UNIVERSIDAD DE CARABOBO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE TELECOMUNICACIONES
DEPARTAMENTO DE SEÑALES Y SISTEMAS

DISEÑO DE UNA HERRAMIENTA COMPUTACIONAL QUE PERMITA EL RECONOCIMIENTO DE PERSONAS A TRAVÉS DE LA VOZ APLICANDO «DEEP LEARNING» Y TRANSFORMADA DE WAVELET.

RUDOLF P. ROSANNA C.
LAMAS R. LUCIANO R.

Bárbula, 5 de diciembre del 2016

UNIVERSIDAD DE CARABOBO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE TELECOMUNICACIONES
DEPARTAMENTO DE SEÑALES Y SISTEMAS

DISEÑO DE UNA HERRAMIENTA COMPUTACIONAL QUE PERMITA EL RECONOCIMIENTO DE PERSONAS A TRAVÉS DE LA VOZ APLICANDO «DEEP LEARNING» Y TRANSFORMADA DE WAVELET.

TRABAJO ESPECIAL DE GRADO PRESENTADO ANTE LA ILUSTRE UNIVERSIDAD DE CARABOBO PARA OPTAR AL TÍTULO DE INGENIERO DE TELECOMUNICACIONES

RUDOLF P. ROSANNA C.
LAMAS R. LUCIANO R.

Bárbula, 5 de diciembre del 2016

UNIVERSIDAD DE CARABOBO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE TELECOMUNICACIONES
DEPARTAMENTO DE SEÑALES Y SISTEMAS

CERTIFICADO DE APROBACIÓN

Los abajo firmantes miembros del jurado asignado para evaluar el trabajo especial de grado titulado «DISEÑO DE UNA HERRAMIENTA COMPUTACIONAL QUE PERMITA EL RECONOCIMIENTO DE PERSONAS A TRAVÉS DE LA VOZ APLICANDO «DEEP LEARNING» Y TRANSFORMADA DE WAVELET.», realizado por los bachilleres RUDOLF P. ROSANNA C., cédula de identidad 20.591.860, LAMAS R. LUCIANO R., cédula de identidad 21.466.585, hacemos constar que hemos revisado y aprobado dicho trabajo.

Firma
Prof. ING. ELIMAR HERNANDEZ
TUTOR

Firma
Prof. ING. AHMAD OSMAN.
JURADO

Firma
Prof. ING. PAULINO DEL PINO.
JURADO

Bárbula, 5 de diciembre del 2016

Dedicatoria

Principalmente a Dios por abrirme caminos de bien y de exitos.
A mis padres y hermano, sin ellos no hubiese alcanzado esta meta.
A mis familiares por su apoyo y cariño incondicional.

RUDOLF P. ROSANNA C.

A Dios, por estar presente en cada uno de mis pasos.
A mi familia, por nunca perder la confianza a pesar de mis fallos.
A todos los que a lo largo de este trabajo me tendieron su mano.

LAMAS R. LUCIANO R.

Agradecimientos

En primer lugar, a Dios por haber guiado cada uno de nuestros pasos durante el desarrollo de esta investigación hasta su exitosa culminación.

A nuestra familia, por su apoyo incondicional, comprensión y motivación en cada instante de nuestra experiencia universitaria.

A nuestra tutora académica, Ing. Elimar Hernández, por su confianza, apoyo, compañía y consejos, siendo pilar fundamental en nuestra formación académica y alcance del Trabajo Especial de Grado. Por su paciencia y esfuerzo, respaldando cada decisión y tropiezo que condujeron a la satisfactoria conclusión de esta etapa de nuestra vida académica. Por ser una fuente de inspiración, a ella todo nuestro respeto, admiración y deseos positivos tanto para su vida personal como profesional.

A la Escuela de Telecomunicaciones, en especial al personal docente que intervino en nuestra formación como profesionales íntegros, por todas las experiencias que aportaron a nuestro crecimiento individual.

A todos aquellos cuya participación y colaboración permitió llevar a cabo diferentes etapas de esta investigación.

A todos, Muchas gracias. . . .

Índice general

Índice de Figuras xi

Índice de Tablas xv

Acrónimos xvii

Resumen xix

I. Introducción 1
- 1.1. Motivación . . . 1
- 1.2. Objetivos. . . . 4
  - 1.2.1. Objetivo General. . . . 4
  - 1.2.2. Objetivos Específicos. . . . 4
- 1.3. Alcances. . . . 5

II. Marco conceptual 7
- 2.1. Formantes (F1, F2, F3, F4) . . . 7
- 2.2. Pitch o Frecuencia Fundamental (F0) . . . 9
- 2.3. Shimmer y Jitter . . . 11
  - 2.3.1. Jitter (absolute) . . . 12
  - 2.3.2. Shimmer (dB) . . . 13
- 2.4. Transformada de Wavelet . . . 13
- 2.5. Redes Neuronales . . . 15
- 2.6. Deep Learning . . . 21
- 2.7. Praat . . . 23

III. Procedimientos de la investigación 27
- 3.1. Fase I. Revisión de Referencias para la Selección de Parámetros Característicos . . . 27
- 3.2. Fase II. Diseño del sistema de adquisición de datos . . . 28
- 3.3. Fase III. Desarrollo del Algoritmo del Identificador de personas . . . 29
- 3.4. Fase IV. Evaluación de la herramienta diseñada . . . 33

IV. Análisis, interpretación y presentación de los resultados **35**
4.1. Herramienta computacional para la autentificación por medio de la voz . . . 35
4.1.1. Pantalla Principal . . . 35
4.1.2. Ventana Menú . . . 37
4.1.3. Ventana Nuevo Usuario . . . 38
4.1.3.1. Registro . . . 38
4.1.3.2. Adquisición de Muestra . . . 40
4.1.3.3. Procesar Muestra . . . 42
4.1.3.4. Graficar . . . 46
4.1.4. Validar Usuario . . . 47
4.1.4.1. Datos Personales . . . 49
4.1.4.2. Adquisición de Muestra . . . 49
4.1.4.3. Procesar Muestra . . . 50
4.1.4.4. Graficar . . . 51
4.1.4.5. Visualización del Resultado . . . 51
4.1.5. Editar Registro . . . 54
4.1.6. Eliminar Registro . . . 55
4.2. Análisis de los ensayos realizados. . . . 57

V. Conclusiones y recomendaciones **63**
5.1. Conclusiones . . . 63
5.2. Recomendaciones . . . 64

A. Algoritmo de procesamiento para la Transformada de Wavelet. **67**

B. Algoritmo de procesamiento para la eliminación de silencios. **69**

C. Algoritmo de procesamiento para Deep Learning. **71**

Referencias Bibliográficas **73**

Anexos

A. Manual de Usuario.

B. Tablas de Cálculos de Parámetros de la voz.

C. Glosario de Términos

D. Algoritmo del diseño de la interfaz gráfica de la herramienta.

E. Herramienta Diseñada para el reconocimiento de personas por medio de la voz.

Índice de figuras

2.1. Representación esquemática del primer formante y del segundo formante de las vocales en un espectrograma. *Fuente: Joaquim Llisterri.[10]* 8
2.2. Aparato fonador humano. *Fuente: Bárcena y Alonso. [12]* 10
2.3. Relación directa entre la resolución temporal y la resolución frecuencial al aplicar STFT. *Fuente: Chávez y Camarera-Ibarrola. [19]* 14
2.4. Proceso de Descomposición de la Señal Original. *Fuente: Acosta. [22]* . 16
2.5. Descomposición de una Señal de 120 Hz hasta 4 Niveles, en función de los Coeficientes de Aproximación. *Fuente: Acosta. [22]* 16
2.6. Coeficientes derivados del Análisis por TW sin modificar. *Fuente: Acosta. [22]* 16
2.7. Función de Transferencia para el Criterio de Hard Thresholding y Soft Thresholding. *Fuente: Acosta. [22]* 17
2.8. Ejemplo de una red neuronal totalmente conectada. *Fuente: Matich. [24]* 18
2.9. Procedimiento a seguir por el Sistema de Reconocimiento de Voz. *Fuente:Propia.* 19
2.10. Etapa de Identificación. *Fuente:Rose, P. [10]* 20
2.11. . Etapa de Verificación. *Fuente:Rose, P. [10]* 20
2.12. Fundamento del Prototipo sugerido en la presente investigación. *Fuente:Propia.* 22
2.13. Principio de funcionamiento «*Deep Learning*». *Fuente: Oscar Chang EC. [26]* 24

3.1. Diagrama del procesamiento de la señal de audio. *Fuente: Propia.* . . 32
3.2. Diagrama del funcionamiento de la herramienta. *Fuente: Propia.* . . . 34

4.1. Pantalla principal de la herramienta. *Fuente:Propia* 36
4.2. Mensaje de confirmación de salida de la herramienta. *Fuente:Propia* . 36
4.3. Ventana Menú de la herramienta. *Fuente: Propia* 37
4.4. Recuadro desplegable en la ventana Menú con acceso a las fases de funcionamiento de la herramienta. *Fuente: Propia* 38
4.5. Ventana para registro de nuevo usuario. *Fuente: Propia* 39
4.6. Sección `Registro` de la ventana de nuevo usuario. *Fuente: Propia* . . 39
4.7. Validación del campo `Cédula` para registro de nuevo usuario. *Fuente: Propia* 40

4.8. Validación del campo `Nombre` para registro de nuevo usuario. *Fuente: Propia* . . . 40
4.9. Validación del campo `Apellido` para registro de nuevo usuario. *Fuente: Propia* . . . 41
4.10. Validación del campo `Género` para registro de nuevo usuario. *Fuente: Propia* . . . 41
4.11. Ventana emergente notificando usuario ya registrado. *Fuente: Propia* . 42
4.12. Vetana emergente notificando la opción de `Grabar` para el registro del nuevo usuario. *Fuente: Propia* . . . 42
4.13. Recuadro desplegable en la ventana Menú con acceso a las faces de funcionamiento de la herramienta. *Fuente: Propia* . . . 43
4.14. Ventana emergente notificando inicio de grabación. *Fuente: Propia* . . 43
4.15. Ventana emergente notificando que la grabación ha finalizado. *Fuente: Propia* . . . 43
4.16. Sección `Procesar Muestra` de la ventana de nuevo usuario. *Fuente: Propia* . . . 44
4.17. Ventana de procesamiento de `Praat Object` y `Praat Picture`. *Fuente: Propia* . . . 44
4.18. Recuadro desplegable en la ventana Menú con acceso a las faces de funcionamiento de la herramienta. *Fuente: Propia* . . . 45
4.19. Fichero de extracción de valores formantes en intervalos etiquetados. *Fuente: Propia* . . . 45
4.20. Ventana emergente verificando sobreescritura del fichero. *Fuente: Propia* . . . 45
4.21. Valores de formantes extraídos del procesamiento de cada audio. *Fuente: Propia* . . . 46
4.22. Sección `Graficar` de la ventana de nuevo usuario. *Fuente: Propia* . . 46
4.23. Visualización temporal de la señal de audio original y la procesada con TW y algoritmos de supresión de silencios. *Fuente: Propia* . . . 47
4.24. Ventana emergente notificando registro exitoso del nuevo usuario. *Fuente: Propia* . . . 47
4.25. Ventana emergente verificando ingreso de nuevo usuario. *Fuente: Propia* . . . 48
4.26. Ventana de validar usuario. *Fuente: Propia* . . . 48
4.27. Sección `Datos Personales` de la ventana de validar usuario. *Fuente: Propia* . . . 49
4.28. Ventana emergente notificando que el usuario no se encuentra registrado. *Fuente: Propia* . . . 49
4.29. Adquisición de la muestra para el proceso de validación. *Fuente: Propia* 50
4.30. Opción `Validar` de la ventana `Praat Object`. *Fuente: Propia* . . . 51
4.31. Fichero de extracción de valores formantes en intervalos etiquetados. *Fuente: Propia* . . . 52

4.32. Valores de formantes extraídos del procesamiento de cada audio. *Fuente: Propia* 53
4.33. Sección `Graficar` de la ventana de validar usuario. *Fuente: Propia* . . 53
4.34. Sección `Validar` de la ventana de validación de usuario. *Fuente: Propia* 53
4.35. Campo de validación al reconocer al usuario registrado. *Fuente: Propia* 54
4.36. Campo de validación al no reconocer al usuario registrado. *Fuente: Propia* 54
4.37. Ventana emergente notificando la repetición del proceso de validación. *Fuente: Propia* 54
4.38. Ventana emergente verificando abandono de la ventana de validación. *Fuente: Propia* 54
4.39. Ventana de edición de un usuario ya registrado. *Fuente: Propia* 55
4.40. Ventana de edición, identificando al usuario registrado. *Fuente: Propia* 55
4.41. Ventana nuevo usuario para la opción de edición de la adquisición de muestras. *Fuente: Propia* 56
4.42. Ventana de eliminar usuario. *Fuente: Propia* 57
4.43. Ventana de eliminación de usuario, identificando el usuario registrado. *Fuente: Propia* 57
4.44. Ventana emergente notificando que el usuario fue eliminado. *Fuente: Propia* 57

Indice de tablas

4.1. Porcentaje de aciertos considerando la misma persona para las etapas de entrenamiento y validación 58
4.2. Porcentaje de aciertos de la herramienta considerando distintas personas para las etapas de entrenamiento y validación 59
4.3. Porcentaje de aciertos de la herramienta realizando las etapas de entrenamiento y validación con frases diferentes. 60
4.4. Porcentaje de aciertos de la herramienta considerando los estados de ánimo del usuario para la etapa de validación. 61

Acrónimos

RNA	**R**ed **N**euronal **A**rtificial
ANN	**A**rtificial **N**eural **N**etworks
IA	**I**nteligencia **A**rtificial
RAH	**R**econocimiento **A**utomático del **H**abla
ASR	**A**utomatic **S**peech **R**econigtion
STFT	**S**hort **T**ime **F**ourier **T**ransform
DWT	**D**iscret **W**avelet **T**ransform
RAL	**R**econocimiento **A**utomático del **L**ocutor
cA	**C**oeficiente de **A**proximación
cD	**C**oeficiente de **D**etalle

DISEÑO DE UNA HERRAMIENTA COMPUTACIONAL QUE PERMITA EL RECONOCIMIENTO DE PERSONAS A TRAVÉS DE LA VOZ APLICANDO «DEEP LEARNING» Y TRANSFORMADA DE WAVELET.

por

RUDOLF P. ROSANNA C. y LAMAS R. LUCIANO R.

Presentado en el Departamento de Señales y Sistemas
de la Escuela de Ingeniería en Telecomunicaciones
el 5 de diciembre del 2016 para optar al Título de
Ingeniero de Telecomunicaciones

RESUMEN

Actualmente existen gran número de aplicaciones basadas en características biométricas para comprobar y validar la identidad de la persona, con destacada aceptación y fiabilidad dadas las ventajas que en materia de seguridad y privacidad generan, puesto que no comprometen la información de los usuarios presentes en las operaciones gestionadas mediante estos sistemas. Así mismo, como consecuencia de la masificación a escala mundial que han experimentado los equipos de telefonía móvil, se ha encontrado en ello uno de los medios más idóneos para implementar masivamente un sistema biométrico de bajo costo a través del análisis

de la voz. Pensando en esto, se llevó a cabo un proyecto dedicado al desarrollo de una herramienta computacional con capacidad para adquirir, procesar, visualizar y analizar señales de voz, con base a la implementación de algoritmos de inteligencia artificial basados en «Deep Learning», Transformada de Wavelet y la conformación de una interfaz de usuario, concebida mediante el software MATLAB, cuyo entorno cumplía con todos los requisitos dispuestos. El diseño del identificador se concentró en 2 etapas de funcionamiento, en primer lugar, el registro del locutor en la base de datos del sistema, a partir de la adquisición de muestras vocales, realizándole un tratamiento previo para lograr eliminar silencios y minimizar efectos de ruido mediante la transformada de Wavelet y de esta forma facilitar la extracción de características que sirvan como fuente de aprendizaje para el reconocimiento de patrones. El estudio de estas características se realizó a través del uso del programa Praat, elegido por ser un software gratuito de gran utilidad para los estudios fonéticos del habla. En segundo lugar, la fase de validación, donde se certifica que la persona es quien dice ser, cuya decisión va por cuenta del bloque de procesamiento inteligente basado en redes neuronales implementando técnicas de Deep Learning. Cabe señalar que se incluyó una sección para la representación gráfica de modo que la experiencia para el usuario fuese de mayor atractivo al tener la posibilidad de contrastar la señal de voz original y la derivada del tratamiento interno para su adecuación para los fines acústicos esperados, que en conjunción con el soporte dado por la revisión bibliográfica, consolidaron a los formantes y pitch como los parámetros característicos de la voz de mayor soporte para distinguir a un individuo de cualquier otro. Asimismo, los ensayos arrojaron que la herramienta es capaz de validar con un porcentaje de acierto superior al 90 %.

Palabras Claves: Deep Learning, Herramienta computacional, Transformada de Wavelet, Verificación de voz, Praat.

Tutor: ING. ELIMAR HERNANDEZ
Profesor del Departamento de Señales y Sistemas
Escuela de Telecomunicaciones. Facultad de Ingeniería

Capítulo I

Introducción

1.1. Motivación

El avance de la tecnología ha permitido integrar los sistemas biométricos en diferentes aplicaciones dado que ofrecen servicios y transacciones seguras. Por ejemplo, los sistemas biométricos se utilizan para el control de acceso en edificaciones, aeropuertos y fronteras, así como en transacciones bancarias, dadas las ventajas que en materia de seguridad y privacidad genera, puesto que no compromete la información de los usuarios involucrados en dichas operaciones. En este sentido, para lograr que estos servicios sean seguros, se requiere la autentificación de las personas que hacen uso de los mismos, a través de diferentes métodos que permitan obtener características biométricas como la huella, la retina,la voz, entre otros, mediante el análisis de aspectos físicos o del comportamiento que le definen como individuo, por ser únicas y con un alto nivel de complejidad para falsificar[1].

Ahora bien, para que una característica biométrica sea efectiva en el reconocimiento de personas debe cumplir con propiedades tales como la universalidad, singularidad, invariabilidad, robustez, entre otras. Sin embargo, es difícil encontrar un sistema de reconocimiento que presente todas las características anteriormente señaladas y aún más a medida que se incrementa la cantidad de usuarios a identificar[2]. Actualmente, existen diversas técnicas para el reconocimiento de personas,

pero hasta ahora, ninguna de estas ha logrado ser absolutamente segura, por lo que existe un grado de vulnerabilidad. Es por esto, que el desarrollo de investigaciones en esta área tienen como propósito mejorar las técnicas actuales, para aumentar el nivel de certeza y seguridad en los sistemas que requieran de autentificación de personas.

Hoy en día, existen diferentes dispositivos electrónicos tales como video cámaras, micrófonos, teléfonos celulares, entre otros, que pueden capturar diferentes características biométricas para su uso en procesos de comparación que permitan identificar a una persona. Así mismo, dado que en la actualidad el manejo y disposición de equipos de telefonía móvil se encuentra masificado a escala mundial, es uno de los medios más idóneos para implementar masivamente un sistema biométrico de bajo costo a través del análisis de la voz.

En este sentido, la voz es una característica biométrica con la capacidad de diferenciar a un individuo de otro tal como lo evidencia un estudio en ciencias fonéticas presentado por la Universidad de Glasgow en el 2009[3], donde se determinó que el ser humano es capaz de reconocer a una persona el 99,9 % de las veces con oír tan sólo 2 palabras. Como complemento a ello, otro estudio relevante fue el realizado en el laboratorio de fonética del CSIC referente a cualidad individual de voz e identificación del locutor, donde se comprobó que la voz humana tiene características determinantes que no se logran disimular ni distorsionar, ya que cada persona tiene características acústicas únicas, caso similar al de las huellas dactilares[4].

Es necesario recalcar que en esta investigación se compararan parámetros de la voz con una base de datos previamente establecida, para determinar si el hablante es quien dice ser. Se genera de esta forma un problema de clasificación que puede ser abordado mediante la implementación de algoritmos basados en inteligencia artificial como las redes neuronales, las cuales en los últimos años han evolucionado, cambiando sus técnicas de aprendizaje tradicional por un aprendizaje profundo (Deep Learning), lo cual le ha dado mayor nivel de sensibilidad y capacidad para predecir y clasificar, siendo de esta forma la herramienta idónea para disminuir la cantidad de errores que pueden producirse al comparar parámetros de la voz para identificar a un individuo.

Es así como, uno de los trabajos que sirvió de punto de partida para la presente investigación es el realizado por Cruz, L. y Acevedo, M. denominado «Reconocimiento de Voz usando Redes Neuronales Artificiales Backpropagation y Coeficientes LPC», donde se propone un algoritmo para el reconocimiento de personas en un canal telefónico y se muestra una metodología para la adquisición, procesamiento y comparación de señales de voz[5], siendo esto un referente de utilidad para la presente investigación.

De igual forma, cabe señalar que algunas características de la voz pueden estar sujetas a variaciones por diferentes factores, como el estado de ánimo, la salud, la entonación, entre otros, lo que dificulta la identificación si no se eligen adecuadamente los parámetros que permitan el reconocimiento[1][6], por lo que esto demanda sistemas biométricos más robustos. La identificación de patología en la voces ha sido tratada en diferentes investigaciones, como lo señala el trabajo realizado por Del Pino P. (2008) denominado «Aplicación de la transformada de Wavelet para el análisis de señales de voz normales y patológicas», en el cual se aborda el problema de la determinación de parámetros característicos obtenidos a partir del procesamiento de la señal de voz mediante la transformada de Wavelet[7], lo que guarda relación con el tema en cuestión y proporciona información relevante en la solución de la problemática. Así mismo, el trabajo de grado realizado por Perdomo, Y. (2015) denominado «Desarrollo de software libre interactivo para realizar análisis espectral de voz» da como aporte una herramienta computacional que permite caracterizar el espectro de señales de voz y destaca en forma concisa un marco conceptual de los principales parámetros que caracterizan a la voz humana[8].

Motivado a esto, en el presente trabajo de investigación se realizará el diseño de una herramienta de entrenamiento computacional que sustrae cualidades del habla, la cual servirá como base para futuras investigaciones científicas en el área de seguridad biométrica con la finalidad de mejorar el desempeño de la misma mediante la combinación de la Transformada de Wavelet y Deep Learning. Ésta, al basarse en técnicas de autenticación por voz, permitirá al usuario realizar operaciones a distancia sin necesidad de estar presente en el lugar, ya que el sistema podrá verificar si se trata o no de la persona que dice ser, lo que fortalece la seguridad en

la transacción.Por otra parte, este trabajo podría utilizarse como fuente de consulta para las técnicas de «Deep Learning» y TW, empleadas en otras investigaciones.

Por último, con el desarrollo de esta herramienta se estará innovando en el área de aprendizaje profundo (Deep Learning) dentro de las investigaciones que se han realizado hasta la fecha en la Escuela de Ingeniería de Telecomunicaciones de la Universidad de Carabobo, abriendo camino a la automatización de procesos mediante inteligencia artificial.

1.2. Objetivos.

1.2.1. Objetivo General.

Diseñar una herramienta computacional que permita el reconocimiento de personas a través del análisis de su voz aplicando técnicas de Deep Learning y Transformada de Wavelet.

1.2.2. Objetivos Específicos.

1. Revisar las referencias para la selección de parámetros característicos en las señales de voz que permitan diferenciar una persona de otra.
2. Diseñar un sistema para la adquisición de señales de voz que permita generar una base de datos.
3. Diseñar el algoritmo que permita identificar a una persona a través de la voz empleando «Deep Learning» y Transformada de Wavelet.
4. Evaluar el desempeño de la herramienta diseñada para la identificación de personas.

1.3. Alcances.

En la presente investigación se desarrollará una herramienta computacional que permita la autentificación de personas aplicando Transformada deWavelet y «Deep Learning» a los parámetros característicos de la voz. La herramienta se enfocará en determinar los parámetros que arrojare mayores porcentajes de acierto en la etapa de validación de la misma. Específicamente, para el caso de la TW se seleccionará el tipo de Wavelet madre y el nivel de descomposición. Para la implementación del aprendizaje profundo se elegirá el tipo de red neuronal, el número de capas y el número de neuronas por capas, y de la voz se elegirán aquellos parámetros o características que permitan diferenciar a una persona de otra. Por otro lado, la herramienta contará con una interfaz gráfica de usuario, la cual contribuirá al manejo de la misma en todas sus respectivas etapas de funcionamiento.

Capítulo II

Marco conceptual

2.1. Formantes (F1, F2, F3, F4)

Desde la perspectiva del carácter biológico, se definen como aquellas frecuencias que se enfatizan debido a la concentración de energía en el tracto vocal, provocadas por los órganos articulatorios que actúan como resonadores, los cuales pueden detallarse a partir del uso de un espectrograma, como los picos que se presentan en la envolvente espectral de la señal de voz. Las frecuencias formantes del tracto vocal que aportan mayor información para el análisis acústico de la voz son las primeras cinco (F1, F2, F3, F4 y F5), de cuya relación es que se deriva uno de los principales aspectos de la voz como lo es el timbre o calidad vocal. [9]

Sin embargo, según sea el interés que se tenga en la manipulación e interpretación de una determinada muestra, cada formante aporta una información específica, esto puede observarse en la Figura 2.1, donde se tiene que las dos primeras frecuencias (F1 y F2) son indispensables para la percepción del timbre vocálico, mientras que las restantes intervienen en identificar ciertas características del sistema responsable de la fonación. De tal modo, en las vocales, la primera formante (F1), controla la amplitud del sonido y depende de la cavidad faríngea, mientras más estrecha ésta, mayor frecuencia y viceversa. La segunda formante (F2) controla la claridad del sonido, cualidad sujeta a la posición de la lengua, si la misma se eleva

en la parte anterior, la frecuencia subirá en relación con la altura y la anterioridad alcanzada; por el contrario, si es en la parte posterior, la frecuencia descenderá en relación inversa con la altura de la lengua. [10]

Finalmente, la tercera formante (F3) se relaciona con la acción de los labios, donde su valor en frecuencia es más alto si éstos están estirados y más bajo si están redondeados. Para el caso de F4 y F5 varían con la anchura y longitud del tracto vocal, cuanto más corto y estrecho el tracto, más agudas serán las mencionadas formantes. [11][10]

Desde el punto de vista acústico, al ponerse en vibración las cuerdas vocales producen una onda compuesta. Si los órganos de la articulación no se moviesen, se tendría siempre un único sonido vocal para cada hablante, sin embargo, la articulación de cada vocal requiere de unas determinadas condiciones de los elementos que intervienen, por lo que se originan cavidades de diferentes formas y volúmenes, de modo que se configura una estructura armónica diferente para cada vocal en la que unos determinados armónicos del tono fundamental se realzan mientras que otros se disipan. A ese conjunto de armónicos que son realzados por el fenómeno de resonancia, recibe el nombre de formante.[12]

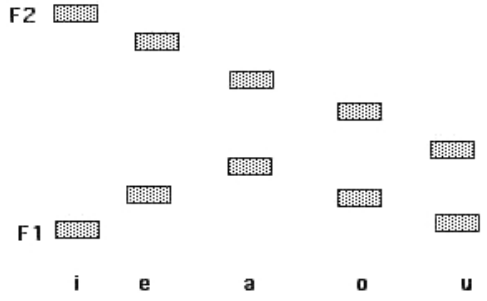

Figura 2.1: Representación esquemática del primer formante y del segundo formante de las vocales en un espectrograma. *Fuente: Joaquim Llisterri.[10]*

En resumen, son tres las etapas que intervienen en la fonación, siendo éstas, un generador de energía, compuesto por las cavidades infraglóticas, un sistema vibrante que representa las cuerdas vocales y un sistema resonante, el cual es responsable de modular el tracto vocal. La estructura biológica descrita anteriormente

se detalla en la Figura 2.2, compuesta en primera instancia por las «Cavidades infraglóticas», conformadas por el Diafragma, Pulmones, Bronquios y Tráquea, es en ellas donde se produce el flujo de aire que posteriormente es modulado, producto del proceso de relajación y contracción del músculo situado por debajo de los pulmones, que no es otro que el diafragma, mientras que los restantes elementos actúan como estructura guía; posteriormente, se encuentra la «Cavidad glótica», compuesta por la Laringe, done una vez que el flujo de aire atraviesa las cuerdas vocales, produce en éstas una vibración variable en frecuencia e intensidad sujeta a 3 factores: masa, longitud y tensión de la glotis.

La vibración se produce únicamente en ciertos tramos de la voz conocidos como segmentos sonoros, en donde la señal adquiere la propiedad de cuasi periodicidad, cuya frecuencia será aquella con la que vibran las cuerdas vocales, parámetro definido como *pitch*. Las frecuencias de resonancia o como comúnmente se les denominan, formantes, son la característica de mayor relevancia, puesto que dependen directamente de la forma y el tamaño del tracto vocal, el cual difiere para cada individuo.

Y finalmente, las «Cavidades supraglóticas», conformada por la Cavidad Nasal, Cavidad Bucal y Faringe, donde ocurre el fenómeno de modulación de la señal de voz.

2.2. Pitch o Frecuencia Fundamental (F0)

Existen numerosas definiciones que se le atribuyen al término pitch, encontrando entre ellas, que representa o corresponde a la frecuencia de abertura y cierre de los pliegues vocálicos. Por otro lado, se puntualiza como la onda sonora simple de frecuencia más baja entre las que conforman una onda sonora compleja. [10]

Para las autoras Gallardo M. Elsa y Asuaje Rosa A., desde la perspectiva del estudio de la producción de voz, la vibración de las cuerdas vocales son el medio fisiológico por el cual se da existencia a la frecuencia fundamental en la señal del habla, por tanto, la misma se refiere entonces a la vibración correspondiente de las

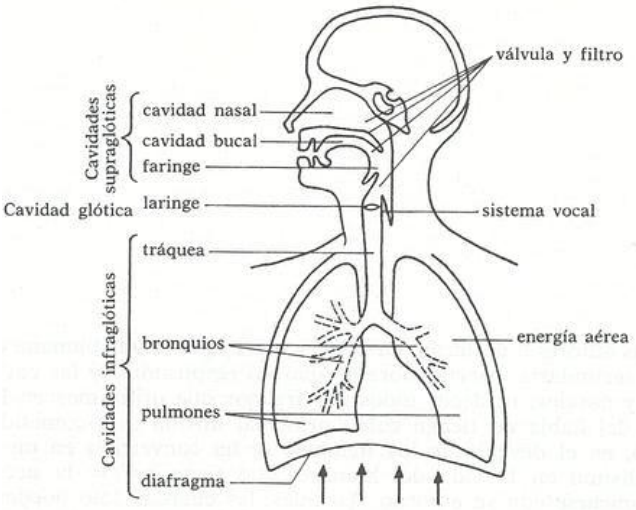

Figura 2.2: Aparato fonador humano. *Fuente: Bárcena y Alonso. [12]*

cuerdas vocales en el momento de la emisión de la voz. De esta cualidad dependen una serie de estudios que permiten el análisis acústico de la voz, como lo es la altura del sonido, donde un sonido alto es aquél con un gran número de vibraciones de las cuerdas vocales, categorizándole así como agudo, mientras que un sonido bajo o grave, por inferencia, es aquél constituido por una baja cantidad de vibraciones. [13]

En función de ello, es posible establecer desde la percepción de la voz, que ésta es percibida como un conjunto de armónicos o frecuencias secundarias, así como variaciones propias de la frecuencia fundamental, responsables en buena parte de la abstracción física del sonido que reconoce el receptor y que posteriormente traduce en una serie de patrones que asocia a un código lingüístico, dando lugar así a un proceso de aprendizaje que conduce a la interpretación del habla y la comunicación. [13]

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. La información que es capaz de proporcionar resulta de interés en la detección de patrones de

vibración de las cuerdas vocales, de modo que sea posible visualizar alguna alteración de las mismas en caso de que se presenten patologías. [12]

El Análisis melódico (Curva Melódica – *Pitch Contour*), es la herramienta de preferencia para apreciar la representación de las variaciones de la frecuencia fundamental de la voz bajo estudio –eje vertical- a lo largo del tiempo –eje horizontal-. La cual está sujeta a una serie de elementos derivados de las características fisiológicas y culturales de cada individuo, entre las que cabe mencionar: Acento, Entonación, Pausas, Velocidad de Ejecución, Ritmo. [10]

2.3. Shimmer y Jitter

El *Jitter* y el *Shimmer* son medidas de perturbación, es decir, se refieren a la variabilidad que se produce ciclo a ciclo en la frecuencia fundamental F0, medida en Hertz (Hz) y en la intensidad medida en dB, respectivamente, de una vocal emitida en forma continua por un sujeto particular. Desde el punto de vista clínico, el *jitter* puede tener lugar cuando las cuerdas vocales sufren alteraciones de su masa, entiéndase situaciones en las que se incrementa o disminuye su volumen. Por otro lado, el *shimmer* se origina a causa de cambios en el tono muscular, producto de desórdenes neurológicos, o por alteraciones aerodinámicas debido a problemas en el tramo bronco-pulmonar. Estas mediciones permiten expresar el grado de estabilidad o inestabilidad del sistema fonatorio durante la producción de voz. [14][15][16]

A nivel matemático, existen diferentes formas de presentar la determinación de los parámetros ya mencionados, según sea el significado físico que le otorgue la relación de variables empleadas durante el proceso, comúnmente suele realizarse de forma relativa. En primera instancia, se tiene el parámetro *shimmer*, cuyo procedimiento se inicia determinando la amplitud pico a pico de la señal en cada período Vpp(i), posteriormente se calcula la variación de la amplitud de cada período con respecto al período anterior en forma porcentual, resultando la siguiente expresión [17][18] :

$$Shimmer(i) = \frac{\mid V_{pp}(i) - V_{pp}(i+1) \mid}{V_{pp}(i)} * 100 \tag{2.1}$$

Por otro lado, para determinar el *jitter* se sigue un procedimiento como el descrito anteriormente, planteando el cálculo de la variación del *pitch* o frecuencia fundamental de una ventana con respecto al *pitch* de la ventana anterior de manera relativa, generando la expresión:

$$Jitter(i) = \frac{\mid pitch(i) - pitch(i+1) \mid}{pitch(i)} * 100 \tag{2.2}$$

Sin embargo, estas no son las únicas formas de expresar dichas características, también puede encontrarse en gran parte de la literatura e investigaciones de índole clínica o científica, sus ecuaciones en forma absoluta.

2.3.1. Jitter (absolute)

Variación ciclo a ciclo de la frecuencia fundamental, la media de la diferencia absoluta entre períodos consecutivos, dado como:

$$Jitter(absoluto) = \frac{1}{(N-1)} \sum_{i=1}^{N-1} \mid T_i - T_{i+1} \mid \tag{2.3}$$

Donde T_i representa las diferentes longitudes periódicas dadas de la frecuencia fundamental y N es el número de períodos de frecuencia fundamental a considerar. [17][18]

2.3.2. Shimmer (dB)

Se expresa como la variabilidad de la amplitud pico a pico en decibelios, es decir, la media absoluta del logaritmo de base 10 de la diferencia entre amplitudes de períodos con secutivos, multiplicado por 20:

$$Shimmer(dB) = \frac{1}{(N-1)} \sum_{i=1}^{N-1} \mid 20log(\frac{A_{i+1}}{A_i}) \mid \tag{2.4}$$

2.4. Transformada de Wavelet

Para llevar a cabo el reconocimiento de voz se requiere de un análisis espectral dinámico, entiéndase con ello, un análisis acústico orientado a extraer las frecuencias que le dan forma a la señal de voz, los formantes, pero en forma análoga determinar en qué instante ocurren, de modo que es indispensable disponer de información tanto del dominio temporal como del dominio frecuencial. Con anterioridad, la herramienta predilecta para ello ha sido la Transformada de Fourier de Tiempo Corto (STFT), el cual consiste en fraccionar la señal de voz en segmentos de corta duración denominados marcos, a la cual se le aplica un enventanado de tal forma que sea posible sólo trabajar con el tramo de interés y la restante señal a los extremos del marco sea suprimida, sin embargo, su aplicación tiene como desventaja el compromiso entre la resolución de frecuencia y la resolución temporal, puesto que si alguna se aumenta, la otra por acción inversa se reduce, tal como se logra apreciar en la Figura 2.3.

Sin embargo, no se requiere de la misma resolución temporal para todas las bandas de frecuencias, para el caso de las frecuencias bajas, dado que experimentan cambios lentos no se requiere de mucho detalle relativo a los cambios temporales, caso contrario a las frecuencias altas por lo que se necesita de una resolución en tiempo mucho mayor. En función a este principio, es que la Teoría de Wavelets, da gran aporte para el análisis de las señales de voz, debido a que los wavelets¸ tienen un mejor desempeño donde las señales experimentan alteraciones bruscas, dado

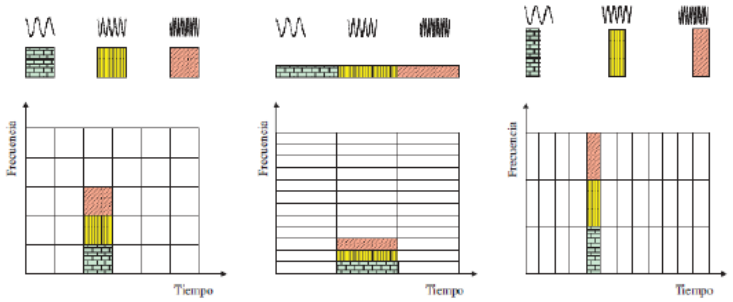

Figura 2.3: Relación directa entre la resolución temporal y la resolución frecuencial al aplicar STFT. *Fuente: Chávez y Camarera-Ibarrola. [19]*

que éstos son más difíciles de expresar con senoides de duración infinita, en consecuencia, para el caso de estudio definido en el presente trabajo de investigación, ante la manipulación de señales no estacionarias como la voz, la «Transformada Discreta de Wavelet» le caracteriza mejor. [19]

La DWT no requiere de segmentar la señal en marcos de tiempo ni aplicación de enventanados, sus características proporcionan información simultánea del dominio del tiempo y de la frecuencia, la cual a diferencia del análisis de Fourier, descompone la señal en una serie de versiones escaladas y desplazadas de la onda referencia, teniendo en forma matemática la siguiente expresión [19]

$$f(t) = \sum_{j,k} b_{j,k} w_{j,k}(t) \tag{2.5}$$

Donde $b_{j,k}$ es el coeficiente que pondera a la función base $w_{j,k}(t)$, que no es otra que el wavelet en la escala j y desplazamiento k, de tal modo que:

$$w_{j,k}(t) = w(2^j t - k) \tag{2.6}$$

El objetivo de aplicar esta herramienta matemática para suprimir el ruido pre-

sente en los datos de la señal de voz grabada y almacenada es el de obtener un proceso lo más eficiente posible preservando las características fundamentales de la voz. La aplicación de la Transformada de Wavelet en diferentes campos de estudio ha permitido concluir respecto a su superioridad respecto a las técnicas tradicionales de tratamiento de ruido como los comparadores o uso de filtros pasa bajos, teniendo como ventaja sobre estos últimos que permite realizar un análisis espectral sin restricciones, al tratar las componentes de la señal con frecuencias dentro de la banda de corte, las cuales podrían tener un significado físico relevante para el estudio, generando así un método de selección de muestras de mayor robustez. Sin embargo ésta no es la única característica a destacar, en especial para el manejo de audio, dada su adaptabilidad para el tiempo-frecuencia, logrando la representación y descripción de fenómenos transitorios y no estacionarios, para la posterior clasificación y detección de propiedades.[20][21][22]

La técnica más conocida se denomina umbral de eliminación de ruido wavelet «Wavelet Thresholding Denoising», cuyo desempeño consiste en la selección de una Wavelet Madre, función referencia para generar la descomposición y extracción de características o coeficientes de la señal de interés, lo cual se realiza en dos etapas, una con un filtro desvanecedor (pasa-bajas) y otro con un filtro de detalles (pasa-altas), de los cuales se derivan los Coeficientes de Aproximación (cA) y Coeficientes de Detalle (cD) respectivamente, tal como se observa en la Figura 2.4 y Figura 2.5. Para cada nivel de tratamiento se aplica el método de umbral seleccionado, en función del cual cada componente se somete a un criterio de evaluación para su consideración como elemento de ruido o información útil, a través de una función de transferencia particular, como se presenta en las Figuras 2.6 y 2.7. [23]

2.5. Redes Neuronales

Las Redes Neuronales Artificiales, por sus siglas en inglés ANN (Artificial Neural Networks), están inspiradas en las redes neuronales biológicas del cerebro humano. Las RNA más allá de la similitud en cuanto a estructura con el sistema nervioso, presentan una serie de cualidades propias del procesamiento cognitivo, da-

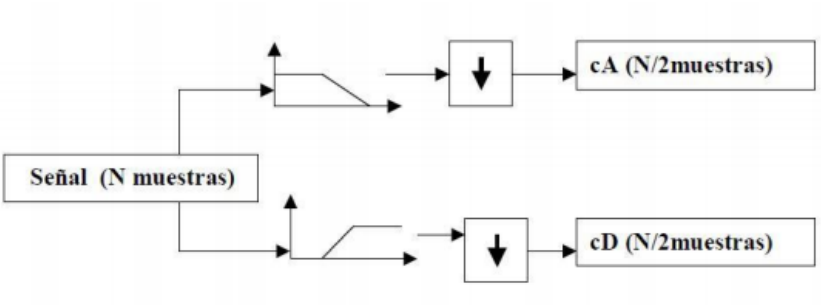

Figura 2.4: Proceso de Descomposición de la Señal Original. *Fuente: Acosta. [22]*

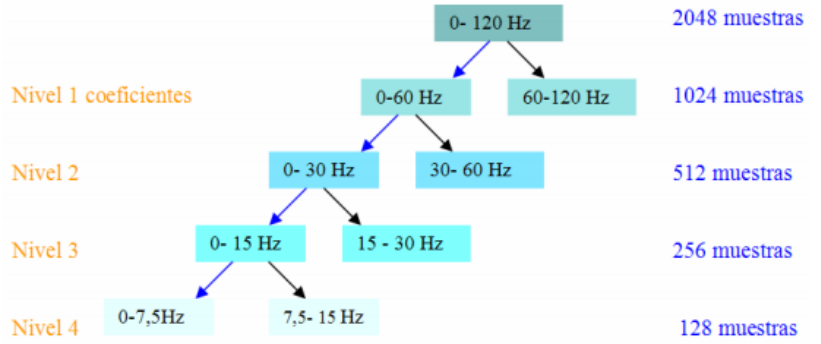

Figura 2.5: Descomposición de una Señal de 120 Hz hasta 4 Niveles, en función de los Coeficientes de Aproximación. *Fuente: Acosta. [22]*

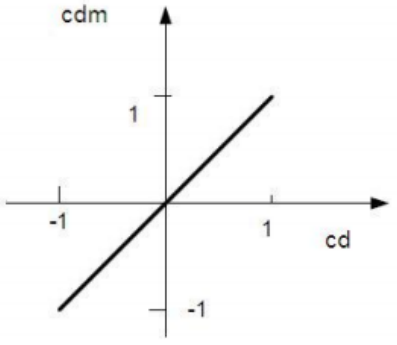

Figura 2.6: Coeficientes derivados del Análisis por TW sin modificar. *Fuente: Acosta. [22]*

do que aprenden de la experiencia, generalizan de ejemplos previos a ejemplos nuevos y extraen e interpretan las características principales de una serie de datos,

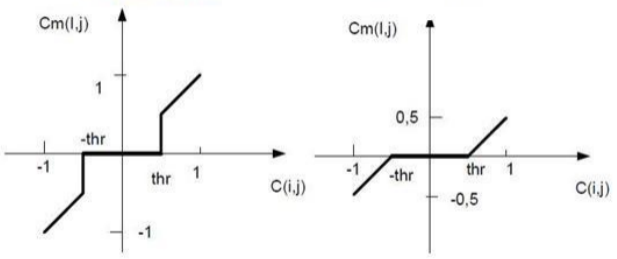

Figura 2.7: Función de Transferencia para el Criterio de Hard Thresholding y Soft Thresholding. *Fuente: Acosta. [22]*

resumiendo así su funcionamiento en 3 experiencias [24]

- *Aprender*, proceso de adquirir conocimiento de un elemento o situación del entorno por medio del estudio y ejercicio, siendo capaz de cambiar su comportamiento en función de las cualidades del medio, se les muestra un conjunto de entradas y por sí mismas se ajustan para producir unas salidas consistentes.
- *Generalizar*, por naturaleza las redes son capaces de ofrecer, dentro de un margen de tolerancia, respuestas correctas a entradas que presenten pequeñas variaciones debido a los efectos propios de agentes externos que distorsionen la toma clara de los datos.
- *Abstraer*, proceso de interpretación y clasificación por separado de las cualidades de un elemento, de modo que los aspectos comunes o relativos tengan una prioridad dentro de su aporte para el cumplimiento de la tarea en particular.

El objetivo del entrenamiento de una RNA radica en lograr que una determinada aplicación, para un conjunto de entradas produzca una serie de salidas deseadas o con el mayor margen de consistencia que sea posible alcanzar del algoritmo empleado. Dicho proceso de aprendizaje consiste en la aplicación secuencial de diferentes vectores de entrada que permitan el ajuste gradual de los pesos de las interconexiones según sea el procedimiento predeterminado, de modo que tras cada

sesión de entrenamiento, los mencionados pesos converjan hacia los valores que hacen que cada entrada genere las respuestas esperadas.[24]

La distribución de neuronas dentro de la red se realiza formando niveles o capas, con un número determinado de dichos elementos en cada una de ellas, esto puede observarse en la Figura 2.8. A partir de su posicionamiento dentro de la red, es posible distinguir 3 tipos de capas [25]

- *De entrada*, es la capa responsable de recibir directamente la información proveniente de las fuentes externas de la red.
- *Ocultas*, son internas a la red y no tienen contacto directo con el entorno exterior. El número de niveles ocultos es variable, del mismo modo en que las neuronas de cada capa oculta pueden interconectadas de distinta manera, lo cual en conjunto con su cantidad, define las diferentes topologías de redes neuronales.
- *De salida*, son las responsables de transferir la información derivada de la red al exterior.

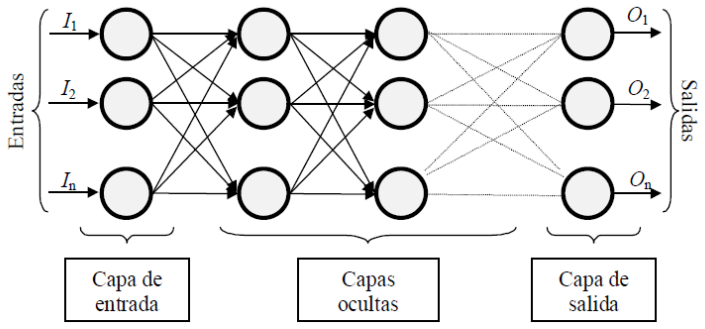

Figura 2.8: Ejemplo de una red neuronal totalmente conectada. *Fuente: Matich. [24]*

Una de las aplicaciones basadas en el uso de Redes Neuronales para el procesamiento de datos, radica en el Reconocimiento Automático del Habla (RAH) o por sus siglas en inglés, ASR (Automatic Speech Recognition), cuyo principal objetivo es la obtención de una representación simbólica discreta de una señal vocal continua, base de la propuesta a desarrollar, apoyada en 2 fases de preparación.

- ***Entrenamiento o aprendizaje.*** Muestras de voz previamente seleccionadas.
- ***Reconocimiento.*** Utilización de los datos adquiridos durante el aprendizaje para obtener una representación discreta a partir de una nueva señal vocal.

En la Figura 2.9, se describe el procedimiento general de un sistema de reconocimiento de voz, el cual consta de cuatro etapas: adquisición de datos, extracción de características, creación del modelo de referencia y la decisión. Por consiguiente, a partir de cada una de estas etapas es posible el reconocimiento de la voz.

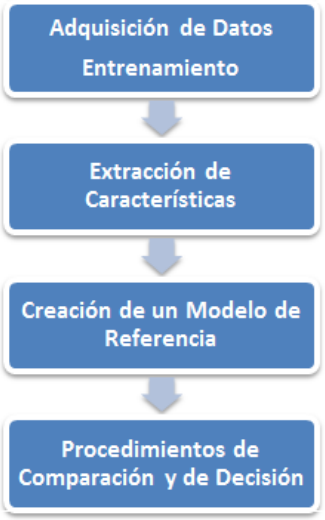

Figura 2.9: Procedimiento a seguir por el Sistema de Reconocimiento de Voz. *Fuente:Propia.*

Para ello, se plantean dos etapas para la estructuración de la herramienta virtual:

- ***Identificación del locutor.*** Comparación entre la muestra de habla de un locutor desconocido y muestras de hablantes conocidos. Objetivo: Determinar si alguna de las muestras de hablantes conocidos previamente proviene del locutor desconocido. Lo cual constituye la etapa de entrenamiento de la red a partir del uso de *Deep Learning*. [Remitirse a Figura 2.10].

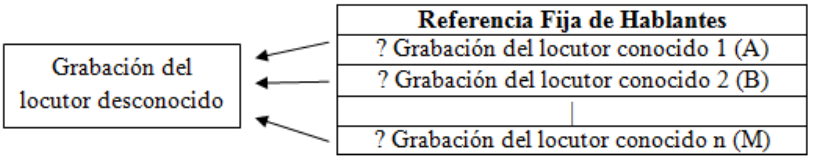

Figura 2.10: Etapa de Identificación. *Fuente:Rose, P. [10]*

- ***Verificación del locutor.*** Comparación entre la muestra de habla de un locutor que dice ser A y las muestras de habla de un conjunto de locutores entre los cuales se encuentra A. Objetivo: Determinar si quien dice ser A es realmente A.[Remitirse a Figura 2.11].

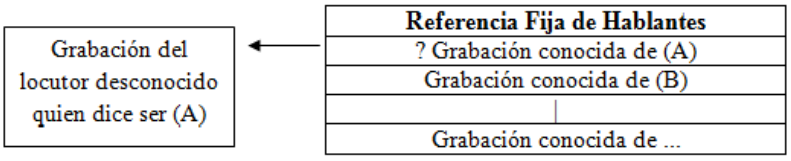

Figura 2.11: . Etapa de Verificación. *Fuente:Rose, P. [10]*

Los sistemas de verificación del locutor pueden clasificarse en función del experimento por el cual se opte para realizar las etapas de entrenamiento y reconocimiento, entiéndase por ello, un sistema dependiente de una pronunciación fonética determinada y un sistema independiente. Para el primer caso, el texto con el que se realiza la inscripción del individuo en la data debe ser el mismo que se pronuncie en la validación, donde para la etapa de entrenamiento un modelo para capturar la muestra pueden ser letras, palabras, frases o textos más elaborados y extensos, por el contrario de los independientes, que no están sujetos a dicha condición puesto que son capaces de realizar la verificación a partir de una tramo de audio desconocido. Sin embargo, para estos últimos la tasa de error es notablemente superior y precisan de tiempos más elevados de entrenamiento y prueba. [12]

Ahora bien, una herramienta del tipo RAL como ya hemos visto, puede tener

dos posibles salidas: válido o inválido. Por tanto, es posible definir dos tipos de errores a partir de los resultados esperados.

- *Falsa Validez.* Aquella situación donde un impostor se hace pasar por otra persona y el sistema responde al engaño aceptando su identidad.
- *Falsa Invalidez.* Prueba donde un usuario auténtico, es decir, que es quien dice ser, es rechazado por el sistema.

Por tal motivo, resulta indispensable validar la calidad del prototipo, teniendo como principales recomendaciones generar dos tipos de pruebas, por un lado un test auténtico donde la fase de entrenamiento y validación la lleva a cabo el mismo locutor, mientras que la restante estará orientada a someter a prueba el modelo de un usuario previamente registrado ante posibles muestras de impostores, para con ello establecer un fundamento estadístico que consolide el desarrollo de la aplicación, incluyendo en ello, las consideraciones propias del entorno, al recurrir a dispositivos móviles o micrófonos para la captura de muestras. [9]

En general, se puede afirmar que un sistema de reconocimiento del locutor posee mayor robustez y precisión según aumente la información de entrada, entiéndase por ello, la longitud del texto, mientras que es mucho más vulnerable a medida que se acorta el contenido fonético del audio suministrado, más ambos modelos tienen sus propias consideraciones y desventajas, puesto que aunque cierto es, que mientras más grande sea la muestra empleada se puede obtener una referencia más solida de su análisis acústico, este proceso implica un tiempo de procesamiento más extenso, y lo que se pretende al desarrollar dichas herramientas, es que su respuesta sea lo más cercana al tiempo real, aspecto de relevancia al enmarcar el algoritmo y fundamentos teóricos del software. [9]

2.6. Deep Learning

La Inteligencia Artificial (IA) se ha desarrollado en función del estudio del comportamiento de los procesos cognitivos humanos, con el propósito de comprender

Figura 2.12: Fundamento del Prototipo sugerido en la presente investigación. *Fuente:Propia.*

el proceso mediante el cual se origina el conocimiento, proveniente de las conexiones entre las redes neuronales que extraen patrones de la realidad y que los clasifica con el fin de generar una reacción cuando se presenten condiciones similares, de este modo, el hombre es capaz de aprender y reproducir posteriormente un idioma, reconocer grupos de objetos, comportamientos en otras personas y hasta asociar una determinada cualidad a la identidad de otro individuo, sea voz, olor, movimientos de pisadas.[26]

Programar una máquina o sistema operativo con algoritmos que funcionen de tal forma que asemejen el comportamiento de las redes neuronales, es lo que se conoce como *Machine Learning* o Aprendizaje Automático. Sin embargo, para que cualquier dispositivo pueda decodificar, separar y analizar grandes cantidades de datos, se requiere que un humano le adiestre a partir de patrones básicos desde los cuales pueda desarrollar una base de conocimiento. Y es allí donde se presenta el nuevo reto para las actuales generaciones, ser capaces de que los dispositivos tengan la cualidad de reconocer patrones con los cientos de posibles combinaciones en que puedan presentarse, a la vez que su capacidad de abstracción por sí sola se

refuerce, sin la presencia de un ser humano, es a esto a lo que se conoce como *Deep Learning*. [27]

Aprendizaje Profundo es el término mediante el cual se hace referencia al uso de un conjunto de algoritmos para realizar abstracciones jerárquicas de alto nivel de información y facilitar el aprendizaje automático (*machine learning*), que le permita a un dispositivo a partir de dichos patrones de datos realizar tareas de mayor envergadura en cuanto a percepción e interpretación, como reconocer el habla, el movimiento, una señal o una imagen. Consecuente con lo anterior, según sea la intención del sistema a capacitar y según sea la arquitectura de Redes Neuronales Artificiales (RNA) a requerir, los algoritmos de entrenamiento se pueden caracterizar principalmente en dos categorías[26] :

- ***Supervisado.*** El entrenamiento es controlado por un agente externo, el cual cumple una función de «guía» a lo largo del proceso de adecuación de la red mediante la comparación continua entre las salidas deseadas y las que se derivan del procesamiento del sistema, generando un comportamiento similar al ensayo y error, tomando el error o diferencia resultante para realimentar la red.
- ***No supervisado.*** El aprendizaje es realizado presentándole a la red los datos directamente, es decir, ya no existe un agente que supervise el entrenamiento, de modo que la red aprende los datos de entrada a medida que va modificando los pesos en función de los datos caracterizados, obteniendo así por cuenta propia sus conclusiones.

2.7. Praat

En la actualidad, el campo de la lingüística abarca gran parte de los desarrollos tecnológicos que involucran el aprendizaje y reconocimiento de voz, por lo que el habla grabada representa el punto de partida para el abordaje de la investigación, de modo que en un trabajo descriptivo, el uso de oscilogramas a fin de ilustrar un

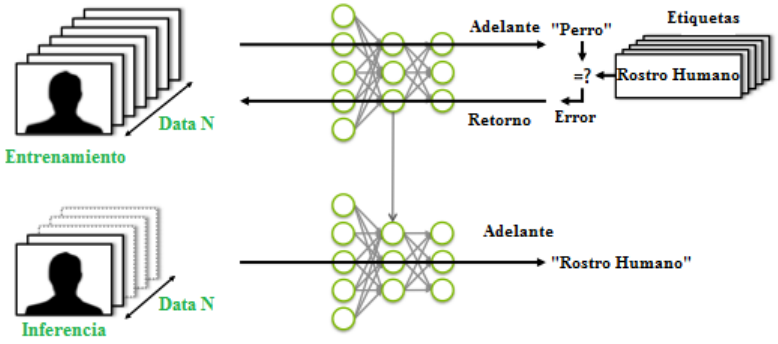

Figura 2.13: Principio de funcionamiento «*Deep Learning*». *Fuente: Oscar Chang EC. [26]*

fenómeno o postular una hipótesis particular en cuanto a la naturaleza de alguna propiedad fonética de la lengua, es probablemente la herramienta más acertada para describir, diferenciar, interpretar y reconocer patrones vinculados a las condiciones acústicas de cada individuo, sujetas claro está, a sus características biológicas e influencia del lenguaje. Sin embargo, por los postulados de Lindblom, B., la voz no es un parámetro estático, todo lo contrario, el hablante se encuentra en constante adaptación según las necesidades que a su juicio le demanden situaciones específicas, por tanto es susceptible a la variación fonética, justificado en los diferentes estilos del habla, propuestos por Labov, W. [10]

Es por ello, que para efectos del análisis acústico, algún parámetro del discurso grabado, como ritmo o la intensidad, es a menudo alterado, de modo que el nuevo sonido obtenido describe diferentes condiciones y resultados, por lo que representa un estudio de continuo aprendizaje y de grandes exigencias en cuanto a herramientas para disponer de los datos, y es allí donde la introducción de la computadora ha generado una revolución virtual en las ciencias del lenguaje respecto al uso de grabaciones de voz, donde la combinación de la arquitectura y los algoritmos de simulación del pensamiento humano, han facilitado el proceso de grabar, anotar y modificar discursos con el uso de unos pocos comandos simples, destacando en particular el software Praat, base del presente trabajo de investigación. [28]

Praat es una herramienta informática gratuita para analizar, sintetizar y mani-

pular el habla, desarrollado desde 1992 por Paul Boersma y David Weenink en el Instituto de Ciencias Fonéticas de la Universidad de Ámsterdam. De gran utilidad para una amplia gama de estudios fonéticos, pues posibilita la observación de características de los parámetros de emisión de la voz, favoreciendo así el análisis acústico, la síntesis articulatoria, el procesamiento estadístico de datos, edición de señales de audio, entre otras tantas por destacar. Una vez obtenida la entrada de audio de interés, es posible manipular una serie de gráficos asociados a ella, como el Oscilograma, donde se hace visible la representación del sonido o forma de onda, Espectrograma, que no es más que la representación de la cantidad de frecuencias altas y bajas disponibles en la señal; el Contorno de Tono (la frecuencia de periodicidad) el cual se asocia a la vibración de las cuerdas vocales, y los Contornos de los Formantes, principales constituyentes del espectrograma, vinculados a las resonancias del tracto vocal. Laos cuales serán los principales instrumentos para la extracción de características de las diferentes muestras dadas por los locutores seleccionados, como fuente de entrada de la Red Neuronal a entrenar. [29][30]

Capítulo III

Procedimientos de la investigación

3.1. Fase I. Revisión de Referencias para la Selección de Parámetros Característicos

Para llevar a cabo el análisis acústico que valide el reconocimiento de una persona sobre otra a partir de su voz, se requiere establecer el conjunto de características que aportan mayor información de cada una de ellas. Por esta razón, la investigación partió de la revisión y la selección de los parámetros de las señales de voz más relevantes para el fin mencionado previamente.

Para ello, se realizó un estudio del proceso de fonación humano, observando a nivel estadístico en la mayor parte de las referencias consultadas, que los autores convergen respecto a la frecuencia fundamental o *pitch* y los formantes como los principales parámetros con mayor capacidad para identificar a un individuo, cuya determinación en el plano espectral es de gran utilidad para distinguir patrones en el habla producto de las características biológicas y sociales que caracterizan a cada sujeto [12][8].

Ahora bien, con el propósito de establecer una valoración concreta de las características que se presentan durante la producción del habla, se toman como punto de referencia los formantes, indispensables en la conformación y percepción de las

vocales, mientras que la frecuencia fundamental o *pitch* se refiere entonces a la vibración correspondiente de las cuerdas vocales en el momento de la emisión de la voz, cualidades espectrales relevantes para la comparación entre individuos, como se describe en el Capítulo II de este documento[9].

En resumen, la fase se fundamentó en la recopilación y consideración bibliográfica relacionada con el tratamiento de señales de voz para la extracción de parámetros característicos que intervienen en la comparación de los patrones de voz propios de cada individuo, planteando bajo un orden de prioridad aquellos rasgos espectrales a evaluar durante su procesamiento.

3.2. Fase II. Diseño del sistema de adquisición de datos

En esta segunda etapa, se realizó el diseño de la interfaz que permite la captura y almacenamiento de la señal de voz asociada a un usuario, para su posterior manipulación y caracterización, cuyo proceso está presente en 2 etapas del funcionamiento general del sistema, una primera etapa definida como inscripción o registro del usuario, donde inicialmente se hace la captura de la voz de la persona, a fin de conformar la base de datos con la cual se entrenará la red «Deep Learning», y en una segunda etapa, denominada validación, donde se obtiene la señal de voz y se verifica si esta corresponde al usuario que se está identificando.

Para comenzar el desarrollo de esta fase se hizo una revisión de los software disponibles que permitieran desarrollar la herramienta computacional, para ello se hizo un estudio haciendo énfasis en su capacidad para adquirir, procesar, visualizar y analizar señales de voz y la existencia de librerías o toolbox que permitieran la implementación de algoritmos de inteligencia artificial basados en «Deep Learning», Transformada de Wavelet y construcción de interfaz de usuarios. Dentro de los software evaluados estuvieron Python, Java, TensorFlow, GNU radio y Matlab. El software elegido fue Matlab, considerando que el mismo cumplía con todos los requisitos necesarios para el desarrollo de la herramienta.

Una vez definido el software con el que se diseñaría la herramienta se definieron los datos a registrar de cada usuario, seleccionando los siguientes: cédula de identidad, nombre, apellido, género y captura de su voz pronunciando una palabra o frase en específico. Para la adquisición de la voz se definieron las cuatro etapas mostradas en el diagrama a continuación, comenzando con la etapa de grabación, donde se decidió grabar al usuario diciendo una frase única, en la que se pudiera medir los parámetros característicos de su voz, específicamente se decidió grabar al usuario diciendo su cédula de identidad. Seguidamente se consideró una etapa de reproducción que permitiera escuchar el audio registrado, para que en caso de que el audio grabado presentara algún problema se pudiese detectar.

Para la realización del proyecto se utilizan archivos de audio o voz en formato WAV ya que es uno de los más utilizados para almacenar sonidos y es un formato sin ningún tipo de compresión de datos y con cuantificación uniforme [31]. Finalmente, una vez capturado de forma satisfactoria la voz del usuario se consideró una etapa para intentar eliminar el ruido presente en el audio grabado.

Hay dos factores importantes que se consideraron durante el proceso de captura: Primero la taza de muestreo, y segundo el número de canales (mono o estéreo). Específicamente, para las aplicaciones de reconocimiento de voz un canal mono es suficiente y dado a que el habla es relativamente de bajas frecuencias (entre 100Hz - 8kHz) [9], se considera que una frecuencia de muestreo de 16kHz es suficiente para lograr capturar la señal de voz sin perder información, sin embargo, se fijó una frecuencia de muestreo de 44100 Hz ya que la mayoría de los equipos actuales de captura de voz trabajan por defecto a esta frecuencia.

3.3. Fase III. Desarrollo del Algoritmo del Identificador de personas

La primera etapa a desarrollar en esta fase es el algoritmo para disminuir el ruido presente en el audio grabado, para ello se decidió hacer uso de la Transformada

De Wavelet mediante la función 'wden' de MATLAB. El formato de esta función es el siguiente:

$$XD = wden(X, TPTR, SORH, SCAL, N, 'wname') \tag{3.1}$$

Donde:

XD es la versión con ruido disminuido de la señal de entrada X.

En el campo **TPTR** se coloca la regla para la selección del umbral; las opciones son: 'minimaxi', 'rigrsure', 'heursure' y 'sqtwolog'.

En **SORH** se escoge el tipo de la función de transferencia a utilizar para el proceso de thresholding; las opciones son 's' que significa soft-thresholding y 'h' que significa hard-thresholding.

En **SCAL** se coloca el nombre de reescala para el umbral; las opciones son: 'one', para que sea sin reescala, 'sln', para reescala usando una sola estimación de nivel de ruido basada en el primer nivel de coeficientes y 'mln' para reescala usando una estimación de nivel de ruido dependiente del mismo.

En **N** se elige el nivel de descomposición de la señal de entrada X para la eliminación del ruido.

Por último, en **'wname'** se coloca el nombre del filtro Wavelet ortogonal a usar. Las opciones son las familias Wavelet: 'Daubechies', 'Coiflets', 'Symlets', 'Discrete Meyer', 'Biorthogonal' y 'Reverse Biorthogonal' junto con sus respectivas variaciones.

A la señal capturada se le aplicó adicionalmente una etapa de eliminación de silencios, para ello, se realiza el cálculo de la energía de la señal en segmentos cortos de 10ms, cuando la energía promedio determinada es menor que el valor umbral fijado, este segmento de señal es eliminado. Las fórmulas utilizadas fueron:

$$En = \sum_{k=1}^{W_n} | X[k] |^2 \, W[n-k] \tag{3.2}$$

$$E_{avg} = \frac{1}{N} \sum_{k=1}^{N} | x[k] |^2 \tag{3.3}$$

El proceso que sigue en la primera parte de esta fase es el siguiente:

1. Descomponer la señal ruidosa hasta un nivel deseado N.
2. Para cada uno de los niveles, seleccionar un umbral y aplicar un algoritmo de Thresholding o también llamado umbralización.
3. Reconstruir la señal.
4. Eliminación de silencios.

Para la elección de los parámetros definitivos de la función wden, fue necesario hacer ensayos variando cada uno de ellos hasta lograr limpiar la señal sin distorsionar significativamente el audio capturado de tal manera de no afectar los parámetros característicos de la voz. De esta forma, a medida que se variaban los parámetros de la función wden se determinaban los parámetros de la voz, para ello se utilizó el programa Praat ya que es un software gratuito de gran utilidad para los estudios fonéticos del habla, que permite la determinación, observación y análisis de los parámetros característicos de la voz. Los parámetros de la voz evaluados fueron: Los cuatro primeros formantes, el pitch, la intensidad, los pulsos, el jitter, el shimmer, el espectro y el contenido armónico.

Por otro lado, para el desarrollo del algoritmo de verificación se comenzó definiendo el modelo de red neuronal bajo la cual se implementaría Deep Learning, para ello se hicieron ensayos considerando los modelos más utilizados [32]. Junto con la elección del modelo se determinó la arquitectura de la red a implementar. De esta forma el procedimiento realizado para definir el algoritmo fue:

1. Definición del modelo de red neuronal bajo el cual se implementaría Deep Learning. Los modelos ensayados fueron: Restricted Boltzmann machines (RBM), los autoencoders (AEN), las Redes Neuronales Convolucionales (CNN), Deep Belief Network (DBN).

2. Definición de la arquitectura de la red, lo cual implicó definir:

   a) Número de capas.

   b) Número de neuronas por capa.

   c) Tipo de aprendizaje.

   d) Función de transferencia.

   e) Algoritmo de entrenamiento.

Durante el desarrollo del algoritmo se realizaron diferentes ensayos considerando variaciones en los parámetros de la transformada de Wavelet TW y de la red Deep Learning hasta conseguir resultados satisfactorios, para ello fue necesario evaluar el clasificador de forma que al cambiar un parámetro por ejemplo de la Transformada Wavelet o de la red se pudiera observar su efecto sobre el rendimiento de éste.

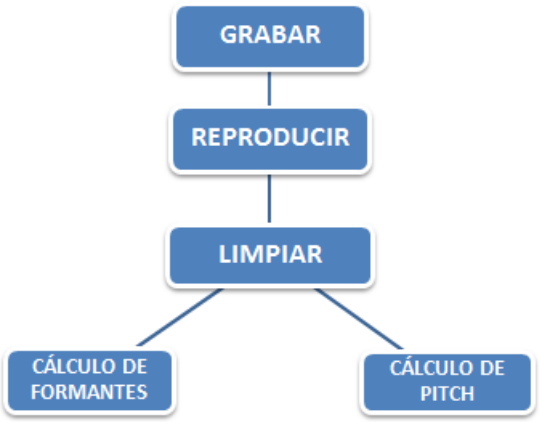

Figura 3.1: Diagrama del procesamiento de la señal de audio. *Fuente: Propia.*

3.4. Fase IV. Evaluación de la herramienta diseñada

En esta última etapa se evaluó el desempeño de la herramienta computacional diseñada. Para ello se evaluó la tasa de éxito o porcentaje de acierto determinada al comparar las predicciones realizadas por la herramienta con el resultado esperado. Esta fue una de las variables más importante a considerar para la selección de los parámetros definitivos que se utilizarían para implementar la transformada de Wavelet y de la Red Deep Learning de la herramienta diseñada.

En Definitiva para la evaluación de la herramienta diseñada se estimó el porcentaje de acierto considerando los siguientes escenarios:

- Caso 1. Evaluación de la herramienta diseñada considerando las etapas de entrenamiento y validación realizadas por la misma persona.
    - Hombres verificando su identidad.
    - Mujeres verificando su identidad.
    - Usuarios verificando su identidad con ruido extra de fondo.
- Caso 2. Evaluación de la herramienta diseñada considerando las etapas de entrenamiento y validación realizadas por personas diferentes.
    - Hombres verificando la identidad de otros Hombres.
    - Hombres verificando la identidad de Mujeres.
    - Mujeres verificando la identidad de otras mujeres.
    - Mujeres verificando la identidad de hombres.
    - Usuarios verificando la identidad de otros usuarios con voces similares.
- Caso 3. Evaluación de la herramienta diseñada considerando la etapa de validación realizada con frases diferentes a la cédula del usuario registrado.
    - Hombres verificando su identidad alternando los dígitos de su número de cédula.

- Mujeres verificando su identidad alternando los dígitos de su número de cédula.
- Usuarios verificando la identidad de otros usuarios con cualquier frase.

- Caso 4. Evaluación de la herramienta diseñada considerando la etapa de validación con distintos estados de ánimo del usuario.
  - Hombres verificando su identidad.
  - Mujeres verificando su identidad.
  - Usuario verificando la identidad de otros usuarios.

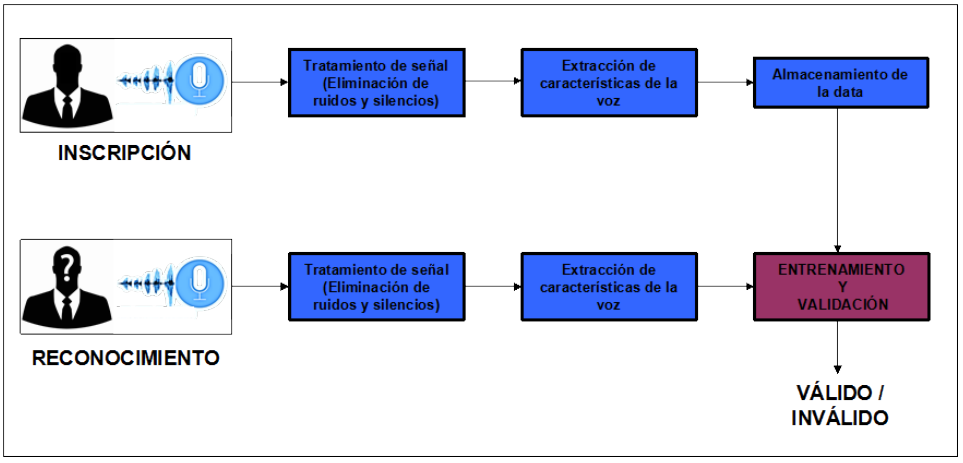

Figura 3.2: Diagrama del funcionamiento de la herramienta. *Fuente: Propia.*

Capítulo IV

Análisis, interpretación y presentación de los resultados

4.1. Herramienta computacional para la autentificación por medio de la voz

La realización del software se fundamentó en el desarrollo de una Interfaz Gráfica de Usuario para la verificación del locutor bajo algoritmos de procesamiento inteligente, respectivamente fase II y fase III del trabajo de investigación, en una versión portable y ejecutable a través de la herramienta «MATLAB» versión R2016a ó R2015b, cuyo funcionamiento se estructuró en seis ventanas, definidas de la siguiente forma:

4.1.1. Pantalla Principal

La aplicación se inicia con una pantalla de bienvenida donde se presenta el logo y nombre respectivos de la herramienta, en la cual se encuentran habilitados los botones de ingreso y cierre de la misma, así como los respectivos íconos para cerrar y minimizar la ventana. La imagen de la ventana principal se observa en la Figura 4.1.

Figura 4.1: Pantalla principal de la herramienta. *Fuente:Propia*

Al presionar el botón `Salir`, se despliega una ventana de diálogo que permite confirmar la salida de la aplicación, dada en la Figura 4.2.

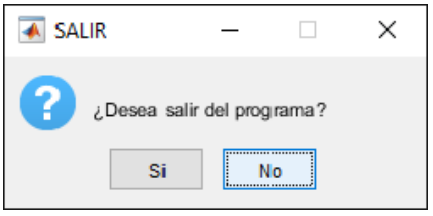

Figura 4.2: Mensaje de confirmación de salida de la herramienta. *Fuente:Propia*

Por el contrario, si se presiona el botón `Ingresar`, se da por terminada la función de la pantalla inicial, por tanto la misma se cierra para dar paso a la ventana `Menú`, como se muestra en la Figura 4.3, a partir de la cual el usuario dispondrá de acceso a cualquiera de las dos fases de funcionamiento mencionadas en la descripción del sistema.

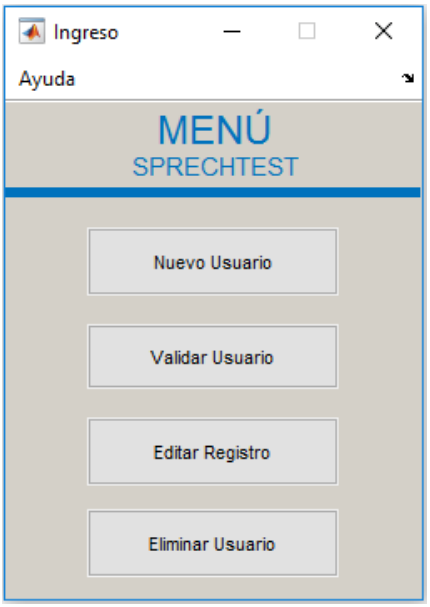

Figura 4.3: Ventana Menú de la herramienta. *Fuente: Propia*

4.1.2. Ventana Menú

Al ingresar al sistema, se despliega el recuadro de acceso a las fases de funcionamiento, para el cual se tiene una sección documentada en la parte superior izquierda, generada con el propósito de servir de orientación respecto al uso de la interfaz y los conceptos asociados al procesamiento y caracterización de señales de voz, a través de un manual de usuario y un archivo de contenido teórico vinculado a la temática y fin de la herramienta. Este recuadro puede visualizarse en la Figura 4.4

Por otro lado, se tienen los botones `Nuevo Usuario` y `Validar Usuario`, mostrados en la Figura 4.3, que dan entrada a las etapas de inscripción y reconocimiento respectivamente, así como la opción `Editar Registro`, a partir de la cual es posible acceder a la data de un locutor particular y modificar los valores de sus parámetros característicos derivados del análisis acústico, y finalmente `Eliminar Usuario`, mediante la cual es posible suprimir las muestras y datos particulares de un individuo previamente registrado.

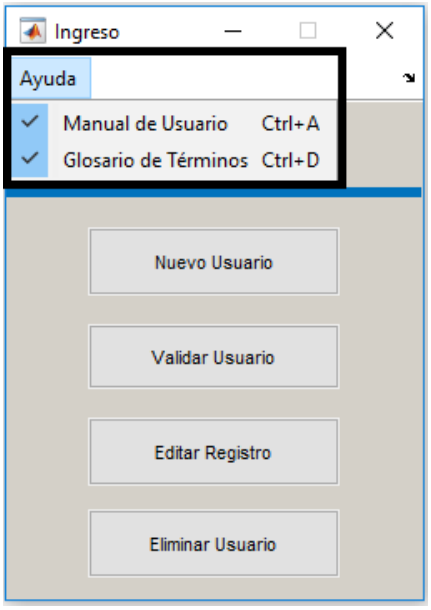

Figura 4.4: Recuadro desplegable en la ventana Menú con acceso a las fases de funcionamiento de la herramienta. *Fuente: Propia*

4.1.3. Ventana Nuevo Usuario

En la ventana corresponde a la fase de inscripción del sistema, tal como se observa en la Figura 4.5, donde nuevamente en la parte superior izquierda se tiene el menú desplegable `Ayuda`, mediante el cual se tiene acceso a la documentación de apoyo al usuario. De esta figura, se observa que la ventana comprende de 4 secciones particulares: `Registro`, `Adquisición de Muestra`, `Procesar Muestra`, `Graficar`, así como los botones `Finalizar Registro` y `Regresar`.

4.1.3.1. Registro

En esta primera sección se habilita el ingreso y captura de los datos personales (Cédula, Nombre, Apellido, Género) asociados a la persona como se indica en la Figura 4.6, donde para cada campo en caso de que existan errores se generarán ventanas emergentes indicándole al usuario cuál ha sido la falla, para la debida corrección que dé paso a la sección de `Adquisición de Muestra`. Para las diferentes

Figura 4.5: Ventana para registro de nuevo usuario. *Fuente: Propia*

eventualidades mencionadas anteriormente, se despliegan los cuadros de diálogo indicados mediante el conjunto de figuras comprendidas entre 4.7 y 4.10, los cuales indican el uso de caracteres inválidos o campos incompletos.

Figura 4.6: Sección `Registro` de la ventana de nuevo usuario. *Fuente: Propia*

Una vez se verifica que los datos son ingresados bajo el formato deseado, se generan dos posibles mensajes emergentes tras pulsar el botón `OK`, de encontrarse previamente registrado el locutor asociado a la cédula entrante se despliega el cuadro de diálogo de la Figura 4.11, retornando el panel a sus condiciones iniciales,

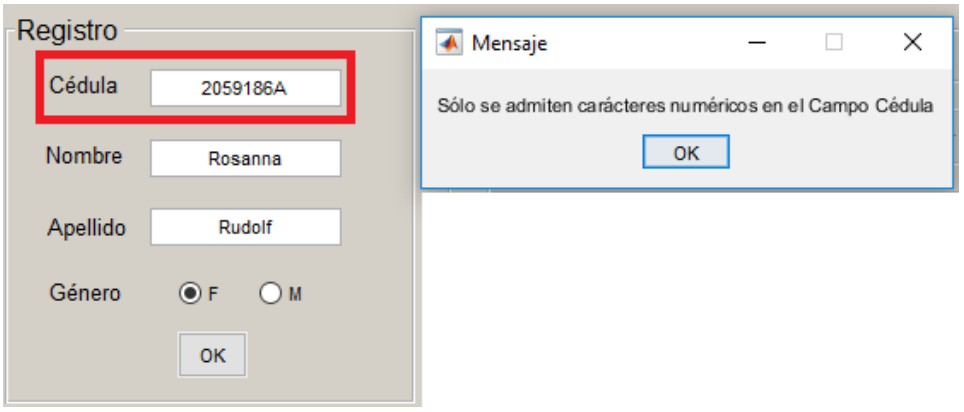

Figura 4.7: Validación del campo `Cédula` para registro de nuevo usuario. *Fuente: Propia*

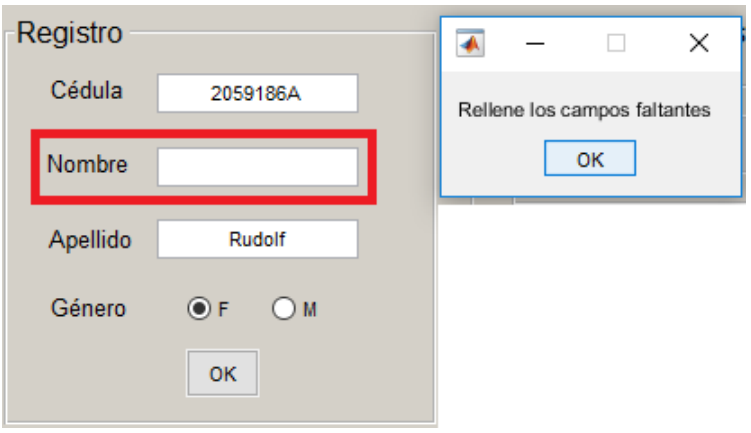

Figura 4.8: Validación del campo `Nombre` para registro de nuevo usuario. *Fuente: Propia*

por el contrario, de no hallarse en la data disponible, se procede con una ventana como la Figura 4.12, permitiendo el manejo de la siguiente sección e inhabilitando la de registro para evitar alterar los datos suministrados.

4.1.3.2. Adquisición de Muestra

Este panel se encuentra estructurado en 3 sub-secciones como se distingue en la Figura 4.13, donde inicialmente sólo se encuentra habilitado el botón `Grabar` contenido en `Grabación 1`, a partir del cual se habilitarán en forma sucesivas los

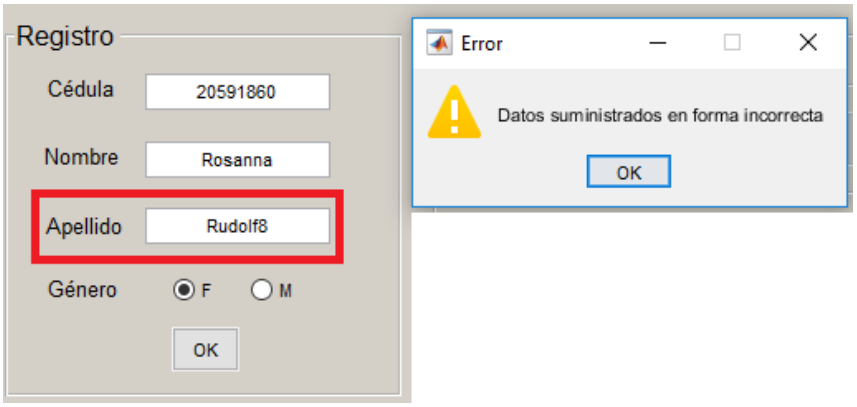

Figura 4.9: Validación del campo `Apellido` para registro de nuevo usuario. *Fuente: Propia*

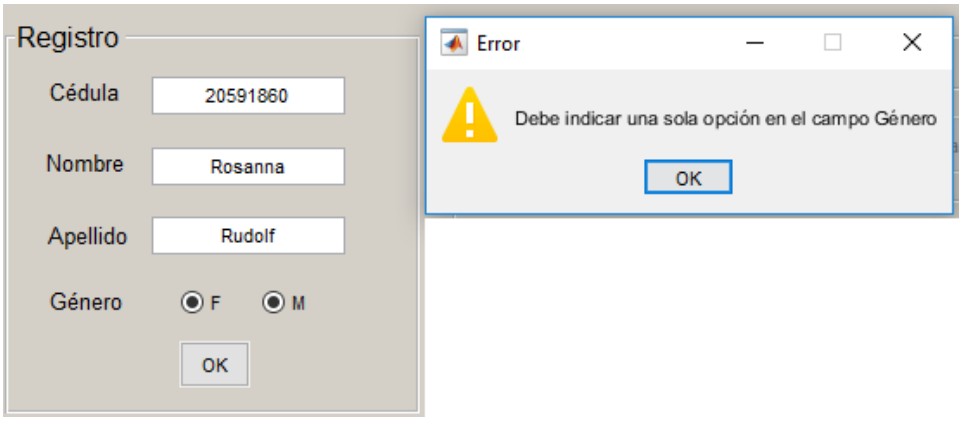

Figura 4.10: Validación del campo `Género` para registro de nuevo usuario. *Fuente: Propia*

restantes según se registren las señales de voz necesarias para su almacenamiento, procesamiento y extracción de características para conformar la base de datos que servirá de entrada a la red de entrenamiento. Además de dicho botón, se encuentran los botones `Reproducir` y `Limpiar Audio`, que permiten escuchar la grabación una vez finalizado el proceso de captura y para el tratamiento de la muestra de voz mediante Transformada de Wavelet, generando el archivo final de mayor fidelidad que será empleado posteriormente para el análisis acústico.

El proceso de grabación es controlado a partir de una serie de mensajes emergentes que le indican al usuario el inicio y finalización de la captura de muestra. Al

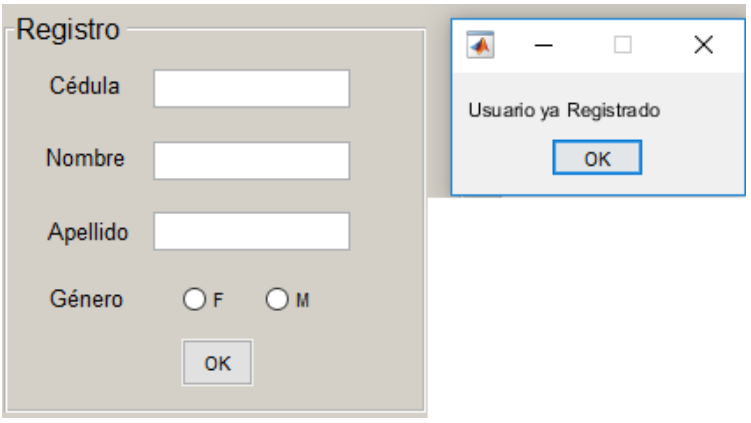

Figura 4.11: Ventana emergente notificando usuario ya registrado. *Fuente: Propia*

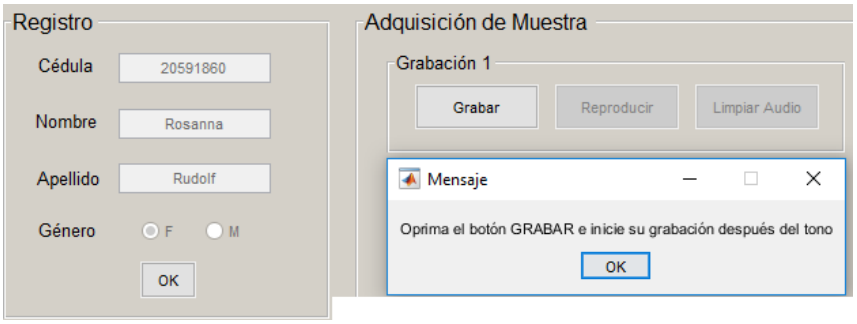

Figura 4.12: Vetana emergente notificando la opción de `Grabar` para el registro del nuevo usuario. *Fuente: Propia*

presionar `Grabar` se despliega la ventana dada en la Figura 4.14, y una vez transcurrido el tiempo de captura de muestra, se indica mediante el recuadro de la Figura 4.15 la finalización, habilitando los botones `Reproducir` y `Limpiar Audio`.

Dicho proceso descrito anteriormente se repite con igual desempeño en las restantes sub-secciones, `Grabación 2` y `Grabación 3`, que tras pulsar el botón `Limpiar Audio` de esta última, se habilita `Procesar Muestra`.

4.1.3.3. Procesar Muestra

Para la etapa de `Procesar Muestra`, mostrada en la Figura 4.16, se tiene el botón `Calcular`, mediante el cual se hace un llamado de la herramienta Praat para

Figura 4.13: Recuadro desplegable en la ventana Menú con acceso a las faces de funcionamiento de la herramienta. *Fuente: Propia*

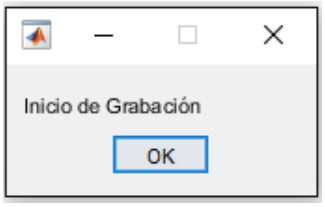

Figura 4.14: Ventana emergente notificando inicio de grabación. *Fuente: Propia*

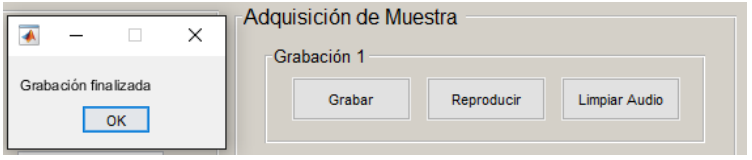

Figura 4.15: Ventana emergente notificando que la grabación ha finalizado. *Fuente: Propia*

procesar y extraer las parámetros característicos de interés de los audios obtenidos anteriormente.

Una vez se hace efectiva la ejecución, se despliegan las ventanas características del software, figura 4.17, sin embargo para los fines de la herramienta, la pantalla

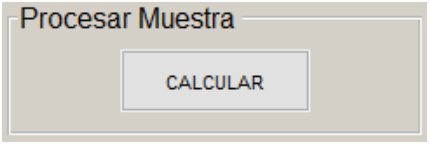

Figura 4.16: Sección `Procesar Muestra` de la ventana de nuevo usuario. *Fuente: Propia*

`Praat Picture` no genera ningún tipo de soporte, por tanto se puede descartar y cerrar, mientras que mediante `Praat Objects` se ejecutaran los ficheros principales vinculados con el procesamiento y extracción de las muestras derivadas del tratamiento acústico.

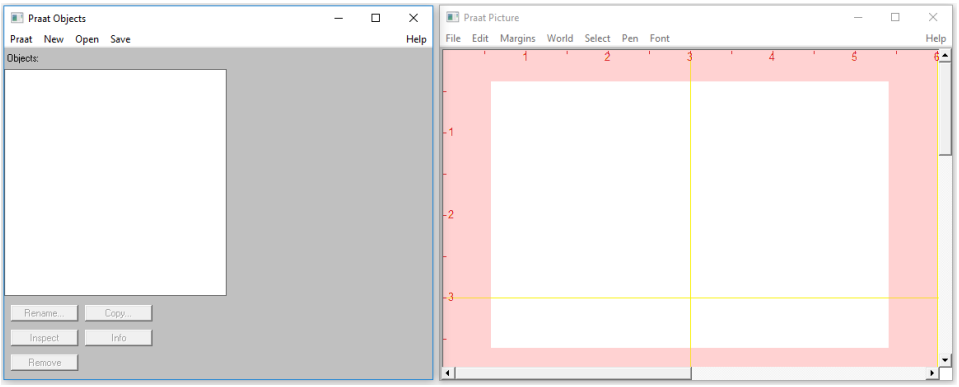

Figura 4.17: Ventana de procesamiento de `Praat Object` y `Praat Picture`. *Fuente: Propia*

En la ventana `Praat Objects`, mediante el menú `Praat` ubicado en la parte superior izquierda se accede al sub-menú `Registro...`, ver Figura 4.18, con el cual se procesan los tres archivos de voz, extrayendo de los mismos los formantes y *pitch* asociados, según una serie de características de diseño para el correcto funcionamiento del fichero, vistas en la Figura 4.19, fijadas por condición de diseño aunque sujetas a modificación según los intereses del usuario. En caso de que el fichero que almacena los datos derivados del procesamiento anterior ya exista, se notificará para validar su sobreescritura, visto en la Figura 4.20.

Finalmente, en pantalla se generan cada uno de los valores extraídos derivados del procesamiento de cada audio, parámetros que posteriormente serán empleados

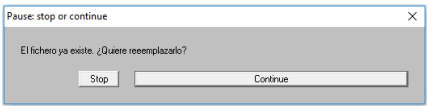

Figura 4.18: Recuadro desplegable en la ventana Menú con acceso a las faces de funcionamiento de la herramienta. *Fuente: Propia*

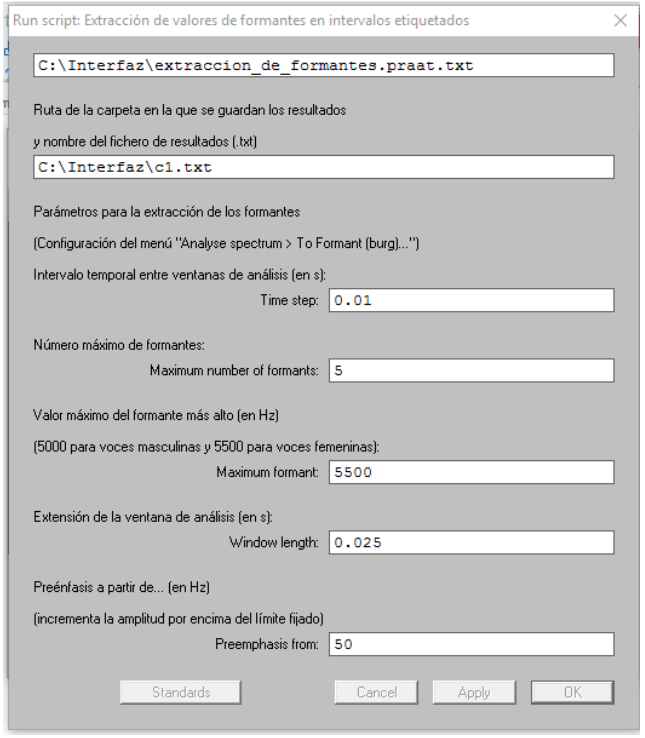

Figura 4.19: Fichero de extracción de valores formantes en intervalos etiquetados. *Fuente: Propia*

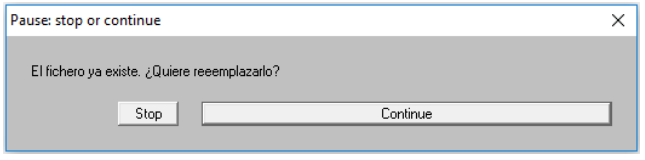

Figura 4.20: Ventana emergente verificando sobreescritura del fichero. *Fuente: Propia*

como fuente de información para el entrenamiento de la red neuronal, indicado en la Figura 4.21. Una vez se dé por cumplida su función, se remite a cerrar la ventana, retornando a la ventna `Registro`, cerrando en forma simultánea el Praat

y habilitando el botón `Finalizar registro`

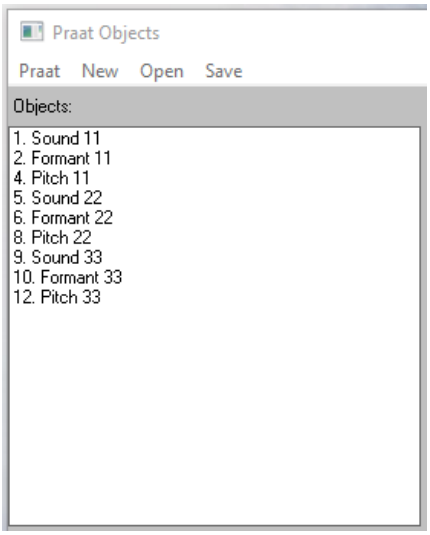

Figura 4.21: Valores de formantes extraídos del procesamiento de cada audio. *Fuente: Propia*

4.1.3.4. Graficar

A partir de este panel, compuesto de un menú desplegable y el botón `Generar` de la ventana `Nuevo Usuario`, ver Figura 4.22, es posible buscar y seleccionar las respectivas señales de voz almacenadas durante la fase Adquisición de Muestra, permitiendo obtener para cada sub-sección una representación gráfica de la muestra de audio previa y posterior a su tratamiento con TW y algoritmo de supresión de silencios, ejemplo de ello se puede observar en la Figura 4.23.

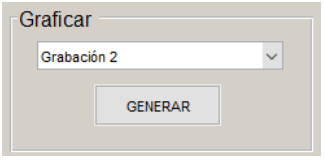

Figura 4.22: Sección `Graficar` de la ventana de nuevo usuario. *Fuente: Propia*

Una vez obtenida la data de parámetros característicos de las señales de voz desde el Praat, se habilita el botón `Finalizar Registro`, mediante el cual se crea

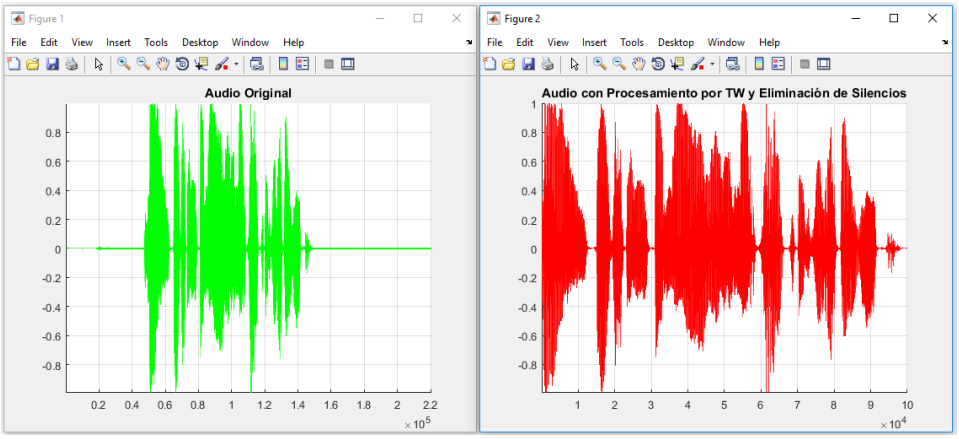

Figura 4.23: Visualización temporal de la señal de audio original y la procesada con TW y algoritmos de supresión de silencios. *Fuente: Propia*

un archivo de datos con referencia a la cédula del usuario, donde para cada locutor se almacenan sus datos personales e información de los formantes y *pitch* de las señales de audio procesadas, completando así la fase de inscripción.e indicando el éxito de la misma mediante un breve mensaje en pantalla, ver Figura 4.24. Por último, se dispone del botón `Regresar`, que en caso de ser empleado, despliega un cuadro de diálogo como el dado en la Figura 4.25, donde el usuario tiene la posibilidad de optar por retornar a la ventana Menú o registrar a otra persona, retornando la ventana `Nuevo Usuario` a sus condiciones iniciales.

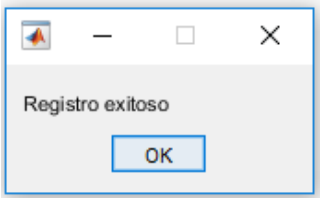

Figura 4.24: Ventana emergente notificando registro exitoso del nuevo usuario. *Fuente: Propia*

4.1.4. Validar Usuario

Para esta etapa se despliega una pantalla responsable del funcionamiento de la segunda fase denominada reconocimiento, cuya distribución se muestra en la

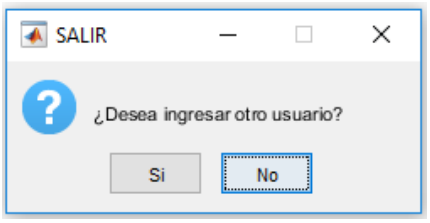

Figura 4.25: Ventana emergente verificando ingreso de nuevo usuario. *Fuente: Propia*

Figura 4.26, la cual contiene el respectivo menú de `Ayuda` en la sección superior izquierda de la ventana, así como se compone de 4 sub-secciones específicas: `Datos Personales`, `Adquisición de Muestra`, `Procesar Muestra`, `Graficar` y `Visualización del Resultado`.

Figura 4.26: Ventana de validar usuario. *Fuente: Propia*

4.1.4.1. Datos Personales

Para esta sección se tiene el campo de ingreso del número de cédula del locutor con previo registro al que se espera verificar, en caso de que se complete la información requerida en forma errónea, se despliega una ventana emergente, como se detalló anteriormente en la Figura 4.26, por otro lado, de resultar un dato válido, se verificará en primera instancia de que el sujeto exista en sistema, en caso afirmativo, se obtienen sus datos y se muestran en el campo `Usuario`, ver Figura 4.27, habilitando `Adquisición de Muestra`, de lo contrario se indica mediante un mensaje, que el mismo no aparece en sistema, observar en la Figura 4.28.

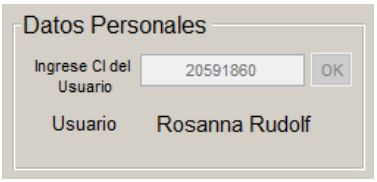

Figura 4.27: Sección `Datos Personales` de la ventana de validar usuario. *Fuente: Propia*

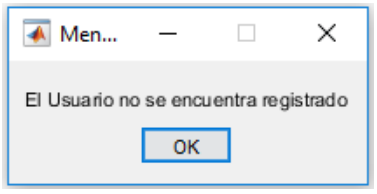

Figura 4.28: Ventana emergente notificando que el usuario no se encuentra registrado. *Fuente: Propia*

4.1.4.2. Adquisición de Muestra

Desde este panel se obtiene la señal de voz mediante la cual se realizará la debida comparación con la data del usuario electo, con el propósito de validar si el locutor que graba es quien indica ser mediante la selección del campo `Cédula`. Al pulsar el botón grabar, se despliega una ventana como la dada en la Figura 4.14, indicándole al usuario que puede generar la muestra, transcurrido el tiempo fijado para ello, se muestra un mensaje de finalización y se habilitan los botones `Reproducir` y

`Limpiar Audio`, para escuchar el archivo de audio generado y procesar mediante los algoritmos de TW y eliminación de ruido el mismo, respectivamente, ver Figura 4.29, habilitando de este modo las secciones `Procesar Muestra` y `Graficar`.

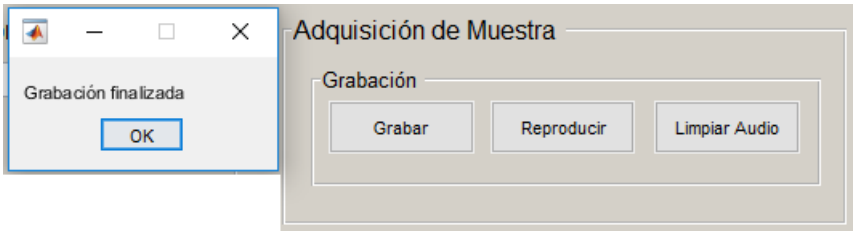

Figura 4.29: Adquisición de la muestra para el proceso de validación. *Fuente: Propia*

4.1.4.3. Procesar Muestra

En este panel al presionar `Calcular`, ver la Figura 4.30, el usuario accede al software Praat, para procesar el archivo de audio generado anteriormente y extraer las características: formantes y *pitch*. Al ejecutarse el programa se despliegan nuevamente las ventanas `Praat Objects` y `Praat Picture`, esta última no cumple ninguna función para el proceso por tanto puede cerrarse, mientras que en la primera ventana, accediendo al menú Praat, se selecciona el sub-menú `Validar...`, ver Figura 4.30, mediante el cual se ingresa a la ventana de configuración de parámetros para el funcionamiento del fichero, cuyos valores ya previamente están ajustados por diseño, sin embargo, están libres para su modificación, esto puede observarse en la Figura 4.31.

En caso de que previamente se haya realizado alguna validación, se generará un mensaje emergente como el de la Figura 4.20, para confirmar la opción de sobreescribir archivo, luego de ello se ejecuta el fichero, originando los siguientes resultados mostrados en pantalla correspondientes a la extracción del *pitch* y formantes, ver figura 4.32, tras lo cual se da por terminada la ejecución del Praat, habilitando el botón validar de la sección `Visualización del Resultado`.

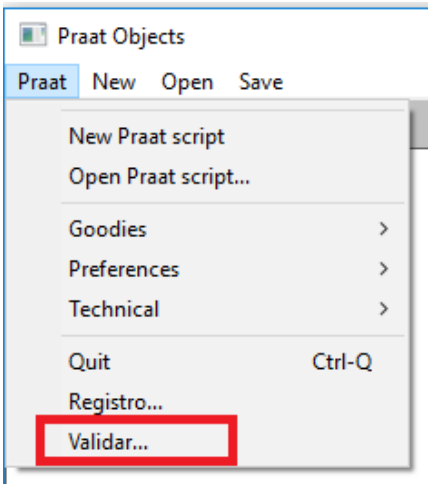

Figura 4.30: Opción `Validar` de la ventana `Praat Object`. *Fuente: Propia*

4.1.4.4. Graficar

En esta sección, compuesta del *checkbox* `Seleccionar Muestra` y el botón `Generar`, se valida la representación gráfica de las muestras de audios, tanta la original como la obtenida tras su tratamiento mediante TW y eliminación de silencio, lo cual no es un paso indispensable para el funcionamiento de la etapa de reconocimiento, es más un soporte en caso de que se quiera realizar una comparación en la eficiencia de los algoritmos de procesamiento de audio, la cual se presenta en la Figura 4.33, mientras que la representación gráfica se genera como se muestra en la figura 4.23.

4.1.4.5. Visualización del Resultado

Por último se tiene la sección para habilitar el entrenamiento de la red neuronal mediante algoritmos de aprendizaje profundo, a partir del botón `Validar`, la cual toma como parámetros de entrada los valores obtenidos de la data registrada del locutor seleccionado a autentificar y del locutor que realiza la prueba, esto puede observarse en la Figura 4.34. Una vez se genera el debido procesamiento, en el recuadro `Resultado` se expresa la decisión del sistema, en caso de que sea satisfactorio a su criterio de evaluación, se indicará según la Figura 4.35, sin embargo,

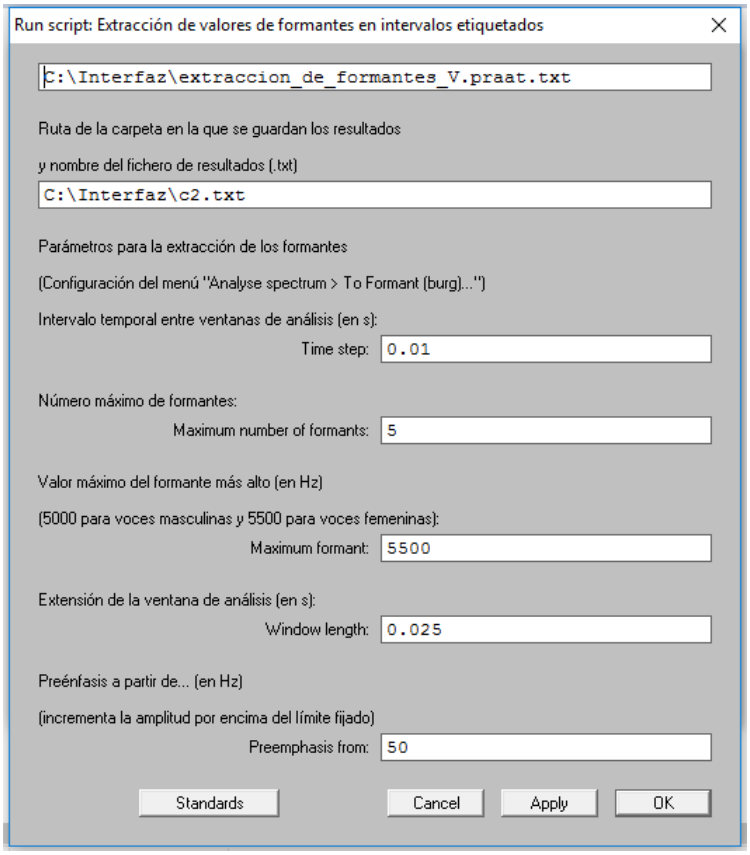

Figura 4.31: Fichero de extracción de valores formantes en intervalos etiquetados.
Fuente: Propia

para el caso contrario, se observará lo mostrado en la Figura 4.36. De igual forma, se despliega un recuadro de diálogo, como el que se observa en la Figura 4.37 mediante el cual el usuario puede indicar si desea repetir la validación para el mismo registro, de tal forma que puede repetir el proceso descrito a lo largo de esta fase, sea para obtener una mejor muestra de voz o con los mismos datos fuentes volver a habilitar el entrenamiento, a modo de intentar comprobar si nuevamente bajo las mismas condiciones el resulta es otro, en caso de que se indique 'No', se reinician los campos de la ventana, de modo que se ingrese otro sujeto para comprobar la identidad.

Finalmente mediante el botón `Regresar`, el usuario a partir del cuadro de diálogo, puede optar por abandonar la pantalla `Validar Usuario` para acceder a la

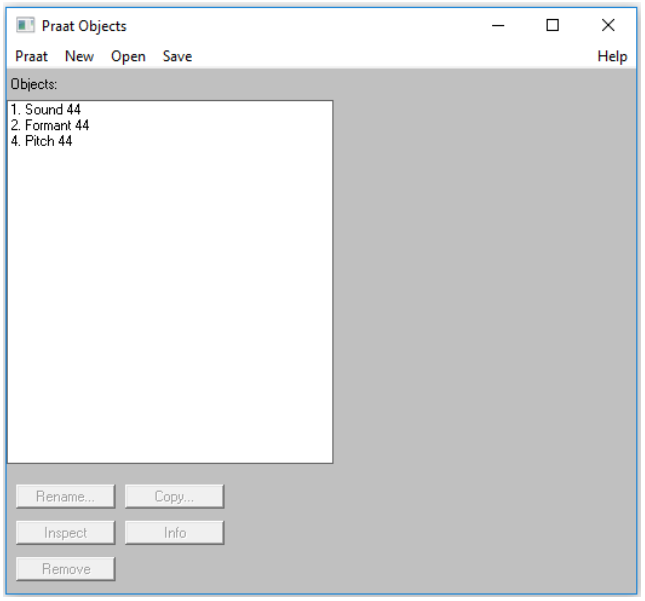

Figura 4.32: Valores de formantes extraídos del procesamiento de cada audio. *Fuente: Propia*

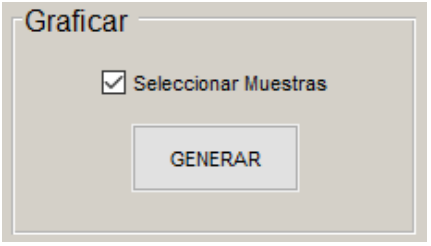

Figura 4.33: Sección `Graficar` de la ventana de validar usuario. *Fuente: Propia*

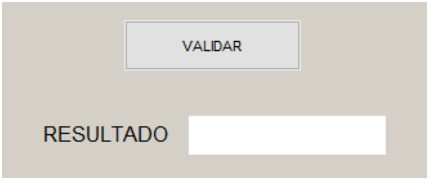

Figura 4.34: Sección `Validar` de la ventana de validación de usuario. *Fuente: Propia*

pantalla Menú para realizar otra tare o mantenerse en la misma, según la Figura 4.38.

Figura 4.35: Campo de validación al reconocer al usuario registrado. *Fuente: Propia*

Figura 4.36: Campo de validación al no reconocer al usuario registrado. *Fuente: Propia*

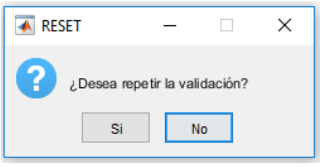

Figura 4.37: Ventana emergente notificando la repetición del proceso de validación. *Fuente: Propia*

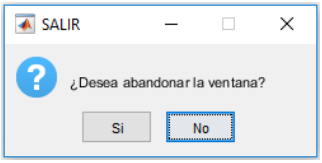

Figura 4.38: Ventana emergente verificando abandono de la ventana de validación. *Fuente: Propia*

4.1.5. Editar Registro

En esta sección se habilita la opción para modificar la data de un usuario previamente registrado, tal como se muestra en la Figura 4.39, donde para verificar la existencia o no de un sujeto, se requiere del ingreso de la cédula de identidad, en caso de que no existan errores de escritura o que el usuario no se encuentre registrado, se habilitarán los campos con sus datos y el botón `Editar`, esto puede observarse en la Figura 4.40, que una vez pulsado, remite a la ventana `Nuevo Usuario` con los datos personales ya cargados y únicamente se le pedirá que realice el proceso descrito desde la sub-sección `Adquisición de Muestras`, como se muestra en la

Figura 4.41.

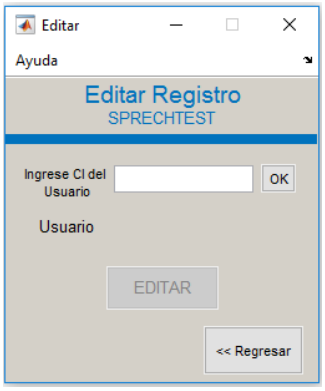

Figura 4.39: Ventana de edición de un usuario ya registrado. *Fuente: Propia*

Figura 4.40: Ventana de edición, identificando al usuario registrado. *Fuente: Propia*

Como método para abandonar la ventana, se encuentra el botón `Regresar`, mediante el cual se despliega un cuadro de diálogo para confirmar el retorno a la pantalla Menú o mantener el funcionamiento de la actual, por lo que se muestra la misma notificación mostrada en la Figura 4.38.

4.1.6. Eliminar Registro

Para esta sección se despliega una pantalla mediante la cual se gestiona la eliminación de un usuario registrado en la base de datos del sistema, tal cual como se muestra en la Figura 4.42, donde se tiene una etapa de ingreso de datos, la cual se

Figura 4.41: Ventana nuevo usuario para la opción de edición de la adquisición de muestras. *Fuente: Propia*

completa con la cédula asociada al locutor que se desee retirar, la cual es verificada y confirmada mediante el botón OK, en caso de que exista un error de transcripción, se desplegará un mensaje emergente o de ser válida, se procederá a habilitar el campo usuario con sus referencias y el botón `Eliminar`, ver Figura 4.43, que una vez pulsado validará eliminar la data existente, asegurando que el proceso sea exitoso mediante un mensaje mostrado en la Figura 4.44.

Una vez culminado dicho proceso, el usuario puede optar por abandonar dicha ventana o mantenerse en ella para gestionar otra selección, mediante el botón `Regresar`.

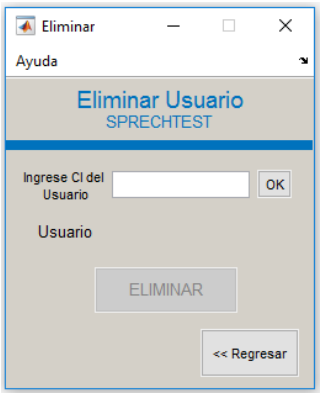

Figura 4.42: Ventana de eliminar usuario. *Fuente: Propia*

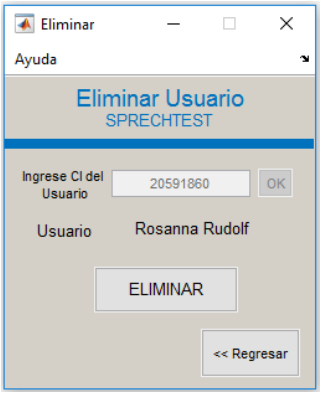

Figura 4.43: Ventana de eliminación de usuario, identificando el usuario registrado. *Fuente: Propia*

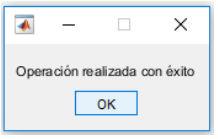

Figura 4.44: Ventana emergente notificando que el usuario fue eliminado. *Fuente: Propia*

4.2. Análisis de los ensayos realizados.

Se realizó un conjunto de pruebas con la herramienta diseñada para comprobar el funcionamiento y efectividad de la misma. Para cada caso evaluado se muestra la tabla resumen de los resultados obtenidos en la etapa de validación. Es importante

señalar que en las tablas siguientes se considera como data original aquella a la que no se le ha hecho un tratamiento previo para disminuir el ruido y los silencios presentes en ella. Con esto se buscó evaluar el beneficio que aporta la aplicación de dicha técnica sobre la data antes de introducirla a la red Deep Learning.

Caso 1. Evaluación de la herramienta diseñada considerando las etapas de entrenamiento y validación realizadas por la misma persona.

- Hombres verificando su identidad (H-H).
- Mujeres verificando su identidad (M-M).
- Usuarios verificando su identidad con ruido extra de fondo (U+R).

Tabla 4.1: Porcentaje de aciertos considerando la misma persona para las etapas de entrenamiento y validación

Caso	Ensayo	Tamaño de la muestra	Datos originales sin procesamiento		Con eliminación de ruido y silencios	
			Nro de errores	% de aciertos	Nro de errores	% de aciertos
Caso 1	H-H	30	1	96,67 %	1	96,67 %
	M-M	30	3	90 %	2	93,33 %
	U+R	12	7	41,66 %	5	58,33 %

En la Tabla 4.1 se reflejan los porcentajes de aciertos obtenidos en cada sub caso, donde se puede notar que para la verificación de identidades masculinas (ensayo H-H) la tasa de éxito es mayor que al verificar identidades femeninas (ensayo M-M). Para el sub caso de usuarios verificando su identidad con ruido extra de fondo el porcentaje de acierto disminuyó notablemente en función a los casos anteriores. Cabe destacar que el resultado de esta última prueba depende de la intensidad del ruido de fondo con respecto a la intensidad de la voz del usuario estudiado, ya que se encontró que la herramienta no validó correctamente en los ensayos donde la intensidad del ruido de fondo fue comparable con la del usuario, considerando este ruido como el resultado de otras personas hablando durante la toma de la muestra del usuario de interés.

Caso 2. Evaluación de la herramienta diseñada considerando las etapas de entrenamiento y validación realizadas por personas diferentes.

- Hombres verificando la identidad de otros Hombres. (H1-H2)
- Hombres verificando la identidad de Mujeres. (H-M)
- Mujeres verificando la identidad de otras mujeres. (M1-M2)
- Mujeres verificando la identidad de hombres. (M-H)
- Usuarios verificando la identidad de otros usuarios con voces similares. (U1-U2)

Tabla 4.2: Porcentaje de aciertos de la herramienta considerando distintas personas para las etapas de entrenamiento y validación

Caso	Ensayo	Tamaño de la muestra	Datos originales sin procesamiento		Con eliminación de ruido y silencios	
			Nro de errores	% de aciertos	Nro de errores	% de aciertos
Caso 2	H1-H2	30	2	93,33 %	1	96,67 %
	H-M	30	0	100 %	0	100 %
	M1-M2	30	3	90 %	2	93,33 %
	M-H	30	0	100 %	0	100 %
	U1-U2	10	1	90 %	1	90 %

Para la Tabla 4.2 se observó un buen porcentaje de aciertos para cada sub caso, siendo los ensayos «H-M y M-H» los más exactos, teniendo la herramienta un 100 % de efectividad en la distinción entre personas de sexos opuestos. Para el caso de los ensayos «H1-H2 y M1-M2» se obtuvieron porcentajes de 93,33 % y 90 % respectivamente, siendo un buen resultado en función al número de muestras tomadas. Para el caso restante (U1-U2) la respuesta también estuvo a favor de la herramienta, ya que a pesar de haber considerado personas con voces parecidas, la misma pudo distinguir en la mayoría de los casos si el usuario era válido o no.

Caso 3. Evaluación de la herramienta diseñada considerando la etapa de validación realizada con frases diferentes a la cédula del usuario registrado.

- Hombres verificando su identidad alternando los dígitos de su número de cédula. (H-C)

- Mujeres verificando su identidad alternando los dígitos de su número de cédula. (M-C)
- Usuarios verificando la identidad de otros usuarios con cualquier frase. (U3-C)

Tabla 4.3: Porcentaje de aciertos de la herramienta realizando las etapas de entrenamiento y validación con frases diferentes.

Caso	Ensayo	Tamaño de la muestra	Datos originales sin procesamiento		Con eliminación de ruido y silencios	
			Nro de errores	% de aciertos	Nro de errores	% de aciertos
Caso 3	H-C	15	5	66,67 %	5	66,67 %
	M-C	15	6	60 %	5	66,67 %
	U3-C	15	0	100 %	0	100 %

En la Tabla 4.3se muestran los resultados obtenidos para casos donde en lugar de validar un usuario diciendo su número de cédula, se validó alternando los dígitos de la misma, notando que en los ensayos "H-C y M-C" el número de errores fue alto en función a la cantidad de muestras tomadas, ya que la herramienta fue diseñada para verificación de usuarios con frase dependiente, por lo que las veces que acertó si la persona era válida fue en los casos donde la frase mantenía la misma estructura de las grabaciones realizadas para el entrenamiento. Para el caso restante (U3-C) el porcentaje de efectividad fue de 100 % ya que la herramienta logro reconocer que en ningún momento se trataba de un usuario valido.

Caso 4. Evaluación de la herramienta diseñada considerando la etapa de validación con distintos estados de ánimo del usuario.

- Hombres verificando su identidad. (H-EA)
- Mujeres verificando su identidad. (M-EA)
- Usuario verificando la identidad de otros usuarios. (U-EA)

Tabla 4.4: Porcentaje de aciertos de la herramienta considerando los estados de ánimo del usuario para la etapa de validación.

Caso	Ensayo	Tamaño de la muestra	Datos originales sin procesamiento		Con eliminación de ruido y silencios	
			Nro de errores	% de aciertos	Nro de errores	% de aciertos
Caso 4	H-EA	8	1	87,5 %	1	87,5 %
	M-EA	8	0	100 %	0	100 %
	U-EA	8	0	100 %	0	100 %

En este caso se evaluó la herramienta en función a los estados de ánimo de los usuarios para verificar si al variar las intensidades de la voz moderadamente se mantenía el buen funcionamiento de la misma, arrojando un porcentaje de aciertos de 87,5 % y 100 % respectivamente para los ensayos «H-EA y M-EA» reflejados en la Tabla 4.4, lo que nos permitió observar que a pesar de que el usuario no siempre hable con la misma intensidad de voz, la herramienta es capaz de detectar si es o no es quien dice ser. Para el último caso (U-EA) se obtuvo un 100 % de aciertos ya que independientemente del estado de ánimo de la persona, la herramienta reconoció que se trataba de un usuario inválido.

Capítulo V

Conclusiones y recomendaciones

5.1. Conclusiones

- Los parámetros de la voz que permitieron validar a un usuario con mayor porcentaje de acierto fueron los cuatros primeros formantes y el pitch, por su característica de dependencia directa de las propiedades de cuasi periodicidad que define los segmentos de voz por causa de las cuerdas vocales, así como la forma y el tamaño del tracto vocal, elementos integrantes del sistema fonatorio humano, los cuales difieren para cada individuo.

- Con los ensayos realizados se verificó la utilidad de la Transformada Wavelet como una herramienta adicional para el procesamiento previo de la data a ser aplicada a la red Deep Learning pues, en el 40 % de los ensayos realizados se logró mejorar el porcentaje de acierto en la validación del usuario y en ningún caso se obtuvo un porcentaje menor con respecto a aquellos que no incluían la Transformada de Wavelet. La Wavelet madre que permitió un mejor comportamiento de la herramienta fue la Daubechies-4.

- Como complemento a los algoritmos de la Transformada de Wavelet, se incluyó eliminación de silencios a la etapa de tratamiento de las señales de voz, como mejora en la calidad de las muestras previo a su procesamiento acústico para la extracción de parámetros característicos, comprobando para los

diferentes ensayos que no sólo resultó útil para eliminar aquellos segmentos al inicio y fin de la grabación que no aportan información, sino que disminuye posibles varaciones en los valores obtenidos de formantes y pitch al no considerarse parte del análisis vocal.

- La presente investigación demostró que las redes neuronales basadas en Deep Learning son apropiadas para implementar sistemas de verificación de usuarios por medio de su voz. El modelo de Red Deep Learning que arrojó mejor porcentaje de acierto en el proceso de validación fueron las Autoencoder (AEN), con una arquitectura de cuatro capas con 80 neuronas cada una y función de activación o transferencia «purelin».

- Los ensayos realizados a la herramienta permiten concluir que la elección de una frase única para identificar a cada usuario mejora el porcentaje de acierto, sujeto a que cada locutor tiene una forma única de pronunciar un texto particular, lo cual justificó la selección de la cédula para ello, facilitando el proceso de extracción de características al mantener la invariabilidad en forma, dando al sistema mayor eficiencia para el tiempo de registro y la toma de decisión en la fase de validación.

5.2. Recomendaciones

1. Verificar el diseño de la herramienta propuesta bajo entorno Python, con el propósito de facilitar el uso y desarrollo de la aplicación mediante software libre.

2. Complementar la disposición del entorno de programación con el uso de TensorFlow, sistema gratuito, de continua actualización y sin requisitos previos para su acceso, basado en aprendizaje automático para el uso de herramientas en materia de procesamiento inteligente a partir de Redes Neuronales y Deep Learning. Actualmente abierto a todo público, como almacén para compartir experiencias y resultados.

3. Uso de un mismo dispositivo para capturar la entrada de voz al sistema, tanto para la etapa de inscripción como validación. Se sugiere para ello el uso de un micrófono de calidad de modo que el proceso de grabación tenga las condiciones más favorables dentro del rango de frecuencias de los sonidos del habla.

4. Comprobar la factilidad de desarrollo de un sistema con formato de aprendizaje no dependiente del texto, de modo que para las fases de registro y autentificación el locutor pueda libremente generar el habla sin la restricción de un proceso intermedio de sugerencia respecto a la información a suministrar, posibilitando acelerar el funcionamiento y el tiempo de respuesta de la herramienta.

5. Realizar el registro con diferentes estados de ánimo del locutor, de modo que la data tenga mayor robustez respecto a las posibles variaciones que experimenta la voz del locutor según su conducta.

6. Se sugiere realizar la fase de entrenamiento con diferentes frases para aumentar el nivel de robustez de la data.

Apéndice A

Algoritmo de procesamiento para la Transformada de Wavelet.

```
[s,Fs] = audioread('1.wav');
Ls = length(s);
w = 'db4'; %Seleccion de Wavelet Madre Daubechies 4
n = 12; %Descomposicion de Orden 12
tptr = 'modwtsqtwolog'; %Criterio de Seleccion del Umbral: Umbral Universal
                        %Definido mediante la teoria de Donoho
sorh = 's'; %Seleccion de la funcion de transferencia para el umbral:
            %Soft-Thresholding
scal = 'mln'; %Seleccion del metodo de reescala: Uso de estimacion del nivel
              %de ruido dependiente del mismo.
[s_den, C]=wden(s,tptr,sorh,scal,n,w); %Aplicar TW.
```

Código 1.1: Algoritmo para la TW.

Apéndice B

Algoritmo de procesamiento para la eliminación de silencios.

```
1
2
3  %Silencio
4  %Corta el silencio en la senal completa
5  function y = silencio(s,Fs)
6  len = length(s); %Tamano del vector
7  d=max(abs(s));
8  s=s/d;
9  avg_e = sum(s.*s)/len; %promedio senal entera
10 THRES = 0.4;
11 y = [0];
12 for i = 1:12500:len-12500 %Generar una segmentacion de mayor valoracion
13 seg = s(i:i+12499);%Segmentos
14 e = sum(seg.*seg)/12500; %Promedio de cada segmento
15 if( e> THRES*avg_e) %Si el promedio energetico es mayor que la senal
16                       %completa por el valor umbral.
17     y=[y;seg(1:end)];% Almacena en y, sino es eliminado como espacio en blanco
18 end;
19 end
```

Código 2.1: Algoritmo para eliminación de silencios.

Apéndice C

Algoritmo de procesamiento para el entrenamiento con Deep Learning.

```

function [sal] = rac2(x, y, x3)
%Variable de uso global
global resp
hiddenSize = 80;
autoenc1 = trainAutoencoder(x,hiddenSize, 'L2WeightRegularization',0.001,
'SparsityRegularization',4,'SparsityProportion',0.05,...
    'DecoderTransferFunction','purelin', 'ShowProgressWindow',false);
features1 = encode(autoenc1,x);
hiddenSize = 80;
autoenc2 = trainAutoencoder(features1,hiddenSize,...
    'L2WeightRegularization',0.006,...
    'SparsityRegularization',4,...
    'SparsityProportion',0.05,...
    'DecoderTransferFunction','purelin',...
    'ScaleData',false, 'ShowProgressWindow',false);
features2 = encode(autoenc2,features1);

hiddenSize = 80;
autoenc3 = trainAutoencoder(features2,hiddenSize,...
    'L2WeightRegularization',0.006,...
    'SparsityRegularization',4,...
    'SparsityProportion',0.05,...
    'DecoderTransferFunction','purelin',...
    'ScaleData',false, 'ShowProgressWindow',false);
features3 = encode(autoenc3,features2);

softnet = trainSoftmaxLayer(features3,y,'LossFunction','crossentropy','ShowProgressWindow',false);
deepnet = stack(autoenc1,autoenc2,autoenc3,softnet);
deepnet = train(deepnet,x,y);
wine_type = deepnet(x3);
sal = wine_type;
if sal > 0.9
    resp = 1;
else
    resp = 0;
end
```

Código 3.1: Algoritmo de entrenamiento con Deep Learning.

Referencias Bibliográficas

[1] Poceros F. Villalobos y J. Pérez E. *Sistema de seguridad por reconocimiento de Voz.* Instituto Politécnico Nacional, México. 2013.

[2] Dr. Raúl Arrabales Moreno. «Deep Learning: Qué es y Por qué va a ser una tecnología clave en el futuro de la Inteligencia Emocional.» En: *Xataca* (2016).

[3] Keith Rey Wui Leung, Allard Jongman y Joan A. Sereno Yue Wang. «Proceedings og the 18th International Congress of Phonetic Sciences». En: *Acoustic characteristics of clearly spoken English tense and lax vowels*. Ed. por The Scottish Consortium for ICPhS 2015. University of Glasgow. Glasgow, Scotland, UK: https://www.internationalphoneticassociation.org/icphs-proceedings /ICPhS2015/proceedings.html, 2015, págs. 10-14.

[4] Consejo Superior de Investigaciones Científicas (CSIC). «Cualidad Individual de Voz e Identificación de Locutor (CIVIL)». En: *Centro de Ciencias Humanas y Sociales* (2014).

[5] L. Cruz y M. Acevedo. «Reconocimiento de Voz usando Redes Neuronales Artificiales Backpropagation y Coeficientes LP». En: *SEPI-Telecomunicaciones ESIME IPN Unidad Profesional "Adolfo López Mateos".* (2008).

[6] P. Del Pino y C. Jiménez. «Identificación de algunos parámetros espectrales que determinan la calidad de la voz.» En: *Ingenieria UC* 11.3 (2004), págs. 7-16.

[7] Paulino Del Pino. «Aplicación de la Transformada de Wavelet para el análisis de señales de voz normales y patológicas.» En: *Ingenieria UC* (2008).

[8] Yanina Perdomo. *Desarrollo de software libre interactivo para realizar análisis espectral de voz.* 2015.

[9] Javier Monzón Alonso. *Desarrollo de una Herramienta de Análisis de la Señal de Voz.* Escuela Técnica Superior de Ingeniería y Sistemas de Telecomunicación. 2013.

[10] Dr. en Filología Románica y profesor titular de Lingüística General. Joaquim Llisterri. *Introducción a Praat.* 2014. Universidad Autónoma de Barcelona (UAB). URL: liceu.uab.es/~joaquim/home.html.

[11] José Alejandro Correa Duarte. *Manual de Análisis Acústico del Habla con Praat.* Series Minor XLIX. Instituto Caro y Cuervo. Bogotá, 2014.

[12] Daniel Moral Bárcena y Jesús Villadangos Alonso. *Procesado Digital de Voz para el Reconocimiento del Hablante Aplicado a Dispositivos Móviles.* Escuela Técnica Superior de Ingenieros Industriales y de Telecomunicación. Pamplona, España. 2012.

[13] Rosa Amelia Asuaje. Elsa Mora Gallardo. *El canto de la palabra: Una iniciación al estudio de la prosodia.* Ed. por grupo de investigación en ciencias fonéticas (GICIFO). Centro de Investigación y atención lingüistica (CIAL). Universidad de Los Andes, Venezuela., 2009.

[14] Dr. Juan Carlos Vallejo. «Determinación de Valores Normales para el Análisis Acústico de la Voz.» En: *Servicio de Otorrinolaringología, Hospital Vozandes Quito.* (2010).

[15] Carlos Monzo, Ignasi Iriondo y Elisa Martínez. «Procedimiento para la Medida y la Modificación del Jitter y del Shimmer aplicado a la Síntesis del Habla Expresiva.» En: *V Jornada en Tecnología del Habla.* (2008.). Universidad Ramón Llull. Barcelona, España.

[16] Paulino Del Pino, Iván Granadillo, Mario Miranda, Carlos Jiménez y José A. Díaz. «Diseño de un sistema de medición de parámetros característicos y de calidad de señales de voz.» En: *Revista Ingeniería UC.* 15.2 (2008.). Desarrollado por el Departamento de Electrónica y Comunicaciones de la Escuela de Ingeniería Eléctrica de la Facultad de Ingeniería, Universidad de Carabobo., págs. 13 -20.

[17] Mireia Farrús, Javier Hernando y Pascal Ejarque. «Jitter and Shimmer Measurements for Speaker Recognition.» En: *Universidad Politécnica de Cataluña, Barcelona, España.* (2000).

[18] Oscar Tosi. «Medición Computarizada de Jitter y Shimmer.» En: *la Revista Logopedia, Foniatría y Audiología.* 7.1 (1987). Michigan State University., págs. 49-49.

[19] Ismael Chávez y Antonio Camarera-Ibarrola. «Wavelets en el Reconocimiento de Voz.» En: *Reunión de Otoño de Potencia, Electrónica y Computación.* (2011).

[20] Christian Duque Sanchez y Mauricio Morales Pérez. *Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones.* Universidad Tecnológica de Pereira, Colombia. 2007.

[21] Francisco Santamaría, Camilo A. Cortés y Francisco J. Roman. «Uso de la Transformada de Ondeletas (Wavelet Transform) en la Reducción de Ruido en las Señales de Campo Eléctrico producidas por Rayos.» En: *Información Tecnológica versión On-line. Bogotá, Colombia.* 23.1 (2012), págs. 65-78.

[22] Ing. Hernando González Acosta. «Reducción de Ruido en Señales de Electrocardiogramas Utilizando la Transformada de Wavelet.» Tesis de mtría. Universidad de Carabobo, Facultad de Ingeniería., 2015.

[23] Michel Misiti, Yves Misiti, Georges Oppenheim y Jean-Michel Poggi. *Wavelet ToolboxTM. User's Guide.* Manual de Usuario para versión R2015b. 2015. URL: www.profesores.elo.utfsm.cl/~mzanartu/IPD414/Docs/wavelet_ug.pdf.

[24] Damián Jorge Matich. «Redes Neuronales: Conceptos Básicos y Aplicaciones.» En: *la Universidad Tecnológica Nacional – Facultad Regional Rosario, Dpto. de Ingeniería Química, Grupo de Investigación Aplicada a la Ingeniería Química (GIAIQ).* (2001).

[25] Prof. Xabier Basogain Olabe. «Redes Neuronales Artificiales y Sus Aplicaciones.» En: *Dpto. Ingeniería de Sistemas y Automática de la Escuela Superior de Ingeniería de Bilbao.* (2014).

[26] ObservatorioTIC EC. «Deep Learning: La Nueva Herramienta de la Informática.» En: *Ministerio de Telecomunicaciones y de la Sociedad de Información.* Ecuador, 2015. URL: https://www.youtube.com/watch?v=GgrhnPGIUxM..

[27] Glen Jhan Pierre Restrepo Arteaga. *Aplicación del Aprendizaje Profundo ("Deep Learning") al Procesamiento de Señales Digitales.* Universidad Autónoma de Occidente, Facultad de Ingeniería, Dpto. de Automática y Electrónica. Santiago de Cali. 2015.

[28] Faustino Núñez Batalla, Rocío González Márquez, M. Belén Peláez González, Irene González Laborda, María Fernández Fernández y Marta Morato Galán. «Análisis acústico de la voz mediante el programa Praat: estudio comparativo con el programa Dr. Speech.» En: *Realizado en conjunto entre la unidad de Servicio de Otorrinolaringología del Hospital Universitario Central de Asturias y la Facultad de Psicología de la Universidad de Oviedo. 26 de marzo de 2014. Oviedo, España.* (2014).

[29] David Weenik. Paul Boersma. *Praat: doing Phonetics by Computer.* Instituto de Ciencia Fonéticas de la Universidad de ámsterdam. 1992. URL: link:http://www.fon.hum.uva.nl/praat/..

[30] Vincent van Heuven. Paul Boersma. «Speak and unSpeak with PRAAT.» En: *Glot International*, 5.9/10 (2001).

[31] Franklin Barzola Tobar y Roberto Cabrera Velasco. *Comparación entre compresión de audio en diferentes formatos de imagenes equivalentes y el formato de compresión mp3.* Escuela Superior Politecnica del Litoral. Guayaquil, Ecuador. 2009.

[32] Gabriel Valverde Castilla. «Deep Learning». En: *VI jornada usuarios* (2014).

Anexo A

Manual de Usuario.

Anexo B

Tablas de Cálculos de Parámetros de la voz.

Anexo C

Glosario de Términos

Anexo D

Algoritmo del diseño de la interfaz gráfica de la herramienta.

Anexo E

Herramienta Diseñada para el reconocimiento de personas por medio de la voz.