This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

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

JUAN CARLOS MORENO-RODRIGUEZ[1], JUAN CARLOS ATENCO-VAZQUEZ[1], JUAN MANUEL RAMIREZ-CORTES[1], RENE ARECHIGA-MARTINEZ [2], PILAR GOMEZ-GIL [3], AND RIGOBERTO FONSECA-DELGADO[4]

[1]Department of Electronics, National Institute of Astrophysics, Optics and Electronics, Luis Enrique Erro 1, Sta. Maria Tonantzintla 72840, Mexico (e-mail: xalatl@inaoep.mx)
[2]New Mexico Tech, 801 Leroy Place, Socorro, NM 87801, USA, Mexico (e-mail: rene.arechiga@nmt.edu)
[3]Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Luis Enrique Erro 1, Sta. Maria Tonantzintla 72840, Mexico (e-mail: pgomez@inaoep.mx)
[4]Electrical Engineering Department, Metropolitan Autonomous University, San Rafael Atlixco 186, 09340 Iztapalapa, CDMX, Mexico(e-mail: rfonseca@izt.uam.mx)

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

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

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, in this paper we introduce 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, six user recognition experiments are presented: a face recognition identifier with accuracy of 99%, two speaker identification systems with maximum accuracy of 100%, 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, have 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], [4], [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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

described in terms of its capacity to 1) always accept a legitimate user while rejecting all impostors (verification systems) or 2) correctly identify the presenting users with the registered identities in the database (identification systems). Many metrics have been defined in order 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, Wong, Furnell and Zhou [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. In a similar fashion 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 years old ($\overline{x} = 29.75$, $\sigma = 10.97$); 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 mentioned three-modal proposed 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. RELATED DATASETS

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 years, 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 music video segments of one minute length. 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 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 felt emotions for each clip. 40 clips were used for each user, giving a total of 240 instances. However, 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 form 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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

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].

III. ACQUISITION PROTOCOL

The experiment protocol consisted in the capture of video, voice and EEG signals of the uttering of a sequence of digits. Prior to the recording session, a 14-channel Emotiv™ 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 Figure 1.

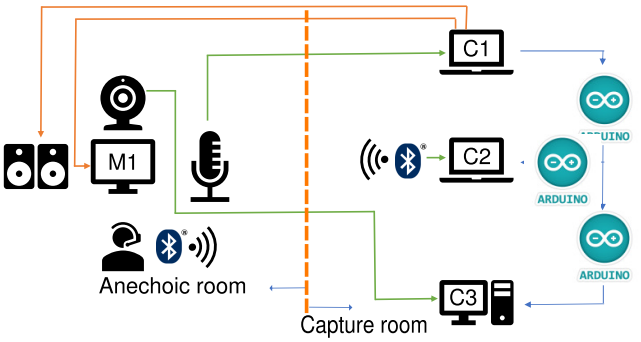


FIGURE 1. Hardware disposition for signal acquisition.

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. The user has to pronounce the displayed number. The difference between sessions lies on 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 breath 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 Figure 2.

IV. MODALITIES AND PHYSICAL RESOURCES

This section describes the functions performed at each recording station (C1, C2 and C3 in figure 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™ MD 421-II Cardioid Dynamic Microphone and a Yamaha™ MG06X Audio Mixing Console connected to the audio input of computer C1. As shown in figure 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™ script. At the beginning of the relaxed with eyes closed (REC) stage, a marker with code 99 is emitted via USB port to this computer's Arduino, which is defined as master in the I2C bus configuration. The marker code will be read from the bus by the other stations' Arduinos to be incorporated to 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 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 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 be generated. Table 1 shows the sequences presented for 4-digit sessions and 10-digit sessions.

Figure 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 user giving a total of 1020 files for the 51 users. The nomenclature for these files is conformed as shown in figure 4.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

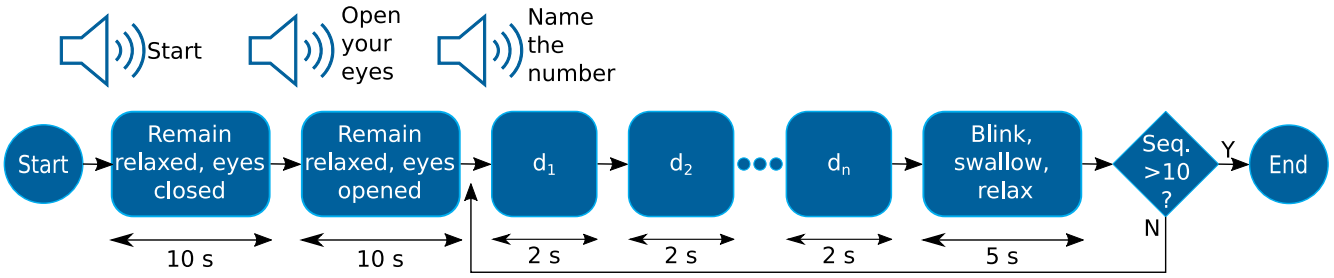


FIGURE 2. Protocol timing diagram. The number of digits *n* of the sequence is either 4 or 10.

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)

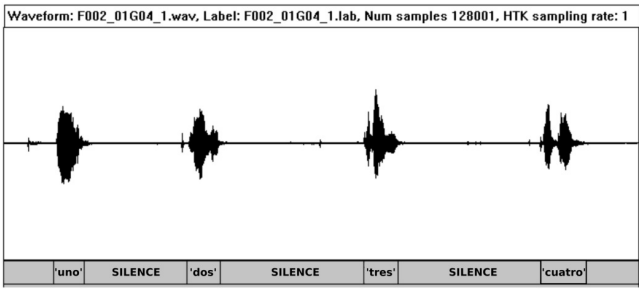


FIGURE 3. A voice sample showing a four-digit Spanish-pronounced sequence uttered by user F002.

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. EDF-format 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 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.

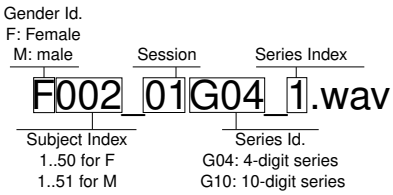


FIGURE 4. Nomenclature configuration for the generated audio files.

As mentioned before, Emotiv's Epoc is a 14-channel, wet electrode wireless headset. In accordance with 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 in the manufacturer's website [24]. Figure 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.

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

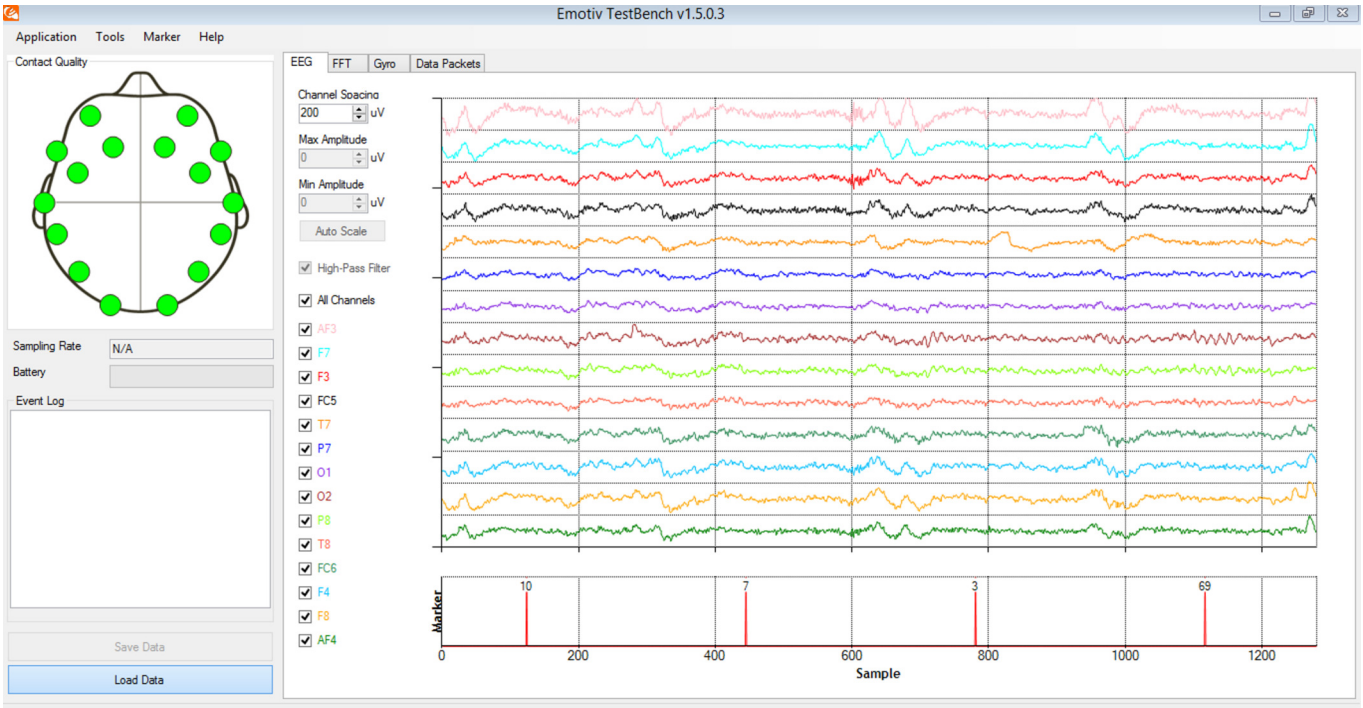


FIGURE 6. EEG signal representation for a given sample from user F021. Second four-digit sequence shown.

.avi-formatted files per user are created, with frame size of 1280x720, and with frame rate of 8fps, one for the 10-digit sequence and one for the 4-digit series. Figure 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.

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 four unimodal identification experiments (one for face recognition, two speaker recognition examples and an EEG identifier) and two bimodal biometric identification exercises, based on face-voice and EEG-voice.

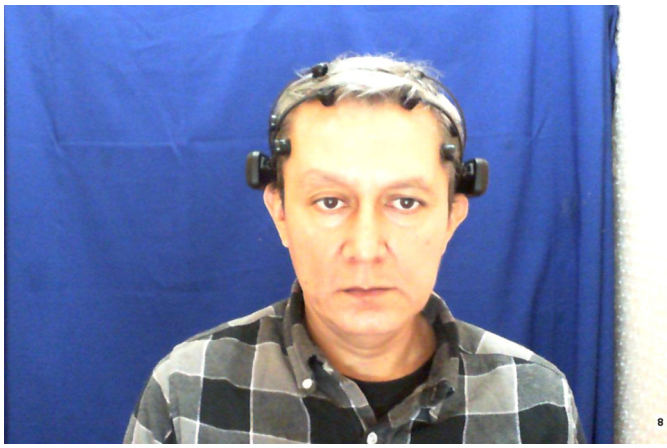


FIGURE 7. An example of a video frame during the presentation of number eight for user M003.

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], Electrocardiogram

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

(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 it 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 two unimodal recognition experiments based on Convolutional Neural Networks (CNN), using the dataset BIOMEX-DB with voice and face information. In the second part, these two individual modalities are fused following a CNN-based bimodal approach at feature level. Results on identification and verification modes are described.

B. SPEAKER RECOGNITION

The first experiment consists of a unimodal speech-based biometric system implemented in a Convolutional Neural Network (CNN) Sincnet framework. The architecture is shown in Table 3. Categorical cross-entropy was used as the cost function within an Adam optimizer. Training of the CNN was carried out with a learning rate of 0.001 in 50 epochs. Voice signals were obtained from the presented database using 39 subjects. The 10-digit utterances were used for training and validation, while the 4-digit utterances were reserved to be used during the testing process.

TABLE 3. Speaker recognition Sincnet architecture

Layers	Filters/Neurons	Size	Activation fcn
Sinc Conv1d	120	251	ReLu
Batch Norm	-	-	-
Max Pooling	-	5	-
Conv1D	32	5	ReLu
Batch Norm	-	-	-
Max Pooling	-	5	-
Conv1D	64	5	ReLu
Batch Norm	-	-	-
Max Pooling	-	5	-
Fully connected	512	-	ReLu
Batch Norm	-	-	-
Fully connected	39	-	Softmax

In every case, the raw speech signals were organized in segments of 200 ms length, and normalized in amplitude. In order to test the noise-related robustness of the network, a process of data augmentation was carried out using noise signals obtained from the MUSAN database [34] at several signal to noise ratio (SNR) values ranging from 0 to 15 dB. The average results corresponding to identification and verification modes obtained are shown in Table 4.

TABLE 4. Speaker recognition results.

SNR (dB)	Identification				Verification	
	Frame Accuracy (%)		Sentence Accuracy (%)		EER (%)	
	Mean	σ	Mean	σ	Mean	σ
0	76.64	5.6	93.43	1.31	15.48	0.36
5	85.33	5.36	98.25	0.52	9.03	0.24
10	86.99	4.87	99.48	0.39	4.66	0.31
15	89.53	4.33	100	0	4.53	0.21
Noiseless	88.07	3.93	100	0	4.36	0.18

C. FACE RECOGNITION

The second experiment consists of a unimodal face biometric system, implemented using an approach similar to the previous case, with a Convolutional Neural Network. The CNN architecture is described in table 5. The images were obtained from the BIOMEX-DB database using the same set of 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 100x100 pixels. The available dataset was further divided in three parts to be used for training, validation, and testing, respectively. Categorical cross-entropy was used as the required cost function. 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.

TABLE 5. 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

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 [35]. 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 in average 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.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

TABLE 6. 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

D. FACE-VOICE BIMODAL BIOMETRICS

A set of bimodal face-speech experiment 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 previous experiments. Table 6 summarizes the average verification results.

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

A. EXPERIMENT DESCRIPTION

There is a general consensus among many authors, such as [36], on the levels at which the fusion of multimodal systems can be carried out. In accordance with 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 [37]. As a subsequent step, fusion at feature level is presented here, according to the scheme depicted in figure 8. A multiple classifier performance evaluation is considered

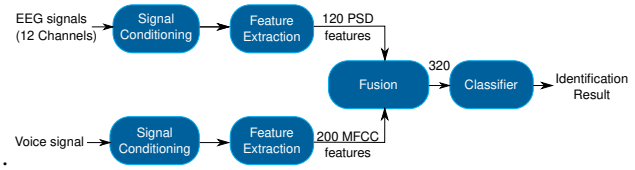


FIGURE 8. Block diagram of the proposed system with fusion at feature level

and presented for comparison purposes. As in the previous section, unimodal cases are evaluated prior to the execution of the bimodal one.

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 biomodal 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 the information of 14 EEG channels of digit-sequences. The first step of signal conditioning consists in 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. High-pass filtering with cut-off frequency of 1 Hz is applied to the 12 signals, followed by a low-pass filtering with 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 by means of 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 [38]. 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 5 windows with 50% overlap. Therefore, for 12 channels, the resulting feature vector has a length equal to 120.

To validate the suitability of the selected feature vector, several classifiers were tested in an identification task, with 75% of the samples reserved for training and the remaining for testing, with a 5-fold validation scheme, obtaining, among others, the results shown in Table 7.

TABLE 7. 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

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.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

TABLE 8. 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

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 of 75% of the samples reserved for training and under the same validation scheme. Results are shown in Table 8.

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 9.

TABLE 9. 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

As it can be appreciated in Table 9 the best performance was obtained by an ANN, made up by an input layer of 320 nodes, a hidden layer with 640 neurons and ReLu activation function, a dropout layer with 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. Figure 9 shows the loss function evolution across epochs, whereas figure 10 shows the accuracy evolution as the network is trained.

For comparative purposes, Figure 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.

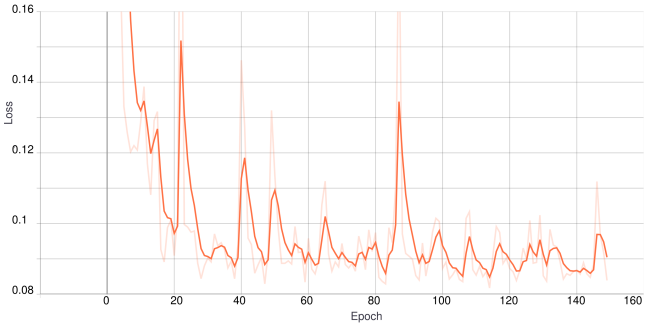


FIGURE 9. Loss function evolution across epoch of the training stage.

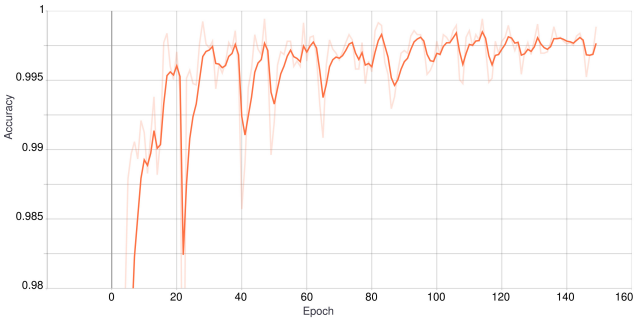


FIGURE 10. Accuracy evolution across epoch of the training stage.

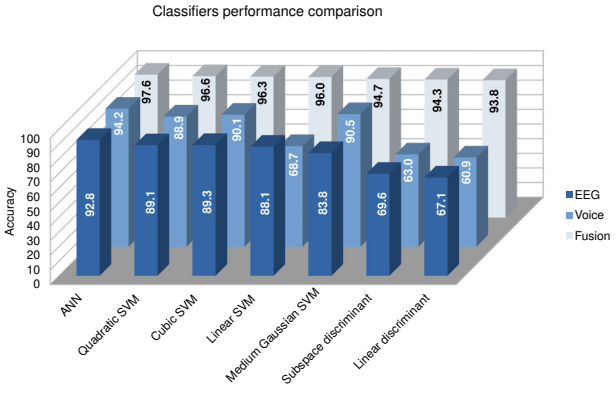


FIGURE 11. Classifier's accuracy comparison: EEG, voice and fusion.

VII. CONCLUSIONS

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 explores 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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

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 are required.

REFERENCES

[1] Orlando Nieves and Vidya Manian. Automatic person authentication using fewer channel eeg motor imagery. In 2016 World Automation Congress (WAC), pages 1–6. IEEE, 2016.

[2] Tze Zhi Chin, A Saidatul, and Z Ibrahim. Exploring eeg based authentication for imaginary and non-imaginary tasks using power spectral density method. In IOP Conference Series: Materials Science and Engineering, volume 557, page 012031. IOP Publishing, 2019.

[3] Rajkumar Saini, Barjinder Kaur, Priyanka Singh, Pradeep Kumar, Partha Pratim Roy, Balasubramanian Raman, and Dinesh Singh. Don't just sign use brain too: A novel multimodal approach for user identification and verification. Information Sciences, 430:163–178, 2018.

[4] Veeru Talreja, Matthew C Valenti, and Nasser M Nasrabadi. Deep hashing for secure multimodal biometrics. IEEE Transactions on Information Forensics and Security, 16:1306–1321, 2020.

[5] Stefanidi Anton, Topnikov Artem, Priorov Andrey, and Kosterin Igor. Modification of vgg neural network architecture for unimodal and multimodal biometrics. In 2020 IEEE East-West Design & Test Symposium (EWDTS), pages 1–4. IEEE, 2020.

[6] JC Zapata, CM Duque, Y Rojas-Idarraga, ME Gonzalez, JA Guzmán, and MA Becerra Botero. Data fusion applied to biometric identification–a review. In Colombian Conference on Computing, pages 721–733. Springer, 2017.

[7] Mahmut Karakaya and Elif T Celik. Effect of pupil dilation on off-angle iris recognition. Journal of Electronic Imaging, 28(3):033022, 2019.

[8] Meenakshi Choudhary, Vivek Tiwari, and U Venkanna. Biometric spoofing: Iris presentation attack detection and contact lens discrimination through score-level fusion. Applied Soft Computing, page 106206, 2020.

[9] Simon Eberz, Kasper B Rasmussen, Vincent Lenders, and Ivan Martinovic. Evaluating behavioral biometrics for continuous authentication: Challenges and metrics. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, pages 386–399, 2017.

[10] Weizhi Meng, Duncan S Wong, Steven Furnell, and Jianying Zhou. Surveying the development of biometric user authentication on mobile phones. IEEE Communications Surveys & Tutorials, 17(3):1268–1293, 2014.

[11] Aditya Sundararajan, Arif I Sarwat, and Alexander Pons. A survey on modality characteristics, performance evaluation metrics, and security for traditional and wearable biometric systems. ACM Computing Surveys (CSUR), 52(2):1–36, 2019.

[12] S Shunmugam and RK Selvakumar. Electronic transaction authentication—a survey on multimodal biometrics. In 2014 IEEE International Conference on Computational Intelligence and Computing Research, pages 1–4. IEEE, 2014.

[13] Mehdi Ghayoumi. A review of multimodal biometric systems: Fusion methods and their applications. In 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pages 131–136. IEEE, 2015.

[14] Hareesh Mandalapu, Aravinda Reddy PN, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, SR Mahadeva Prasanna, and Christoph Busch. Audio-visual biometric recognition and presentation attack detection: A comprehensive survey. IEEE Access, 9:37431–37455, 2021.

[15] Juan Carlos Moreno Rodríguez, Juan Carlos Atenco Vazquez, Rigoberto Fonseca-Delgado, Juan Manuel Ramirez-Cortes, Pilar Gomez-Gil, and Rene Arechiga-Martinez. Mendeley Data - BIOMEX-DB, 2021.

[16] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis; using physiological signals. IEEE transactions on affective computing, 3(1):18–31, 2011.

[17] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. IEEE transactions on affective computing, 3(1):42–55, 2011.

[18] Soheil Rayatdoost, David Rudrauf, and Mohammad Soleymani. Expression-guided eeg representation learning for emotion recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3222–3226. IEEE, 2020.

[19] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language, 60:101027, 2020.

[20] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. Proc. Interspeech 2017, pages 2616–2620, 2017.

[21] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. Proc. Interspeech 2018, pages 1086–1090, 2018.

[22] Germán A Pressel Coretto, Iván E Gareis, and H Leonardo Rufiner. Open access database of eeg signals recorded during imagined speech. In 12th International Symposium on Medical Information Processing and Analysis, volume 10160, page 1016002. International Society for Optics and Photonics, 2017.

[23] Juan Carlos Moreno-Rodriguez, Juan Manuel Ramirez-Cortes, Rene Arechiga-Martinez, Pilar Gomez-Gil, and Juan Carlos 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, pages 645–657. Springer, 2020.

[24] Inc. Emotiv. EDF files - EmotivPRO, 2018.

[25] Puja Bharati and Ankita Pramanik. Deep learning techniques—r-cnn to mask r-cnn: A survey. In Computational Intelligence in Pattern Recognition, pages 657–668. Springer, 2020.

[26] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. A survey on deep learning for multimodal data fusion. Neural Computation, 32(5):829–864, 2020.

[27] Saumya Borwankar, Shrey Bhatnagar, Yash Jha, Shraddha Pandey, and Khushi Jain. Improved automatic speaker verification system using deep learning. In International Conference on Information and Communication Technology for Intelligent Systems, pages 523–531. Springer, 2020.

[28] David Belo, Nuno Bento, Hugo Silva, Ana Fred, and Hugo Gamboa. Ecg biometrics using deep learning and relative score threshold classification. Sensors, 20(15):4078, 2020.

[29] Mousumi Sardar, Subhashis Banerjee, and Sushmita Mitra. Iris segmentation using interactive deep learning. IEEE Access, 8:219322–219330, 2020.

[30] Shefali Arora, MPS Bhatia, and Harshita Kukreja. A multimodal biometric system for secure user identification based on deep learning. In International Congress on Information and Communication Technology, pages 95–103. Springer, 2020.

[31] Rami M Jomaa, Hassan Mathkour, Yakoub Bazi, and Md Saiful Islam. End-to-end deep learning fusion of fingerprint and electrocardiogram signals for presentation attack detection. Sensors, 20(7):2085, 2020.

[32] Nada Alay and Heyam H Al-Baity. Deep learning approach for multimodal biometric recognition system based on fusion of iris, face, and finger vein traits. Sensors, 20(19):5523, 2020.

[33] Yang Liu and Ajay Kumar. Contactless palmprint identification using deeply learned residual features. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2(2):172–181, 2020.

[34] David Snyder, Guoguo Chen, and Daniel Povey. MUSAN: A Music, Speech, and Noise Corpus, 2015. arXiv:1510.08484v1.

[35] K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Pattern Anal. Mach. Intelligence, 27(5):684–698, 2005.

[36] Maneet Singh, Richa Singh, and Arun Ross. A comprehensive overview of biometric fusion. Information Fusion, 52:187–205, 2019.

[37] Juan Carlos Moreno-Rodriguez, Juan Manuel Ramirez-Cortes, Juan Carlos Atenco-Vazquez, and Rene Arechiga-Martinez. Eeg and voice bimodal biometric authentication scheme with fusion at signal level. In 2021 IEEE

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3100035, IEEE Access

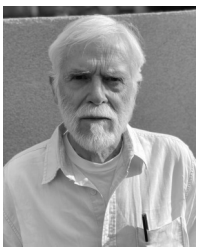
Mexican Humanitarian Technology Conference (MHTC), pages 52–58. IEEE, 2021.

[38] Bacary Goudiaby, Alice Othmani, and Amine Nait-Ali. Eeg biometrics for person verification. In Hidden Biometrics, pages 45–69. Springer, 2020.

JUAN CARLOS MORENO-RODRIGUEZ was born in Puebla, Mexico in 1971. He received the B.Sc. degree in 1995 and his M.Sc. degree in 1998 in Electronic Engineering from Universidad de las Américas-Puebla. He is currently pursuing the Ph.D. degree in electronics at National Institute of Astrophysics, Optics and Electronics, Mexico. He has worked as an Assistant Professor in the departments of Computer Systems and Electronics at Tecnologico Nacional de Mexico and Universidad Iberoamericana, respectively. His research interest includes biometrics, machine learning and signal processing.

JUAN 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. He is currently a Ph.D. student at the Electronics Department, INAOE, in Mexico. His research interests include signal processing, biometric systems, embedding systems, neural networks and applications.

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



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. from University of New Mexico, all in electrical engineering. He is currently an Associate Professor at the Electrical Engineering Department of University of New Mexico. His research interests include digital signal processing applied to speech recognition and thunderstorms.

PILAR GOMEZ-GIL received the B.Sc. degree from the Universidad de las Americas A.C, Mexico, the M.Sc. and Ph.D. degrees from Texas Tech University, USA, all in computer science. She is currently a Titular Researcher in the Computer Science Department at the National Institute of Astrophysics, Optics, and Electronics (INAOE), Mexico. She is 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 of the National Polytechnic School, Ecuador, the M.Sc. and Ph.D. from the National Institute of Astrophysics, Optics, and Electronics, Mexico. He is a professor at the Metropolitan Autonomous University, Iztapalapa at Mexico City. His research interest includes classification and prediction of time series, resource allocation, and artificial intelligence applications.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/