Deep Learning Convolutional Network for Bimodal Biometric Recognition with Information Fusion at Feature Level

Juan Carlos Atenco Vázquez, Juan Carlos Moreno Rodríguez and Juan Manuel Ramírez Cortés

***Abstract*—Biometric recognition has been an extensively researched field in recent years due to the growth of its applications in daily activities. State of the art work in biometrics proposes the implementation of multimodal systems that employ one or more traits to increase the security of the system since it is more difficult for an impostor to acquire, falsify or forge multiple samples of different traits from an enrolled user. In this paper, we propose the implementation of a Deep Learning bimodal network that combines voice and face modalities. Voice feature extraction was done with a SincNet arquitecture and face image features were extracted with a set of convolutional layers. The feature vectors of both modalities are combined within the network with two methods: averaging or concatenation. The averaged/concatenated vector is further processed with a fully connected layer to output a bimodal vector that contains discriminatory information of an individual. The bimodal vector is used with a fully connected layer with the softmax function to perform the identification task. The verification task is performed by matching the bimodal vector with a template to obtain a score that must be used to either accept or reject an user's identity. We compared the results yielded by both fusion methods implemented in our proposed network for both recognition tasks. Both methods achieved an accuracy as high as 99 % in the identification task and an Equal Error Rate (EER) as low as 0.14 % for verification. These results were obtained by combining BIOMEX-DB and VidTimit databases.**

***Index Terms*—Multimodal biometrics, Deep Learning, Speaker recognition, Face recognition.**

I. Introducción

En años recientes ha habido un enorme crecimiento de las aplicaciones que requieren autenticar la identidad de un individuo, ya sean servicios digitales proporcionados por entidades públicas o privadas, procesos en ciencias forenses o tareas de seguridad, entre otros. Para autenticar la identidad de un individuo se han empleado métodos tradicionales como memorización de contraseñas o números de identificación personal (PIN), credenciales físicas o dispositivos electrónicos que contienen la información de un usuario, etc. Estos métodos conllevan diversos riesgos de seguridad para los usuarios tales como: olvido de contraseña o PIN, robo o extravío de credencial o dispositivo de identificación, falsificación de credenciales físicas, o sencillamente que un tercero averigüe la contraseña por cualquier medio. [1].

Biometría es el estudio de la medición y el análisis de los rasgos físicos y de comportamiento que permiten distinguir entre individuos. A estos rasgos se les conoce como rasgos biométricos. Un sistema biométrico utiliza información extraída de uno o varios rasgos como: voz, rostro, huella digital, etc, con el fin de autenticar la identidad de un usuario. Esto permite que las personas inscritas dentro del sistema no necesiten memorizar información o disponer de alguna credencial o dispositivo para autenticar su identidad, dado que un rasgo de su cuerpo o su comportamiento son inherentes a ellas y no pueden ser extraviados, transferidos o robados [2]. Un sistema biométrico está compuesto de varios módulos que realizan el siguiente proceso para autenticar una identidad: el sensor adquiere la información de un rasgo biométrico para su posterior digitalización, después se extraen características discriminativas mediante algún método matemático, un conjunto de tales características es almacenado en el sistema biométrico y servirá como patrón de referencia; en cada ocasión en que un usuario haga uso del sistema las características extraídas de su muestra biométrica son comparadas con uno o varios patrones de referencia para generar una calificación numérica, finalmente un módulo utiliza dicha calificación para tomar la decisión de aceptar o rechazar la identidad de un usuario [3]. La figura 1 muestra los módulos que componen un sistema biométrico.

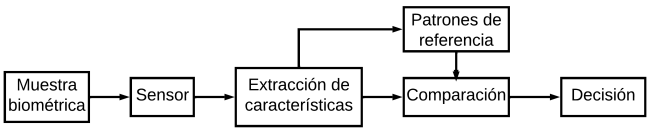

Fig. 1. Diagrama de bloques de un sistema biométrico.

Todos los sistemas biométricos poseen un cierto grado de vulnerabilidad en menor o mayor medida. Un impostor puede tratar de imitar los rasgos fisiológicos o comportamentales, usar algún dispositivo para falsificar dicho rasgo o simplemente utilizar grabaciones de audio o video para tratar de engañar al sistema [4]. Una estrategia relevante en el diseño de sistemas biométricos es el uso de sistemas multimodales, orientados a reforzar la seguridad del sistema y disminuir significativamente la posibilidad de reconocer erróneamente a un impostor. Para desarrollar un sistema multimodal se pue-

Juan Carlos Atenco Vazquez is with Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro 1, Sta. Maria Tonantzintla, San Andres Cholula, Puebla, Mexico. E-mail:atencovaz@inaoep.mx.

Juan Carlos Moreno Rodríguez is with Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro 1, Sta. Maria Tonantzintla, San Andres Cholula, Puebla, Mexico. E-mail:xalatl@inaoep.mx.

Juan Manuel Ramírez Cortés is with Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE), Luis Enrique Erro 1, Sta. Maria Tonantzintla, San Andres Cholula, Puebla, Mexico. E-mail:jmram@inaoep.mx.

Corresponding author: Juan Carlos Atenco Vázquez.

den combinar/fusionar diversas fuentes de información de las siguientes maneras: combinar los datos de varios rasgos adquiridos con los sensores, las características adquiridas mediante diferentes métodos, las calificaciones generadas al comparar las muestras de varios rasgos con sus respectivos patrones o las decisiones tomadas después de procesar varios rasgos [5]. El uso de dos o más rasgos biométricos para autenticar una identidad refuerza la seguridad del sistema, complicándole a un impostor la posibilidad de validar falsamente su identidad.

En este trabajo se propone el desarrollo y evaluación de una red neuronal convolucional bimodal (BiCNN) que lleva a cabo las tareas biométricas de identificación y verificación. Proponemos fusionar las modalidades de voz y rostro mediante dos métodos: concatenar o promediar sus correspondientes vectores de características. La fusión se lleva a cabo dentro de la arquitectura de la red, esto permite que los vectores de características extraídos se ajusten conforme a la optimización de los parámetros de la red durante la etapa de entrenamiento. Empleamos el operador Local Binary Pattern (LBP) como descriptor de textura para las imágenes de rostro y aprovechar su propiedad de ser invariante ante la rotación y los cambios de iluminación de una imagen. Para procesar la modalidad de voz incorporamos la capa convolucional de una sola dimensión llamada SincNet que extrae características frecuenciales a partir de señales de voz segmentadas; las ventajas del uso de esta capa convolucional es que permite el procesamiento directo de audio sin necesidad de emplear previamente un método de extracción de características, además contiene menos parámetros entrenables que una capa convolucional de dos dimensiones y permite una convergencia más rápida. Para realizar el entrenamiento y pruebas a nuestro modelo creamos una base de datos virtual que combina las bases de datos BIOMEX-DB y VidTimit. Incrementamos la cantidad de datos biométricos disponibles aumentando los datos de voz y rostro agregando ruido a las señales de voz y aplicando transformaciones a las imágenes.

Se seleccionaron las modalidades de voz y rostro para nuestro sistema bimodal por su fácil adquisición y por tener un grado mínimo de intrusión para los participantes. Los datos de estas modalidades son fácilmente manipulables, lo que facilita la implementación de un esquema de aumento artificial de datos para prevenir el sobreajuste del modelo. Otro factor importante a considerar es que ambos rasgos no son mutuamente dependientes, por lo tanto es posible combinar bases de datos con características similares para aumentar la cantidad de información disponible o como es el caso de nuestro trabajo completar información faltante, esto se explicará más adelante.

Por otra parte, la elección de un modelo de aprendizaje profundo (DL por sus siglas en inglés) para implementar nuestra propuesta se justifica en que este tipo de redes han superado varias de las limitaciones que se encuentran en aprendizaje de máquina (ML por sus siglas en inglés), siendo una de las más importantes la extracción de características. Los métodos de ML requieren que los datos con que se entrenarán y evaluarán sus respectivos modelos sean procesados para extraer sus características mediante algún procedimiento matemático. Sin embargo, estos procedimientos pueden no ser compatibles con ciertos tipos de datos biométricos, con diferentes bases de datos o si la información presenta alguna alteración como ruido o que hayan sufrido alguna transformación [6]. Aprendizaje profundo ofrece la posibilidad de que el mismo modelo extraiga las características de la información que recibe como entrada y ajuste sus parámetros durante la etapa de entrenamiento mejorando el desempeño de reconocimiento final. Por lo anterior es importante mencionar que la fusión multimodal de vectores de características dentro de la arquitectura de red también se ve beneficiada de esta optimización.

El trabajo se organiza de la siguiente manera: la sección II presenta un panorama de trabajos importantes de biometría multimodal enfocados a fusión de vectores de características mediante modelos de redes neuronales o aprendizaje de máquina, en la sección III se explican los detalles más importantes del entrenamiento de la BiCNN, la sección IV presenta las condiciones de evaluación de desempeño de la red y los correspondientes resultados, finalmente la sección V presenta las conclusiones.

II. Trabajo relacionado

Las modalidades de voz y rostro presentan características relevantes dentro de las propiedades buscadas en un sistema biométrico, y continúan siendo objeto de investigación a través de enfoques unimodales. En particular, algunos de los enfoques modernos se concentran en el uso de técnicas de aprendizaje de máquina e inteligencia artificial con métodos basados en aprendizaje profundo para biometría basada en voz [7], [8], y biometría basada en rasgos faciales con diversas variantes tales como combinación de imágenes térmicas y visuales [9], empleando aprendizaje 'one shot' con aumento de datos [10] o mapeo de las imágenes hacia un espacio euclidiano usando redes convolucionales [11].

En la literatura de biometría multimodal se pueden encontrar diversos métodos para fusionar la información de varios rasgos biométricos. Estos métodos dependen de las características de los rasgos considerados y de los clasificadores empleados. La tabla I lista algunos trabajos relevantes sobre biometría multimodal.

III. Desarrollo del sistema biométrico bimodal

El sistema biométrico bimodal propuesto en el presente trabajo está compuesto por una red convolucional de dos entradas donde cada una corresponde a un rasgo biométrico. Cada entrada se conecta a una subred que procesa de manera independiente la información de cada rasgo y entrega un vector de características. Los dos vectores son concatenados y forman un *vector bimodal* el cual se procesa con una capa completamente conectada para seguir extrayendo más información importante y reducir su dimensión. Finalmente, se tiene una capa completamente conectada que servirá como un clasificador softmax. A continuación se describe a detalle la arquitectura de la BiCNN.

TABLA I
REVISIÓN DE TRABAJOS SOBRE BIOMETRÍA MULTIMODAL.

Autor	Modalidades	Método	Base de datos	Resultados
Talreja et al. 2017 [12]	Iris y rostro.	Concatenación de vectores extraídos mediante red neuronal basada en arquitectura VGG-19.	CASIA-Webface, ND-Iris-0405 y WVU-Multimodal.	Verificación: Genuine Acceptance Rate (GAR) 99.65 %.
Zhang et al. 2018 [13]	Iris y área periférica del ojo.	Concatenación de vectores ponderados extraídos de red neuronal con unidades maxout.	CASIA-IrisV4 y CASIA-CSIR2015.	Verificación: EER 1.88 %.
Xin et al. 2018 [14]	Rostro, huella dactilar y venas de los dedos.	Concatenación de vectores Fisher calculados a partir de vectores de características.	Adquirida por los autores.	Identificación: Exactitud 50 sujetos 88 %. 20 sujetos 90 %. 15 sujetos 93 %.
Khryashchev et al. 2018 [15]	Rostro y voz.	Fusión de scores por modelos de Mezclas Gaussianas Mixtas (GMM), Modelos Universales Alternativos (UBM) y red neuronal.	Adquirida por los autores.	Identificación: Exactitud 95 %.
Abozaid et al. 2019 [16]	Rostro y voz	Extracción características mediante varios métodos. La fusión de información se llevó a cabo a nivel de características y a nivel de scores.	Adquirida por los autores.	Verificación: EER 0.62 %.
Olazabal et al. 2019 [17]	Rostro y voz.	Los vectores de características de las dos modalidades fueron utilizados para entrenar un clasificador de K vecinos más próximos (KNN).	CSUF-SG5.	Verificación: EER 8.04 %.
Alay et al. 2020 [6]	Iris, rostro y venas de los dedos.	Concatenación de vectores de características extraídos por una red VGG-16 y fusión de scores.	SDUMLA-HMT, IT Delhi y FERET.	Identificación: Exactitud Concatenación 99.39 %. Fusión de scores 100 %.
Zhang et al. 2020 [18]	Rostro y voz.	Fusión de scores generados por la comparación de características LBP y un modelo GMM mediante una suma ponderada.	XJTU	Verificación: True Acceptance Rate (TAR) 100 %. False Rejection Rate (FRR) 0 %. False Acceptance Rate (FRR) 0 %
Alkeem et al. 2021 [19]	Electrocardiograma (ECG), rostro y huella dactilar.	Concatenación de vectores de características extraídos por una red neuronal multitarea. También se hicieron pruebas con fusión de scores por diferentes métodos.	Base de datos virtual creada a partir de las bases de datos ECG-ID, PTBECG, Faces95 y FVC2006.	Identificación: Exactitud Fusión de características 98.97 %. Regla de la suma 98.95 %. Regla del producto 96.55 %.
Leghari et al. 2021 [20]	Huella dactilar y firma digital.	Concatenación de vectores de características extraídas por una red convolucional. La concatenación ocurre en dos puntos de la arquitectura.	Adquirida por los autores.	Identificación: Exactitud Fusión temprana 99.1 %. Fusión posterior 98.35 %.
Costa-Filho et al. 2022 [21]	Rostro y voz.	La fusión se llevó mediante la normalización y suma de scores de cada modalidad generados mediante comparación de vectores.	MOBIO	Verificación: Área bajo una curva ROC. 0.98
Kamlaskar et al. 2022 [22]	Iris y huella digital.	Fusión mediante correlación canónica y análisis de componentes principales (PCA).	SDUMLA-HMT	Identificación: Exactitud 100 %. Verificación: EER 0.176 %.

TABLA II
ARQUITECTURA DE LA BICNN.

Capas de procesamiento de rostro	Filtros/Neuronas	Dimensiones	Paso	Función de activación
Convolución 2D	32	3x3	1x1	LeakyReLU
Normalización de batch	-	-	-	-
Max Pooling 2D	-	2x2	1x1	-
Convolución 2D	64	5x5	1x1	LeakyReLU
Normalización de batch	-	-	-	-
Max Pooling 2D	-	2x2	1x1	-
Completamente conectada	512	-	-	LeakyReLU
Normalización de batch	-	-	-	-
Capas de procesamiento de voz				
SincNet	120	251	-	LeakyReLU
Normalización de batch	-	-	-	-
Max Pooling 1D	-	5	1	-
Convolución 1D	32	5	1	LeakyReLU
Normalización de batch	-	-	-	-
Max Pooling 1D	-	5	1	-
Convolución 1D	64	5	1	LeakyReLU
Normalización de batch	-	-	-	-
Max Pooling 1D	-	5	1	-
Completamente conectada	512	-	-	LeakyReLU
Normalización de batch	-	-	-	-
Fusión y salida				
Promediado/Concatenación	-	-	-	-
Completamente conectada	512	-	-	LeakyReLU
Dropout	0.5	-	-	-
Normalización de batch	-	-	-	-
Completamente conectada	45	-	-	Softmax

A. *Arquitectura de la Red*

La tabla II muestra los detalles de las capas que componen la arquitectura de nuestra red.

El procesamiento de las imágenes LBP de rostro se llevó a cabo con un conjunto de capas convolucionales de dos dimensiones. El número de filtros se incrementa con cada capa al igual que las dimensiones del kernel de convolución para capturar información más detalladamente. A los mapas de características se les aplica una operación de max pooling para reducir su tamaño y excluir información no relevante. Finalmente los mapas generados por la última capa convolucional son procesados por una capa completamente conectada para obtener un vector de características de dimensión fija de 512 puntos; este vector contiene la información discriminatoria de una imagen de rostro.

Las señales de voz fueron procesadas con la capa convolucional SincNet [23]. Esta capa está definida como un banco de filtros pasabanda que extraen características frecuenciales de las señales de voz. Los filtros pasabandas son definidos con funciones sinc en el dominio del tiempo como se muestra en la ecuación 1.

$$g[n, f_1, f_2] = 2f_1 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \quad (1)$$

donde f_1 y f_2 son las frecuencias de corte alta y baja respectivamente de los filtros pasabandas. Los valores de estas frecuencias son optimizados durante la fase de entrenamiento. Para nuestros experimentos las frecuencias de corte del banco de filtros fueron inicializadas en forma logarítmica de la misma manera en que se definen para obtener los MFCC [24].

Las características extraídas por la capa SincNet son procesadas con bloques de capas convolucionales en una dimensión,

aplicando una operación de max pooling para descartar la información menos relevante. Finalmente se utiliza una capa completamente conectada de 512 neuronas para obtener un vector que contenga la información de voz de un individuo.

Los vectores de rostro y voz se combinan mediante uno de dos esquemas: concatenación o promediado. Posteriormente empleamos una capa completamente conectada para reducir las dimensiones del vector de fusión a un tamaño fijo, esta capa posee 512 neuronas con un dropout de 0.5. La salida de la BiCNN es una capa completamente conectada que actúa como un clasificador softmax cuyo número de neuronas es igual a la cantidad de individuos inscritos en el sistema biométrico a los cuales se les denomina usuarios legítimos. Es importante hacer énfasis en que el vector de longitud 512 entregado por la penúltima es el que contiene las características de los rasgos de voz y rostro, en adelante nos referiremos a él como vector bimodal y su utilidad se describirá en la siguiente sección.

Todas las capas convolucionales, excepto la capa de salida, tienen normalización de batch como regularizador y como función de activación Leaky ReLU siguiendo las conclusiones expuestas en [25]. En este estudio se comparan el tiempo de convergencia de una red neuronal y su desempeño de reconocimiento con la base de datos MNIST cuando la arquitectura utiliza la función ReLU o Leaky ReLU. Siendo Leaky ReLU la función de activación con la que la red converge en menos épocas y entrega resultados muy similares a la función ReLU.

B. Preprocesamiento de Datos

En [4] se listan varios trabajos de biometría bimodal de voz y rostro implementados con diferentes métodos de fusión. Varios de estos trabajos contemplaron poblaciones relativamente pequeñas para sus experimentos. Esta situación ocurre debido a que existen pocas bases de datos multimodales de acceso público que contengan una gran población en la que todos los individuos tengan completa su información biométrica. Muchos de los autores citados recurrieron a generar sus propias bases de datos que se ajustaran a los requerimientos de sus estudios. El proceso de adquisición de datos biométricos requiere de muchos recursos por lo que no siempre es posible tener poblaciones grandes. Otra alternativa es combinar dos bases de datos emparejando a los individuos de cada una.

Para entrenar nuestra red bimodal utilizamos la información de voz y rostro de la base de datos BIOMEX-DB que fue generada por nuestro grupo de investigación [26]. Sin embargo, debido a que algunos sujetos no contaban con datos de la modalidad de rostro se tomó la decisión de completar esta información con el conjunto de imágenes de la base de datos bimodal VidTimit [27].

BIOMEX-DB contiene información de las modalidades de electroencefalograma (EEG), voz y rostro. La población total es de 51 sujetos conformada por 25 mujeres y 26 hombres. Los datos de voz consisten en grabaciones de pronunciaciones en español de cadenas de dígitos ordenadas aleatoriamente. Cada sujeto posee 20 archivos de audio, la mitad corresponde a pronunciaciones de cadenas de 10 dígitos, mientras que la otra mitad a cadenas de 5 dígitos. Los datos de rostro consisten en videos de rostro grabados mientras los participantes pronunciaban sus cadenas de dígitos. La base de datos VidTimit contiene información de voz y rostro y tiene una población de 43 individuos. Los datos de voz son pronunciaciones de 10 oraciones cortas en inglés y los datos de rostro son imágenes extraídas de videos grabados mientras pronunciaban tales oraciones. Es importante aclarar que 12 sujetos de BIOMEX-DB no cuentan con datos de rostro, por esta razón fue necesario completar esta información faltante con las imágenes de rostro de 12 sujetos de VidTimit como se mencionó anteriormente. De esta manera se conformó una población de 51 sujetos para nuestro estudio, los cuales se dividieron aleatoriamente en dos conjuntos: 45 usuarios legítimos y 6 impostores.

Los datos de voz fueron normalizados para que los valores de las señales se encontraran en el rango de [-1,1]. También se eliminaron silencios y pausas entre pronunciaciones con el objetivo de conservar solo la información de voz. El aumento artificial de las señales de voz se llevó a cabo de la siguiente manera: se agregó ruido de fondo tomado muestras de la base de datos MUSAN [28], el ruido se agregó en valores de 0 y 5 decibeles (dB) en términos de razón señal a ruido (SNR).

Las imágenes de rostro de cada individuo de BIOMEX-DB fueron extraídas de sus correspondientes videos. Posteriormente se utilizó el detector de rostro de la biblioteca OpenCV [29] para recortar el área de interés y obtener una imagen final de 128x128 pixeles en escala de grises. Finalmente se extrajeron características LBP con 8 pixeles de vecindad y un radio de 1. Estas características son robustas ante cambios de postura del rostro y variaciones en la iluminación de la imagen, además han sido empleadas con éxito en conjunto con redes convolucionales en tareas de clasificación que involucran imágenes de rostro [30], [31], [32]. Se empleó el mismo procedimiento de detección, recorte de rostro y obtención de características LBP para las imágenes tomadas de VidTimit. Para aumentar la cantidad de imágenes se realizaron dos transformaciones con la biblioteca imgaug [33]: en la primera se modificó la iluminación de cada imagen y en la segunda fueron rotadas aleatoriamente en ángulos entre [-45,45] grados, es importante mencionar que las transformaciones se aplicaron a las imágenes en escala de grises y el operador LBP fue aplicado a las imágenes resultantes de las transformaciones.

C. Entrenamiento de la Red Bimodal

Se utilizó la biblioteca Keras para definir la arquitectura y entrenar la red bimodal. El entrenamiento se llevó a cabo por 50 épocas con una tasa de aprendizaje de 0.001 utilizando el optimizador Adam. Se empleó como función de costo la entropía cruzada categórica. Como datos de entrenamiento y validación utilizamos las muestras de voz de pronunciaciones de cadenas de 10 dígitos y las imágenes de rostro de sus videos correspondientes. Para entrenar a la red se deben crear pares compuestos de una imagen con un segmento de señal de voz. Para mejorar la generalización de nuestro modelo, en cada época se crean de manera aleatoria 8 batches de 32 pares cada uno para entrenamiento y 8 batches de 16 pares cada uno para validación. Esto se hace en concordancia con el hecho de que la capa convolucional SincNet acepta segmentos de una señal de voz sin procesar con duración de varios milisegundos. Considerando que los videos de BIOMEX-DB tienen una tasa

de 8 cuadros por segundo, cada segmento de una señal de voz tiene una duración de 125 milisegundos. Finalmente se definió un monitoreo del valor de precisión de validación en cada época para que solo se guarden los valores de los parámetros que maximicen esta métrica.

La figura 2 muestra el diagrama a bloques de la red bimodal propuesta.

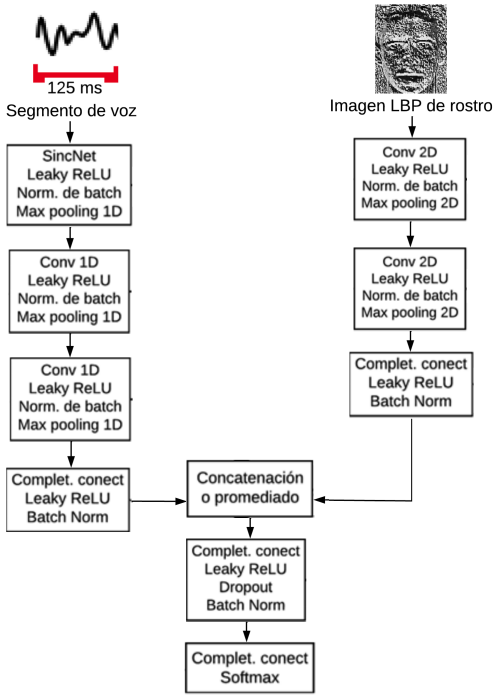

Fig. 2. Diagrama de bloques de la red bimodal.

IV. Desarrollo Experimental y Resultados

Uno de los objetivos principales de este trabajo es comparar el desempeño de nuestra BICNN al emplear dos métodos de fusión de características: concatenación y promediado.

Para llevar a cabo la evaluación se usaron los datos de pronunciaciones de cadenas de 5 dígitos y las imágenes de rostro de sus correspondientes videos.

Para evaluar el sistema bimodal utilizamos los datos biométricos resultantes del aumento artificial de datos junto con las muestras sin modificar, esto nos permitirá evaluar el desempeño de la red bajo diferentes condiciones de ambas modalidades. En algunos de los trabajos consultados de la sección II, los autores mencionan que también emplearon un esquema de aumento de datos para entrenar sus modelos, sin embargo, sus evaluaciones no ofrecen información acerca de si realizaron pruebas solo con sus datos originales o si incluyeron datos del conjunto generado artificialmente y como podrían afectar el desempeño de reconocimiento. Por esta razón presentamos de manera esquemática resultados que demuestran el efecto que tienen las muestras de voz y rostro generadas artificialmente en identificación y verificación. Asumiendo que estas muestras son una aproximación a condiciones de operación reales.

A. *Métricas para Evaluar Identificación*

En la tarea de identificación un usuario entrega sus datos biométricos al sistema y este último entregará el número de identificación o etiqueta del usuario legítimo inscrito en el sistema biométrico cuyas características se asemejen más a las del usuario. Es una comparación uno a muchos.

En nuestro trabajo el procedimiento anterior se lleva a cabo alimentando el clasificador softmax de salida con el vector bimodal, el resultado es una distribución normalizada de probabilidad de la cual podemos determinar el número de identificación del usuario legítimo cuyas características de voz y rostro se asemejen más a las encontradas en el vector bimodal mediante la función argmax.

Nuestros experimentos de identificación fueron realizados en modalidad de conjunto cerrado, lo que significa que el sistema asume que cualquier usuario está inscrito en él y su función es determinar su identidad y no verificarla o validarla. Por lo tanto, no se hicieron pruebas con datos pertenecientes a impostores ya que no contribuyen a evaluar correctamente el desempeño de la red. Tomando en cuenta lo descrito anteriormente, la métrica para medir el desempeño fue la exactitud definida en la ecuación 2.

$$Exactitud = \frac{Pc}{N}, \tag{2}$$

donde Pc es el número de veces en que el sistema predijo correctamente la identidad de un usuario, mientras que N es el número total de pruebas realizadas.

B. *Métricas para Evaluar Verificación*

La tarea de verificación consiste en comparar las características biométricas de un usuario con las de un patrón almacenado en el sistema que representa las características de un usuario legítimo específico, el resultado es la validación o rechazo de la identidad del usuario.

En nuestros experimentos las características biométricas que se compararon fueron los vectores bimodales. Para crear el patrón de un usuario legítimo se entregó a la BiCNN un conjunto de 10 pares de muestras de voz y rostro elegidas aleatoriamente, el patrón es generado al promediar los 10 vectores bimodales resultantes. Para llevar a cabo la verificación de identidad se compara el patrón de un usuario legítimo específico con un vector bimodal generado por un usuario que quiere validar su identidad. En este caso esta comparación se llevó a cabo mediante la función similitud coseno, la cual entrega un valor que entre mayor sea se considera que el patrón y el vector muestra pertenecen al mismo usuario legítimo. Adicionalmente se emplea un valor umbral para comparar con la similitud coseno y tomar la decisión de aceptar o rechazar la validación de dicha identidad. La ecuación 3 corresponde a la función similitud coseno.

$$similitud = cos(\theta) = \frac{\mathbf{P} \cdot \mathbf{M}}{\| \mathbf{P} \| * \| \mathbf{M} \|} \tag{3}$$

donde $\mathbf{P}$ es el patrón de un usuario legítimo y $\mathbf{M}$ es el vector muestra de un usuario que quiere verificar o validar su identidad.

Para medir el desempeño de un sistema biométrico en la tarea de verificación se emplea comúnmente la métrica conocida como Equal Error Rate (EER). Esta es definida como el punto en que el número de ejemplos positivos clasificados como negativos es igual al número de ejemplos negativos clasificados como positivos [34].

C. Resultados de Identificación

Los resultados de la tarea de identificación se muestran en la tabla III, en cada una de las condiciones de evaluación se realizaron 1000 pruebas para calcular el porcentaje de exactitud.

TABLA III
RESULTADOS DE IDENTIFICACIÓN EN EXACTITUD (%) DE LA RED BIMODAL CON PROMEDIADO/CONCATENACIÓN DE VECTORES DE VOZ Y ROSTRO.

SNR (dB)	Transformaciones		
	Sin transformación	Iluminación	Rotación
Sin ruido	98.4 / 99.3	97.5 / 99	84.5 / 91.4
0	98.9 / 99	97.9 / 99	83.5 / 89.6
5	99 / 99.1	97.7 / 99	85 / 90.5

Los resultados muestran que para la gran mayoría de condiciones se obtuvo una precisión mayor al 97 %, mientras que las condiciones que involucran imágenes con rotación entregaron valores menores entre 83 y 92 %. Se observa que la fusión por concatenación produce mejores resultados en comparación a promediado, esta brecha es más significativa en las condiciones que incluyen rotación de imágenes. Este último factor puede indicar que las diferentes condiciones de una imagen de rostro pueden afectar en mayor medida al valor de exactitud que el ruido de las señales de voz.

D. Resultados de Verificación

La evaluación de la tarea de verificación se llevó a cabo haciendo 1000 pruebas con muestras de usuarios legítimos y otras 1000 con muestras del conjunto de impostores para cada condición contemplada en este estudio. La tabla IV muestra los resultados en términos de EER entregados por el sistema biométrico bimodal.

TABLA IV
RESULTADOS DE VERIFICACIÓN EN TÉRMINOS DE EER (%) DE LA RED BIMODAL CON PROMEDIADO/CONCATENACIÓN DE VECTORES.

SNR (dB)	Transformaciones		
	Sin transformación	Iluminación	Rotación
Sin ruido	0.2 / 0.61	0.93 / 0.8	5.66 / 3.71
0	0.15 / 0.75	0.92 / 0.96	4.97 / 4.22
5	0.14 / 0.57	1.01 / 0.71	5.81 / 3.77

Se puede observar que en la gran mayoría de los casos los valores de EER son menores al 1 % lo cual es indicativo de un buen desempeño de verificación. De manera análoga a la tarea de identificación las condiciones que involucran rotación de imágenes obtuvieron valores significativamente mayores en comparación con el resto. Al comparar los valores de EER de ambos métodos de fusión de características se observa que promediado entregó resultados más competentes en esta instancia, siendo que en las condiciones en que las imágenes no fueron transformadas en las que fusión por promediado obtuvo mejores valores.

Al analizar las diferencias que hay entre valores de EER nuevamente se aprecia que la brecha es más notable al comparar entre transformaciones de imágenes que entre los diferentes valores de SNR; por lo que es posible que también en la tarea de verificación las condiciones de las imágenes de rostro tengan mayor influencia en los resultados que el ruido de las muestras de voz.

E. Comparación de Resultados

En la última etapa de evaluación se entrenaron algunos sistemas unimodales de voz y rostro con el fin de demostrar el mejor desempeño que entregan los sistemas multimodales en comparación con los primeros. Las aproximaciones elegidas para esta comparación son ampliamente utilizadas en la literatura, su estructura las hace compatibles para ser entrenadas con diferentes bases de datos y facilitan la reproducción de los resultados de este trabajo. Aproximaciones multimodales podrían no ser implementables con diferentes bases de datos dado que los métodos de extracción de características podrían no ser compatibles con determinados datos biométricos [20]. En el caso de implementaciones con DL otro factor a considerar es que una arquitectura diseñada para ser entrenada con una base de datos con determinadas características podría no converger con otras bases de datos.

Se escogieron 2 sistemas unimodales por cada rasgo biométrico, las condiciones de entrenamiento y evaluaciones fueron las mismas que para la BiCNN: se tomaron los datos biométricos de BIOMEX-DB y VidTimit y se consideraron la misma cantidad de usuarios legítimos y de impostores.

Los sistemas de reconocimiento de rostro considerados son una red ResNet con 4 capas residuales [35], esta arquitectura se eligió dado que agregando más capas residuales no se consiguieron resultados significativamente mejores. El segundo sistema de reconocimiento de rostro está basado en Eigenfaces, empleamos la definición que se encuentra en la biblioteca OpenCV con los parámetros descritos en [36]. Por otro lado los sistemas de reconocimiento de hablante elegidos fueron la implementación original de la red SincNet [23] y un sistema basado en X-vectors como está descrito originalmente en [37]. En las referencias [38], [39] se puede consultar sobre los sistemas de reconocimiento de rostro empleados en este estudio, en [40], [41] se halla información sobre los sistemas de reconocimiento de hablante.

La tabla V muestra una comparación de los mejores resultados de nuestros experimentos de identificación con los obtenidos con los sistemas unimodales.

En la modalidad de identificación de rostro todos los modelos tuvieron una exactitud superior al 97 % para los casos de imágenes sin transformar y con cambio de iluminación. Para

TABLA V
COMPARACIÓN DE RESULTADOS DE EXACTITUD (%) DE LA TAREA DE IDENTIFICACIÓN.

Sistema biométrico	Transformaciones		
	Sin transformación	Iluminación	Rotación
Red Bimodal (Promediado)	99	97.9	85
Red Bimodal (Concatenación)	99.3	99	91.4
CNN ResNet	100	99.5	93.5
Eigenfaces	100	98.66	85.33
		SNR (dB)	
	Sin ruido	0	5
Red Bimodal (Promediado)	98.4	98.9	99
Red Bimodal (Concatenación)	99.3	99	99.1
SincNet	100	88	96.33
X-vectors	95.11	83	92

TABLA VI
COMPARACIÓN DE RESULTADOS DE EER (%) DE LA TAREA DE VERIFICACIÓN..

Sistema biométrico	Transformaciones		
	Sin transformación	Iluminación	Rotación
Red Bimodal (Promediado)	0.14	0.92	4.97
Red Bimodal (Concatenación)	0.57	0.71	3.71
CNN ResNet	0.54	1.43	3.58
Eigenfaces	2.96	12.82	27.77
		SNR (dB)	
	Sin ruido	0	5
Red Bimodal (Promediado)	0.2	0.15	0.14
Red Bimodal (Concatenación)	0.61	0.75	0.57
SincNet	1.81	13.39	5.55
X-vectors	1.72	4.92	2.37

la condición de rotación hay una disminución considerable de los valores de exactitud en las cuales solo la red bimodal con concatenación de características y el modelo CNN ResNet entregaron resultados apenas superiores al 90 %. En este análisis nuestro modelo bimodal y el modelo CNN ResNet tienen resultados similares.

Con respecto a identificación de hablante cabe destacar que los dos métodos de fusión de la red bimodal tuvieron resultados competentes superiores al 95 % en comparación con las propuestas unimodales cuando se emplean señales de voz sin ruido. La mayor diferencia destaca en los dos casos de ruido en las señales de voz, donde los dos métodos de fusión de la red bimodal entregaron valores de exactitud muy superiores a los sistemas unimodales. Esto nos indica que nuestra propuesta bimodal obtiene un alto desempeño de identificación a pesar del ruido, mientras que los otros sistemas sufren de una disminución notoria en sus respectivos desempeños.

La tabla VI muestra la comparación de desempeño de la tarea de verificación en términos de EER.

En verificación de rostro la red bimodal con sus dos métodos de fusión y el modelo CNN ResNet entregaron resultados con poca diferencia entre sus valores de EER para todas las transformaciones de imagen. En la condición de rotación todos los modelos mostraron un incremento de EER, esta es una situación similar a lo visto en identificación. Eigenfaces siendo un método no basado en redes neuronales mostró el peor desempeño en todas las condiciones por una diferencia significativa.

En cuanto a verificación en la modalidad de voz, la red bimodal con promediado obtuvo los mejores resultados en los tres casos evaluados, mientras que la fusión por concatenación demostró un desempeño ligeramente menor. Ambas redes bimodales superaron a las aproximaciones SincNet y X-vectors por un margen considerable en los casos en que las señales de voz están contaminadas con ruido. Estos resultados muestran evidencia de que nuestra red tiene un mejor desempeño de verificación en comparación con las otras aproximaciones cuando opera con muestras de voz con ruido agregado a estos valores de SNR.

Finalmente, acumulamos todos los scores empleados para evaluar las condiciones de cada sistema biométrico para generar las correspondientes curvas ROC (Receiver Operating Characteristic). La figura 3 muestra las curvas ROC de todos los modelos evaluados.

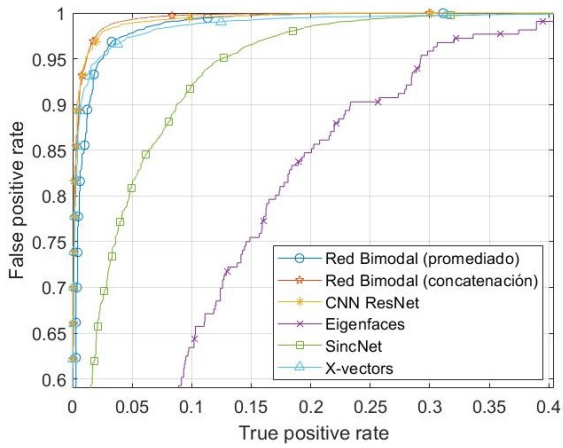

Fig. 3. Comparación de curvas ROC.

Los valores de EER generales obtenidos mediante las curvas ROC son los siguientes: Red Bimodal (promediado) 3.27 %, Red Bimodal (concatenación) 2.15 %, CNN ResNet 2.36 %, Eigenfaces 17.83 %, SincNet 9.31 % y X-vectors 3.51 %. Estos valores demuestran que nuestra propuesta bimodal en sus dos variantes entregó resultados competentes y en la mayoría de los casos superiores a los sistemas unimodales considerados. El método de fusión mediante concatenación de vectores fue el mostró el mejor desempeño de verificación, siendo este resultado consistente con lo obtenido en las anteriores evaluaciones.

La tabla VII muestra una breve comparación de nuestros resultados con otros trabajos relevantes de biometría bimodal de voz y rostro. Seleccionamos trabajos cuyo número de sujetos en sus experimentos sea similar o no mucho mayor al nuestro. Como se puede observar nuestros resultados son competentes con los presentados por otros autores que implementaron otras técnicas y emplearon otras bases de datos.

TABLA VII
RESULTADOS DE VARIOS TRABAJOS SOBRE BIOMETRÍA BIMODAL DE VOZ Y ROSTRO.

Autor	Clasificador	Base de datos y población	EER (%)
Gofman et al. 2018 [42]	Máquina de soporte vectorial (SVM).	CSUF-SG5: 27.	20.59
Abozaid et al. 2019 [16]	Varios clasificadores.	Adquirida por los autores: 100.	0.65
Olazabal et al. 2019 [17]	KNN.	CSUF-SG5: 27.	8.04
Ramachandra et al. 2019 [43]	FaceNet + Red residual dilatada (DRN).	SWAN: 88.	3.1
Antipov et al. 2020 [44]	Regresión logística.	NIST SRE19: 47.	2.78
Sadjadi et al. 2020 [45]	No especificado.	NIST SRE19: 42.	2.78
Zhang et al. 2020 [18]	Comparación de características LBP y GMM	XJTU: 102.	0
Liu et al. 2021 [46]	CNN de 2 entradas	Deep Lip (base de datos virtual): 150.	1.11
Fenu et al. 2022 [47]	ResNet-50.	VoxCeleb1-Test: 40. MOBIO: 150. AveRobot: 111.	5.6 H, 5.3 M 7.6 H, 9 M 29.9 H, 31.5 M
Nuestra propuesta.	CNN de 2 entradas.	BIOMEX-DB y VidTimit: 51.	Promediado: 3.27 general, 0.14 mejor Concatenación: 2.15 general, 0.57 mejor

V. CONCLUSIONES

Se desarrolló un sistema biométrico bimodal para llevar a cabo las tareas de reconocimiento de identificación y verificación. Este sistema consiste en una red convolucional de dos entradas donde cada una procesa los datos biométricos de cada modalidad. La entrada correspondiente a la modalidad de rostro procesa imágenes obtenidas mediante el operador LBP para que las capas convolucionales extraigan características que aprovechen las propiedades de robustez ante cambios de iluminación y de rotación de dicho operador. Por otro lado el uso de la capa SincNet para extraer características frecuenciales de las señales de voz tiene la ventaja de que es posible utilizar las señales de manera directa, otra propiedad importante de esta capa es que tiene menos parámetros entrenables lo que favorece una convergencia más rápida.

Las capas que extraen características de cada modalidad entregan un vector de longitud 512 que contiene la información más importante sobre la identidad de un individuo. En nuestro trabajo se estudian dos métodos de fusión de características: promediado o concatenación de vectores de características. Finalmente, el vector bimodal resultante de la fusión se emplea en los procedimientos de identificación y verificación.

Creamos una base de datos virtual combinando la información de las bases de datos bimodales BIOMEX-DB y VidTimit. Empleamos un esquema de aumento de datos para incrementar la cantidad de imágenes de rostro y de señales de voz realizando transformaciones a las imágenes de rostro y añadiendo ruido a las señales de voz a diferentes valores de SNR en decibeles. Empleamos las condiciones contempladas en el proceso de aumento de datos para evaluar el desempeño de reconocimiento de nuestra BiCNN.

Los resultados de nuestras evaluaciones demostraron que la red bimodal con fusión mediante concatenación obtuvo mejores resultados de reconocimiento en la mayoría de las condiciones consideradas tanto en identificación como verificación.

En la segunda etapa de evaluación se entrenaron propuestas unimodales populares con la misma base de datos virtual y se realizaron pruebas bajo las mismas condiciones que la BiCNN. Si bien desde la perspectiva de la modalidad de rostro los resultados son muy similares a la red unimodal basada en ResNet, nuestra propuesta demostró tener un desempeño superior al sistema de Eigenfaces y a las dos aproximaciones unimodales de voz en la mayoría de condiciones consideradas.

Finalmente presentamos una breve comparación de nuestros resultados con aquellos obtenidos en otros trabajos que emplearon otras técnicas, se observó que nuestra propuesta entregó valores de EER competentes y comparables con los demás trabajos en los cuales la cantidad de población considerada en los experimentos es similar a la nuestra.

LIMITACIONES Y TRABAJO FUTURO

Las limitaciones de nuestro trabajo son las siguientes: debido a que contamos con poca información de nuestra base de datos no pudimos realizar experimentos para agregar o remover individuos de nuestro sistema biométrico, por lo que en el estado actual sería necesario reentrenar la red si se quieren llevar a cabo estas operaciones. Como trabajo a futuro se pueden recabar más datos biométricos para realizar las operaciones de agregar nuevos individuos. También se puede considerar dentro del procedimiento de reconocimiento el contenido lingüístico de las señales de voz y el movimiento de los labios durante las pronunciaciones de dígitos.

AGRADECIMIENTOS

Los dos primeros autores agradecen a CONACYT (Consejo Nacional de Ciencia y Tecnología-México) por la beca otorgada para la realización de estudios de doctorado.

REFERENCIAS

[1] T. Sabhanayagam, V. P. Venkatesan, and K. Senthamaraikannan, "A comprehensive survey on various biometric systems," *International Journal of Applied Engineering Research*, vol. 13, no. 5, pp. 2276–2297, 2018.

[2] S. K. S. Modak and V. K. Jha, "Multibiometric fusion strategy and its applications: a review," *Information Fusion*, vol. 49, pp. 174–204, 2019.

[3] W. Dahea and H. Fadewar, "Multimodal biometric system: A review," *International Journal of Research in Advanced Engineering and Technology*, vol. 4, no. 1, pp. 25–31, 2018.

[4] H. Mandalapu, A. R. PN, R. Ramachandra, K. S. Rao, P. Mitra, S. M. Prasanna, and C. Busch, "Audio-visual biometric recognition and presentation attack detection: A comprehensive survey," *IEEE Access*, vol. 9, pp. 37431–37455, 2021.

[5] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187–205, 2019.

[6] N. Alay and H. H. Al-Baity, "Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits," *Sensors*, vol. 20, no. 19, p. 5523, 2020.

[7] S. Shakil, D. Arora, and T. Zaidi, "Feature based classification of voice based biometric data through machine learning algorithm," *Materials Today: Proceedings*, vol. 51, pp. 240–247, 2022.

[8] N. D. AL-Shakarchy, H. K. Obayes, and Z. N. Abdullah, "Person identification based on voice biometric using deep neural network," *International Journal of Information Technology*, pp. 1–7, 2022.

[9] N. K. Benamara, E. Zigh, T. B. Stambouli, and M. Keche, "Towards a robust thermal-visible heterogeneous face recognition approach based on a cycle generative adversarial network," *Int. J. Interact. Multimed. Artif. Intell*, 2022.

[10] D. M. Jiménez-Bravo, Á. Lozano Murciego, A. Sales, L. Augusto Silva, and D. H. De La Iglesia, "Edge face recognition system based on one-shot augmented learning," 2022.

[11] A. Alcaide, M. A. Patricio, A. Berlanga, A. Arroyo, and J. J. Cuadrado Gallego, "Lipsnn: A light intrusion-proving siamese neural network model for facial verification," 2022.

[12] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, "Multibiometric secure system based on deep learning," in *2017 IEEE Global conference on signal and information processing (globalSIP)*, pp. 298–302, IEEE, 2017.

[13] Q. Zhang, H. Li, Z. Sun, and T. Tan, "Deep feature fusion for iris and periocular biometrics on mobile devices," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2897–2912, 2018.

[14] Y. Xin, L. Kong, Z. Liu, C. Wang, H. Zhu, M. Gao, C. Zhao, and X. Xu, "Multimodal feature-level fusion for biometrics identification system on iomt platform," *IEEE Access*, vol. 6, pp. 21418–21426, 2018.

[15] V. V. Khryashchev, A. I. Topnikov, A. F. Stefanidi, and A. L. Priorov, "Bimodal person identification using voice data and face images," in *Eleventh International Conference on Machine Vision (ICMV 2018)*, vol. 11041, pp. 296–303, SPIE, 2019.

[16] A. Abozaid, A. Haggag, H. Kasban, and M. Eltokhy, "Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion," *Multimedia tools and applications*, vol. 78, pp. 16345–16361, 2019.

[17] O. Olazabal, M. Gofman, Y. Bai, Y. Choi, N. Sandico, S. Mitra, and K. Pham, "Multimodal biometrics for enhanced iot security," in *2019 IEEE 9th annual computing and communication workshop and conference (CCWC)*, pp. 0886–0893, IEEE, 2019.

[18] X. Zhang, D. Cheng, P. Jia, Y. Dai, and X. Xu, "An efficient android-based multimodal biometric authentication system with face and voice," *IEEE Access*, vol. 8, pp. 102757–102772, 2020.

[19] E. Al Alkeem, C. Y. Yeun, J. Yun, P. D. Yoo, M. Chae, A. Rahman, and A. T. Asyhari, "Robust deep identification using ecg and multimodal biometrics for industrial internet of things," *Ad Hoc Networks*, vol. 121, p. 102581, 2021.

[20] M. Leghari, S. Memon, L. D. Dhomeja, A. H. Jalbani, and A. A. Chandio, "Deep feature fusion of fingerprint and online signature for multimodal biometrics," *Computers*, vol. 10, no. 2, p. 21, 2021.

[21] C. Costa-Filho, J. Negreiro, and M. Costa, "Multimodal biometric system based on autoencoders and learning vector quantization," in *XXVII Brazilian Congress on Biomedical Engineering: Proceedings of CBEB 2020, October 26–30, 2020, Vitória, Brazil*, pp. 1611–1617, Springer, 2022.

[22] C. Kamlaskar and A. Abhyankar, "Feature level fusion framework for multimodal biometric system based on cca with svm classifier and cosine similarity measure," *Australian Journal of Electrical and Electronics Engineering*, pp. 1–14, 2022.

[23] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, IEEE, 2018.

[24] A. Mahmood and K. Utku, "Speech recognition based on convolutional neural networks and mfcc algorithm," *Advances in Artificial Intelligence Research*, vol. 1, no. 1, pp. 6–12, 2021.

[25] A. K. Dubey and V. Jain, "Comparative study of convolution neural network's relu and leaky-relu activation functions," in *Applications of Computing, Automation and Wireless Systems in Electrical Engineering*, pp. 873–880, Springer, 2019.

[26] J. C. Moreno-Rodriguez, J. C. Atenco-Vazquez, J. M. Ramirez-Cortes, R. Arechiga-Martinez, P. Gomez-Gil, and R. Fonseca-Delgado, "Biomex-db: A cognitive audiovisual dataset for unimodal and multimodal biometric systems," *IEEE Access*, vol. 9, pp. 111267–111276, 2021.

[27] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *International conference on biometrics*, pp. 199–208, Springer, 2009.

[28] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015. arXiv:1510.08484v1.

[29] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[30] M. Wang, Z. Wang, and J. Li, "Deep convolutional neural network applies to face recognition in small and medium databases," in *2017 4th International Conference on Systems and Informatics (ICSAI)*, pp. 1368–1372, IEEE, 2017.

[31] P. Ke, M. Cai, H. Wang, and J. Chen, "A novel face recognition algorithm based on the combination of lbp and cnn," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 539–543, IEEE, 2018.

[32] Q. Xu and N. Zhao, "A facial expression recognition algorithm based on cnn and lbp feature," in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, pp. 2304–2308, IEEE, 2020.

[33] A. B. Jung, K. Wada, J. Crall, S. Tanaka, J. Graving, C. Reinders, S. Yadav, J. Banerjee, G. Vecsei, A. Kraft, Z. Rui, J. Borovec, C. Vallentin, S. Zhydenko, K. Pfeiffer, B. Cook, I. Fernández, F.-M. De Rainville, C.-H. Weng, A. Ayala-Acevedo, R. Meudec, M. Laporte, *et al.*, "imgaug." https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020.

[34] J.-M. Cheng and H.-C. Wang, "A method of estimating the equal error rate for automatic speaker verification," in *2004 International Symposium on Chinese Spoken Language Processing*, pp. 285–288, IEEE, 2004.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[36] I. Aliyu, M. A. Bomoi, and M. Maishanu, "A comparative study of eigenface and fisherface algorithms based on opencv and sci-kit libraries implementations.," *International Journal of Information Engineering & Electronic Business*, vol. 14, no. 3, 2022.

[37] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5329–5333, IEEE, 2018.

[38] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, p. 342, 2020.

[39] A. Verma, A. Goyal, N. Kumar, and H. Tekchandani, "Face recognition: a review and analysis," *Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021*, pp. 195–210, 2022.

[40] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[41] A. Q. Ohi, M. F. Mridha, M. A. Hamid, and M. M. Monowar, "Deep speaker recognition: Process, progress, and challenges," *IEEE Access*, vol. 9, pp. 89619–89643, 2021.

[42] M. Gofman, N. Sandico, S. Mitra, E. Suo, S. Muhi, and T. Vu, "Multimodal biometrics via discriminant correlation analysis on mobile devices," in *Proceedings of the International Conference on Security and Management (SAM)*, pp. 174–181, The Steering Committee of The World Congress in Computer Science, Computer . . . , 2018.

[43] R. Ramachandra, M. Stokkenes, A. Mohammadi, S. Venkatesh, K. Raja, P. Wasnik, E. Poiret, S. Marcel, and C. Busch, "Smartphone multimodal biometric authentication: Database and evaluation," *arXiv preprint arXiv:1912.02487*, 2019.

[44] G. Antipov, N. Gengembre, O. L. Blouch, and G. L. Lan, "Automatic quality assessment for audio-visual verification systems. the love submission to nist sre challenge 2019," *arXiv preprint arXiv:2008.05889*, 2020.
[45] S. O. Sadjadi, C. S. Greenberg, E. Singer, D. A. Reynolds, L. P. Mason, J. Hernandez-Cordero, *et al.*, "The 2019 nist audio-visual speaker recognition evaluation.," in *Odyssey*, pp. 259–265, 2020.
[46] M. Liu, L. Wang, K. A. Lee, H. Zhang, C. Zeng, and J. Dang, "Exploring deep learning for joint audio-visual lip biometrics," *arXiv preprint arXiv:2104.08510*, 2021.
[47] G. Fenu and M. Marras, "Demographic fairness in multimodal biometrics: A comparative analysis on audio-visual speaker recognition systems," *Procedia Computer Science*, vol. 198, pp. 249–254, 2022.

Juan Carlos Atenco-Vázquez He received the B.Sc. degree in 2015 from the Puebla Institute of Technology (ITP), Mexico, and the M.Sc. degree in 2018 from the National Institute of Astrophysics, Optics, and Electronics (INAOE), Mexico. He is currently a Ph.D. student at the Department of Electronics, INAOE, in Mexico. His research interests include signal processing, biometric systems, embedding systems, neural networks and applications.

Juan Carlos Moreno-Rodríguez He received the B.Sc. degree in 1995 and his M.Sc. degree in 1998 in Electronic Engineering from Universidad de las Américas-Puebla. He received the Ph.D. degree in electronics in 2021 at National Institute of Astrophysics, Optics and Electronics, Mexico. He has worked as an Assistant Professor in the departments of Computer Systems and Electronics at Tecnologico Nacional de Mexico and Universidad Iberoamericana, respectively. His research interest includes biometrics, machine learning and signal processing.

Juan Manuel Ramírez-Cortés He received the B.Sc. degree from the National Polytechnic Institute, Mexico, the M.Sc. degree from the National Institute of Astrophysics, Optics, and Electronics (INAOE), Mexico, and the Ph.D. from Texas Tech University, all in electrical engineering. He currently holds a researcher position at INAOE. He is member of the Mexican national research system (SNI), level 2. His research interests include signal and image processing, biometric, neural networks, fuzzy logic, and digital systems.