Revista Ingeniería UC, Vol. 24, No. 2, Agosto 2017 170 - 184

Development of free software for the spectral analysis of pathological voices

Yanina Perdomo[a], Ahmad Osman[*,b], Jesús Jiménez[b]

[a]*Escuela de Telecomunicaciones. Facultad de Ingeniería, Universidad de Carabobo, Valencia, Venezuela.*
[b]*Centro de Análisis y Tratamiento de Señales. Facultad de Ingeniería, Universidad de Carabobo, Valencia, Venezuela.*

Abstract.-

The creation of free software for the spectral analysis of pathological voices using efficient algorithms for the computation of the *Discrete Fourier Transform* (DFT), is considered. For this, the state of the art review of the *Fast Fourier Transform* (FFT) algorithms was performed and the implementation was proceeded, in code Fortran, of the most significant. Then, through Python's importable modules, a graphical interface for the interactive use of the algorithms was designed. From the results of the FFT, some spectral parameters were determined: power spectrum, formant estimation, pitch estimation, spectral envelope extraction and spectrogram. Finally, we performed tests of estimation of parameters of some speech samples, obtaining results highly approximated to Praat software.

Keywords: Spectral Analysis; Voice Signal; Discrete Fourier Transform;

Desarrollo de software libre para el análisis espectral de voces patológicas

Resumen.-

Se plantea la creación de un software libre para el análisis espectral de voces patológicas usando algoritmos eficientes para el cómputo de la *Transformada Discreta de Fourier* (TDF). Para esto, se realizó la revisión del estado del arte de los algoritmos de la *Transformada Rápida de Fourier* (TRF) y se procedió a la implementación, en código Fortran, de los más significativos. Luego, se diseñó una interfaz gráfica para el uso interactivo de los algoritmos, mediante módulos importables a Python. Se determinaron, a partir de los resultados de la FFT, algunos parámetros espectrales como: espectro de potencia, estimación de formantes, estimación de pitch, extracción de la envolvente espectral y espectrograma. Finalmente, se realizaron pruebas de estimación de parámetros de algunas muestras de voz, obteniendo resultados altamente aproximados a los del software *Praat*.

Palabras clave: Análisis Espectral Señal de Voz Transformada Discreta de Fourier

Recibido: febrero 2017
Aceptado: mayo 2017

1. Introducción.

La voz humana ha pasado a ser un importante objeto de investigación en distintas especialidades, ya que se ha demostrado que mediante el análisis de la señal de voz producida por las personas se pueden determinar características como el sexo, la edad e incluso, investigaciones recientes en el área de la ingeniería biomédica demuestran que es posible identificar patologías en las personas. Lo anterior constituye un método alternativo, no invasivo y además menos costoso, que permite la detección temprana e incluso la evaluación a distancia del progreso de determinadas patologías [1, 2, 3]

El diagnóstico está basado en la extracción de las características de las señales de voz que

*Autor para correspondencia
Correo-e: aosman@uc.edu.ve (Ahmad Osman)

pueden determinar la calidad de la misma, esto se realiza a través de técnicas de procesamiento digital de señales. Habitualmente, los parámetros de interés son el espectro de potencia, la frecuencia fundamental y los formantes, perceptibles en el dominio de la frecuencia y a partir de los cuales pueden ser determinadas otras características más específicas [4].

Existen diferentes herramientas para realizar el análisis espectral de las señales de voz. Sin embargo, en esta investigación se utiliza la Transformada Discreta Fourier como herramienta base; cuyo rendimiento ha sido probado en distintas investigaciones aplicadas a la extracción de parámetros de señales de voz bajo distintas condiciones de análisis. Ahora bien, las señales de voz son no estacionarias, sin embargo, su análisis se realiza en intervalos cortos en los cuales se puede asumir que la señal es estacionaria. La técnica usada para este tipo de análisis se conoce como Transformada de Fourier de Tiempo Reducido (STFT) [4, 5, 6, 7].

En este sentido, la TDF puede ser determinada a través de algoritmos más eficientes, estos se conocen, de manera colectiva, como algoritmos FFT, cuya ventaja principal es la eficiencia derivada de la reducción del número de operaciones y del tiempo de cómputo. Ahora bien, los parámetros espectrales de las señales de voz, se pueden estimar mediante técnicas basadas en la FFT. Por esta razón se consideró conveniente desarrollar un software libre para realizar análisis espectral de voz patológicas que emplee el cómputo de la FFT como cálculo fundamental [8, 9, 10].

Para lograr este fin, se realizó una revisión del estado del arte de los algoritmos de cómputo de la Transformada Discreta de Fourier(TDF). Posteriormente se realizó un ejercicio de selección de los algoritmos más significativos, para luego implementarlos en función de la realización del análisis espectral de las señales. de voz. Adicionalmente, se construyó una interfaz gráfica, para el manejo interactivo de cada uno de los algoritmos codificados y de los parámetros espectrales que se desean calcular.

2. Características de las señales de voz

De acuerdo con la acción de las cuerdas vocales, las señales de voz pueden ser clasificadas en vocales y no vocales. En las señales vocales, que consisten en la generación de sonidos melódicos, el tracto vocal tiene un comportamiento similar al de una cavidad resonante; para este tipo de señales, el sonido se produce como efecto de la vibración de las cuerdas vocales, las cuales modifican el área de la traquea al abrirse y cerrarse produciendo un tren de pulsos casi periódico. En el dominio de la frecuencia estas señales están conformadas por armónicos, como consecuencia de su casi periodicidad en el dominio del tiempo, y una envolvente espectral debida al tracto vocal. En las señales no vocales, que consisten en la generación de sonidos sordos, las cuerdas vocales permanecen abiertas y el aire fluye libremente por el tracto vocal, por lo que están formadas por una contribución desordenada de componentes frecuenciales y presentan una aleatoriedad similar a la del ruido blanco [11, 12].

2.1. Frecuencia fundamental y pitch

La frecuencia fundamental es la frecuencia de vibración de las cuerdas vocales y se puede interpretar como el número de veces que estas se abren y cierran por segundo. También es conocida como tono habitual, pues es el nivel óptimo en el cual la voz es producida sin esfuerzo y sin tensión en la laringe. Esta frecuencia varía de acuerdo al sujeto y las características de longitud, grosor y tensión de sus cuerdas vocales, sin embargo los valores típicos en adultos se encuentran entre $137Hz$ para los hombres y $207Hz$ para las mujeres. Su determinación resulta de interés para identificar los patrones de vibración de las cuerdas vocales, permitiendo detectar alguna alteración de los mismos en el caso de que existan patologías [13, 14, 15, 16].

El pitch o tono percibido, es la percepción del oyente a cambios de frecuencia, es un fenómeno psicológico. En contraste con la frecuencia fundamental, ésta última, es la manifestación de un fenómeno físico y se ve afectada por las características del tracto vocal del sujeto [17].

2.2. Formantes

Los formantes del habla se corresponden con las resonancias de baja frecuencia en el tracto vocal, son los máximos que se producen en la envolvente del espectro de potencia, y pueden ser identificados en un espectrograma. Esta característica sólo se observa en las señales vocales, pues las no vocales tienen una estructura ruidosa como se describió anteriormente. Los tres primeros formantes de la señal de voz, denotados como $F1$, $F2$ y $F3$ contienen suficiente información acerca de la señal de voz y son considerados como la fuente principal de información espectral [12, 18, 19].

3. La Transformada Discreta de Fourier en la estimación de parámetros espectrales de voz.

La Transformada Discreta de Fourier (TDF) puede obtenerse a partir de *La Transformada de Fourier en Tiempo Discreto* (TFTD). Esta última viene dada por las siguientes expresiones:

$$x\,[n] = \frac{1}{2\pi}\int_{2\pi} X(e^{j\omega})e^{j\omega n}\,d\omega \tag{1}$$

$$X(e^{j\omega}) = \sum_{n=-\infty}^{+\infty} x\,[n]\,e^{-j\omega n} \tag{2}$$

definidas para una secuencia no periódica de duración finita $x\,[n]$. La ecuación (1), se conoce como ecuación de síntesis, mediante la cual es posible reconstruir la señal de voz como una combinación lineal de exponenciales complejas. La ecuación 2, representa la ecuación de análisis, obtenida a través de la proyección ortogonal del vector secuencia $x[n]$ sobre una base de vectores ortogonales de un subespacio de funciones exponenciales complejas a diferentes frecuencias [20].

En la ecuación (2), $X(e^{j\omega})$ es una función contínua y periódica cada 2π en el dominio de la frecuencia; es por ello que, para efectos prácticos, se define de la *Transformada Discreta de Fourier*, a fines de obtener una representación del espectro en forma discreta que permita el análisis y la implementación de los algoritmos de cómputo para la extracción de parámetros de la voz. La expresión de la *Transformada Discreta de Fourier* se muestra a continuación:

$$X\,[k] = \sum_{n=0}^{N-1} x\,[n]\,W_N^{kn},$$

donde

$$W_N = e^{-j\frac{2\pi}{N}}$$

El término W_N^{kn}, es una función periódica de kn con periodo N, por lo que existirán N raíces sobre la circunferencia unitaria del plano complejo, que también serán periódicas. Cada una de estas raíces recibe el nombre de factor de rotación debido a que la multiplicación de un número por una de ellas cambia la fase de ese número sin cambiar su magnitud.

3.1. El Cómputo de la Transformada Discreta de Fourier en la estimación de parámetros espectrales de voz

El cálculo de la TDF de una secuencia de entrada $x[n]$ de N muestras, por el método directo requiere N^2 multiplicaciones complejas y $N(N-1)$ sumas complejas, si la función es expresada en función de operaciones con números reales, requiere $4N^2$ multiplicaciones reales y $N(4N-2)$ sumas reales [21, 20].

La TDF tiene una enorme capacidad para mejorar su eficiencia aritmética, debido a la periodicidad, simetría y ortogonalidad de las funciones base y la relación con la convolución, por esta razón la aplicación del método directo para determinar la TDF es considerado básicamente ineficiente porque no explota las propiedades de simetría y periodicidad del factor de fase W_N. Es por ello, que en la práctica han surgido técnicas para el análisis de señal de manera eficiente como lo es *la Transformada Rápida de Fourier* (FFT), aligerando el cómputo de la Transformada Discreta de Fourier, pues el número de cálculos aumenta en función del cuadrado de la cantidad de muestras N a transformar, de hecho, antes de la invención de esta técnica, el análisis de Fourier era una herramienta principalmente teórica. La potencia real del método FFT es que, a menudo, la división se puede aplicar de forma recursiva a

los subproblemas, lo que conduce a una reducción del orden de complejidad [21, 22, 23, 24, 25].

En la presente investigación se realizó la revisión y selección de los algoritmos más significativos, a través de un análisis de rendimiento relevancia y eficiencia mediante simulaciones de pruebas con miras a codificarlos en Python. Se adaptó la codificación de los algoritmos de cómputo más usados para la estimación de parámetros espectrales de voz pertenecientes a tres familias: los algoritmos para un número de muestras N potencia entera de 2, que incluyen los algoritmos de diezmado en tiempo(DIT) y en frecuencia(DIF); los algoritmos de factores primos(PFA) y los algoritmos derivados del análisis de Goertzel [20].

En una primera selección se determinó usar los siguientes algoritmos:

1. DIF base-2
2. DIT base-2
3. PFA
4. Goertzel
5. TDF directa

Fueron escogidos los algoritmos de base 2, porque además de ser eficientes, comparados con el algoritmo directo por definición de la TDF, son los algoritmos clásicos de FFT y la mayoría de las aplicaciones prácticas de la TDF están basados en los mismos.

Asimismo, se escogió el Algoritmo de Factores Primos (PFA) el cual es aplicable cuando el número de muestras de la señal de entrada puede descomponerse en factores relativamente primos, y el algoritmo de Goertzel de segundo orden, que es aplicable a secuencias de cualquier longitud; pues ambos suponen una mejora en eficiencia con respecto a la TDF por definición.

Luego de esta selección preliminar, se realizaron simulaciones de prueba para cada uno de los algoritmos seleccionados, se disponía de códigos desarrollados en MATLAB y Fortran, sin embargo para la aplicación final se decidió realizar la adaptación definitiva a Fortran, ya que es el lenguaje de programación estándar de mayor eficiencia en la computación científica. Por esta razón fueron empleadas adaptaciones de los algoritmos disponibles en [23], escritos en FORTRAN 77. El procedimiento para las simulaciones realizadas consistió en determinar la TDF de secuencias aleatorias, mediante los algoritmos bajo prueba y comparar los resultados con la TDF directa de la definición y las determinadas mediante la función que ofrece el paquete para computación científica con Python, Numpy [26].

Una vez ejecutado lo anterior, se realizó un estudio de las posibilidades existentes para codificar los algoritmos en Python: traducción directa del lenguaje original a Python o la creación de módulos de extensión importables a Python. No se decidió realizar la traducción desde Fortran a Python, si no que fueron creados módulos de extensión que son importables al código Python. Este es un paradigma de programación bastante utilizado en la actualidad que combina lenguajes de diferentes niveles para obtener mayor eficiencia.

Python es un lenguaje de alto nivel que ofrece un entorno de programación interactivo que simplifica y acelera el desarrollo de modelos computacionales pero la velocidad para resolver el modelo puede resultar desfavorable. Por esta razón se usó librerías de bajo nivel y así obtener un mejor rendimiento, esto se debe a que, de esta manera, se ejecuta código en un módulo de extensión, lo cual es equivalente a que la máquina virtual de Python ejecute código máquina directamente en vez de código de bytes de alto nivel. Esto disminuye el gasto de recursos del intérprete, mientras se ejecutan las operaciones dentro del módulo de extensión

El software desarrollado tiene dos opciones de análisis: *local* y *stft*. En caso de que el análisis sea *local* (la longitud de la ventana es igual al número de muestras de la señal de entrada), se usa la función `AnalisisLocal`, código desarrollado en esta investigación que procede a enventanar la señal, ajustar la longitud a la siguiente potencia de dos, realizar el *zero padding*, determinar la TDF y calcular los parámetros que hayan sido solicitados por el usuario.

Ahora bien, si el análisis es mediante la técnica *stft* se usa la función definida en Python, en la cual se determina el número de tramas a partir

de la longitud de la ventana y del parámetro de solapamiento (`noverlap`), luego se procede para cada trama de manera similar a la función `AnalisisLocal`. El código desarrollado para esta función es una adaptación del código MATLAB que se encuentra en [18].

Cabe destacar que el *zero padding* realizado a las señales bajo estudio después de ser enventanadas, se realizó insertando las muestras de valor cero en la mitad de la señal/trama bajo análisis, esto se conoce como *zero padding* de fase cero, el cual es la opción correcta cuando se utilizan ventanas de fase cero como las empleadas en esta investigación [18].

3.2. La técnica del enventanado y la Transformada de Fourier en Tiempo Reducido (TFTR) en la estimación de parámetros espectrales de voz

En el análisis espectral de señales de voz, casi siempre se analiza un segmento corto de señal (de 10 a 40ms) en lugar de toda la señal. Por lo tanto, el primer paso en el análisis tiempo-frecuencia de una señal de audio consiste en la segmentación de la misma; la forma correcta de extraer los segmentos cortos es multiplicar la señal por una función ventana [18, 20].

Una función ventana no es más que una envolvente específica que se aplica a la señal que se desea analizar. En general, la mayoría de las ventanas usadas en análisis espectral, tienen características pasa bajos en el dominio de la frecuencia y una forma muy similar a la curva de Gauss en el tiempo. El objetivo principal de usar una ventana de desvanecimiento es evitar las discontinuidades abruptas en los bordes durante la segmentación de la señal [18, 20].

Las dos características más importantes relacionadas a las ventanas de una longitud dada son: el ancho de su banda de paso (lóbulo principal) y la atenuación en su banda de rechazo (lóbulos secundarios). El ancho del lóbulo principal impone un límite de la mínima distancia entre dos picos que en el dominio de la frecuencia, ya que si éstos están más cerca que el ancho del lóbulo principal, serán integrados en un solo pico. Por supuesto, incrementando la longitud de la ventana se producen lóbulos principales estrechos, lo que ayuda con la resolución de picos espectrales muy cercanos.

Para una ventana de longitud M muestras, el tipo de ventana controla la supresión del lóbulo lateral (a expensas de la resolución cuando M es fijo) y la longitud controla la resolución en frecuencia. El beneficio principal de la elección de una buena función ventana en el análisis de Fourier es la minimización de los lóbulos laterales que causan diafonía en el espectro estimado de una frecuencia a otra [18, 27].

Cabe destacar que la resolución del espectro obtenido a partir de una señal enventanada, se ve limitada por el principio de incertidumbre de Heisenberg, de manera que no es posible lograr alta resolución en ambos dominios (temporal y frecuencial) de manera simultánea; por lo tanto la fijación de los parámetros de la función ventana dependerá del dominio que interese estudiar, ya que una ventana estrecha tendrá una alta resolución temporal, mientras que la resolución frecuencial será muy baja; y el empleo de una ventana ancha, tendrá una baja resolución temporal, mientras que la resolución frecuencial será alta [27].

De acuerdo con la literatura consultada, las ventanas mayormente empleadas son: la ventana rectangular (cuya característica principal es que su amplitud es constante) la ventana de Hanning (atenúa la señal en los bordes), Hamming (similar a la ventana de Hanning) y Blackman; comparadas entre sí, la ventana rectangular y de Hanning tienen un lóbulo principal muy definido y baja atenuación de frecuencias parásitas, mientras que la ventana de Hanning y Blackman presentan características similares entre sí y una mayor atenuación de frecuencias parásitas que las mencionadas anteriormente [20].

En algunos casos, es útil incorporar la técnica del enventanado haciendo uso de la TFTR cuya implementación computacional constituye una poderosa herramienta de propósito general para procesamiento de señales de voz. Ésta provee una clase particularmente útil de las distribuciones tiempo-frecuencia que especifican la amplitud compleja en función del tiempo y la frecuencia

para cualquier señal. En el análisis de señales de voz el ajuste de los parámetros de TFTR se hace pensando en la medición de las características de la voz en un intervalo de tiempo reducido, procurando lograr el equilibrio entre la resolución de los armónicos y la detección tanto del pitch como de las variaciones de los formantes [27].

4. Técnicas espectrales usadas para la extracción de los parámetros fundamentales de las señales de voz

Para llevar a cabo la extracción de los parámetros fundamentales de las señales de voz, se consideró, en inicio, un modelo matemático aproximado del proceso de producción de la voz. Luego, se consideró la aplicación específica de técnicas para la extracción de los parámetros: formantes y pitch haciendo uso de los algoritmos de cómputo de la TDF seleccionados en la sección anterior.

4.1. Modelo de Producción de la Voz

En la Figura 1 se muestra el modelo (simplificado) del sistema de producción del habla, en este modelo la señal de voz $x[n]$ es considerada como la salida de un sistema lineal, que es excitado por un tren de impulsos en el caso de voz sonora o ruido en el caso de voz sorda. Los parámetros del modelo vienen dados por la selección sordo/sonoro, el tono (en caso de voz sonora) y los parámetros del filtro $H(z)$: la ganancia G y los coeficientes a_k [28].

$H(z)$ es un filtro que modela los efectos de la glotis, el tracto vocal y los labios en la producción de la voz. De esta manera, las contribuciones del modelo (excitación y filtro) están relacionadas en el dominio del tiempo por la siguiente ecuación de convolución:

$$x[n] = u[n] * h[n] \tag{3}$$

4.2. Extracción de Formantes Mediante LPC

La técnica de predicción lineal consiste en estimar el valor actual de una señal $x[n]$ como una

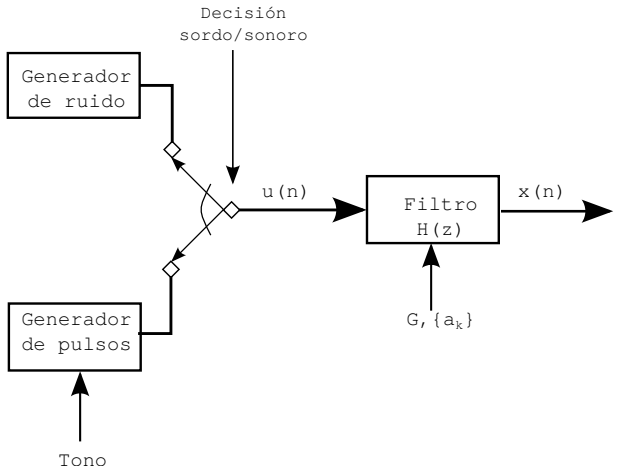

Figura 1: Modelo Lineal de Producción de Voz. Fuente [28]

combinación lineal de las muestras anteriores. El valor estimado se escribe como

$$\hat{x}(n) = -\sum_{k=1}^{N_{LP}} a_k x[n-k], \tag{4}$$

donde N_LP es el orden de predicción y a_k son los coeficientes de predicción.

Tomando en cuenta la ecuación (4), la señal de voz puede definirse como:

$$x(n) = -\sum_{k=1}^{N_{LP}} a_k x[n-k] + e[n], \tag{5}$$

donde $e[n]$ representa el error de predicción entre el valor real $x[n]$ y el valor estimado$\hat{x}(n)$. En el dominio z la ecuación anterior se plantea como

$$E(z) = A(z)X(z), \tag{6}$$

donde $A(z)$ es la función de transferencia y viene dada por:

$$A(z) = 1 + \sum_{k=1}^{N_{LP}} a_k z^{-k}, \tag{7}$$

A partir de lo anterior se puede establecer una relación en donde se puede considerar un sistema en el cual existe la señal $e[n]$ como entrada y $x[n]$ como salida a partir del sistema inverso a (6), en donde

$$X(z) = \frac{1}{A(z)}E(z) = H(z)E(z) \tag{8}$$

Por lo tanto, asumiendo que la señal $x[n]$ obedece al modelo (1), en el cual $e[n]$ es considerada la señal $u[n]$, $A(z)$ será el filtro inverso al filtro $H(z)$. $H(z)$ es el filtro que modela el tracto vocal, las frecuencias de resonancia de este filtro son los formantes. Además, gráficamente, los máximos de la respuesta del tracto vocal se corresponden también con los formantes [4, 19, 28].

Para la estimación de los formantes fue creada una función en Python que recibe como parámetros de entrada la señal, la longitud de la misma, la frecuencia de muestreo, el número de coeficientes a determinar y el nombre del algoritmo eficiente de TDF. El procedimiento se describe de la siguiente manera: Se realiza el preénfasis, para reducir el rango dinámico del espectro de la señal de voz, esto se hace mediante un filtro pasa altos, cuya función de transferencia es

$$H_p(z) = 1 - az^{-1}$$

el valor de a usualmente es fijado entre $0,9$ y 1, en este caso, se utilizó $a = 0,9$, este filtro es implementado mediante la función `lfilter` del módulo `Scipy` de Python. Seguidamente se hace la determinación de los coeficientes mediante la función `lpc`, implementada igualmente desde un módulo de Python, la cual estima los coeficientes de predicción lineal usando el método de auto-correlación mediante la recursión de Levinson-Durbin. Para determinar el número de coeficientes se utilizó la siguiente relación

$$N = 2 + \frac{fs}{1000}$$

donde fs es la frecuencia de muestreo.

Luego de la determinación de los coeficientes, las frecuencias formantes se estiman con base en la relación entre los formantes y los polos del filtro del tracto vocal $H(z)$ (4.2). Recordando que el denominador de la función de transferencia está relacionado con los coeficientes de predicción lineal, éste se factoriza como:

$$1 + \sum_{k=1}^{N_{LP}} a_k z^{-k} = \prod_{i=1}^{N_{LP}} (1 - c_i z^{-1})$$

Donde c_i es un conjunto de números complejos donde cada par de polos conjugados representa una resonancia a la frecuencia:

$$\hat{F}_i = \left(\frac{fs}{2\pi}\right) \arctan\left[\frac{Im(c_i)}{Re(c_i)}\right]$$

Ahora bien, la raíz representa un formante si se cumple la siguiente condición:

$$\sqrt{Im(c_k)^2 + Re(c_k)^2} \geq 0,7$$

Para implementar este método en Python, se definieron las funciones `detcoef` y `detformantes`, la primera para realizar la determinación de los coeficientes de predicción lineal y la segunda para realizar la determinación de los formantes haciendo uso de la primera. [19]

Así mismo también se codificó un algoritmo para estimar la envolvente espectral haciendo uso de la función `detcoef` y de la función `Algoritmo`, que determina la FFT mediante los algoritmos eficientes.

4.3. Extracción del Pitch mediante Cepstrum

Basado en el modelo (1), la señal en el dominio del tiempo viene dada por la convolución de ambas contribuciones: la excitación y el tracto vocal. El análisis Cepstrum, es una técnica diseñada para separar estas componentes mediante una transformación de la señal $x[n]$ a un dominio en el que la convolución es una suma. Partiendo de (3), se toma la Transformada de Fourier (TF), luego se extrae la magnitud y se transforma a escala logarítmica para luego tomar la Transformada Inversa de Fourier (TIF). El diagrama de bloques se muestra en la Figura 2.

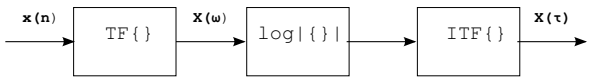

Figura 2: Transformación al Dominio Cepstral. Fuente [19]

Esta útima transformación (TIF), toma la función de vuelta en el dominio del tiempo, pero no es el mismo de la señal original, de hecho es una medición de la tasa de cambio de las magnitudes espectrales. Este dominio es el llamado Cepstrum y el eje del tiempo se denomina eje quefrency.

La extracción de la frecuencia fundamental de la voz mediante este análisis, consiste en: una vez realizada la transformación, ubicar el máximo pico en el cepstrum de potencia. La posición en el eje quefrency de este máximo está relacionada con la frecuencia fundamental de la voz mediante la siguiente expresión:

$$f_0 = \frac{fs}{qf_{max}}, \tag{9}$$

donde fs es la frecuencia de muestreo en Hertz y qf_{max} es la posición del máximo en el eje quefrency [19, 29]

Para la extracción del pitch en el dominio Cepstral, se creo una función llamada `Pitchceps`, los parámetros de entrada a la función creada para esta determinación son: la TDF de la señal/trama bajo estudio y la frecuencia de muestreo, luego se realizan las operaciones descritas en (4.3). Una vez que se ha tomado la Transformada Inversa de la señal, se determina su magnitud y se procede a buscar el valor máximo en el rango de 70-500 Hz, que se corresponderá con el pitch una vez hecha la transformación desde la escala quefrency a la escala lineal, mediante la relación (9).

4.4. El espectrograma en las señales de voz

El espectrograma es una representación de la distribución de energía de una señal en el plano tiempo-frecuencia, como una función que depende de ambos dominios, obtenida mediante la Transformada de Fourier de Tiempo Reducido (STFT), descrita anteriormente [30].

La gran ventaja del uso de espectrogramas para analizar señales de voz es que esta herramienta permite la representación de la frecuencia de cada componente armónico, la intensidad de cada uno y el instante del fenómeno, todo en el mismo gráfico. El eje de las abscisas corresponde al dominio del tiempo, el eje de las ordenadas corresponde a la frecuencia y la intensidad se muestra mediante escala de grises.

El espectrograma también se ve afectado por el principio de incertidumbre mencionado en secciones anteriores, de esta manera se condiciona la resolución tanto frecuencial como temporal que pueda obtenerse, existiendo así dos tipos de espectrograma: de banda estrecha y de banda ancha, el primero presenta mejor resolución frecuencial mientras que el segundo presenta mejor resolución temporal, la elección de cierto tipo de espectrograma dependerá de la finalidad del análisis realizado [31].

Los parámetros implicados en el espectrograma son la longitud de ventana, tipo de ventana, tamaño de salto y longitud de la TDF. La eficiencia de los resultados obtenidos dependerá de la elección de estos parámetros al momento de calcular el mismo [30, 18].

5. Pruebas de desempeño del «*softvoz*» versión alfa

5.1. Interfaz gráfica para el manejo de los algoritmos

Se planteó una interfaz que permitiera facilitar la realización de pruebas y el análisis de resultados referentes a la extracción de los parámetros de las voces. Para esto, se propuso un esquema sencillo que consistiera en cargar una señal de voz, seleccionar una porción, fijar los parámetros de análisis: ventana, longitud de ventana, algoritmo TDF, seleccionar los parámetros que se deseen extraer y la visualización de los resultados. Con este esquema en mente, se decidió incluir cuatro etapas que son importar y preparar señal para el análisis, parámetros de análisis, parámetros a determinar y resultados, como se muestra en la Figura 3.

5.2. Aplicación del software a muestras de voz

Se procedió a realizar pruebas de extracción de características con señales de voz producidas de manera natural, que consistieron en la pronunciación de manera sostenida de las cinco vocales, por parte de dos sujetos, uno de sexo masculino y de 24 años de edad y otro de sexo femenino de 54 años de edad, esto generó un total de 10 señales de voz.

Para cada una de las señales el experimento consistió en tomar una muestra de cinco segundos y extraer tanto el pitch como los formantes, mediante análisis de tiempo reducido, usando como herramienta el software desarrollado y otro

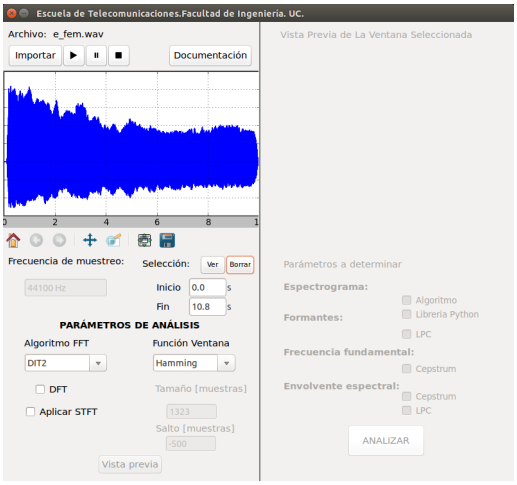

Figura 3: Ventana principal de la aplicación desarrollada.

software libre para análisis de señales de voz denominado Praat, para luego comparar los resultados. En el caso del análisis en tiempo reducido, se usó Hamming de longitud 1323 muestras 30ms como función ventana y avance/salto de 500 muestras. Los resultados mostrados corresponden a la media observada en la muestra de tiempo estudiada.

En las secciones siguientes se muestra, mediante tablas, los resultados obtenidos para cada uno de los parámetros, se usó como criterio, para realizar la comparación entre los resultados obtenidos, la diferencia porcentual entre las determinaciones, tomando como referencia el resultado obtenido mediante Praat.

5.2.1. Resultados de Extracción de Pitch

Los resultados en la extracción de pitch para cada uno de los sujetos considerados en este análisis, se muestra en las Tablas 1 y 2.

Tabla 1: Resultados de determinación de pitch mediante análisis de tiempo reducido para voz de sujeto de sexo masculino de 24 años de edad. Fuente: Propia.

Vocal	Pitch Estimado [Hz]	Pitch (Praat) [Hz]	Dif [%]
a	145,50	148,54	2,04
e	161,49	164,13	1,61
i	218,00	218,30	0,14
o	133,47	148,01	9,82
u	128,93	155,81	16,18

Se observa que, en el caso del sujeto de sexo masculino (ver Tabla 1), para todas las vocales la diferencia se produce por defecto y en general, se produce una media de diferencia de 5,96%. Se logra menos de 3% de diferencia en la determinación del pitch para la emisión de las tres primeras vocales (a,e,i); y para las últimas dos vocales (o,u) la diferencia aumenta. La mínima diferencia se produce en la emisión de la vocal «i», en el caso de la «o» la diferencia aumenta, pero aún así no supera el 10%, por lo que se considera dentro del límite de tolerancia; sin embargo en el caso de la vocal «u», la diferencia aumenta casi al doble de este último.

Se decidió variar la longitud de la función ventana a 1764 muestras (40 ms) y avance/salto de 650 muestras, aproximadamente 63% de solapamiento, para la determinación del pitch en el caso de la emisón de la vocal «u», obteniendo un valor de pitch estimado de 153,21 [Hz], lo que reduce la diferencia de 16,18% a 0,39%. De esta manera se evidencia que la variación del tamaño de la ventana influye en la determinación de los resultados.

Tabla 2: Resultados de determinación de pitch mediante análisis de tiempo reducido para voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia

Vocal	Pitch Estimado [Hz]	Pitch (Praat) [Hz]	Dif [%]
a	186,94	187,87	0,49
e	189,82	190,34	0,27
i	213,15	213,71	0,26
o	187,56	201,45	6,89
u	187,01	211,75	11,68

En el caso del sujeto de sexo femenino de 54 años de edad (ver Tabla 2), la media de diferencia general es de 3,92%, las diferencias mínimas, que no superan el 1% al igual que en el caso del sujeto de sexo masculino, se obtienen en la estimación de pitch para las tres primeras vocales (a,e,i), produciéndose la diferencia mínima en la señal correspondiente a la emisión de la vocal «i». Las diferencias aumentan para las últimas dos vocales (o,u) nuevamente y el valor de diferencia obtenido en la emisión de la vocal «o» se encuentra

en un rango tolerable. El valor de pitch obtenido en la emisión de la vocal «u», presenta mayor diferencia, aunque no tan elevado como en el caso del sujeto de sexo masculino.

Variando la longitud de la ventana, para la determinación del pitch en la emisión de la vocal «u», se obtiene un valor de 202,95 Hz, lo que disminuye la diferencia a 4,16%. Para esto la longitud de la ventana se incrementó a 2205 muestras y se ajustó el salto a 771 muestras (aproximadamente 65% de solapamiento).

5.2.2. Resultados en estimación de formantes

En las Tablas 3 y 4 se muestra la estimación del primer formante. En el caso del sujeto de sexo masculino, se obtuvo una media de diferencia porcentual absoluta de 5,61% y, en general, la diferencia se encuentra por debajo del 10%, el mínimo valor de diferencia de estimación se encontró en la vocal «a» y el máximo en la vocal «i». En el caso, del sujeto de sexo femenino el valor medio de diferencia porcentual absoluto obtenido fue de 8,98%, obteniendo la mayor diferencia en la estimación del primer formante en la vocal «i».

Tabla 3: Resultados de extracción de primer formante mediante análisis de tiempo reducido para voz de sujeto de sexo masculino de 24 años de edad. Fuente: Propia

Vocal	F1 Estimado [Hz]	F1 (Praat) [Hz]	Dif [%]
a	688,20	708,81	2,91
e	345,03	369,11	6,52
i	357,51	388,84	8,06
o	449,46	484,98	7,32
u	355,04	343,96	-3,22

En las Tablas 5 y 6 se muestra la estimación del segundo formante. En el caso del sujeto de sexo masculino, una media de diferencia absoluta de 18,81%, obteniendo la mayor diferencia en la estimación realizada en la pronunciación de la vocal «e». En el caso del sujeto de sexo femenino se observó una media de diferencia absoluta de 27,90%, obteniendo la mayor diferencia en el caso de la letra «u».

Tabla 4: Resultados de extracción de primer formante mediante análisis de tiempo reducido para voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia

Vocal	F1 Estimado [Hz]	F1 (Praat) [Hz]	Dif [%]
a	766,05	811,77	5,63
e	420,35	471,10	10,77
i	323,40	383,46	15,66
o	430,85	445,34	3,25
u	360,46	396,77	9,15

Tabla 5: Resultados de extracción de segundo formante mediante análisis de tiempo reducido para voz de sujeto de sexo masculino de 24 años de edad. Fuente: Propia

Vocal	F2 Estimado [Hz]	F2 (Praat) [Hz]	Dif [%]
a	1483,49	1473,98	-0,65
e	1154,85	2424,38	52,37
i	1962.04	2348,80	16,47
o	968,32	1026,03	5,62
u	884,17	743,45	-18,93

Tabla 6: Resultados de extracción de segundo formante mediante análisis de tiempo reducido para voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia

Vocal	F2 Estimado [Hz]	F2 (Praat) [Hz]	Dif [%]
a	1380,17	1435,77	3,87
e	1841,94	2287,81	19,49
i	1875,45	2556,17	26,63
o	900,69	844,04	-6,71
u	1209,09	661,44	-82,80

En las Tablas 7 y 8 se muestra la estimación del tercer formante. La mayor diferencia se observó en la estimación realizada en la señal correspondiente a la vocal «e»del sujeto de sexo masculino, en cuya señales se obtuvo una media de diferencia absoluta de 20,51%. En el caso del sujeto de sexo femenino se observó una diferencia absoluta en promedio de 6,46%, observando la mayor diferencia en la vocal «o».

En las Tablas 9 y 10 se muestra la estimación del cuarto formante. Acá se observa la mayor diferencia porcentual en la vocal «i»en el caso del sujeto de sexo masculino, y en promedio la

Tabla 7: Resultados de extracción de tercer formante mediante análisis de tiempo reducido para voz de sujeto de sexo masculino de 24 años de edad. Fuente: Propia

Vocal	F3 Estimado [Hz]	F3 (Praat) [Hz]	Dif [%]
a	2446,40	2425,55	-0,86
e	1154,85	3009,02	61,62
i	2649,48	3301,01	19,74
o	2276,98	2478,77	8,14
u	2288,45	2606,05	12,19

Tabla 8: Resultados de extracción de tercer formante mediante análisis de tiempo reducido para voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia.

Vocal	F3 Estimado [Hz]	F3 (Praat) [Hz]	Dif [%]
a	2328,24	2483,68	6,26
e	2619,49	2748,71	4,70
i	2702,79	2803,67	3,60
o	2424,27	2778,01	12,73
u	2513,32	2645,91	5,01

diferencia absoluta es de 15,05%. En el caso del sujeto de sexo femenino la media de diferencia absoluta fue de 12,17% y la mayor diferencia se observó en la vocal «o».

Tabla 9: Resultados de extracción de cuarto formante mediante análisis de tiempo reducido para voz de sujeto de sexo masculino de 24 años de edad. Fuente: Propia.

Vocal	F4 Estimado [Hz]	F4 (Praat) [Hz]	Dif [%]
a	3665,86	3766,54	2,67
e	3075,41	4046,44	24,00
i	3358,03	4817,64	30,30
o	3359,49	4010,20	16,23
u	3376,34	3447,27	2,06

Para todas las vocales, las mejores estimaciones (aquellas con menor diferencia) fueron la del primer y tercer formante, en el caso del sujeto de sexo femenino y la estimación del primer formante en el caso del sujeto de sexo masculino. Todas con un diferencia absoluta media menor a 10%.

En las Tablas 11 y 12 se muestran las diferencias absolutas entre las estimaciones realizadas con

Tabla 10: Resultados de extracción de cuarto formante mediante análisis de tiempo reducido para voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia.

Vocal	F4 Estimado [Hz]	F4 (Praat) [Hz]	Dif [%]
a	3272,40	3449,28	5,13
e	3498,09	3833,26	8,74
i	3305,33	3805,56	13,14
o	3257,05	4316,98	24,55
u	3596,52	3965,17	9,30

Tabla 11: Diferencia absoluta porcentual en las estimaciones realizadas para el sujeto de sexo masculino. Fuente: Propia.

Vocal	Dif F1 [%]	Dif F2 [%]	Dif F3 [%]	Dif F4 [%]	Media [%]
a	2,91	0,65	0,86	2,67	1,77
e	6,52	52,37	61,62	24,00	36,13
i	8,06	16,47	19,74	30,30	18,64
o	7,32	5,62	8,14	16,23	9,33
u	3,22	18,93	12,19	2,06	9,10
Media	5,61	18,81	20,51	15,05	14,99

Tabla 12: Diferencia absoluta porcentual en las estimaciones realizadas para el sujeto de sexo femenino. Fuente: Propia.

Vocal	Dif F1 [%]	Dif F2 [%]	Dif F3 [%]	Dif F4 [%]	Media [%]
a	5,63	3,87	6,26	5,14	5,22
e	10,77	19,49	4,70	8,74	10,93
i	15,66	26,63	3,60	13,14	14,76
o	3,25	6,71	12,73	24,55	11,81
u	9,56	82,77	5,01	9,30	26,66
Media	8,98	27,90	6,46	12,17	13,88

Praat y las realizadas con el software desarrollado. Para el sujeto de sexo masculino, los formantes con menor diferencia fueron los obtenidos para las vocales «a», «o» y «u»; en el caso del sujeto de sexo femenino fueron los obtenidos para las vocales «a» y «e». Las estimaciones con mayor diferencia fueron las realizadas para la vocales «e» y «u», para el sexo masculino y femenino, respectivamente. Para todos los formantes,las estimaciones con menor diferencia, y por tanto mayor aproximación, fueron las realizadas para la vocal «a», en ambos sujetos. En el caso del sujeto de

sexo masculino con una diferencia absoluta media de 1,77% y en el sujeto de sexo femenino 5,22%.

Adicionalmente, se realizó un análisis local en la señal de la vocal «u» emitida por el sujeto de sexo femenino, ya que en el segundo formante obtenido para la misma se observó la mayor diferencia porcentual (82,77%).En este nuevo análisis la determinación de los formantes se realizó de manera gráfica, a partir de la envolvente espectral determinada mediante LPC. En principio, se realizó el análisis a una trama de un segundo, cuya envolvente obtenida se muestra en la Figura 4 y los resultados obtenidos se muestran en la Tabla 13. Ahora bien, al observar la Figura 4 resulta evidente que el segundo formante no es perceptible, esto puede deberse a que la resolución en frecuencia no es lo suficientemente buena como para resolver dos picos tan cercanos,por tanto el máximo correspondiente al segundo formante está siendo sumado al máximo que corresponde al primer formante. Sin embargo, si se comparan los valores de diferencia absoluta porcentual obtenidos para cada formante (ver Tabla 13) con los obtenidos con el análisis anterior (Tabla 12), existe una disminución en la diferencia obtenida en la estimación del primer y cuarto formante.

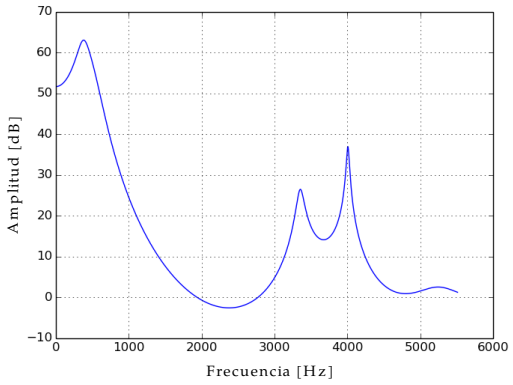

Figura 4: Envolvente espectral obtenida mediante LPC, trama 1s. Vocal «u», sujeto de sexo femenino de 54 años de edad. Fuente: Propia

De la misma manera, se realizó la extracción de formantes a partir de la envolvente espectral, pero esta vez con una trama de 30 ms. La envolvente espectral obtenida se muestra en la Figura 5 y los resultados obtenidos se encuentran en la Tabla 14. Se observa que en esta nueva gráfica (5), son perceptibles los cuatro primeros formantes, asimismo es posible estimar el valor del segundo formante con una diferencia porcentual absoluta de 15,34%. Con este análisis, mejora la estimación el segundo y cuarto formante; la diferencia absoluta para los otros dos formantes aumenta, sin embargo, la media de diferencia absoluta para los formantes estimados es de 11,65%, lo que implica una disminución de un poco más de la mitad del valor conseguido en el primer análisis 26,66%.

Tabla 13: Resultados de estimación de formantes mediante análisis local a trama de 1s de señal de voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia.

Técnica	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]
Envolvente	381	0	3595	4007
Praat	393,94	659,76	3359,41	3997,83
$\|Dif\|$ [%]	3,28	100%	7,01	0,23

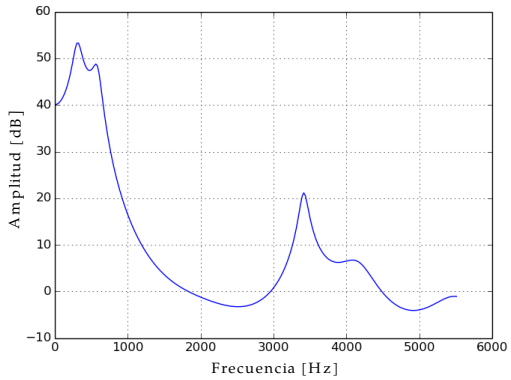

Figura 5: Envolvente espectral obtenida mediante LPC, Trama 30ms. Fuente: Propia

Tabla 14: Resultados de estimación de formantes mediante análisis local a trama de 30ms de señal de voz de sujeto de sexo femenino de 54 años de edad. Fuente: Propia.

Técnica	F1 [Hz]	F2 [Hz]	F3 [Hz]	F4 [Hz]
Envolvente	315	560	3400	4100
Praat	396,77	661,44	3125,66	4024,67
$\|Dif\|$ [%]	20,61	15,34	8,78	1,87

5.2.3. Estimación de parámetros con señales grabadas en ambiente controlado

Se decidió realizar la estimación de parámetros a dos señales de voz, que consistían en la emisión de la vocal «a» sostenida durante cinco segundos, pertenecientes a dos sujetos diferentes, capturadas bajo ambiente controlado y con micrófono de alta directividad, el procedimiento para realizar la estimación fue similar al usado en las pruebas anteriores, a continuación se presentan los resultados obtenidos para cada una de las señales.

Con respecto a la estimación del pitch, los resultados se muestran en la Tabla 15, donde es evidente que la mayor diferencia absoluta es de $0,12\%$

Tabla 15: Resultados de estimación de pitch mediante análisis de tiempo reducido a voces bajo ambiente controlado. Fuente: Propia

Sujeto	Pitch Estimado [Hz]	Pitch (Praat) [Hz]	Dif [%]
1	208,46	208,21	-0,12
2	153,53	153,48	0,10

Asimismo, los resultados en extracción de formantes se muestran en las Tablas 16, 17, 18 y 19.

Tabla 16: Resultados de extracción de primer formante mediante análisis de tiempo reducido a voces bajo ambiente controlado. Fuente: Propia

Sujeto	F1 Estimado [Hz]	F1 (Praat) [Hz]	Dif [%]
1	821,01	851,09	3,53
2	601,10	607,02	0,98

Tabla 17: Resultados de extracción de segundo formante mediante análisis de tiempo reducido a voces bajo ambiente controlado. Fuente: Propia

Sujeto	F2 Estimado [Hz]	F2 (Praat) [Hz]	Dif [%]
1	1326,38	1343,54	1,28
2	963,25	955,59	-0,80

En general, para todos los formantes extraídos en el sujeto 1, la menor diferencia absoluta es de $1,28\%$ y la mayor es $6,47\%$, con una media absoluta de $3,48\%$ y para el sujeto 2, la menor diferencia absoluta obtenida es de $0,70\%$ y la mayor es de $1,44\%$ con una media de $0,98\%$.

Tabla 18: Resultados de extracción de tercer formante mediante análisis de tiempo reducido a voces bajo ambiente controlado. Fuente: Propia

Sujeto	F3 Estimado [Hz]	F3 (Praat) [Hz]	Dif [%]
1	2915,37	3116,92	6,47
2	2574,86	2612,59	1,44

Tabla 19: Resultados de extracción de cuarto formante mediante análisis de tiempo reducido a voces bajo ambiente controlado. Fuente: Propia

Sujeto	F4 Estimado [Hz]	F4 (Praat) [Hz]	Dif [%]
1	3795,58	3898,12	2,63
2	3126,97	3149,12	0,70

Para ambos parámetros, pitch y formantes, es evidente que las diferencias obtenidas son menores comparadas con las estimaciones de los apartados 5.2.1 y 5.2.2, por tanto las condiciones de captura de la señal de voz también influyen en los resultados.

6. Conclusiones.

Se inició el desarrollo de un software libre, bajo lenguaje Python, que permite realizar el análisis de señales de voz pregrabadas, mediante análisis local o en tiempo reducido, usando Transformada Discreta de Fourier; diseñado para que el usuario tenga la libertad de fijar los parámetros implicados en el análisis de las señales de voz, con la posibilidad de estimar el pitch, los formantes, espectrograma, y envolvente de la señal.

Las características espectrales fundamentales de las señales de voz son: la frecuencia fundamental y los formantes. Éstas, pueden ser estimadas a partir de distintas técnicas de extracción, de las cuales resultaron relevantes: el análisis en dominio cepstral, para la estimación de la frecuencia fundamental, y el análisis de predicción lineal para los formantes.

Los algoritmos más significativos de FFT, son los algoritmos de base dos en sus versiones Diezmado en Frecuencia (DIF) y Diezmado en Tiempo (DIT), ambos diseñados para cuando el número de muestras de la entrada es una potencia entera de dos; estos son los algoritmos clásicos de FFT por lo que la mayoría de las aplicaciones prácticas están basadas en los mismos.

La implementación de los algoritmos de FFT para realizar el análisis espectral de voz es factible mediante el uso de técnicas que consideren la característica de no estacionariedad de las señales de voz, esto es realizar el análisis en tramas de tiempo reducido en un rango de $10 - 40ms$, intervalo en el que se considera que las características de la voz no varían de manera considerable.

Se comprobó que la variación de los parámetros de análisis: función ventana, longitud y solapamiento, influye en los resultados obtenidos en la estimación de las características esenciales de las señales de voz.

Con respecto a la estimación del pitch, se obtuvieron estimaciones con un mínimo margen de diferencia con respecto a otra aplicación de código libre desarrollada con el mismo fin. Con respecto a la extracción de formantes, se comprobó que la mejor estimación se realiza a través de la inspección visual de la envolvente espectral de la señal de voz.

Se vislumbra que la ejecución de módulos en lenguaje *Fortran* a través de *Python*, abre posibilidades de nuevas formas de ejecución de código a la luz de mejorar el rendimiento en el análisis de la voz.

Referencias

[1] R. J. Moran, R. B. Reilly, P. Chazal, and P. D. Lacy. Telephony-based voice pathology assessment using automated speech analysis. *IEEE Transactions on biomedical engineering*, 53(3):468–477, Marzo 2006.

[2] F. Martínez. Trastornos del habla y la voz en la enfermedad de parkinson. *Revista de Neurología*, 51(9):542–550, 2010.

[3] A. Tsanas, M. Little, P. McSharry, and L. Ramig. Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests. *IEEE Transactions on biomedical engineering*, 57(4):884–893, 2010.

[4] P. Del Pino. Identificación de los parámetros espectrales que determinan la calidad de la voz. Tesis de Maestría, Universidad de Carabobo, Venezuela, 2003.

[5] C. Jiménez, J.A. Díaz, P. Del Pino y H. Rothman. Aplicación de la transformada de wavelet para el análisis de señales de voz normales y patológicas. *Revista Ingeniería UC*, 15(1):7–13, 2008.

[6] J. Bernal, P. Gómez y J. Bobadilla. Una visión práctica en el uso de la transformada de Fourier como herramienta para el análisis espectral de la voz. *Estudios de Fonética Experimental. Universitat de Barcelona*, 10:77–105, 1999.

[7] P. Del Pino, I. Granadillo, M. Miranda, C. Jiménez y J. A. Díaz. Diseño de un sistema de medición de parámetros característicos y de calidad de señales de voz. *Revista Ingeniería UC*, 15(1):13–20, 2008.

[8] P. S. R Diniz, E. A. B. Da Silva, and S. L Netto. *Digital Signal Processing: System Analysis and Design*. Cambridge University Press, 2010.

[9] V Madisetti. *Digital Signal Processing Fundamentals*. CRC Press, 2010.

[10] D. Sundararajan. *Digital Signal Processing: Theory and Practice*. World Scientific, 2003.

[11] M.C.E. Boquera. *Servicios avanzados de telecomunicación*. Díaz de Santos, 2003.

[12] D. Salcedo and A. Teixeira. Diseño de un sistema de reconocimiento del habla para controlar dispositivos eléctricos. *Tekhne. Revista de la Facultad de Ingeniería. UCAB.*, 10:92–106, 2007.

[13] J. A. Díaz, C. Sapienza, H. Rothman y Y. Natour. Algoritmo robusto para la detección de la frecuencia fundamental de la voz basado en el espectograma. *Revista Ingenier´a UC.*, 10(3):7–16, 2003.

[14] I. Cobeta, F. Núñez y S. Fernández. *Patología de la voz:*. Marge books, 2013.

[15] M. C. A. Jackson-Menaldi. *La voz normal*. Editorial Médica Panamericana, 1992.

[16] C. Suárez, L. M. Gil-Garcedo, J. Marco, J.E. Medina, P. Ortega y J. Trinidad. *Tratado de Otorrinolaringología y Cirugía de Cabeza y Cuello*. Ed. Médica Panamericana.

[17] L.J. Raphael, G.J. Borden, and K.S. Harris. *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*. Communication sciences. Lippincott Williams & Wilkins, 2007.

[18] J.O.I.I.I. Smith. *Spectral Audio Signal Processing*. W3K, 2011.

[19] Med Kammoun, Dorra Gargouri, Mondher Frikha, and Ahmed Ben Hamida. Cepstrum vs. lpc: A comparative study for speech formant frequencies estimation. *Gests*, 19(1), 2006.

[20] Alan. V. Oppenheim, Ronald W. Schafer y John R. Buck. *Tratamiento de Señales en Tiempo Discreto*. Pearson Educación,S.A., 2 edition, 2000.

[21] C. Gasquet and P. Witomski. *Fourier Analysis and*

Applications. Springer-Verlag, 1998.

[22] John G. Proakis and Dimitris G. Manolakis. *Digital Signal Processing*. Prentice-Hall International, Inc, 3 edition, 1996.

[23] C. Burrus, F. Matteo, et al *Fast Fourier Transforms*. Connexions. Rice University., 2012.

[24] P. Duhamel and M. Vetterli. Fast fourier transforms: A tutorial review and a state of the art. *Signal Processing*, 19(4):259–299, Abril 1990.

[25] M. J Roberts. *Señales y Sistemas*. McGraw-Hill, 2005.

[26] Pearu Peterson. F2py: a tool for connecting fortran and python programs. *Int. J. Computational Science and Engineering*, 4(4):296–305, 2009.

[27] Christian Duque Sanchez, Mauricio Morales Pérez, et al. Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones. Trabajo Especial de Grado, Universidad Tecnológica de Pereira, Pereira, Colombia, 2007.

[28] J. R. Deller, J. G. Proakis, and J. L. Hansen. *Discrete-Time Processing of Speech Signal*. Prentice Hall, 1987.

[29] F. Rodríguez y T. Loreto. Implementación de un detector de frecuencia fundamental de voz en tiempo real usando la plataforma stm32f407 discovery de stmicroelectronics. Trabajo Especial de Grado, Escuela de IngenierÃa de Telecomunicaciones, Universidad de Carabobo, Venezuela, 2014.

[30] S. A. Fulop. *Speech Spectrum Analysis*. Signals and Communication Technology. Springer, 2011.

[31] F. N. Batalla y C. S. Nieto. *Espectrografía clínica de la voz*. Universidad de Oviedo, 1999.