Sistema de identificación biométrico basado en reconocimiento de voz mediante coeficientes cepstrales para detección de *spoofing* en llamadas telefónicas

Alberto Karel Guzman Zumaeta
20160663@aloe.ulima.edu.pe
https://orcid.org/0009-0009-4298-0215
Universidad de Lima, Perú

Recibido: 28 de agosto del 2023 / Aceptado: 31 de octubre del 2023
doi: https://doi.org/10.26439/interfases2023.n018.6625

RESUMEN. Los delitos informáticos en los sistemas telemáticos de las empresas perjudican a la sociedad porque ocasionan un clima de incertidumbre en los clientes, quienes tienen la percepción de que el sistema informático encargado de gestionar el servicio o producto a consumir no es tan seguro como para confiar su dinero o hacer transacciones de forma remota. Uno de los delitos informáticos más extendidos es el *spoofing*, el cual consiste en suplantar la identidad de una persona o una entidad. El objetivo es implementar un sistema de reconocimiento de voz, como una aplicación móvil, para que permita identificar casos de suplantación de voz por *spoofing* mediante llamadas telefónicas. Para este propósito, se utilizaron los coeficientes cepstrales en la escala de Mel (MFCC) como clasificadores para la limpieza de anomalías en los audios, así como redes neuronales de retro propagación para el sistema de identificación de usuarios que trabaja en conjunto dentro de un aplicativo móvil. En las pruebas realizadas, el sistema propuesto tuvo una tasa de éxito del 83,5 %. Para diseñar las 20 entidades necesarias en el trabajo de investigación, se utilizó un conjunto de 2000 audios. Estos audios se dividieron en grupos de 100, donde cada grupo correspondía a un autor diferente. Es decir, se contó con 100 audios de voz provenientes de cada uno de los 20 autores distintos, lo que permitió crear y probar las entidades del sistema de manera representativa y diversa. Se concluye que el sistema es exitoso en el ámbito de seguridad, ya que tiene una tasa de aceptación óptima y un sistema robusto para los diferentes tipos de *spoofing* que se ha logrado recopilar en este trabajo de investigación.

PALABRAS CLAVE: biometría de voz / coeficientes cepstrales en las frecuencias de Mel / prevención de *spoofing*

BIOMETRIC IDENTIFICATION SYSTEM BASD ON VOICE RECOGNITION USING CEPSTRAL COEFFICIENTS FOR SPOOFING DETECTION IN TELEPHONE CALLS

ABSTRACT. Computer crimes in the telematic systems of company's harm society because they cause a climate of uncertainty in customers, who have the perception that the computer system, in charge of managing the service or product to be consumed, is not so secure as to trust its money or make transactions remotely. One of the most widespread computer crimes is Spoofing, which consists of impersonating the identity of a person or entity. The objective is to implement a voice recognition system as a mobile application to identify cases of voice impersonation by Spoofing through telephone calls. For this purpose, the Mel scale cepstral coefficients (MFCC) were used as a classifier for cleaning anomalies in the audios, as well as back-propagation neural networks for the user identification system that works together within a mobile application. In the tests carried out, the proposed system had a success rate of 83.5% with 20 entities that were designed by the author out of a total of 2000 audios with 100 corresponding audios from each author for the respective research work. It is concluded that the system is successful in the field of security since it has an optimal acceptance rate and must have a robust system for the different types of Spoofing that has been collected in this research work.

KEYWORDS: voice biometrics / Mel frequency cepstral coefficients / spoofing prevention.

1. INTRODUCCIÓN

El delito informático llamado *spoofing* hace referencia al uso de técnicas a través de las cuales un atacante, generalmente con usos maliciosos o de investigación, se hace pasar por una entidad distinta, a través de la falsificación de los datos en comunicación (Fuertes et al., 2010). A nivel mundial, según la cadena de televisión BBC News, (Zorro, 2022), en los 12 meses previos al mes de agosto de 2022, se realizaron alrededor de 10 millones de llamadas fraudulentas en todo el mundo utilizando el servicio de ISpoof.cc, donde los atacantes se hacían pasar por corporaciones confiables para obtener información confidencial. Europol, agencia de la unión europea para la cooperación policial, estima una pérdida financiera de más de 100 millones de libras esterlinas.

Las llamadas telefónicas afectadas con *spoofing* generan anualmente quince millones de dólares americanos en pérdidas (Mustafa et al., 2014), lo cual trae consigo la pérdida de reputación y el descontento de usuarios de diferentes empresas del mundo. Estos problemas de fraudes podrían ser solucionados con métodos de autentificación del usuario.

Según Cabeza (2023), a nivel nacional, en el primer trimestre de este año se tiene -igual que el año pasado - un total de 2445 casos que fueron víctimas de delitos cibernéticos. El 2022 cerró con 3946 delitos informáticos denunciados a Divindat (División de investigación de Delitos de Alta Tecnología de la Policía Nacional del Perú - PNP). El año pasado se obtuvo un récord de 229 detenidos y se espera mantener o elevar esta cifra.

Dado el avance tecnológico de las modalidades para transacciones comerciales existe una mayor incidencia en el *spoofing,* práctica que utiliza diversos mecanismos para vulnerar la seguridad de los sistemas.

El presente trabajo de investigación tiene por objetivo ofrecer una alternativa de solución a la suplantación y plantea para ello la implementación de un sistema robusto de reconocimiento de voz. Con este sistema se evitan/ evitarían problemas de identificación de personas, tomando en cuenta que la suplantación de voz se ve agravada con el uso de diferentes técnicas como grabaciones de voz, uso de sintetizadores de voz, imitación de voz, entre otras posibilidades de fraude. Martínez Mascorro y Aguilar Torres (2013) realizaron una comparación de diferentes técnicas de reconocimiento de voz y llegaron a la conclusión de que la más efectiva es el MFCC, con un 97,77 % de tasa efectiva de reconocimiento de voz.

2. ESTADO DEL ARTE

En esta sección describimos trabajos que se orientan a la identificación por voz y las diferentes complicaciones que pueden ocurrir en sus aplicaciones. Para el caso de las llamadas telefónicas donde se utiliza un sisteidentificación de voz, se tiene que obtener el audio que transmite la voz del usuario. Por ello, el sistema que se analiza se centra

en reconocer el audio que identifica al usuario mediante diversos estudios de reconocimientos de audios.

2.1 Sistemas de identificación de voz con el fin de prevenir el spoofing.

Este segmento trata de enfocarse en posibles soluciones. Al tratar de contrarrestar el delito informático *Spoofing* por medio de grabaciones de audios, se tratará de observar cómo los veinte diferentes patrones de la biometría de voz influyen respecto de las grabaciones de audios. A su vez, se tratará de relacionar los diferentes tipos de algoritmos ocultos de Márkov y analizar cómo influyen en la detección de audios en las grabaciones.

Singh et al. (2016) mencionan que disfrazar la voz implica hacerse pasar por otra persona, ocultar la identidad del hablante o hacer ambas cosas a la vez. Este proceso puede llevar a la identificación de dos voces: una correspondería al autor original de la voz disfrazada, mientras que la otra sería la persona a la que se intenta imitar o reemplazar.. La metodología utilizó el conocimiento de sonidos usados típicamente para hallar la variante voluntaria más susceptible por parte del hablante. Como también se trata de analizar la voz de ambos objetivos, se estudian suplantaciones de voz realizadas por un imitador experto, centrándose específicamente en mediciones y averiguando el tipo de alcance y de manipulación que realiza el experto a nivel de fonemas individuales Los resultados producirán un patrón de referencia, lo que permitirá identificar a los imitadores de voz expertos como una categoría distinta. Este patrón será considerado como un estándar de excelencia en su nivel., En conclusión, los autores pudieron detectar tanto la voz que se quiere imitar (ya que está guardada en la base de datos del sistema) y la voz del imitador, ya que se relaciona con la voz que dice ser. Esto generará un sistema de autentificación muy peculiar con alta aceptación para futuros trabajos.

Le et al. (2019), tienen como objetivo separar las voces naturales humanas de las voces reproducidas por cualquier tipo de dispositivo de audio en el contexto de una interacción en la interfaz de un usuario con su voz. La metodología utilizada implica recopilar información de diversos conjuntos de datos del mundo real para construir modelos predictivos basados en *Deep Neural Network* (DNN), los cuales se han desarrollado utilizando diferentes combinaciones de funciones de audio. Los resultados confirman la viabilidad de la tarea: la combinación de las incrustaciones de audio extraídas de la red SoundNet y VGGish produce una precisión de su clasificación de aproximadamente un 90 %. Se concluye que es un conjunto de datos a gran escala en condiciones bien controladas -para adaptarse mejor al entorno del hogar- y de esta forma permitir entrenar la predicción basada en modelos DNN con una tasa de acierto de 90,43 %.

Kinnunen et al. (2012) tienen como objetivo identificar la importancia de las vulnerabilidades que deben tomar en cuenta los sistemas frente a los ataques de *spoofing*

por medio de voz. La metodología que implementaron fue convertir automáticamente las expresiones propias en sonidos, como si fueran expresados por otro hablante. Se implementó un sistema de conversión de voz de dos tipos de características: coeficientes de 30 mel-cepstrales (MCEP) y pares de espectro de líneas (LSP). Los resultados demostraron que aumentó la tasa de falsos positivos del 3,24 % al 17,33 %. Se concluye que un oyente podría juzgar las voces convertidas, mas no el sistema; por lo tanto, se necesitan soluciones para la discriminación del habla natural y no natural, a fin de que puedan ser reconocidas por alguna alteración o no.

Alegre et al. (2013) tienen como objetivo presentar una contramedida para el *spoofing* basada en análisis de señales de voz y usando patrones binarios locales seguidos de una clase única de enfoque de clasificación. La metodología busca capturar diferencias en la textura espectro-temporal del hablante genuino y del discurso falso, y se realiza sobre tres enfoques diferentes de suplantación: conversión de voz, síntesis de voz y señales artificiales. Después de este proceso, el programa muestra patrones binarios que ofrecen precisión en la detección de voces alteradas por un programa. Los resultados dieron una detección confiable contra ataques de *spoofing* por medio de voz sintetizada, señales artificiales y ataques que no están optimizados en su totalidad. Se concluye que existe una necesidad de desarrollar contramedidas para la suplantación, ya que los sistemas biométricos son susceptibles a una amplia gama de ataques que podrían afectar significativamente la seguridad de las empresas.

2.2 Técnicas de autentificación basados en MFCC.

Este segmento se enfoca en sistemas que implementa el algoritmo MFCC, el cual se aplica para diferentes sistemas de autentificación de usuarios. Este algoritmo pasa por diferentes procesos para limpiar los diferentes ruidos que obtiene al momento de grabar el audio emitido, para así tener un reconocimiento óptimo de las señales de voz.

Reconocimiento de locutor basado en MFCC y Redes Neuronales BP

Wang y Lawlor (2017) tienen como objetivo realizar un sistema de identificación de usuario con MFCC (uno de los métodos más exitosos debido a que generalmente el algoritmo se basa en un sistema auditivo humano), y las redes neuronales, porque solucionan las complicaciones que existen en algunas regiones al reconocer las frecuencias, lo que llevaría una menor eficacia en el sistema. La metodología que usan es un método de reconocimiento de hablantes basado en MFCC y redes neuronales de retro propagación. Los resultados demostraron que el reconocimiento es exitoso cuando el número de hablantes cuestionables es menor. Cuando el número de hablantes aumenta, el reconocimiento disminuye. Se concluye, por tanto, que el sistema es factible cuando la cantidad de hablantes no es muy grande. Se hace necesario mejorar la capacitación de las redes neuronales, que están en constante desarrollo y tienen un papel importante en el campo de reconocimiento de voces.

3. MARCO TEÓRICO

3.1. MFCC:

Es una técnica muy usada para la autentificación de autores de voz. Devuelve como coeficientes a los representantes de audios. Este algoritmo, al momento de implementarse, extrae características de la señal de audio para una identificación de usuario competente, para que el proceso sea de muy alta calidad, para que entre a ciertas fases de optimización, limpiado de ondas, división de segmentos, entre otras. Esto genera que se elimine la información que no es relevante al momento de captar la señal de audio. Por ejemplo, la acentuación, tono, volumen, emociones, ruidos extraños de anomalías, entre otros.

3.1.1 Definición de MFCC:

Martínez Mascorro y Aguilar Torres (2013), definen los coeficientes cepstrales en la escala de Mel (MFCC) como una representación de la amplitud del espectro del habla de manera compacta. Por esta razón, se ha vuelto una técnica muy usada de extracción de características. El MFCC es una técnica de parametrización de la voz, cuyo objetivo es tener una representación apropiada, robusta y compacta para obtener un modelo estadístico de un grado alto de precisión.

Figura 1

Imagen sacada del artículo original, indicando el proceso de obtención de los coeficientes MFCC

Nota. De Martínez et al. (2013).

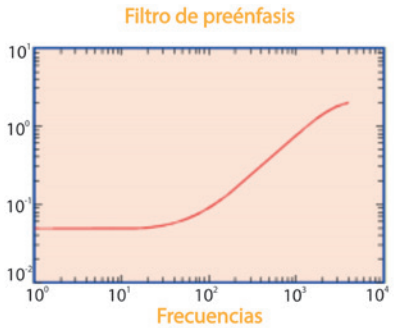
3.1.2 Cálculo de MFCC

Rueda (2011) señala que, en primer lugar, se aplica un filtro de pre-énfasis a la señal para contrarrestar la pendiente espectral negativa, que describe el grado de decaimiento de la amplitud espectral. Se obtiene directamente de realizar una regresión lineal, obteniendo así la pendiente de la recta para suavizar el espectro, provocando una reducción en

las inestabilidades de cálculo relacionadas con las operaciones aritméticas de precisión finita. De esta forma ayuda a las etapas posteriores de análisis a modelar los aspectos importantes del espectro de la voz. Por ejemplo, en la Figura 2 se muestra el dominio frecuencial del filtro de pre-énfasis.

Figura 2

Frecuencia de filtro de pre-énfasis

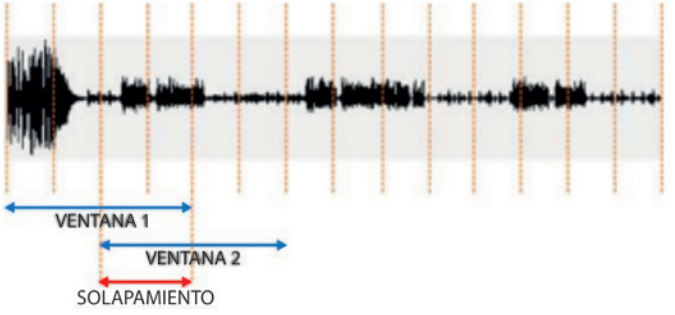


Nota. De Rueda (2011)

Las señales de voz, al ser un proceso aleatorio y no estacionario, provocan una dificultad al momento de analizarlas. Para resolver esta dificultad se dividen las tramas en cortos plazos de tiempo para ser más específicos en milisegundos (ms) y esto convierte a las señales en casi-estacionarias. En la Figura 3 se observa cómo se dividen las tramas en segmentos, los cuales, por lo general, se dividen en 20 milisegundos, ya que el sistema de señales de voz se vuelve muy pesado para procesar la información. Por otro lado, se lleva a cabo otro proceso para mantener la continuidad de la información de la señal: se muestran bloques solapados, de tal manera que los eventos de transición no se pierden

Figura 3

División de tramas de las señales de voz



Nota. De Morejón S. (2011)

Luego las tramas se dividen y se aplica una función de ventana, que sirve para dar una mejor acentuación a la parte central de la trama, eliminando los bordes de la señal para su análisis. Existen diferentes tipos de ventaneo que son muy utilizados. Entre ellos están: *Rectangular*, *Hanning*, *Hamming*, *Barlett* y *Blackman*. Se utilizará la función *Hamming*, ya que se adecua en esta sección para limpiar el audio.

Se usará *Hamming*, que tiene la siguiente denotación:

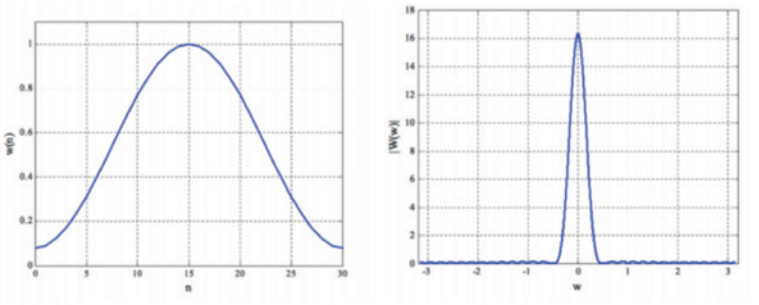
$$w(n) = \frac{27}{50} - \frac{23}{50} \cdot cos\ cos\ \left(\frac{2\pi n}{N}\right) \quad (1)$$

Donde N es el largo de cada cuadro o segmento de análisis

Donde n es el número de muestras

Figura 4

Aplicación de cómo influye la ventana de Hamming



Nota. De Toro Cerón (2018)

Una vez obtenida la Transformada Discreta de Fourier (DFT) de cada una de las tramas, se aprovecha la amplitud del espectro. Esto permite una optimización de la descomposición de la transformada a unas más simples y se transforma en valores de ceros y unos. Luego, las transformadas más simples se agrupan en otras de nivel superior y nuevamente tienen que pasar por el proceso. Así sucesivamente, hasta llegar al nivel más alto. Al final del proceso, se requiere reorganizar los resultados obtenidos para transferir la información al dominio de Mel mediante el banco de filtros, lo que ayuda a aclarar la incertidumbre sobre este dominio.

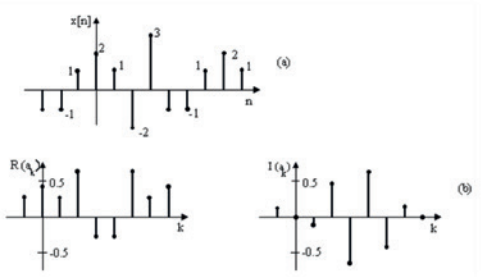
En la etapa de la transformada discreta de Fourier se usará la siguiente denotación:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi kn}{N}} \quad (2)$$

Dada una señal en tiempo discreto x(n) con N muestras su transformada X(k) está dada por esta ecuación (La ecuación se explica a más detalle en la página 4 ecuación número (1)

Figura 5

Representación gráfica de una señal discreta en el tiempo

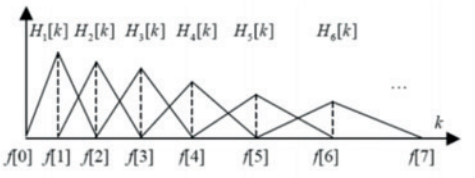


Nota. De Fraga (2001)

El banco de filtros de Mel es altamente recomendable ya que proporciona una expresión matemática que permite calcular el valor de cada filtro. Aunque en la Figura 6 se muestran los filtros con formas triangulares, es importante tener en cuenta que estos filtros pueden adoptar otras formas según los tipos de ventanas mencionados previamente.

Figura 6

Grafica del banco de filtros de Mel



Nota. De Aguilar, (2020)

En la escala Mel, el sistema está en función del comportamiento psico-acústico humano. Se mapea la frecuencia actual al *pitch* que percibe. Esta escala es lineal por debajo de 1 kHz y logarítmica por encima del umbral.

Luego, es importante tener en cuenta que los MFCC constituyen un esquema de análisis segmental en el que se recopilan los coeficientes de energía del espectro de un banco de filtros cuyas frecuencias centrales están distribuidas de manera uniforme en la escala Mel. El comportamiento del sistema psico-acústico humano se representa con la siguiente ecuación:

Donde:

$$Mel(f) = 2595 * (1 + \frac{f}{700}) \quad (3)$$

f corresponde con la frecuencia representada en el eje de escala lineal.

Una vez que la envolvente del espectro de la señal de voz es multiplicada por el banco, se calcula la energía correspondiente en cada uno de los filtros: Como las tramas están centradas la energía de estas se calcula de la siguiente manera:

$$s_t[m] = ln\ ln\left(\sum_{n=0}^{N-1} |x_t[n]|^2 H_m[n]\right), 0 \leq m \leq M \quad (4)$$

Donde:

- $x_t\,[n]$, es la transformada discreta de Fourier (DFT) de la t-ésima trama de la señal de voz de entrada.
- $H_m\,[n]$, es la respuesta en frecuencia de la n-esima del filtro de banco de audio.
- n, es el tamaño de la ventana de la transformación.
- M, es el número total de filtros.

Después de obtener la energía, es necesario calcular su logaritmo. Este paso lleva la energía al dominio de la potencia espectral logarítmica. Una consecuencia de trabajar en este dominio es que los filtros de las bandas adyacentes generan coeficientes espectrales que están altamente correlacionados entre sí, debido a su alto grado de correlación.

Una vez obtenido el logaritmo de la señal, finalmente se aplica la transformada de coseno discreta (DCT), que se utiliza en diversas aplicaciones de compresión de datos debido a una propiedad denominada "compactación de la energía". Este proceso tiende a concentrar una cantidad considerable de información de la señal en los coeficientes de baja frecuencia, por lo que se necesita un menor número de coeficientes para representarla. Este proceso se utiliza para eliminar la dependencia o correlación estadística.

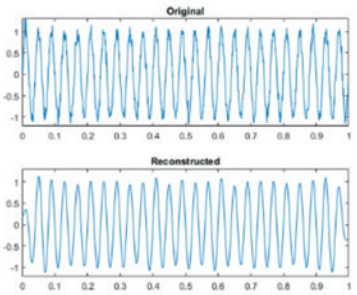
Shukla et al. (2019) indicaron que f (0, 1, ..., N-1) denota una secuencia de datos discretos de señal f, n indica el número de muestras, F (0,1, ..., N-1) denota coeficientes de transformadas de coseno.

$$F(u) = \propto (u) \sum_{i=0}^{N-1} cos\ cos\left[\frac{\pi . u}{2N}(2i+1)\right] f(i) \quad (5)$$

En la Figura 7 se muestra la reconstrucción de la onda que empieza desde la frecuencia 0. Por otro lado, se observa que las ondas se vuelven más limpias, lo que resulta en una mejor ondulación y una percepción mejorada al momento de comparar en el MFCC. Los resultados de este vector generan la cantidad de coeficientes deseados por trama, concluyendo así el proceso MFCC.

Figura 7

Representación de cómo influye la transformada de coseno discreta en MATLAB



3.2. Api Twilio

Twilio es una plataforma de Comunicaciones como Servicio (CPaaS) que permite a los desarrolladores de software realizar y recibir llamadas telefónicas, así como enviar y recibir mensajes de texto, entre otras funciones de comunicación, como mensajería instantánea en redes sociales. Normalmente, las llamadas salientes de Twilio se realizan mediante una biblioteca auxiliar. Cuando se necesita realizar una llamada, la aplicación envía una solicitud de publicación para iniciar la llamada dentro de Twilio, que a su vez realiza la llamada al teléfono celular. Una vez que la llamada está conectada, Twilio solicita instrucciones a la aplicación mediante TwiML (lenguaje de marcado de Twilio). Por otro lado, Twilio recibe llamadas entrantes a través de solicitudes web desde la aplicación, luego envía una solicitud HTTP a su aplicación para recibir instrucciones (TwiML) y finalmente reproduce un archivo de audio pregrabado.

Figura 8

Proceso de Twilio básico en una llamada telefónica

Nota. https://en.wikipedia.org/wiki/Twilio

4. METODOLOGÍA

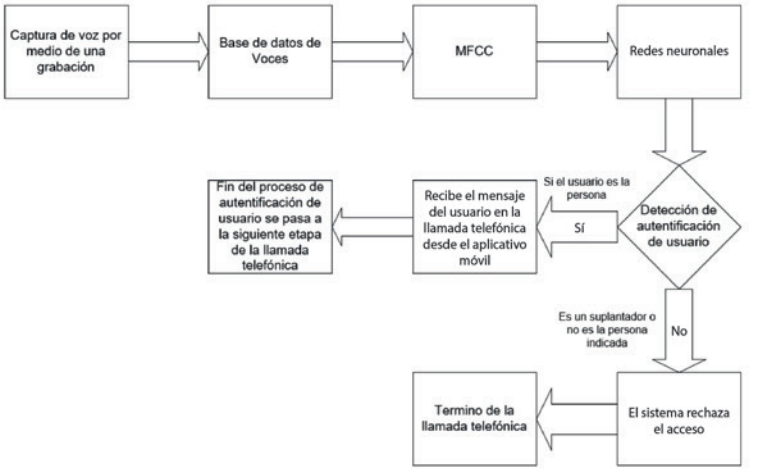
4.1 Propuesta de solución:

El objetivo de este trabajo de investigación es brindar autentificación instantánea de personas en el servicio de llamadas telefónicas por medio de la biometría de voz para

hacer frente a los ataques de *Spoofing*. El sistema se centra en la etapa de validación de identidad del usuario. De esta manera, se presentará un diagrama de bloques de esta propuesta, añadiendo un aplicativo que sirva como un identificador del suplantador.

Figura 9

Metodología de la propuesta de solución de la investigación en diagrama de bloques



El proceso comienza con la captura de voz del usuario mediante una grabación, que luego se almacena en una base de datos creada para la investigación. Las grabaciones, de tres segundos de duración, se someten al algoritmo MFCC para eliminar ruidos. Después de esta limpieza, se realiza una etapa de retroalimentación y, finalmente, las grabaciones se procesan a través de un segmento de redes neuronales.

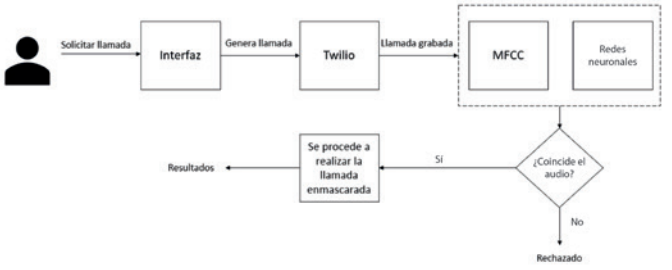
Se utilizarán redes neuronales en *Anaconda Navigator*, en el lenguaje de programación llamado Python, para autentificar al usuario una vez entrenadas las redes neuronales. Las redes neuronales se usan porque el algoritmo MFCC no garantiza al 100% la autenticidad y tampoco detecta si existe un intento de suplantación de la persona. El uso de redes neuronales servirá al segmento de detección de autentificación de usuario y permitirá que se tome una decisión: si es un suplantador de voz o es una persona que no dice ser, el sistema rechazará el acceso y dará por terminada la llamada telefónica; en cambio, si el sistema detecta la autenticidad del usuario, dará permiso para que se pueda comunicar con el receptor de la llamada telefónica y, en este caso, el proceso terminará cuando concluya la llamada telefónica. Una vez que se obtenga el sistema, se procederá a implementarlo en una aplicación móvil para que se trabaje en el ámbito de telecomunicaciones, específicamente en el área de llamadas telefónicas.

El aplicativo tendrá una interfaz gráfica que solicitará el nombre del usuario, el teléfono celular que posee y el teléfono celular al que desea llamar. Una vez completados

estos pasos, se procederá a realizar una primera llamada con Twilio, la cual será guardada. Una vez guardada la llamada, se procederá a realizar el mismo proceso para el audio generado, debido a que se tendrá que comparar este audio con los datos entrenados por las redes neuronales. Si el audio no coincide con una de las entidades de los datos, se procede a rechazar el proceso generando un error y avisando que no es la persona que dice ser. Si el audio coincide, se procederá a ejecutar una segunda llamada: Twilio actuará como intermediario en una llamada enmascarada entre el receptor y el emisor, proporcionando un sistema que previene los ataques de suplantación de identidad (*spoofing*).

Figura 10

Diagrama de bloques del aplicativo que se realizará en el sistema propuesto



4.2 Experimentación:

Luego de tener una base de datos de audios óptima para su respectivo uso, se tendrá como objetivo resolver la suplantación de las personas que imitan la voz del usuario requerido. Por ello, al sistema se le añadirá una aplicación móvil para que funcione con llamadas telefónicas y detecte al usuario en tiempo real. Se realizó la experimentación en Anaconda, y se tuvo como plantilla un repositorio en GitHub que utilizaba el algoritmo MFCC, modificando lo que se desea implementar con datos reales generados por el autor de esta investigación. Se tuvo que usar un convertidor estándar de los archivos .M4A, que se guardaban directamente de la grabadora de voz, a .WAV. Esto sucede porque los convertidores de archivos tienen diferentes velocidades de bits.

En algunos casos el algoritmo no detectaba bien la voz del usuario, por lo que se tenía que volver a grabar al usuario y volver a realizar la conversión necesaria. El sistema deberá tener datos estandarizados (formato de audio, audios de tres segundos, que sea una voz directa, entre otros), para que funcione correctamente la comparación.

Dado que la base de datos está estandarizada, es complicado usar una grabación ya que el sistema tendría diferentes características. Por ejemplo, existen diferentes tipos de grabadoras y micrófonos que estabilizan la voz, ya sea en un audio analógico o digital. Como el sistema tendrá en consideración la velocidad de bits del audio emitido, si no se obtiene la velocidad estandarizada, el sistema detectará, por defecto, que se trata de una grabación.

Por otro lado, se utiliza un promedio de la amplitud máxima de las ondas sonoras de cada longitud de onda de cada audio establecido por usuario. Esto da como resultado un arreglo de promedios, con lo cual el algoritmo detecta la aproximación de la voz del usuario. La ventaja de usar esta implementación es que es una herramienta de gran utilidad al momento de extraer parámetros de una señal de voz. Por otro lado, según el artículo de los autores Wang y Lawlor (2017), las redes neuronales dan flexibilidad en el sistema y una capacidad de procesar información incierta para luego verificar con más claridad al usuario indicado; por ello se utilizan complementariamente: para tener un sistema más robusto de verificación de usuario.

Una vez obtenida la base de datos necesaria, se procederá a utilizar *Anaconda Navigator*, que es una interfaz gráfica de usuario GUI con un potencial enorme, pues puede gestionar de manera avanzada paquetes relaciones a ciencia de datos con Python. Esto permitirá facilitar la implementación de redes neuronales para su correcta retroalimentación. Después de esto, se intentará comprobar cuántas veces el sistema consigue detectar si existe una suplantación o no.

5 PRUEBA DE CONCEPTO

Se procederá a contratar el servicio de API de Twilio, que permitirá agregar la autentificación. Twilio es una plataforma de comunicaciones en la nube, que más de dos millones de desarrolladores utilizan para la participación de sus clientes. El servicio que ellos brindan incluye SMS, mensajería de redes sociales, llamadas telefónicas, entre otros.

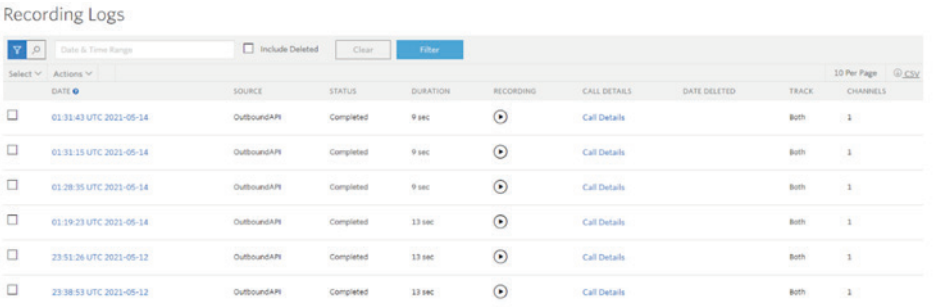
Figura 11

Interfaz gráfica del aplicativo donde se iniciará las llamadas

Se diseñó una interfaz gráfica en Qt Designer. El programa requiere tres datos: el nombre de la persona, el número telefónico del usuario y, finalmente, el número telefónico de la persona que quiere realizar la llamada. Una vez realizada la llamada, Twilio procede a grabarla y esta grabación servirá para comprobar si la persona que está solicitando la llamada es la que dice ser. Si la grabación no coincide con los audios entrenados, se rechazará y terminará el proceso. Por otro lado, si el audio coincide, Twilio hará una llamada enmascarada tanto al receptor como al emisor, para que puedan comunicarse sincrónicamente.

Figura 12

Lista de grabaciones hechas para las pruebas solicitadas



Se tendrá un repositorio de las llamadas grabadas, el cual indica su respectivo estado: si resultó exitosa, si tuvo una complicación al momento de la grabación o en la misma llamada. Por otro lado, en cada llamada grabada se encripta en formato json un archivo .wav binario, que se deberá desencriptar para proceder a la etapa en la que se tuvieron que realizar todos los audios. La grabación procederá a realizar los mismos pasos previos que tuvieron los audios anteriores; es decir, los coeficientes cepstrales de Mel y las redes neuronales.

5.1. Resultados

La propuesta final de la investigación es un sistema automático que espera la aceptación o rechazo de la llamada entrante. Luego, se descarga manualmente el audio grabado, ya que el sistema envía un archivo .json encriptado junto con un archivo binario .wav a la galería de llamadas grabadas. El proceso de desencriptación se llevará a cabo durante la fase de ajuste del sistema. Una vez descargado el audio, se realizan los procedimientos previos por los que todos los audios pasaron en el trabajo de investigación.

Se realizaron audios de prueba utilizando grabaciones de llamadas telefónicas, principalmente de los 400 audios anteriores. Dado que no se dispone de todas las entidades, se reproducirá y grabará el audio en la llamada entrante. El objetivo es comparar estos resultados con los obtenidos previamente. Se llevará a cabo para verificar las

métricas del sistema. Además, se está realizando una llamada enmascarada para que, al aceptarse, el API de Twilio actúe como intercomunicador entre el receptor y el emisor. A continuación se presentan los resultados.

Figura 13

Registro de todas las pruebas realizadas

Resultados	
1	[80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80 80]
2	[20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20 20]
3	0.835
4	[[20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
	[0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
	[0 0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
	[0 0 0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
	[0 0 1 0 18 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0]
	[0 0 0 0 0 20 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
	[0 0 0 0 0 0 20 0 0 0 0 0 0 0 0 0 0 0 0 0]
	[0 0 0 0 0 0 0 18 0 0 0 0 0 0 1 0 0 1 0 0]
	[0 0 0 0 0 0 0 0 19 0 0 0 0 0 0 0 0 1 0 0]
	[0 0 0 0 0 0 0 0 0 19 0 0 0 0 0 0 1 0 0 0]
	[0 0 1 0 0 0 1 0 16 0 0 0 0 2 0 0 0 0 0 0]
	[0 0 0 0 0 0 1 0 0 0 0 19 0 0 0 0 0 0 0 0]
	[0 0 1 0 0 0 0 0 0 0 5 0 2 12 0 0 0 0 0 0]
	[0 0 5 1 0 0 1 0 5 0 1 0 0 7 0 0 0 0 0 0]
	[0 0 0 0 0 0 0 0 1 0 0 0 0 1 17 0 0 1 0 0]
	[0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 17 0 0 0 0]
	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 20 0 0 0]
	[0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 19 0 0]
	[0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 19 0]
	[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 20]]
5	(0.8415302586538005, 0.835, 0.8132053971556372, None)
6	(0.835, 0.835, 0.835, None)
7	(0.8415302586538005, 0.835, 0.8132053971556371, None)

1. Se obtuvo un registro de los audios entrenados y con ellos se creó una lista con el número de audios utilizados (en total, ochenta por persona). La lista tendrá veinte números, dado que existen veinte entidades en el sistema.
2. Luego, se obtuvo un registro de los audios de pruebas y con ellos se creó una lista con el número de audios utilizados (en total 20 por persona). La lista tendrá veinte números, dado que existen veinte entidades en el sistema.
3. Se obtuvo la tasa de aceptación de las pruebas de testeo con un 83,5 %.
4. Se creó una matriz de confusión, donde están representadas las métricas que se obtienen al momento de ser evaluadas entre ellas. Las diagonales representan la cantidad de aciertos (veinte será el número mayor, ya que solo se obtienen veinte muestras de cada una).

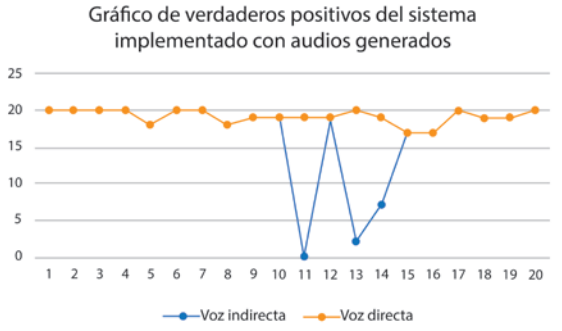
5. Se obtuvo un soporte, en el cual se calculan las métricas para cada etiqueta y se encuentra la media no ponderada. No se tiene en cuenta el desequilibrio de las etiquetas.
6. Se obtuvo un soporte, en el cual se calculan las métricas globalmente, contando el total de verdaderos positivos, falsos negativos y falsos positivos.
7. Se obtuvo un soporte, en el cual se calculan las métricas para cada etiqueta y encuentra el promedio ponderado por soporte (el número de instancias reales para cada etiqueta).

5.2 Analizando los resultados

Se da una variación en los resultados, dado que no se está realizando con voz directa (es decir, entre usuario y micrófono directamente), si no que se está reproduciendo el audio de prueba grabado de voz directa en un equipo de sonido Sony MHC-V72D interconectado con un Roland DJ-202 de la marca Serato. Se escogieron equipos de alta calidad para que puedan reproducir el audio sin ninguna alteración en el proceso. El micrófono de celular que se utilizó fue de un Samsung Galaxy S8 Plus, que recibía la llamada para luego grabarla y tenerla en el repositorio. Para explicar la variación de los resultados se presentará un gráfico donde se representan los verdaderos positivos del sistema en voz directa y voz indirecta, respectivamente.

Figura 14

Comparación de verdaderos positivos en pruebas de voz directa e indirecta



Nota. Se comparan los verdaderos positivos obtenidos en pruebas de voz directa e indirecta, representando la cantidad en el eje vertical y los individuos en el eje horizontal.

De la Figura 14 podemos concluir que, a pesar de ser audios idénticos, pero adaptados al sistema en diferentes condiciones (voz directa e indirecta), existen algunas entidades que llegaron a bajar su reconocimiento. Incluso existe una entidad (para ser más específicos, la entidad once) a la que el sistema no le reconoce ni un solo audio. Estos cambios surgen por la grabación en plena llamada telefónica, ya que (aun implementando el

algoritmo MFCC que se utiliza para ser lo más limpio posible) existen algunas anomalías que influyen y perjudican este proceso: de tener 95,75 %, con el aplicativo móvil se obtiene un 83,5 % de tasa de aceptación.

6 DISCUSIÓN

En el presente trabajo de investigación se utilizó como clasificador la matriz MFCC. Una vez que se obtiene los coeficientes cepstrales, se pasa a una red neuronal para la obtención de los resultados. Este proceso separa los audios de voz directa e indirecta. Así mismo, existen cambios en la tasa de aceptación, dado que el audio emitido por un reproductor de sonido no será igual que emitido por el mismo usuario, a pesar de que fueron los mismos audios de pruebas que se tenían en voz directa. Actualmente, el sistema que utiliza el aplicativo móvil ha logrado identificar a la persona con una tasa de aprobación de 83,5 % con respecto a los 1 600 audios que se obtuvieron para entrenar la red neuronal y los otros cuatrocientos audios grabados por medio de llamadas telefónicas que se obtuvieron para testear su efectividad entre las redes neuronales.

Existen experiencias previas que utilizan el algoritmo MFCC junto a las redes neuronales de retro propagación, como indica el artículo de Wang y Lawlor (2017); sin embargo, ningún trabajo de investigación implementa una aplicación móvil para tratar de combatir el delito informático llamado *spoofing*. El valor agregado del sistema propuesto es la capacidad de detectarlo, en tiempo real, para poder mitigar el riesgo de este delito a través de una aplicación móvil que pasa a una etapa de retroalimentación por medio de las redes neuronales.

El sistema propuesto con voz directa tiene un reconocimiento de 95,75 % y con voz indirecta un 83,5 %. Este es un resultado esperable, ya que, al utilizar voces grabadas de una llamada telefónica, se pierde calidad y con ello disminuye la posibilidad de su correcta detección.

La implementación del algoritmo MFCC se complementa con las redes neuronales, debido a que limpia los ruidos externos y también extrae las características de los datos para que posteriormente se entrenen con la red neuronal. Con ello se logra una exitosa tasa de aceptación del 83,5 % con voz indirecta. Adicionalmente, el sistema propuesto en esta investigación tiene como valor agregado aumentar el grado de seguridad ante el riesgo de suplantación de identidades por medio de llamadas telefónicas, debido a que el sistema brinda reconocimiento en tiempo real del hablante.

7 CONCLUSIONES

Debido a la fuente de origen, los audios emitidos pueden ser considerablemente afectados en cuanto a su tasa de aceptación. El estudio comprobó que la captación de un audio

por medio de voz directa tiene un mejor desempeño al momento de verificar la identidad de un usuario, si se compara con el audio emitido por una voz indirecta. Estos detalles en la captación de audios influyen mucho, debido a que un simple detalle puede alterar la tasa de aceptación (por ejemplo, la distancia entre el micrófono y el usuario, si es una grabación de voz o si existe ruido en el entorno donde se graba la voz, entre otros).

De los resultados obtenidos se desprende que mientras mayor sea la cantidad de personas o entidades existentes en la base de datos, será más difícil identificar de manera correcta a una persona. Este resultado coincide con la investigación de Wang y Lawlor (2017), quienes también concluyeron que mientras más entidades existan en el sistema, será más difícil reconocer a la persona, porque las redes neuronales tendrán que verificar con más usuarios y esto resulta una desventaja para el sistema propuesto. Una forma de contrarrestar esto es aumentar la cantidad de audios para que las redes neuronales tengan un mejor entrenamiento. En el presente trabajo de investigación solo se cuenta con 100 audios por cada entidad, pero si se aumentara la cantidad, se produciría un aumento en su tasa de exactitud.

Se obtienen resultados significativos al considerar los diversos cambios existentes. Si se busca obtener un rendimiento óptimo, es crucial implementar redes neuronales en el sistema. Debido a que el sistema opera en tiempo real, es necesario realizar un seguimiento continuo durante el proceso. Estos procesos mejorarán con el tiempo, ya que el sistema se volverá más complejo a medida que aumente el número de usuarios que requieran el servicio de autenticación, dependiendo de la empresa que adopte este sistema propuesto en la investigación.

REFERENCIAS

Alegre, F., Amehraye, A., & Evans, N. (2013). A one-class classification approach to generalised speaker verification spoofing countermeasures using local binary patterns. *IEEE 6th International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013*. https://doi.org/10.1109/BTAS.2013.6712706

Cabeza, Y. (2023). *Denuncias por ciberdelincuencia se incrementan en un 150% en el 2023: mayoría son por fraude*. https://www.infobae.com/peru/2023/09/09/denuncias-por-ciberdelincuencia-se-incrementan-en-un-150-en-el-2023-mayoria-son-por-fraude/

Fraga, L. (2001). *La transformada discreta de Fourier y la transformada rápida de Fourier*. https://delta.cs.cinvestav.mx/~fraga/Cursos/PDI/tdf.pdf .

Fuertes, W., Zapata, P., Ayala, L., & Mejía, M. (2010). *Plataforma de experimentación de ataques reales a redes IP utilizando tecnologías de virtualización*. https://repositorio.espe.edu.ec/bitstream/21000/6057/1/AC-RIC-ESPE-034343.pdf.

Kinnunen, T., Wu, Z. Z., Lee, K. A., Sedlak, F., Chng, E. S., & Li, H. (2012). Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 4401-4404. https://doi.org/10.1109/ICASSP.2012.6288895

Le, T., Gilberton, P., & Duong, N. Q. K. (2019). Discriminate natural versus loudspeaker emitted speech. *arXiv*, 1901.11291.

Martínez Mascorro, G. A., & Aguilar Torres, G. (2013). Reconocimiento de voz basado en MFCC, SBC y Espectrogramas. *Ingenius* (10), 12-20. https://doi.org/10.17163/ings.n10.2013.02

Morejón S. (2011). *Segmentación de audio y de locutores para recuperación de información multimedia y su aplicación a videos de información turística*. 118-170. https://repositorio.uam.es/bitstream/handle/10486/6734/39702_20110603LeticiaRueda.pdf?sequence=1&isAllowed=y

Mustafa, H., Xu, W., Sadeghi, A. R., & Schulz, S. (2014). You can call but you can't hide: Detecting caller ID spoofing attacks. *Proceedings of the International Conference on Dependable Systems and Networks*. https://doi.org/10.1109/DSN.2014.102

Rueda, L. (2011). *Mejoras en reconocimiento del habla basadas en mejoras en la parametrización de la voz*. https://repositorio.uam.es/bitstream/handle/10486/6734/39702_20110603LeticiaRueda.pdf?sequence=1&isAllowed=y

Shukla, S., Ahirwar, M., Gupta, R., Jain, S., & Rajput, D. S. (2019). Audio Compression Algorithm using Discrete Cosine Transform (DCT) and Lempel-Ziv-Welch (LZW) Encoding Method. *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Prespectives and Prospects, COMITCon 2019*. https://doi.org/10.1109/COMITCon.2019.8862228

Singh, R., Gencaga, D., & Raj, B. (2016). Formant manipulations in voice disguise by mimicry. *4th International Conference on Biometrics and Forensics* (IWBF), pp. 1-6, https://doi.org/10.1109/IWBF.2016.7449675

Toro Cerón, L. G. (2018). *Análisis de Estrés en la Voz Utilizando Coeficientes Cepstrales de Mel y Máquina de Vectores de Soporte*. https://bibliotecadigital.usb.edu.co/entities/publication/41b81de7-886a-4763-bd62-386dbddad29b.

Wang, Y., & Lawlor, B. (2017). Speaker recognition based on MFCC and BP neural networks. *28th Irish Signals and Systems Conference, ISSC 2017*, 0-3. https://doi.org/10.1109/ISSC.2017.7983644

Zorro, M. (2022). *Irish arrests in global anti-fraud operation*. BBC News NI. https://www.bbc.com/news/articles/czq3d1ld6l9o