

Diseño de un sistema de medición de parámetros característicos y de calidad de señales de voz

Paulino Del Pino, Iván Granadillo, Mario Miranda, Carlos Jiménez, José A Díaz
Universidad de Carabobo, Facultad de Ingeniería, Escuela de Ingeniería Eléctrica,
Departamento de Electrónica y Comunicaciones.
Email: pdelpi@uc.edu.ve

Resumen

Este trabajo presenta el desarrollo de una aplicación para la extracción de parámetros característicos y de calidad de señales de voz con el objetivo de calificarlas como voces sanas o patológicas. El sistema desarrollado permite calcular cinco parámetros característicos: Pitch, Formantes, Energía, Jitter, y Shimmer; y tres parámetros de calidad en el dominio de la frecuencia: relación valor pico - valor medio en frecuencia (PMR), relación señal a ruido a frecuencias medias (SNRM), y relación señal a ruido a frecuencias altas (SNRH). La herramienta desarrollada, confirmó el comportamiento de calidad esperado en lo que respecta a los parámetros extraídos en el dominio de la frecuencia, adicionalmente se encontraron diferencias significativas en algunos de los parámetros característicos en el dominio del tiempo, de las voces sanas y patológicas, que sugieren su utilización para calificar señales voz como sanas o patológicas.

Palabras clave: Señal de voz, parámetros característicos, parámetros de calidad, voz sana, voz patológica.

System design for extraction of characteristics and quality parameter signal voices

Abstract

This work presents the development of an application for the extraction of characteristic and quality parameters of voice signals with the objective of qualifying them as healthy or pathological voices. The developed system allows to calculate five characteristic parameters: Pitch, Formant, Energy, Jitter, and Shimmer; and three quality parameters in the domain of the frequency: relationship value pick - half value in frequency (PMR), signal to noise relations at mediums frequencies (SNRM), and signal to noise relation to high frequencies (SNRH). The developed tool, it confirmed the prospective behavior of quality in the parameters extracted in the domain of the frequency, additionally they were significant differences in some of the characteristic parameters in the domain of the time, of the healthy and pathological voices that suggest their use to qualify signals voice as health or pathological.

1. INTRODUCCIÓN

La señal de voz, dada su importancia en el proceso de comunicación, ha sido objeto de muchos estudios a lo largo de la historia. En el siglo XX, varios investigadores desarrollaron diversos modelos que explican como se produce el fenómeno del habla, y en los últimos años, con el desarrollo de la tecnología en computación e informática [1], se ha hecho más fácil la implementación de aplicaciones que se basan en

estos modelos, como reconocedores de voz, telefonía digital, etc. Otra aplicación del análisis de señales de voz se produce en el área médica [2,3], en éstas, el estudio de los parámetros de la voz se realiza a través de un análisis matemático eficiente, rápido y cuantificable, proporcionando al usuario una medición confiable del comportamiento del sistema fonatorio del paciente, para así conocer el estado del mismo. El objetivo de este trabajo es el diseño de una aplicación para la medición de parámetros característicos y de calidad

de la voz, mediante técnicas del procesamiento digital de señales de voz. Entre los parámetros que se calculan con este sistema de medición se encuentran cinco parámetros característicos (pitch, formantes, energía, shimmer y jitter) y tres de calidad en el dominio de la frecuencia: PMR, SNRM y SNRH, [4]. Para la obtención de los mismos, se programaron distintos algoritmos descritos en la literatura investigada referente al análisis de voz [5].

2. METODOLOGÍA

El sistema de medición, ver figura 1, consta de las tres partes o subsistemas siguientes:

a) **Adquisición de los datos (voz):** Este subsistema realiza la captura de la señal de voz mediante un micrófono de alta directividad conectado a la entrada de audio de línea del computador. La señal de voz se graba a una frecuencia de muestreo de 44100 Hz y se digitaliza a 16 bits, se registran cinco segundos de voz en cada captura por lo que se tienen para cada archivo 220500 muestras. Los datos son almacenados en un archivo tipo wav. Se debe señalar que el sistema solicita que el paciente emita en forma continua y sostenida el sonido correspondiente a la vocal /a/. Para la adquisición de los datos se crearon varios programas e interfaces que permiten obtener la información a través de dos vías: grabando una nueva muestra de voz, o cargando una señal de voz previamente grabada.

b) **Subsistema de medición:** En este se realiza el procesamiento digital requerido a los datos obtenidos en el subsistema anterior y se aplican los algoritmos correspondientes para la determinación de los parámetros deseados: Frecuencia Fundamental, Formantes, Contorno de energía total, Shimmer, Jitter, PMR, SNRM y SNRH.

c) **Resultados y Reportes:** este subsistema corresponde a la interfaz gráfica para el despliegue de los resultados, valores de los parámetros deseados, y reportes en forma numérica o gráfica, imprimible, que permiten el análisis de los mismos para su correspondiente utilización, por ejemplo por un médico para el diagnóstico de una patología del sistema fonador.

La Figura 2, muestra la interfaz de usuario principal en la que se muestran las opciones de operación

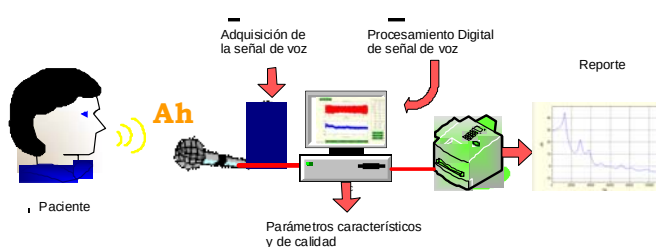


Figura 1. Diagrama de bloques del sistema diseñado.

de usuario, estas son: a) Grabar un archivo de voz, o cargar un archivo existente, b) analizar el archivo seleccionado, c) reproducirlo, d) observar los resultados gráficos de los análisis realizados, e) presenta los valores numéricos de los parámetros extraídos y f) generación de un reporte para impresión de los resultados obtenidos.

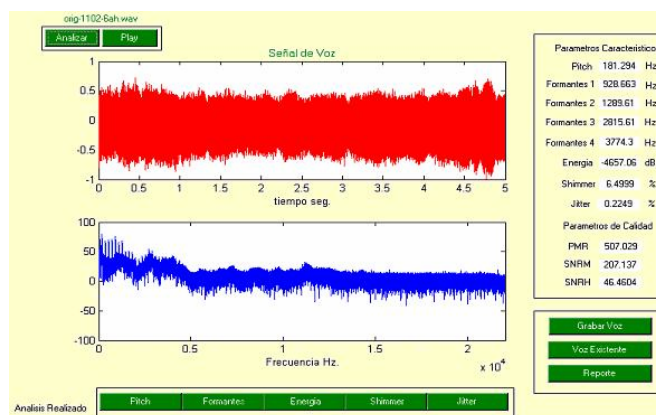


Figura 2. Interfaz de usuario principal del sistema de medición de parámetros característicos y de calidad de señales de voz.

Oprimiendo el botón de grabar voz aparece una interfaz de grabación como la mostrada en la figura 3, en ella se introduce el nombre del archivo/paciente y al oprimir el botón de grabar se inicia el proceso de grabación, en este momento se muestra en la pantalla el mensaje Grabando, luego de transcurrido un tiempo de cinco segundos se muestra el mensaje Señal Grabada, para indicar que la grabación ha finalizado. El botón Play permite reproducir la señal para que el usuario pueda verificar si esta fue grabada de manera correcta (que no haya silencios y que no haya saturación), de no ser así, se puede grabar la señal nuevamente hasta obtener el resultado deseado, y por último, el botón de Aceptar carga los datos grabados en la interfaz principal.



Figura 3. Interfaz de grabación de archivo de voz.

La adquisición de una voz previamente grabada se hace a través del botón Voz existente de la interfaz principal, al oprimir este botón se despliega una ventana en la que se puede seleccionar un archivo de audio en formato wav y convertirlo en un vector de datos.

3. ALGORÍTMOS DE PROCESAMIENTO DE LAS SEÑALES DE VOZ

3.1. Frecuencia Fundamental

Este parámetro es determinado mediante el algoritmo de aproximación por autocorrelación, para ello se divide la señal en ventanas de 1323 muestras (30ms a una frecuencia de muestreo de 44,1 KHz.), con un desplazamiento de 500 muestras entre ventanas (este valor de desplazamiento fue escogido experimentalmente para evitar cambios bruscos de los resultados entre dos frames consecutivos). Se filtra la señal para eliminar las altas frecuencias, y dejar la parte correspondiente a la excitación, para esto se utiliza un filtro pasa bajo con una frecuencia de corte entre 800 y 900 Hz de respuesta impulsiva finita (FIR) y de 986 coeficientes. Se recorta la señal en su parte central para acentuar su periodicidad, al solo dejar los máximos y los mínimos (clipping). Después de analizar varias voces se observó que se obtenían mejores resultados con un nivel de recorte entre 20% y 35%, por lo que se estableció el nivel de recorte en 25%. A continuación, se calcula la función de autocorrelación de la señal, se determina la posición del máximo en la señal, tarea que se realiza mediante una función que busca la posición del máximo en un intervalo de muestras dado. Este intervalo es entre 126 muestras (350Hz) y 880 muestras (50Hz) y finalmente se calcula la frecuencia fundamental como el cociente entre la frecuencia de muestreo y la posición del máximo.

La figura 3, muestra un ejemplo de resultado gráfico para la frecuencia fundamental en el tiempo,

correspondiente a la señal de voz para la vocal /a/, determinado según el algoritmo descrito.

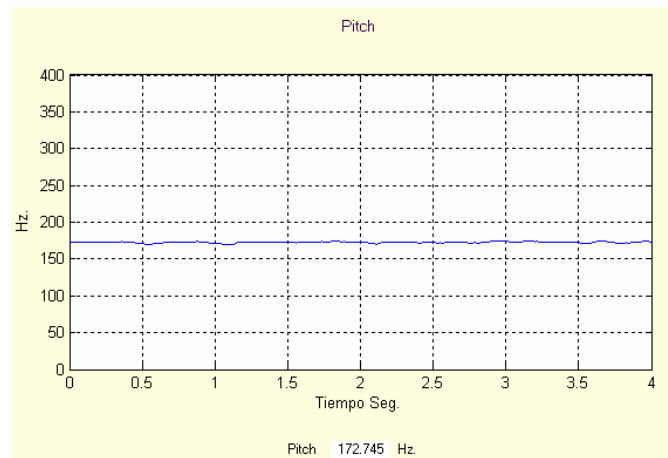


Figura 4. Frecuencia fundamental de una señal de voz reportada por el sistema desarrollado correspondiente a la vocal /a/.

3.2. Formantes

El algoritmo utilizado para la extracción de este parámetro es el de análisis de predicción lineal (filtrado inverso), de esta manera se aproxima la función de transferencia del tracto vocal en cada instante de tiempo y se realiza la búsqueda de picos en el espectro de frecuencia. El orden de predicción se estimó como $N = F_s + 4$, para sonidos vocalizados, con F_s como la frecuencia de muestreo en KHz, en este caso con $F_s = 44$ KHz, se tiene un orden de estimación del predictor de $N = 48$. El procedimiento seguido es el siguiente: se divide la señal en ventanas de 1323 muestras (30ms), se crea un vector de autocorrelación, calculando la autocorrelación de 48 muestras de la señal, se genera la matriz de autocorrelación. Mediante un programa que toma el vector de autocorrelación y genera una matriz tipo Toeplitz a partir de éste, posteriormente se calcula la matriz inversa de la matriz de autocorrelación y se multiplica por el vector de autocorrelación para obtener los coeficientes de predicción lineal, se filtra la señal de modo inverso y se restan la señal original y la señal filtrada para obtener la parte de la señal que corresponde al tracto bucal, se calcula la transformada rápida de Fourier (FFT) de la señal resultante para obtener su espectro de frecuencia, se calcula la posición y el valor del máximo pico en cada banda de frecuencia del espectro, se realiza la conversión de las posiciones de los máximos a frecuencia, se desplaza la ventana 500 muestras y se repite el proceso hasta completar el análisis en toda la señal de voz.

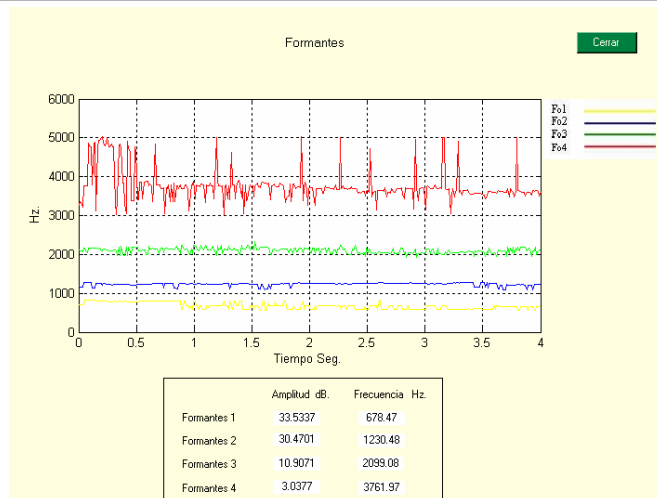


Figura 4a. Formantes correspondientes a la emisión de la vocal /a/.

En la figura 4a, puede observarse los formantes y en la figura 4b, el espectro en frecuencia para un instante de tiempo, determinados mediante el algoritmo descrito y reportados por la aplicación desarrollada para el caso de la vocal /a/.

3.3. Contorno de Energía Total

Para la obtención del contorno de energía total, se realizó un programa que calcula la energía en cada periodo, utilizando el teorema de Parseval, para hacer esto, primero se eleva cada término de la señal al cuadrado y luego se convierte a decibelios (dB), y por último se suman. Este proceso se aplica para cada periodo de la señal y se guarda el resultado en un vector. Después de calcular la energía para todos los periodos, se calcula su valor promedio y la variación relativa a éste, según la ecuación 1.

$$\Delta en_{(i)} = \frac{|\Delta en_{(i)} - en|}{en} * 100 \quad (1)$$

En la figura 5, se observa el resultado reportado por la aplicación desarrollada para este parámetro, correspondiente a la emisión de la voz de la vocal /a/.

3.4. Shimmer

El parámetro shimmer se calcula de manera relativa, es decir primero se mide la amplitud pico a pico de la señal en cada periodo $V_{pp}(i)$ y después se calcula la variación de la amplitud de cada periodo con res-

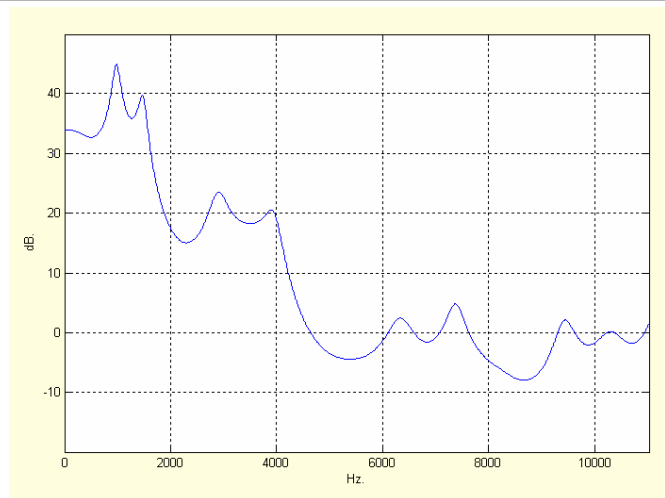


Figura 4b. Espectro en frecuencias para un instante de tiempo.

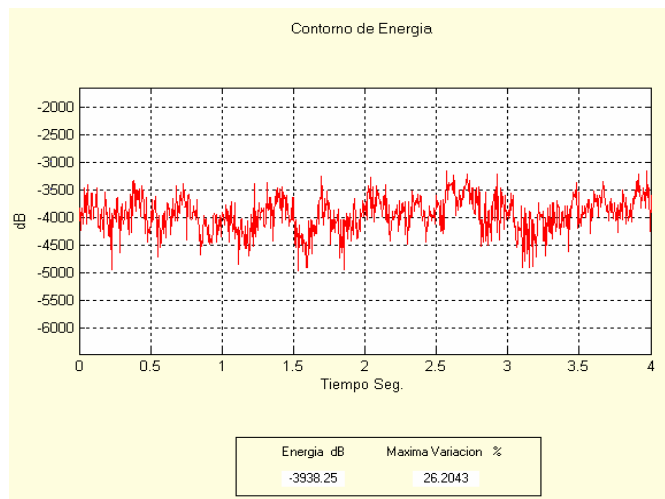


Figura 5. Contorno de energía correspondiente a la vocal /a/.

pecto al periodo anterior de manera porcentual, ecuación (2).

$$Shimmer(i) = \frac{|V_{pp}(i) - V_{pp}(i+1)|}{V_{pp}(i)} * 100 \quad (2)$$

En la figura 6, se aprecia este parámetro reportado por el sistema desarrollado para la vocal /a/.

3.5. Jitter

Para determinar el jitter se aplica un procedimiento similar al del shimmer, se calcula la variación del pitch o frecuencia fundamental, de una ventana con respecto al pitch de la ventana anterior de manera

relativa, ecuación (3).

$$Jitter(i) = \frac{|pitch(i) - pitch(i+1)|}{pitch(i)} * 100 \quad (3)$$

En la figura 7, se observa el reporte para este parámetro.

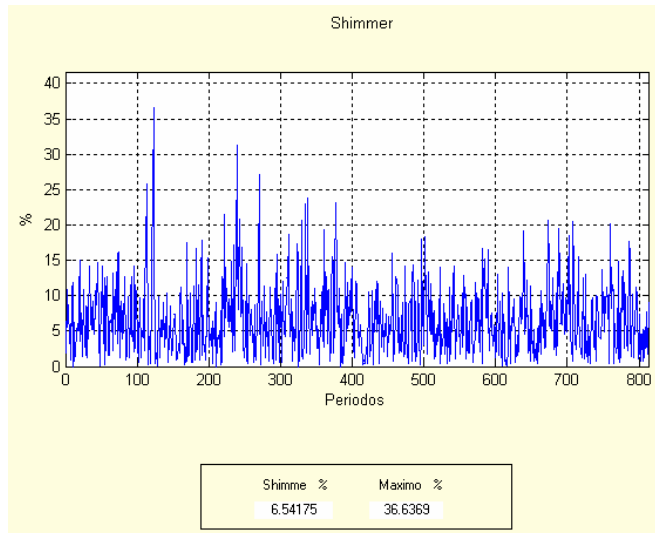


Figura 6. Shimmer correspondiente a la emisión de una vocal /a/.

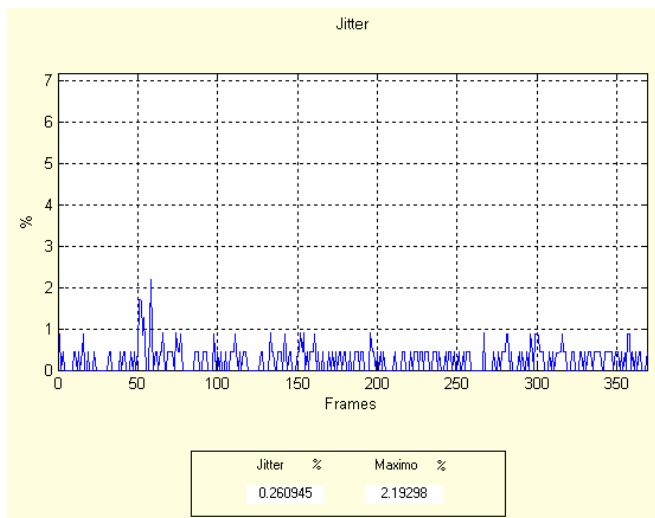


Figura 7. Jitter correspondiente a la emisión de una vocal /a/.

3.6. PMR

Es el primer parámetro de calidad en el dominio de la frecuencia y consiste en el cálculo de la relación valor pico a valor medio del espectro de amplitud de la señal de voz, ecuación (4).

$$PMR = \frac{\text{máximo}|S(\omega)|}{\text{medio}|S(\omega)|} \quad (4)$$

donde $S(\omega)$ es el espectro de amplitud de la señal de voz calculado mediante la transformada discreta de Fourier.

3.7. SNRM y SNRH

Estos también son parámetros de calidad extraídos del dominio de la frecuencia. Para el cálculo de la relación señal a ruido a frecuencias medias SNRM y altas SNRH, es necesario estimar el ruido que se encuentra presente en la señal. Para esto, se utiliza un filtro de mediana de longitud N (N=6 normalmente), y se restan la señal original de la señal filtrada. Después se calculan los espectros en frecuencia de la señal sin ruido y del ruido, se calcula la energía en cada banda de frecuencia (media, y alta) de ambos espectros, y se dividen, ecuaciones (5) y (6).

$$SNRM = \frac{\text{Energía}_{\text{señal}}(500-1500\text{Hz.})}{\text{Energía}_{\text{ruido}}(500-1500\text{Hz.})} \quad (5)$$

$$SNRH = \frac{\text{Energía}_{\text{señal}}(1500-2000\text{Hz.})}{\text{Energía}_{\text{ruido}}(1500-2500\text{Hz.})} \quad (6)$$

Los parámetros PMR, SNRM y SNRH son reportados en forma numérica por el sistema para su posterior análisis.

4. RESULTADOS

Los parámetros característicos antes descritos fueron calculados para un conjunto de señales de voz calificadas previamente como sanas o patológicas por especialistas del Department of Communication Sciences and Language Disorders of the University of Florida USA, específicamente se utilizaron doce voces calificadas como sanas y doce voces calificadas como patológicas, todas ellas señales que corresponden a la emisión en forma sostenida de la vocal /a/. Para efectos de validación se utilizó un programa desarrollado por Phillipous C, Loizou, University of Texas at Dallas, denominada Colea [6], utilizando la aplicación desarrollada y la antes mencionada se calcularon y compararon las máximas diferencias porcentuales de los parámetros comunes a ambas aplicaciones y que se resumen en la tabla 1.

Adicionalmente, los valores de los parámetros de calidad en el dominio de la frecuencia (PMR,

SNRM y SNRH), determinados con la herramienta desarrollada corresponden, con diferencias menores al 1%, a los reportados en un trabajo realizado anteriormente [4], esto permite decir que el sistema desarrollado es confiable en la extracción de los parámetros objeto de estudio.

Tabla 1. Comparación de parámetros entre la aplicación desarrollada y otra aplicación de referencia.

	Diferencia Porcentual (%)
Frecuencia Fundamental	10.71
Frecuencia 1 ^{er} Formante	2,23
Frecuencia 2 ^{do} Formante	4,61
Frecuencia 3 ^{er} Formante	6,28
Frecuencia 4 ^{to} Formante	4,42
Amplitud 1 ^{er} Formante	5,21
Amplitud 2 ^{do} Formante	2,64
Amplitud 3 ^{er} Formante	3,16
Amplitud 4 ^{to} Formante	2,16

En general los resultados obtenidos son los mostrados anteriormente en forma gráfica y numérica en la interfaz principal, sin embargo como aporte adicional al estudio de parámetros de calidad se estudio el comportamiento estadístico de estos parámetros en los grupos de voces sanas y voces patológicas mediante un análisis de varianza de unifactorial ANOVA, con un nivel de significancia de $\alpha = 0,05$ previa confirmación de que los datos seguían una distribución normal para determinar si existían diferencias significativas importantes de estos parámetros en los grupos de voces sanas y patológicas[7]. Este estudio arrojó que para los parámetros amplitudes medias del primer, segundo y cuarto formante existen diferencias significativas para los grupos de voces sanas y patológicas. A manera de ilustrar este comportamiento la tabla 2,

Tabla 2. Amplitud media del primer formante para los dos grupos de voces.

Voces	
Sanas (dB)	Patológicas (dB)
40.28	30.2
41.03	39.36
45.95	36.69
40.78	39.33
44.42	27.79
37.34	34.52
44.45	41.45
38.76	36.3
43.2	33.13
40.86	41.61
47.11	34.35
38.31	33.14

resume los valores de amplitud del primer formante para las voces sanas y patológicas.

El análisis de varianza de una dimensión para este parámetro se resume en la tabla 3. De esta tabla se observa de cómo $F > F_{crítico}$ y para un nivel de significancia del 5%, por lo que no queda duda de que los dos grupos son diferentes en lo que respecta a este parámetro.

En la figura 8, podemos ver un gráfico de cajas que ilustra la distribución de las amplitudes del primer formante en los dos grupos de voces. De este gráfico podemos observar lo fuertemente diferenciados que están, no solamente sus valores medios sino el conjunto de datos que constituyen la población del grupo.

Tabla 3. Análisis de varianza de un factor para el parámetro amplitud del 1er Formante.

Análisis de varianza						
Origen de las variaciones	Suma de cuadrados	Grados de libertad	Promedio de los cuadrados	F	Probabilidad	Valor crítico para F
Entre grupos	232,068204	1	232,068204	16,3461072	0,00054307	4,3009436
Dentro de los grupos	312,337392	22	14,1971542			
Total	544,405596	23				

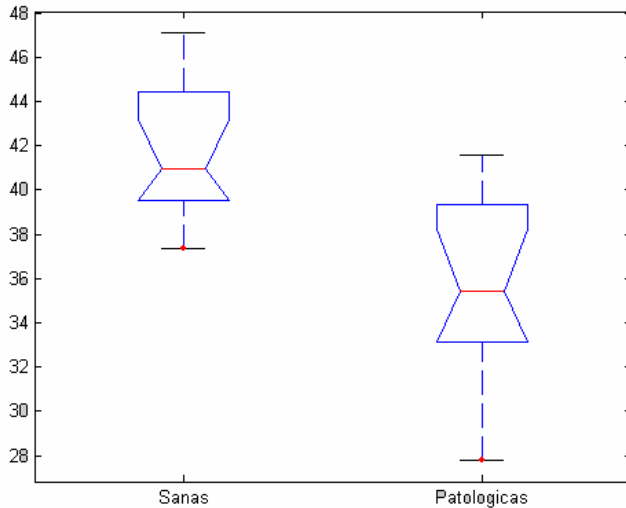


Figura 8. Gráfico de cajas correspondiente al ANOVA del parámetro amplitud del primer formante.

Siguiendo el mismo proceso de análisis de varianza de un factor para cada uno de los parámetros extraídos con la aplicación desarrollada se concluye que las amplitudes del segundo y cuarto formante se comportan de manera similar que la amplitud del primer formante, el mismo comportamiento se encuentra también en el valor medio de la energía, el Jitter y el Shimmer, es decir existen diferencias significativas entre los grupos de voces sanas y patológicas en lo que respecta a estos parámetros por lo que los mismos son discriminadores de voces sanas y voces patológicas.

El análisis realizado a estos parámetros en el dominio del tiempo, en conjunto con el análisis de los parámetros de calidad (PMR, SNRM, y SNRH) en el dominio de la frecuencia, aportan a manera de referencia una serie de valores para voces sanas, basados en los valores medios y la estadística mostrada de estos parámetros determinados para el grupo de voces de referencia y calificación médica conocida

5. CONCLUSIONES

De acuerdo a los resultados obtenidos, se puede concluir que la implementación de los algoritmos de obtención de parámetros seleccionados, hacen de la aplicación desarrollada una herramienta sencilla, amigable, eficiente y confiable al momento de analizar las señales de voz.

Del análisis realizado al grupo de 12 voces sanas y 12 voces patológicas, se puede concluir que los pará-

Tabla 4. Valores de referencia para voces sanas

Parámetro	Valor de Referencia
Amplitud F1	Mínimo 39.5 dB
Amplitud F2	Mínimo 35 dB
Amplitud F4	Mínimo 18 dB
Energía	Mínimo -5000 dB
Jitter	Máximo 1.5 %
Shimmer	Máximo 6.5 %
PMR	Mínimo 294
SNRM	Máximo 190
SNRH	Mínimo 45

metros amplitud del primer, segundo y cuarto formante, la energía media, el Jitter y el Shimer existen diferencias significativas entre los dos grupos de voces, por lo que estos representan un aporte en la obtención de parámetros para la clasificación de voces sanas y patológicas.

Aún cuando se proponen valores umbrales y sus tendencias para la calificación de voces como sanas y patológicas basadas en los valores medios de los parámetros y su estadística ANOVA, se requiere hacer un estudio con un grupo más amplio de voces calificadas a priori por especialistas médicos del área.

Es recomendable que se realice un estudio con un grupo de voces clasificado de acuerdo a las patologías para obtener mejores resultados, así como utilizar otros sistemas para comparar los resultados.

La calificación como sana o patológica de una voz analizada con la herramienta aquí desarrollada requiere que al menos un parámetro se ubique en la calificación patológica, de cualquier manera la herramienta es un elemento auxiliar al examen realizado por el especialista.

5. REFERENCIAS

- [1] Rodríguez C., M, Cortazar M., Ismael, Tapias T., Daniel, Relaño G., José. *Estado del arte en tecnologías de voz* Comunicaciones de Telefónica I+D, España, Número 20 (Marzo 2001).
- [2] Gonzáles J., Cervera T., Miralles J. L., *Análisis Acústico de la voz: Fiabilidad de un conjunto de parámetros multidimensionales*. Acta Otorrinolaringol. Esp; 53: 256-268. (2002).

- [3] Alonso, J B., Ferrer, M. A., de León, J. y Travieso, C. *Cuantificación de la calidad de la voz para su evaluación clínica por medio de análisis acústico*. Universidad de las Palmas de Gran Canaria. IV Jornadas de tecnología del habla, Zaragoza Del 8 al 10 de Noviembre de 2006.
- [4] Del Pino, P., Díaz J. A., Jiménez, C. y Rothman, H. *Identificación de algunos Parámetros Espectrales que determinan la Calidad de la voz*, Revista Ingeniería UC, Vol 11, Nº 3. Universidad de Carabobo, Venezuela. (Diciembre 2004).
- [5] Atal, Bishnu S., Miller, Joanne L., Kent, Raymond D., editors. *Papers in Speech Communication: Speech Processing*. Acoustical Society of America, Woodbury, New York, (1991).
- [6] Loizou, Philip. "COLEA; A MATLAB software tool for speech analysis". University of Arkansas at Little Rock. <http://www.utdallas.edu/~loizou/speech/colea.htm>, (consultado en el año 2005).
- [7] Lyman Ott, R. "An Introduction to statical methods and data análisis" Edition four, Ed. Duxbury Press 1992.