2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Towards the Classification of Error-Related Potentials using Riemannian Geometry

Yichen Tang[1,3], *Student Member, IEEE*, Jerry J. Zhang[1], *Student Member, IEEE*,
Paul M. Corballis[2], and Luke E. Hallum[1], *Member, IEEE*

***Abstract*— The error-related potential (ErrP) is an event-related potential (ERP) evoked by an experimental participant's recognition of an error during task performance. ErrPs, originally described by cognitive psychologists, have been adopted for use in brain-computer interfaces (BCIs) for the detection and correction of errors, and the online refinement of decoding algorithms. Riemannian geometry-based feature extraction and classification is a new approach to BCI which shows good performance in a range of experimental paradigms, but has yet to be applied to the classification of ErrPs. Here, we describe an experiment that elicited ErrPs in seven normal participants performing a visual discrimination task. Audio feedback was provided on each trial. We used multi-channel electroencephalogram (EEG) recordings to classify ErrPs (success/failure), comparing a Riemannian geometry-based method to a traditional approach that computes time-point features. Overall, the Riemannian approach outperformed the traditional approach (78.2% versus 75.9% accuracy, $p < 0.05$); this difference was statistically significant ($p < 0.05$) in three of seven participants. These results indicate that the Riemannian approach better captured the features from feedback-elicited ErrPs, and may have application in BCI for error detection and correction.**

I. INTRODUCTION

A key goal of brain-computer interface (BCI) research is to use signals derived from the electroencephalogram (EEG) to interface with a device [1]. Typical BCIs use machine learning to translate EEG activity into control signals [1]. Because both humans and machines make mistakes [2], [3], [4], error detection and correction are vital for improving the utility of a BCI [2], [5]. The error-related potential (ErrP) is family of event-related potential (ERP) components that can be derived from EEG activity. ErrPs are usually evoked when a person recognises an error, regardless of whether that error was made by the person, someone else, or the BCI [2], [3], [4], [6]. ErrPs were originally described in the context of cognitive psychology [3], [6], and were thereafter adopted for BCIs for the error detection and correction and for online refinement of decoding algorithms [2], [7]. ErrPs comprise a characteristic series of voltage deflections, including a frontocentral negativity followed by a positivity, and then a parietal positivity [2], [8]. However, based on the source of error information, these components may have different manifestations and nomenclature [2], [8]. When the person recognises their own error, the frontal components are termed error-related negativity (ERN) and the positive deflection (Pe), while the ErrP generated by feedback is termed the feedback-related negativity (FRN). The later frontal (P3A [2], [6], [9]) and parietal (P3B [8], [10]) positivities are cognitive components not specific to error monitoring or feedback.

Traditionally, ERP classification for BCI involves computation of features at different points in time on a number of EEG channels [1], which are then concatenated to form a high-dimensional feature vector. By contrast, a new approach that uses Riemannian geometry [1] has often outperformed traditional methods [1], [11]. Instead of extracting a predefined feature vector, these methods map ERPs onto a Riemannian manifold – often by computing a covariance matrix on individual trials – and then classifying samples directly on the manifold, or transforming samples into vectors for classification [1], [11]. Although this classification was originally designed for capturing spatial features for use in a motor-imagery paradigm, these methods can potentially be adapted to a wide range of BCI paradigms [11], [12] by changing how the covariance matrices are computed (such as the adoption of prototype ERP responses, discussed below).

To our knowledge, Riemannian geometry-based methods have not yet been used for feature extraction and classification of ErrPs. Therefore, here we apply such a method combined with logistic regression to extract and classify ErrPs during a two-alternative forced-choice (2AFC) visual-disctimination task. We compare outcomes to a traditional feature extraction and classification method.

II. METHODS

A. Experiment and EEG Recording

1) Participants: We recorded behavioural responses and multi-channel EEG from seven participants (six males; age range: 20 to 25 years old), all with normal or corrected-to-normal vision. Participants provided informed consent. Experimental protocols were approved by the University of Auckland Human Participants Ethics Committee.

2) Procedure: Each participant sat for five blocks of 60 trials (300 trials total); blocks were separated by short breaks. Participants performed a visual discrimination task at fixation. Each trial began with a 1-s presentation of a white crosshairs on a grey background, followed by two temporal intervals. The first interval contained a circular target (diameter = 0.5°of visual angle) with luminance = L1;

*This work was partly supported by a University of Auckland Faculty of Engineering Research Development Fund award to L.E.H.

[1]Yichen Tang, Jerry J. Zhang, and Luke E. Hallum are with Department of Mechanical Engineering, University of Auckland, Auckland, New Zealand

[2]Paul M. Corballis is with School of Psychology, University of Auckland, Auckland, New Zealand

[3]Yichen Tang is the corresponding author. Email: ytan415@aucklanduni.ac.nz

the second interval contained a circular target with luminance = L2. After the two intervals, the participant responded with a keypress to indicate whether the 1st or 2nd interval contained the target at higher luminance (we randomised this experimental parameter across trials). The trial structure is detailed in Fig. 1. We used a one-up/one-down staircase to adjust target luminance on a trial-by-trial basis. This staircase procedure ensured that the task was sufficiently challenging to elicit approximately equal numbers of correct and incorrect responses. On each trial, participants were allowed 2 s to provide a response before time-out; the few trials on which participants gave no response were treated as incorrect. Each block began with five practice trials which we later discarded from our analysis. We instructed participants to always fixate the display centre and to blink as infrequently as possible during the experiment.

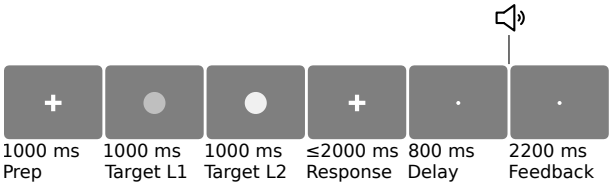

Fig. 1: Trial structure. On each trial, the participant prepared ("Prep.") by fixating a crosshairs (width = 1°of visual angle). Then, a circular target (diameter = 0.5°) appeared in each of two 1-s temporal intervals, followed by the reappearance of the crosshairs. The participant responded with a keypress to indicate whether the 1st or 2nd target was higher luminance. Audio feedback was provided 800 ms after the participant's response.

3) Visual stimuli and feedback: We presented visual stimuli using two 24-inch, gamma-corrected liquid-crystal displays (60 Hz; ColorEdge CG247X; EIZO Corp., Hakusan, Japan) [13], each reflected into its eye through a 45° mirror stereoscope. Stimuli at fixation were identical on both displays. (For separate experimental purposes which we describe in a companion paper [14], dichoptic stimuli were presented in annuli surrounding fixation. These dichoptic stimuli were absent from participants' awareness, and are of no consequence to the present results.) We provided correct (incorrect) feedback using 150-ms pure tone at 700 Hz (200 Hz). We used two loudspeakers (Edifier r1700bt; Edifier International Ltd., Beijing, China), each 0.55 m from its ear and 30° off the sagittal plane. The sound level of tones was adjusted to approximately 70 dB SPL at the ear. We used a phototransistor (TEPT4400; Vishay Intertechnology, Inc., Malvern, PA, United States) and a sound sensor (XC-4438; Jaycar, Rydalmere, NSW, Australia) together with two Arduino UNO R3 boards to synchronise visual and audio signals with EEG recordings.

4) EEG recording and preprocessing: We used a BioSemi ActiveTwo AD-box (ADC-17; ActiveTwo; Biosemi, Amsterdam, Netherlands) to record and amplify multi-channel EEG at a rate of 2048 samples per second for each of 32 scalp channels. We positioned the scalp electrodes according to the international 10-20 system; the channels were Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, PO7, P3, POz, PO3, O1, Oz, O2, PO4, P4, PO8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, AF4, Fp2, Fz, and Cz. Raw recordings were, first, bandpass filtered (1 Hz to 100 Hz) and then notch filtered (at 50Hz and 100Hz). We re-referenced recordings to the average across channels and used independent component analysis (ICA) to remove any EOG contamination [15]. Recordings were epoched from 0.5 s before to 2 s after the onset of audio feedback, baseline-subtracted (i.e., we subtracted the average of the recording from -0.5 to 0 s from each data point in each epoch), and downsampled to 256 samples per second per channel. We labelled each epoch to indicate the participant's success or otherwise on the corresponding trial (see *Visual stimuli and feedback*). In total, we recorded three hundred epochs from each participant.

B. Classification and Evaluation

1) Cross-validation: To evaluate classifiers, we used 10-fold cross-validation (CV), repeated 10 times (i.e., "10-by-10-fold" CV). We performed this 10-by-10-fold CV within participant [16]. On each repeat, the dataset was randomly shuffled into ten equally sized parts, each with a balanced proportion of the "success" and "failure" classes. For each fold, the classifier was trained on 90% of the data and its performance was quantified using the remaining 10%, calculating accuracy in the standard fashion: the proportion of the predicted classes of the epochs that matched the true classes. CV was implemented using Python (version 3.8.3). Specifically, we used scikit-learn (v0.23.1) [17], an open-source library that implements common machine-learning algorithms.

2) Within- and between-participant comparison of classifier accuracy: To compare classification methods within participant, we used the Nadeau & Bengio corrected t-test as described by Bouckaert & Frank [18]. To compare classification methods across participants, we used a permutation test as follows. For each participant, we computed median accuracy separately for each of the two classifiers being compared, and then calculated the difference between these medians ("classifier 1 minus classifier 2"). This metric was summed across all participants. We then z-scored this metric against a null distribution and computed the p-values. The null distribution comprised 1000 null metrics, each of which was computed in the same way as described above after shuffling the labels ("classifier 1" and "classifier 2") on accuracies used in the computation.

3) Chance-level accuracy: We computed the chance-level accuracy for each participant using a shuffle test based on the benchmark approach (described below). We randomly shuffled the classes for all epochs, tested the benchmark approach using the CV framework stated above, and recorded the average CV accuracy for 100 random shuffles. Then, we recorded the average and the 97.5 percentile of the shuffled accuracies as the classification chance level and the chance threshold for each participant.

4) Riemannian geometry-based feature extraction and classification: To extract ErrP features, we adopted methods introduced by Barachant and colleagues [11], [12], [19], and used Barachant's Python (pyRiemann, v0.2.6) implementation of these methods [20]. In brief, on each fold of a 10-fold cross-validation, we used training data to construct covariance matrices. We projected these covariance matrices onto a space tangential to the manifold, defined by the geometric mean of all matrices. This projection vectorized matrices. These feature vectors were used for training an L2-regularized logistic regression classifier implemented in scikit-learn [17]. To construct covariance matrices, we first windowed our training epochs, using only 100 to 600 ms (i.e., 128 samples per channel), where 0 ms is the onset of feedback. On each channel (32 channels total), we separately averaged "failure" and "success" epochs, giving two "prototype" matrices, each 32 rows-by-128 columns. We concatenated these prototype matrices with each single trial taken from the training set, giving a 96-by-128 "super trial" matrix. Each super trial was used to compute a 96-by-96 covariance matrix. Parts of this matrix captured the covariance between channels within the given trial; other parts of the matrix captured covariance between the given trial and the prototype trials. Using these training prototype matrices, we then applied the same procedure to single test trials; we used the feature vectors generated by test trials to assess classifier performance.

5) Benchmark feature extraction and classification: To help evaluate the Riemannian geometry-based feature extraction, we developed a benchmark. For each epoch, for each electrode, we computed a feature vector comprising eight feature. Features were (1) mean and (2) standard deviation computed during each of four temporal windows: 100-200 ms, 200-300 ms, 300-400 ms, and 400-600 ms, where 0 ms refers to the onset of audio feedback. These features were concatenated and each feature was scaled by its maximum absolute value in the training set. Again, we used the logistic regression classifier to perform classification; this classifier outperformed a range of other classifiers, including a support vector machine, and a linear disciminant analysis (data not shown). This benchmark – specifically, our use of windows – was adapted from recent work by others [21] and [22].

III. Results

A. Participants' behavioural performance

Participants performed the fixation task capably. To estimate each participant's threshold, we averaged across blocks the last 10 reversals of each block's staircase, and then transformed the contrasts back into non-logged Weber contrasts [23]. Thresholds and performance are shown in TABLE I.

B. ERP components

We recorded robust ERPs from all participants. In all participants we observed a FRN at the fronto-central scalp areas (centered at Fz) in failure epochs compared with success epochs (failure minus success), peaking between 100 to 200 ms. Following this negativity, in 5 of 7 participants, we also observed a frontally distributed positivity which peaked around 300 ms (P3A [8], [10]). These two ERP components are exemplified in Fig. 2. These observations were broadly consistent with the ErrP described in the literature [2], [8], which we discuss below (see *Discussion*).

TABLE I: The discrimination thresholds (in Weber contrast %) and behavioural performance (% correct behavioural responses) for all participants (P1 to P7). The percentage of correct responses was calculated across all five blocks.

	P1	P2	P3	P4	P5	P6	P7
Discrimination threshold	0.338	0.779	0.542	1.129	0.161	0.269	0.382
Behavioural performance	57.0	55.7	57.3	55.7	59.7	59.7	57.3

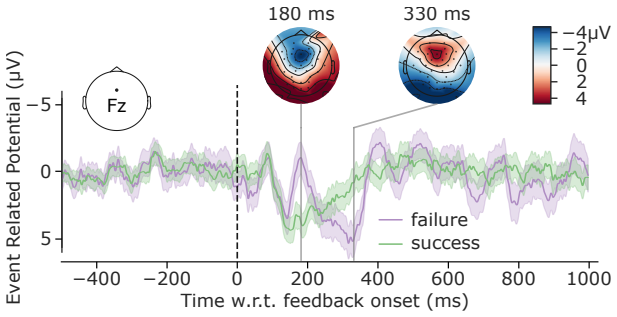

Fig. 2: Example ERPs, participant P5. ERPs recorded on electrode Fz (inset), showing the average of trials on which the participant received feedback indicating success on the discrimination task (green) and feedback indicating failure (purple). Shaded regions show 95% confidence intervals computed via bootstrapping. The scalp maps show "failure minus success" at 180 ms (left) and 330 ms (right). For this participant, the negativity at 180 ms was localised to a frontal-central region surrounding electrode Fz; the positivity at 330 ms showed similar spatial organisation.

C. Classification approach performance

We used 10-by-10-fold cross-validation to quantify the accuracy of all classification approaches (see *Cross-validation*); all approaches performed at rates above chance. Overall, the Riemannian geometry-based approach outperformed the benchmark. For experimental participants P2, P6, and P7, we saw significantly higher performance using the Riemannian approach than that of the benchmark: P2, 86.8% versus 80.7% (t=2.326, p=0.022); P6, 81.6% versus 76.9% (t=2.136, p=0.035); P7, 88.7% versus 83.8% (t=2.016, p=0.047). In two other participants (P1 and P5), the Riemannian approach outperformed the benchmark, but the differences did not reach statistical significance. Across all participants, the Riemannian approach also reached an overall accuracy of 78.2% which was statistically significantly greater than the benchmark's performance of 75.9% by the benchmark (z=4.028, p=5.632e-05). Data are shown in Fig. 3.

IV. Discussion

Our results demonstrate that the Riemannian geometry-based approach to classifying the FRN outperforms a tra-

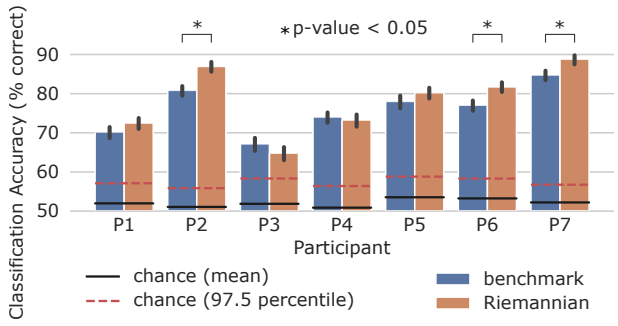

Fig. 3: Comparison of cross-validated classifier accuracy. In 3 of 7 experimental participants, the performance of the Riemannian geometry-based method was statistically significantly greater than that of the benchmark.

ditional approach. Because both of these approaches used the same classifier (logistic regression), it seems that the Riemannian approach was better able to extract the salient features contained in the FRN. A further advantage of the Riemannian approach is that it represents both spatial and temporal information without much feature engineering (i.e., the postulation of features and specification of temporal windows, as is necessary in the traditional approach).

ErrPs elicited by error feedback have a characteristic morphology; a frontocentral FRN appears between 200 and 300 ms after feedback, followed by a P3A appearing after 300 ms [8]. In our recordings, we observed a positivity in five of seven participants (P2, P4 through P7) after 300 ms. We observed a negativity surrounding site Fz (e.g., Fig. 2) in all participants. However, this peaked somewhat early, between 100 and 200 ms, that is, where one might expect to find the N1 and P2 components of the auditory evoked response (AER) [24]. Because we provided feedback using pure tones at 700 and 200 Hz, it is possible that frequency-related differences in N1/P2 contributed to the negativity we observed. However we suspect this contribution is small. Picton et al. [25] measured N1 and P2 amplitudes as a function of frequency, using tone bursts of 250, 500, and 1000 Hz at 87 ± 3 dB SPL. Picton's data indicate that our use of 200 and 700 Hz for feedback may contribute a small negativity to our recordings at around 100 to 175 ms. However, the tone bursts used by Picton were high-intensity compared to ours, and may not accurately model AERs in our participants. In ongoing work, we aim to separate the FRN from any AERs.

References

[1] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.

[2] R. Chavarriaga, A. Sobolewski, and J. d. R. Millán, "Errare machinale est: the use of error-related potentials in brain-machine interfaces," *Frontiers in Neuroscience*, vol. 8, p. 208, 2014.

[3] T. Zeyl, "Adaptive brain-computer interfacing through error-related potential detection," Ph.D. dissertation, University of Toronto, 2016.

[4] G. Schalk, J. R. Wolpaw, D. J. McFarland, and G. Pfurtscheller, "EEG-based communication: presence of an error potential," *Clinical Neurophysiology*, vol. 111, no. 12, pp. 2138–2144, 2000.

[5] L. C. Parra, C. D. Spence, A. D. Gerson, and P. Sajda, "Response error correction-a demonstration of improved human-machine performance using real-time EEG monitoring," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, no. 2, pp. 173–177, 2003.

[6] W. H. Miltner, C. H. Braun, and M. G. Coles, "Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection," *Journal of Cognitive Neuroscience*, vol. 9, no. 6, pp. 788–798, 1997.

[7] A. Llera, M. A. van Gerven, V. Gómez, O. Jensen, and H. J. Kappen, "On the use of interaction error potentials for adaptive brain computer interfaces," *Neural Networks*, vol. 24, no. 10, pp. 1120–1127, 2011.

[8] M. Ullsperger, A. G. Fischer, R. Nigbur, and T. Endrass, "Neural mechanisms and temporal dynamics of performance monitoring," *Trends in Cognitive Sciences*, vol. 18, no. 5, pp. 259–267, 2014.

[9] W. J. Gehring, B. Goss, M. G. Coles, D. E. Meyer, and E. Donchin, "A neural system for error detection and compensation," *Psychological Science*, vol. 4, no. 6, pp. 385–390, 1993.

[10] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical Neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.

[11] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: a review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1753–1762, 2016.

[12] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Riemannian geometry applied to BCI classification," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2010, pp. 629–636.

[13] L. E. Hallum and S. L. Cloherty, "Liquid-crystal display (LCD) of achromatic, mean-modulated flicker in clinical assessment and experimental studies of visual systems," *Plos one*, vol. 16, no. 3, p. e0248180, 2021.

[14] J. J. Zhang, Y. Tang, S. C. Dakin, and L. E. Hallum, "Balanced, orientation-dependent dichoptic masking in cortex of visually normal humans measured using electroencephalography (EEG)," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Nov 2021.

[15] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[16] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining Practical Machine Learning Tools and Techniques*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2004, pp. 3–12.

[19] A. Barachant and M. Congedo, "A plug&play P300 BCI using information geometry," *arXiv preprint arXiv:1409.0107*, 2014.

[20] A. Barachant, "pyriemann." [Online]. Available: https://github.com/alexandrebarachant/pyRiemann/tree/v0.2.6

[21] I. Kakkos, E. M. Ventouras, P. A. Asvestas, I. S. Karanasiou, and G. K. Matsopoulos, "A condition-independent framework for the classification of error-related brain activity," *Medical & Biological Engineering & Computing*, vol. 58, no. 3, pp. 573–587, 2020.

[22] F. M. Schönleitner, L. Otter, S. K. Ehrlich, and G. Cheng, "Calibration-free error-related potential decoding with adaptive subject-independent models: A comparative study," *IEEE Transactions on Medical Robotics and Bionics*, vol. 2, no. 3, pp. 399–409, 2020.

[23] A. B. Cobo-Lewis and Y. Yei-Yu, "Selectivity of cyclopean masking for the spatial frequency of binocular disparity modulation," *Vision Research*, vol. 34, no. 5, pp. 607–620, 1994.

[24] R. Näätänen and T. Picton, "The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure," *Psychophysiology*, vol. 24, no. 4, pp. 375–425, 1987.

[25] T. W. Picton, D. L. Woods, and G. Proulx, "Human auditory sustained potentials. II. Stimulus relationships," *Electroencephalography and Clinical Neurophysiology*, vol. 45, no. 2, pp. 198–210, 1978.