IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 23, 2021, accepted July 13, 2021, date of publication July 26, 2021, date of current version August 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3100035

BIOMEX-DB: A Cognitive Audiovisual Dataset for Unimodal and Multimodal Biometric Systems

JUAN CARLOS MORENO-RODRIGUEZ[1], (Graduate Student Member, IEEE),
JUAN CARLOS ATENCO-VAZQUEZ[1],
JUAN MANUEL RAMIREZ-CORTES[1], (Senior Member, IEEE),
RENE ARECHIGA-MARTINEZ[2], PILAR GOMEZ-GIL[3],
AND RIGOBERTO FONSECA-DELGADO[4]
[1]Department of Electronics, National Institute of Astrophysics, Optics and Electronics, San Andrés Cholula 72840, Mexico
[2]Department of Electrical Engineering, New Mexico Tech, Socorro, NM 87801, USA
[3]Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, San Andrés Cholula 72840, Mexico
[4]Electrical Engineering Department, Metropolitan Autonomous University, Iztapalapa 09340, Mexico

Corresponding author: Juan Carlos Moreno-Rodriguez (xalatl@inaoep.mx)

This work was supported in part by the Mexican National Council for Science and Technology (CONACyT) under Grant 777594 and Grant 70847.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT Multimodal biometric schemes arise as an interesting solution to the multidimensional reinforcement problem for biometric security systems. Along with the performance dimension, these systems should also comply with required levels for other conditions such as permanence, collectability, and circumvention, among others. In response to the demand for a multimodal and synchronous dataset, we introduce in this paper an open-access database of synchronously recorded electroencephalogram signals (EEG), voice signals, and video feed from 51 volunteers, 25 female, 26 male, captured for, but not limited to, biometric purposes. A total of 140 samples were collected from each user when pronouncing single digits in Spanish, giving a total of 7140 instances. EEG signals were captured using a 14-channel Emotiv™ Epoc headset. The resulting set becomes a valuable resource when working on unimodal biometric systems, but significantly more for the evaluation of multimodal variants. Furthermore, the usefulness of the collected signals extends to being exploited by projects in brain-computer interfaces and face recognition to name just a few. As an initial report on data separability of the related samples, five user recognition experiments are presented: a face recognition identifier with an accuracy of 99%, a speaker identification system with accuracy of 94.2%, a bimodal face-speech verification case with Equal Error Rate around 2.64, an EEG identification example, and a bimodal user identification exercise based on EEG and voice modalities with a registered accuracy of 97.6%.

INDEX TERMS Biometrics, face recognition, speaker recognition, electroencephalography, brain–computer interfaces, image classification, multiple signal classification, classification algorithms.

I. INTRODUCTION

Biometrics, as "the measuring and statistical analysis of people's physical and behavioral attributes" [1] for individual recognition, has become the reference solution in terms of security [2], especially when compared to other validation methods such as token presentation or password verification. However, several articles, such as [3]–[5] and [6] among others, have manifested the limitations and weaknesses of biometric systems based on a single physical trait or biosignal to perform recognition. This trait or biosignal is known as the system's modality, with each modality producing different behavior and performance. For example, iris-based systems are considered to provide some of the best performance levels, even though they may be affected by pupil dilation and gaze angle [7]. Furthermore, iris biometrics may be vulnerable to spoofing such as the use of textured contact lenses [8].

The most desirable performance of a biometric system is described in terms of its capacity to 1) always accept a legitimate user while rejecting all impostors (verification

The associate editor coordinating the review of this manuscript and approving it for publication was G. R. Sinha.

systems) or 2) correctly identify the presenting users with the registered identities in the database (identification systems). Many metrics have been defined to evaluate how adequate a system is. Among the most widely used metrics are Accuracy, False Acceptance Rate (FAR), False Rejection Rate (FRR), Receiver Operating Characteristic (ROC), and Equal Error Rate (EER). All these metrics describe the system's performance according to efficiency [9]. However, efficiency is not the only characteristic defining a biometric system. Many authors, such as Meng *et al.* [10], agree in defining a wider classification, including the following seven desirable characteristics: Universality, Uniqueness, Permanence, Collectability, Performance, Acceptability, and Circumvention. Hence, even though efficiency as a metric for performance may be considered the most important characteristic in most cases, a high-performance system will have reduced utility in a security application if the modality can be easily forged or if it lacks universality. Unfortunately, sources such as [10], [11], and [12] fail to provide a quantitative method for attributes' evaluation other than performance. To overcome the limitations inherent to single modality systems and in order to take advantage of different modalities' strengths, the use of multimodal biometric systems has been proposed and tested as a reliable alternative [13].

When approaching the design of a multimodal biometric system, a critical decision is the selection of the most suitable modalities. There is not a universal solution for all recognition systems. Since each modality presents different attribute-compliance levels, the adequate combination should be selected considering, among other factors, the reinforcement of one modality's weakness by another modality's strength and always focusing on the specific application for which the system is being designed.

This paper presents a multimodal dataset, intended to be used for multimodal biometric system evaluation. Three modalities were considered due to their particular characteristics: voice, video feed, and electroencephalography (EEG) signals. Similarly as discussed in [14] for audio-visual biometric systems, the selection of the aforementioned modalities aims to take advantage "…of complimentary biometric information present between voice and face cues", and goes a step beyond by cross-relating to EEG biometric information present in the process of generating visually-evoked potentials, imagining speech and uttering-articulation. A total of 51 users volunteered, all Spanish-speaking Latinos, 26 males and 25 females, with ages between 16 and 61 ($\bar{x} = 29.75$, $\sigma = 10.97$); 43 claimed to be right-handed, 5 left-handed and 3 declared being ambidextrous. 45 volunteers are Mexican, 2 Ecuadorians, and 1 each from Colombia, Costa Rica, Venezuela, and Cuba.

In terms of utility, our dataset can be used for evaluation of unimodal biometric systems (Text-dependent and Text-independent for voice, Visually-evoked potentials and uttered speech for EEG, static and dynamic face recognition, to cite some examples), for bimodal systems (static and dynamic Audio-Visual biometric systems, EEG-Voice password-based systems, etc.) as well as for the proposed three-modal experiment. But the technical contribution of this work extends beyond the borders of biometrics to touch fields such as brain-computer interfaces (BCI) and automated-lip reading and, in a more general sense, applications where voice, video, and EEG samples are required and digit-limited vocabulary is not a restriction. The dataset can be openly accessed at http://dx.doi.org/10.17632/s7chktmb6x.1 [15].

II. LITERATURE REVIEW AND RELATED DATASETS

As the EEG modality becomes more important in the field of biometrics, the number of relevant studies increases. This section presents some of the works that in our opinion reflect outstanding and useful aspects in the development of our research.

Many multimodal datasets that include EEG signals were originally conceived to perform emotion recognition functions. DEAP, a database for emotion analysis using physiological signals [16], presents EEG and peripheral physiological signals for 32 users (ages between 19 and 37, 50% female) and video recordings for 22 of the involved subjects. The reported peripheral signals are galvanic skin response (GSR), respiration amplitude, skin temperature, electrocardiogram, blood volume by plethysmograph, electromyograms (EMG) of Zygomaticus and Trapezius muscles, and electrooculogram (EOG). The participants were asked to watch 40 one-minute music video segments. Each segment was rated by the participant's self-assessment of the levels of arousal, valence, liking, and dominance induced by the exposition to each music video segment. Hence, given the 40 samples for the 22 video-included users, a total of 880 one-minute instances of the mentioned signals are available. This data set, as well as MAHNOB-HCI [17], are widely used and are considered as references in the area.

Similarly, Rayatdoost *et al.* [18] reported an approach for emotion recognition and the collection of the required data, namely EEG signals from 64 channels, GSR, respiratory effort, EOG, and EMG signals, as well as video records of eye gaze and facial expressions for 60 subjects (ages between 17 and 67, 31 male). As for the previously mentioned datasets, volunteers were exposed to 1-2 minutes-long video excerpts (in this case, from commercial movies and user-generated material) and were asked to report their emotions for each clip. 40 clips were used for each user, giving a total of 240 instances. However, a high level of noise was reported for 13 users, reducing the used set to 47 out of the 60 available volunteers' data. Besides, no public access to the data is explicitly found in the reported paper.

VoxCeleb, as reported in [19], represents an impressive effort to curate datasets involving voice and video. So far, this project has made public two datasets: VoxCeleb1 [20] and VoxCeleb2 [21], both originally meant to perform speaker recognition experiments. These sets use a fully automated pipeline to extract utterances from YouTube videos. VoxCeleb1 selected 1,251 celebrities (690 male) from which over 100,000 utterances are collected (with an average of 18 videos and 116 utterances per person of interest). VoxCeleb2 increases the volume of the first version by a factor greater than 5, gathering a total of 1,128,246 utterances from 6,112 persons of interest extracted from 150,480 YouTube videos.

On the other hand, intended for BCI purposes, Ref. [22] introduced an open-access database of EEG signals recorded for imagined and pronounced speech of two sets of phonetic emissions: the first one containing the Spanish vowels /a/, /e/, /i/, /o/ and /u/; the second for the Spanish commands *"arriba"* (up), *"abajo"* (down), *"derecha"* (right), *"izquierda"* (left), *"adelante"* (forward) and *"atras"* (backward). Their collected data gather audio and EEG registers for each word on the vocabulary repeated 50 times for 15 subjects; a six-channel acquisition system was used for the EEG signals. This database has already been tested by the authors of this paper for biometric purposes [23].

The novelty in the database that we present in this article resides in the selection of the three involved modalities and the possibility of combining them in synchronous or asynchronous biometric schemes. In addition, it opens the door to linking cognitive studies for biometric and non-biometric applications. To the best of our knowledge, this is the first multimodal dataset based on EEG, voice, and video mainly intended for biometric purposes.

III. ACQUISITION PROTOCOL

The experiment protocol consisted of the capture of video, voice, and EEG signals while uttering a sequence of digits. Prior to the recording session, a 14-channel Emotiv$^{\text{TM}}$ Epoc wireless EEG headset was carefully set on each user. Before the start of the recording session, the user was instructed on the procedure and then taken to the recording room. An anechoic chamber was conditioned to minimize the possible presence of acoustic noise in the voice registers. The volunteers sat in front of a screen at a distance of approximately one meter.

Three computers were used for data acquisition, one for each modality. Markers were emitted by the number-presenting computer (C1) and communicated to the EEG (C2) and video (C3) recording computers using Arduinos connected to them. The proposed array is shown in Fig. 1.

Two different sessions were recorded for each user. For both of them the sequence of events was established as follows: 1) The volunteer is asked to wait for an acoustic signal indicating the start of the recording session. 2) After the signal is emitted, the user must stay as still as possible, while relaxing with eyes closed for a period of 10 seconds, until the next acoustic signal. 3) Now, with eyes opened, the user must stay relaxed for a second period of 10 seconds. 4) After this, another signal is emitted and a series of non-sequential whole numbers between 0 and 9 is presented on the screen and the user has to pronounce the displayed number. The difference between sessions lies in the length of the numbers' series: for the first session, ten digits are presented, whilst for the second, four digits per chain are presented. 5) After a series is completed, the user is granted a relaxation period of 5 seconds to breathe and swallow. 6) The next series is presented. 7) steps 4 to 6 are repeated until 10 sequences are completed. This procedure is depicted in Fig. 2.

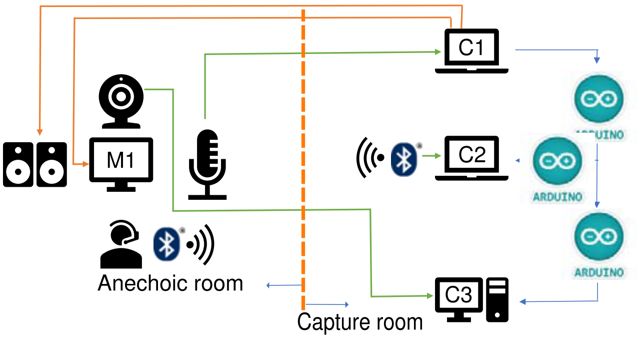


FIGURE 1. Hardware disposition for signal acquisition.

IV. MODALITIES AND PHYSICAL RESOURCES

This section describes the functions performed at each recording station (C1, C2, and C3 in Fig. 1) and provides some relevant information on the physical resources employed for the task.

A. VOICE SIGNAL

Uttered digits were recorded at an anechoic room using a Sennheiser$^{\text{TM}}$ MD 421-II Cardioid Dynamic Microphone and a Yamaha$^{\text{TM}}$ MG06X Audio Mixing Console connected to the audio input of computer C1. As shown in Fig. 1, C1 controls the audio signals, visual instructions, and digits' display at the anechoic room; it also generates event-synchronization markers to be read by computers C2 and C3.

These tasks are coded using a Matlab$^{\text{TM}}$ script. At the beginning of the REC stage, a marker with code 99 is emitted via USB port to this computer's Arduino, which is defined as the master in the I2C bus configuration. The marker code will be read from the bus by the other stations' Arduinos to be incorporated into their respective signals, as will be explained in further sections. The start-beep signal is also emitted and the instruction to "remain relaxed with eyes closed until next beep" is shown in the monitor. After ten seconds, a second marker, with code 89, is generated at the beginning of the relaxed with eyes opened (REO) stage, a beep commands the volunteer to open his/her eyes while the screen message is changed to show the present stage. Ten seconds later, a beep sound is emitted to announce the beginning of the uttering stage, and digits are presented on screen, changing after two-second intervals; for each digit, a marker is generated, coded 1-9 according to the digit presented and coded 10 when zero is presented.

As a result, 20 monoaural audio files are created per user, one for each series of digits, with a sampling frequency of 16 kHz. If digit separation is performed later, a total

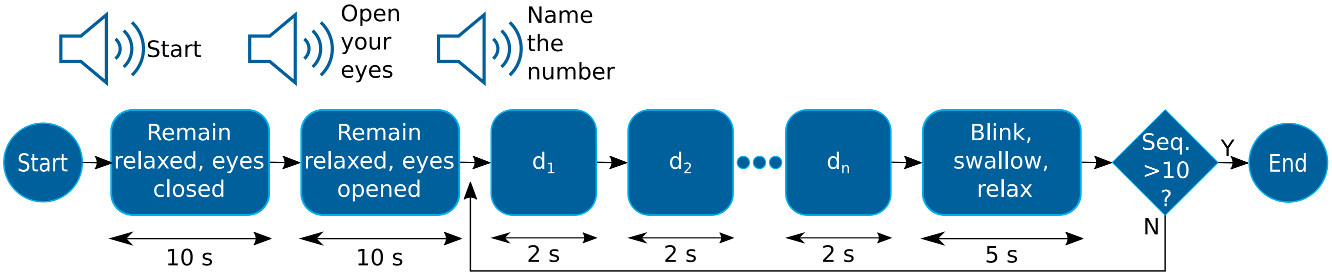


FIGURE 2. Protocol timing diagram. The number of digits *n* of the sequence is 4 or 10.

of 140 number samples can be obtained per user; 40 from the 4-digit sequences and 100 from the 10-digit sequences. Considering all 51 users, a total of 7140 audio files were generated. Table 1 shows the sequences presented for 4-digit sessions and 10-digit sessions.

TABLE 1. Digit sequences for (a) 10-digit series (b) 4-digit series.

Series	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
1	7	9	0	2	1	5	8	6	4	3
2	1	7	0	3	8	4	6	5	2	9
3	6	8	2	5	3	0	9	1	4	7
4	9	4	2	1	0	3	8	7	5	6
5	2	0	9	1	3	7	5	4	6	8
6	8	6	1	5	7	0	3	9	2	4
7	3	5	6	8	1	2	4	7	9	0
8	4	3	5	6	9	7	0	8	2	1
9	0	8	2	1	3	9	7	4	6	5
10	5	3	1	6	7	0	4	9	8	2

(a)

Series	d1	d2	d3	d4
1	1	2	3	4
2	5	3	2	9
3	1	0	7	3
4	9	6	4	7
5	5	4	2	1
6	8	3	9	6
7	7	0	6	8
8	9	5	2	3
9	0	6	4	7
10	8	1	5	0

(b)

Fig. 3 shows an example of a graphic representation for one audio file (e.g., F002_01G04_1.wav). As previously established, 20 audio files were generated by each user giving a total of 1020 files for the 51 users. The nomenclature for these files is conformed as shown in Fig. 4.

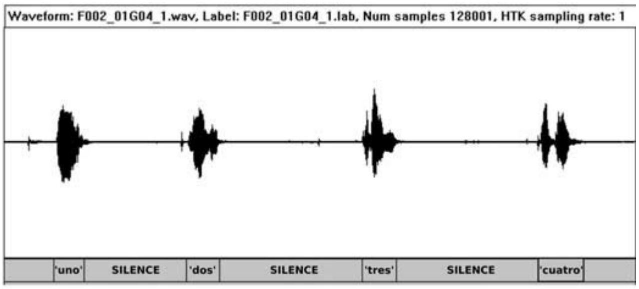

Waveform: F002_01G04_1.wav, Label: F002_01G04_1.lab, Num samples 128001, HTK sampling rate: 1
'uno'
SILENCE
'dos'
SILENCE
'tres'
SILENCE
'cuatro'

FIGURE 3. A voice sample showing a four-digit Spanish-pronounced sequence uttered by user F002.

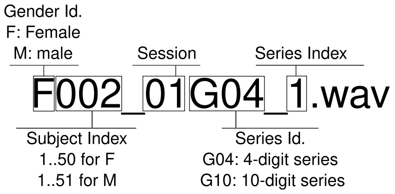


FIGURE 4. Nomenclature configuration for the generated audio files.

B. EEG SIGNAL

EEG signals were transmitted from the headset to terminal C2 via Bluetooth. Markers emitted by terminal C1 were read from the Arduino via the USB port. Both markers and signals are incorporated into the output files. European Data Format (EDF) files were created by Emotiv's Headset TestBench software. Two files per user are generated, one for the 10-digit series and another for the 4-digit sequences. These files were also converted to comma-separated values (CSV)-format files using the same TestBench software and they are available as well, along with the EDF files, for reference and use. Signal segmentation can be easily achieved to obtain REO, REC, and single-digit elements using the markers as segment boundaries.

As mentioned before, Emotiv's Epoc is a 14-channel, wet electrode wireless headset. Under the 10-20 electrode placement system, the following channels are available: AF3, F3, F7, FC5, T7, P7, O1, O2, P8, T8, FC6, F8, F4, and AF4. Signals are generated with a sampling rate of 128 samples per second. The information contained in the EDF and CSV files can be consulted on the manufacturer's website [24]. Fig. 5 shows the structure for the files nomenclature.

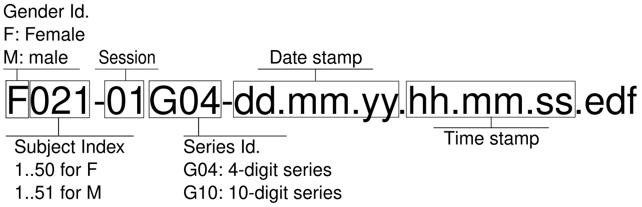
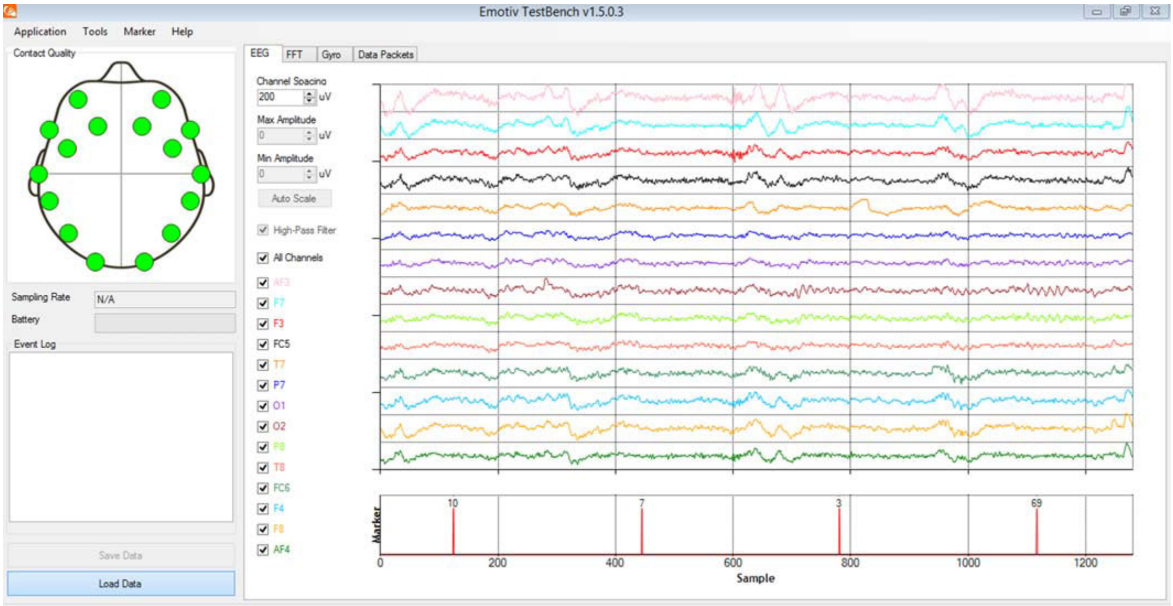


FIGURE 5. Nomenclature configuration for the generated EDF files. G04 files contain the 10 four-digit series, whilst G10 contain the 10 ten-digit ones.

Fig. 6 shows a time frame for one particular signal capture, as presented by Emotiv's TestBench software. On the upper-left, a representation of the position of the electrodes is



Channel Spacing 200 uV

Max Amplitude 0 uV

Min Amplitude 0 uV

☑ High-Pass Filter

☑ All Channels

☑ AF3
☑ F7
☑ F3
☑ FC5
☑ T7
☑ P7
☑ O1
☑ O2
☑ P8
☑ T8
☑ FC6
☑ F4
☑ F8
☑ AF4

Sampling Rate N/A

Battery

Event Log

FIGURE 6. EEG signal representation for a given sample from user F021. Second four-digit sequence shown.

shown. Green-colored circles stand for electrodes with good contact. On the right side, a representation of the channels' signals along time is presented; the red pulses at the bottom of the graphic represent the markers for the digit presentation on screen.

C. VIDEO SIGNAL

Computer C3 receives the video stream from the webcam located at the anechoic room and the markers generated by computer C1. Markers are embedded into the video file and appear in the bottom left corner. As for EEG signals, two .avi-formatted files per user are created, with a frame size of 1280 × 720, and at a frame rate of 8fps, one for the 10-digit sequence and one for the 4-digit series. Fig. 7 shows a sample of one frame from a captured video. Along with the .avi files, Matlab's .mat files with time-stamped markers are included. Due to users' privacy restrictions, video signals are unavailable for 12 users. Table 2 summarizes the dataset information and content.

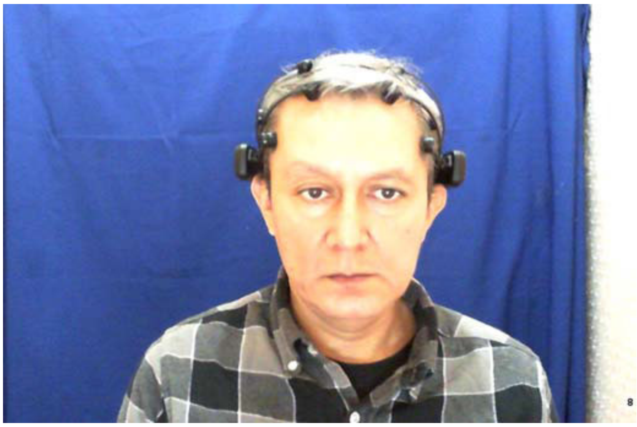


FIGURE 7. An example of a video frame during the presentation of number eight for user M003.

TABLE 2. Dataset content summary.

Modality	Users included	Files per user	File description
EEG	51	2	One file includes ten 10-digit series, a REO sequence and a REC sequence. The other file includes ten 4-digit series, a REO and a REC sequence.
Voice	51	20	10 files include 10-digit series audios and 10 files include 4-digit series audios.
Video	39	2	One file includes the video recordings of the ten 10-digit series and the other the video feed of the ten 4-digit series.

For validation purposes and initial study on data separability, the following sections present three unimodal identification experiments (one for face recognition, a speaker identification example, and an EEG identifier) and two bimodal biometric identification exercises, based on face-voice and EEG-voice.

V. EXPERIMENTAL EVALUATION, CASE I: FACE-VOICE RECOGNITION

A. INTRODUCTION

An initial set of experiments using BIOMEX-DB aiming to explore data characteristics is presented as follows. The first group of experiments is based on Deep Learning models (DL), which have been proven to provide very good results in a variety of applications in the fields of artificial intelligence, machine learning, and pattern recognition [25], and specifically in data fusion [26]. DL techniques have been successfully used in unimodal biometric approaches using several modalities such as speech [27], ECG [28], or iris [29], as well as in multimodal cases using a variety of traits such as iris/face [30], or fingerprint/ECG [31]. A relevant characteristic of DL models is their ability to extract and process features directly from raw biometric data [32], although more complex information can be extracted using deeper models, as is the case with deeply learned residual features [33]. In general, DL techniques achieve very high performance in both identification and verification cases [14], but with the associated complexity cost.

In the first part of this section, we present a unimodal recognition experiment based on CNNs, using the dataset BIOMEX-DB with face information. In the second part, face and voice modalities are fused following a CNN-based bimodal approach at a feature level. Results on identification and verification modes are described in subsection B for face recognition. Section C presents the results of a verification exercise for the fused modalities. A variety of speaker recognition systems can also be designed from the voice data; further experiments contemplate the use of a recognition framework based on the pronunciation of a personalized password formed by a certain combination of digits, combining speech and speaker recognition as a cancellable biometric scheme.

B. FACE RECOGNITION

The experiment consists of a unimodal face biometric system, with a Convolutional Neural Network. The CNN architecture is described in table 3. The images were obtained from the BIOMEX-DB database using 39 subjects. In this experiment, 30 still frames per subject were extracted at random moments from each video. The images were preprocessed through a series of operations including tilt alignment, color to grayscale conversion, and scaling down to 100 × 100 pixels. The available dataset was further divided into three parts to be used for training, validation, and testing, respectively. Categorical cross-entropy was used as the required cost function. The training was carried out with a learning rate of 0.001, and network convergence was reached after 30 epochs on average. The CNN output delivers the probability that the image under analysis corresponds to the pattern learned during the training stage. The label with the highest probability value is considered the best match for a specific trial.

The evaluation corresponding to the verification mode was carried out with a feature extraction process using the last CNN hidden layer. Therefore, each image in the database is represented by a feature vector with a dimension of 512 elements, and the whole set is used for training the network. An impostor's set was obtained from the Yalefaces dataset [34]. Cosine distance was used as the score to determine whether an input sample corresponds or not to the claimed identity. Testing on identification mode was performed using a similar approach over the available dataset. The CNN assigns an identity to each subject according to the minimum Cosine distance rule. In identification mode, the results obtained when a set of 100 trials was executed indicated an accuracy with a mean of 99.51% and a standard deviation of 0.69. The results corresponding to the verification mode with a set of 10 trials exhibited a mean EER of 1.08% with a standard deviation of 0.19.

TABLE 3. Face recognition CNN architecture.

Layers	Filters/Neurons	Size	Activation fcn
Conv2D	32	3x3	ReLu
Batch Norm	-	-	-
Max pooling	-	2x2	-
Conv2D	64	5x5	ReLu
Batch Norm	-	-	-
Max pooling	-	2x2	-
Fully connected	512	-	ReLu
Batch Norm	-	-	-
Fully connected	39	-	Softmax

C. FACE-VOICE BIMODAL BIOMETRICS

A set of bimodal face-speech experiments is then carried out following a direct concatenation of feature vectors previously normalized, aiming to have initial results which can be used for comparison purposes in further approaches. For that purpose, the CNN architecture, as well as training and testing conditions, are kept the same as in the previous experiment. Table 4 summarizes the average verification results.

TABLE 4. Bimodal verification results.

SNR (dB)	EER (%)	
	Mean	σ
0	4.39	0.55
5	2.71	0.34
10	1.99	0.18
15	1.75	0.25
Noiseless	2.67	0.35

VI. EXPERIMENTAL EVALUATION, CASE II: EEG-VOICE RECOGNITION

A. EXPERIMENT DESCRIPTION

There is a consensus among many authors, such as [35], on the levels at which the fusion of multimodal systems can be carried out. Under a biometric system pipeline, fusion can be applied at sensor level (aka signal level), feature level, score level, rank level, and decision level. This experiment is part of a performance analysis, intended to evaluate accuracy variations across different fusion levels for an EEG/voice-based bimodal biometric system. Results from a previous experiment with fusion at signal level can be looked at in [36]. As a subsequent step, fusion at feature level is presented here, according to the scheme depicted in Fig. 8. A multiple classifier performance evaluation is considered and presented for comparison purposes. As in the previous section, unimodal cases are evaluated before the execution of the bimodal one.

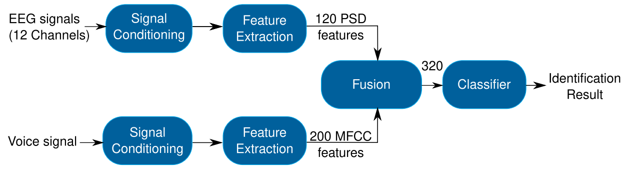


FIGURE 8. Block diagram of the proposed system with fusion at feature level.

To preserve gender balance, 50 users are included in the experiment (F001 to F025 and M001 to M025). A single-digit utterance exercise is proposed. Therefore, the sample set is made up of a total of 7,000 digit instances (140 per user). For the bimodal case, each audio file is associated with its respective EEG file. The individual digits are extracted from the original database in the signal conditioning stages.

B. EEG RECOGNITION

As mentioned in previous sections, the available EDF files contain information from 14 EEG channels of digit sequences. The first step of signal conditioning consists of the selection of channels. 12 out of the 14 available channels are selected, namely: F3, F7, FC5, T7, P7, O1, O2, P8, T8, FC6, F8, and F4; the decision to dispense with channels AF3 and AF4 was due to their content of eye-blinking artifacts and to a certain degree to decrease computational cost. High-pass filtering with a cut-off frequency of 1 Hz is applied to the 12 signals, followed by low-pass filtering with a cut-off frequency of 50 Hz. Next, a Common Average Reference (CAR) re-reference is applied to the signals. Finally, segmentation of the digit sequences to obtain single-digit samples and discarding the REO, REC, and relaxing pause segments is achieved employing a Matlab script using the digit markers contained in the EDF files as segment delimiters.

Feature extraction methods for EEG signals can be classified into three main types: time-domain, frequency-domain, and time-frequency domain [37]. For this experiment, the feature vector for the processed EEG signals is formed by the Power Spectral Density (PSD) of the beta and gamma sub-bands for all the selected channels, each one segmented on five windows with 50% overlap. Therefore, for 12 channels, the resulting feature vector has a length equal to 120.

Several classifiers were tested in an identification task to validate the suitability of the selected feature vector, with 75% of the samples reserved for training and the remaining for testing, with a 5-fold validation scheme. The most relevant results are shown in Table 5.

TABLE 5. Classifiers' accuracy comparison for EEG features.

Classifier	Accuracy (%)	
	Mean	σ
ANN	92.8	0.67
Cubic SVM	89.4	0.12
Quadratic SVM	89.4	0.18
Linear SVM	88.2	0.12
Medium Gaussian SVM	83.8	0.10
Weighted KNN	77.8	0.31
Fine KNN	73.8	0.27
Subspace discriminant	69.7	0.24
Cosine KNN	67.8	0.31
Linear discriminant	67.1	0.34

TABLE 6. Classifiers' accuracy comparison for voice features.

Classifier	Accuracy (%)	
	Mean	σ
ANN	94.2	0.64
Fine KNN	92.0	0.17
Weighted KNN	90.7	0.21
Medium gausian SVM	90.5	0.19
Cubic SVM	90.1	0.27
Quadratic SVM	88.9	0.22
Cosine KNN	88.2	0.13
Linear SVM	68.7	0.23
Subspace discriminant	63.0	0.21
Linear discriminant	60.9	0.16

TABLE 7. Classifiers' accuracy comparison for fused features.

Classifier	Accuracy (%)	
	Mean	σ
ANN	97.6	0.63
Quadratic SVM	96.6	0.12
Cubic SVM	96.3	0.18
Linear SVM	96.0	0.10
Medium Gaussian SVM	94.7	0.08
Subspace discriminant	94.3	0.10
Linear discriminant	93.8	0.08
Fine KNN	93.7	0.10
Weighted KNN	92.5	0.24
Cosine KNN	91.2	0.23

C. SPEAKER IDENTIFICATION

In terms of signal conditioning for the voice files, the 20 sequences of digits from each user are first segmented to obtain 140 audio files of 2.5 seconds length for each of the subjects. After the segmentation process, each audio file is normalized and then processed by a voice detection function which eliminates the silences in order to extract the features in the subsequent stages exclusively from voice segments of the signal. Once treated, for the resulting voice files, Mel frequency cepstral coefficients (MFCCs) and their respective delta coefficients are calculated. A Hanning window of 40 ms with 20 ms overlap is used for the extraction of 20 MFCCs and 20 delta coefficients, for a total vector length of 40. The number of feature vectors (windows) per file is variable since only the voice segments are considered for the extraction process, being the shortest one a five-windows sample and the longest, a 94-windows one.

By the addition of a fixed-length feature vector restriction, only the first five windows of all the samples are considered to obtain 200-long coefficients vectors, resulting from the concatenation of the 5 MFCCs vectors. As for the EEG case, the resulting set is tested with several classifiers under the same conditions with 75% of the samples reserved for training and under the same validation scheme. Results are shown in Table 6.

TABLE 8. Classification report for the feature-level fusion analyzed cases, using ANNs as classifiers.

	EEG				Voice				Fusion			
User	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
F001	1	1	1	34	0.917	0.971	0.943	34	0.971	1	0.986	34
F002	0.854	1	0.921	35	0.939	0.886	0.912	35	1	0.914	0.955	35
F003	1	1	1	35	1	0.886	0.939	35	0.946	1	0.972	35
F004	0.879	0.829	0.53	35	0.895	0.971	0.932	35	1	0.914	0.955	35
F005	0.914	0.914	0.914	35	0.833	0.714	0.769	35	1	0.857	0.923	35
F006	0.97	0.914	0.941	35	1	1	1	35	0.919	0.971	0.944	35
F007	0.941	0.914	0.928	35	0.97	0.914	0.941	35	0.895	0.971	0.932	35
F008	0.919	0.971	0.944	35	0.868	0.943	0.904	35	0.972	1	0.986	35
F009	0.972	1	0.986	35	1	0.914	0.955	35	0.972	1	0.986	35
F010	0.971	0.971	0.971	35	0.938	0.857	0.896	35	0.946	1	0.972	35
F011	0.971	0.971	0.971	35	0.97	0.914	0.941	35	0.972	1	0.986	35
F012	0.938	0.857	0.896	35	0.882	0.857	0.87	35	0.944	0.971	0.958	35
F013	0.969	0.886	0.925	35	0.971	0.943	0.957	35	1	0.971	0.986	35
F014	0.972	1	0.986	35	0.892	0.943	0.917	35	1	0.943	0.971	35
F015	1	0.943	0.971	35	0.972	1	0.986	35	1	1	1	35
F016	1	0.971	0.986	35	0.919	0.971	0.944	35	1	1	1	35
F017	1	0.971	0.986	35	0.861	0.886	0.873	35	1	0.943	0.971	35
F018	1	0.943	0.971	35	0.944	0.971	0.958	35	0.944	0.971	0.958	35
F019	0.969	0.886	0.925	35	0.861	0.886	0.873	35	0.944	0.971	0.958	35
F020	0.795	0.886	0.838	35	1	0.943	0.971	35	0.972	1	0.986	35
F021	0.97	0.914	0.941	35	0.889	0.914	0.901	35	1	0.943	0.971	35
F022	1	1	1	35	1	0.971	0.986	35	0.972	1	0.986	35
F023	1	1	1	35	0.972	1	0.986	35	0.944	0.971	0.958	35
F024	0.914	0.914	0.914	35	0.829	0.971	0.895	35	0.921	1	0.959	35
F025	0.939	0.886	0.912	35	0.892	0.943	0.917	35	0.971	0.943	0.957	35
M001	0.917	0.943	0.93	35	1	1	1	35	1	1	1	35
M002	0.946	1	0.972	35	0.895	0.971	0.932	35	1	0.943	0.971	35
M003	0.972	1	0.986	35	0.895	0.971	0.932	35	1	1	1	35
M004	0.816	0.886	0.849	35	0.921	1	0.959	35	0.97	0.914	0.941	35
M005	0.838	0.886	0.861	35	0.971	0.943	0.957	35	1	0.971	0.986	35
M006	1	0.914	0.955	35	0.914	0.914	0.914	35	0.971	0.971	0.971	35
M007	1	0.857	0.923	35	1	0.971	0.986	35	1	0.943	0.971	35
M008	0.875	1	0.933	35	0.971	0.971	0.971	35	0.972	1	0.986	35
M009	0.943	0.943	0.943	35	1	1	1	35	1	1	1	35
M010	1	1	1	35	1	0.829	0.906	35	1	1	1	35
M011	0.895	0.971	0.932	35	0.971	0.943	0.957	35	0.944	0.971	0.958	35
M012	0.727	0.914	0.81	35	1	0.943	0.971	35	1	1	1	35
M013	0.944	0.971	0.958	35	0.943	0.943	0.943	35	1	1	1	35
M014	0.917	0.943	0.93	35	0.966	0.8	0.875	35	1	0.943	0.971	35
M015	0.732	0.857	0.789	35	1	1	1	35	0.972	1	0.986	35
M016	0.906	0.829	0.866	35	0.971	0.971	0.971	35	1	0.971	0.986	35
M017	0.971	0.943	0.957	35	1	1	1	35	0.972	1	0.986	35
M018	0.88	0.629	0.733	35	0.972	1	0.986	35	0.971	0.971	0.971	35
M019	0.886	0.886	0.886	35	0.971	0.971	0.971	35	1	0.971	0.986	35
M020	0.914	0.914	0.914	35	0.972	1	0.986	35	0.972	1	0.986	35
M021	0.97	0.914	0.941	35	0.971	0.971	0.971	35	1	1	1	35
M022	0.944	0.971	0.958	35	1	0.914	0.955	35	1	1	1	35
M023	0.97	0.914	0.941	35	0.81	0.971	0.883	35	0.921	1	0.956	35
M024	0.919	0.971	0.944	35	0.941	0.914	0.928	35	1	1	1	35
M025	0.824	0.8	0.812	35	0.921	1	0.959	35	0.971	0.971	0.971	35
Accuracy			0.928	1749			0.942	1749			0.976	1749
Macro avg	0.931	0.928	0.928	1749	0.944	0.942	0.942	1749	0.977	0.976	0.976	1749
Weighted avg	0.931	0.928	0.928	1749	0.944	0.942	0.942	1749	0.977	0.976	0.976	1749
Cohen Kappa score			0.926				0.940				0.975	

D. EEG-VOICE BIMODAL BIOMETRICS

After the unimodal evaluation, both EEG and voice feature vectors are then fused by concatenation to form a resulting vector with 320 elements to be fed as input to the classification stage. As well as for the single modalities cases, the same classifiers were tested, producing the results shown in Table 7.

As it can be appreciated in Table 7 the best performance was obtained by an ANN, made up of an input layer of 320 nodes, a hidden layer with 640 neurons and ReLu activation function, a dropout layer with a dropout coefficient of 0.25, and a Softmax-activated output layer with 50 output nodes. The network was set to be trained with an Adam optimizer and a sparse categorical cross-entropy as loss function. To preserve consistency for the network performance evaluation, a 4-fold validation scheme is selected, with 75% of the available samples for training and the remaining 25% for testing. The ANN is trained across 150 epochs. Fig. 9 shows

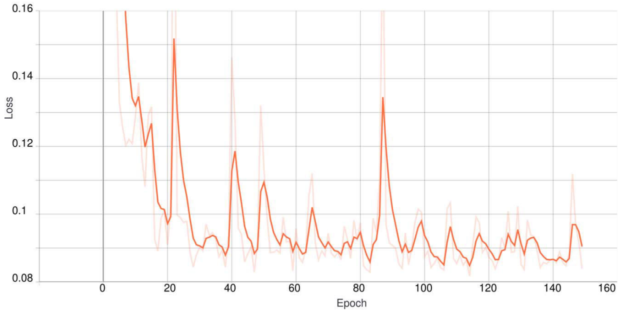


FIGURE 9. Loss function evolution across epoch of the training stage.

the loss function evolution across epochs, whereas Fig. 10 shows the accuracy evolution as the network is trained.

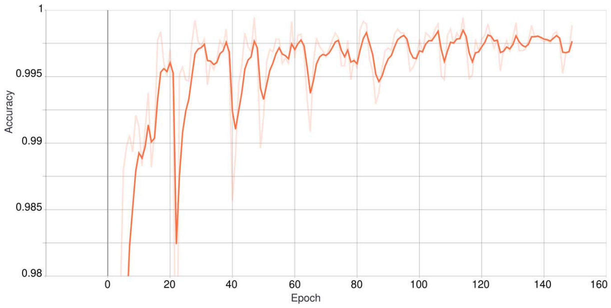


FIGURE 10. Accuracy evolution across epoch of the training stage.

For comparative purposes, Fig. 11 summarizes the results obtained for the best-evaluated classifiers. As expected, the obtained results confirm the achievement of higher accuracies when bimodal systems are attempted. To complete the comparative analysis, Table 8 presents the classification report for both unimodal cases and the fused one for the ANN classifiers.

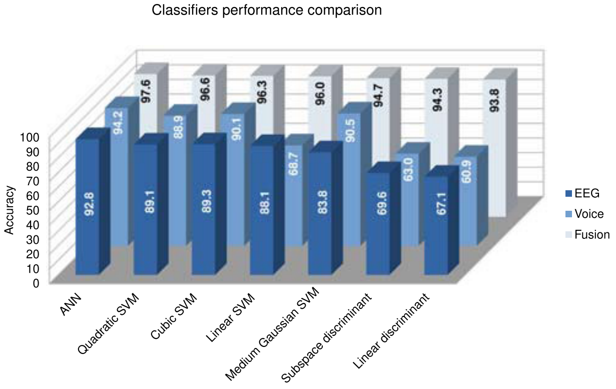


FIGURE 11. Classifiers' accuracy comparison: EEG, Voice and Fusion.

VII. CONCLUSION

An open-access database of synchronously recorded EEG, voice and video signals to be used in biometric projects has been introduced, and a collection of experiments explored data separability of each modality. As previously established, the main underlying justification for multimodal biometrics is the improvement of one or more of the system's desired characteristics. The experimental cases presented in this article validate this argument taking into consideration mainly the performance dimension. The presented database gathers three modalities with different characteristics whose objective is to create a robust recognition system, in which the weakness of a modality is compensated by another modality's strength. In particular, the database relies on the proven collectability and acceptance of voice recognition, the universality and circumvention of EEG, and the permanence and collectability of video stream modalities. Furthermore, when modalities are synchronously used, the robustness of liveness detection increases. The database represents a rich source for multimodal biometric investigation projects and in general for any project in which the use of video feed, voice samples, or EEG signals is required.

REFERENCES

[1] O. Nieves and V. Manian, ''Automatic person authentication using fewer channel EEG motor imagery,'' in *Proc. World Automat. Congr. (WAC)*, Jul. 2016, pp. 1–6.

[2] T. Z. Chin, A. Saidatul, and Z. Ibrahim, ''Exploring EEG based authentication for imaginary and non-imaginary tasks using power spectral density method,'' *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 557, Jun. 2019, Art. no. 012031.

[3] R. Saini, B. Kaur, P. Singh, P. Kumar, P. P. Roy, B. Raman, and D. Singh, ''Don't just sign use brain too: A novel multimodal approach for user identification and verification,'' *Inf. Sci.*, vols. 430–431, pp. 163–178, Mar. 2018.

[4] V. Talreja, M. C. Valenti, and N. M. Nasrabadi, ''Deep hashing for secure multimodal biometrics,'' *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1306–1321, 2021.

[5] S. Anton, T. Artem, P. Andrey, and K. Igor, ''Modification of VGG neural network architecture for unimodal and multimodal biometrics,'' in *Proc. IEEE East-West Design Test Symp. (EWDTS)*, Sep. 2020, pp. 1–4.

[6] J. Zapata, C. Duque, Y. Rojas-Idarraga, M. Gonzalez, J. Guzmán, and A. B. Botero, ''Data fusion applied to biometric identification—A review,'' in *Proc. Colombian Conf. Comput.* Cham, Switzerland: Springer, 2017, pp. 721–733.

[7] M. Karakaya and E. T. Celik, ''Effect of pupil dilation on off-angle iris recognition,'' *J. Electron. Imag.*, vol. 28, no. 3, 2019, Art. no. 033022.

[8] M. Choudhary, V. Tiwari, and U. Venkanna, ''Iris anti-spoofing through score-level fusion of handcrafted and data-driven features,'' *Appl. Soft Comput.*, vol. 91, Jun. 2020, Art. no. 106206.

[9] S. Eberz, K. B. Rasmussen, V. Lenders, and I. Martinovic, ''Evaluating behavioral biometrics for continuous authentication: Challenges and metrics,'' in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 386–399.

[10] W. Meng, D. S. Wong, S. Furnell, and J. Zhou, ''Surveying the development of biometric user authentication on mobile phones,'' *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1268–1293, 3rd Quart., 2015.

[11] A. Sundararajan, A. I. Sarwat, and A. Pons, ''A survey on modality characteristics, performance evaluation metrics, and security for traditional and wearable biometric systems,'' *ACM Comput. Surv.*, vol. 52, no. 2, pp. 1–36, May 2019.

[12] S. Shunmugam and R. Selvakumar, ''Electronic transaction authentication—A survey on multimodal biometrics,'' in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2014, pp. 1–4.

[13] M. Ghayoumi, ''A review of multimodal biometric systems: Fusion methods and their applications,'' in *Proc. IEEE/ACIS 14th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2015, pp. 131–136.

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

[15] J. C. M. Rodríguez, J. C. A. Vazquez, R. Fonseca-Delgado, J. M. Ramirez-Cortes, P. Gomez-Gil, and R. Arechiga-Martinez, "Mendeley data—BIOMEX-DB," INAOE, San Andrés Cholula, Mexico, Tech. Rep. 1, 2021.

[16] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans. Affective Comput.*, vol. 3, no. 1, pp. 18–31, Oct./Mar. 2012.

[17] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012.

[18] S. Rayatdoost, D. Rudrauf, and M. Soleymani, "Expression-guided EEG representation learning for emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3222–3226.

[19] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Comput. Speech Lang.*, vol. 60, Mar. 2020, Art. no. 101027.

[20] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech*, Aug. 2017, pp. 2616–2620.

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech*, Sep. 2018, pp. 1086–1090.

[22] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in *Proc. 12th Int. Symp. Med. Inf. Process. Anal.*, Jan. 2017, Art. no. 1016002.

[23] J. C. Moreno-Rodriguez, J. M. Ramirez-Cortes, R. Arechiga-Martinez, P. Gomez-Gil, and J. C. Atenco-Vazquez, "Bimodal biometrics using EEG-voice fusion at score level based on hidden Markov models," in *Intuitionistic and Type-2 Fuzzy Logic Enhancements in Neural and Optimization Algorithms: Theory and Applications*. Cham, Switzerland: Springer, 2020, pp. 645–657.

[24] *EDF Files—EmotivPRO*, Emotiv, San Francisco, CA, USA, 2018.

[25] P. Bharati and A. Pramanik, "Deep learning techniques—R-CNN to mask R-CNN: A survey," in *Computational Intelligence in Pattern Recognition*. Singapore: Springer, 2020, pp. 657–668.

[26] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Comput.*, vol. 32, no. 5, pp. 829–864, May 2020.

[27] S. Borwankar, S. Bhatnagar, Y. Jha, S. Pandey, and K. Jain, "Improved automatic speaker verification system using deep learning," in *Proc. Int. Conf. Inf. Commun. Technol. Intell. Syst.* Singapore: Springer, 2020, pp. 523–531.

[28] D. Belo, N. Bento, H. Silva, A. Fred, and H. Gamboa, "ECG biometrics using deep learning and relative score threshold classification," *Sensors*, vol. 20, no. 15, p. 4078, Jul. 2020.

[29] M. Sardar, S. Banerjee, and S. Mitra, "Iris segmentation using interactive deep learning," *IEEE Access*, vol. 8, pp. 219322–219330, 2020.

[30] S. Arora, M. Bhatia, and H. Kukreja, "A multimodal biometric system for secure user identification based on deep learning," in *Proc. Int. Congr. Inf. Commun. Technol.* Singapore: Springer, 2020, pp. 95–103.

[31] R. M. Jomaa, H. Mathkour, Y. Bazi, and M. S. Islam, "End-to-end deep learning fusion of fingerprint and electrocardiogram signals for presentation attack detection," *Sensors*, vol. 20, no. 7, p. 2085, Apr. 2020.

[32] 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, Sep. 2020.

[33] Y. Liu and A. Kumar, "Contactless palmprint identification using deeply learned residual features," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 2, no. 2, pp. 172–181, Apr. 2020.

[34] K.-C. Lee, J. Ho, and D. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.

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

[36] J. C. Moreno-Rodriguez, J. M. Ramirez-Cortes, J. C. Atenco-Vazquez, and R. Arechiga-Martinez, "EEG and voice bimodal biometric authentication scheme with fusion at signal level," in *Proc. IEEE Mex. Humanitarian Technol. Conf. (MHTC)*, Apr. 2021, pp. 52–58.

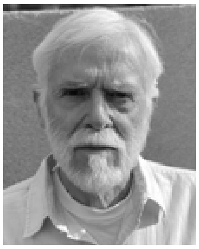
[37] B. Goudiaby, A. Othmani, and A. Nait-Ali, "EEG biometrics for person verification," in *Hidden Biometrics*. Singapore: Springer, 2020, pp. 45–69.

[38] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," 2015, *arXiv:1510.08484*. [Online]. Available: http://arxiv.org/abs/1510.08484

JUAN CARLOS MORENO-RODRIGUEZ (Graduate Student Member, IEEE) was born in Puebla, Mexico, in 1971. He received the B.Sc. and M.Sc. degrees in electronic engineering from the Universidad de las Américas-Puebla, in 1995 and 1998, respectively. He is currently pursuing the Ph.D. degree in electronics with the National Institute of Astrophysics, Optics and Electronics, Mexico. He has worked as an Assistant Professor with the departments of computer systems and electronics at the Tecnologico Nacional de Mexico and Universidad Iberoamericana. His research interests include biometrics, machine learning, and signal processing.

JUAN CARLOS ATENCO-VAZQUEZ was born in Puebla, Mexico, in 1991. He received the B.Sc. degree from the Puebla Institute of Technology (ITP), Mexico, and the M.Sc. degree from the National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico, where he is currently pursuing the Ph.D. degree with the Department of Electronics. His research interests include signal processing, biometric systems, embedding systems, and neural networks and its applications.

JUAN MANUEL RAMIREZ-CORTES (Senior Member, IEEE) 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. degree from Texas Tech University, all in electrical engineering. He currently holds a researcher position at INAOE. He is a member of Mexican National Research System (SNI), level 2. His research interests include signal and image processing, biometric, neural networks, fuzzy logic, and digital systems.



RENE ARECHIGA-MARTINEZ received the B.Sc. degree from the National Polytechnic Institute, Mexico, the M.Sc. degree from Stanford University, and the Ph.D. degree from The University of New Mexico, all in electrical engineering. He is currently an Associate Professor with the Department of Electrical Engineering, New Mexico Tech. His research interests include digital signal processing applied to speech recognition and thunderstorms.

PILAR GOMEZ-GIL received the B.Sc. degree in computer science from the Universidad de las Americas, A.C., Mexico, and the M.Sc. and Ph.D. degrees in computer science from Texas Tech University, USA. She is currently a Titular Researcher with the Department of Computer Science, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico. She is a member of the Mexican National Research System (SNI), level 1. Her research interests include artificial neural networks, time series prediction, image processing, and pattern recognition.

RIGOBERTO FONSECA-DELGADO received the B.Sc. degree from the Faculty of System Engineering, National Polytechnic School, Ecuador, and the M.Sc. and Ph.D. degrees from the National Institute of Astrophysics, Optics and Electronics, Mexico. He is currently a Professor with Metropolitan Autonomous University, Iztapalapa, Mexico. His research interests include classification and prediction of time series, resource allocation, and artificial intelligence applications.