DETERMINACIÓN DE LA RELACIÓN ARMÓNICO-RUIDO (HNR) EN LA CLASIFICACIÓN DE VOCES DISFÓNICAS

RESUMEN
En el presente trabajo se establece una metodología para la estimación de una medida que establezca la relación existente entre el componente de ruido de las señales de voz versus el componente armónico, de tal forma que entregue la mayor discriminancia en problemas de clasificación de voces disfónicas. Mediante la aplicación de la anterior metodología se obtienen tasas de clasificación superiores al 80% usando las vocales del idioma español a modo de conjunto de datos.

PALABRAS CLAVES: Voces disfónicas, Razón Armónico Ruido.

ABSTRACT
In this work we establish a method for the estimation the ratio betweeen the noise energy of speech signals and the energy of harmonic component, namely Harmonics to Noise Ratio, so that gives better performance in disphonic voice clasification problems. By means of application of last method we obtained classification rates between 70% and 80% respect to bayesian classifier, depending on database used.

KEYWORDS: *Dysphonic voices, Harmonics to Noise Ratio.*

FRANKLIN A. SEPÚLVEDA
Docente Catedrático
Universidad Nacional de Colombia sede Manizales
fasepulvedas@unal.edu.co

GERMÁN CASTELLANOS
Ph.D en telecomunicaciones.
Docente
Universidad Nacional de Colombia sede Manizales
gcastell@telesat.com.co

1. INTRODUCCIÓN

Los parámetros basados en la relación entre la energía armónica y la energía de ruido tienen amplia aplicación en la clínica de la voz por su estrecha relación con muchas disfonías [9, 7, 4]. Entre ellos sobresale el HNR (*Harmonics-to-Noise-Ratio*), el cual es un parámetro importante a la hora de diagnosticar patologías de la laringe [4]; debido a que cuando existe alguna enfermedad en las cuerdas vocales y la laringe, la cantidad medida de HNR en señales de voz humana está relacionado con grado de enfermedad [4]. Existen varias formas de calcularlo, dependiendo de la forma en que se halla el componente armónico y el componente de ruido de la señal, sin llegar todos ellos a un mismo valor estimado.

Para calcularlo hay métodos en el dominio de la frecuencia, en el dominio del tiempo [2] y en los dominios conjuntos de tiempo-frecuencia [4]. En [12] se obtiene la parte no periódica de la señal basado en un procedimiento iterativo usando las técnicas de *cepstrum* y LPC (*Linear Predictive Coding*) combinadas, lo que conlleva a incrementos en el coste computacional. En [5] se usa un procedimiento de *denoising* basado en umbralización de los coeficientes *wavelet* en la separación del ruido de la señal de voz, para luego obtener la relación señal a ruido de la voz; el procedimiento usado allí es especialmente útil en aquellos casos en los que el ruido es de tipo blanco gaussiano. Otra opción consiste en detectar periodicidades en series de tiempo a partir de la CWT, y lo que se hace es construir un esqueleto sobre el plano tiempo-frecuencia manteniendo solo aquellos componentes *wavelet* que son máximos locales respecto a las escalas para cada instante de tiempo, pero descartando del análisis aquellos que pertenecen a singularidades de la señal, tal como se hace en [11]. Tal método es costoso en términos de cómputo, además de encontrarse problemas al tratar de establecer trayectorias de máximos a través de las escalas.

En el cálculo del HNR (razón armónico ruido, *harmonics-tonoise ratio*) o su recíproco NHR (*noise-to-harmonics ratio*) en [7] se considera que la energía de ruido se encuentra en la región de los 2800 a los 5800 Hz, dado que la mayor parte del contenido de ruido de la voz pertenecen a altas frecuencias [7]. En [6] se toma la energía del ruido como aquel existente en la banda de los 1500 a los 4500 Hz; pero en los anteriores casos no se tiene en cuenta si los umbrales escogidos son o no convenientes en la determinación de voces patológicas.

El presente trabajo tiene como objetivo mostrar que existe la posibilidad de mejorar la estimación del HNR, determinando un umbral adecuado de frecuencias para su el cálculo, de tal forma que sea eficiente en propósitos de clasificación de voces disfónicas, para lo cual se usa un algoritmo de búsqueda respecto al parámetro frecuencia de corte que controla el valor del HNR tal que maximice el valor de la función de costo denotada por la tasa de clasificación.

2. MATERIALES Y MÉTODOS

2.1 Cálculo del HNR

El objetivo consiste en encontrar el valor de la frecuencia de corte *fc* tal que optimice la función de

Fecha de Recepción: 31 Enero de 2006
Fecha de Aceptación: 18 Mayo de 2006

costo, la cual corresponde al conjunto extractor de características y clasificador; para ello se utiliza un algoritmo de búsqueda, el cual se encarga de encontrar el valor óptimo respecto a la función de costo. El esquema general del procedimiento realizado se representa en la figura (1).

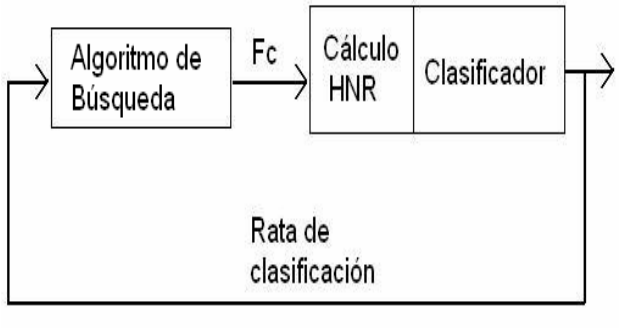

Fig 1. Esquema del procedimiento

En el cálculo del HNR se usa la siguiente definición

$$HNR = \log_2 \frac{E_h}{E_r} \quad (1)$$

donde *Eh* corresponde a la energía armónica de la señal de voz comprendida en la banda de frecuencias 0 a *fc*, y la energía de ruido *Er* es la energía existente en la banda de frecuencia normalizada *fc* hasta 1. La anterior definición se puede implementar directamente ya que para el caso en particular la frecuencia de muestreo con la que se trabaja es de 22050 Hz y además las señales fueron tomadas bajo condiciones controladas de ruido ambiental.

2.2 El clasificador Bayesiano

En el presente trabajo se tienen dos clases, voces normales ω_1 y voces disfónicas ω_2. A cada una de ellas se el asigna una probabilidad de ocurrencia, lo que se traduce en que las funciones que determinan la pertenencia a una u otra clase son las probabilidades a posteriori, $p(\omega_1/x)$ y $p(\omega_2/x)$. Es contraproducente asumir una función de densidad de probabilidad para $p(\omega_1/x)$ y $p(\omega_2/x)$, por lo que su determinación se realiza a través de su estimación basada en kernels. Además para la evaluación del clasificador bayesiano se usa la técnica de validación cruzada para grupos de 1 elemento *(Leave-one-out).*

2.2.1 Estimación de la función de densidad de probabilidad basada en kernels.

En el campo de la estimación de funciones de densidad de probabilidad (FDP) existen tres alternativas diferentes a considerar: los métodos paramétricos, los no-paramétricos y los semiparamétricos. En el primer caso se asume que los datos son generados por un modelo específico de función de densidad de probabilidad. Entonces los valores de los parámetros de dicho modelo son ajustados a los datos. Desafortunadamente en muchos casos una escogencia a priori de la FDP para los modelos no es conveniente dado que dicha función podría ser una representación errónea de la verdadera FDP [1]. Una buena alternativa es usar métodos no paramétricos, dentro de los que sobresalen los métodos basados en ventanas de Parzen (basados en Kernels) [8], [3]. El estimador *kernel* está dado por

$$\hat{p}_{\ker}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad (2)$$

donde a la función *K* se le denomina *Kernel* y además para asegurar una buena estimación en la ecuación 2 se debe cumplir la propiedad

$$\int K(t)dt = 1 \quad (3)$$

Se asume que se tiene una muestra de tamaño *n*, en donde cada observación es un vector de dimensión *d*, X_i, $i=,....,n$. El caso simple del estimador *kernel* multivariable es el producto *kernel*, dado por [8]

$$\hat{p}_{\ker}(x) = \frac{1}{nh_1 \cdots h_n}\sum_{i=1}^{n}\left(\prod_{j=1}^{d} K\left(\frac{x - X_i}{h}\right)\right) \quad (4)$$

donde X_{ij} es la *j−ésima* componente de la *i−ésima* observación. Si se usa un kernel *normal*, entonces se puede usar la regla de referencia normal para el caso multivariable,

$$h_{j_{Ker}} = \left(\frac{4}{n(d+2)}\right)^{\frac{1}{d+4}} \sigma_j ; \quad j = 1,\ldots,d \quad (5)$$

en donde σ_j corresponde a la varianza de los datos, y un estimador conveniente para σ_j puede ser usado.

2.2.2 Evaluación del Clasificador

Una vez se tiene el clasificador, es necesario evaluar su utilidad midiendo el porcentaje de observaciones que fueron clasificadas correctamente. Esto genera una estimación de la probabilidad de casos correctamente clasificados. Se tienen dos métodos para estimar la probabilidad de éstos: prueba de muestras independientes (ITS) y validación cruzada (CV).

En la *prueba de muestras independientes*, si el conjunto de muestras es grande, se puede dividir en un conjunto de entrenamiento y un conjunto de validación. Se usa el conjunto de entrenamiento para construir el clasificador y se clasifican las observaciones del conjunto de validación usando la regla de clasificación. La proporción de observaciones correctamente clasificadas es el porcentaje de clasificación estimado. Como el clasificador no ha usado el total de los patrones en el conjunto de entrenamiento, el porcentaje de clasificación estimado no está sesgado.

En la *validación cruzada* con *k*-particiones, el concepto básico consiste en dividir el conjunto de muestras en *k* particiones de tamaño $k_p = N_s/k$, con N_s siendo el número

total de patrones. Una partición es reservada como conjunto de validación mientras los restantes *k-1* son usadas como conjunto de entrenamiento.

La ventaja de este método es su poca sensibilidad a la partición de los datos. Cada dato logra estar en el conjunto de validación una vez, y *k-1* veces en el conjunto de entrenamiento. La desventaja de este método es que el algoritmo de entrenamiento debe ser evaluado *k* veces, elevando el tiempo de cálculo.

El método usado en el presente trabajo corresponde al LOO (*Leave-one-out*) el cual es una validación cruzada de *k* particiones tomada en el límite donde $k = N_s$. Esto significa que *ks* veces, de forma independiente, el modelo es entrenado sobre todos los datos exceptuando 1 punto. Luego, la validación es hecha para tal punto. Nuevamente el error promedio es calculado y usado para evaluar el modelo.

2.3 El Algoritmo de Búsqueda

Lo que se desea es encontrar aquella frecuencia de corte óptima tal que se obtenga la mayor tasa de clasificación de voces disfónicas respecto a un único parámetro HNR; por lo cual el algoritmo corresponde a un algoritmo de búsqueda unidimensional. No se pueden usar derivadas, ya que el cálculo de la derivada para cada iteración implica evaluar como mínimo dos puntos, lo que aumenta notoriamente el tiempo necesario para el algoritmo converja, además no se tienen pruebas acerca de que la función de costo sea diferenciable; entonces se usará algún algoritmo que no requiera del cálculo de derivadas. Entonces se requiere usar un método a partir del cual se pueda obtener un intervalo de incertidumbre que no dependa de las particularidades de la función con la que se está trabajando, por ejemplo, podría proponerse trabajar con el método de la sección áurea [10]; pero para usar este método se requiere que la función sea unimodal, o al menos que se tengan buenas posibilidades de que la función sea unimodal. Para ello se toman algunos datos de prueba para verificar, su gráfica se observa en la figura 2. De ella se puede observar que la función no cumple con la propiedad de ser unimodal. Por ello, se debe usar el método de búsqueda uniforme. Para limitar la cantidad de cálculos el método se desarrolla en tres iteraciones donde para iteración se escoge un intervalo diferente de incertidumbre, cuidando de que dentro de este intervalo la función que pretendemos optimizar (la función de validación cruzada) corresponda a una función cuasi-convexa, lo cual se verifica visualmente.

2.4 Base de datos

Es de interés reciente el análisis de señales de voz en la determinación patologías, por ello en el presente trabajo se incluye una base de datos de prueba de algoritmos de la cual hacen parte tanto voces normales como patológicas, todas ellas, voces de personas adultas.

- *Sujetos*. La muestra representativa de la población seleccionada, fue evaluada de forma subjetiva por el especialista en fonoaudiología, y una vez realizado el diagnóstico inicial se llevaron a cabo las respectivas sesiones de grabación con aquellas personas que amablemente colaboraron. Las voces recolectadas para ésta base de datos hacen parte de las clases: normal o disfónicas. La muestra de análisis para la presente base de datos consta de un total de 91 registros de voz, clasificados de la forma ilustrada en la tabla I. De los 91 registros, 42 corresponden a voces del sexo masculino y 49 al sexo femenino. Del total de 91, 40 corresponden al tipo normal y 51 de ellas son voces disfónicas.

- *Recolección de señales de voz*. Las señales fueron tomadas bajo condiciones de bajo nivel de ruido ambiental usando un micrófono Shure SM58, el cual es dinámico unidireccional (cardiode) diseñado para vocalistas profesionales. Posee un filtro esférico incorporado que reduce los ruidos causados por el viento y el aliento. La dispersión polar de cardiode que posee, aisla la fuente sonora principal a la vez que reduce los ruidos de fondo. La distancia del micrófono al hablante está entre 10 y 15 cm. Las propiedades de los archivos de audio generados para cada una de las señales son las siguientes: Formato *.wav, frecuencia de muestreo 22000Hz, 16 bits por muestra, monofónica y la evaluación de las voces fue hecha por el especialista en la materia, el fonoaudiólogo.

3. DISCUSIÓN Y RESULTADOS

Se presentan los resultados obtenidos para la vocal /i/ en las figuras (2 y 3), en donde los intervalos de incertidumbre corresponden a los siguientes: 0.1 a 0.9, 0.15 a 0.35, para la primera y segunda iteración respectivamente. El algoritmo finaliza en la tercera iteración habiendo tomado pasos de 0.0001 lo que se traduce en una resolución de 1.1025 Hz para la frecuencia de muestreo a la que se está trabajando, 22050 Hz. El valor óptimo corresponde a 0.268 en frecuencia normalizada. Además de puede observar que se obtienen ratas de clasificación superiores al 80 %, lo cual corresponde a un buen resultado, dado que se está usando un *set* de características de dimensión 1 y corresponde a la característica del HNR.

Resultados adicionales para las otras vocales se muestran en las figuras 4, 5 y 6. Además se realizó una prueba adicional en la cual se usaron a manera de base de datos los registros de todas las vocales como si fuesen un único conjunto de datos cuyo resultado se observa en la gráfica 7.

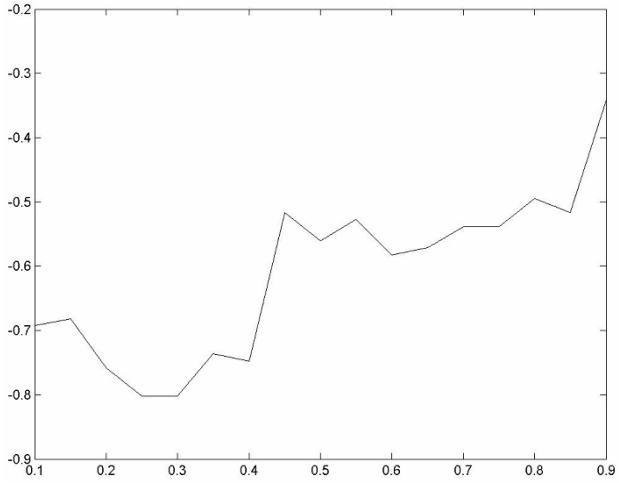

Fig 2 Tasa de clasificación invertida para la vocal /i/ rango de 0.1 a 0.9

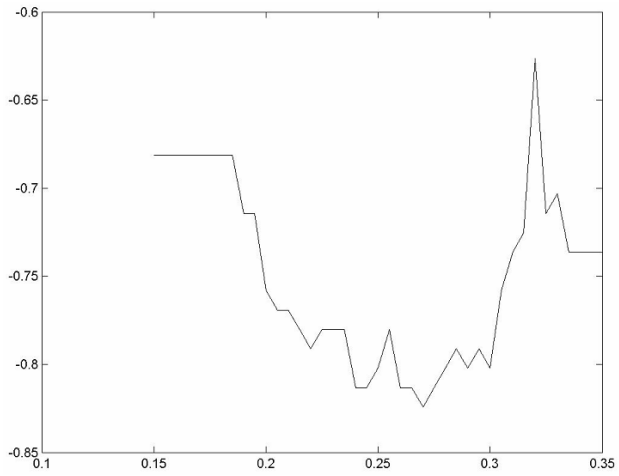

Fig. 3 Tasa de clasificación invertida para la vocal /i/ rango de 0.15 a 0.35

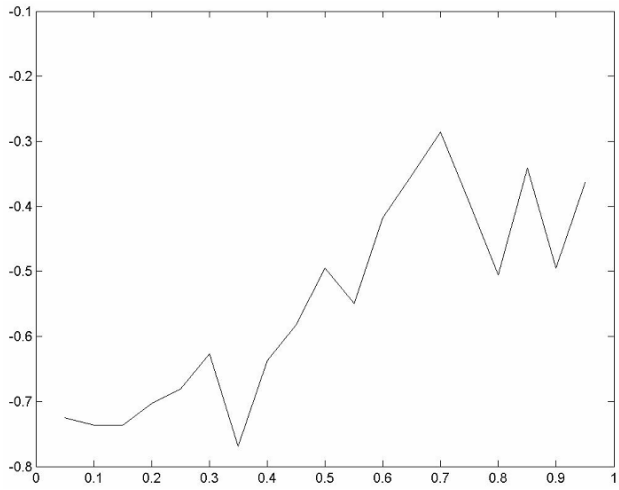

Fig. 4 Tasa de clasificación invertida para la vocal /a/

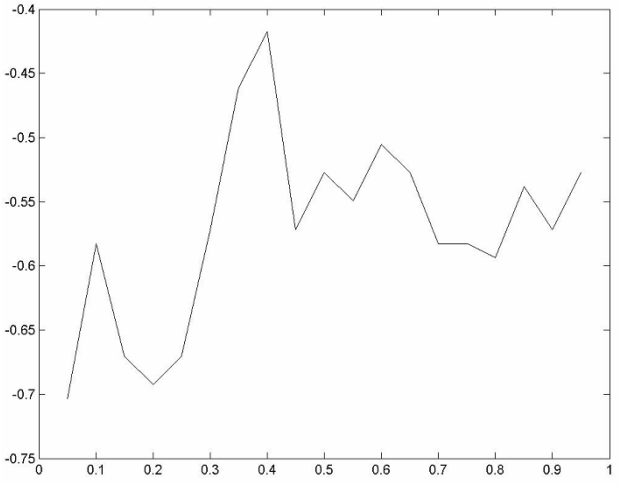

Fig. 5 Tasa de clasificación invertida para la vocal /e/

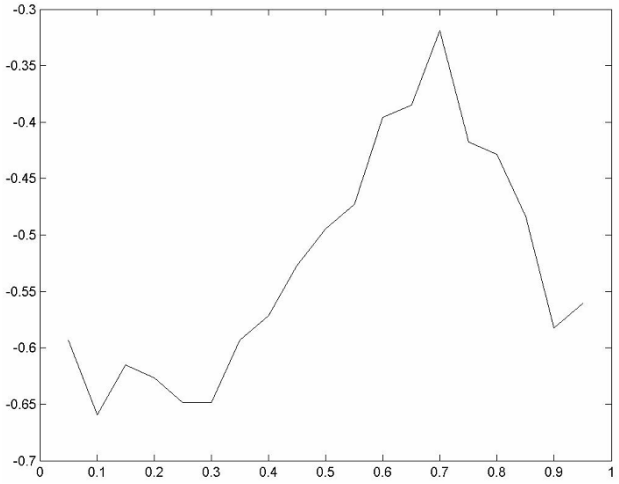

Fig. 6 Tasa de clasificación invertida para la vocal /u/

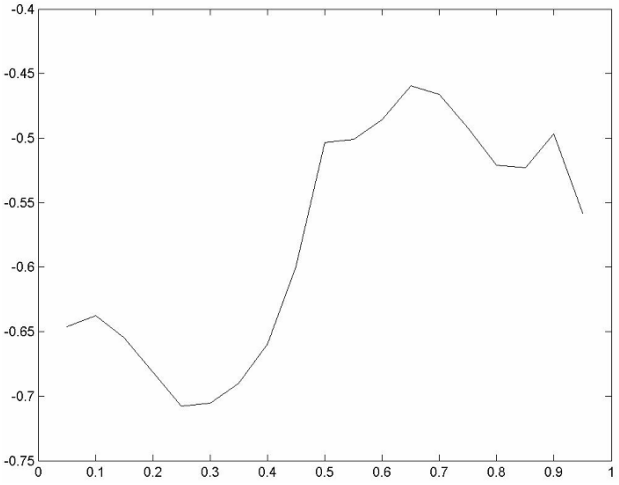

Fig. 7 Tasa de clasificación invertida usando todas las vocales

4. CONCLUSIONES

Mediante el presente trabajo de muestra que se pueden obtener características acústicas que describan el componente de ruido de una señal de voz, de buen poder discriminante en la determinación de voces difónicas, sin necesidad de recurrir a técnicas sofisticadas y de alto coste computacional; además se demuestra que se ahorra complejidad en la implementación si se realiza una selección adecuada del umbral de frecuencia. En las gráficas mostradas en la sección anterior se pueden apreciar tasas de clasificación superiores al 80%, lo cual es bueno si se tiene en cuenta que se una sola característica.

5. BIBLIOGRAFÍA

[1] ARCHAMBEAU, C. and M. Verleysen, "Fully nomparametric probability density function estimation with finite gaussian mixture models," in *5th International Conference on Advances on Pattern Recognition*, 2003.

[2] BOERSMA, P., "Accurate short - term analysis of the fundamental frequency and the harmonics to noise ratio of sample sound." pp. 100 –104, 1993.

[3] DUDA, P. H. R. and D. Stork, *Pattern Classification*. John Waley and Sons.

[4] FROHLICH, H. and D. Michaelis, "Acoustic breathiness measures in the description of pathologic voices," Universitat Gottingen, Tech. Rep., 1999.

[5] GALLOWAY, Kirsta *Estimation by means of wavelet analysis of the signal-to-noise ratio of dysphonic voices*, 1997.

[6] GONZÁLES, M. J. y T. Cervera, "Análise acústica da voz captada na faringe próximo á fonte glótica através da microfone acoplado ao fibrolaringoscópido," *Revista Brasileira de Otorrinolaringología*, vol. 67, 2001.

[7] GONZÁLES, M. J y T. Cervera, "Análisis acústico de la voz: Fiabilidad de un conjunto de parámetros multidimensionales," *Acta Otorrinolaringológica Española*, No. 53, pp. 256–268, 2002.

[8] MARTINEZ W. L. and A. R. Martinez, *Computational Statistics Handbook with Matlab*, A. K. Peters, Ed. Chapman and Hall/CRC, 2002.

[9] MICHAELIS, F. D. and H. W. Strube, "Selection and combination of acooustic features for the description of pathologic voices," Drittes Phisykalisches Institut, Tech. Rep., 1998.

[10] PIERRE, D. A., *Optimization Theory with Applications*. Dover Publications, 1986.

[11] POLYGIANNAKIS J. and X. Moussas, "On signal-noise decomposition of time-series using the continuous wavelet transform: application to sunspot index," *Monthly Notices of the Royal Astronomical Society*, vol. 343, pp. 725–734, 2003.

[12] YEGNANARAYANA, V. and C. d'Alessandro, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. on speech and auudio proc.*, vol. 6, no. 1, 1998.

[13] ZHAO Shou-Guo, e. a., "Study on vocal folds vibration characteristics based on hnr in transmitted sound signals," Institute of Microelectronics, Tsinghua University, Beijing, Tech. Rep., 2000.