

Evaluación de técnicas clásicas de reducción de ruido en señales de voz

D.R. Tomassi, L. Aronson, C.E. Martínez, D.H. Milone, M.E. Torres y H.L. Rufiner
Facultad de Ingeniería Universidad Nacional de Entre Ríos, dtomassi@ciudad.com.ar

Resumen—El presente trabajo evalúa la inteligibilidad y la calidad de señales de voz luego de ser procesadas por un conjunto de técnicas clásicas de reducción de ruido en habla. La inteligibilidad se mide en porcentaje de palabras repetidas correctamente en una prueba subjetiva de reconocimiento, y se discuten las confusiones fonéticas más frecuentes en términos de matrices de confusión. La calidad de las señales obtenidas se evalúa en forma subjetiva y también objetiva a partir de un conjunto de medidas 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 clave— inteligibilidad, evaluación de calidad del habla, algoritmos de reducción de ruido, prótesis auditivas.

I. INTRODUCCIÓN

EL desarrollo de algoritmos de reducción de ruido para señales de voz se ha visto notablemente intensificado en los últimos años. La evaluación de esos métodos, sin embargo, pocas veces obedece un criterio uniforme entre los distintos autores. Esta falta de uniformidad se manifiesta no sólo en la diversidad de medidas y métodos de evaluación utilizados, sino también en las distintas técnicas de referencia sobre las cuales se comparan los nuevos resultados. Por otra parte, pocas veces se tienen en cuenta los posibles efectos de las características particulares del idioma y dialectos en la validación del desempeño tanto de estrategias de procesamiento como de los métodos para su evaluación. Cuando los nuevos métodos de procesamiento introducidos están destinados a asistir a pacientes con alguna discapacidad auditiva, la situación es aún más desfavorable, ya que muchas veces las medidas disponibles sólo han sido validadas con sujetos normo-oyentes. Advirtiendo esta situación, el presente trabajo representa un primer paso hacia la construcción de una base de referencia de algoritmos de reducción de ruido en habla, basada en una evaluación comparativa de su desempeño en el marco del Español Rioplatense. Para ello, se diseñó una batería de señales de voz contaminadas con distintos tipos e intensidades de ruido (a partir de diferentes relaciones señal-ruido o SNR), la que fue procesada con un conjunto reducido de algoritmos clásicos de supresión de ruido en habla. Las señales obtenidas con cada técnica fueron luego evaluadas en términos de calidad e inteligibilidad del habla.

II. TÉCNICAS DE REDUCCIÓN DE RUIDO

Para este trabajo se escogió un conjunto de estrategias de reducción de ruido que son frecuentemente utilizadas para comparar el desempeño de nuevas estrategias. A continuación se describen brevemente cada una de ellas:

A. Sustracción Espectral:

Esta técnica intenta estimar la señal limpia a partir del espectro de la señal ruidosa, sustrayéndole una versión estimada del ruido que es generalmente actualizada durante los intervalos de ausencia de voz [2]. Existen varias modificaciones a la formulación básica [4], [12]. La alternativa seguida aquí es la correspondiente a [1].

B. Filtro de Wiener:

Este método intenta minimizar el error cuadrático medio entre la señal limpia y la estimada. Entre las aproximaciones más comunes se encuentran las motivadas por la sustracción espectral [12], los enfoques iterativos [4] y las que emplean estimadores recursivos de la relación señal – ruido. En este trabajo se empleó un estimador para la SNR del tipo desarrollado en [6].

C. Estimadores de Ephraim y Malah:

Estos estimadores minimizan el error cuadrático medio para la amplitud o para el logaritmo de la amplitud del espectro de la señal limpia, suponiendo que tanto el ruido como la señal de interés son procesos independientes que siguen una distribución gaussiana [6], [7]. Además de las particularidades de la expresión del estimador, un aspecto importante de estas técnicas es la forma en que estiman la SNR en cada segmento de la señal [3].

III. MATERIAL DE PRUEBA

Para las pruebas se empleó la *Batería de Evaluación de Pacientes con Prótesis Auditiva (BEPPA)*, desarrollada especialmente en forma conjunta por la Fundación Arauz, y la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos (no publicada aún). El corpus consta de un listado de consonantes en contexto vocálico, conjuntos de monosílabos y transiciones formánticas, y conjuntos de frases de uso cotidiano. Todas las señales fueron grabadas en una cámara anecoica, y contaminadas con ruido blanco (WHITE) y murmullo (BABBLE). Las grabaciones corresponden a dos sujetos argentinos nativos, uno de sexo masculino y otro de sexo femenino. Los archivos de ruido se tomaron de la base de datos NOISEX [11], y fueron adicionados computacionalmente en distintas SNRs¹.

IV. MÉTODOS

Las pruebas efectuadas comprenden la evaluación de la inteligibilidad y de la calidad del habla obtenida con las distintas estrategias de procesamiento consideradas.

¹ Sin tener en cuenta el conocido efecto Lombard

A. Evaluación de inteligibilidad:

La evaluación de inteligibilidad se efectuó a través de pruebas subjetivas con sujetos normo-oyentes. Los individuos participantes fueron incluidos en la experiencia luego de comprobar a través de una audiometría tonal que su audición es normal. Las señales contaminadas con ruido blanco y murmullo a SNRs de -5, 0, y 5 dB, y posteriormente procesadas por uno de los algoritmos de reducción de ruido, fueron presentadas en forma binaural a los participantes. El material presentado consistió en una secuencia de palabras de cada uno de los subconjuntos que constituyen la batería. La reproducción del material se efectuó a una intensidad de 65 dB SPL (Sound Pressure Level), y tuvo lugar dentro de una cámara anecoica.

Para determinar el desempeño de cada algoritmo se tomaron diez registros para cada tipo y condición de ruido. Se compararon las respuestas con las referencias de las elocuciones presentadas y se determinaron los índices de error de reconocimiento correspondientes a cada algoritmo en cada condición de ruido. Se construyeron también matrices de confusión de consonantes a fin de visualizar los fonemas más comprometidos con cada uno de los algoritmos. Estos arreglos fueron construidos sólo en base a los registros en los cuales era identificable una sustitución, supresión o inserción de fonemas. Los casos en los cuales la respuesta dada por el oyente difirió notablemente de la referencia presentada no fueron considerados.

B. Evaluación de calidad:

La evaluación de la calidad se realizó mediante una prueba subjetiva y una prueba objetiva. En la prueba subjetiva, los oyentes fueron consultados para calificar el habla procesada en cada condición de ruido, de acuerdo a cuatro dimensiones: claridad, ruido residual, confort, y criterio global de aceptación. En todos los casos se utilizó una escala de calificación decimal, siendo de 10 puntos la mejor calificación posible en todos los casos.

La evaluación objetiva de la calidad del habla se efectuó contemplando cuatro medidas:

Log Area Ratio (LAR): Es una medida basada en la diferencia entre los coeficientes de predicción lineal obtenidos para la señal limpia y para la señal procesada [9]. Su buena correlación con la aceptación general de los algoritmos en pruebas subjetivas tanto con normo-oyentes como con pacientes hipoacúsicos ha sido reportada por algunos autores [10].

Log-Likelihood Ratio (LLR): Esta medida es también conocida con el nombre de *distancia de Itakura*, y al igual que la anterior, se basa en la diferencia entre los coeficientes de predicción lineal para la señal limpia y para la señal procesada [9], pero con una ponderación diferente de los espectros obtenidos.

SegSNR: esta medida se basa en la promediación de la relación señal-ruido obtenida para cada segmento. Si bien ha mostrado no tener una gran correlación con los resultados subjetivos en muchos casos, es una de las medidas más empleadas en el reporte de resultados de nuevas técnicas.

Qc: Esta medida brinda una estimación de la similitud de las representaciones internas de la señal

TABLA I

PORCENTAJE DE PALABRAS MAL RECONOCIDAS

	WHITE			BABBLE		
	-5 dB	0dB	5 dB	-5 dB	0 dB	5 dB
PSS	15,7	5,95	1,90	23,6	10,7	4,36
WIENER	15,2	5,71	4,50	12,1	5,27	2,38
LogSTSA	7,10	3,70	1,05	13,1	4,76	2,38

limpia y de la procesada, basándose en un modelo psicoacústico del proceso periférico de audición [5], [8]. Algunos autores [10] han reportado su buena correlación con datos subjetivos asociados a pacientes hipoacúsicos.

V. RESULTADOS

El porcentaje de palabras mal reconocidas para cada algoritmo se resume en la Tabla I. 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 significativamente menores. 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 similar en ruido blanco y con murmullo, mientras que PSS y LogSTSA muestran performances sensiblemente mejores en ruido blanco.

En la Figura 1 se muestran dos ejemplos de matrices de confusión obtenidas en la evaluación de inteligibilidad. En la construcción se han sumado las confusiones originadas con las distintas intensidades de un mismo tipo de ruido para cada par de fonemas. La correspondencia entre la numeración de los ejes y los respectivos fonemas se muestra en la Tabla II. La última fila de cada matriz muestra consonantes insertadas por el oyente, en tanto que la última columna muestra consonantes que fueron suprimidas en esta respuesta. 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/, /g/, oclusiva sonora; y en menor medida para /b/ y /d/, oclusivas sonoras. 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 el elevado porcentaje de reconocimiento de las consonantes fricativas en ruido de murmullo. Esta situación disminuye con ruido blanco, volviéndose especialmente significativas las confusiones del fonema fricativo /f/ (no mostrado en la figura). Esto evidencia la influencia de las características espectrales del ruido en la inteligibilidad alcanzada. Por último, es necesario señalar que si bien la figura parece mostrar índices de confusión similares para los distintos algoritmos evaluados, es importante recordar que tales arreglos están construidos sólo sobre las confusiones identificables. El número de elocuciones no repetidas por los oyentes por resultarles incomprensibles es mayor en el caso de sustracción espectral, al tiempo que menor para las reglas derivadas en [6], [7].

TABLA II

CORRESPONDENCIA ENTRE LOS EJES DE LAS MATRICES DE CONFUSIÓN Y LOS FONEMAS REPRESENTADOS EN LA FIG. 1.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
/f/	/s/	/j/	/z/	/ch/	/b/	/d/	/g/	/N/	/D/	/G/	/p/	/t/	/k/	/n/	/m/	/N/	/ñ/	/r/	/rr/	/l/	/ll/

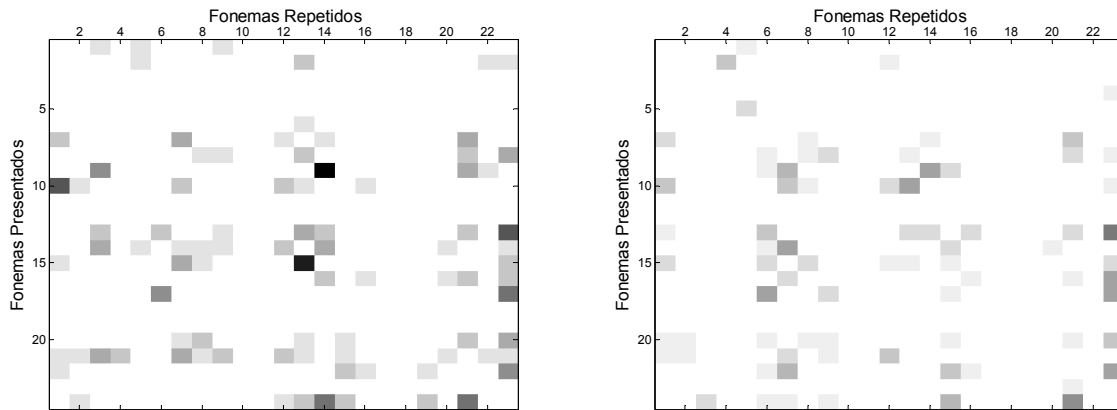


Fig. 1: Matriz de confusión para PSS (izq.) y STSA (der.) en ruido murmullo.

TABLA III

RESULTADOS DE LAS PRUEBAS SUBJETIVAS DE EVALUACIÓN DE LA CALIDAD DEL HABLA.

	APRECIACIÓN GLOBAL						CLARIDAD					
	WHITE			BABBLE			WHITE			BABBLE		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
PSS	4.00	5.29	6.71	3.00	4.86	6.71	3.71	5.22	6.29	3.14	4.71	7.00
WIENER	5.43	6.29	7.70	4.43	6.43	7.70	5.14	6.00	7.57	4.43	6.29	7.43
LogSTSA	6.25	7.12	8.50	8.50	5.00	6.44	7.89	6.12	6.87	8.50	5.00	6.11
	CONFORT						RUIDO RESIDUAL					
	WHITE			BABBLE			WHITE			BABBLE		
	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB	-5 dB	0 dB	5 dB
PSS	3.86	4.71	6.57	3.43	4.14	6.00	3.86	4.86	6.85	2.57	4.43	5.43
WIENER	5.14	6.43	7.00	5.00	6.00	7.29	3.71	5.29	7.57	3.86	6.00	7.00
LogSTSA	6.25	7.12	8.12	5.22	6.22	7.44	5.5	6.87	7.87	5.00	6.22	6.78

Los resultados de las pruebas subjetivas de evaluación de calidad se muestran en la Tabla III. Como puede apreciarse, 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 SNR no negativa.

Para la evaluación objetiva con las medidas consideradas anteriormente, fueron incorporadas además señales contaminadas por ruidos en relaciones de 10 y 15 dB. También se incluyó otra variante de los algoritmos de Ephraim-Malah (STSA), y una técnica basada en una transformación de la señal en cosenos discretos (DCT). Los resultados se muestran en la Figura 2, para contaminación con murmullo. Puede verse que para el caso de LAR, LogSTSA fue encontrada superior a PSS y a WIENER en 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 es mejor ponderada que WIENER, a diferencia de lo ocurrido en las pruebas subjetivas. Además, la diferencia de performance entre LogSTSA y PSS permanece aproximadamente constante en todas las condiciones, mientras que el desempeño relativamente inferior atribuido a WIENER es más evidente a alta SNR. Nótese también que las técnicas STSA y DCT muestran el

mejor desempeño para SNRs menores a 5 dB. Para LLR, PSS resulta mejor que WIENER y que LogSTSA en el rango comprendido entre -5 y 5 dB. Aún más, en la mayor parte de este intervalo WIENER es superior a LogSTSA. Todo esto contradice los resultados obtenidos en las pruebas subjetivas. Para el caso de SegSNR, puede apreciarse que todas las técnicas obtienen una calificación similar, con excepción de WIENER, que es peor valorado. Si bien LogSTSA es también el que obtiene mejor desempeño con esta medida, la calificación relativa entre PSS y WIENER contradice los resultados subjetivos. Por otra parte, la similitud en la valoración de las distintas técnicas parece que no permite realizar un ordenamiento confiable de sus desempeños. Con Qc, por su parte, PSS es valorada pobremente, en tanto que el resto de las técnicas obtiene una calificación similar más alta. Dentro de ellas, WIENER es mejor valorada que STSA, lo cual también contradice los resultados subjetivos obtenidos con normo-oyentes.

VI CONCLUSIONES

Estas pruebas iniciales con BEPPA permiten 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

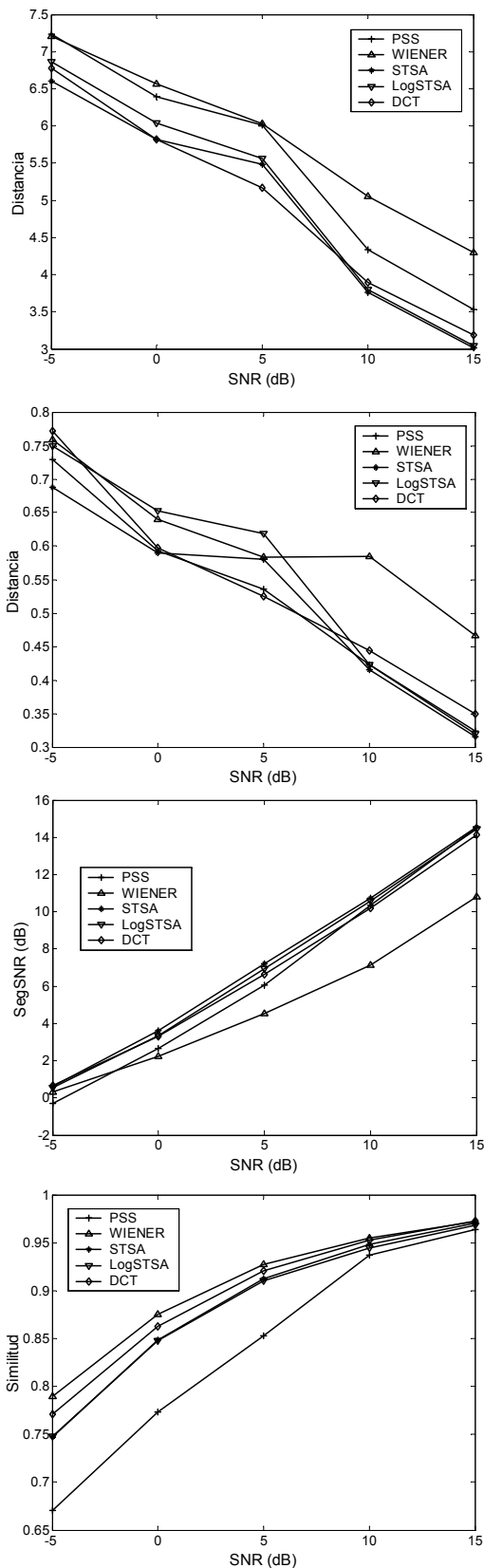


Fig. 2 (arriba): Resultados de la evaluación objetiva de la calidad del habla procesada con los algoritmos considerados. Arriba, estimación de la calidad para distintas SNR utilizando LAR como medida. Más abajo, comparación usando LLR. Luego, comparación por medio de SegSNR. Abajo, por último, estimación de calidad empleando Qc. En todos los casos los resultados mostrados corresponden a habla femenina contaminada con ruido babble.

subjetivas y objetivas de calidad, sugieren la exploración de nuevas técnicas. A la vez, esta 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. Por otra parte, deben ser profundizados los estudios acerca de la influencia de la estructura fonética de la batería de señales sobre los índices de reconocimiento. Ésto es también importante para futuras exploraciones de medidas objetivas para predecir la inteligibilidad del habla.

REFERENCIAS

- [1] M. Berouti, R. Schwartz y J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise", *Proc. of the International Conference on Acoustics, Speech and Signal Processing*, pp. 208-211, 1979.
- [2] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction". *IEEE Trans. On Acoustics, Speech, and Signal Processing*, vol. 27, No2, pp. 113-120, 1979.
- [3] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor", *IEEE Trans. On Speech and Audio Processing*, vol. 2, pp. 345-349, 1994.
- [4] J. Deller, J. Proakis y J. Hansen, "Discrete-Time Processing of Speech Signals", Prentice Hall, 1993.
- [5] T. Dau, D. Puschel y A. Kohlrausch, "A Quantitative Model of the Effective Signal Processing in the Auditory System. Part I: Model Structure". *Journal of the Acoustical Society of America*, vol. 99, No 6, pp. 3615-3622, 1996.
- [6] Y. Ephraim y D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short Time Spectral Amplitude Estimator", *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109-1121, 1984.
- [7] Y. Ephraim y D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator". *IEEE Trans. On Acoustics, Speech and Signal Processing*, vol. 33, pp. 443-445, 1985.
- [8] M. Hansen y B. Kollmeier, "Continuous Assessment of Time-Varying Speech Quality". *Journal of the Acoustical Society of America*, vol. 106, No 5, pp. 2888-2899, 1999.
- [9] J.H.L. Hansen y B. Pellom, "An Effective Quality Evaluation Protocol for Speech Enhancement Algorithms", en *Proc. of the International Conference on Spoken Language Processing*, vol. 7, pp. 2819-2822, 1998.
- [10] M. Marzinzik y B. Kollmeier, "Predicting the Subjective Quality of Noise Reduction Algorithms for Hearing Aids", *Acta Acustica*, vol. 89, pp. 521-529, 2003.
- [11] A. Varga y H. Steeneken, "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*, vol. 12, No 3, pp. 247-251, 1993.
- [12] S.V. Vaseghi, "Advanced Digital Signal Processing and Noise Reduction", 2nd Edition, John Wiley & Sons, 2000.