Longitudinal Assessment of EEG Biometrics under Auditory Stimulation: A Deep Learning Approach

Sherif Nagib Abbas Seha
Dept. of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
sherif.seha@mail.utoronto.ca

Dimitrios Hatzinakos
Dept. of Electrical and Computer Engineering
University of Toronto
Toronto, Canada
dimitris@comm.utoronto.ca

***Abstract*—In this paper, we propose a Deep Learning (DL) approach for the longitudinal assessment of EEG signals under auditory stimulation for a biometric authentication system. Longitudinal assessment involves recordings from 13 subjects over three sessions where the average time-span between the last session and the first two is almost a year. The proposed DL approach encodes the EEG data into an embedding space where the distance between cross-session features from the same subjects is minimized and the distance between features from different subjects is maximized. Also, we adopt an encoder with a custom convolution layer that extracts improved functional connectivity features over the standard convolution. The achieved results show improved recognition rates with a significant reduction in the acquisition time compared with other DL frameworks and BCI techniques.**

***Index Terms*—brainwaves (EEG), biometric authentication, deep learning, triplet loss, session-invariant representations**

I. Introduction

One of the most challenging aspects of brainwaves in biometric recognition applications is the high time-variability of EEG signals, specifically, recordings that are conducted on different days. Although many studies in literature showed that biomarkers from EEG are subject-unique, many physiological and non-physiological factors affect the time-permanence of the EEG patterns like brain state, muscle or eye movement artifacts, electrodes re-setting, and power-line interference. To increase the EEG repeatability, previous works indicated that brainwaves under stimulation (visual or auditory) are more consistent where the random ongoing brainwave oscillations are minimized. Conventional approaches to estimate the Evoked Potential (EP) or the Event-Related Potential (ERP) in response to the presented stimulus is to average synced frames over multiple trials, however, this approach requires significantly long recordings.

Faster approaches to estimate the EP/ERP responses involve learning universal or user-specific spatial filters that either maximize the correlation between the epochs and a template or maximize the covariance between epochs under the same task. These techniques were applied successfully for task identification in Brain Computer Interface (BCI) applications [1]–[3] and recently for biometric tasks [4], [5]. However, these techniques focus only on minimizing intra-subject variability while ignoring the inter-subject separability. The last aspect is considered crucial especially for biometric systems. Furthermore, these techniques act only as de-noising filters and require an additional feature extraction technique for proper user identity classification.

In this paper, we propose a Deep Learning (DL) approach that can tackle the aforementioned issues by training an encoder to minimize a Triplet Loss (TL) objective function. When trained on multi-session data, the encoder model learns to extract session-invariant features that improve the repeatability of the evoked potentials in brainwaves. Additionally, we propose a Custom Convolutional (CC) layer in the encoder model that outperforms the standard convolution by extracting superior functional connectivity-based representations. Combining these approaches leads to an improved EEG biometric system compared to previous DL and BCI approaches.

II. Related Work

In the context of EEG biometrics, different frameworks have been proposed to improve the time-permanence of evoked potentials mainly to achieve high recognition rates using short EEG recordings. These approaches can be divided into three categories: 1) averaging multiple trials [6]–[12], user-specific spatial filtering [4], [5], and deep learning [13]–[19] as presented hereafter.

A. Averaging or Fusion of Multiple Trials

Conventional means to improve the SNR of the EP/ERP signal rely mainly on averaging EEG epochs which are synced to the presented stimulus. This approach assumes that the evoked potentials are stationary across trials and mixed with white Gaussian noise which represents the spontaneous EEG [20]. Different representations of EP/ERP signals are then employed as features including temporal-features [6]–[8], spectral-features [12], and auto-regressive modelling [12], [21]. Other approaches extract features directly from the EEG epochs and then apply a score fusion or majority voting scheme to improve the recognition rates [10].

While these frameworks achieved high recognition rates, they required significantly long EEG segments. For instance, in [7], an acquisition time of $+100\ s$ was required to achieve a perfect recognition rate which is impractical for a biometric system.

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada

B. Subject-Unique Spatial Filtering

BCI applications adopted advanced techniques that achieved a fast and reliable detection of EP signals which significantly improved the information transfer rate. These methods apply a set of spatial filters which are estimated using statistical analysis of the EEG epochs. These filters enhance EP detection either by maximizing the correlation between epochs and a grand-averaged template using Canonical Correlation Analysis (CCA) [1], [2] or by maximizing the inter-trial covariance of epochs under the same task/stimulation using Task-related Component Analysis (TRCA) [22].

These approaches were found effective especially when applied to individual users; i.e. learning a subject-unique set of filters [3]. This approach was re-purposed recently for biometric applications and achieved high recognition rates in relatively shorter authentication times ($\approx$ 10-30 s) [4], [5].

C. Deep Learning

Although various DL models have been previously proposed, most of them used the standard Cross-Entropy (CE) loss to learn time-permanent and subject-unique features [13], [15], [17], [18]. An interesting approach was adopted in [23] to learn invariant representations of EEG using Generative Adversarial Network (GAN). The generator/encoder in the GAN model was trained to learn session-invariant representations of a specific attribute (e.g. task, session, subject) by hiding the true label from the discriminator. This approach was evaluated for biometrics to learn session-invariant features which outperformed classical approaches using only 0.5 s EEG epochs [16].

DL approaches improved the repeatability of EP/ERP across sessions, hence, achieved reliable performance with shorter testing times (0.5 - 1 s) [13], [16]. This inspired us to evaluate a different approach that adopts triplet loss to learn session-invariant features. This approach is compared with other session-invariant feature processing techniques like CCA, TRCA, and GAN.

III. Proposed Framework

A. Database

The protocol used for recording the EEG data under auditory stimulation is described in our previous work in [5]. In summary, two auditory stimulations to elicit steady-state Auditory Evoked Potential (AEP) responses are adopted namely; m40 and m80. Each stimulus has four auditory components that are modulated at frequencies that span the lower and the higher gamma bands. EEG data was recorded at a sampling frequency of 12 kHz from 7 channels that mainly capture the fronto-temporal and central brain activity where the AEP dominates [24]. The channels used are Fz, Cz, T3, T4, C3, C4, and Oz. Three different sessions, S1, S2, and S3, were collected from 13 subjects each on a different day (each session lasted for 5 min). The average time-span between S1 and S2, S2 and S3, S1 and S3 is 14 days, 337 days and 351 days, respectively.

B. Pre-processing and Synchronization

Here, we used the same pre-processing and synchronization pipeline from our previous work in [5]. Briefly, the baseline drifts in a 0.5 s window of the EEG signal are estimated using a Savitzky-Golay filter of polynomial order 3, then, the estimated baseline is removed from the EEG signal. Next, IIR notch filters at 60 and 120 Hz are applied to remove power-line interference. After that, high frequencies beyond 120 Hz are filtered out using a Butterworth IIR filter.

For synchronization, the processed EEG signal is divided into 0.5 s segments and synced to the four auditory components in each stimulus to extract 4 synced EEG epochs (more details about the synchronization step is provided in [5]). Finally, the synced epochs are downsampled to 250 Hz to reduce the computational time required for training.

C. Proposed DL Model: The Encoder Structure

The full structure for the encoder model is provided in Figure 1. The details of each block are discussed hereafter.

Input layer: the input to the encoder model is the synced epochs per auditory component as described in Section III-B. The dataset for training is defined as $\{(\mathbf{X}_n^{(c)}, y_n, s_n)\}$, where $\mathbf{X}_n^{(c)} \in \mathbb{R}^{P \times T^c \times 1}$ is the nth EEG epoch that is synced to the auditory component c, y and s denote the subject and session ID, respectively. The input to the encoder is represented as a 3D tensor where P and T^c are the numbers of the EEG channels ($P = 7$) and the temporal samples, respectively. The last dimension of size 1 represents the number of input channels for the encoder.

Custom Convolution (CC) block: the standard 2D convolution layer scans the EEG channels according to their order as provided by the input layer. However, this does not consider the spatial distance between electrodes, hence, fails to capture inter-connections between the nearby EEG channels. Instead, the neighbouring channels in different brain regions are grouped to capture improved spatially-connected patterns in the EEG signals. To achieve this, the input EEG channels are divided into 5 groups connecting different brain lobes; temporal-parietal, frontal-parietal-occipital, frontal-temporal-occipital as shown in Figure 1b. This allows for better characterization of the functional connectivity-based features. Adding this layer showed higher recognition rates over the Standard Convolution (SC) as described in Section III-B.

Convolution block: this block comprises four identical sub-blocks that extract various spatiotemporal patterns from the previous CC block. Each sub-block consists of three basic layers; standard 2D convolutional, RELU activation, and Batch Normalization (BN).

Reduction block: the main function of this block is to reduce the number of the temporal samples and the encoder channel dimensions before the dense layer in the next block. This provides better model generalization for unseen data [25]. This block consists of a 2D max-pooling layer that reduces the the number of the temporal samples by three. After that, a convolution block is added with the half number of filters (kernels).

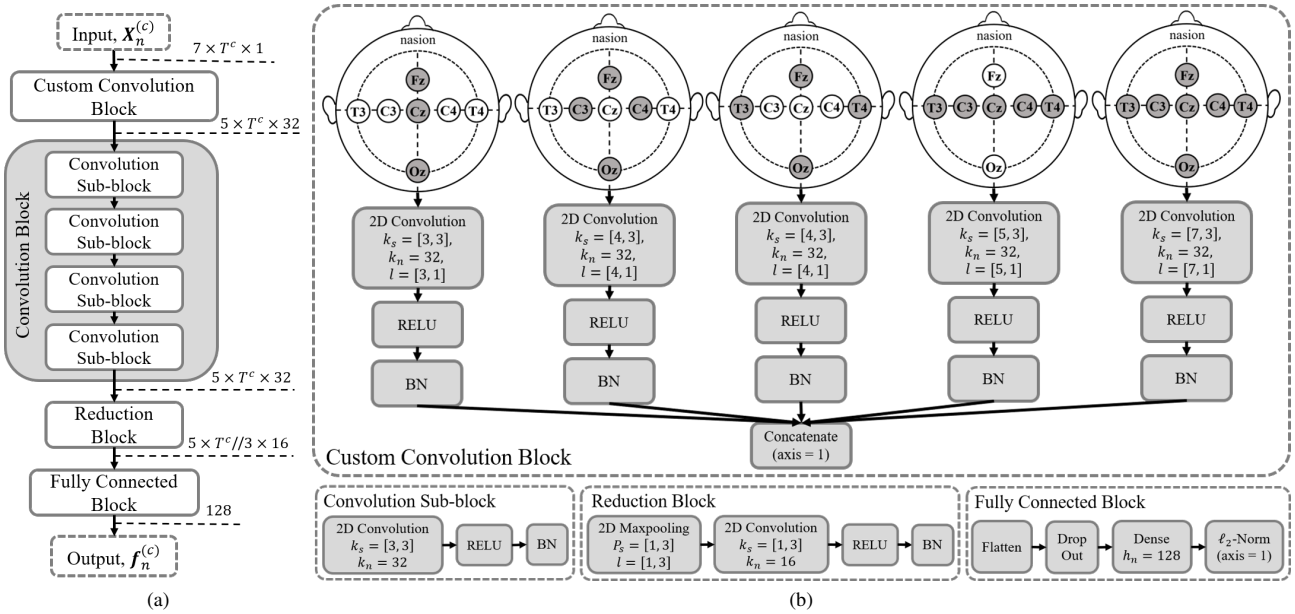


Fig. 1: (a) The full structure of the proposed encoder, (b) Layers of each block (k_s is the kernel size, k_n is the number of kernels, P_s is the pool size, h_n is the embedding vector size, and l is the stride)

Fully connected block: this block acts as a feature extraction stage where the 3D output from the previous blocks is transformed to a 1D vector, also known as *embedding*. This block consists of four basic layers: flatten, dropout, dense layer, and finally an l_2 normalization layer. The size of the embedding vector, h_n, is a crucial hyper-parameter of the encoder as it represents the model capacity to cluster different classes (i.e. user IDs). We tested four values for h_n: 32, 64, 128, 256 and we found empirically that $h_n = 128$ achieves the best accuracy with no significant improvement using 256 vector components.

D. Triplet Loss

Calculation of the TL involves defining three types of examples: Anchor (A), Positive (P), and Negative (N) [26]. The A and P examples are from the same subject across different sessions while the N example is from a different subject. A distance function is computed between the embeddings of A and P, d_P, and another distance function between the A and N embeddings d_N. The TL function is given by the following equation:

$$\mathcal{L} = \sum_{n=1}^{N_b} \left[\left\| \mathbf{f}_A^{(n)} - \mathbf{f}_P^{(n)} \right\|_2^2 - \left\| \mathbf{f}_A^{(n)} - \mathbf{f}_N^{(n)} \right\|_2^2 + \alpha \right]_+ \quad (1)$$

where $d_P = \|\mathbf{f}_A - \mathbf{f}_P\|_2^2$ and $d_N = \|\mathbf{f}_A - \mathbf{f}_N\|_2^2$ using the squared Euclidean distance, α is a margin that separates between d_P and d_N, and n is the example index in the training batch with size N_b. Minimizing the TL function, $\mathcal{L}$, to zero means that the average inter-subject distance of the embedded features, d_N, is greater than the average intra-subject distance of the embedded features, d_P, by a minimum value of α.

IV. Model Training and Results

As mentioned earlier, each stimulus has four auditory components, therefore four models were trained; one for each component. The models were trained to minimize the triplet loss using online mining for hard triplets [27] where two sessions were used for training and the third was used for testing. Since TL convergence takes a significantly long time, the encoder was briefly pre-trained for 4 epochs with CE loss, then, trained for 64 epochs with TL (batch size = 32). A dropout rate of 0.5 was applied during training and the margin α was set to 0.5. Adam optimizer was used for training with an initial learning rate of $1e-2$. During training, the Correct Recognition Rate (CRR) was computed every epoch using k-NN ($k = 1$) to save the model with the best testing CRR. A cloud GPU from Google Colab was used for training with a computational time of 4.5 $s/epoch$, i.e. over 15,600 training examples ($2 \times 13 \times 600$).

After training, a template was created using linear Support Vector Machine (SVM) from the embedded features [25]. In detail, the trained encoders transformed the training synced frames into embeddings. Then, the embeddings from each encoder/frequency component were concatenated together to form one feature vector. Using the final feature vector, a linear SVM model was trained as a template for multi-class classification (for identification under *one-vs-all* setup) and binary classification (for verification). The performance was evaluated using CRR and Equal Error Rate (EER) in identification and verification modes, respectively.

To perform testing, cross-validation using hold-one-session out was conducted. In other words, two sessions were used for model training and SVM template creation and the third was used for testing. This step was performed three times where a different session was selected for testing each

TABLE I: Comparison between our proposed system (CC+TL) and previous DL and BCI frameworks. The mean CRR\EER are reported in each cell in % (white and shaded rows are for the m40 and m80 protocols, respectively)

T_s	Test session = S1				Test session = S2				Test session = S3			
(s)	**CC+TL**	**DC+GAN**	**CCA**	**TRCA**	**CC+TL**	**DC+GAN**	**CCA**	**TRCA**	**CC+TL**	**DC+GAN**	**CCA**	**TRCA**
0.5	**82.8\7.5**	74.7\11.2	52.0\22.1	52.0\21.2	**94.0\2.3**	89.0\5.7	59.7\17.4	59.8\17.4	**85.4\5.3**	67.8\11.8	51.2\20.2	53.8\19.2
	88.8\4.1	79.5\9.0	59.6\17.5	58.8\17.6	**85.3**\5.6	83.1**5.4**	57.3\19.2	55.9\20.2	**91.8\3.4**	80.7\9.5	59.4\18.9	57.0\20.3
1	**85.0\6.6**	77.4\10.8	62.6\17.5	62.7\16.8	**97.4\1.1**	92.2\4.3	74.7\12.5	74.2\12.3	**91.6\3.3**	71.5\10.1	65.6\15.4	67.4\14.6
	91.7\3.0	82.4\8.4	70.9\12.3	71.3\12.4	**87.7**\4.5	86.4**3.8**	67.7\14.7	65.9\16.2	**95.3\2.0**	86.0\8.1	74.8\12.5	69.6\14.6
2	**85.6\6.0**	78.2\10.6	73.2\12.9	72.3\12.9	**99.0\0.4**	94.0\3.5	87.3\7.9	85.5\8.1	**94.8\1.9**	72.3\9.0	77.8\11.3	79.2\10.6
	93.0\2.5	84.2\8.5	79.2\8.0	81.3\8.0	87.6\3.9	**87.6\2.8**	72.6\11.3	70.3\13.0	**97.7\1.2**	88.8\7.7	85.4\8.0	77.5\10.2
3	**85.6\6.0**	78.2\10.7	76.8\11.0	75.8\10.9	**99.4\0.2**	94.7\3.0	91.3\6.3	90.5\6.3	**96.0\1.4**	72.5\8.7	82.8\9.8	82.8\9.6
	93.5\2.3	84.7\8.5	83.6\5.9	85.6\6.2	87.5\3.6	**88.0\2.4**	74.1\10.0	72.2\11.5	**98.5\1.0**	89.8\7.3	89.2\6.4	79.7\8.9
4	**85.5\5.8**	78.3\10.7	80.0\9.7	79.1\9.9	**99.6\0.2**	95.2\2.8	93.9\5.4	93.3\5.4	**96.7\1.2**	72.8\8.5	84.8\8.9	85.8\8.4
	93.7\2.2	85.1\8.5	86.1\5.4	87.8\5.5	87.3\3.5	**88.1\2.3**	73.6\9.0	73.5\10.4	**98.9\0.9**	90.6\7.3	91.7\5.4	80.0\8.3
5	**85.6\5.8**	78.3\10.6	81.3\9.0	80.2\9.1	**99.7\0.1**	95.2\2.6	95.4\4.9	94.1\4.7	**97.1\1.0**	72.8\8.4	86.4\8.6	86.7\8.0
	93.9\2.1	85.2\8.4	86.8\4.7	88.7\5.0	87.2\3.5	**88.1\2.1**	74.2\8.4	73.8\10.0	**99.2\0.8**	91.0\7.2	93.1\4.8	81.1\8.2
10	**85.5\5.6**	78.4\10.7	84.9\7.7	83.0\7.8	**99.9\0.0**	95.7\2.2	98.3\3.9	97.4\3.2	**97.8\0.6**	73.1\8.1	89.0\8.4	89.6\8.0
	94.6\1.8	85.7\8.2	89.5\3.8	90.8\4.7	86.7\3.3	**88.2\1.8**	73.7\8.2	75.0\9.5	**99.8\0.6**	91.6\7.2	94.7\4.5	82.4\7.6

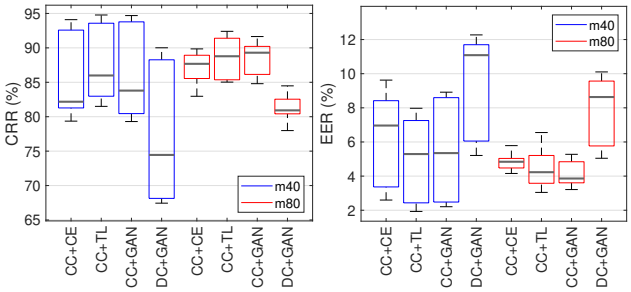
time. Using the trained encoders, the testing synced frames were transformed into embeddings, concatenated together and finally, the class ID was predicted using the SVM template. Additionally, each encoder/SVM model was trained five times with different weight initializations to minimize the effect of random initialization and the mean CRR and EER values were reported.

A. Evaluation of the Proposed Encoder Model with TL

In this section, we evaluate our proposed framework, CC+TL, for session-invariant feature extraction and compare it with previous works including; 1) **DC+GAN:** here, we used the same encoder with Depth-wise Convolution (DC) and adversary training coefficient, $\lambda = 0.01$, as proposed in [16], however, we used the SVM template for testing as described above, 2) **CCA:** We followed the same approach for steady-state AEP estimation as described in [5], however, the template and CCA filter weights were estimated using both training sessions and SVM was used for classification, 3) **TRCA:** Similar to CCA, but the set of subject-unique spatial filters was estimated using TRCA as in [4].

As shown in Table I, DL approaches for learning session-invariant features, either using TL or adversary training (GAN), significantly outperformed BCI techniques, especially at short testing times, e.g. $T_s = 0.5\ s$. Additionally, our proposed model, CC+TL, achieved an improved average cross-session performance over the DC+GAN framework. In terms of performance metrics, our model showed 3.5 - 4.5% lower EER, and 7.5 - 10% higher CRR for m40 and m80 responses, respectively, at $T_s = 0.5\ s$. Also, increasing T_s helps in emerging the response and reducing the unrelated neural oscillations leading to higher performance, however, no significant improvement was achieved above 3 s. Acquisition times $> 0.5\ s$ were performed by averaging multiple embeddings using random sub-sampling without replacement from the test session (more details in [5]).

Figure 2 shows the performance of our encoder model with CC under different training approaches: CE, TL, and GAN (denoted as CC+CE, CC+TL, and CC+GAN in Figure 2).



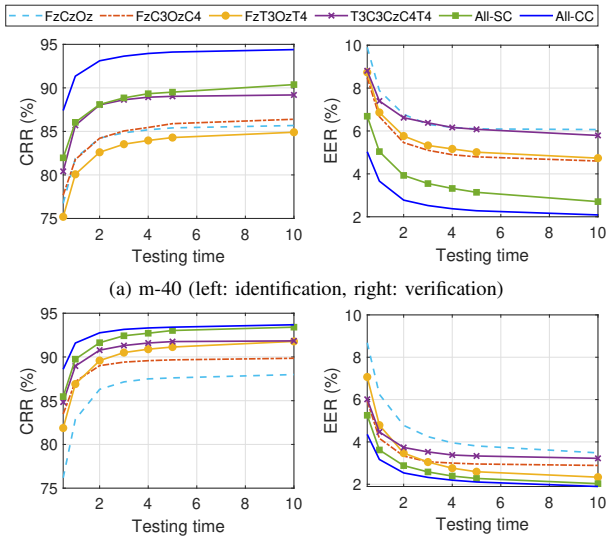
(a) Identification mode (b) Verification mode

Fig. 2: Comparison between different training approaches for our CC encoder model (CE, TL, GAN) at $T_s = 0.5\ s$

The box plot shows the distribution of performance across different testing sessions and random model initializations. Even with adversary training, our encoder model achieved significantly better CRR and EER values compared to the DC+GAN encoder employed in [16], however, both TL and GAN training approaches achieved approximately similar performance using our proposed encoder with custom convolution. Besides, training with TL showed relatively higher performance over CE loss, especially for the m40 response.

B. Evaluation of Reduced Set of EEG Channels

In this section, the performance of the proposed encoder with TL is evaluated under a reduced set of EEG channels as shown in Figure 3. In detail, the selected subsets of channels in the CC block (Figure 1b) are assessed individually as they are passed directly to the convolution block in Figure 1a without custom convolution. As expected, lower performance was achieved using a lower number of channels, however, acceptable performance can be achieved using four or five channels, especially in verification mode. For instance, the subset Fz-T3-Oz-T4 and Fz-C3-Oz-C4 achieved low EER values ($\approx 5\%$) at $T_s = 2\ s$. Additionally, combining these subsets in a custom convolution layer generates better functional connectivity features that outperformed the standard convolution of all the channels (denoted as All-SC in Figure 3). In the All-SC test, a standard 2D convolution (with



(a) m-40 (left: identification, right: verification)

(b) m-80 (left: identification, right: verification)

Fig. 3: Performance of the proposed model using TL under a reduced set of channels

$k_s = [3, 3]$ and valid pad) replaced the CC layer to have similar output.

V. Conclusion

In this paper, we proposed a new deep learning approach that extracts session-invariant features by adopting triplet loss as an objective function. Optimizing TL on multi-session data ensures better session-invariant representations in the feature space by maximizing inter-subject features and minimizing inter-session features for the same subject. Additionally, adopting a custom convolutional layer improves the derivation of functional connectivity-based features by selecting channel subsets that link different brain regions. Using a combination of these configurations, our proposed model outperforms previous deep learning approaches with adversary training and BCI-techniques with subject-unique spatial filtering. Our model showed significantly better performance, in terms of EER and CRR, in relatively shorter authentication time. Although channel subsets were selected empirically in this paper, future work will investigate automatic approaches for channel subset selection using binary neural networks [28].

References

[1] X. Chen, Y. Wang, S. Gao, T.-P. Jung, and X. Gao, "Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface," *Journal of Neural Engineering*, vol. 12, no. 4, p. 046008, 2015.

[2] Z. Lin, C. Zhang, W. Wu, and X. Gao, "Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs," *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2610–2614, 2006.

[3] M. Nakanishi, Y. Wang, Y.-T. Wang, and T.-P. Jung, "A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials," *Plos One*, vol. 10, no. 10, p. e0140703, 2015.

[4] H. Zhao, Y. Wang, Z. Liu, W. Pei, and H. Chen, "Individual identification based on code modulated visual evoked potentials," *IEEE Transactions on Information Forensics and Security*, 2019.

[5] S. N. A. Seha and D. Hatzinakos, "EEG-based human recognition using steady-state AEPs and subject-unique spatial filters," *IEEE Transactions on Information Forensics and Security*, pp. 1–1, 2020.

[6] R. Das, E. Maiorana, and P. Campisi, "EEG biometrics using visual stimuli: A longitudinal study," *IEEE Signal Processing Letters*, vol. 23, no. 3, pp. 341–345, 2016.

[7] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "Permanence of the CEREBRE brain biometric protocol," *Pattern Recognition Letters*, vol. 95, pp. 37–43, 2017.

[8] S.-K. Yeom, H.-I. Suk, and S.-W. Lee, "Person authentication from neural activity of face-specific visual self-representation," *Pattern Recognition*, vol. 46, no. 4, pp. 1159–1169, 2013.

[9] B. C. Armstrong, M. V. Ruiz-Blondet, N. Khalifian, K. J. Kurtz, Z. Jin, and S. Laszlo, "Brainprint: Assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics," *Neurocomputing*, vol. 166, pp. 59–67, 2015.

[10] E. Maiorana and P. Campisi, "Longitudinal evaluation of EEG-based biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 5, pp. 1123–1138, 2017.

[11] M. V. Ruiz-Blondet, Z. Jin, and S. Laszlo, "CEREBRE: A novel method for very high accuracy event-related potential biometric identification," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 7, pp. 1618–1629, 2016.

[12] D. Vinothkumar, M. G. Kumar, A. Kumar, H. Gupta, M. Saranya, M. Sur, and H. A. Murthy, "Task-independent EEG based subject identification using auditory stimulus," in *Proc. Workshop on Speech, Music and Mind*, vol. 2018, 2018, pp. 26–30.

[13] S. N. A. Seha and D. Hatzinakos, "Human recognition using transient auditory evoked potentials: a preliminary study," *IET Biometrics*, vol. 7, no. 3, pp. 242–250, 2018.

[14] E. Maiorana, "EEG-based biometric verification using siamese CNNs," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 3–11.

[15] R. Das, E. Maiorana, and P. Campisi, "Visually evoked potential for EEG biometrics using convolutional neural network," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 951–955.

[16] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Adversarial deep learning in EEG biometrics," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 710–714, 2019.

[17] T. Yu, C.-S. Wei, K.-J. Chiang, M. Nakanishi, and T.-P. Jung, "EEG-based user authentication using a convolutional neural network," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 1011–1014.

[18] E. Maiorana, "Deep learning for EEG-based biometric recognition," *Neurocomputing*, vol. 410, pp. 374–386, 2020.

[19] M. Wang, H. El-Fiqi, J. Hu, and H. A. Abbass, "Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3259–3272, 2019.

[20] T. W. Picton, *Human auditory evoked potentials*. Plural Publishing, 2010.

[21] E. Maiorana, D. La Rocca, and P. Campisi, "On the permanence of EEG signals for biometric recognition," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 1, pp. 163–175, 2015.

[22] M. Nakanishi, Y. Wang, X. Chen, Y.-T. Wang, X. Gao, and T.-P. Jung, "Enhancing detection of ssveps for a high-speed brain speller using task-related component analysis," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 1, pp. 104–112, 2017.

[23] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, "Learning invariant representations from EEG via adversarial inference," *IEEE Access*, vol. 8, pp. 27 074–27 085, 2020.

[24] K. Saupe, E. Schröger, S. K. Andersen, and M. M. Müller, "Neural mechanisms of intermodal sustained selective attention with concurrently presented auditory and visual stimuli," *Frontiers in Human Neuroscience*, vol. 3, p. 58, 2009.

[25] J. S. Kang, Y. Lawryshyn, and D. Hatzinakos, "Neural network architecture and transient evoked otoacoustic emission (TEOAE) biometrics for identification and verification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1858–1867, 2020.

[26] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[28] H. Qin, R. Gong, X. Liu, X. Bai, J. Song, and N. Sebe, "Binary neural networks: A survey," *Pattern Recognition*, p. 107281, 2020.