ISSN 1870-4069

Identificación de un individuo mediante la voz utilizando redes convolucionales y umbrales de aceptación

José Luis Medina Jiménez, Héctor Rodríguez Rangel,
Gloria Ekaterine Peralta Peñuñuri, Mario Alberto Román Garay,
Luis Alberto Morales Rosales

Tecnólogico Nacional de México Campus Culiacán,
División de posgrado,
México

Conacyt-Universidad Michoacana de San Nicolás de Hidalgo,
Facultad de Ingeniería Civil,
México

{jose_medinaj,mario_roman}@itculiacan.edu.mx,
{hector.rr,gloria.pp}@culiacan.tecnm.mx,
lamorales@conacyt.mx

Resumen. En la actualidad el uso de usuario y contraseña como método tradicional de acceso es un problema debido a que son fáciles de olvidar y no proporciona la certeza de que los datos de los usuarios están seguros. Por otro lado, el robo de identidad es otro de los problemas principales que se desea evitar. El robo de identidad en México y en el mundo ha crecido en los últimos años debido a la pandemia, ya que el uso de medios digitales ha aumentado por las medidas sanitarias que se han tomado. La exploración de la seguridad biométrica se ha ampliado gracias al avance de la inteligencia artificial y de los métodos de extracción de características, esto debido a que los rasgos biométricos son únicos y no cambian. Uno de los principales problemas de trabajar con la voz, es el timbre, en particular el tono, ya que de esta manera pueden existir voces muy similares en sus frecuencias capaces de confundir a los sistemas de identificación. En el artículo se describe un sistema de identificación de personas mediante la voz utilizando los espectrogramas. Además, mediante la aplicación de distintos umbrales se realizó una comparativa de estos, para reducir el impacto que tiene el timbre de voz al momento de identificar personas, llegando a obtener una precisión del 93 %.

Palabras clave: Acceso biométrico, reconocimiento de voz, aprendizaje profundo, red neuronal convolucional, VGGVox.

Identification of an Individual by Voice Using Convolutional Networks and Acceptance Thresholds

Abstract. Currently the use of username and password as a method traditional access is a problem because they are easy to forget and do not provide the

certainty that user data is safe. On the other hand, identity theft is another major problem you want to avoid. Identity theft in Mexico and in the world has grown in recent years due to the pandemic, since the use of digital media has increased due to the sanitary measures that have been taken. The exploration of biometric security has been expanded by the advancement of artificial intelligence and feature extraction methods, because biometric features are unique and do not change. One of the main problems of working with the voice is the timbre, in particular the tone, since in this way there can be voices that are very similar in their frequencies, capable of confusing the identification systems. The article describes a system for identifying people by voice using spectrograms. In addition, through the application of different thresholds, a comparison of these was made, to reduce the impact of the voice timbre when identifying people, reaching an accuracy of 93%.

Keywords: Biometric access, voice recognition, learning deep, convolutional neural network, VGGVox.

1. Introducción

Conforme ha crecido la era digital, el robo de identidad ha aumentado junto con ella, de tal manera que la falsificación de datos de personas ha afectado a millones. Durante el 2020, cuando dio lugar el inicio de la pandemia, el confinamiento aumentó el uso de internet, así como aumentaron los casos de falsificación con respecto a años anteriores. Según datos del Banco de México (Banxico), el país se encuentra en el octavo lugar en el mundo en delito de robo de identidad y es el segundo lugar en Latinoamérica, estimando que el 67 % es por perdida de documentos[1].

Actualmente, los métodos tradicionales como el usuario y contraseña ya no son suficientes, ya que son fáciles de olvidar o de robar para la suplantación de identidad, por ello los rasgos biométricos se utilizan en distintos sectores de la industria. La finalidad de utilizar los datos biométricos como sistema de acceso, es aumentar la seguridad y privacidad del usuario, ya que estos rasgos son universales y únicos [2].

La Biometría se define como las características únicas con las que cuenta un individuo. Estas se clasifican en dos categorías: físicas y de comportamiento. Las físicas engloban aquellas como la huella dactilar, iris, geometría de la mano y rostro y, por otro lado, tenemos las características de comportamiento, entre las cuales están, la firma, la escritura, la voz, el andar [4].

El reconocimiento de voz es una gran alternativa para la nueva modalidad que se encuentra actualmente la sociedad, ya que a diferencia de la huella dactilar, no es necesario tener contacto directo con los sensores.

El aprendizaje profundo se ha vuelto ampliamente utilizado para el reconocimiento de voz, debido a que estas técnicas permiten entrenamientos más rápidos y manejo de grandes datos, haciendo más eficiente la tarea de clasificación de voz [3]. Las redes profundas más utilizadas son redes convolucionales, redes recurrentes, memoria a corto plazo o LSTM por sus siglas en inglés, otras arquitecturas de aprendizaje profundo utilizadas son los autocodificadores y las redes generativas-adversarias (GAN) [4].

Los sistemas de reconocimiento de voz se pueden visualizar o separar en dos distintos enfoques. Por un lado, está el enfoque de los modelos tradicionales, llamados así, ya que pertenecen a modelos matemáticos-estadísticos, antes de la llegada de las redes neuronales profundas.

Estos modelos son los HMM (Hidden Markov Model), GMM (Gaussian Mixture Model) algoritmos basados en SVD (Singular Value Decomposition), DCT (Discrete Cosine Transform), Enfoque de agrupamiento iterativo, índice de probabilidad, entre otros derivados de los mencionados anteriormente. Ahora, por parte del enfoque del Aprendizaje profundo, algunos ejemplos son, DNN (Deep Neural Network), CNN (Convolutional Neural Network), SVM (Support Vector Machines) y entre otros derivados de estos mismos.

A manera general, un sistema biométrico tiene cuatro etapas. La primera es la de registro del usuario mediante el uso de un sensor apropiado ("Enrollment"). La segunda etapa se encarga de extraer las características del dato biométrico de entrada ("Feature Extractor") guardándola como plantilla en una base de datos.

La siguiente etapa es la nueva entrada del dato biométrico, donde se le realiza el proceso de extracción de característica y se compara con la que está guardada en la base de datos ("Matcher"). Por último, se hace la toma de decisión de verificación o autentificación de la persona [5].

La identificación de voz es una rama del reconocimiento de voz en donde se procesa una muestra de voz desconocida y compara con una base de datos de personas establecidas, es decir, una medición de 1 contra N personas. La voz desconocida se identifica como la que mejor se adapte, de esta manera la entrada para la identificación de voz es una voz desconocida y la salida es el nombre o la identificación del usuario [7].

Uno de los principales problemas que existe en la identificación de personas mediante la voz, es el timbre que tiene la voz, debido a que esta característica trabaja con frecuencias. Al tener una comparativa de 1 contra N personas, entre más grande sea el valor de N o dicho de otro modo entre más personas, sea con las que se compare la voz de entrada, los sistemas de identificación pueden confundirse detectando de manera errónea a otra persona.

Ante estas problemáticas, en este artículo se propone la implementación de un sistema de identificación de voz empleando Redes Neuronales Convolucionales utilizando como datos de entrada espectrogramas, es decir, trabajamos en el dominio de la frecuencia considerando el timbre de voz. Además, se hace un análisis de distintos umbrales para encontrar el más adecuado que permita identificar de manera correcta mediante la voz a un individuo.

Para ello, se realizaron dos pruebas distintas: 1) la identificación se realiza repitiendo la misma frase de registro del individuo y 2) se utiliza la frase de registro del individuo para comparar su similitud de la voz con una frase distinta mencionada.

A partir de estas pruebas, se llega a la conclusión que el uso de umbrales bajos para la medición de la similitud de la voz permite una mayor fiabilidad en la identificación de personas mediante la voz, obteniendo así un 93 % precisión.

El resto del artículo está organizado de la siguiente manera: Estado del arte, Materiales y Métodos, Resultados, Conclusiones y Trabajos Futuros.

2. Estado del arte

Durante los años 30, Francés McGehee se inspiró en un caso de secuestro y asesinato para el desarrollo del reconocimiento de voz, ya que uno de los afectados reconoció la voz del secuestrador. La idea de McGehee fue realizar una investigación de que tan fiable es el oído humano, que posteriormente esto daría partida a un tipo de investigación forense y psicológica [8].

Actualmente, este tipo de investigación de reconocimiento de voz sigue gracias a la Inteligencia Artificial, enfocado en técnicas de Aprendizaje-Máquina y Aprendizaje Profundo. Esto debido a que se han explorado distintos extractores de características tales como la Transformada Discreta de Wavelet donde principalmente sé utilizando los Modelos Gaussianos Mixtos (GMM) y el Perceptrón Multicapa (MLP) para el reconocimiento de voz [8].

Por otro lado, tenemos los Coeficientes Cepstrales en las Frecuencias de Mel utilizándose en los mismos modelos anteriores y además en DNN, SVM y CNN [8]. Existen otras técnicas de extracción con Vectores X, espectrogramas, características espectrales dinámicas normalizadas, Coeficientes cepstrales temporales basados en la energía de Teager, además de las combinaciones de estas [8].

En el trabajo de [9] proponen un sistema de verificación de voz dependiente de texto, donde se utiliza una DNN supervisada para la extracción de características a nivel cuadro, de esta manera cada cuadro se va apilando en un vector de izquierda a derecha, correspondiendo al número de voces de entrenamiento por cada cuadro.

En el artículo [10] se utiliza una modificación de la DNN, donde se emplea la combinación de un autocodificador con DNN, esto con la finalidad de mejorar el audio, donde la función principal del autoencoder es la eliminación del ruido y la reverberación.

Un modelo enfocado en CNN propuesto en [11] diseñado para optimizar el proceso de identificación de voz, basado en el dataset TIMIT. Para el preprocesamiento de datos se utilizan espectrogramas para mejorar las fuentes acústicas. Esta CNN contiene varias capas de convolución aplicando varios filtros a diferentes secciones locales de entrada y a su vez esa capa le sigue una capa de agrupación máxima, de esta manera emite una versión de más baja resolución de las activaciones de la capa de convolución eliminando la activación total del filtro.

Dentro de la categoría de las CNN existen las redes convolucionales 3D, donde el kernel encargado de la convolución se mueve en 3 direcciones. En el trabajo [12] se propone la utilización de las 3D-CNN, donde estas 3 dimensiones, además del dominio de la frecuencia como en otros trabajos, también se enfoca en el dominio del tiempo y el tamaño de enunciados que existe en un audio, con la finalidad de construir un modelo más robusto para la problemática de los cambios que existen en la voz.

Las Redes Siamesas están basadas en un modelo CNN, propuesto en [7], este trabajo consiste en dos entradas simultáneas aprendiendo una función de similitud y así muestra lo idénticas que son las dos entradas. Por lo tanto, el objetivo de la red siamesa no es clasificar la voz, sino distinguir o conocer la similitud entre las dos voces de entrada.

En el trabajo [13] proponen la utilización de la Máquina de Soporte Vectorial, enfocado en tratar el reconocimiento de voz como un problema de clasificación binaria.

Tabla 1. Características de micrófono HyperX QuadCast.

Características	Valores
Velocidad de muestreo/bits	48kHz/16-bit
Patrones polares	Estéreo, omnidireccional, cardiode, bidireccional
Respuesta de frecuencia	20Hz-20kHz
Sensibilidad	-36dB(1V/Pa a 1kHz)

Durante el proceso de reconocimiento de voz, se aplica el clasificador entrenado en distintos puntos para reconocer si la voz coincide o no. Existen algunas técnicas más recientes donde se proponen variantes y mejoras de extractores de características, ya que estas definen que tan bien van a detectar a la persona que habla.

En el artículo [14] se menciona como se combina el uso de los MFCC con características basadas en el tiempo, convirtiéndola en MFCCT, con la finalidad de mejorar la precisión de los sistemas de Identificación de Voz.

3. Materiales y métodos

Durante el desarrollo del sistema de identificación de voz, fue necesario utilizar un dispositivo de entrada de audio y un servido que se encargue de procesar y realizar el trabajo de identificación de la voz. Se describirá a continuación el hardware utilizado y posteriormente se mostrará a detalle los procesos que se realizan por parte del sistema para realizar la tarea de identificación de voz.

3.1. Hardware

Los sistemas de reconocimiento de voz como método de seguridad para cualquier tipo de acceso, siendo una buena alternativa o inclusive para realizar la fusión de sistemas que utilicen otro tipo de reconocimiento biométrico y así robustecer el método de acceso. En cualquier sistema biométrico es necesario contar con el sensor o dispositivo de entrada para capturar el rasgo de biometría a utilizar. En el caso del reconocimiento de voz se debe contar con un micrófono para capturar la voz.

Para tener un sistema controlado y no exista variaciones en los audios, se utilizó un micrófono HyperX QuadCast, ya que este cuenta con características ideales para tener mayor control en las ganancias de los audios y así evitar la mayoría de los ruidos ambientales o inclusive las voces de fondo que puedan entorpecer el proceso de reconocimiento de voz.

Para el sistema de reconocimiento de voz, es necesario que el audio cuente con ciertas características, como una tasa de 16 – Bit, audio mono y que tenga una frecuencia de 16 kHz, para esto se muestra las características del micrófono en la Tabla 1, la cual cuenta con los requisitos necesarios para nuestro sistema.

El procesamiento de software se realiza con un servido, la cual cuenta con las características necesarias para realizar la tarea de identificación de voz sin ningún problema. Las características esenciales para realizar el proceso se muestran en la Tabla 2.

Tabla 2. Características del servido.

Nombre	Características
Sistema Operativo	Windows 11
RAM	16 GB
Almacenamiento	500 GB SSD
Procesador	Intel Core i7 9° Gen
Tarjeta Grafica	Nvidia RTX 2060

3.2. Sistema de identificación de voz

El Sistema de Identificación de voz se divide en dos etapas, la etapa de registro de usuario y la etapa de identificación mediante la voz, las cuales a continuación se van a explicar sus procesos.

Etapa de registro de usuario

En la etapa de Registro, se introduce la voz de la persona a la base de datos con sus respectivos pre-procesamientos y se almacena, tal como se muestra en la Figura 1.

– **Frase de registro:** En la etapa de registro el audio llega en "crudo", es decir, que llega en un formato tipo "WAV", a una frecuencia de 16 kHz, una tasa de muestreo a 16-Bits y en canal mono. Debido a que el modelo fue entrenado con audios con las características descritas anteriormente, es necesario que los registros y las entradas de audio cumplan con este requisito.

  Ahora, para realizar el registro fue necesario que el usuario usara una frase en concreto de una duración alrededor de 5 segundos. La frase utilizada fue "El reconocimiento de voz es la llave de acceso al futuro como contraseña", esto con la finalidad que se hicieran pruebas con frases iguales y frases distintas y verificar si existía una variación en precisión entre usar la misma frase a una distinta.

– **Extractor de espectrogramas:** A partir de obtener el audio en "crudo", durante la etapa de extracción de espectrograma se realiza un preprocesamiento en el audio en el cual mediante la utilización de la Transformada Rápida de Fourier pasa del dominio del tiempo a dominio de la frecuencia.

  Una vez transformado el audio al dominio de la frecuencia, se extrae una imagen del audio, el cual representa el espectrograma como se puede observar en la Figura 2, esto para posteriormente pasar al modelo utilizado para obtener el vector de características, es decir, el "embedding".

– **Generación de embeddings:** El modelo utilizado para la extracción del vector de características fue del trabajo [15] llamada VGGVox, el cual está basada en una red convolucional VGG-M enfocada en la clasificación de imágenes, se hace una pequeña modificación donde se recibe como entrada las imágenes de espectrogramas extraídas de los audios. En la Figura 3 se muestra la estructura general de la Red neuronal Convolucional VGGVox.

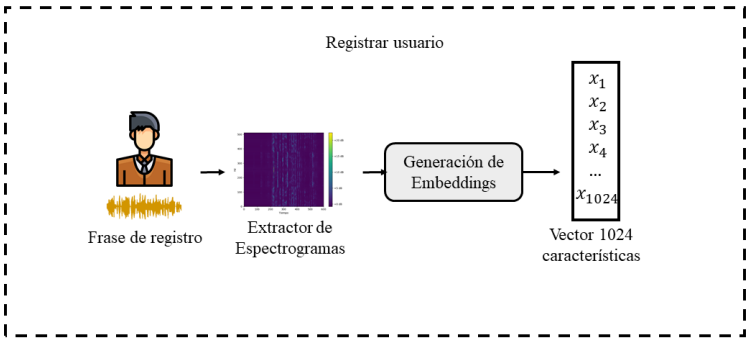

Fig. 1. Diagrama de registro de voz.

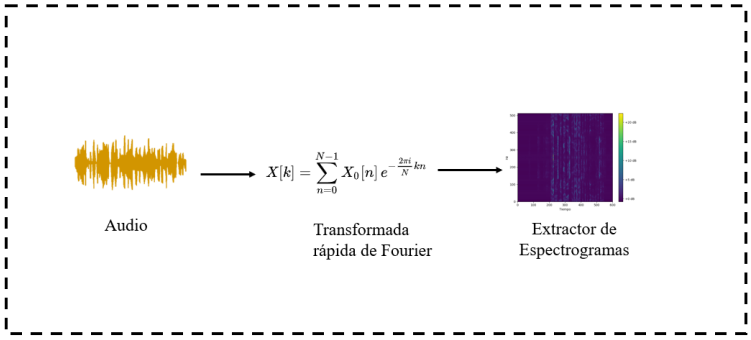

Fig. 2. Espectrogramas.

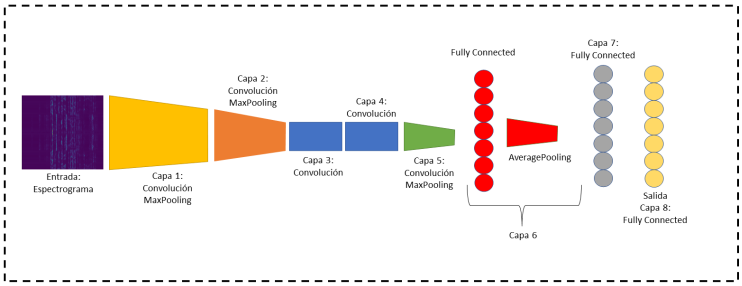

Fig. 3. Red Neuronal Convolucional VGGVox.

En la Tabla 3 se detalla la arquitectura de la VGGVox donde los autores hacen una modificación de la VGG-M, agregando a la capa “fully connected” (fc6) una capa de “average pool” el cual tiene un kernel que varía dependiendo de los segundos del audio esto para tener la flexibilidad de meter audios de diferentes tiempos. Observando la Tabla 3 de la arquitectura, en la última capa la dimensión de filtro es la salida entregada de la red VGGVox.

Tabla 3. Arquitectura VGGVox.

Capa	Kernel	Dimensión de filtro	Filtros	Stride	Tamaño de datos
conv1	7x7	1	96	2x2	256x148
mpool1	3x3	-	-	2x2	126x73
conv2	5x5	96	256	2x2	62x36
mpool2	3x3	-	-	2x2	30x17
conv3	3x3	256	384	1x1	30x17
conv4	3x3	384	256	1x1	30x17
conv5	3x3	256	256	1x1	30x17
mpool5	5x3	-	-	3x2	9x8
fc6	9x1	256	4096	1x1	1x8
apool6	1xn	-	-	1x1	1x1
fc7	1x1	4096	1024	1x1	1x1
fc8	1x1	1024	1251	1x1	1x1

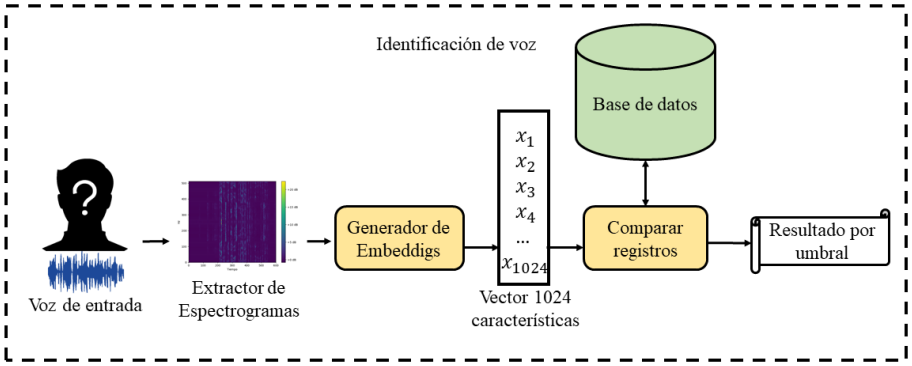

Fig. 4. Diagrama de identificación de voz.

– **Vector de características:** Una vez que el usuario realiza el proceso de grabar la frase con su voz, al pasar por el proceso de extracción de espectrograma y se obtiene un vector de características denominado “embedding” el cual cuenta con 1024 características extraídas de la imagen del espectrograma.

Etapa de identificación de usuario

La etapa de identificación mostrada en la Figura 4, llega una voz de entrada al sistema y este se encargará de clasificar esta voz entre los usuarios registrados en la etapa de registro, los cuales estos fueron almacenados en una base de datos.

– **Voz de entrada:** En la etapa de voz de entrada, tiene el mismo proceso que pasa al registrar un usuario, la persona introduce una frase con su voz, este pasa por una etapa de extracción del espectrograma y se extrae el “embedding”.

Durante la etapa de pruebas se recolectaron alrededor 768 audios de muestra de 128 personas, es decir, 6 audios por cada persona. De los 6 audios recopilados por personas, 3 eran con la finalidad de decir la misma frase que se hizo de registro y los otros 3 con frases distintas, para posteriormente separar estos dos grupos y someterlos al sistema.

– **Comparar registros:** Al tener el vector de características de la voz de entrada llega la etapa de comparar registros, es decir, identificar a que voz coincide con las registradas en la base de datos, es por eso que en esta etapa, al ser vectores de mismo tamaño de características y solo trabajar en un mismo plano (frecuencias) mediante la utilización de la ecuación 1 es considerada una de las métricas más utilizadas y de menor costo computacional al momento de comparar similitud entre vectores:

$$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{1}$$

donde y representa el vector de entrada y x representa las voces registradas, esta última iterándose para medirse una a una con la voz a identificar y se guarda cada distancia en un vector para la siguiente etapa. La finalidad de la medición de la distancia euclidiana es que entre más cercana a 0 sea el valor obtenido, significa que más acertada es la identificación.

– **Resultado por umbral:** Al obtener el vector de distancias es necesario determinar si esa persona identificada es la correcta o no, es por eso que mediante un umbral se determina la identificación de la persona.

Al ir aumentando el tamaño de personas registradas, este umbral debe ir variando, es decir, ir disminuyendo este parámetro conforme va aumentando el registro de personas para así tener mayor tasa de precisión y evitar la identificación errónea.

Debido a que estamos utilizando la medida de distancia euclidiana para identificar a la persona, entre menor sea el umbral, significa que es más similar es la voz registrada a la de entrada al sistema, ya que el umbral va en función a los valores que se obtienen con la distancia euclidiana. Existen otras métricas que ayudan a conocer que tan fiable es el sistema de identificación.

La Exactitud que es interpretada por la tasa de porcentaje de la cantidad de predicciones positivas que fueron correctas, y se obtienen con la fórmula 2 ahora bien, para obtener el valor de precisión, el cual representa el porcentaje de casos positivos detectados, donde este porcentaje nos indica que tan fiable es el valor detectado como positivo, se define con la ecuación 3:

$$\text{exactitud} = \frac{(VP + VN)}{(VP + FP + FN + VN)}, \tag{2}$$

$$\text{precision} = \frac{VP}{(VP + FP)}. \tag{3}$$

Tabla 4. Matriz de confusión con audios diciendo la misma frase de registro.

Umbral	Verdaderos positivos	Falsos positivos	Verdaderos negativos	Falsos negativos
0.1	136	9	85	154
0.12	175	20	76	113
0.14	210	29	72	73
0.16	237	40	63	44
0.18	248	51	55	30
0.2	254	68	44	18
0.22	259	77	37	11
0.24	262	86	30	6
0.26	264	93	26	1
0.28	265	101	18	0
0.3	265	106	13	0

4. Resultados

En la Tabla 4 mediante la utilización de las 4 opciones que ofrece la matriz de confusión, se exploraron distintos umbrales para la toma de decisión de similitud de audios. Se obtuvieron los Verdaderos Positivos (VP) que son aquellos que el sistema detecta como verdaderos y su valor real es verdadero, Falsos Positivos (FP) se interpretan como los audios aceptados como una persona distinta, Verdaderos Negativos (VN) estos resultados se obtienen de personas que no están en el sistema y que realmente los rechaza y por último Falsos Negativos (FN) que son aquellos audios que si debieron asignarle un valor verdadero, pero que fueron rechazados por el sistema.

A partir de la Tabla 4 de los 4 puntos de la matriz de confusión, se determinaron las métricas por cada umbral de exactitud, precisión y F1 las cuales se pueden observar en la Tabla 5. En la Tabla 6 se muestra los resultados de la matriz de confusión para frases distintas con el mismo tamaño de audios, además que en la Tabla 7 se muestran las métricas obtenidas.

Al comparar resultados entre la Tabla 4 y 6 de matriz de confusión con la misma frase y frase distinta, se puede observar que existe una diferencia de aproximadamente un 20 % entre los Verdaderos positivos, es decir, que al utilizar la misma frase existe mayor exactitud a la hora de reconocer correctamente a la persona.

Por otro lado, un punto importante de la matriz de confusión en los sistemas biométricos son los falsos positivos, ya que estos determinan cuantas personas identificó erróneamente.

Al comparar ambas pruebas conforme va creciendo el umbral en el sistema empieza aumentar los falsos positivos, teniendo que la diferencia entre el umbral más bajo, no hay una diferencia significativa, pero al utilizar un umbral alto, como es el caso de 0.3 la diferencia entre ellos es casi de 15 %. Por lo tanto, no es conveniente en los sistemas biométricos una tasa alta en falsos positivos, ya que esto implica asignar erróneamente un valor correcto a una persona.

Por parte de las métricas de exactitud y precisión en ambas pruebas de las Tablas 5 y 7, se puede observar que el sistema es preciso con umbrales bajos, esto debido a que entre más bajo sea el umbral más cerca está de ser exacta a la voz de la base de datos.

Tabla 5. Métricas obtenidas de la matriz de confusión de audios diciendo la misma frase de registro.

Umbral	Exactitud	Precisión	F1
0.1	57.55 %	93.79 %	62.53 %
0.12	65.36 %	89.74 %	72.46 %
0.14	73.44 %	87.87 %	80.46 %
0.16	78.13 %	85.56 %	84.95 %
0.18	78.91 %	82.94 %	85.96 %
0.2	77.60 %	78.88 %	85.52 %
0.22	77.08 %	77.08 %	85.48 %
0.24	76.04 %	75.29 %	85.06 %
0.26	75.52 %	73.95 %	84.89 %
0.28	73.70 %	72.40 %	83.99 %
0.3	72.40 %	71.43 %	83.33 %

Tabla 6. Matriz de confusión con audios diciendo distinta frase a la de registro.

Umbral	Verdaderos positivos	Falsos positivos	Verdaderos negativos	Falsos negativos
0.1	60	5	90	227
0.12	96	21	86	179
0.14	125	41	78	138
0.16	149	64	69	100
0.18	173	93	55	61
0.2	189	112	47	34
0.22	196	128	34	24
0.24	200	141	26	15
0.26	205	149	21	7
0.28	207	156	15	4
0.3	208	160	11	3

La precisión en este caso indica la fiabilidad o certeza con la que el sistema está clasificando el valor positivo, es decir, para evitar que el sistema se confunda por la similitud en las personas en su timbre de voz, es mejor utilizar umbrales bajos para así asegurar mayor precisión a la hora de identificar a la persona de manera correcta. Es por eso que, el usar el umbral de 0.1 nos entrega un desempeño del 93 %.

Por último, la métrica F1 simplifica las medidas de precisión y exhaustividad, donde va de 0 a 1, o 0 a 100 en porcentaje, siendo 1 (100) el mejor caso. El decir la misma frase tiene mejor rendimiento en el sistema, evaluándose con la medida F1 debido a que hay mayores verdaderos positivos a la hora evaluarse el sistema.

En la Tabla 8 se compara el desempeño del trabajo de este artículo con respecto los trabajos de otros autores mostrados en el estado del arte. Se pueden observar las distintas características de entrada y los tipos de extractores de características, ya que estos elementos son clave para obtener una tasa alta de desempeño.

Tabla 7. Métricas obtenidas de la matriz de confusión de audios diciendo distinta frase a la de registro.

Umbral	Exactitud	Precisión	F1
0.1	39.27 %	92.31 %	34.09 %
0.12	47.64 %	82.05 %	48.98 %
0.14	53.14 %	75.30 %	58.28 %
0.16	57.07 %	69.95 %	64.50 %
0.18	59.69 %	65.04 %	69.20 %
0.2	61.78 %	62.79 %	72.14 %
0.22	60.21 %	60.49 %	72.06 %
0.24	59.16 %	58.65 %	71.94 %
0.26	59.16 %	57.91 %	72.44 %
0.28	58.12 %	57.02 %	72.13 %
0.3	57.33 %	56.52 %	71.85 %

Tabla 8. Sistemas desarrollados del estado del arte.

Autores	Base de datos utilizada	No. De hablantes	Entrada	Modelo	Desempeño
Variani et al. (2014) [9]	NA	646	Características energéticas del marco	DNN	EER : 2.00 (Por 20 expresiones)
Lukic et al. (2016) [11]	TIMIT	630	Espectrograma de datos de voz	CNN	AC: 97
Plchot et al. (2016) [10]	PRISM: Fisher corpora Switch Board SRE	13916 1991 2740	MFCC, PNCC	DNN auto encoder	NA
Chung et al. (2017) [15]	Voxceleb	1251	Espectograma	CNN	AC: 80.5 EER: 7.8
Torfi et al. (2018) [12]	WVU-Multimodal 2013	1083	Marco de la MFEC	3D-CNN	EER: 21.1
Dhakal et al. (2019) [13]	ELSDSR	22	Estadística, característica de Gabor y basada en CNN	SVM RF DNN	AC: 98.07 AC: 99.41 AC: 98.14
Jahangir et al. (2020) [14]	LibriSpeech	50 Hombres 50 Mujeres	Expresiones de los hablantes	MLDNN	AC: 92.9
Este trabajo (2022)	propio	128	Espectograma	CNN	AC: 93

Además, existe una gran diferencia entre las bases de datos utilizadas por los diferentes autores, ya que contienen diferencias acerca del ruido ambiental en los audios, el número de audios, algunos cuentan con grandes variedades de etnias, edades, nacionalidades y en contraste otros solo consideran el idioma Inglés-Americano. Por lo tanto, es difícil determinar cuál sistema de reconocimiento de voz es mejor. Existe variabilidad en las consideraciones de cada trabajo, por lo que tener un sistema único capaz de reconocer a una persona por la voz sigue siendo un reto abierto.

5. Conclusiones

El reconocimiento de voz es complejo debido a que existen diversos factores como el ruido ambiental que puede introducir frecuencias no deseadas al audio y confundir al sistema. Por otro lado, los factores de las emociones, el timbre de voz, también son importantes a considerar este tipo de problemáticas, ya que a pesar de que el

timbre es una de las características que hacen única a la voz, gracias a que con esta se mide la calidad con la que es producida, englobando la parte de la entonación donde mayormente ha tenido investigación el reconocimiento de voz por estar en el dominio de las frecuencias. La forma de como articulas las palabras y la intensidad, son clave al momento de utilizar este tipo de reconocimiento, debido a que al realizar pruebas, el sistema reconocía más fácilmente a las personas que hablaban de manera más clara y a una velocidad modulada.

A nivel hardware se trató de tener un sistema controlado, ya que la sensibilidad con la que se maneja el micrófono es importante, debido a que una alta sensibilidad puede afectar con mucho ruido al momento de hablar y los ruidos ambientales. Además, la distancia con la que hablas al micrófono va de la mano con la sensibilidad, ya que es más fácil para el reconocimiento de voz mantener una distancia corta al micrófono debido a que detecta mejor las frecuencias principales de tu voz.

La voz es un tema amplio de investigar, por los diversos factores que influyen con solo decir una frase, la variabilidad que existe de una persona a otra en es inmensa por las distintas características que puede contar la voz.

5.1. Trabajo futuro

A partir de los resultados obtenidos, y enfocados en el estado del arte, se puede explorar diversos extractores de características que existen, inclusive combinarse para aumentar la fiabilidad del sistema a la hora de identificar mediante la voz y así reducir los falsos positivos que arroja el sistema.

Por otro lado, una vez mejorado el sistema de identificación de voz, se propone agregar una nueva funcionalidad donde se detecten audios falsos, es decir, detectar cuando se reproducen audios de voz en la entrada del micrófono para así robustecer y atacar este tipo de problemas comunes en los sistemas de reconocimiento de voz.

Referencias

1. Banxico. Informe anual sobre el ejercicio de las atribuciones conferidas por la Ley para la Transparencia y Ordenamiento de los Servicios Financieros (2021)
2. Gayathri, M., Malathy, C., Prabhakaran, M.: A review on various biometric techniques, its features, methods, security issues and application areas. In: International Conference On Computational Vision and Bio Inspired Computing, vol. 1108, pp. 931–941 (2019) doi: 10.1007/978-3-030-37218-7_99
3. Boles, A. Rad, P.: Voice biometrics: Deep learning-based voiceprint authentication system. In: 12th System Of Systems Engineering Conference (SoSE). IEEE, pp. 1–6 (2017) doi: 10.1109/SYSOSE.2017.7994971
4. Minaee, S., Abdolrashidi, A., Su, H., Bennamoun, M., Zhang, D.: Biometrics recognition using deep learning: A survey. ArXiv Preprint ArXiv:1912.00271 (2019) doi: 10.48550/arXiv.1912.00271
5. Jain, A., Nandakumar, K.: Biometric authentication: System security and user privacy. Computer, vol. 45, no. 11, pp. 87–92 (2012)

6. Muckenhirn, H., Doss, M. M., Marcell, S.: Towards directly modeling raw speech signal for speaker verification using CNNS. In: IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), pp. 4884–4888 (2018) doi: 10.1109/ICASSP.2018.8462165
7. Tandel, N. H., Prajapati, H. B., Dabhi, V. K.: Voice recognition and voice comparison using machine learning techniques: A Survey. In: 6th International Conference On Advanced Computing And Communication Systems (ICACCS), pp. 459–465 (2020) doi: 10.1109/ICACCS48705.2020.9074184
8. Hanifa, R., Isa, K., Mohamad, S.: A review on speaker recognition: Technology and challenges. Computers & Electrical Engineering, vol. 90, pp. 107005 (2021) doi: 10.1016/j.compeleceng.2021.107005
9. Variani, E., Lei, X., McDermott, E., Moreno, I. L., Gonzalez-Dominguez, J.: Deep neural networks for small footprint text-dependent speaker verification. In: IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), pp. 4052–4056 (2014) doi: 10.1109/ICASSP.2014.6854363
10. Plchot, O., Burget, L., Aronowitz, H., Matejka, P.: Audio enhancing with DNN autoencoder for speaker recognition. In: IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP), pp. 5090–5094 (2016) doi: 10.1109/ICASSP.2016.7472647
11. Lukic, Y., Vogt, C., Dürr, O., Stadelmann, T.: Speaker identification and clustering using convolutional neural networks. In: IEEE 26th International Workshop On Machine Learning For Signal Processing (MLSP), pp. 1–6 (2016) doi: 10.1109/MLSP.2016.7738816
12. Torfi, A., Dawson, J., Nasrabadi, N. M.: Text-independent speaker verification using 3d convolutional neural networks. In: IEEE International Conference On Multimedia And Expo (ICME). IEEE, pp. 1–6 (2018) doi: 10.1109/ICME.2018.8486441
13. Dhakal, P., Damacharla, P., Javaid, A. Y., Devabhaktuni, V.: A near real-time automatic speaker recognition architecture for voice-based user interface. Machine learning and knowledge extraction, vol. 1, no. 1, pp. 504–520 (2019) doi: 10.3390/make1010031
14. Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M., Ali, I.: Text-independent speaker identification through feature fusion and deep neural network. IEEE Access, vol. 8, pp. 32187–32202 (2020) doi: 10.1109/ACCESS.2020.2973541
15. Nagrani, A., Chung, J. S., Zisserman, A.: Voxceleb: a large-scale speaker identification dataset. ArXiv Preprint ArXiv:1706.08612 (2017) doi: 10.21437/Interspeech.2017-950