Asociación Argentina
de Mecánica Computacional

Mecánica Computacional Vol XXVI, pp.2152-2161
Sergio A. Elaskar, Elvio A. Pilotta, Germán A. Torres (Eds.)
Córdoba, Argentina, Octubre 2007

MODELOS DE AUDICIÓN Y ANÁLISIS TIEMPO –FRECUENCIA PARA LA EVALUACIÓN DE TÉCNICAS DE MEJORA DE LA SEÑAL DE VOZ

Elizabeth Vera de Payer, Juana Armesto, Marina Voitzuk

Laboratorio de Procesamiento de Señales – Departamento Matemática
Facultad de Ciencias Exactas, Físicas y Naturales – Universidad Nacional de Córdoba- Argentina
Av. Vélez Sarzfield 1601 – (5000) Córdoba-
epayer@com.uncor.edu - juanaarmesto@arnet.com.ar - mvoitzuk@yahoo.com.ar

Palabras Clave: relación señal-ruido, análisis tiempo-frecuencia

Resumen: En este trabajo se analiza un problema que presentan varias técnicas usuales de mejora de la señal de voz utilizadas cuando se dispone de un sólo micrófono con una relación señal/ruido (SNR) por debajo de los 5 dB y se desea aumentar la inteligibilidad como sucede en los sistemas portátiles compactos tales como audífonos y teléfonos celulares. Es práctica común que ellas accionen sobre la relación señal/ruido aun a costa de provocar distorsión de la señal e introducir un ruido residual. Pero la dificultad radica en que, si bien la SNR es una medida objetiva muy fácil de computar, ella refleja sobre todo, la calidad percibida pero no la inteligibilidad de la señal de voz.
El problema de separar las nociones de calidad percibida e inteligibilidad es debida, en parte, a la imposibilidad de aislar y caracterizar aquellas cualidades de la señal de voz que son propias de cada una de ellas. Esta es la razón por la cual sobre todo cuando se quiere evaluar la inteligibilidad, se utilizan medidas subjetivas como el ¨diagnostic rhyme test¨ (DRT), basadas en la opinión de grupos de escuchas donde se presentan palabras que difieren sólo en la consonante principal. La desventaja de este tipo de test es que puede presentar un sesgo dependiente del oyente por lo que se han realizado esfuerzos para lograr establecer un protocolo de evaluación de calidad total para estos algoritmos de mejora. El principal escollo reside en la no uniformidad de la distorsión provocada por el ruido, tanto en tiempo como en frecuencia. Por esta razón, el objetivo aquí planteado es analizar la vinculación de la relación señal/ruido con algunas características de la señal de voz.
La introducción de modelos que simulan la acción de la membrana basilar ha permitido reformular los algoritmos de mejora logrando con ello perfeccionar su desempeño, como así también desarrollar medidas objetivas de la calidad total que toman en cuenta las propiedades del sistema de audición periférico, incluyendo la percepción de sonoridad, banda de frecuencia y el fenómeno de enmascaramiento.
En el presente trabajo se explota la no estacionariedad de la señal de voz analizando en el dominio tiempo-frecuencia como se modifican con la relación señal/ruido, algunos parámetros importantes sobre los distintos fonemas y en sus respectivas bandas de frecuencia, usando para ello la distancia de Jensen-Rényi generalizada y una versión adaptada de la distancia Itakura-Saito.
Se observa que no siempre es posible establecer un ordenamiento que haga corresponder mayor valor de SNR con menor distancia entre sus representaciones tiempo-frecuencia sobre las distintas bandas de frecuencia. Esta situación se agudiza en las consonantes. Estos hechos pueden interpretarse como que la relación señal/ruido si bien da indicaciones globales respecto a las características de la señal limpia comparada con la señal ruidosa, no siempre refleja con precisión los efectos del ruido sobre las bandas críticas.

1. INTRODUCCIÓN

La calidad total de la señal de voz sobre los sistemas de comunicación se ve generalmente reducida por ruido interferente el cual puede tomar la forma de ruido ambiente, reverberación, otras voces o ruido de los canales eléctricos entre otros.

Esto impone la necesidad de implementar técnicas de mejora para la reducción de ruido las cuales dependen:

1) del números de canales disponibles (técnicas de canal simple o multicanal)
2) de como el ruido se mezcla con la señal de voz (ruido aditivo, multiplicativo, convolucional)
3) de la relación estadística entre el ruido y la señal (incorrelado o aun ruido independiente, ruido correlacionado como eco y reverberación)

En este contexto, a la calidad total la podemos considerar como formada por dos factores: la inteligibilidad, asociada a la posibilidad de comprender lo hablado, y la calidad percibida que está vinculada a no generar cansancio en el auditorio y resultar agradable al oyente. Los métodos usuales de mejora están dirigidos al aumento de la SNR con lo que en general se logra actuar sobre la calidad percibida (Rix et al, 2006).

El caso que nos ocupa es el de mejora de la inteligibilidad, disponiendo de un solo micrófono para una señal contaminada con ruido ambiente, aditivo e incorrelado, y una relación señal/ruido por debajo de los 5 dB, situación que se presenta en sistemas portátiles tales como audífonos y teléfonos celulares.

El problema de separar las nociones de calidad percibida e inteligibilidad es debida, en parte, a la dificultad de aislar y caracterizar aquellas cualidades de la señal de voz que son propias de cada una de ellas (Ephraim et al, 2003). Esta es la razón por la cual sobre todo cuando se quiere evaluar la inteligibilidad, se utilizan medidas subjetivas como el ¨diagnostic rhyme test¨ (DRT), basadas en la opinión de grupos de escuchas a quienes se presentan palabras que difieren sólo en la consonante principal. La desventaja de este tipo de test es que puede presentar un sesgo dependiente del oyente por lo que se han realizado esfuerzos para lograr establecer un protocolo de evaluación de calidad total para estos algoritmos de mejora (Hansen y Pellom, 1998). Nuevos métodos de medición objetiva están siendo desarrollados, como es el caso del PEMO-Q (Huber y Kollmeier, 2006) para la valoración de la calidad de audio, que se basa en aplicar un modelo del sistema auditivo periférico a un par de señales, una de referencia y otra de test, y tomar los coeficientes de correlación de las salidas a fin de medir la similitud perceptual de las señales. Debido a la gran fidelidad de las señales de referencia, es razonable interpretar cualquier desviación perceptible como una disminución de la calidad total.

El objetivo de estas mediciones es el de cuantificar la degradación sufrida por la señal contaminada respecto a la señal original y poder predecir la calidad subjetiva observada por un panel de oyentes. El principal escollo reside en la no uniformidad de la distorsión provocada por el ruido tanto en el tiempo, esto es sobre los distintos fonemas, como en frecuencia considerando las bandas críticas.

De aquí la importancia de establecer como se modifican algunos parámetros destacados de la señal de voz debido al impacto diferenciado del ruido en el dominio tiempo-frecuencia.

2. SISTEMA DE AUDICIÓN

La percepción de una señal de audio es el resultado de la acción conjunta de varios

fenómenos fisiológicos y psicológicos cuyas vinculaciones no están todavía totalmente comprendidas. Sin embargo algunos modelos que describen estos efectos han sido implementados con éxito, permitiendo el desarrollo de nuevas técnicas de procesado de voz.

La cóclea, elemento básico en el sistema auditivo, se comporta como un banco de filtros pasabanda cuasi Q-constante, los que están finamente sintonizados a una frecuencia particular y que describen el nivel de umbral de sonido necesario para activar una determinada porción de la membrana basilar en comunicación con las neuronas auditivas que conducen los estímulos al cerebro (Giguére y Woodland, 1994). Esto es, la resolución en frecuencia del oído humano no es uniforme y se lo describe usualmente en base a las llamadas bandas críticas. A partir de este concepto se han desarrollado varios modelos del sistema de audición que han podido usarse muy satisfactoriamente tanto en el codificado de voz como en la optimización de procesadores de audio (Kollmeier, 2005).

Uno de los modelos más difundidos es el filtro Gammatone introducido por Johannesma (1972) del cual se han realizado varias versiones (Lyon, 1982; Seneff, 1988; Irino y Patterson, 1997).

En el modelo propuesto por Patterson y Holdsworth la respuesta al impulso de cada filtro del banco se expresa:

$$g(t) = at^{N-1}\exp(-2\pi bERB(f_c)t).\cos(2\pi f_c t + \phi)) \qquad (1)$$

con:

N : orden del filtro

f_c : frecuencia central

$ERB(f_c)$: ancho de banda rectangular equivalente del filtro de audición

a, b : constantes reales de normalización

El ERB (Equivalent Rectangular Bandwidth) es una medida psicoacústica del ancho de banda del filtro de audición para cada punto a lo largo de la cóclea y equivale a las bandas críticas. Para niveles de potencia moderados corresponde a la expresión:

$$ERB(f_c) = 24.7(4.37\frac{f_c}{1000} + 1) \quad \text{(Glasberg y Moore, 1990)} \qquad (2)$$

En el presente trabajo se utilizará el modelo de Patterson y Holdsworth de 18 canales desarrollado por M. Slanley (1993).

Este modelo simplificado deja de lado algunos fenómenos no lineales tales como la co-resonancia de la membrana tectoria y mecanismos de realimentación por las células ciliares externas. Sin embargo, permite simular en una primera aproximación, la selectividad en frecuencia del sistema auditivo periférico.

3. ANALISIS CONJUNTO TIEMPO FRECUENCIA

Tomando en consideración la no estacionariedad de la señal de voz, se hace necesario la utilización de técnicas tiempo-frecuencia. Se elige inicialmente la Distribución Zhao-Atlas-Marks (ZAM) o Distribución Cono (Zhao, Atlas y Marks, 1990; Ho y Marks, 1992) que, siendo una representación tiempo-frecuencia (RTF) perteneciente a la Clase de Cohen (Cohen, 1989), presenta la ventaja de suprimir los términos interferentes entre aquellas componentes que tienen la misma frecuencia central. Esta propiedad permite un buen análisis visual. Si bien los gráficos de estas RTF permiten observar algunas características de las

señales estudiadas, es posible además realizar mediciones a fin de traducir en números las diferencias apreciadas en las imágenes.

Una aproximación al problema explota la analogía entre la densidad de energía de la señal y las densidades probabilísticas (Flandrin, 1984). Así como $|x(t)|^2$ y $|X(f)|^2$ se comportan como densidades unidimensionales en tiempo y frecuencia respectivamente de la energía de la señal, algunas RTF de la forma $C_x(t,f)$ pueden considerarse que actúan como densidades bidimensionales de energía en el plano tiempo-frecuencia

En Teoría de la Información, la entropía sirve de base para mediciones de divergencia y distancia entre densidades probabilísticas como así también para cuantificar la complejidad y el grado de incerteza de las mismas. Este hecho ha motivado la aplicación de fórmulas similares para las RTF aunque sin perder de vista que estas últimas no se comportan exactamente como densidad de probabilidades debido a la no positividad de la mayoría de ellas, lo que impide la aplicación de la entropía de Shannon. Williams, Brown y Hero (1991) propusieron el uso de la entropía de Rényi generalizada para señales de energía unitaria:

$$H_\alpha(C_x) = \frac{1}{1-\alpha}\log_2 \iint C_x^\alpha(t,f)\,dt\,df \tag{3}$$

Estudios empíricos mostraron muy buenos resultados para α=3 abarcando un importante conjunto de señales aun en el caso en que la RTF tome localmente valores negativos, gozando de algunas propiedades realmente destacadas.

Una condición a cumplir por la RTF para que la entropía de Rényi de orden α esté bien definida es que sea real y satisfaga:

$$\iint C_x^\alpha(t,f)\,dt\,df > 0 \tag{4}$$

Tanto la Distribución Cono como la Distribución de Wigner-Ville (DWV) (Cohen, 1989) de las señales de voz satisfacen este requerimiento, al menos para $\alpha = 3$.

Las RTF con *kernels* pasa-bajo conducen normalmente a estimaciones de la entropía de Rényi más robustas que cuando se usa la Distribución de Wigner-Ville debido a la atenuación de los términos interferentes. Como contrapartida, esto se logra a costa de que los niveles de entropía tengan un sesgo dependiente de la señal (Aviyente y Williams, 2003), a pesar de lo cual se utilizará inicialmente la Distribución Cono.

Una manera usual de derivar mediciones de distancia entre densidades probabilísticas es a partir de la diferencia de Jensen (Rao y Nayak, 1985) que adaptada con utilización de la media geométrica y la entropía de Rényi conduce a la distancia de Jensen-Rényi:

$$J(C_1,C_2) = H_\alpha\sqrt{C_1.C_2} - \frac{H_\alpha(C_1)+H_\alpha(C2)}{2} \tag{5}$$

la que a su vez, puede generalizarse para su aplicación para RFT no positivas (Aviyente et al, 2004).

A los efectos de realizar comparaciones, es de interés aplicar además otras medidas de distancia entre las RTF como la euclidiana (6) o algunas de sus variantes:

$$D_{L2} = \left[\int\int |C_1(t,\omega) - C_2(t,\omega)|^2\,dt.d\omega\right]^{1/2} \tag{6}$$

Sin embargo, existe un problema de dimensionalidad: para una señal a tiempo discreto de N puntos, la RTF tiene N.M puntos (si asumimos que el eje de frecuencias tiene M puntos), por lo que aumenta considerablemente la carga computacional.

Una solución es considerar que así como las funciones de densidad de probabilidades pueden ser caracterizadas por algunos de sus momentos, en particular el valor medio, la variancia, la asimetría y el apuntamiento, de la misma forma, la representación conjunta tiempo-frecuencia de señales no estacionarias, a pesar de no ser en rigor una densidad probabilística, pueda ser también caracterizada por sus momentos de bajo orden (Tacer y Loughlin, 1998; Davidson y Loughlin, 2000). Los momentos conjuntos tiempo-frecuencia de la representación $C_x(t,\omega)$ pueden calcularse a partir de la expresión:

$$< t^n, \omega^m > = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} t^n \omega^m C_x(t,\omega)\,dt\,d\omega \qquad (7)$$

Por otro lado, como sucede en particular con la Distribución de Wigner-Ville, algunos de éstos, si se trabaja con la forma analítica de la señal, tienen una interpretación física concreta que los hace más interesantes, como es el caso de los momentos de primer orden en frecuencia y en el tiempo que se corresponden con la frecuencia instantánea (8) y retardo de grupo (9) respectivamente. Así mismo se tomará el momento conjunto de segundo orden (10).

$$fm(t) = \frac{\int_{-\infty}^{\infty} f.C_x(t,f)\,df}{\int_{-\infty}^{\infty} C_x(t,f)\,df} \quad (8) \qquad tm(f) = \frac{\int_{-\infty}^{\infty} t.C_x(t,f)\,dt}{\int_{-\infty}^{\infty} C_x(t,f)\,dt} \quad (9) \qquad m_1^1 = \int t.\omega\,.C(t,\omega)\,dt.d\omega \qquad (10)$$

Resultan también de interés las llamadas propiedades marginales que representan la densidad espectral de energía (11) y la potencia instantánea (12), obtenidas como distribuciones marginales de $C_x(t,f)$, como así también la energía total (13).

$$margt(f) = \int_{-\infty}^{\infty} C_x(t,f)\,dt \quad (11) \qquad margf(t) = \int_{-\infty}^{\infty} C_x(t,f)\,df \quad (12) \qquad E = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} C_x(t,f)\,dt\,df \qquad (13)$$

Cuando se intenta cuantificar el grado de complejidad de una señal, mediciones basadas en los momentos segundo orden como son el producto duración en el tiempo-ancho de banda, no aportan resultados. En este caso, también la entropía nos brinda una solución ya que es una medida de la información esperada a través de todos los valores que puede tomar un proceso aleatorio, luego es mayor mientras mayor es el grado de incerteza, lo que se relaciona con que la señal sea más compleja (Baraniuk y Flandrin, 2001). De aquí que se espere poder medir la complejidad de la señal a través de la entropía de su RTF ya que es ésta y no la señal, la que se comporta como una función de densidad de probabilidades. A medida que la señal es más compleja, también lo es su RTF y en consecuencia, mayores los valores de entropía.

4. MEDICIÓN DE DISTANCIA Y VECTOR DE CARACTERÍSTICAS

A los fines de establecer la correspondencia entre el impacto de ruido medido por su SNR y la vinculación con su RTF de los distintos fonemas y bandas de frecuencia, se tomará una

frase contaminada por distintos niveles de ruido y se medirá con la distancia de Jensen-Rényi generalizada la diferencia de las RTF entre las señales ruidosas y limpias.

Además se formarán vectores de características constituidos por elementos que se desprenden del análisis conjunto tiempo-frecuencia: frecuencia instantánea, retardo de grupo, marginales en tiempo y frecuencia y entropía de Rényi.

Los momentos son previamente normalizados para estabilizar la variancia y reducir el rango dinámico. El proceso se repite para la misma frase contaminada con distintos niveles de ruido. La distancia entre los vectores se mide con una versión adaptada de la distancia Itakura- Saito (1969)

$$d_{I-S} = \frac{(b-a)*M*(b-a)'}{a*M*a'} \qquad (14)$$

con *a:* vector que representa a la señal limpia
b: vector que representa a la señal contaminada.
M: matriz de autocorrelación del vector *a*

5. PARTE EXPERIMENTAL

Se tomó la frase **¨las cuentas cerraron bien¨** dicho por cinco hablantes femeninos adultos. Se generaron artificialmente señales ruidosas adicionando a la señal limpia, ruido blanco Gaussiano con una relación señal/ruido global de 0, 2, 4 y 10 dB.

Se seleccionaron segmentos particulares de la frase, como vocales (/a/) (en ¨las¨, 768 muestras), consonantes (/l/) (en ¨las¨, 768 muestras), (/n/ en ¨cerraron¨, 512 muestras) y los diptongos ue (en ¨cuentas¨ , 768 muestras) y ie (en ¨bien¨, 512 muestras). Para apreciar la incidencia del ruido sobre cada fonema, se calculó la SNR en cada tramo.

	0dB	**2dB**	**4dB**	**10dB**
/l/	0.0798	2.2046	4.718	10.254
/a/	8.386	10.467	12.568	18.592
ue	6.6384	8.7632	10.872	16.101
/n/	-4.6106	-2.4858	-0.3516	5.1712
ie	0.2854	3.4103	5.6868	11.614

Tabla 1: Relaciones Señal-Ruido en los tramos seleccionados

En la Tabla 1 se puede observar un impacto fuertemente diferenciado del ruido en los distintos tramos. Las señales contaminadas se han creado artificialmente incorporando a la señal limpia ruido blanco Gaussiano, con una relación señal/ruido global de 0 dB, 2 dB, 4 dB y 10 dB. Sin embargo según puede apreciarse, no se refleja de la misma forma en los distintos tramos, lo cual era de esperar si se toma en cuenta que la potencia de ruido es casi constante mientras que la señal original presenta potencia variable a través de los distintos fonemas. Así, para el tramo representativo del fonema /a/, la relación señal/ruido es muy elevada presentando los valores más bajos para los que corresponden a las consonantes /l/, /n/. Los diptongos presentan valores intermedios.

Se pasó a analizar con técnicas tiempo-frecuencia los distintos fonemas seleccionados tanto de la señal limpia como ruidosa, usando para ello la Distribución Cono.

En la Figura 1 se observa que aunque globalmente la frase se contaminó con ruido Gaussiano con SNR de 0 dB, la influencia en el fonema /a/ es mucho menor que en el fonema /l/ considerando la representación en el dominio tiempo-frecuencia en coincidencia con la mayor relación señal/ruido del fonema /a/ frente a los restantes.

Para medir las diferencias de las RTF de la señal limpia y las señales ruidosas de cada fonema se utilizó la distancia de Jensen-Rényi (J-R) generalizada (Tabla 2).

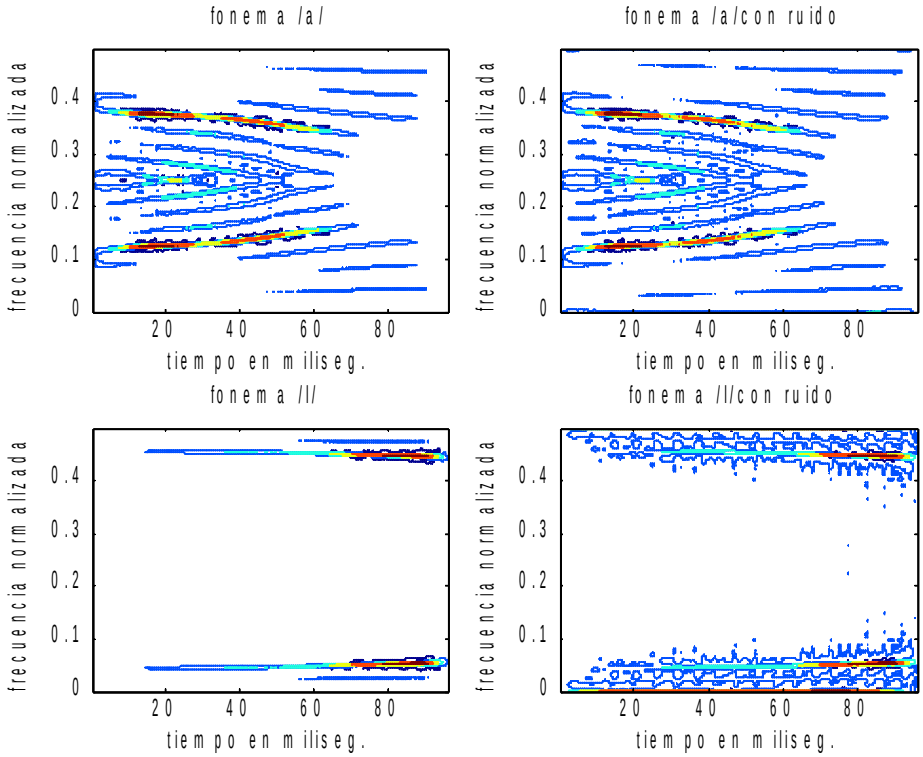

Figura 1: Distribución Cono de los fonemas /a/ y /l/ sin ruido y con ruido a 0 dB

Comparando la Tabla 1 y la Tabla 2 puede apreciarse que existe una correspondencia directa entre mejor relación señal/ruido y menor distancia de Jensen-Rényi.

A fin de ratificar esta coincidencia, se formaron vectores representativos de las RTF correspondientes. Si bien la Distribución Cono permite una mejor visualización de las características de la señal, esto es a costa de no gozar de algunas propiedades destacadas que sí posee la Distribución de Wigner- Ville y que resultan de interés en un análisis cuantitativo.

El problema que ésta presenta de los términos interferentes puede disminuirse si se trabaja sobre la señal analítica lo que permite además una interpretación física importante de sus momentos de primer orden. En consecuencia, se calcularon vectores de características formados por mediciones realizadas sobre la Distribución de Wigner-Ville de los distintos fonemas: los momentos (8), (9), (10) previamente normalizados, los marginales (11), (12) y la entropía de Rényi (3) para $\alpha = 3$. Se calcularon las distancias Itakura-Saito adaptada de los vectores correspondientes a la señal limpia y a la señal ruidosa.

	0dB	2dB	4dB	10dB
/l/	133	43.7	14.9	1.144
/a/	9.674	4.58	1.60	0.241
ue	9.87	4.64	1.69	0.321
/n/	329	128	50.42	3.469
ie	91.6	37.80	13.70	1.381

Tabla 2: Distancia de Jensen –Rényi (J-S)

	0dB	2dB	4dB	10dB
/l/	133.5	33.30	12.06	1.747
/a/	9.309	6.124	0.946	0.483
ue	15.32	7.951	2.153	0.502
/n/	174	99.01	53.22	8.469
ie	87.01	47.9	34.33	0.852

Tabla 3. Distancia Itakura- Saito (I-S)

De los resultados, resumidos en la Tabla 3 se observa una relación similar a lo acontecido en la Tabla 2: una correspondencia entre mejor SNR y menor distancia entre vectores de características.

A continuación se analizó la SNR y ambas mediciones de distancia sobre las distintas bandas críticas de cada fonema obtenidas a partir de un filtro Gammatone de 18 canales.

	/l/	/a/	ue	/n/	ie
B1	-31.7	-14.1	-23.6	-40.5	-21.3
B2	-29.3	0.094	-8.15	-34.0	-15.6
B3	-24.6	7.01	6.95	-31.8	-15.4
B4	-9.65	12.96	-12.3	-21.9	-14.1
B5	-15.3	17.3	-3.55	-22.4	-23.3
B6	-22.2	19.3	8.50	-15.4	-20.6
B7	-24.6	21.2	8.88	-11.7	-18.9
B8	-16.1	20.9	2.27	2.24	-16.4
B9	-14.1	14.6	-0.32	-0.87	-10.2

	/l/	/a/	ue	/n/	ie
B10	-14.7	13.17	3.56	-5.31	-2.69
B11	-13.7	-3.87	17.07	3.01	2.20
B12	-0.94	7.52	23.01	9.98	14.51
B13	8.99	-0.01	28.99	3.65	10.26
B14	12.81	6.56	12.30	4.26	5.12
B15	-4.30	19.77	16.51	-6.54	8.02
B16	3.75	3.014	17.85	8.62	5.16
B17	2.50	-15.9	10.36	2.31	-13.1
B18	-17.9	-18.4	-14.9	-17.1	-21.8

Tabla 4: SNR en las distintas bandas de frecuencia de los fonemas considerados con SNR global de 2 dB

Se observa que la SNR (Tabla 4) presenta importantes oscilaciones dependiendo de la banda de frecuencia considerada. Así, para frecuencias altas y medias los mejores valores de SNR se dan para los tramos representativos de /a/, no obstante lo cual, esta situación no se mantiene para las bandas de frecuencias bajas (B11 en adelante).

A los fines de justificar la variación de SNR en las distintas bandas de frecuencia, se determinaron en cada uno de los tramos de la señal limpia, las frecuencias donde el espectrograma toma valores máximos, interpretando que corresponden a concentración de energía. En la Tabla 5 se indican las frecuencias centrales en cada banda de frecuencia del banco de filtros Gammatone y en la Tabla 6, los picos del espectrograma de cada fonema.

B1	**B2**	**B3**	**B4**	**B5**	**B6**
3437	2968	2531	2156	1843	1562
B7	**B8**	**B9**	**B10**	**B11**	**B12**
1343	1125	937	781	656	531
B13	**B14**	**B15**	**B16**	**B17**	**B18**
437	343	281	218	156	93

Tabla 5: frecuencias centrales en Hz

	F1-B	**F2-B**	**F3-B**	**F4-B**
/l/	437(B13)	218(B16)	625(B11)	2093(B4)
/a/	1093 (B8)	1875(B5)	1375(B7)	1687(B6)
ue	468(B13)	656(B11)	218(B16)	1343(B7)
/n/	593(B12)	250(B16)	1187(B8)	787(B10)
ie	562(B12)	281(B6)	781(B10)	2781(B2)

Tabla 6: Picos del espectrograma en Hz

Se pasó a analizar las distancias tanto la Jensen-Rényi como la Itakura–Saito sobre cada banda crítica. En la Figura 2 se representa para SNR global de 2 dB para los fonemas /a/ y /l/

Se observa un comportamiento regular en el caso del fonema /a/, donde hay correspondencia entre las mejores SNR en las frecuencias centrales (por encima de los 500 Hz.) y menores valores de distancia. Sin embargo, para el fonema /l/ los valores de distancia presentan bruscas oscilaciones que no dependen de una variación similar de la SNR.

6. CONCLUSIONES

La primera observación importante es la ratificación que el impacto del ruido no es uniforme sino que depende del fonema que se considera Así, los de mayor energía (las vocales y los diptongos) sufren menor efecto contaminante que las consonantes que son de menor energía, tal como puede apreciarse en la Tabla 1.

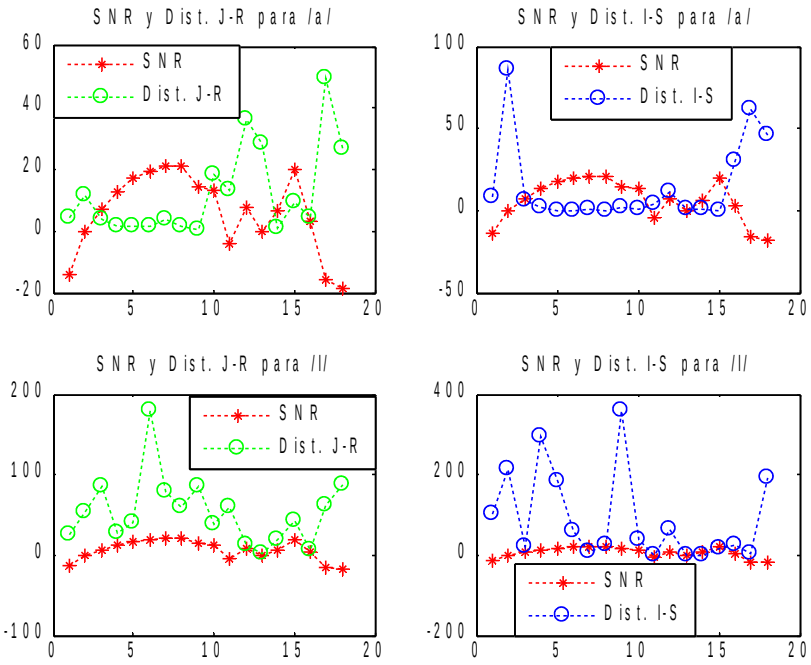

Figura 2: SNR y distancias Jensen-Rényi e Itakura- Saito para las distintas bandas de frecuencia de /a/ y /l/

Analizados los distintos fonemas con técnicas tiempo-frecuencia comparando tanto gráficamente (Figura 1) como en forma cuantitativa con distintas medidas de distancia (Tabla 2 y Tabla 3) efectivamente se observa que hay correspondencia entre mejor SNR y menor distancia.

Es de destacar la coherencia de los resultados obtenidos usando tanto la distancia Jensen-Rényi generalizada y como la distancia Itakura-Saito adaptada para vectores formados con características básicas de la representación tiempo-frecuencia.

Cuando se analizan las distintas bandas de frecuencia de cada fonema, aparece nuevamente un impacto diferenciado de la SNR (Tabla 4) que puede justificarse a partir de la distinta concentración de energía en cada fonema de acuerdo a la banda de frecuencia considerada (Tabla 5 y Tabla 6). Sin embargo se advierte que no siempre es posible establecer un ordenamiento que haga corresponder mayor valor de SNR con menor distancia entre sus representaciones tiempo-frecuencia. Esta situación se agudiza en las consonantes (Figura 2).

Estos hechos pueden interpretarse como que la relación señal/ruido si bien da indicaciones globales respecto a las características de la señal limpia comparada con la señal ruidosa, no siempre refleja con precisión los efectos del ruido sobre las bandas críticas.

El análisis realizado debe complementarse con un trabajo futuro que permita establecer cual es la vinculación entre la distorsión de las características tiempo-frecuencia y el grado de deterioro de la inteligibilidad de la señal de voz cuando ésta se contamina con ruido, medida a través de la opinión de grupos de oyentes u otras medidas objetivas adecuadas.

REFERENCIAS

S. Aviyente y W. J. Williams. Entropy based detection on the time- frequency plane. *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Proc. ICASSP´2003,* IV: 441- 444, 2003.

S.Aviyente, L.Brakel, R.Kushwada, M.Snodgrass, H.Shevin y W.Williams. Characterization of Event Related Potentials Using Information Theoretic Distance Measures. *IEEE Trans. On Medical Engineering*, 51:737-743, 2004.

R.G.Baraniuk y P. Flandrin. Measuring time-frequency information content using the Rényi entropies. *IEEE Transactions on Information Theory,* 47:1391-1409, 2001.

L. Cohen . Time-Frequency Distributions: A review. *Proceeding IEEE,* 77: 941-981, 1989.

K.L. Davidson y P.J. Loughlin. Instantaneous spectral moments. *Journal of the Franklin Institute,* 337: 421-436, 2000.

Y.Ephraim, H. Lev-Ari y W.Roberts. A brief survey of speech enhancement. *The Electronic Handbook*, 1:1-17, CRC Press, 2003.

P. Flandrin. Some features of time-frequency representations of multicomponent signals. *IEEE Int. Conf. Acoust., Speech, Signal Processing,* 1:41- 44, 1984.

C.Giguére y P. Woodland. A computational model of the auditory periphery for speech and hearing research. *J. Acoust. Soc. Am*., 95: 331-349, 1994.

B.Glasberg y B.C. Moore. Derivation of auditory filter shape from notched-noise data. *Hear Res.,* 47:103-138, 1990.

J.Hansen y B.Pellom. An effective Quality Evaluation Protocol for Speech Enhancement Algorithms. *ICSLP-98: Inter. Conf. On Spoken Language Processing - Sydney, Australia,* 7: 2819-2822, 1998.

S.Ho y R.Marks II. Some Properties of the Generalized Time Frequency Representation with Cone-Shaped Kernel. *IEEE Transactions on Signal Processing,* 40:1735-1745, 1992.

R.Huber y B.Kollmeier. PEMO-Q . A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Trans. on Audio, Speech and Language Processing,* 14:1902- 1911, 2006.

T. Irino y R.D.Patterson. A time-domain, level dependent auditory filter: The Gammachirp. *J. Acoust. Soc. Am., 1*01: 412-419, 1997.

F.Itakura y S. Saito. Speech analysis-synthesis system based on the partial autocorrelation coefficient. Proceedings of the Acoustical Society of Japan Meeting, 1969.

P.Johannesma. The pre- response stimulus ensemble of neurons in the cochlear nucleus. *Symposium on Hearing Theory , IPO, Eindhoven Holland ,* 1972.

B. Kollmeier. Auditory Models for Audio Processing- Beyond the Current Perceived Quality?. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics – New Paltz (N.Y.),* 178-182, 2005.

R.F. Lyon. A computational model of filtering, detection and compression in the cochlea. *Proc. of IEEE ICASSP,* 1:1148 -1151, 1983.

C.R.Rao y T.N.Nayak. Cross-entropy dissimilarity measures and characterizations of quadratic entropy. *IEEE Trans. Inform. Theory,* 13: 589-593, 1985.

A.W.Rix, J.G.Beerends, D.S.Kim, P.Kroon y O.Ghitza. Objective Assessment of Speech and Audio Quality – Technology and Applications. *IEEE Trans. On Audio, Speech and Language Processing,* 14:1890-1901, 2006.

S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics,* 16: 55-76, 1988.

M. Slanley. An efficient Implementation of the Patterson-Holdsworth auditory filter bank. *Apple Computer Technical Report,* 35 : 1-41, 1993.

B. Tacer y P.J. Loughlin. Non-Stationary Signal Classification Using the Joint Moments of Time .Frequency Distribution. *Pattern Recognition,* 31:1635-1641, 1998.

W. Williams,M. Brown y A. Hero. Uncertainty, information and time-frequency distributions. *Proc. SPIE Int. Soc.Opt.Eng.,* 1566: 144-156, 1991.

Y.Zao, L.Atlas y R. Marks II. The Use of Cone-Shaped Kernels for Generalized Time-Frequency Representations of Nonstationary Signals. *IEEE Trans. on Acoustics, Speech and Signal Processing,* 38: 1084-1091, 1990.