retardo admisible para la comunicación en tiempo real, el análisis de la señal de voz se efectúa sobre tramos cortos segmentados a medida que se produce el habla. La transformación elegida se aplica sobre cada uno de estos segmentos, se procesan los coeficientes resultantes de la transformación para suprimir el ruido, y finalmente se invierte la transformación para obtener la representación temporal de la señal mejorada, bajo un esquema de solapamiento y adición de esos segmentos [18].

Las distintas técnicas de reducción de ruido están caracterizadas por los enfoques seguidos para suprimir el ruido a partir de los coeficientes obtenidos mediante la transformación aplicada. Los algoritmos clásicos comprenden diferentes variantes de sustracción espectral [1, 2], técnicas de filtrado óptimo y adaptativo [5, 6], y la estimación de la señal limpia adoptando modelos estadísticos [7, 8]. Todas estas técnicas requieren conocer las características del ruido. Dado que los dispositivos que implementan estos enfoques generalmente disponen de un único micrófono para capturar la señal, estas características usualmente se actualizan sólo durante los intervalos de silencio del hablante. Se han propuesto diversos métodos para estimar estos intervalos [10]. Todos ellos suponen que las características del ruido varían mucho más lentamente que las de la señal de voz. Recientemente, sin embargo, se ha volcado el interés hacia nuevos enfoques que permiten estimar continuamente el ruido de fondo, aun en presencia de voz, permitiendo una mejor estimación de ruidos no-estacionarios [19].

Para este trabajo se escogió un conjunto de técnicas clásicas de reducción de ruido que son frecuentemente utilizadas para comparar el desempeño de nuevas estrategias. La estimación del ruido para todos los algoritmos se actualiza sólo durante los períodos de silencio del hablante, asumiendo que se conocen esos intervalos.

a. Sustracción espectral

Un enfoque intuitivo para la supresión de ruido es substraer una estimación del mismo al habla contaminada. Las técnicas de sustracción espectral realizan esta tarea sobre la representación de la señal ruidosa en el dominio de Fourier [1]. Apoyándose en la mayor importancia de la magnitud en la percepción del habla, el enfoque consiste en estimar la magnitud del espectro de la señal limpia, y asignarle la fase de la señal contaminada. En cada tramo de análisis, la estimación de la magnitud del habla limpia se obtiene restando una estimación del espectro de magnitud del ruido al espectro de magnitud de la señal ruidosa. Debido a que no se substrae el espectro real del ruido sino sólo una versión estimada, es necesario introducir una rectificación para asegurar que los valores de magnitud obtenidos sean positivos.

Las técnicas de sustracción espectral aplican una atenuación dependiente de la relación señal-ruido (SNR, del inglés *Signal to Noise Ratio*) del tramo analizado [10]. Sin embargo, la rectificación introducida y las fluctuaciones de las características reales del ruido con respecto al valor estimado, originan la aparición de una

distorsión conocida como ruido musical [4]. Este ruido se presenta como picos espectrales aislados que aparecen aleatoriamente sobre frecuencias cambiantes entre los distintos tramos de análisis, tomando características altamente no estacionarias. Se han introducido distintas alternativas a fin de reducir estos efectos. La técnica implementada para este trabajo responde a la siguiente expresión, la cual emplea espectros de potencia en lugar de magnitud [2]:

$$\left|\hat{S}_{PSS}(k,n)\right|^2 = \begin{cases} \left|Y(k,n)\right|^2 - \alpha(n)\left|\hat{N}(k,n)\right|^2, & \text{si } \left|Y(k,n)\right|^2 \geq \alpha(n)\left|\hat{N}(k,n)\right|^2 \\ \beta\left|Y(k,n)\right|^2, & \text{si } \left|Y(k,n)\right|^2 < \alpha(n)\left|\hat{N}(k,n)\right|^2 \end{cases} \quad (1)$$

donde $\hat{S}_{PSS}(k,n)$ [1], $Y(k,n)$ y $\hat{N}(k,n)$ son las DFT de la señal limpia estimada, la señal contaminada y la estimación del ruido, respectivamente, correpondientes a la frecuencia discreta *k*, en el tramo *n*. β es un factor que adjudica a las componentes espectrales rectificadas un valor proporcional a las componentes respectivas del habla ruidosa y α es un factor de sobre-sustracción dependiente de la SNR del tramo analizado. La inclusión de estos factores reduce la excursión de los picos espectrales facilitando su enmascaramiento por bandas adyacentes de frecuencias.

Otras alternativas utilizan un factor de sobre-sustracción dependiente de la frecuencia [4], distintas potencias de la magnitud espectral, y la descomposición en bancos de filtros en lugar de la DFT convencional [10]. Trabajos más recientes intentan la incorporación de modelos de percepción auditiva [20], y el empleo de mejores estimadores espectrales [21].

b. Filtrado de Wiener

El filtro de Wiener es una técnica relacionada con las estrategias de filtrado óptimo que intenta minimizar el error cuadrático medio entre la señal limpia y la estimada. Este filtro es el mejor estimador lineal que cumple con esta tarea, y es también óptimo entre los estimadores no lineales cuando los procesos involucrados responden a modelos gaussianos [3, 4]. En el dominio espectral, y teniendo en cuenta que la señal de voz sólo puede considerarse localmente estacionaria, el filtro de Wiener puede expresarse de la siguiente manera:

$$H_W(k,n) = \frac{\hat{\Gamma}_S(k,n)}{\hat{\Gamma}_S(k,n) + \hat{\Gamma}_N(k,n)} \quad (2)$$

donde $\hat{\Gamma}_S(k,n)$ y $\hat{\Gamma}_N(k,n)$ representan estimadores de corta duración de la densidad espectral de potencia de la señal limpia y del ruido, respectivamente. Usualmente se toma $\hat{\Gamma}_S(k,n) = \left|S\left(k,n\right)\right|^2$ y $\hat{\Gamma}_N(k,n) = \left|N\left(k,n\right)\right|^2$.

Dado que no se conoce la densidad espectral de potencia de la señal de voz limpia, en la práctica el filtro debe reducirse a alguna aproximación de la fórmula anterior. Una alternativa simple es utilizar estimadores similares a los utilizados en sustracción espectral. En este caso ambas técnicas quedan vinculadas por una relación de cuadrados. Otro enfoque consiste en estimar el espectro de la señal limpia de forma iterativa, modelando la voz como la respuesta de un sistema autorregresivo [5, 6]. Un tercer enfoque consiste en reformular la expresión (2) en términos de la SNR, y emplear un estimador de la SNR, por ejemplo, como los usados en las reglas de Ephraim-Malah [22]. A pesar de su simpleza, estos estimadores han mostrado tener muy buenas propiedades [23]. La implementación escogida para este trabajo responde a este último tipo.

c. Reglas de supresión de Ephraim-Malah

En lugar de plantear un problema de optimización directamente sobre la señal limpia y la estimada, los estimadores de Ephraim-Malah minimizan el error cuadrático medio sobre la *magnitud* o sobre el *logaritmo de la magnitud* de los coeficientes en la representación de Fourier del habla limpia [7, 8]. El enfoque supone que estos coeficientes, al igual que los del ruido, son variables aleatorias estadísticamente independientes cuyas componentes reales e imaginarias pueden modelarse mediante una distribución gaussiana con media cero. Los estimadores obtenidos logran una elevada reducción de ruido introduciendo menos distorsión, lo que resulta en una disminución del ruido musical.

Para este trabajo se escogió la técnica que responde al siguiente problema de optimización[2] [8]:

$$\hat{S}_{LogSTSA}(k,n) = \arg\min_{\hat{S}} E\left\{\left[\log\left|S(k,n)\right| - \log\left|\hat{S}(k,n)\right|\right]^2 \mid Y(k,n)\right\} \quad (3)$$

[1] El acrónimo *PSS* indica sustracción espectral en potencia (del Inglés *Power Spectral Subtraction*)

[2] El acrónimo *LogSTSA* indica logaritmo de la magnitud del espectro de corta duración (del Inglés *Log-Short-Time Spectral Amplitude*). También suele usarse simplemente *LSA* (del Inglés *Log-Spectral Amplitude*).

donde *E* denota la operación de esperanza. El estimador resultante está dado por la expresión:

$$\left|\hat{S}_{LogSTSA}(k,n)\right| = \frac{\xi(k,n)}{1+\xi(k,n)} \exp\left[\frac{1}{2}\int_{v(k,n)}^{\infty} \frac{e^{-t}}{t} dt\right] \left|Y(k,n)\right| \quad (4)$$

donde $v(k,n) = \xi(k,n)/(1+\xi(k,n))\gamma(k,n)$ y $\gamma(k,n) = |Y(k,n)|^2 / |N(k,n)|^2$. Aquí, $\xi(k,n)$ y $\gamma(k,n)$ representan la SNR *a priori* y *a posteriori*, respectivamente. La eliminación del ruido musical se debe principalmente al estimador propuesto para la SNR *a priori*, $\xi(k,n)$, que muestra un buen compromiso entre su varianza y la velocidad de respuesta a cambios de la señal [23]. Este estimador está dado por [7]:

$$\xi(k,n) = \alpha \frac{\left|\hat{S}(k,n-1)\right|^2}{\left|N(k,n-1)\right|^2} + (1-\alpha) P\left[\gamma(k,n)-1\right] \quad (5)$$

donde *P[•]* es un operador de rectificación y α un factor de actualización.

En la deducción de estos estimadores se supone que existe un aporte de habla en todas las componentes espectrales. Esto no es necesariamente cierto, puesto que existirán tramos de silencio y otros segmentos dominados por fonemas sonoros que mostrarán una concentración de la energía en pocas frecuencias del espectro. Advirtiendo esto, se han propuesto distintas alternativas para incluir la incertidumbre en la presencia de voz [7, 24]. En la versión más simple, se considera un valor constante para la probabilidad *a priori* de ausencia de voz, igual para todas las frecuencias [7]. Estrategias recientes plantean la estimación de esta probabilidad *a priori* segmento a segmento, de acuerdo a información obtenida del plano tiempo-frecuencia [24].

Al igual que en la sustracción espectral, el espectro de fase de la señal limpia estimada por estos métodos es igual al de la señal ruidosa. Otros enfoques similares incluyen transformaciones alternativas a la DFT. En [9], se emplea la DCT, suponiendo también modelos gaussianos para las representaciones obtenidas. Trabajos recientes proponen también el empleo de modelos super-gaussianos, tales como la distribución gamma o la laplaciana [25].

Material de Prueba

El material de habla utilizado para la evaluación de los distintos algoritmos se tomó de la *Batería de Evaluación para Pacientes con Prótesis Auditiva* (BEPPA). Este corpus es un desarrollo conjunto de la Fundación Arauz y la Facultad de Ingeniería de la UNER, destinado a asistir la calibración de audífonos digitales y dispositivos de implante coclear. El material está dividido en varios conjuntos: (a) *consonantes en contexto vocálico,* constituido por una lista con todas las consonantes de uso corriente en el Español Rioplatense, precedidas y sucedidas por la vocal /a/ y con acentuación en la segunda sílaba; (b) *transiciones vocálicas*, constituido por listas de 19 palabras cada una, conteniendo todas las transiciones entre vocales posibles en el español de uso cotidiano; (c) *monosílabos*, constituido por listas de 10 monosílabos cada una; y (c) *oraciones de uso cotidiano*, constituido por listas de 10 oraciones cada una, incluyendo oraciones enunciativas, exclamativas e interrogativas. El material, con excepción del conjunto (a), está compuesto por palabras y frases de uso cotidiano del Español Rioplatense. Esto posibilita que el resultado de las pruebas subjetivas no esté sesgado por el nivel sociocultural de los sujetos participantes, siempre que éstos sean hablantes nativos de este idioma. Todas las señales fueron grabadas en una cámara anecoica, y contaminadas con ruido aditivo blanco (WHITE) y murmullo (BABBLE), tomados de la base de datos NOISEX [26]. La adición de ruido se efectuó computacionalmente con SNRs de -5, 0, 5, 10 y 15 dB, sin tener en cuenta el efecto Lombard. Se tomaron los conjuntos correspondientes a un hablante femenino y un hablante masculino, ambos adultos nativos de la región del Río de La Plata.

Evaluación de la inteligibilidad del habla

La inteligibilidad del habla es una expresión de la dificultad encontrada para su comprensión. La evaluación de la inteligibilidad de las señales procesadas por las técnicas de reducción de ruido seleccionadas se efectuó por medio de pruebas de reconocimiento de palabras. En estas pruebas participaron sujetos normoyentes con edades comprendidas entre los 20 y los 25 años, todos hablantes nativos del idioma Español Rioplatense. Los participantes fueron incluidos en la experiencia luego de comprobar que su audición era normal a través de una audiometría tonal y una logoaudiometría. Sólo se tuvieron en cuenta para esta prueba señales contaminadas con ruido blanco y murmullo con SNRs de -5, 0 y 5 dB. No se utilizaron las relaciones de 10 y 15 dB dado que en experiencias previas los sujetos mostraron reconocer la totalidad de las palabras presentadas en estas condiciones.

Se tomaron 10 registros del desempeño de cada algoritmo en cada condición de ruido y para cada hablante, cada uno incluyendo un listado de todos los conjuntos del corpus. El material procesado se presentó a los sujetos mediante auriculares en forma biaural, con una intensidad de 65 dB SPL. Todas las pruebas tuvieron lugar en una cámara anecoica. La instrucción dada a los participantes consistió en repetir la elocución reproducida al final de la misma. Las respuestas obtenidas se compararon con los listados de referencia de las señales presentadas,

determinando el porcentaje de reconocimiento de palabras brindado por cada técnica en cada una de las condiciones de ruido consideradas. Se construyeron también matrices de confusión de consonantes a fin de visualizar cuáles son las más comprometidas con cada uno de los algoritmos.

Evaluación de la calidad del habla

La calidad del habla puede entenderse como una expresión de la aceptación que muestra el oyente. Distintos factores influyen sobre la percepción de la calidad del habla, tales como la claridad, naturalidad, el contenido de ruido de fondo, disconfort al escucharla, etc. [12]. No hay un acuerdo general entre los autores sobre cuáles son todos los factores que influyen sobre la calidad percibida de la voz, ni sobre la relación que existe entre ésta y la inteligibilidad. Se acepta, no obstante, que la inteligibilidad es una condición necesaria (aunque no suficiente) para obtener una buena calidad de habla [12, 3]. En este trabajo, la calidad de las señales obtenidas con las distintas técnicas de reducción de ruido seleccionadas se evaluó por medio de pruebas subjetivas y de un grupo de medidas objetivas escogidas al efecto. Estas últimas son herramientas computacionales que, si bien se asume que no sustituyen la valoración de los oyentes, intentan predecir estas calificaciones y resultan útiles para efectuar evaluaciones preliminares de las distintas técnicas, permitiendo reducir los costos y el tiempo requerido por las pruebas con personas [13].

a. Pruebas subjetivas para la evaluación de la calidad del habla

Las pruebas subjetivas se realizaron como complemento de las pruebas de inteligibilidad, con la participación de los mismos sujetos y en idénticas condiciones experimentales. La instrucción dada a los participantes fue calificar el habla procesada por cada algoritmo en cada condición de ruido de acuerdo a cuatro aspectos vinculados con la percepción de calidad: claridad, apreciación del ruido residual, confort, y aceptación general de la señal escuchada [27]. En todos los casos se utilizó una escala de calificaciones decimal, siendo de 10 puntos la mejor calificación posible para cada aspecto requerido.

b. Medidas objetivas para la evaluación de la calidad del habla

Las medidas objetivas que evalúan la calidad del habla intentan predecir la calificación que daría un grupo de oyentes en base a medir alguna distancia entre la señal limpia y la obtenida luego del procesamiento. Aunque esta distancia debería ser perceptualmente significativa, las medidas analíticas clásicas se han centrado principalmente sobre diferencias simples en las representaciones temporales o espectrales de ambas señales. Para este trabajo se escogieron las siguientes medidas:

- *SNR por tramos (SegSNR)*: esta medida se centra sobre la diferencia entre la forma de onda de la señal limpia y de la estimada. Se calcula promediando las SNR obtenidas para cada tramo de análisis, donde la SNR de cada segmento puede tomar valores entre -10 y 35 dB. Se supone que fuera de este rango la variación de la SNR no está acompañada de una variación en la percepción del habla. La medida está dada entonces por [3]:

$$SegSNR = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1} s^2(n)}{\sum_{n=Nm}^{Nm+N-1}\left(s(n)-\hat{s}(n)\right)^2} \tag{5}$$

  Aquí s y $\hat{s}$ representan la señal original y la distorsionada o procesada respectivamente, M el número de tramos analizados y N la cantidad de muestras en cada segmento.

- *Log-Likelihood Ratio (LLR)*: esta medida también se la conoce como *distancia de Itakura*, y es una de las más utilizadas en la evaluación de la calidad del habla [28]. Se centra sobre la diferencia entre las representaciones espectrales del habla limpia y la procesada, modelando ambas señales como la respuesta de un sistema autorregresivo [3]:

$$LLR(m) = \log\frac{\vec{a}_{\hat{s}}(m)R_s(m)\vec{a}_{\hat{s}}^{\,T}(m)}{\vec{a}_s(m)R_s(m)\vec{a}_s^{\,T}(m)} \tag{6}$$

  Aquí $\vec{a}_s$ y $\vec{a}_{\hat{s}}$ representan los vectores de coeficientes de predicción lineal para la señal limpia y la procesada, respectivamente, y R_s es la matriz de correlación aumentada de la señal limpia. Puede mostrarse que la medida asigna más peso a las diferencias en los picos espectrales que a aquellas que ocurren en los valles. Esto guarda relación con la percepción del habla, ya que el sistema auditivo es más sensible a errores en la estimación de formantes.

· *Log-Area Ratio (LAR)*: Esta medida también se basa en la diferencia entre los espectros correspondientes a modelos autorregresivos de la señal limpia y de la estimada. En lugar de usar los coeficientes de predicción lineal, esta medida usa los coeficientes de reflexión *r* [3]:

$$LAR(m) = \left[\frac{1}{M} \sum_{i=1}^{M} \left(\log \frac{1+r_s(m)}{1-r_s(m)} - \log \frac{1+r_{\hat{s}}(m)}{1-r_{\hat{s}}(m)} \right)^2 \right]^{\frac{1}{2}} \quad (7)$$

Se ha reportado que esta medida es la que ofrece mayor correlación con los resultados de pruebas subjetivas, entre todas aquellas basadas en el análisis de predicción lineal [28]. Otros trabajos han mostrado que también presenta una buena correlación con la aceptación general de algoritmos por parte de sujetos hipoacúsicos, especialmente a baja SNR [29].

· *Qc*: esta medida da cuenta de la correlación entre las representaciones internas de la señal limpia y de la señal procesada. Estas representaciones perceptuales se obtienen mediante un modelo psicoacústico del sistema auditivo periférico que tiene en cuenta la descomposición de la señal por un banco de filtros *gammatone*, procesos de compresión dinámica y la adaptación de célula ciliadas [30, 31]. Algunos trabajos han reportado que logra una buena correlación con la percepción y valoración del ruido residual por parte de sujetos hipoacúsicos [29].

Resultados y discusión

Los porcentajes de palabras reconocidas logrados con cada algoritmo se resumen en la Tabla 1. Como puede apreciarse, en ruido blanco estos porcentajes son similares para la técnica de sustracción espectral (PSS) y el filtro de Wiener (WIENER), en tanto que los correspondientes a la regla de Ephraim y Malah (LogSTSA) son mayores. El mejor desempeño de esta última es más significativo para la condición de -5 dB. En murmullo, por su parte, WIENER y LogSTSA tienen desempeños comparables, en tanto que el de PSS es significativamente inferior. Comparando la influencia de ambos tipos de ruido puede notarse además que el desempeño de WIENER es mejor con murmullo, mientras que PSS y LogSTSA muestran desempeños mejores en ruido blanco.

TABLA 1: Porcentaje de reconocimiento de palabras

	RUIDO BLANCO			**RUIDO MURMULLO**		
	-5 dB	**0 dB**	**5 dB**	**-5 dB**	**0 dB**	**5 dB**
PSS	84.3	94.1	98.1	76.4	89.3	95.6
WIENER	84.8	94.3	95.5	**87.9**	94.7	**97.6**
LogSTSA	**92.9**	**96.3**	**98.9**	86.9	**95.2**	**97.6**

En la Fig. 1 se muestran las matrices de confusión de consonantes obtenidas en la evaluación de inteligibilidad. Para su construcción se tuvieron en cuenta los errores de reconocimiento registrados con consonantes en contexto vocálico, transiciones vocálicas y monosílabos. También se han sumado las confusiones registradas con las distintas intensidades de un mismo tipo de ruido para cada par de fonemas. Para hacer referencia a los fonemas se utilizó la grafía más cercana a su pronunciación. Las filas corresponden a los fonemas reproducidos durante las pruebas, y las columnas a los fonemas reportados por los participantes. La última fila de cada matriz muestra consonantes insertadas por los oyentes, en tanto que la última columna muestra consonantes que fueron suprimidas en sus respuestas. Por ejemplo, si la elocución presentada es /asa/ y el oyente responde /asta/, se considera que ha insertado el fonema /t/. De forma similar, si la elocución presentada es /chal/ y el oyente reporta haber oído /cha/, se considera que ha suprimido el fonema /l/. No se tuvieron en cuenta los casos en los que los participantes no brindaron ninguna respuesta. Por otra parte, las matrices exhiben sólo las confusiones entre fonemas, eliminándose las diagonales correspondientes a los aciertos. La figura muestra que para la contaminación con murmullo, los algoritmos estudiados presentan dificultades principalmente para la discriminación de las consonantes oclusivas sordas /p/, /t/ y /k/; la oclusiva sonora /g/; y en menor medida para las oclusivas sonoras /b/ y /d/. Vale decir que para todos los casos es también importante la omisión de fonemas registrada en condiciones elevadas de ruido, comprometiendo principalmente los oclusivos y nasales. Por otra parte, es también significativa la inserción de fonemas registrada con sustracción espectral. Puede además notarse una baja confusión de las consonantes fricativas en ruido de murmullo. Esta situación empeora con ruido blanco, volviéndose especialmente significativas las confusiones del fonema fricativo /f/. Esto resalta el efecto de las características espectrales del ruido sobre la inteligibilidad alcanzada. Por último, es necesario recordar que si bien la figura parece mostrar índices de confusión similares

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
D. R. Tomassi, L. Aronson, C. E. Martínez, D. H. Milone, M. E. Torres & H. L. Rufiner; "Evaluación de técnicas clásicas de reducción de ruido en señales de voz"
Revista de la Sociedad Argentina de Bioingeniería, Vol. 11, No. 2, pp. 19-28, 2005.

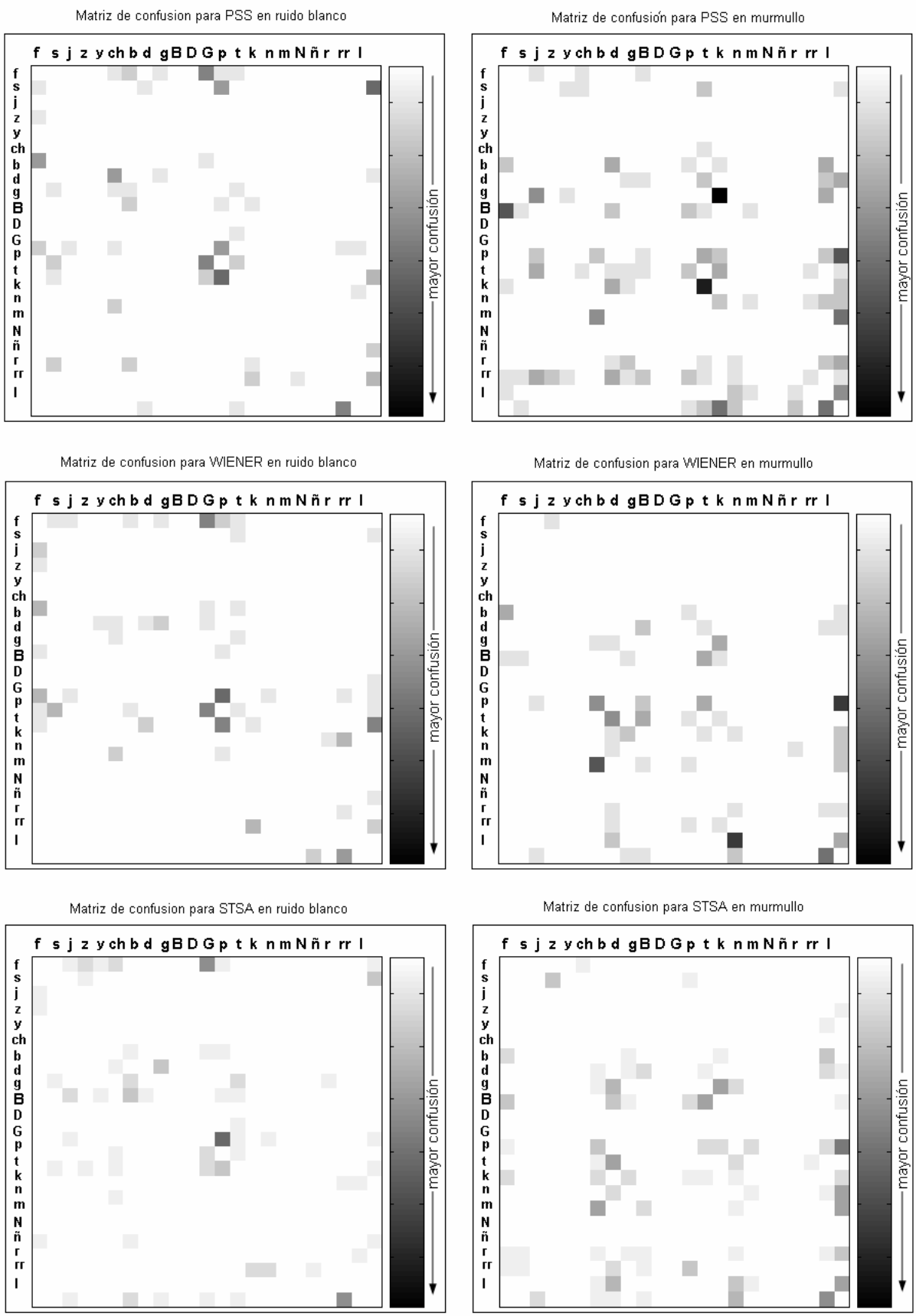

***Fig. 1**: Matrices de confusión de consonantes. Ver comentarios en el texto.*

T**ABLA** 2: Resultados de la evaluación subjetiva de calidad del habla

APRECIACIÓN GENERAL												
	Ruido Blanco						**Ruido Murmullo**					
	-5 dB		**0 dB**		**5 dB**		**-5 dB**		**0 dB**		**5 dB**	
	μ	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
PSS	4.1	1.10	5.4	0.84	6.7	1.06	3.1	0.57	4.9	1.29	6.6	1.07
WIENER	5.5	1.58	6.1	1.52	7.8	1.23	4.2	2.25	**6.3**	1.49	**7.7**	0.67
LogSTSA	**6.2**	1.40	**7.0**	1.25	**8.5**	0.53	**5.0**	1.05	**6.5**	1.08	**7.8**	0.63

CLARIDAD												
	Ruido Blanco						**Ruido Murmullo**					
	-5 dB		**0 dB**		**5 dB**		**-5 dB**		**0 dB**		**5 dB**	
	μ	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
PSS	3.7	1.16	5.4	0.70	6.2	1.75	3.4	0.97	4.7	0.82	6.9	1.20
WIENER	5.3	1.16	6.0	1.33	7.7	1.34	4.4	1.35	**6.3**	0.95	**7.5**	0.85
LogSTSA	**6.1**	1.37	**6.9**	0.88	**8.5**	0.85	**5.0**	0.94	**6.1**	1.20	**7.9**	0.88

CONFORT												
	Ruido Blanco						**Ruido Murmullo**					
	-5 dB		**0 dB**		**5 dB**		**-5 dB**		**0 dB**		**5 dB**	
	μ	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
PSS	3.9	0.99	4.8	1.03	6.5	1.18	3.5	0.71	4.1	0.74	6	1.05
WIENER	5.2	1.62	6.4	1.51	6.9	1.45	**5.0**	1.15	**6.1**	1.20	**7.3**	1.25
LogSTSA	**6.1**	1.60	**7.1**	1.10	**8.2**	0.79	**5.1**	1.29	**6.3**	1.42	**7.3**	0.82

RUIDO RESIDUAL												
	Ruido Blanco						**Ruido Murmullo**					
	-5 dB		**0 dB**		**5 dB**		**-5 dB**		**0 dB**		**5 dB**	
	μ	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**	**μ**	**σ**
PSS	3.9	0.99	4.8	1.23	6.9	1.29	2.7	1.06	4.3	1.77	5.4	0.97
WIENER	3.9	0.99	5.3	1.89	7.6	1.07	3.8	1.75	**6.1**	0.88	**7.0**	0.94
LogSTSA	**5.5**	1.43	**6.9**	1.37	**7.9**	0.74	**5.0**	1.76	**6.3**	1.25	6.8	2.04

para los distintos algoritmos evaluados, tales arreglos sólo contemplan los casos en los que los participantes emitieron una respuesta. El número de elocuciones no repetidas por los oyentes por resultarles incomprensibles es mayor en el caso de sustracción espectral, y menor para el algoritmo de Ephraim-Malah.

Los resultados de las pruebas subjetivas de evaluación de calidad se muestran en la Tabla 2. Se incluye el valor medio y la desviación standard de las calificaciones obtenidas por cada algoritmo. El algoritmo de Ephraim y Malah fue encontrado, en general, superior en todos los aspectos, seguido del filtro de Wiener y por último la técnica de sustracción espectral. Puede también apreciarse que WIENER y LogSTSA alcanzan calificaciones similares en condiciones de contaminación con murmullo con SNRs no negativas.

Para la evaluación objetiva con las medidas consideradas anteriormente, fueron incorporadas además señales contaminadas por ruidos en relaciones de 10 y 15 dB. Los resultados se muestran en la Figura 2, incluyendo sólo los correspondientes a las señales del hablante femenino contaminadas con murmullo. Puede verse que para el caso de LAR, LogSTSA fue encontrada superior a PSS y a WIENER para todas las SNR consideradas (debe tenerse en cuenta que una menor distancia representa una mayor similitud entre la señal procesada y la señal limpia de referencia). Sin embargo, PSS obtiene una mejor calificación que WIENER, a diferencia de lo ocurrido en las pruebas subjetivas. Además, la diferencia de performance entre PSS y WIENER permanece aproximadamente constante en todas las condiciones, mientras que el desempeño superior atribuido a LogSTSA es más evidente a alta SNR. Esto también contradice la concurrencia que muestran los resultados subjetivos para las dos últimas técnicas cuando la SNR aumenta. Para LLR, PSS resulta mejor que WIENER y que LogSTSA en el rango comprendido entre -5 y 5 dB. Esto se opone claramente a los resultados obtenidos en las pruebas subjetivas. Para el caso de SegSNR, PSS y LogSTSA obtienen calificaciones similares en todo el rango de SNRs consideradas, mientras que WIENER recibe en general una valoración inferior. Si bien LogSTSA es también el que obtiene mejor desempeño con esta medida, la calificación relativa entre PSS y WIENER contradice los resultados subjetivos, en tanto que la diferencia con PSS no es significativa. Con Qc, por su parte, WIENER recibe mejor calificación que con los métodos anteriores, en tanto que PSS obtiene la peor valoración. Estos resultados se corresponden mejor con las pruebas subjetivas. No obstante, el bajo desempeño otorgado a LogSTSA, apenas superior a PSS, tampoco está de acuerdo con la calificación de los oyentes.

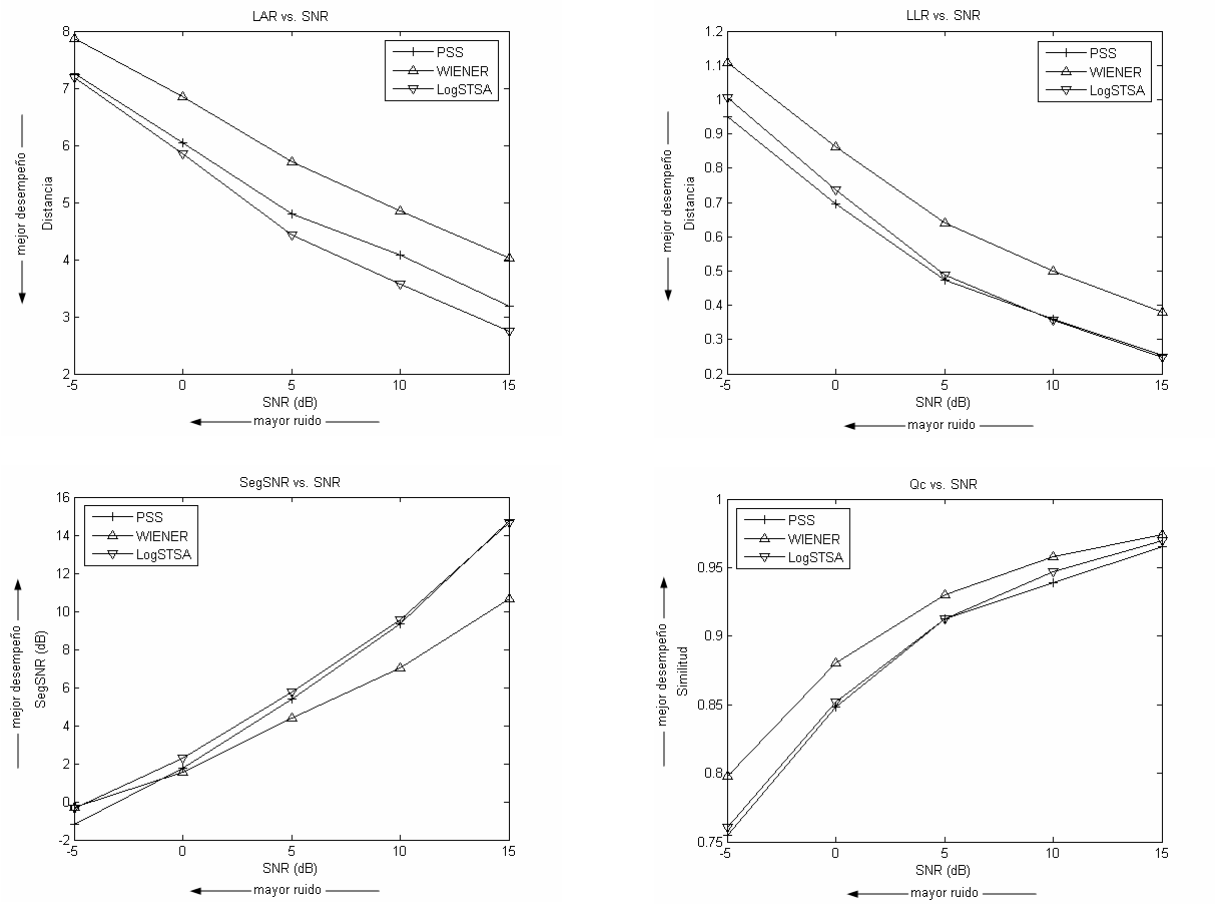

***Fig. 2**: Resultados de la evaluación de la calidad del habla por medio de las medidas objetivas consideradas. Arriba: LAR (izq.) y LLR (der.). Abajo: SegSNR (izq.) y Qc (der.). Ver comentarios en el texto.*

Conclusiones y trabajos futuros

Los resultados obtenidos en las pruebas con sujetos normoyentes muestran que el algoritmo de Ephraim-Malah es la técnica de reducción de ruido con mejor desempeño entre las evaluadas. La superioridad de este algoritmo se refleja en los mayores porcentajes de reconocimiento de palabras alcanzados, y en la mayor aceptación mostrada por los oyentes en la casi totalidad de los criterios y condiciones consideradas. Las medidas objetivas incluidas, en tanto, no alcanzan una buena correlación con estas valoraciones. En consecuencia, no resulta apropiado evaluar el desempeño de las distintas técnicas de reducción de ruido con una sola de estas medidas. Mayores estudios son necesarios para establecer cuáles de ellas y en qué grado se ajustan a la aceptación de los oyentes. Estas pruebas iniciales con BEPPA permiten, no obstante, comprender la complejidad de los procedimientos de evaluación de desempeño de algoritmos de reducción de ruido en habla, y representan un paso inicial hacia la elaboración de un protocolo confiable de evaluación destinado a pacientes hipoacúsicos. La divergencia entre las mediciones subjetivas y objetivas de calidad, sugieren la exploración de nuevas técnicas que contemplen mejor los aspectos perceptuales del proceso de audición. A la vez, la falta de correlación con una medida tan utilizada como LLR enfatiza la necesidad de establecer protocolos universales y realistas de evaluación, a fin de escoger técnicas de referencia para nuevos desarrollos. Asimismo, es necesario considerar los posibles efectos de la variabilidad de juicio de los oyentes, y el número y experiencia de éstos en pruebas de este tipo. Actualmente se está trabajando también en la incorporación de nuevas medidas objetivas para predecir la inteligibilidad del habla y en la ampliación del material de la batería de pruebas. Trabajos posteriores estarán destinados a repetir la evaluación de las distintas técnicas incorporando pacientes hipoacúsicos.

Referencias

1. Boll SF. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics, Speech, and Signal Processing 1979; 27 (2): 521-534.
2. Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise. Proc. IEEE, Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-79: 208-211.
3. Deller JR, Proakis JG, Hansen JHL. *Discrete-Time Processing of Speech Signals*. New Jersey, USA: Prentice Hall, Inc; 1993.
4. Vaseghi SV. *Advanced Digital Signal Processing and Noise Reduction, 2nd Ed.* West Sussex, England: John Wiley & Sons, Ltd; 2000.
5. Lim JS, Oppenheim AV. All-pole modeling of degraded speech. IEEE Trans. Acoustics, Speech, and Signal Processing 1978; 26 (3): 197-210.
6. Hansen JHL, Clements MA. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Processing 1991; 39: 795-805.
7. Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error short-time spectral amplitud estimator. IEEE Trans. Acoustics, Speech, and Signal Processing 1984; 32 (6): 1109-1121.
8. Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitud estimator. IEEE Trans. Acoustics, Speech, and Signal Processing 1985; 33 (2): 443-446.
9. Soon IY, Koh SN, Yeo CK. Noisy speech enhancement using discrete cosine transform. Speech Communication 1998; 24: 249-257.
10. Diethorn EJ. Subband noise reduction methods for speech enhancement. En: Huang Y, Benesty J (Editors), *Audio Signal Processing for Next-Generation Multimedia Systems*. Boston, USA: Kluwer Academic Publishers; 2004: 91-115.
11. Hansen JHL, Pellom B. An effective evaluation protocol for speech enhancement algorithms. Proc. IEEE, Int.Conf. on Spoken Language Processing, ICSLP-98; 7: 2819-2822.
12. Voiers WD, Sharpley AD, Panzer IL. Evaluating the effects of noise on voice communication systems. En: David GM (Editor), *Noise Reduction in Speech Applications*. Boca Raton, USA: CRC Press LLC; 2002.
13. Schmidt-Nielsen A. Intelligibility and acceptability testing for speech technology. En: Syrdal A, Bennett R, Greenspan S (Editors), *Applied Speech Technology*. Boca Raton, USA: CRC Press LLC; 1994.
14. Barnwell TP. Objective measures for speech quality testing. J. Acoust. Soc. Am. 1979; 66 (6): 1658-1663.
15. Chu WC. Speech quality assessment. En Chu WC, *Speech Coding Algorithms. Foundation and Evolution of Standardized Coders*. New Jersey, USA: John Wiley & Sons, Inc; 2003: 501-506.
16. Kates JM. Signal processing for hearing aids. En: Kahrs M, Brandenburg K (Editors), *Applications of Digital Signal Processing to Audio and Acoustics*. New York, USA: Kluwer Academic Publishers, 2002: 235-277.
17. Eatwell GP. Single-channel speech enhancement. En: David GM (Editor), *Noise Reduction in Speech Applications*. Boca Raton, USA: CRC Press LLC; 2002.
18. Rabiner LR, Schafer RW. *Digital Processing of Speech Signals*. New Jersey, USA: Prentice Hall, 1978.
19. Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech and Audio Processing 2001; 9 (5): 504-512.
20. Virag N. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech and Audio Processing 1999; 7 (2): 126-137.
21. Hu Y, Loizou PC. Incorporating a psychoacoustical model in frequency domain speech enhancement. IEEE Signal Processing Letters 2004; 11 (2): 270-273.
22. Cohen I. Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech and Audio Processing 2005; 13 (5): 870-881.
23. Cappé O. Elimination of the musical noise phenomenon with the Ephraim and Malah Noise Supressor. IEEE Trans. Speech and Audio Processing 1994; 2 (2): 345-349.
24. Cohen I, Berdugo B. Speech enhancement for non-stationary noise environments. Signal Processing 2001; 81: 2403:2418.
25. Martin R. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Trans. Speech and Audio Processing 2005; 13 (5): 845-856.
26. Varga A, Steeneken H. Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 1993; 12 (3): 247-251.
27. Arehart KH, Hansen JHL, Gallant S, Kalstein, L. Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners. Speech Communication 2003; 40: 575-592.
28. Hansen JHL, Nandkumar S. Objective speech quality assessment and the RPE-LTP coding algorithm in different noise and language conditions. J. Acoust. Soc. Am. 1995; 97 (1): 609-627.
29. Marzinzik M, Kollmeier B. Predicting the subjective quality of noise reduction algorithms for hearing aids. Acta Acustica 2003; 89: 521-529.
30. Dau T, Puschel D, Kohlrausch A. A quantitative model of the effective signal processing in the auditory system. Part I: model structure. J. Acoust. Soc. Am. 1996; 99 (6): 3615-3622.
31. Hansen M, Kollmeier B. Continuous assessment of time-varying speech quality. J. Acoust. Soc. Am 1999; 106 (5): 2888-2899.