Enhancing Biometric Security with Bimodal Deep Learning and Feature-level Fusion of Facial and Voice Data

Khaled Merit and Mohammed Beladgham

Tahri Mohammed University of Bechar, Bechar, Algeria

https://doi.org/10.26636/jtit.2024.4.1754

Abstract — Recent research in biometric technologies underscores the benefits of multimodal systems that use multiple traits to enhance security by complicating the replication of samples from genuine users. To address this, we present a bimodal deep learning network (BDLN or BNet) that integrates facial and voice modalities. Voice features are extracted using the SincNet architecture, and facial image features are obtained from convolutional layers. Proposed network fuses these feature vectors using either averaging or concatenation methods. A dense connected layer then processes the combined vector to produce a dual-modal vector that encapsulates distinctive user features. This dual-modal vector, processed through a softmax activation function and another dense connected layer, is used for identification. The presented system achieved an identification accuracy of 99% and a low equal error rate (EER) of 0.13% for verification. These results, derived from the VidTimit and BIOMEX-DB datasets, highlight the effectiveness of the proposed bimodal approach in improving biometric security.

Keywords — biometric recognition, deep learning, multimodal systems, SincNet, voice modality

1. Introduction

Recently, there has been considerable growth in the number of applications that require the verification of an individual's identity. This is applicable to digital services provided by public or private entities and various other processes, such as forensic sciences or security tasks. Traditional methods of identity verification include memorizing passwords, relying on physical credentials or electronic devices containing user information. However, these methods pose security risks. For example, people may forget their password, lose their credentials or identification device, have their physical credentials forged, or discover their password through unauthorized means [1].

The scientific study of measuring and analyzing distinctive physical and behavioral traits that enable individual identification is known as biometrics. A biometric system utilizes data extracted from one or more biometric features, such as voice, face recognition, or fingerprints, to authenticate a user's identity. This approach eliminates the need for individuals to memorize information, possess credentials, or carry devices for authentication, as a unique characteristic of their body or behavior is inherent to them and cannot be lost, transferred, or stolen [2].

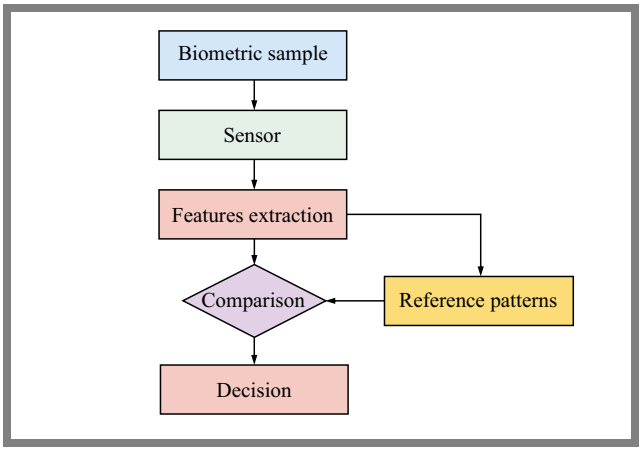

Fig. 1. Block diagram of a biometric system.

A biometric system is made up of several interconnected modules that perform the following steps to authenticate an individual's identity. The sensor acquires and digitizes biometric information, mathematical methods are used to extract discriminatory characteristics, and a set of these characteristics is stored in the biometric system to serve as a reference standard. Each time a user accesses the system, the characteristics extracted from their biometric sample are compared with one or more reference standards to generate a numerical rating, and finally a module uses this rating to decide whether to accept or reject the user's identity [3]. Figure 1 illustrates the modules that make up a biometric system.

Biometric systems are inherently vulnerable to various degrees of deception, either through imitation of physiological or behavioral traits, the use of devices to falsify these traits, or the use of audio or video recordings to deceive the system [4].

To address this issue, the use of multimodal systems is a relevant strategy in the design of biometric systems. By combining various sources of information, such as combining data from multiple traits acquired by sensors, characteristics acquired by different methods, ratings generated by comparing samples of various traits with their respective patterns, or decisions made after processing several traits [5], the security of the system can be significantly strengthened. Thanks to this, the likelihood of erroneously recognizing an impostor

can be significantly reduced. Furthermore, using two or more biometric features to authenticate an identity can further enhance the system safety, making it more challenging for an imposter to validate one's identity falsely.

In addition to using established techniques, this paper presents a novel approach to feature-level fusion by integrating SincNet for voice feature extraction and local binary patterns (LBPs) for face features, optimizing their combination through deep learning techniques. This work introduces and evaluates a biometric neural network, termed BiCNN, designed for identification and verification tasks. The integration of voice and facial modalities is achieved through the combination or averaging of their corresponding feature vectors within the network architecture.

This merging process is optimized during the training phase, where the extracted feature vectors are adjusted according to network parameters. Face images are analyzed using the LBP algorithm, which is invariant to image rotation and lighting changes, while the speech modality employs SincNet, a single-dimensional convolutional layer that extracts frequency characteristics from segmented speech signals.

SincNet offers advantages such as direct audio processing, fewer trainable parameters, and faster convergence. To train and assess the proposed model, we constructed a virtual data set merging BIOMEX-DB and VidTimit datasets, expanding the available biometric data by injecting noise into voice signals and applying transformations to images. This integration and its evaluation under various conditions represent a significant advance over previous studies that often focus on single-modality or simpler fusion techniques.

The voice and facial modalities of the bimodal system were chosen for their ease of acquisition and minimal intrusion into the participants. These modalities allow for simple data manipulation, enabling the implementation of artificial data augmentation schemes to prevent model overfitting. Additionally, it is essential to note that these traits are not mutually dependent, allowing the combination of databases with similar characteristics to increase the amount of available information, as demonstrated in our work, to fill in missing data.

In contrast, the utilization of a deep learning (DL) model for this proposal is justified by its ability to overcome limitations found in machine learning (ML) models. One of such limitations is the requirement for feature extraction, where ML methods require mathematical procedures to extract data characteristics, which may not be compatible with certain types of biometric data or databases, or if the information is altered by noise or transformation [6].

Deep learning models can automatically extract characteristics from input data and adjust their parameters during the training process to improve recognition performance. Furthermore, the multimodal fusion of feature vectors within the network architecture also benefits from this optimization.

This research is presented as follows. Section 2 outlines a comprehensive survey of significant multimodal biometric projects that focus on the integration of feature vectors through neural network models or machine learning. Section 3 provides the essential aspects of BiCNN training. Sections 4 and 5 outline the criteria for assessing network performance and provide the corresponding outcomes. Lastly, Section 6 provides the conclusions and outlook.

2. Related Work

The biometric system requires relevant characteristics from the modalities of voice and face, which continue to be the focus of research through unimodal methods. Machine learning techniques and computational intelligence with deep learning-based approaches are utilized in modern approaches for voice-based biometrics [7], [8]. Additionally, facial trait-based biometrics employs various techniques, such as combining thermal and visual images [9], employing one-shot learning with data augmentation [10], or mapping images to a Euclidean space using convolutional networks [11].

The literature on multimodal biometrics demonstrates different approaches to combining data from multiple biometric features. The selection of these approaches is influenced by the traits being considered and the classification techniques employed. Table 1 provides an overview of significant studies in this field.

3. Development of the Bimodal Biometric System

The bimodal biometric system proposed in this paper comprises a convolutional neural network with two inputs, each corresponding to a distinct biometric feature. Each input is connected to a subnet that independently processes information related to each trait and delivers an attribute vector.

The two vectors are combined (concatenated) to generate a bimodal vector that is refined using a densely connected layer to extract additional significant information and reduce its dimensionality.

The network was trained using the Keras library with an Adam optimizer, a learning rate 0.001, and a step size of 50 epochs. Training involved eight batches of 32 pairs each for training and validation. The convolutional layers for face processing used filters of increasing size (32, 64) and kernel dimensions (3×3, 5×5) with Leaky ReLU activations and batch normalization. Voice processing utilized the SincNet architecture with 120 filters and a kernel size of 251, followed by additional 1D convolutional layers.

The result is a densely connected layer that functions as a SoftMax classifier. The architecture of the suggested BiCNN is listed in the Tab. 2, while Fig. 2 shows the configuration of proposed bimodal network.

3.1. Feature Extraction

An objective of this research is to integrate facial and vocal records at the attribute level to realize the multimodal aspect of the proposed biometric architecture. Recent research suggests

Tab. 1. Review of works on multimodal biometrics.

Ref.	Modalities	Method	Database	Results
[6]	Iris, face, and veins of the fingers	Concatenation of feature vectors extracted by a VGG16 network and fusion of scores	SDUMLA-HMT: 106, IT Delhi and FERET	Identification accuracy: concatenation 99.39%, fusion of scores 100%
[12]	Face and iris	Concatenation of vectors extracted by neural network based on VGG-19 architecture	CASIA-webface: 10576, ND-Iris-0405: 1356, WVU-multi-modal: 2264	Verification: genuine acceptance rate (GAR) 99.67%
[13]	Iris and area peripheral of the eye	Concatenation of extracted weighted vectors neural network with maxout units	CASIA-IrisV4, CASIA-CSIR 2015	Verification: EER 1.88%
[14]	Face, fingerprint, and veins of the fingers	Concatenation of computed Fisher vectors from feature vectors	Self-acquired: 51	Identification accuracy: 50 participants 88.01%, 20 part. 90.01%, 15 part. 93.01%
[15]	Voice and face	Fusion of scores by mixed Gaussian mixture models (GMM), alternative universal models (UBM) and neural network	Self-acquired	Identification accuracy: 95%
[16]	Face and voice	Extraction of characteristics by means of various methods. The fusion of information led toc out at the level of characteristics and at the level of scores	Self-acquired	Verification: EER 0.62%
[17]	Voice and face	The feature vectors of the two modes were used to train a K-classifier nearest neighbors (KNN)	CSUF-SG5: 28	Verification: EER 8.05%
[18]	Face and voice	Fusion of scores generated by the comparison of LBP features and a GMM model by a weighted sum	XJTU: 103	Verification:true positive rate (TPR) 100%, type I error rate 0%, type II error rate 0%
[19]	Fingerprint and face, electro-cardiogram (ECG)	Concatenation of extracted feature vectors by a multi-tasking neural network. They were also made tests with fusion of score by different methods	Virtual database: 58, created from the ECG-ID databases, PTBECG, Faces95, and FVC2006	Identification accuracy: fusion of characteristics 98.97%, rule of the sum 98.95%, product rule 96.55%
[20]	Digital signature and fingerprint	Concatenation of extracted feature vectors by a convolutional network. Concatenation occurs on two points of architecture	Self-acquired: 280	Identification accuracy: early modality fusion (EMF) 99.11%, late modality fusion (LMF) 98.36%
[21]	Voice and face	The fusion was carried out by normalizing and adding of scores of each modality generated by comparison of vectors	MOBIO	Verification: area under the ROC curve 0.98
[22]	Iris and digital fingerprint	Fusion by canonical correlation and principal component analysis (PCA)	SDUMLA-HMT	Identification accuracy 100%, verification EER 0.176%
[23]	Fingerprint, fingervein, palmprint	Multimodal biometric model, U-Net with attention, feature fusion	Union DB1: 400, DB2: 500, DB3: 5500	Identification: EER 0.098%, EER 0.024%, EER 0.117%
[24]	Ear and palm vein	Adaptive 2D Gabor filter, curvature detection, morphological operations, sensor- and feature-level fusion	PUT vein database, custom ear database	Accuracy 97.65%, EER 2.15%
[25]	Online signatures, fingerprints	Empirical modal decomposition (EMD) for signatures, minutiae extraction for fingerprints, score-level fusion	MYCT-100, SVC2004, FVC2004	Best EER 1.69% with min-max normalization

incorporating a module within the network structure that effectively integrates the obtained characteristics from all biometric features [13]. This fusion block generates a singular descriptor that encapsulates an individual's most critical identity information.

LBP face images were processed using a series of two-dimensional convolutional layers. In each layer, the number of filters increased, as did the dimensions of the convolutional mask, in order to capture more detailed information. A "max pooling" process was utilized for feature maps to decrease

Tab. 2. Architecture of BICNN.

Layers	Filters/neurons	Dimensions	Step	Activation function
Face processing layers				
2D convolution	32	3×3	1×1	LReLU
Batch normalization	–	–	–	–
MaxPool 2D	–	2×2	1×1	–
2D convolution	64	5×5	1×1	LReLU
Batch normalization	–	–	–	–
MaxPool 2D	–	2×2	1×1	–
Densely connected	512	–	–	LReLU
Batch normalization	–	–	–	–
Voice processing layers				
SincNet	120	251	–	LReLU
Batch normalization	–	–	–	–
MaxPool 1D	–	5	1×1	–
1D convolution	32	5	1×1	LReLU
Batch normalization	–	–	–	–
MaxPool 1D	–	5	1×1	–
1D convolution	64	5	1×1	LReLU
Batch normalization	–	–	–	–
MaxPool 1D	–	5	1×1	–
Densely connected	512	–	–	LReLU
Batch normalization	–	–	–	–
Fusion and exit				
Averaging/concatenation	–	–	–	–
Densely connected	512	–	–	LReLU
Dropout	0.5	–	–	–
Batch normalization	–	–	–	–
Densely connected	45	–	–	Softmax

their size and eliminate irrelevant information. Ultimately, the final convolutional layer generated maps that were processed by a fully connected layer to produce a fixed-dimensional characteristic vector of 512 points, which contained the discriminatory information of the face image.

Voice signals were processed with the SincNet convolutional layer [26]. This layer is characterized by a bank of bandpass filters that extract frequency characteristics from voice signals. The bandpass filters are defined with sinc functions in the time domain, as illustrated in Eq. (1).

$$g[n, f_1, f_2] = 2\, f_1\, \mathrm{sinc}(2\pi f_2 n) - 2\, f_1\, \mathrm{sinc}(2\pi f_1 n)\,, \qquad (1)$$

where f_1 and f_2 are the high and low cut-off frequencies of the bandpass filters, the values of these frequencies are optimized during the training phase. For our experiments, the cut-off frequencies of the filter bank were initialized in logarithmic form, to obtain the MFCCs [27].

The SincNet layer extracts features, which are subsequently processed using one-dimensional convolutional blocks, culminating in a max pooling operation to eliminate irrelevant information. Lastly, a fully connected layer comprising 512 neurons generates a voice information vector for everyone.

3.2. Bimodal Data Fusion

The combination of face and voice vectors is implemented by concatenation or averaging. A feature combination block was also realized through a weighted average regulated by an optimization variable ρ. This variable is adjusted during the training phase, and its measure indicates which mode contributes more significantly to the individuality of the person. Equation (2) illustrates the weighted average utilized in the bimodal embedding.

$$V_{SF} = \rho\, V_F + (1 - \rho)\, V_S\,, \qquad (2)$$

where V_S and V_F represent the face and speech vectors, correspondingly, and V_{SF} is the resultant multimodal embedding. The initial value of ρ was set to 0.5. Following the fusion step, a densely connected layer is used to reduce the dimensionality of the resulting fusion vector to a fixed size, comprising 512 neurons and featuring a dropout rate of 0.5.

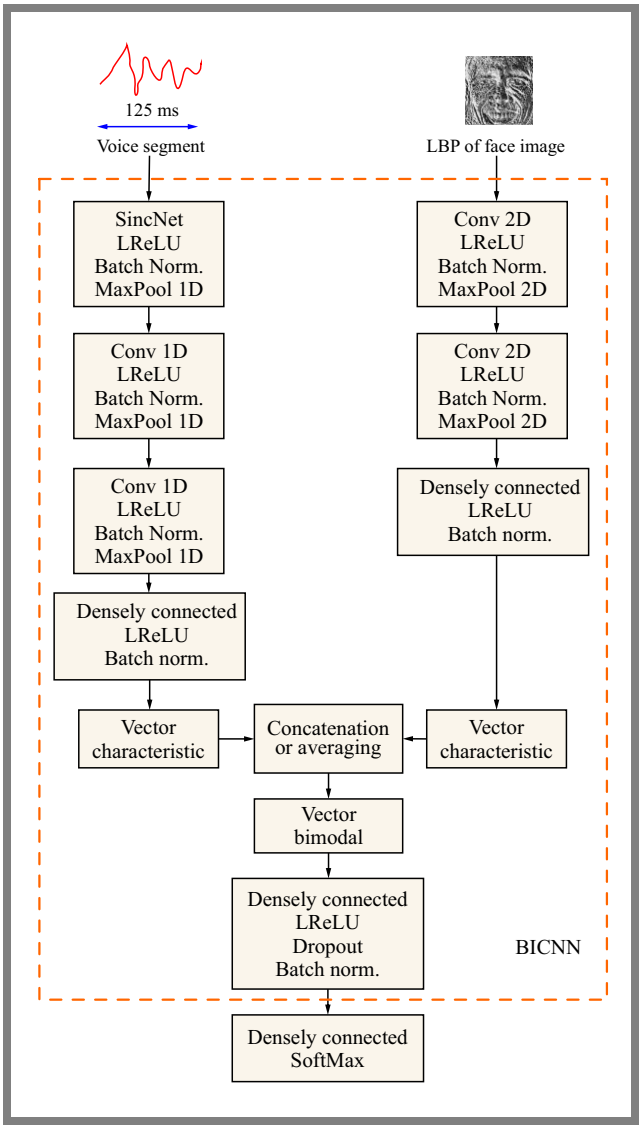

Fig. 2. Schematic diagram of the bimodal convolutional network (BNet).

3.3. Outputs

The final result of BiCNN is a fully connected layer that functions as a SoftMax classifier, incorporating an equal number of neurons as the legitimate user count registered in the biometric system. It is crucial to note that the 512-dimensional vector provided by the penultimate layer contains the features of both voice and face characteristics and will hereafter be referred to as the bimodal vector.

All convolutional layers, apart from the output layer, have batch normalization as a regulator and employ the LReLU activation function according to the findings described in [28]. This research examines the convergence time of a neural network and its recognition performance when utilizing the MNIST database. It evaluates the activation functions of differences between the ReLu and LReLU. The results show that LReLU enables the network to converge more quickly while maintaining performance comparable to that of the ReLu function.

3.4. Pre-processing Data

Several bimodal voice and face biometrics works implemented with different fusion methods are listed in [4]. A part of these works considered relatively small populations for their experiments. This situation occurs because few publicly accessible multimodal databases contain a large population in which all individuals have complete biometric information. Many of the cited authors resorted to generating databases that fit the requirements of their studies. The process of acquiring biometric data requires many resources, so it is sometimes only possible to have large populations. Another alternative is to combine two databases by matching the individuals of each.

3.5. Datasets

To develop our bimodal network, we used voice and face data from the BIOMEX-DB database [29]. However, as some subjects lacked facial data, we opted to supplement this information with a set of images from the VidTimit bimodal database [30].

The BIOMEX-DB database contains information on EEG, voice, and face modalities. It consists of a total of 51 subjects, including 25 women and 26 men. The voice data comprises recordings of English pronunciations of strings of digits, with each subject having 20 audio files, 10 of which correspond to 10-digit string pronunciations, and the remaining ten correspond to 5-digit strings. Face data consists of face videos recorded while participants uttered their strings of digits. The database has a population of 43 people and includes voice and face information. The voice data comprises pronunciations of 10 short English sentences, while the face data consists of images extracted from videos recorded while pronouncing these sentences.

It is crucial to note that 12 BIOMEX-DB subjects do not have face data, which requires the completion of this missing information with face images of 12 VidTimit subjects, as previously mentioned. Consequently, a population of 51 subjects was formed for the study, which was randomly divided into two sets: 45 legitimate users and six impostors.

Although our current study is based on the BIOMEX-DB and VidTimit datasets comprising 51 subjects, future work will involve larger and more diverse datasets such as VoxCeleb and MOBIO to further validate the findings and enhance the generalizability of the results.

3.6. Data Processing and Augmentation

The standardization of voice data was performed to confirm that the signal values were within the range of $-1, \ldots, 1$. Speech signals were processed to remove silences or pauses between pronunciations. To improve speech signals, artificial enhancement was carried out incorporating background noise sampled from the MUSAN database [31] and adjusting the noise level by 0 and 5 dB according to the signal-to-noise ratio.

The face images of all individuals in BIOMEX-DB were extracted from their corresponding videos using the OpenCV

Tab. 3. Experimental setup and training parameters.

Parameters	Value
Training epochs	60
Batch size	32
Learning rate	0.001
Optimizer	Adam
Loss function	CCE

[31] library's face detector to crop the area of interest into a final 128×128 grayscale pixel image. The LBP features with eight neighborhood pixels and a radius of 1 were then extracted from these images. These features are robust to changes in facial posture and variations in image illumination and have been successfully used in conjunction with convolutional networks for face image classification tasks [33]–[35].

The same procedure for detection, face cropping, and extraction of LBP features was employed for the images obtained from VidTimit. Two transformations were carried out using the Imgaug library to increase the number of images [36]. First, the lighting of each image was modified, and second, they were randomly rotated at angles between $-45°, \ldots, 45°$. It is important to note that these transformations were applied to the grayscale images and the LBP operator was applied to the images resulting from the transformations.

3.7. Hyperparameters

The network was implemented using Keras with the Adam optimizer, with a learning rate of 0.001 and was trained over 50 epochs. A batch size of 32 was chosen to optimize learning efficiency while managing computational resources. Table 3 outlines the experimental setup and training parameters.

3.8. Architectural Choices

SincNet was chosen for voice processing because of its unique capability to directly process raw audio signals without extensive processing. This architecture utilizes sinc functions as learnable filters, allowing it to effectively capture important frequency information while reducing the number of trainable parameters. By focusing on the temporal and spectral characteristics of the waveform, SincNet enhances the robustness of feature extraction, leading to improved performance under challenging conditions, such as background noise or varying speaker characteristics. This efficiency is particularly advantageous in scenarios where computational resources are limited or where real-time processing is essential.

For face recognition, local binary patterns (LBP) were selected due to their effectiveness in encoding texture information while being invariant to changes in lighting and facial expressions. LBP operates by thresholding neighborhood pixels and generating a binary code, which allows the creation of a histogram that summarizes the texture features. This method is computationally efficient and has been demonstrated to provide a strong performance in various face classification tasks. By maintaining the integrity of the feature under various conditions, LBP enables the system to achieve higher accuracy in face recognition, especially when dealing with variations in pose, illumination, and occlusion.

The integration of these two architectural choices: SincNet for voice and LBP for face recognition form a robust bimodal biometric system capable of managing the complexities inherent in multimodal data.

3.9. Bimodal Network Training

The data set was partitioned into training, validation (evaluation) and test sets using a 65/05/30 ratio, which provided a distinct validation set for optimizing the model's parameters and enhancing its performance. During the training phase, we used specific neural network architectures along with their configurations and hyperparameters. Important training parameters, such as the learning rate, batch size, number of epochs, and optimizer, were carefully chosen to maximize the efficiency of the training process.

We developed and trained the bimodal network using the Keras library. Training was carried out over 50 epochs, with a learning rate of 0.001 and the Adam optimizer. Categorical cross entropy (CCE) served as the loss function. The models were trained on the training dataset and evaluated on the validation set to keep track of their performance, allowing for any necessary adjustments. Ultimately, the models were evaluated on a separate test set to determine their effectiveness in unseen data.

For training and validation purposes, voice samples of 10–digit string pronunciations and the corresponding face images from the associated videos were utilized. Each training instance required pairs of an image and a voice segment. To promote the generalization of the proposed model, we randomly generated eight batches of 32 pairs for training and eight batches of 16 pairs for validation. This was designed with the understanding that the SincNet convolutional layer can process segments of raw speech signals that last several milliseconds. The BIOMEX-DB videos have a frame rate of 8 frames per second, and each speech segment lasts 125 ms. Additionally, a monitoring mechanism was implemented to track validation accuracy during each epoch, ensuring that only those parameters that maximized this metric were retained.

4. Experimental Evaluation

One of the key objectives of this work is to introduce a novel approach in the field by comparing the performance of proposed BICNN using two methods of feature fusion: concatenation and averaging.

To carry out the evaluation, the pronunciation data of 5-digit strings and the facial images of their corresponding videos were applied.

To assess the bimodal system, we use the biometric data resulting from the artificial increase in the data together with the unmodified samples. This will permit us to assess the performance of the network under diverse conditions for both

modalities. In some of the related works, the authors mention that they also used a data augmentation scheme to train their models, however, their evaluations do not offer information about whether they tested only with their original data or if they included data from the artificially generated set and how they could affect the reconnaissance performance.

For this reason, we present results that demonstrate the effect of artificially generated voice and face samples on identification and verification in a schematic way, assuming that these samples approximate actual operating conditions.

4.1. Identification Evaluation

In the identification task, a user delivers his biometric data to the system, and the latter will provide the identification number or label of the legitimate user registered in the biometric system whose characteristics are more closely resemble those of the user. It's a one-to-many comparison.

In this work, the above procedure is carried out by feeding the output SoftMax classifier with the bimodal vector. The result is a normalized probability distribution from which we can determine the identification number of the legitimate user whose voice and face characteristics resemble more closely those found in the bimodal vector using the argmax function.

The identification experiments were performed in closed set mode, which means that the system assumes that any user is enrolled in it and that its function is to determine their identity and not to verify or validate it. Therefore, data belonging to impostors were not tested, as they do not contribute to correctly evaluating the network's performance. Considering what was described above, the metric for measuring performance was the accuracy defined as:

$$Accuracy = \frac{P_c}{N}, \tag{3}$$

where P_c is the number of times the system successfully predicted a user's identity and N is the total number of tests performed.

4.2. Verification Evaluation

The verification process involves evaluating the biometric features of a user against a set of characteristics stored in the system that correspond to a legitimate user. The outcome of this comparison is the confirmation or denial of the user's identity.

In experimental efforts, we evaluated biometric features by contrasting bimodal vectors. To fashion a legitimate user's pattern, ten pairs of randomly chosen voice and face samples were supplied to the BiCNN, and the pattern was generated by averaging the resulting ten bimodal vectors.

To carry out identity verification, we compared the pattern of a specific legitimate user with a bimodal vector generated by an individual seeking to validate their identity. In this instance, we employed the cosine similarity function to make the comparison, which returns a value that increases as the pattern and sample vector become more similar. Furthermore, we utilized a threshold value to assess cosine similarity and decide to accept or reject the validation of the said identity. The cosine similarity function is:

$$Similarity = \cos(\theta) = \frac{P.M}{\|P\| * \|M\|}, \tag{4}$$

where P is the pattern of a legitimate user and M is the sample vector of a user who wants to verify or validate their identity.

The equal error rate (EER) metric is typically utilized to assess the effectiveness of a biometric system in the verification task. This metric refers to the point at which the number of precisely classified negative examples equals the number of correctly positively categorized examples [37].

Tab. 4. Identification results (in percent) in the precision of the bimodal network with (A – averaging, C – concatenation) of voice and face vectors.

SNR [dB]	No transformations		Transformations			
			Lighting		Rotation	
	A	C	A	C	A	C
Noise free	98.5	99.4	97.6	99.1	84.6	91.5
0	99.0	99.1	98.0	99.1	83.6	89.7
5	99.1	99.2	97.8	99.1	85.1	90.6

5. Results and Commentary

5.1. Identification Outcomes

The results of the identification task are shown in the tab. 4, in each of the evaluation conditions, 1000 tests were performed to calculate the percentage of accuracy.

The results show that accuracy was greater than 97% for most conditions, while conditions involving images with rotation delivered lower values between 83% and 92%.

Concatenation fusion produces better results than averaging. This gap is more significant under conditions that include image rotation. The latter factor indicates that different conditions of a face image affect the accuracy value more than the noise of voice signals. To provide a comprehensive comparison, we evaluated the proposed BDLN against other state-of-the-art multimodal and unimodal techniques such as VGG-19, Fisher vectors, and traditional SVM classifiers across datasets such as CASIA and MOBIO. The results consistently show superior performance in terms of accuracy and EER. Additionally, real-world testing scenarios were simulated by introducing varying levels of noise and transformations.

5.2. Verification Outcomes

The verification task was evaluated by performing 1000 tests with samples from legitimate users and another 1000 with samples from the set of impostors for each condition contemplated in this research. Table 5 illustrates the results in terms of EER delivered by the bimodal biometric system.

Tab. 5. Verification outcomes (in percent) in terms of equal error rate EER for the bimodal distribution network upon vector (A – averaging, C – concatenation).

SNR [dB]	No transformations		Transformations			
			Lighting		Rotation	
	A	C	A	C	A	C
Noise free	0.19	0.60	0.92	0.79	5.65	3.70
0	0.14	0.74	0.91	0.95	4.96	4.21
5	0.13	0.56	1.0	0.70	5.80	3.76

In most cases, the EER values are less than 1%, which is indicative of good verification performance. Similarly, the identification task conditions involving image rotation obtained significantly higher values than the rest. When comparing the EER values of both methods of characteristic fusion, the averages delivered more competent results in this case. Under the conditions in which the images were not transformed, averaged fusion obtained better values. Although the current study focuses on the concatenation and averaging of feature vectors, future work will explore more advanced fusion methods such as principal component analysis (PCA), linear discriminant analysis (LDA), and deep feature fusion (DFF). Preliminary experiments with PCA have already shown potential improvements in verification accuracy.

When analyzing the differences between EER values, it is again seen that the gap is more noticeable when comparing image transformations than between the different SNR values. Therefore, it is possible that also in the verification task, the conditions of the face images have a more significant influence on the results than the noise of the voice samples.

5.3. Comparison of Results

In the last evaluation stage, some unimodal voice and face systems were trained in order to demonstrate better performance delivered by multimodal systems compared to the first ones. The approaches chosen for this comparison are widely used in literature, their structure makes them compatible to be trained with different databases, and facilitates the reproduction of the results of this work. Multimodal approaches may not be implementable with different databases, since feature extraction methods might not be compatible with specific biometric data [20]. In the case of implementations with DL, another factor to consider is that an architecture designed to be trained with a database with specific characteristics may not converge with other databases.

Two unimodal systems were chosen for each biometric feature. The training and evaluation conditions were the same as for BiCNN. The biometric data of BIOMEX-DB and VidTimit were taken, and the same number of legitimate users and impostors were considered.

The face recognition systems considered are a ResNet network with four residual layers [38]. This architecture was chosen since adding more residual layers did not achieve significantly better results. The second face recognition system is based on eigenfaces, we use the definition found in the OpenCV library with the parameters described in [39]. However, the voice recognition systems chosen were the original implementation of the SincNet network [26] and a X-vector-based system as initially described in [40]. In references [41], [42] you can consult the face recognition systems used in this study, and in [43], [44] you can find information on voice recognition systems.

Tab. 6. Comparison of the accuracy results of the identification task.

Biometric system	No transformations	Transformations	
		Lighting	Rotation
BNet (averaged)	99.1%	98%	85.1%
BNet (concatenation)	99.4%	99.1%	91.5%
ResNet CNN	100%	99.6%	93.6%
Eigenfaces (PCA algorithm)	100%	98.76%	85.43%
SNR	Noise free	0 dB	5 dB
BNet (averaged)	98.5%	99%	99.1%
BNet (concatenation)	99.4%	99.1%	99.2%
SincNet (SNC)	100%	88.1%	96.43%
X-vectors (XVEC)	95.21%	83.1%	92.1%

5.4. Identification

Table 6 provides a comparison of the best results of our identification experiments with those obtained with unimodal systems.

In the face identification modality, all modes had an accuracy of more than 97% for untransformed images and with changing illumination. In the rotation condition, the accuracy values significantly decreased. Only the bimodal network with feature concatenation and the ResNet CNN model delivered results of just over 90%. In this analysis, our bimodal model and the ResNet CNN model have similar results.

In terms of speaker identification, it is worth mentioning that the two techniques for merging the bimodal network exhibited outstanding performance, achieving accuracy levels greater than 95% when noiseless voice signals were used. However, when noise was present in the voice signals, the two methods of merging the bimodal network demonstrated significantly higher accuracy values compared to unimodal systems.

This suggests that the proposal for a bimodal system can maintain a high level of identification performance even in the presence of noise. In contrast, other systems experience a noticeable decrease in performance under similar circumstances.

Tab. 7. Comparison of the EER results of the verification task.

Biometric system	No transformations	Transformations	
		Lighting	Rotation
BNet (averaged)	0.13%	0.93%	4.98%
BNet (concatenation)	0.56%	0.72%	3.72%
ResNet CNN	0.55%	1.44%	3.59%
Eigenfaces (PCA algorithm)	2.97%	12.83%	27.78%
SNR	Noise free	0 dB	5 dB
BNet (averaged)	0.21%	0.16%	0.13%
BNet (concatenation)	0.62%	0.76%	0.56%
SincNet (SNC)	1.82%	13.4%	5.56%
X-vectors (XVEC)	1.73%	4.93%	2.38%

5.5. Verification

Table 7 summarizes an evaluation of the effectiveness of the verification task in terms of the equal error rate (EER).

In face verification, the bimodal network with its two fusion methods and the ResNet CNN model delivered results with little difference between their EER values for all image transformations. In the rotation condition, all models showed an increase in EER. This is a situation similar to that seen in identification. Eigenfaces, being a non-neural network-based method, showed the worst performance in all conditions by a significant difference.

In the domain of voice verification, the bimodal network that utilized averaging emerged as the top performer in all three evaluation scenarios, while the concatenation fusion method demonstrated slightly inferior results. Both bimodal networks consistently outperformed the SincNet and X-vector approaches by a significant margin when dealing with speech samples that had noise added to them. These findings suggest that the proposed network demonstrates superior performance in verification tasks when working with speech samples that have noise added to them at these specific SNR values.

Finally, we accumulated all the scores used to evaluate the conditions of each biometric system to generate the corresponding receiver operating characteristic (ROC) curves. Figure 3 shows the ROC curves of all the models evaluated.

The general EER values obtained using the ROC curves are as follows: bimodal network (BNet) averaged 3.28%, bimodal network (BNet) concatenation 2.16%, ResNet CNN 2.37%, eigenfaces 17.84%, SincNet (SNC) 9.32%, and X-vectors (XVEC) 3.52%. These values show that the bimodal proposal in its two variants delivered expected results and, in most cases, superior to the considered unimodal systems. The fusion method by concatenation of vectors showed the best verification performance and this result is consistent with what was obtained in the previous evaluations.

5.6. Comparison with Related Works

Table 8 shows a brief comparison of our results with other relevant work on bimodal voice and face biometrics. We select papers whose number of subjects in their experiments is similar or not much higher than ours. As can be seen, the results are comparable to those presented by other authors who implemented other techniques and used other databases. Furthermore, Fig. 3 presents the ROC curves for our experimental results, offering a visual comparison of performance across different methods and conditions.

Figure 4 compares the percentages of equal error rate (EER) of several studies focusing on the efficiency of biometric or authentication systems. The EER is a critical performance metric that marks the intersection point between the false acceptance rate (FAR) and the false rejection rate (FRR). A lower EER percentage indicates a more accurate and reliable system.

The methodology proposed in the current analysis provides an EER of only 0.13%, compared to 0.56% for a concatenation approach. This represents a significant improvement over other systems, especially in contrast to much higher EERs reported in previous literature, such as 20.59% for [45], 25.70% for [51], and 19.90% for [53]. Even state-of-the-

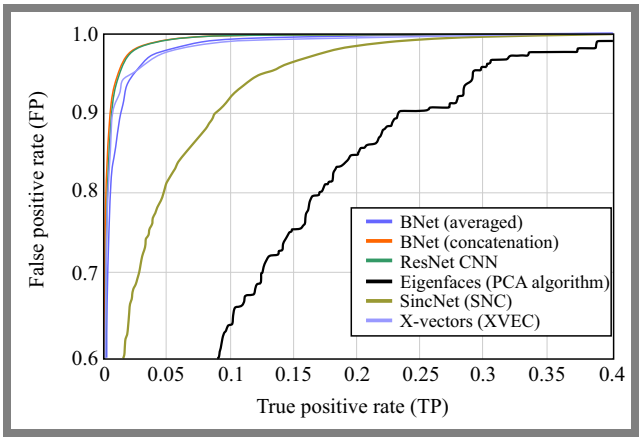

Fig. 3. Comparison of ROC curves.

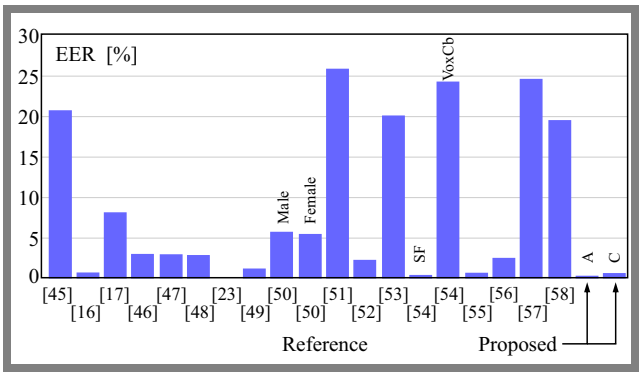

Fig. 4. Comparison of the equal error rate (EER) between different studies.

Tab. 8. The EER score of the proposed model along with the state-of-the-art methods on bimodal voice and face biometrics.

Paper	Classifier	Database	EER
[16]	Various classifiers	Acquired by the authors: 100	0.65%
[17]	KNN	CSUF-SG5: 27	8.04%
[18]	Comparison of LBP and GMM characteristics	XJTU: 102	0%
[45]	Support vector machine (SVM)	CSUF-SG5: 27	20.59%
[46]	FaceNet + dilated residual network (DRN)	SWAN: 88	3.01%
[47]	Logistic regression	NIST SRE19: 47	2.78%
[48]	ResNet	NIST SRE19: 42	2.78%
[49]	CNN with two inputs	Deep lip (virtual database): 150	1.11%
[50]	ResNet-50	VoxCeleb1-Test: 40, MOBIO: 150, AveRobot: 111	5.7% male, 5.4% female, 7.7% male, 9.1% female 30% male, 31.6% female
[51]	Single-branch network (Git)	VoxCeleb1 (VC1): 40	25.7%
[52]	Fusion (JCA) with U-BLSTM and J-BLSTMs	VoxCeleb1	2.214%
[53]	Multi face-voice association learning with keynote speaker diarization (MFV-KSD)	FAME 2024	19.9%
[54]	Universal transformer, vision transformer (ViT), multimodal prototypical network loss	SpeakingFaces (SF), VoxCeleb (VoxCb)	0.27%, 24.1%
[55]	Gammatonegram for voice and VGGFace2 for face	VoxCeleb2	0.62%
[56]	Real additive angular marlin loss, attention-based fusion	SpeakingFaces	2.5%
[57]	Voice-face aligner (VFAligner)	VoxCeleb2	24.40%
[58]	Fuse after align (FAA) framework, multimodal encoder	VoxCeleb	19.3%
Proposed	BNet (two stream-CNN)	BIOMEX-DB, VidTimit: 51	BNet (averaged): 3.28% general, 0.13% best, BNet (concatenation): 2.16% general, 0.56% best

art approaches, such as [16] (0.65%) and [49] (1.11%), fail to match the performance of the proposed method.

Furthermore, [18] achieved an EER of 0.00%, which is notable for its precision but still does not diminish the significance of the low EER values in the proposed solution, especially considering the broader scope and versatility of the system presented in this paper.

6. Conclusion and Future Work

The evaluation results show that the bimodal network with fusion by concatenation obtained better recognition results in most of the conditions considered in both identification and verification.

In the second evaluation stage, where popular unimodal proposals were trained with the same virtual database and were tested under the same conditions as BiCNN, in case of the face modality, the results are very similar to the unimodal network based in ResNet. Despite that, our proposal demonstrated superior performance to the eigenfaces system and the two unimodal voice approaches in most conditions considered.

Our proposal delivers decent EER values and is comparable with the other works in which the amount of population considered experiments is like ours. Although our research demonstrates the technical viability of bimodal biometric systems, it is imperative to address ethical concerns such as privacy, data security, and user consent. Biometric data are inherently sensitive and proper measures must be taken to ensure their protection.

Additionally, to further validate the system's robustness and scalability, we plan to incorporate more diverse datasets that closely mimic real-world conditions. Testing on larger databases such as VoxCeleb and MOBIO will allow us to evaluate the system's performance across varied environments. This will help improve its generalizability and reliability, making it more suitable for practical applications in diverse real-world scenarios.

References

[1] S.A. Abdulrahman and B. Alhayani, "A Comprehensive Survey on the Biometric Systems Based on Physiological and Behavioral Characteristics", *Materials Today: Proceedings*, vol. 80, pp. 2642–2646, 2023 (https://doi.org/10.1016/j.matpr.2021.07.005).

[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 (https://doi.org/10.1016/j.inffus.2018.11.018).

[3] D. Patel, S. Patel, A.A. Thadeshwar, and R. Chaturvedi, "Multimodal Biometric Systems: A Review", *International Journal of Advanced Research in Computer Science*, vol. 9, no. 2, pp. 361–365, 2018 (https://doi.org/10.26483/ijarcs.v9i2.5742).

[4] H. Mandalapu *et al.*, "Audio-visual Biometric Recognition and Presentation Attack Detection: A Comprehensive Survey", *IEEE Access*, vol. 9, pp. 37431–37455, 2021 (https://doi.org/10.1109/access.2021.3063031).

[5] M. Singh, R. Singh, and A. Ross, "A Comprehensive Overview of Biometric Fusion", *Information Fusion*, vol. 52, pp. 187–205, 2019 (https://doi.org/10.1016/j.inffus.2018.12.003).

[6] N. Alay and H.H. Al-Baity, "Deep Learning Approach for Multimodal Biometric Recognition System Based on Fusion of Iris, Face, and Finger Twenty Traits", *Sensors*, vol. 20, no. 19, art. no. 5523, 2020 (https://doi.org/10.3390/s20195523).

[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 (https://doi.org/10.1016/j.matpr.2021.05.261).

[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*, vol. 15, no. 2, pp. 789–795, 2023 (https://doi.org/10.1007/s41870-022-01142-1).

[9] N.K. Benamara, E. Zigh, T.B. Stambouli, and M. Keche, "Towards a Robust Thermal-visible Heterogeneous Face Recognition Approach Based on a Generative Cycle Adversarial Network", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 132–145, 2022 (https://doi.org/10.9781/ijimai.2021.12.003).

[10] D.M. Jiménez-Bravo *et al.*, "Edge Face Recognition System Based on One-shot Augmented Learning", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 6, pp. 31–44, 2022 (https://doi.org/10.9781/ijimai.2022.09.001).

[11] A. Alcaide *et al.*, "LIPSNN: A Light Intrusion-proving Siamese Neural Network Model for Facial Verification", *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 4, pp. 121–131, 2022 (https://doi.org/10.9781/ijimai.2021.11.003).

[12] V. Talreja, M.C. Valenti, and N.M. Nasrabadi, "Multibiometric Secure System Based on Deep Learning", *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Montreal, Canada, 2017 (https://doi.org/10.1109/globalsip.2017.8308652).

[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 (https://doi.org/10.1109/tifs.2018.2833033).

[14] Y. Xin *et al.*, "Multimodal Feature-level Fusion for Biometrics Identification System on IoMT Platform", *IEEE Access*, vol. 6, pp. 21418–21426, 2018 (https://doi.org/10.1109/access.2018.2815540).

[15] V.V. Khryashchev, A.I. Topnikov, A.F. Stefanidi, and A.L. Priorov, "Bimodal Person Identification Using Voice Data and Face Images", *Eleventh International Conference on Machine Vision (ICMV 2018)*, Munich, Germany, 2019 (https://doi.org/10.1117/12.2523138).

[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 (https://doi.org/10.1007/s11042-018-7012-3).

[17] O. Olazabal *et al.*, "Multimodal Biometrics for Enhanced IoT Security", *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, USA, 2019 (https://doi.org/10.1109/ccwc.2019.8666599).

[18] X. Zhang *et al.*, "An Efficient Android-based Multimodal Biometric Authentication System with Face and Voice", *IEEE Access*, vol. 8, pp. 102757–102772, 2020 (https://doi.org/10.1109/access.2020.2999115).

[19] E. Al Alkeem *et al.*, "Robust Deep Identification Using ECG and Multimodal Biometrics for Industrial Internet of Things", *Ad Hoc Networks*, vol. 121, art. no. 102581, 2021 (https://doi.org/10.1016/j.adhoc.2021.102581).

[20] M. Leghari *et al.*, "Deep Feature Fusion of Fingerprint and Online Signature for Multimodal Biometrics", *Computers*, vol. 10, no. 2, art. no. 21, 2021 (https://doi.org/10.3390/computers10020021).

[21] C.F.F. Costa-Filho, J.V. Negreiro, and M.G.F. Costa, "Multimodal Biometric System Based on Autoencoders and Learning Vector Quantization", *Brazilian Congress on Biomedical Engineering*, Vitoria, Brazil, 2020 (https://doi.org/10.1007/978-3-030-70601-2_236).

[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*, vol. 20, no. 2, pp. 205–218, 2023 (https://doi.org/10.1080/1448837x.2022.2129147).

[23] Z. Zhang, H. Lu, P. Sang, and J. Wang, "MultiBioGM: A Hand Multimodal Biometric Model Combining Texture Prior Knowledge to Enhance Generalization Ability", in: *Biometric Recognition (CCBR 2023)*, pp. 106–115, 2023 (https://doi.org/10.1007/978-981-99-8565-4_11).

[24] V. Gurunathan and R. Sudhakar, "Multimodal Biometric System Using Palm Vein and Ear Images", *Proceeding of International Conference on Computer Visions and Robotics*, pp. 439–451, 2023 (https://doi.org/10.1007/978-981-99-4577-1_36).

[25] T. Hafs, H. Zehir, A. Hafs, and A. Nait-Ali, "Multimodal Biometric System Based on the Fusion in Score of Fingerprint and Online Handwritten Signature", *Applied Computer Systems*, vol. 28, no. 1, pp. 37–49, 2023 (https://doi.org/10.2478/acss-2023-0006).

[26] M. Ravanelli and Y. Bengio, "Speaker Recognition from Raw Waveform with SincNet", *2018 IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, 2018 (https://doi.org/10.1109/slt.2018.8639585).

[27] Y. Badr, P. Mukherjee, and S.M. Thumati, "Speech Emotion Recognition using MFCC and Hybrid Neural Networks", *Proceedings of the 13th International Joint Conference on Computational Intelligence*, pp. 366–373, 2021 (https://doi.org/10.5220/0010707400003063).

[28] 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, 2019 (https://doi.org/10.1007/978-981-13-6772-4_76).

[29] D.B. Jadhav, G.S. Chavan, V.C. Bagal, and R.R. Manza, "Review on Multimodal Biometric Recognition System Using Machine Learning", *Artificial Intelligence and Applications*, vol. 20, pp. 1–7, 2023 (https://doi.org/10.47852/bonview3202593).

[30] C. Sanderson and B.C. Lovell, "Multi-region Probabilistic Histograms for Robust and Scalable Identity Inference", in: *Advances in Biometrics (Conference Proceedings)*, pp. 199–208, 2009 (https://doi.org/10.1007/978-3-642-01793-3_21).

[31] D. Snyder, D. Povey, and G. Chen, "MUSAN: A Music, Speech, and Noise Corpus", *ArXiv*, 2015 (https://doi.org/10.48550/arxiv.1510.08484).

[32] A. Zelinsky, "Learning OpenCV–Computer Vision with the OpenCV Library", *IEEE Robotics & Automation Magazine*, vol. 16, no. 3, p. 100, 2009 (https://doi.org/10.1109/mra.2009.933612).

[33] M. Wang, Z. Wang, and J. Li, "Deep Convolutional Neural Network Applies to Face Recognition in Small and Medium Databases", *2017 4th International Conference on Systems and Informatics (ICSAI)*, Hangzhou, China, 2017 (https://doi.org/10.1109/icsai.2017.8248499).

[34] P. Ke, M. Cai, H. Wang, and J. Chen, "A Novel Face Recognition Algorithm Based on the Combination of LBP and CNN", *2018 14th IEEE International Conference on Signal Processing (ICSP)*, Beijing, China, 2018 (https://doi.org/10.1109/icsp.2018.8652477).

[35] Q. Xu and N. Zhao, "A Facial Expression Recognition Algorithm Based on CNN and LBP Feature", *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, Chongqing, China, 2020 (https://doi.org/10.1109/itnec48623.2020.9084763).

[36] A.B. Jung *et al.*, "Imgaug", GitHub: San Francisco, USA, 2020 (https://github.com/aleju/imgaug).

[37] J.-M. Cheng and H.-C. Wang, "A Method of Estimating the Equal Error Rate for Automatic Speaker Verification", *2004 International Symposium on Chinese Spoken Language Processing*, Hong Kong, China, 2004 (https://doi.org/10.1109/CHINSL.2004.1409642).

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016 (https://doi.org/10.1109/cvpr.2016.90).

[39] 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, pp. 30–40, 2022 (https://doi.org/10.5815/ijieeb.2022.03.04).

[40] D. Snyder *et al.*, "X-vectors: Robust DNN Embeddings for Speaker Recognition", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018 (https://doi.org/10.1109/icassp.2018.8461375).

[41] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face Recognition Systems: A Survey", *Sensors*, vol. 20, no. 2, art. no. 342, 2020 (https://doi.org/10.3390/s20020342).

[42] A. Verma, A. Goyal, N. Kumar, and H. Tekchandani, "Face Recognition: A Review and Analysis", in: *Computational Intelligence in Data Mining (Conference Proceedings)*, pp. 195–210, 2022 (https://doi.org/10.1007/978-981-16-9447-9_15).

[43] Z. Bai and X. L. Zhang, "Speaker Recognition Based on Deep Learning: An Overview", *Neural Networks*, vol. 140, pp. 65–99, 2021 (https://doi.org/10.1016/j.neunet.2021.03.004).

[44] 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 (https://doi.org/10.1109/access.2021.3090109).

[45] M. Gofman *et al.*, "Multimodal Biometrics via Discriminant Correlation Analysis on Mobile Devices", *International Conference on Security and Management (SAM)*, Las Vegas, USA, 2018.

[46] R. Ramachandra *et al.*, "Smartphone Multimodal Biometric Authentication: Database and Evaluation", *ArXiv*, 2019 (https://doi.org/10.48550/arXiv.1912.02487).

[47] 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*, 2020 (https://doi.org/10.48550/arXiv.2008.05889).

[48] S.O. Sadjadi *et al.*, "The 2019 NIST Audio-visual Speaker Recognition Evaluation", *The Speaker and Language Recognition Workshop: Odyssey 2020*, Tokyo, Japan, 2020 (https://doi.org/10.21437/odyssey.2020-37).

[49] M. Liu *et al.*, "Exploring Deep Learning for Joint Audio-visual Lip Biometrics", *ArXiv*, 2021 (https://doi.org/10.48550/arXiv.2104.08510).

[50] 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 (https://doi.org/10.1016/j.procs.2021.12.236).

[51] M.S. Saeed *et al.*, "Single-branch Network for Multimodal Training", *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 2023 (https://doi.org/10.1109/ICASSP49357.2023.10097207).

[52] G.P. Rajasekhar and J. Alam, "Audio-visual Speaker Verification via Joint Cross-attention", *International Conference on Speech and Computer*, Dharwad, India, 2023 (https://doi.org/10.1007/978-3-031-48312-7_2).

[53] R. Tao *et al.*, "Multi-stage Face-voice Association Learning with Keynote Speaker Diarization", *ArXiv*, 2024 (https://doi.org/10.48550/arXiv.2407.17902).

[54] M. Abdrakhmanova *et al.*, "One Model to Rule Them All: A Universal Transformer for Biometric Matching", *IEEE Access*, vol. 12, pp. 96729–96739, 2024 (https://doi.org/10.1109/ACCESS.2024.3426602).

[55] A. Farhadipour, M. Chapariniya, T. Vukovic, and V. Dellwo, "Comparative Analysis of Modality Fusion Approaches for Audio-visual Person Identification and Verification", *ArXiv*, 2024 (https://doi.org/10.48550/arXiv.2409.00562).

[56] C. Wang, H. Zhu, and L. Xu, "Research on the Improvement of the Target Speaker Recognition System Based on Dual-Modal Fusion", *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, Zhuhai, China, 2024 (https://doi.org/10.1109/cvidl62147.2024.10603613).

[57] Y. Jiang *et al.*, "Target Speech Diarization with Multimodal Prompts", *ArXiv*, 2024 (https://doi.org/10.48550/arXiv.2406.07198).

[58] C. Peng, L. He, and D. Su, "Fuse after Align: Improving Face-voice Association Learning via Multimodal Encoder", *ArXiv*, 2024 (https://doi.org/10.48550/arXiv.2404.09509).

Khaled Merit, Ph.D.
Laboratory of TIT, Department of Electrical Engineering
https://orcid.org/0000-0002-7762-1898
E-mail: merit.khaled@univ-bechar.dz
Tahri Mohammed University of Bechar, Bechar, Algeria
https://www.univ-bechar.dz

Mohammed Beladgham, Ph.D., Full Professor
Laboratory of TIT, Department of Electrical Engineering
https://orcid.org/0000-0002-2371-6859
E-mail: beladgham.mohammed@univ-bechar.dz
Tahri Mohammed University of Bechar, Bechar, Algeria
https://www.univ-bechar.dz