Journal of Neuroscience Methods 311 (2019) 338–350

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

Application of identity vectors for EEG classification

Christian Ward*, Iyad Obeid

Department of Electrical Engineering, Temple University, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
Electroencephalogram
I-Vector
Universal Background Model
Unsupervised Learning
Subject Verification

ABSTRACT

Background: Finding an optimal EEG subject verification algorithm is a long standing goal within the EEG community. For every advancement made, another feature set, classifier, or dataset is often introduced; tracking improvements in classification without a consistent benchmark, such as a classifier-feature pairing tested on a publicly available dataset, makes it difficult to understand how and why these improvements occur.
New Method: Following on previous biometric experiments, I-Vectors and Gaussian Mixture Model-Universal Background Models are compared to an established Mahalanobis classifier. A second experiment then addresses the impact of epoch duration as a function of classification performance across all three classifiers.
Results: The experimental classification results indicate that I-Vectors are more robust than the other classifiers displaying less sensitivity to epoch duration, data composition, and feature selection.
Comparison with Existing Methods: This I-Vector based approach is compared against commonly used EEG classifiers, such as Mahalanobis and Gaussian mixture models. These classifiers are benchmarked using the publicly available PhysioNet database converted into three feature sets, spectral coherence, power spectral density, and cepstral coefficients.
Conclusions: The experimental results suggests I-Vectors provide reliable baseline performance by leveling the field between feature set and datasets making them well suited for EEG signal processing tasks.

1. Introduction

Signal processing techniques for electroencephalogram (EEGs) have typically been focused on applications such as seizure detection, brain-computer interface (BCIs), and biometrics. Specifically, investigators have sought techniques for both data optimization and modeling that are optimized for EEG signals. By tuning data parameters such as channel count, feature type, and recording duration, investigators have successfully classified data according to subject (Delpozo-Banos et al., 2015) and waveform (Ruiz-blondet et al., 2016). As the feature space of EEG continues to be better understood, more nuanced classification techniques will become feasible to implement.

One such technique, originally developed in the speech processing community, is Identity Vectors (I-Vectors). They were developed to advance speaker verification by building on the established joint factor analysis (JFA) technique (Kenny et al., 2008). JFA separated speech into subject factors, channel factors, and noise factors (Kenny, 2005). Reviews of the process showed that “channel factors estimated using JFA... also contain information about speakers,” which spurred the development of I-Vectors, as detailed in Section 2.1 (Dehak et al., 2011a,b). I-Vectors showed promise in subject verification (Garcia-Romero and Espy-Wilson, 2011; Kenny et al., 2013) as well as accent (Behravan et al., 2016), age (Bahari et al., 2012), language (Dehak et al., 2011a,b), and gender (Senoussaoui et al., 2011) classification tasks.

The work presented here is consistent with the goals of the EEG processing community in that it supports techniques for subject classification (Marcel et al., 2007; La Rocca et al., 2014; Delpozo-Banos et al., 2015; Fraschini et al., 2015), BCI applications (Schröder et al., 2005; Kindermans et al., 2014), session variability (Brigham and Kumar, 2010; Yang et al., 2016; Marcano et al., 2018), and noise classification (Lawhern et al., 2016). The most logical applications of I-Vectors are those focused on biometrics (La Rocca et al., 2014; Fraschini et al., 2015; Yang et al., 2016) because they align with I-Vectors 's original intent of speaker verification, defined here as one-to-one matching of an unknown recording to a known recording. The aim of this work is to develop I-Vectors for EEGs as a subject verification tool in order to support the development of nuanced classification tasks related to BCI (Ahn and Jun, 2015), seizure detection and prediction (Ramgopal et al., 2014) and, eventually, cohort retrieval across EEG databases (Ward and Obeid, 2016).

The hypothesis under test is that I-Vectors are well-suited to EEGs

* Corresponding author at: 1947 N 12th St, Philadelphia, PA 19122, USA.
E-mail address: christian.ward@temple.edu (C. Ward).

https://doi.org/10.1016/j.jneumeth.2018.09.015
Received 12 June 2018; Received in revised form 21 August 2018; Accepted 6 September 2018
Available online 19 September 2018
0165-0270/

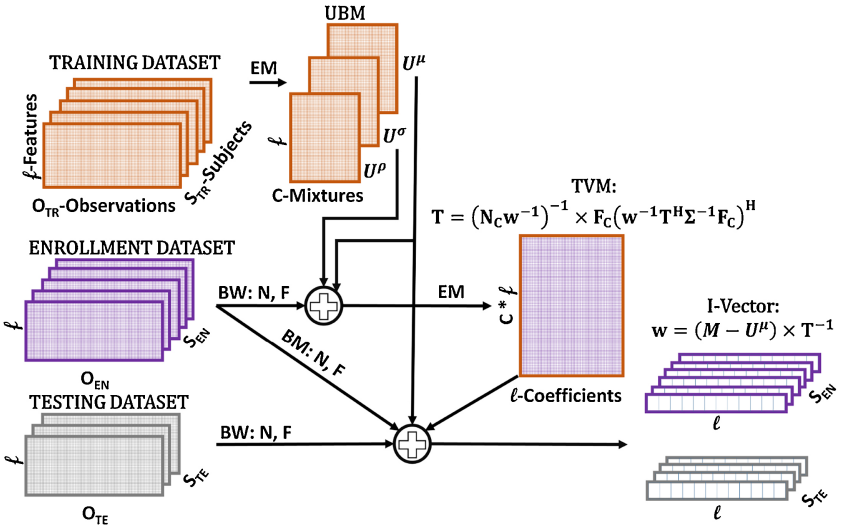

Fig. 1. The process for producing Identity Vectors (I-Vectors) and their associated total variability matrixs (TVMs) relies on calculating the Baum-Welch (BW) statistics of the datasets combining them with the Universal Background Model (UBM) and expectation maximization (EM) to model the characteristics of the data space from the Universal Background Model (UBM). Thus Identity Vectors (I-Vectors) only need the enrollment or testing data as the total variability matrix (TVM) transformational map of the modeled data space.

considering the strength of I-Vectors in identifying statistical patterns in speech data and the similarities between speech and EEG signals. For example, both signal types are built hierarchically: phonemes → words → sentences in speech (Davis and Mermelstein, 1980; Lee et al., 1992), and waveforms → events → statuses in EEG (Tatum and Tatum, 2014). However speech signals are typically single-channel whereas EEGs are mostly multi-channel. Furthermore, the number of EEG electrodes and their locations may vary substantially as dictated by a specific patient's medical condition or the needs of a particular research protocol.

One of the most commonly used datasets in EEG classification research is the PhysioNet EEG Motor Movement/Imagery (PhysioNet) dataset (see Section 2.5) (Goldberger et al., 2000). This dataset is freely available online and contains 64-channel data from 109 subjects, each performing 14 trials of either actual movements, imagined movements, or resting. The database has been extensively used for evaluating subject verification techniques (Su and Robbins, 2013; La Rocca et al., 2014; Rodrigues et al., 2016; Fraschini et al., 2015; Yang et al., 2016). Specifically, La Rocca et al's work (La Rocca et al., 2014), in which excellent subject classification was shown for the resting state trials, has been drawn on as a basis for other experiments such as (Fraschini et al., 2015; Yang et al., 2016).

La Rocca et al. evaluated the use of power spectral density (PSD) and spectral coherence (COH) features for channel-based subject classification. They also introduced a methodology in which optimal "channel sets" were constructed by testing individual channels and then adding them iteratively to the set, but only if they increased classification performance. Using these channel sets they achieved 100% Correct Recognition Rate (CRR) with COH features classified by their Mahalanobis distances. This confirmed their hypothesis that COH features could improve CRR over the more commonly used PSD features for biometric applications (Lotte et al., 2007).

This same process has been adapted for the current work in order to test two classifiers, Gaussian Mixture Model Universal Background Models (GMM-UBMs) (Ward et al., 2016) and I-Vectors, as well as a new feature type, cepstral coefficients (CEP) (Harati et al., 2016). These features are a variant of the dominant speech feature set Mel-frequency CEP (Dehak et al., 2007). The efficacy of the two algorithms and new feature set are evaluated by using La Rocca's results as a benchmark. Following on from this, we extend La Rocca's work in two significant ways.

First, La Rocca's methods only used PhysioNet's resting trials while this work incorporates all the trials, mixing motion and resting trial data. As GMM-UBM and I-Vector techniques are known to work well on large and varied speech datasets (Senoussaoui et al., 2010), it is hypothesized that their performance will improve as epoch size decreases providing more data samples. In a one minute recording with 10 s epochs there are 10 samples, but the same recording produces 60 samples with 1 s epochs. Conversely, the Mahalanobis distance performance will likely struggle given its reliance on pooled covariance matrices (see Section 2.5).

Secondly, La Rocca used a static epoch duration of 10 s to partition the channel data into testing and training sets. However, it is known that the optimal epoch size may vary significantly according to specific combinations of subject identification technique and data characteristics (Ruiz-blondet et al., 2016; Rodrigues et al., 2016; La Rocca et al., 2012). We therefore tested the classifiers over a range of epoch durations, with the goal of identifying classifiers that were relatively insensitive to epoch duration.

2. Methods

This section presents an introduction to I-Vectors as well as the classification techniques of cosine similarity, GMM-UBM, and Mahalanobis distance. As with La Rocca, all feature sets used in this work were drawn from the PhysioNet dataset. In addition to the PSD and COH feature sets that were evaluated by La Rocca, a third feature, CEP, was also tested. Finally, we introduce two protocols; Experiment 1 follows La Rocca's original methodology whereas Experiment 2 varies epoch duration and introduces the multiple trial datasets.

2.1. Identity vectors

I-Vectors are statistically robust unsupervised representations of data in a multi-variate space. They evolved from joint factor analysis to address subject verification for speech processing (Kenny et al., 2008; Dehak et al., 2011a,b). By modeling the f dimensional feature space with a Universal Background Model (UBM) it is possible to constrain and control the dimensionality of the resultant I-Vectors via a total variability matrix (TVM). This is achieved without the need for labeled data or pre-processing to improve data quality. Although an overview of the I-Vector mathematics is presented here, the reader is directed to the the work of Kenny et al. (2008) for a more rigorous treatment.

Generally speaking, I-Vectors are created in a two-step process, Fig. 1. First, *training data* is used to create a UBM that captures the statistical properties common to the data space. Secondly, *enrollment data* is used to create I-Vectors by using a TVM to isolate statistical deviations from the UBM that are common to the enrollment dataset. Examples of enrollments may include subjects that share a medical condition or that perform a common motor task. The efficacy of I-

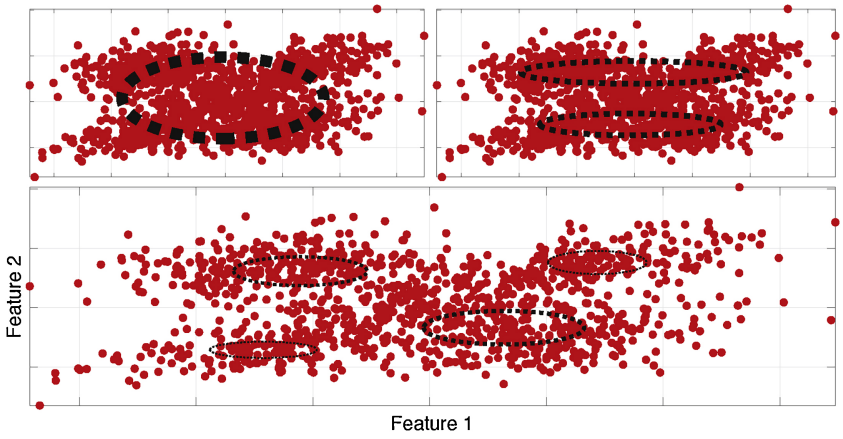

Fig. 2. A visual example of how Universal Background Models (UBMs) adapt based upon the desired mixture size. In each plot the base mixture, red dots, is identical but the number of mixtures in the Universal Background Model (UBM) increases from 1 (top left) to 2 (top right) to 4 (bottom). The relative mixture weights are denoted by the thickness of the dotted black lines.

Vectors for classification is evaluated by using the TVM to produce I-Vectors from a set of *testing data*. The enrollment and test I-Vectors can be compared using common metrics such as cosine distance or Mahalanobis distance.

2.1.1. Universal Background Models

The first step in creating I-Vectors is to generate a UBM. This process relies on expectation maximization (EM) to build Gaussian Mixture Models (GMMs) that capture the overall statistical properties from the training data (Reynolds, 2009). Within each mixture, an independent normal distribution, consisting of a mean μ and variance σ, is resolved for each of the f features. Thus the training data is mapped into C three-dimensional mixtures (including a weight ρ for each mixture) to collectively form the UBM:

$$\boldsymbol{U} = \{\boldsymbol{U}^{\mu}, \boldsymbol{U}^{\sigma}, \boldsymbol{U}^{\rho}\}_{c=1:C} \tag{1}$$

This parametric modeling is limited by the number of features, but the number of mixtures can be varied to optimize performance. As an example, Fig. 2 shows toy data with two features and their corresponding UBMs with 1, 2, and 4 clusters, respectively. In this example, while the number of mixtures can be increased beyond 4, it would be counterproductive to do so since the underlying data only contains four distributions; additional mixtures would split existing ones and result in overlapping distributions with decreasing weights. However in real-world applications, the number of underlying distributions is not known *a priori*, meaning that a range of mixtures sizes should be considered. Mixture size testing typically follows a base 2 progression: $1 \rightarrow 2 \rightarrow 4 \rightarrow 8$.

2.1.2. Total variability matrix

Once the UBM $\boldsymbol{U}$ has been created, a TVM $\boldsymbol{T}$ must be produced from enrollment data. This matrix is used to transform enrollment and test data into I-Vectors. In both cases, the principal equation is

$$\boldsymbol{M} = \boldsymbol{U}^{\mu} + \boldsymbol{T}\boldsymbol{w} \tag{2}$$

where $\boldsymbol{M}$ is a supervector of Baum-Welch (BW) statistics built from a subset $\boldsymbol{e}_{n\ldots m}$ of the dataset's E epochs; $\boldsymbol{U}^{\mu}$ are all feature means in the UBM; $\boldsymbol{T}$ is the TVM; and $\boldsymbol{w}$ is the I-Vector itself. The BW statistics are estimations of the mean and variance first developed to produce Hidden Markov Models (HMMs) (Rabiner, 1989).

The TVM $\boldsymbol{T}$ represents the variances of the C mixtures in the UBM and is of size $[f \times C, l]$, where l is a dimensional constraint to I-Vector length. The value of l must be less than the number of unique subjects in the enrollment or testing datasets to avoid overfitting. This value sets the length of the I-Vectors contributing to dimensionality reduction when $l < f*(n-m)$.

Similar to the UBM, EM is used to generate a TVM by iteratively building from a previously 'solved' $\boldsymbol{T}$. This estimation process relies on the weighted mean $\boldsymbol{A}$ and variances $\boldsymbol{K}$. Each mixture c has a matrix in $\boldsymbol{A}$ of size $[f, l]$ and in $\boldsymbol{K}$ of size $[l, f]$.

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{A}_1^{-1}*\boldsymbol{K}_1 \\ \vdots \\ \boldsymbol{A}_C^{-1}*\boldsymbol{K}_C \end{bmatrix} \tag{3}$$

$$\boldsymbol{A}_c = \sum_{t=1}^{E} \hat{\boldsymbol{N}}_c(t)\boldsymbol{w}^{-1}(t) \tag{4}$$

$$\boldsymbol{K}_c = \sum_{t=1}^{E} \hat{\boldsymbol{F}}_c(t)^*(\boldsymbol{w}^{-1}(t)^*\boldsymbol{T}^{H}*\boldsymbol{\Sigma}^{-1}*\hat{\boldsymbol{F}}_c(t))^{H} \tag{5}$$

However, $\boldsymbol{A}$ and $\boldsymbol{K}$ are dependent on $\boldsymbol{w}$, $\boldsymbol{T}$, $\boldsymbol{\Sigma}$ and the BW statistics $\hat{\boldsymbol{N}}$, $\hat{\boldsymbol{F}}$ (where superscript H denotes the Hermitian transpose and $\hat{}$ indicates an aggregate statistics matrix). This requires initializing a $\boldsymbol{w}$ with a normal distribution Norm(0, 1) of length l. A randomized weighting of $\boldsymbol{U}^{\sigma}$ seeds an initial $\boldsymbol{T}$ where $\boldsymbol{U}^{\sigma}$ is vectorized such that the variances are grouped by mixture $(\sigma_1^1 \ldots \sigma_f^1 \ldots \sigma_f^C)$. The $\boldsymbol{\Sigma}$ matrix is a $[C \cdot f \times C \cdot f]$ diagonalization of the $\boldsymbol{U}^{\sigma}$ values.

$$\boldsymbol{w}_{\text{init}} = \text{Norm}[(0),\ (1)]_l \tag{6}$$

$$\boldsymbol{T}_{\text{init}} = \text{Norm}\left[\binom{0}{0}, \binom{1}{1}\right]_{l,C\times f} * \overrightarrow{\boldsymbol{U}^{\sigma}_{C\times f}} \tag{7}$$

The updates of these terms are driven by $\boldsymbol{w}$ and $\boldsymbol{T}$ along with the BW statistics, which are estimations of the means ($\boldsymbol{N}$) and variances ($\boldsymbol{F}$) of the enrollment data epochs E. The BW 0th ($\boldsymbol{N}$) and 1st ($\boldsymbol{F}$) order statistics are generated from the evaluation of $\boldsymbol{U}$ against the E epochs in the enrollment data. The higher order statistic must be offset by the preceding orders resulting in a centered 1st order statistic $\widetilde{\boldsymbol{F}}$. Each statistic models the f features in each of the C mixtures, resulting in $C \times f$ matrices. Each epoch is evaluated to generate initial probabilities based on $\boldsymbol{U}$ for $\boldsymbol{N}$ and $\boldsymbol{F}$.

$$\hat{\boldsymbol{N}} = \begin{bmatrix} \boldsymbol{N_1} & & \\ & \ddots & \\ & & \boldsymbol{N_C} \end{bmatrix} \tag{8}$$

$$\hat{\boldsymbol{F}} = \begin{bmatrix} \widetilde{\boldsymbol{F}}_1 \\ \vdots \\ \widetilde{\boldsymbol{F}}_C \end{bmatrix} \tag{9}$$

$$\widetilde{\boldsymbol{F}}_c = \sum_{t=1}^{E} P(c|\boldsymbol{e}_t, \boldsymbol{U}_c)(\boldsymbol{e}_t - \boldsymbol{U}_c^{\mu}) \tag{10}$$

$$\boldsymbol{N}_c = \sum_{t=1}^{E} P(c|\boldsymbol{e}_t, \boldsymbol{U}_c) \tag{11}$$

$$\boldsymbol{F}_c = \sum_{t=1}^{E} P(c|\boldsymbol{e}_t, \boldsymbol{U}_c)\boldsymbol{e}_t \tag{12}$$

Iterated estimates for w are then computed according to

$$\boldsymbol{w} = (\boldsymbol{I} + \boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{N}}\boldsymbol{T})^{-1}\boldsymbol{T}^t\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{F}} \tag{13}$$

where $\boldsymbol{I}$ is an $[l \times l]$ identity matrix and the other terms are the most recently updated ones for the given mixture size. The process of generating $\boldsymbol{T}$ completes once the likelihoods converge (Kenny et al., 2005), typically 20 iterations of Eqs. (3)–(13) for EEGs. Eqs. (8)–(12) that generate the initial estimate of $\boldsymbol{w}$ from Eq. (13) underlie the I-Vector extraction process. Using this extraction process, solving for the BW estimates, on the enrollment and testing datasets coupled with the finalized $\boldsymbol{T}$ and $\boldsymbol{\Sigma}$ produces the enrollment and testing I-Vectors.

Typically there is one additional step involving linear discriminant analysis (LDA) to separate the feature space produced by $\boldsymbol{T}$. Using the enrollment data Fischer's LDA is performed to enhance the l dimensions of the I-Vectors. This length control does not need to be the same for the TVM and LDA, but it is kept consistent to constrain the variables in the experiments.

2.2. Gaussian Mixture Model-Universal Background Models

GMM-UBMs were developed prior to I-Vectors and are equally reliant on UBMs through a process known as maximum *a posteriori* (MAP) adaptation (Kenny et al., 2007). Instead of constraining all the mixtures of a UBM to a single matrix, the means and variances of each mixture are combine with the BW statistics of the enrollment data. The mixing between the two statistics is determined by user controlled MAP relevance factors $\alpha_c^{\mu,\sigma,\rho}$ for each dimension of the UBM. These adapted distributions are then scored by comparing the log-likelihood ratios of them against those of the test epochs. A comprehensive guide to this technique is explained in by Marcano et al.'s work (Marcano et al., 2018).

2.3. Data

The PhysioNet dataset contains 109 subjects following computer-prompted motion/motion imagery trials at the New York State Department of Health's Wadsworth Center (Schalk et al., 2004). Each subject from the PhysioNet dataset completed 14 experimental runs: two one-minute calibrations runs of resting Eyes Open (EO) and resting Eyes Closed (EC), and three two-minute runs of four different test trials. Signals were recorded from 64 electrodes using a 10–20 layout sampled at 160 Hz. The data is publicly available through the PhysioNet website (Goldberger et al., 2000).

Numerous feature sets derived from the PhysioNet data have been introduced in the literature to suit specific applications (Fraschini et al., 2015; Rodrigues et al., 2016; Yang et al., 2016). For example, La Rocca et al. used PSD and COH features for their EEG analyses. In this work, we evaluate those two feature sets as well as a cepstrum-based feature set that conforms to the standards of the Temple University Hospital Electroencephalogram Corpus (TUHEEG Corpus) (Obeid and Picone, 2016).

2.3.1. Power spectral density and spectral coherence connectivity

Both of the feature sets used by La Rocca are based on computing Welch's average modified periodogram over a given epoch duration. Their work uses 10 s non-overlapping epochs while the proposed work adds 5, 2, and 1 s epochs. Regardless of initial epoch size, each epoch is filtered by a 1 second Hanning window with a 0.5 s overlap and then resampled from 160 Hz to 100 Hz. This requires 100 points for the Fourier Transform to produce the 1 Hz resolution bins from 1 Hz to 40 Hz that serve as the features.

The original recordings used 64 channels, but La Rocca reduced this to 56 channels. The removal of these 8 electrodes (AF8, AF7, T10, T9, FT8, FT7, Oz, and Iz) assured the set 'constitute[s] a montage common to different available datasets that we can use in future analysis (La Rocca et al., 2014)'. With the epochs being channel independent there are 56 epochs each with 40 features for each set (10, 5, 2, 1) of PSD features. The final features are scaled by a logarithmic transformation, according to the La Rocca protocol (La Rocca et al., 2014).

The COH features represent the cross-spectral density between these same 56 channels. Producing the COH features requires computing the inter-element spectral coherence based on the cross-spectral density, $S_{i,j}(f)$ between signals. Where i and j are the indices of the elements and f is the specific frequency bin,

$$\mathrm{COH}_{i,j}(f) = \frac{\|S_{i,j}(f)\|^2}{S_{i,j}(f)\cdot S_{i,j}(f)} \tag{14}$$

this results in 1540 cross-spectral channels each with 40 features. These final feautres are scaled by Fisher's Z transform to normalize their distributions (La Rocca et al., 2014).

2.3.2. Cepstrum coefficients

Cepstrum coefficients have long been used in speech processing (Davis and Mermelstein, 1980) and are the dominant feature set paired with I-Vector classification (Greenberg et al., 2014). Given their prominence in speech, they were adapted for use with EEGs and a variant has recently shown promise on classification (Harati et al., 2016). Following the methodology of Harati, the first 9 cepstrum features and their associated deltas and delta-deltas, removing the delta-delta for the first cepstrum coefficient, are used to build a 26 element feature vector.

These features originate from 9 cepstrum coefficients drawn from the data as 1 s Hamming windows with 90% overlap producing 100 ms frames. These time signals are Fourier transformed and filtered with 1 s Hamming window to smooth their frequency representation which enables the delta and delta-delta energy calculations to be performed. The delta calculations are the first and second derivatives of changes between the original 9 coefficients between adjacent (prior to current) feature vectors.

Given the amount of data produced by this approach only the central 100 ms epoch is used in these experiments. This ensures there is no data overlap between the epochs when the data is Hamming filtered in the time domain. Each epoch is natively one second in duration, so building larger epochs requires taking the average of the epochs over the desired epoch duration. Finally, the channels are paired down into a differential Trans-Cranial Montage that is commonly used by neurologists, Fig. 3. This results in 22 channels each with 26 features for each epoch.

2.4. Classifiers

Each of the feature sets used a unique classification scheme while operating on identical copies of enrollment testing data based upon La Rocca's proposed 6-fold cross validation (CV) (La Rocca et al., 2014). When operating on datasets containing more than six epochs, 6 non-repeated epochs were chosen at random to produce the testing datasets. The training data supplied to the GMM-UBM and I-Vector classifiers for generating the UBMs were identical to the enrollment data, such that no testing data was used for training or enrollment.

For the *Mahalanobis* classifier, the Mahalanobis distance was measured between each subject's channel based enrollment data with a channel pooled covariance matrix against the testing data's epochs. This was necessary because the number of enrollment epochs used in La Rocca's CV was too small to build a distribution model. The GMM-UBM and I-Vector classifiers only used data from the channel of interest to build the training, enrollment, and testing datasets for Experiment 1, making them unable to pool information across the subjects. For Experiment 2 their training, enrollment, and testing data was pooled across all channels for each trial putting all three classifiers on equal

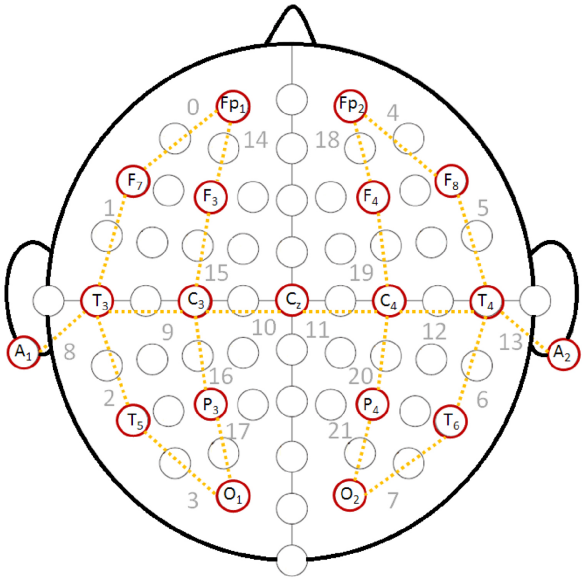

Fig. 3. The Transcranial Montage used by the cepstrum features. The dotted lines represent the channel pairings with the numbers representing the montage channels. Each electrode is labeled as its most common name within the 10/20 configuration profile (Jurcak et al., 2007).

footing and reducing computational complexity.

In the *GMM-UBM* classifier, the log likelihood distance was measured between the MAP-adapted UBMs to the testing data (Reynolds, 2009). Finally, the *I-Vector* classifier was based upon the cosine distance between pairs of enrollment and testing I-Vectors. The elements of each I-Vector represent the weights in their l-dimensional space; comparing the weights of a test I-Vector against an enrollment I-Vector provides a similarity score on the range $[-1, 1]$.

$$\cos(\Theta_{w_1,w_2}) = \frac{w_1^t w_2}{||w_1||*||w_2||} \tag{15}$$

2.5. Performance metrics

Classification efficacy was quantified using two complimentary metrics. Following the methodology of La Rocca, Correct Recognition Rate (CRR) was used to measure the likelihood of the test data correctly matching to the corresponding enrollment model. Despite being a widely adopted metric, CRR does not account for "near-miss" matches, in which a test signal is incorrectly classified by a narrow margin. Such cases may correspond to test signals that incorrectly match to an epoch that is statistically similar to the correct one.

In many applications, such near-miss errors may be considered less problematic than "complete-miss" errors. The degree of error is quantified by the equal error rate (EER). In signal detection theory, the EER is the point at which False Positive and False Negative errors are equal. In general, lower EERs correspond to models with better overall separability. While not frequently cited in EEG biometric experiments (Yang et al., 2016), this classification is critical for those in the broader EEG classification community undertaking similar experiments (Del Pozo-Banos et al., 2014).

$$\text{EER} = \frac{\text{false positive rate} + \text{false negative rate}}{2} \tag{16}$$

$$\text{mEER} = \frac{\text{EER}}{\text{epochs}} \tag{17}$$

2.6. Experimental protocol

I-Vector performance for unsupervised EEG classification was evaluated using two separate experiments. Experiment 1 replicated the methodology of La Rocca to compare the three classifiers (Mahalanobis, GMM-UBM, and I-Vector) applied to three different feature sets (PSD, COH, and CEP). Specifically, the La Rocca methodology performs classification on individual channels and then selectively combines the best performing channels to produce an optimal subject-specific classifier. In Experiment 2, we compared classifier performance using a novel *channel agnostic* approach in which all channel data for a given subject-trial were collapsed into a single *super-channel* of data for each subject-trial. In this experiment, we also probed the effect of epoch duration on classification performance.

2.6.1. Experiment 1

Experiment 1 followed the La Rocca protocol and was carried out using only the resting EC and EO calibration trials. First, a channel-specific CRR was computed from the match matrix $\boldsymbol{S}$, where matches are indicated by 1 and 0 otherwise, for each iteration ψ of the 6-fold cross validation scheme, as detailed by La Rocca(La Rocca et al., 2014). These CRRs were then averaged to produce a resultant mean Correct Recognition Rate (mCRR) that represented the mean CRR over all folds and subjects. Next, these mCRR were combined in a process called *match-score fusion*, producing an optimal subject-specific channel set. The set was built by sequentially adding channels based on their ordered mCRR, but only if they improved the resultant mCRR for the whole set.

$$\text{CRR} = \left(\frac{1}{\Psi}\sum_{\psi=1}^{\Psi} \boldsymbol{S}[\psi, \psi]\right) \times 100 \tag{18}$$

$$\text{mCRR} = \frac{\text{CRR}}{\text{epochs}} \tag{19}$$

A similar averaging calculation was used to compute mean Equal Error Rate (mEER).

Given the relatively small number of training samples, a pooled covariance matrix was used to build the subject-specific models for Mahalanobis distance classification. For training the UBMs, mixture sizes from 2 to 512 (by powers of two) were chosen. The length of the I-Vectors l was set for 100, with a maximum of 50 iterations for the EM to resolve the TVM.

2.6.2. Experiment 2

Experiment 2 focused on the impact of trial differentiation and epoch size, and was limited to PSD features. CEP features were omitted because they performed poorly in Experiment 1; COH features were omitted because their channel set is two orders of magnitude larger (1540 channels v. 56 channels), making it computationally prohibitive. PSD features are commonly used in the EEG machine learning literature and are therefore appropriate for further analysis (Lotte et al., 2007).

Each subject-trial was treated as a distinct subject with the goal of verifying each specific subject-trial. This level of classification merged the channel data to produce a *super-channel* of data for each subject. The channel data was still split into training/enrollment and testing epochs via the prescribed CV of Experiment 1, but with no channel level classification. This was necessary as the individual channel data was insufficient to support the Mahalanobis models, noted in Experiment 1, which could have provided it an advantage. Thus merging the channel data allowed testing of a deeper subject pool based upon the differences of each trial.

The epochs themselves were drawn randomly without replacement for each trial assuring the same epochs in each trial were selected. When the number of epochs exceeded 6 as epoch duration decreased the selection set grew, but only 6 of the epochs were chosen. This built a

unique training/enrollment testing split for each trial to maintain 6-fold cross validation from Experiment 1. With models generated for each subject-trial the classification task becomes harder because (1) the number of subjects is increasing and (2) the contents of those subjects is varied by their tasks and temporal effects between recording trials.

In addition to expanding the subject pool, the impact of epoch duration was explored by producing PSD features for 5 s, 2 s, and 1 s epochs. This increased the number of epochs in the training and enrollment datasets while simultaneously testing how well each classification technique performs with respect to epoch duration. To mitigate any bias introduced by changing the epoch size (and thereby the amount of data) three unique PhysioNet datasets, in terms of content and scale were used. The first set was limited to the four motion trials. The second set used the first four motion trials plus the two resting trials, and the final set included data from all 14 trials. This is in contrast to La Rocca's work, which limited itself to resting state trials only.

As with Experiment 1, results were measured using CRR and EER. By pooling all trial data for a given subject, the reported scores produced a generalized subject-specific mCRR and mEER, independent of the type of trial. The code to run these experiments can be found at the following link: https://github.com/izlandman/jneumeth-2018.

3. Results

The results are presented as a function of performance, CRR and EER. Experiment 1 details the results of replication La Rocca's features and establishing a baseline for their classification protocol. Experiment 2 builds upon Experiment 1 by tracking changes in performance as a function of epoch duration and dataset composition.

3.1. Experiment 1

The results from Experiment 1 are summarized in Figs. 4–6. Specifically, the CRR results for the Mahalanobis, GMM-UBM, and I-Vector classifiers are shown in Fig. 4–6, respectively, while the EER results are shown in Fig. 7. Each of the CRR figures contains two plots. The left plots recreate La Rocca's presentation of the results of match score fusion using mCRR. The right plots show the distribution of the underlying CRRs used to build the mCRR. These distribution plots highlight the mean (central white dot), standard deviation (colored bar), and range (markers) of CRRs for each algorithm and dataset. For clarity the GMM-UBM and I-Vector plots all use a UBM of 32 mixtures as minimal performance gains occured with larger mixtures.

Fig. 4 shows the results of the Mahalanobis classifier. Classification of the COH and PSD features achieved a mCRR of 95% or better by the 6th channel, whereas the cepstrum features reached mCRR of 28.13% for EC and 32.57% for EO. For the GMM-UBM classifier (Fig. 5), the COH and PSD features achieved a mCRR of 90% or better by the 6th channel while the cepstrum based features only reached mCRR of 8.10% for EC and 9.33% for EO. Finally, for the I-Vector classifier (Fig. 6), COH and PSD based features were correctly classified with mCRR of 90% or better by the 6th channel, and the cepstrum based features reached mCRR of 4.13% for EC and 4.13% for EO. Fig. 7 displays the mEER results on the EO data set. The three best mEERs all corresponded to PSD features (6.64% PSD-GMM, 6.86% PSD-IVEC, and 8.60% PSD-MHAL). Note that the CEP features all yielded mEERs values over 40%.

3.2. Experiment 2

Figs. 8–10 summarize the results from Experiment 2. In these plots, the CRR and EER means, and standard deviation are presented as a function of epoch size (bottom axis) and mixture size (top axis). Note that Mahalanobis classifiers do not use mixtures, and therefore performance is independent of mixture size.

Fig. 8 shows classifier performance for datasets built from the first four motion trials. This makes the dataset 8 times larger than the EO and EC from Experiment 1. The best reported CRR of 91.4% occurs using 2s epochs with a GMM-UBM based on 16 mixtures. For 1 s epochs a CRR of 87.46% is achieved using I-Vectors with 256 mixtures. For 5 s epochs a CRR of 89.79% is reached using GMM-UBM with 64 mixtures.

Fig. 9 shows classifier performance for datasets built from the first four motion trials plus EO and EC. The resulting dataset is therefore 10 times larger than the individual EO and EC datasets from Experiment 1. The best reported CRR of 87.92% occurs using 1s epochs with I-Vectors based on 512 mixtures. For 2s epochs a CRR of 84.2% is reached using I-Vectors with 512 mixtures. For 5 s epochs a CRR of 87.05% is reached using GMM-UBM with 64 mixtures.

Fig. 10 shows classifier performance for all data in the PhysioNet database (26 times more data than Experiment 1). The best reported CRR of 90.52% occurs using 1s epochs with I-Vectors based on 512 mixtures. For 2 s epochs a CRR of 80.42% is reached using I-Vectors with 512 mixtures. For 5 s epochs a CRR of 84.05% is reached using GMM-UBM with 64 mixtures.

4. Discussion

This work seeks to use I-Vectors as a means for increasing the robustness and flexibility of subject verification tasks using EEG data. This analysis has required the introduction of an additional performance metric (EER) beyond the CRR, which is commonly used in similar experiments. This discussion first examines the results of Experiment 1 in

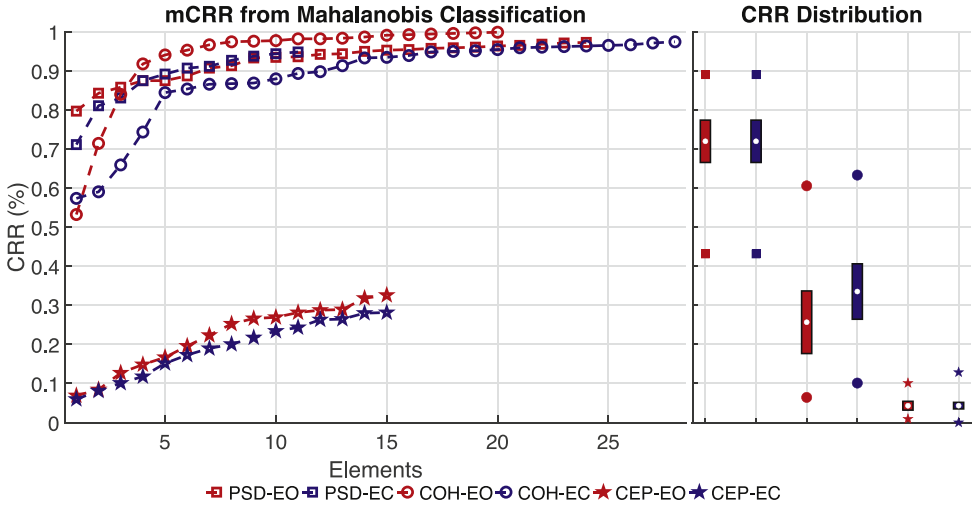

Fig. 4. The results of Experiment 1 provide the match-score fusion mean Correct Recognition Rate (mCRR) (*y*-axis) in the left plot as a function of the number included elements (*x*-axis) for Mahalanobis classification. Elements in these plots refer to channels for power spectral density (PSD) and cepstral coefficients (CEP) and channel pairings for cepstral coefficients (CEP) features. The right plot shows the distribution of the 6-fold cross validation Correct Recognition Rates (CRRs) (*y*-axis) with mean (dot), standard deviation (bar), and full range (marker) for each classifier-feature pairing (*x*-axis). (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

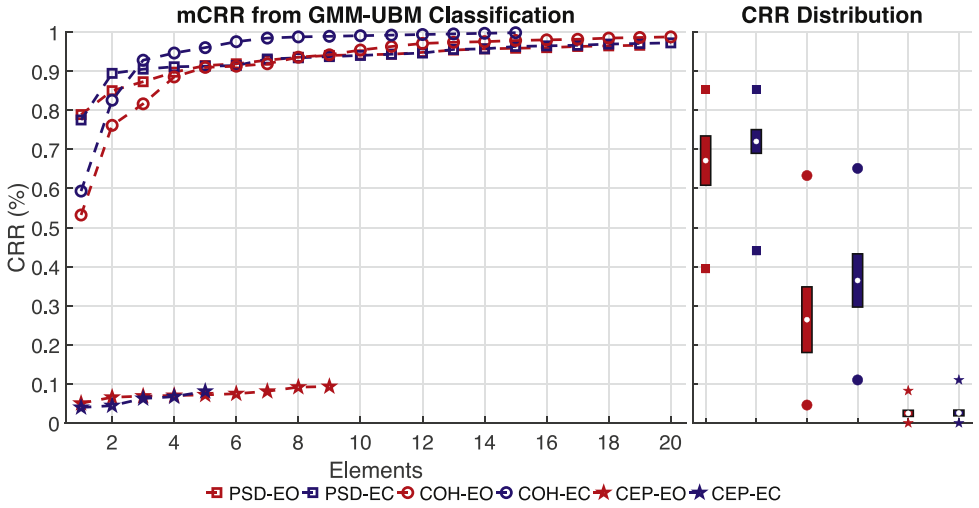

Fig. 5. Experiment 1 Gaussian Mixture Model Universal Background Model (GMM-UBM) classification Correct Recognition Rate (CRR) (*y*-axis) versus element count and feature set (*x*-axis). (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

terms of CRR and EER as a function of classifier-feature pairings. Next, we examine Experiment 2's common feature set, PSD, for each classifier as a function of epoch duration and dataset composition. Lastly, we review the influence of feature sets, datasets, and experimental parameters in constraining the performance of these experiments.

4.1. Experiment 1

Replicating the original La Rocca experiment with an additional feature set (CEP) and evaluation criteria (EER and mEER) helps contextualize the use cases of the features and classification techniques. The match-score fusion based mCRR for the Mahalanobis trials (Fig. 4) were equivalent to those reported by La Rocca et al. (2014). The difficulty in perfectly replicating La Rocca's work could be attributed to using all 109 subjects whereas they used 108, the process of averaging to the mean ear electrodes, or differences when using a 'proper anti-aliasing low-pass filter' to restrict signals to 50 Hz. However, each classifier-feature pair's mEER (Fig. 7) suggests La Rocca's COH features require a considerable performance trade-off to achieve a 5% improvement in mCRR over PSD features for the original Mahalanobis classification.

The range of CRRs (the secondary plots of Figs. 4–6) confirms La Rocca's finding that aggregating channels is a simple method for improving performance relative to the mCRR. Figs. 4–6 indicate that, regardless of trial, it takes at least Mahalanobis 13 elements, GMM-UBM 6 elements, and I-Vectors 11 elements to reach 90% mCRR for COH and PSD feature sets. Despite these successes, the best performance of the CEP features barely reaches 30% mCRR with the Mahalanobis classifier and performs even worse for the other classifiers. However, within these benchmarks there are trade-offs between mCRR and mEER for the COH and PSD features based on the classifier.

The GMM-UBM classifier provides equivalent mCRR performance to the Mahalanobis classifier while improving the mEER of PSD by 1.96% and of COH by 10.70%. The I-Vector classifier provides similar performance to the other two classifiers, but fails to significantly improve either mEERs, with PSD decreasing by 1.74% and COH increasing by 0.57%. This may explain why Mahalanobis distance is not frequently used for EEG classification despite its acceptable classification performance (Lotte et al., 2007; Maiorana et al., 2016). It also suggests that I-Vectors are an acceptable alternative given their mCRR on the dominant feature set, PSD.

Each feature set appears to be clustered according to its mEER. Notably, the mEER of PSD being 15% stronger than other features (Fig. 7) for each classifier suggests that feature selection is the most critical component in deciding EER. Conversely the mCRR present with no such trend, other than the overall poor performance of the CEP

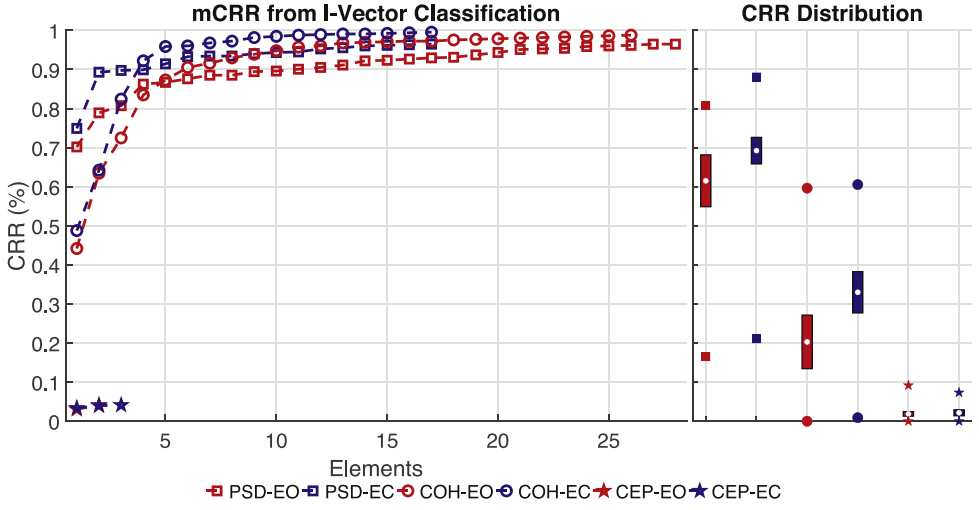

Fig. 6. Experiment 1 Identity Vector (I-Vector) classification Correct Recognition Rate (CRR) (*y*-axis) versus element count and feature set (*x*-axis). (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

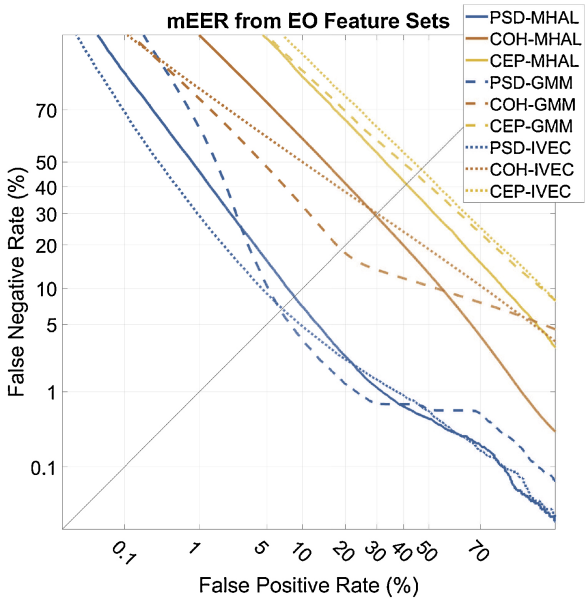

Fig. 7. Experiment 1 operating characteristic (ROC) curves based on the Eyes Open (EO) trial false negatives (*y*-axis) versus false positives (*x*-axis). The reported best in class mean Equal Error Rates (mEERs) range from 6.64%, PSD-IVEC, to 46.97%, CEP-IVEC, indicated by the diagonal. (For interpretation of the references to color in the text, the reader is referred to the web version of this article.)

features and that additional channels improves mCRR. Despite this, classifier choice does appear to marginally improve the mEERs on these small datasets which was not addressed in La Rocca's original work.

Overall, the results suggest that I-Vectors can perform on an equivalent level as the other algorithms regardless of feature set (CEP, COH, and PSD) and data source (EO or EC). This performance does not suggest that the I-Vector technique should replace others, but rather that it warrants additional testing and consideration in future EEG classification tasks such as BCI, seizure, and our eventual goal of cohort retrieval.

4.2. Experiment 2

Expanding the original experiments by incorporating smaller epochs and larger datasets tests the modeling process of each algorithm. The inclusion of multiple distinct motion trials and the resting trials increases the diversity of the testing datasets and producing smaller epochs tests the modeling of each classifier. Typically, EEG biometric experiments maintain consistency of the datasets to mitigate the complexities inherent to the brain (Campisi et al., 2014). The speech community has addressed similar issues through tools like I-Vectors that address variations in the recording channel (landline, mobile phone, microphone, and other speech modifiers) and residual noise to identify the speaker (Dehak et al., 2011a,b; McLaren and Van Leeuwen, 2012).

This is immediately apparent in the performance decrease seen when adding the resting trials, Fig. 9, to the first four motion trials,

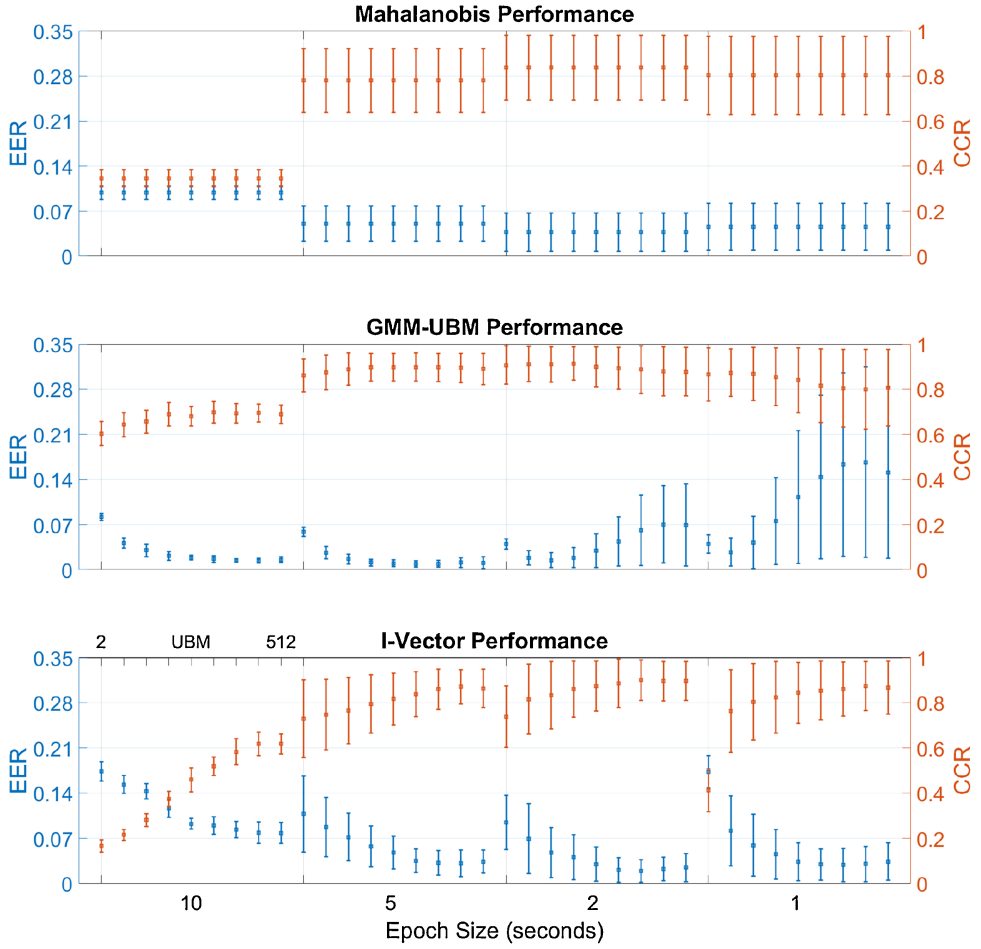

Fig. 8. Experiment 2 classification performance on the first four PhysioNet EEG Motor Movement/Imagery (PhysioNet) motion trials. The equal error rate (EER) (*y*-axis left) and Correct Recognition Rate (CRR) (*y*-axis right) are evaluated as a function of epoch duration (*x*-axis) for each classifier. The Universal Background Model (UBM) mixtures are shown as individual distributions within each epoch duration.

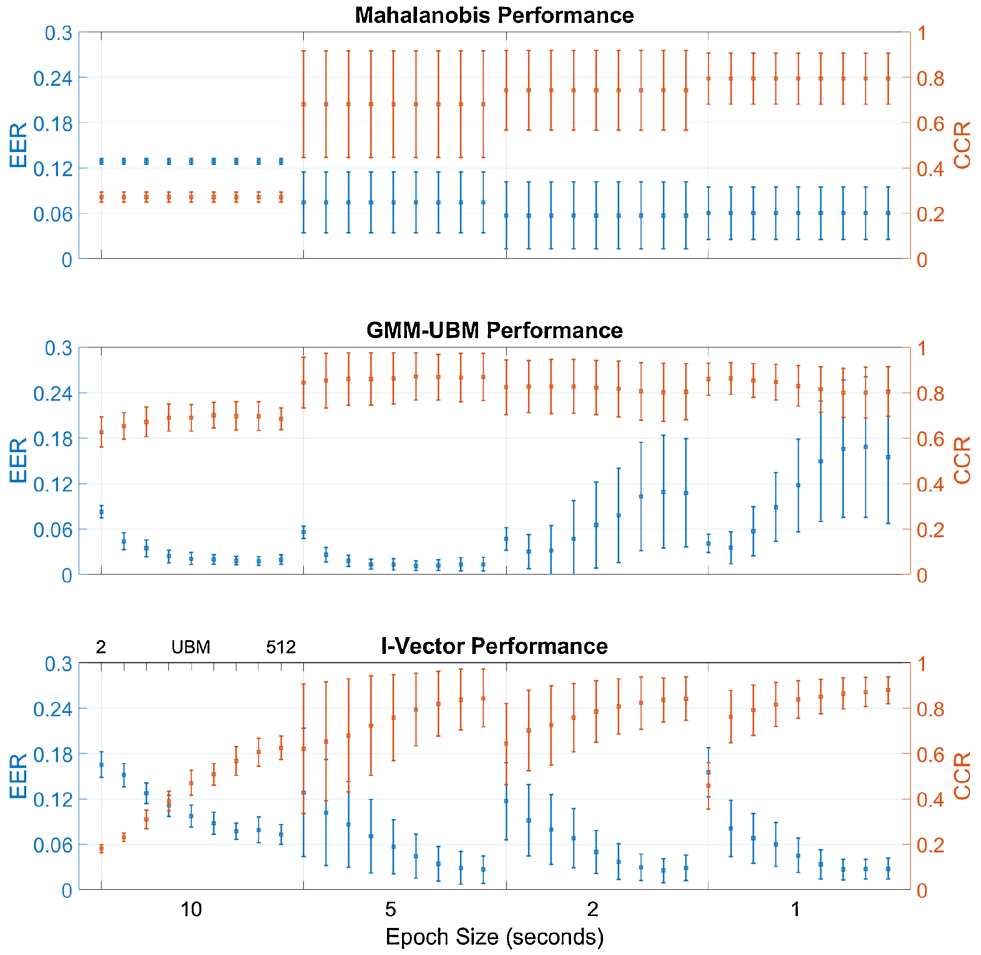

Fig. 9. Experiment 2 classification performance on the first four PhysioNet EEG Motor Movement/Imagery (PhysioNet) motion trials and the resting trials Eyes Open (EO) and Eyes Closed (EC). The equal error rate (EER) (*y*-axis left) and Correct Recognition Rate (CRR) (*y*-axis right) are evaluated as a function of epoch duration (*x*-axis) for each classifier.

Fig. 8. The addition of the resting trials causes a decrease in CRR compared to the pure motion trials for all algorithms. Additionally, the EERs of the Mahalanobis and GMM-UBM classifiers are worse for the larger dataset. Both of these occurrences are likely attributed to the presence of rest periods in the motion trials being classified as from the resting trials. Furthermore, a false positive being more likely than a true positive drives up the EER. This is more likely with the expanded datasets because classification occurs across multiple recordings from the same target subject, but the within trial subject remains the only correct answer.

Across the three datasets, Figs. 8–10 show that the 10 s epoch performance decreases for all classifiers as the amount of data increases. However, smaller epochs provide varying levels of improvement to the Mahalanobis and I-Vector classifiers. As the epoch duration shrinks, more epochs become available in the enrollment data to build the subject-trial models which is the likely driver of the improving performance. This is an interesting result because it does not fully align with the view that longer sessions produce more stable subject verification results, as presented by Maiorana et al. (2016).

Their work, however, limited feature generation to 5 s epochs with 40% overlap, which were evaluated after epoch-based fusion scoring algorithms, similar to the match-score fusion, for a given recording duration (10 s to 90 s). Thus they did not produce a 90 s epoch, but rather evaluate a series of 5 s epochs drawn from fewer subjects (50) using only within-trial data (EO or EC). Their work clearly validates the longitudinal stability of the approach, with days and weeks between recording sessions, but that is beyond the scope of introducing I-Vectors to the EEG community. In contrast, the scope of this work is to introduce and establish I-Vectors as a tool that could remove the requirements of specifically matched datasets, arbitrary epoch lengths between experiments, and channel/epoch based fusion schemes from EEG classification.

It is within these two smallest epoch durations, Figs. 11 and 12, that performance peaks in terms of CRR and EER for both GMM-UBM and I-Vector classificaiton. The I-Vector CRR and EER improve with each subsequent mixture across all epoch sizes which is most evident in Figs. 10 and 11. The reported I-Vector CRR and EER (80.42%, 3.11% and 90.52%, 1.96%) exceed those of the GMM-UBM (80.25%, 5.26% and 90.10%, 3.48%) for 2 and 1 s epochs, respectively. The performance margin is smaller in Figs. 9 and 12, but I-Vectors (84.20%, 2.854% and 87.92%, 2.75%) are again superior to GMM-UBMs (82.7%, 4.67% and 86.24%, 3.52%) for the two smallest epochs. It is only when the resting EO and EC datasets are removed that GMM-UBM outperforms the I-Vectors for the 2 s epochs.

The seemingly varied GMM-UBM EERs for 2 s and 1 s epochs probably occurs because at most there are 512 mixtures of 40 features working to account for numerous subject-trial variations. Adjusting the MAP relevance factor, r, could help correct this behavior by relying more on the enrollment data. The smaller epochs make for deeper training and enrollment datasets, so it is likely the UBM are producing articulated mixtures that the MAP adaptation is unable to generate sufficiently diverse models because the relevance factor α_c is too small. It is difficult to optimize the relevance factor because it applies to the weight of each mixture, in the UBM, produced during BW estimation,

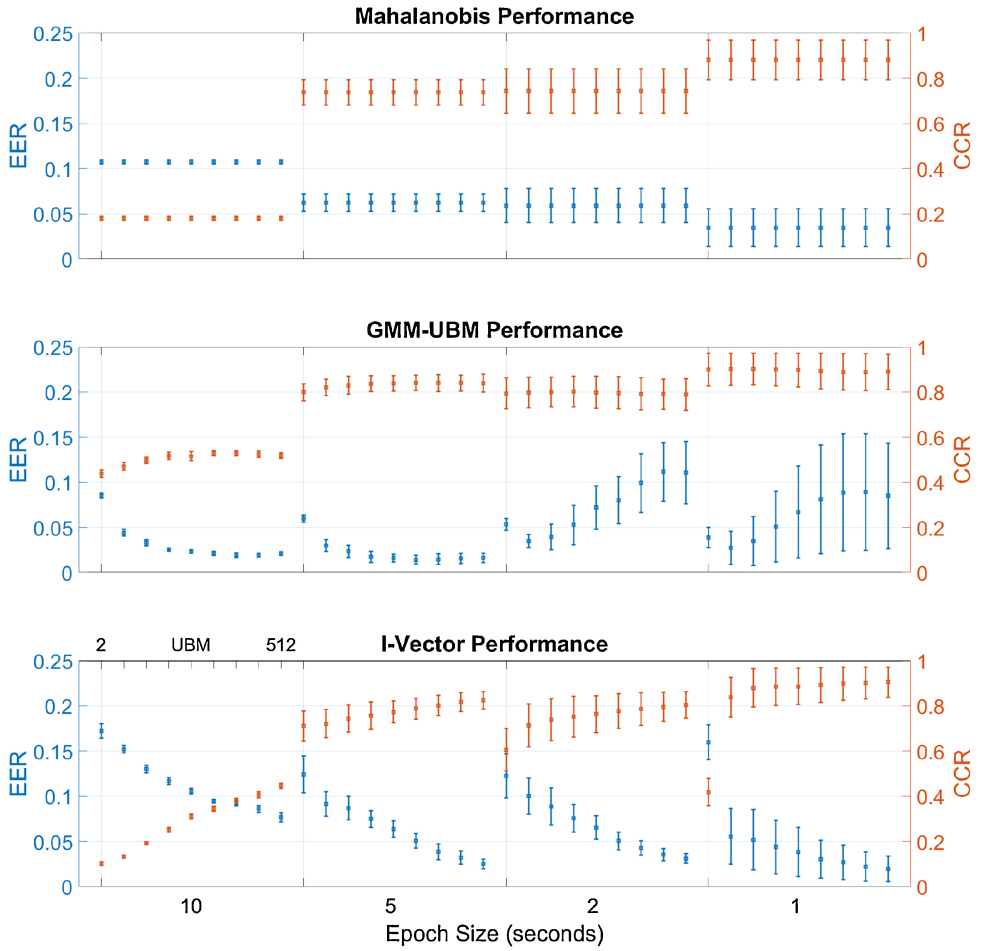

Fig. 10. Experiment 2 classification performance on all 14 PhysioNet EEG Motor Movement/Imagery (PhysioNet) trials, motion and resting. The equal error rate (EER) (*y*-axis left) and Correct Recognition Rate (CRR) (*y*-axis right) are evaluated as a function of epoch duration (*x*-axis) for each classifier.

where $\boldsymbol{N}_c$ comes from Eq. (11).

$$\alpha_c = \frac{\boldsymbol{N}_c}{\boldsymbol{N}_c + r} \tag{20}$$

Furthermore it is could also be attributed to the epoch duration being less than the duration (4 s) of the tasks in the motion trials. This means epochs could straddle rest and motion tasks. As all three classifiers are pooled data across the channel, they could encounter difficulty if an epoch splits the tasks. The GMM-UBM and I-Vector classifiers are built on the sensitivities of the individual means and variances, by adapting the underlying UBMs, which means they may underperform if the UBMs fail to capture all modes of the data. The Mahalanobis distance combines the variances making it less sensitive to the variance of each model and more reliant on the means.

It could be that the combined weighting of the Mahalanobis distance prioritizes different facets of the features than those of the UBMs giving rise to different sets of outliers between the classifiers. Given that outliers are likely to appear when epochs overlap tasks in the recordings, 5 s and 2 s results could be the true measure of the classifiers strength. This may explain why there is minimal improvement, or often a decrease, in classification performance between 5 s and 2 s epochs as the means are not changing, only the variances. Even so, shorter epochs cannot be disregarded as the root cause because, for event classification tasks, short epochs have proven successful (Radha et al., 2014; Wulsin

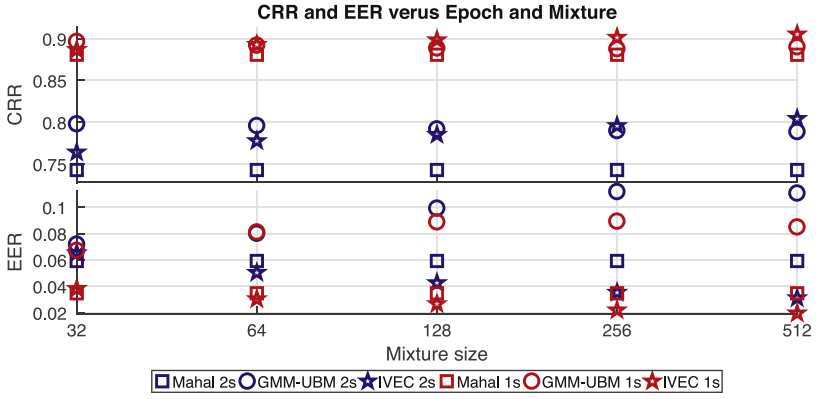

Fig. 11. Mean classifier performance, Correct Recognition Rate (CRR) and equal error rate (EER), as a function of epoch duration and mixture size limited to the five largest mixtures using all 14 trials. The color of the markers represents 2 s epochs (blue) and 1 second epochs (red). The shape of the markers represents the algorithms with Mahalanobis (square), Gaussian Mixture Model Universal Background Model (GMM-UBM) (circle), Identity Vector (I-Vector) (star). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

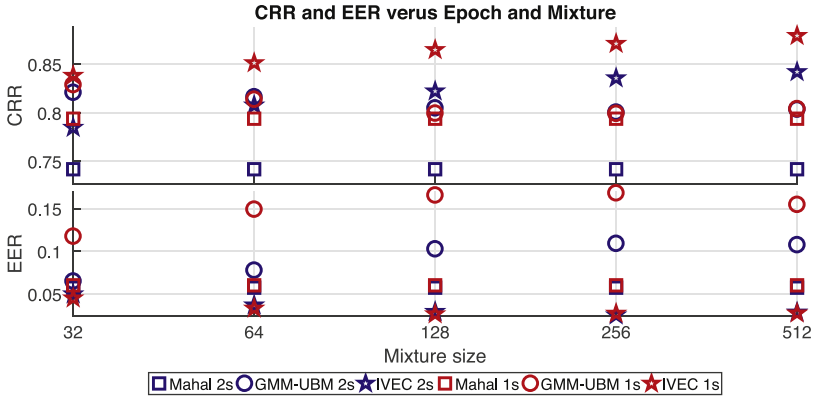

Fig. 12. Mean classifier performance, Correct Recognition Rate (CRR) and equal error rate (EER), as a function of epoch duration and mixture size limited to the five largest mixtures using the 4 motion trials and 2 resting trials. The color of the markers represents 2 second epochs (blue) and 1 second epochs (red). The shape of the markers represents the algorithms with Mahalanobis (square), Gaussian Mixture Model Universal Background Model (GMM-UBM) (circle), Identity Vector (I-Vector) (star). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and Litt, 2011).

There are no such trends with the GMM-UBM classifier, as its EER grows for each mixture size in the 2 s and 1 s epoch sets, which is the opposite of the 10 s and 5 s epochs. Its CRRs shows minimal improvement, and often even a regression, at these larger mixture sizes. This suggests the subject-trial models are generating scores from non-matching subjects that exceed those found when matching subjects. The models may be unable to overcome these edge cases despite increasing the mixtures suggesting the problem is fundamental to the data and not the modeling process. While none of the algorithms avoid performance problems over all epochs and datasets, only the I-Vectors show a continued progression towards improved CRR and EER for all datasets, epochs, and mixture sizes. I-Vectors would therefore be best suited to generalizing to additional data, similar to its adaption from speech to EEGs.

4.3. Constraints

La Rocca's findings showed COH features could improve subject verification via individual channels and through their match-score fusion when compared to PSD features. Reducing the required amount of data to perform subject verification is an important goal for developing biometric-based EEG applications (Rodrigues et al., 2016; Fraschini et al., 2015; Delpozo-Banos et al., 2015), which is often achieved by incorporating novel features, new algorithms, or unique dataset configurations. La Rocca focused on novel features by replacing PSD features for COH features given the influence of each channel's 'dynamic relationship with other regions' of the brain (La Rocca et al., 2014).

Using this substantially larger channel set improved match-score fusion classification performance to 100%, but channel reduction was exchanged for an increase channel search space. Both feature sets were constructed from the same raw data and classified using the same Mahalanobis distance classifier. This suggests that the features were the driving factor in classification improvement. However, by replicating the original work in Experiment 1, GMM-UBMs and I-Vectors provided equal if not better performance compared to Mahalanobis classifiers where COH features were outperformed by PSD features. As the epoch sizes became smaller and the datasets increased in the Experiment 2, only I-Vectors maintained their performance. This suggests the epoch duration and dataset size were the strongest performance factors for PSD features.

Experiment 2 highlighted the impact of epoch duration and dataset size/composition in classification performance. Needing to control for dataset composition, dataset size, and epoch duration increases the difficulty of finding an optimal feature set and classification algorithm pairing (Marcel et al., 2007; Bogaarts et al., 2016). La Rocca's work mitigated this with a static epoch duration, brain regions, and controlled datasets because they understood their dataset and classification goal. Other research agrees that performance can be improved and maintained over time through external knowledge and preparation of the data (Maiorana et al., 2016; Armstrong et al., 2015; Brigham and Kumar, 2010). However, some research accepts the risks and tries to work through them without external influence(Marcel et al., 2007; Yang et al., 2016).

Such conditions are not always present, which makes understanding the impact of epoch duration and dataset composition critical for experimental success. The presence of *a priori* knowledge is only useful if it is properly understood, otherwise it creates an unequal testing ground that impedes the development of robust features and algorithms. When given an environment lacking in this knowledge, I-Vector classification is able to mitigate these shortfalls. This reduces the impact of epoch duration, dataset composition, and algorithm performance so that the focus can be on feature selection and the level of discrimination (subject, trial, channel, etc.) leading to fundamental knowledge gains about EEG behavior (Ward and Obeid, 2016).

The results of Experiment 2 showed that I-Vector EERs decreased and CRRs increased in terms of both mean and variance with each subsequent mixture size. Although the experiment stopped with a UBM of 512 mixtures, the speech community has tested mixtures of size 2048 (Greenberg et al., 2014) and 4096 (Behravan et al., 2015). It is possible that larger mixture sizes could produce stronger classification results for Experiment 1 and Experiment 2. This would bringing their performance into line with the best case results of La Rocca's original experiment. Therefore future experiments should increase the set of mixtures tested, even though it will result in an increase in computational time.

5. Conclusion

The results of this work have indicated that, although the choice of feature set affects subject verification, epoch duration is a stronger predictor of performance. A poorly chosen epoch duration could negatively impact an otherwise acceptable feature set. This can make it difficult to select an appropriate classifier-feature pairing without having deep *a priori* knowledge of the dataset or running extensive experiments to determine the optimal configuration. Such a process would need to be repeated with each new dataset, because EEG subject verification lacks a consistent baseline classification scheme to act as a benchmark.

Despite indications that smaller epochs are equivalent and in some instances better than their counterparts the source data is small and controlled. To say this applies to all EEG circumstances would be premature. Further experimentation using more diverse data is necessary to grasp the influence of epochs on a given classification task. None of the data was longitudinal and the PhysioNet database provides no information to characterize the subjects (age, gender, handedness, etc.). A comprehensive study drawing from detailed subjects is the next step and is planned using the TUHEEG (Obeid and Picone, 2016) which contains annotations and documentation for all subjects.

I-Vectors could be the missing benchmark tool capable of linking

EEG subject identification/verification tests together. Such a tool is necessary because of the multimodal nature of EEG data itself and the amount of variability from subject to subject in larger datasets (Ahn and Jun, 2015). While not exhaustively tested in this work, I-Vectors showed they provided robust subject verification on a common dataset using a common feature set by overcoming uncertainties introduced by epoch duration and dataset composition. This capability should enable experiments to focus on fundamental questions without needing to spend time optimizing low-level data configurations, thereby leading to more insightful experimental results.

Declaration of interest

The authors declare that there are no financial and/or personal relationships with other people or organizations that could inappropriately influence their work.

Acknowledgment

The authors would like to thank the Office of the Assistant Secretary of Defense for Health Affairs which supported them indirectly through the Neurosensory and Rehabilitation Research Award Program under Award No. W81XWH-15-1-0045.

References

Ahn, M., Jun, S.C., 2015. Performance variation in motor imagery brain–computer interface: a brief review. J. Neurosci. Methods 243, 103–110.
Armstrong, B.C., Ruiz-Blondet, M.V., Khalifian, N., Kurtz, K.J., Jin, Z., Laszlo, S., 2015. Brainprint: assessing the uniqueness, collectability, and permanence of a novel method for ERP biometrics. Neurocomputing 166, 59–67.
Bahari, M.H., McLaren, M., Van Hamme, H., Van Leeuwen, D.A., 2012. Speaker age estimation using i-vectors. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 1, September 506–509.
Behravan, H., Hautamäki, V., Kinnunen, T., 2015. Factors affecting i-vector based foreign accent recognition: a case study in spoken Finnish. Speech Commun. 66 (February), 118–129.
Behravan, H., Hautamäki, V., Siniscalchi, S.M., Kinnunen, T., Lee, C.H., 2016. I-Vector modeling of speech attributes for automatic foreign accent recognition. IEEE/ACM Trans. Speech Lang. Process. 24 (1), 29–41.
Bogaarts, J.G., Gommer, E.D., Hilkman, D.M.W., van Kranen-Mastenbroek, V.H.J.M., Reulen, J.P.H., 2016. Optimal training dataset composition for SVM-based, age-independent, automated epileptic seizure detection. Med. Biol. Eng. Comput. 54 (August (8)), 1285–1293.
Brigham, K., Kumar, B.V.K.V., 2010. Subject identification from electroencephalogram (EEG) signals during imagined speech. 2010 Fourth IEEE Int. Conf. Biometrics Theory, Appl. Syst. IEEE, September 1–8.
Campisi, P., La Rocca, D., Rocca, D.L., 2014. Brain waves for automatic biometric-based user recognition. IEEE Trans. Inf. Forensics Secur. 9 (5), 782–800.
Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. 28 (August (4)), 357–366.
Dehak, N., Dumouchel, P., Kenny, P., 2007. Modeling prosodic features with joint factor analysis for speaker verification. IEEE Trans. Audio. Speech Lang. Process. 15 (7), 2095–2103.
Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011a. Front-end factor analysis for speaker verification. IEEE Trans. Audio. Speech. Lang. Process. 19 (may (4)), 788–798.
Dehak, N., Torres-Carrasquillo, P.A., Reynolds, D.A., Dehak, R., 2011b. Language recognition via i-vectors and dimensionality reduction. INTERSPEECH, no. August 857–860.
Del Pozo-Banos, M., Alonso, J.B., Ticay-Rivas, J.R., Travieso, C.M., 2014. Electroencephalogram subject identification: a review. Expert Syst. Appl. 41 (15), 6537–6554.
Del Pozo-Banos, M., Travieso, C.M., Weidemann, C.T., Alonso, J.B., 2015. EEG biometric identification: a thorough exploration of the time-frequency domain. J. Neural Eng. 12 (5).
Fraschini, M., Hillebrand, A., Demuru, M., Didaci, L., Marcialis, G.L., 2015. An EEG-based biometric system using eigenvector centrality in resting state brain networks. IEEE Signal Process. Lett. 22 (6), 666–670.
Garcia-Romero, D., Espy-Wilson, C.Y., 2011. Analysis of I-vector length normalization in speaker recognition systems. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 249–252.
Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E., 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101 (June (23)), E215–20.
Greenberg, C.S.C., Bansé, D., Doddington, G.R., Garcia-Romero, D., Godfrey, J.J., Kinnunen, T., Martin, A.F., McCree, A., Przybocki, M., Reynolds, D.A., 2014. The NIST 2014 speaker recognition i-Vector machine learning challenge. Proc. Speak. Lang. Recognit. Work., no. June 224–230.
Harati, A., Golmohammadi, M., Lopez, S., Obeid, I., Picone, J., 2016. Improved EEG event classification using differential energy, in 2015 IEEE Signal Process. Med. Biol. Symp. – Proc., no. December 2015. IEEE, December 1–4.
Jurcak, V., Tsuzuki, D., Dan, I., 2007. 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. Neuroimage 34 (February (4)), 1600–1611.
Kenny, P., Boulianne, G., Dumouchel, P., 2005. Eigenvoice modeling with sparse training data. IEEE Trans. Speech Audio. Process. 13 (May (3)), 345–354.
Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007. Speaker and session variability in GMM-based speaker verification. IEEE Trans. Audio. Speech Lang. Process. 15 (May (4)), 1448–1460.
Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of interspeaker variability in speaker verification. IEEE Trans. Audio. Speech. Lang. Process. 16 (July (5)), 980–988.
Kenny, P., Stafylakis, T., Ouellet, P., Alam, J., Dumouchel, P., 2013. PLDA for speaker verification with utterances of arbitrary duration. 2013 IEEE Int. Conf. Acoust. Speech Signal Process., no. 1. IEEE, May 7649–7653.
Kenny, P., 2005. Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montr. CRIM-06/08-13. pp. 1–17.
Kindermans, P.J., Schreuder, M., Schrauwen, B., Müller, K.R., Tangermann, M., 2014. True zero-training brain-computer interfacing – an online study. PLOS ONE 9 (7).
La Rocca, D., Campisi, P., Scarano, G., 2012. EEG biometrics for individual recognition in resting state with closed eyes. Int. Conf. Biometrics Spec. Interes. Gr., no. Figure, vol. 1 1–12.
La Rocca, D., Campisi, P., Vegso, B., Cserti, P., Kozmann, G., Babiloni, F., De Vico Fallani, F., 2014. Human brain distinctiveness based on EEG spectral coherence connectivity. IEEE Trans. Biomed. Eng. 61 (9), 2406–2412.
Lawhern, V., Slayback, D., Wu, D., Kass, M., 2016. Efficient labeling of EEG signal artifacts using active learning. Proc. – 2015 IEEE Int. Conf. Syst. Man, Cybern. SMC, 2015 3217–3222.
Lee, C.-H., Giachin, E., Rabiner, L.R., Pieraccini, R., Rosenberg, A., 1992. Improved acoustic modeling for large vocabulary continuous speech recognition. Comput. Speech Lang. 6 (April (2)), 103–127.
Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. Neural Eng. 4 (June (2)), R1–R13.
Maiorana, E., La Rocca, D., Campisi, P., 2016. On the permanence of EEG signals for biometric recognition. IEEE Trans. Inf. Forensics Secur. 11 (1), 163–175.
Marcano, J.L., Bell, M.A., Beex, A.L., 2018. Classification of ADHD and non-ADHD subjects using a universal background model. Biomed. Signal Process. Control 39, 204–212.
Marcel, S., Millán, J.D.R., Millan, J.D.R., 2007. Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. IEEE Trans. Pattern Anal. Mach. Intell. 29 (4), 743–748.
McLaren, M., Van Leeuwen, D., 2012. Source-normalized LDA for robust speaker recognition using i-Vectors from multiple speech sources. IEEE Trans. Audio. Speech Lang. Process. 20 (3), 755–766.
Obeid, I., Picone, J., 2016. The Temple University Hospital EEG data corpus. Front. Neurosci. 10 (MAY), 196.
Rabiner, L.R., 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. pp. 257–286.
Radha, M., Garcia-Molina, G., Poel, M., Tononi, G., 2014. Comparison of feature and classifier algorithms for online automatic sleep staging based on a single EEG signal. Conf. Proc.... Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf., vol. 2014 1876–1880.
Ramgopal, S., Thome-Souza, S., Jackson, M., Kadish, N.E., Sánchez Fernández, I., Klehm, J., Bosl, W., Reinsberger, C., Schachter, S., Loddenkemper, T., 2014. Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy. Epilepsy Behav. 37, 291–307.
Reynolds, D.A., 2009. Gaussian mixture models. Encycl. Biometr. 2, 659–663.
Rodrigues, D., Silva, G.F.A., Papa, J.P., Marana, A.N., Yang, X.-S.S., 2016. EEG-based person identification through Binary Flower Pollination Algorithm. Expert Syst. Appl. 62 (November), 81–90.
Ruiz-blondet, M.V., Jin, Z., Laszlo, S., 2016. CEREBRE: a novel method for very high accuracy event-related potential biometric identification. IEEE Trans. Inf. Forensics Secur. 6013 (July (c)), 1–13.
Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R., 2004. BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans. Biomed. Eng. 51 (June (6)), 1034–1043.
Schröder, M., Lal, T.N., Hinterberger, T., Bogdan, M., Hill, N.J., Birbaumer, N., Rosenstiel, W., Schölkopf, B., 2005. Robust EEG channel selection across subjects for brain–-computer interfaces. EURASIP J. Adv. Signal Process. 2005 (19), 3103–3112.
Senoussaoui, M., Kenny, P., Dehak, N., Dumouchel, P., Canada, C., 2010. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. Odyssey Speak Lang. Recognit. Work.
Senoussaoui, M., et al., 2011. Mixture of PLDA models in I-vector space for gender-independent speaker recognition. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, no. August 25–28.
Su, K., Robbins, K.A., 2013. A framework for content-based retrieval of EEG with applications to neuroscience and beyond. Proc.... Int. Jt. Conf. Neural Networks. Int. Jt. Conf. Neural Networks 1–8.
Tatum, W.O., Tatum IV, W.O., 2014. Handbook of EEG Interpretation, 2nd ed. Demos Medical, New York.

Ward, C.R., Obeid, I., 2016. Feasibility of Identity Vectors for use as subject verification and cohort retrieval of electroencephalograms. 2016 IEEE Signal Process. Med. Biol. Symp. IEEE, December, vol. 1 1–5.
Ward, C.R., Picone, J., Obeid, I., 2016. Applications of UBMs and I-vectors in EEG subject verification. 2016 38th Annu Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE, August 748–751.
Wulsin, D.F., Litt, B., 2011. An unsupervised method for identifying regions that initiate seizures on intracranial EEG. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS 3091–3094.
Yang, S., Deravi, F., Hoque, S., 2016. Task sensitivity in EEG biometric recognition. Pattern Anal. Appl. 1–13.