Spring 2016

# Application of DTW Barycenter Averaging to Finding EEG Consensus Sequences

Levi C. Davis
*University of Colorado, Boulder*, levi.davis@colorado.edu

# Application of DTW Barycenter Averaging to Finding EEG Consensus Sequences

Levi C. Davis

Department of Psychology and Neuroscience

University of Colorado at Boulder

Undergraduate Thesis

Thesis Advisor:

Dr. Tim Curran

Department of Psychology and Neuroscience

Defense Committee:

Dr. Tim Curran, Department of Psychology and Neuroscience

Dr. Randall O'Reilly, Department of Psychology and Neuroscience

Dr. Willem Schreüder, Department of Computer Science

Defended April 4, 2016

**Abstract**

DTW Barycenter Averaging (DBA) has proven to be a useful tool for calculating consensus sequences, but it has not yet been applied to real electroencephalography (EEG) data. This study tests DBA on real EEG sequences using the modification proposed by Kotas et al. (2015), and proposes several further modifications to the initial sequence selection process to improve the method's efficacy for EEG analysis. Errors in peak latency and peak amplitude measures for a single EEG component, namely N250, were measured to test each method. Three of the proposed DBA variations produced consensus sequences that were significantly more accurate for replicating features of single-trial N250 components than the widely-used Event-Related Potential (ERP) technique. Potential implications include the uncovering of previously obscured effects in EEG data and providing more accurate descriptions of the prototypical electrophysiological responses to external events.

**Introduction**

Electrophysiology, the study of the nervous system through measurement of the electrical properties of tissues, provides the advantage of an extremely high temporal resolution when compared to other methods of studying the brain. Electrophysiological data can be obtained with millisecond resolution, allowing neurological events to be accurately pinpointed in time (Luck, 2005). Electroencephalography (EEG) is particularly advantageous, as its non-invasive nature allows researchers to easily study the temporal aspects of human cognition (Luck, 2005).

Unfortunately, EEG has a relatively poor signal-to-noise ratio; thus, obtaining meaningful results requires averaging across a large data pool (Luck, 2005). The most common method for averaging is known as the Event-Related Potential (ERP) technique (Luck, 2005). In order to obtain an ERP, short segments of EEG data are temporally-aligned based on important external events, such as stimulus presentation or response time; then, the data at each time point is averaged across trials and/or subjects (Luck, 2005). The resulting ERP data is comprised of several positive and negative deflections in voltage, referred to as ERP components,

representing consistent neurological reactions to the stimulus or task being studied (Luck, 2005).

However, there is a potentially confounding factor in the ERP method. When averaging temporal sequences, it is important that analogous features do not vary in timing between sequences (Huang and Jansen, 1984; Luck, 2005). If any latency differences exist across the trials or subjects being averaged, the calculated ERP waveform may not accurately represent the nature of the underlying neurological activity. Typically, components will have a lower magnitude than the events that they represent, as well as a greater temporal duration (Luck, 2005). To further complicate matters, temporally proximal components of opposite charge can summate and cancel, becoming indistinguishable in the ERP waveform (Luck, 2005). These obscured components could contain important cognitive features that have yet been unobserved due to the cancellation effect. Indeed, it may be that significant effects have gone unnoticed in previous EEG studies due simply to variations in temporal latency of an important component.

For these reasons, while the ERP waveform can certainly provide insight into the electrophysiological response to an event, it does not constitute a *consensus sequence* for that response; in other words, the ERP cannot be considered a prototype of which each single EEG response is a variation. In some cases, the ERP may not resemble the original EEG waveforms at all. If researchers hope to uncover the prototypical electrophysiological response to a certain event, an alternate method is needed. *Figure 1* illustrates the differences between ERP and a theoretical consensus sequence using simple simulated EEG waveforms.

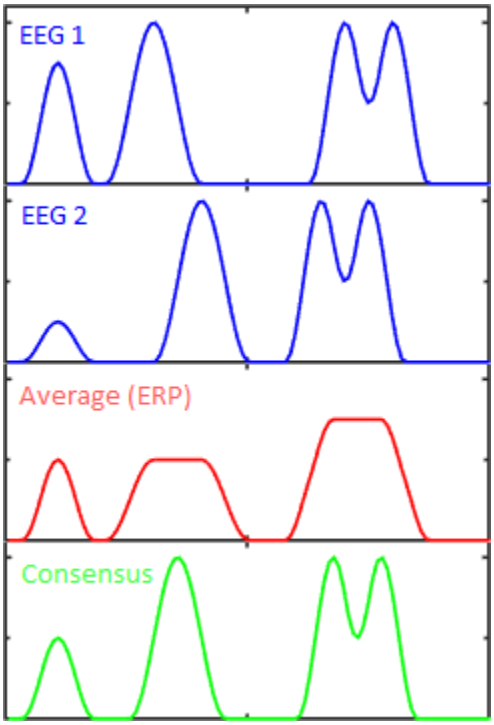


***Figure 1*: Simulated EEG data (blue) and the corresponding ERP (red) and consensus (green) waveforms. Left: an isolated component with constant latency will be accurately averaged in the ERP waveform. Middle: an isolated component with varying latency will be flattened and broadened in the ERP waveform. Right: overlapping components with varying latency will result in blending and/or cancellation in the ERP waveform, potentially obscuring the underlying components.**

In order to supply an alternative to the ERP method of EEG data analysis, further insight into the

temporal features of the underlying sequence must be identified. In the field of signal analysis, a technique known as dynamic time warping (DTW) is used to nonlinearly align two temporally-varying sequences with each other and provide a measurement of their similarity (Sakoe and Chiba,1978).

DTW has been applied to EEG data in various forms throughout the past few decades (Huang and Jansen, 1985; Picton, Hunt, Mowrey, Rodriguez, and Maru, 1988; Wang, Begleiter, and Porjesz, 2001; Casarotto, Bianchi, Cerutti, and Chiarenza, 2005; Aarabi, Kazemi, Grebe, Moghaddam, and Wallois, 2009; Assecondi et al., 2009), however DTW itself does not produce a prototypical consensus sequence. Several methods have been proposed that build upon DTW to construct a consensus (Myers and Rabiner, 1981; Gupta, Molfese, Tamman, and Simos, 1996; Keogh and Pazzani, 2001; Petitjean, Ketterlin, and Gançarski, 2011), and notable among these is DTW barycenter averaging (DBA) (Petitjean et al., 2011). DBA uses a modified version of DTW to align a set of sequences with each other and compute the average of the aligned sequences, iteratively improving the alignment in an attempt to converge on a consensus sequence (Petitjean et al., 2011). This method has been shown to outperform other contemporary DTW-based averaging methods (Petitjean and Gançarski, 2012). Kotas, Leski and Moroń (2015) applied DBA to simulated EEG data, but to our knowledge there has not yet been a published application of this relatively new technique to real EEG data. The present article attempts to address this.

There are several complications when applying DBA to EEG data. As noted by Kotas et al. (2015), DBA tends to amplify noise in the input sequences, which is prevalent in EEG waveforms. The authors proposed a method to resolve this amplification: When computing the DTW alignment path, it was proposed that the alignment costs should be redefined, such that short subsequences are matched together rather than single data points. This modification to the DBA algorithm will be used in the application of DBA herein.

In addition to noise amplification, we found that the DBA method is very sensitive to initial sequence selection. Prior to iterative alignment, a single sequence needs to be selected to act as a tentative consensus; the original authors proposed that this sequence could be arbitrarily selected from the input sequences (Petitjean et al., 2011), but in the present study

we found that the temporal features of the initial EEG sequence tended to persist into the final consensus waveform, regardless of how typical or atypical those features were. So, in this study we will propose several alternate solutions for selecting an initial sequence, and evaluate their relative performance. Each proposed modification will be described further in the *Methods* section, under the *Algorithms* heading.

In order to evaluate the performance of the various techniques, the present study will examine peak amplitude and latency of a specific ERP component. Since peak amplitude and latency can be especially altered in ERPs relative to their EEG counterparts (Luck, 2005), this will be a strong measure of any improvements provided by DBA. In particular, the negative component occurring at approximately 250ms (N250) will be examined. In ERP, the N250 is seen as a small negative deflection occurring between two large positive components. If there is any latency variability in N250 or the surrounding components, then the small deflection seen in ERP is likely a result of the cancellation of a larger EEG component. For these reasons, N250 is a prime candidate for DBA analysis, which may be able to reduce cancellation, and therefore preserve its features more faithfully than standard ERP averaging.

## Methods

### Data set

The data set used in this study originates from an unpublished EEG study (Curran, N.d.). This data set is similar to one used in a previous study (Scott, Tanaka, Sheinberg and Curran, 2008), in which the N250 component was found to be significantly amplified after training subjects to differentiate visual stimuli. This amplification should allow reliable identification of the N250 component in subject ERPs and single-trial EEGs.

Data from 28 subjects was studied; 12 subjects with 92 trials each and 16 subjects with 93 trials each. A single EEG channel, E65 (corresponding to lateral parieto-occipital cortex), was analyzed. Data had previously been low-pass and high-pass filtered to reduce the influence of noise, and trials containing non-neurological artifacts were discarded. Waveforms ran from

approximately 0ms to 500ms relative to stimulus presentation; additional 120ms buffers before and after the 500ms duration of interest were included in the averaging processes, to ensure stable waveforms throughout the duration of interest. Additionally, data was compressed by discarding every other data point in order to reduce computational complexity. The original sampling rate was 250Hz; compression reduced the sampling rate to 125Hz, giving a temporal resolution of 8ms per data point.

## Algorithms

Five variations of DBA averaging were tested on the data set, in addition to standard ERP averaging. All versions of DBA used the noise-reduction improvement proposed by Kotas et al. (2015). The authors did not specify a value for the sampling distance used in their algorithm: Through trial and error, a sampling window of 24ms (3 data samples) before and after the central point was selected for this study based on visual results.

The first DBA variation ($DBA_0$) is the default proposed by the original authors (Petitjean et al., 2010), in which the initial sequence is selected arbitrarily from one of the input sequences. Specifically, the first non-discarded trial for each subject was used as the initial sequence for this method.

The second variation ($DBA_{ERP}$) calculates the ERP of all input sequences and uses that as the initial sequence for DBA averaging. The ERP contains minimal noise while still retaining a similar structure to the input sequences, and therefore seems like a strong starting point for alignment.

The third variation ($DBA_{DTW}$) uses the ERP as a reference to select one of the initial sequences. Each input sequence is compared to the ERP using standard DTW, and the sequence with the lowest score (i.e., the sequence most similar to the ERP) is selected. It is hoped that this sequence will be the most prototypical of the input sequences, due to its similarity to the ERP, while still retaining original EEG features which are lost in the ERP.

The fourth variation ($DBA_{LEN}$) similarly uses the ERP as a reference for initial sequence selection, but it utilizes a variation of DTW to score each sequence. Instead of a standard DTW score, the modified DTW returns the length of the path through the scoring matrix. The

sequence that requires the shortest matrix path is hoped to bear the closest temporal resemblance to the ERP, while ignoring differences in the magnitudes of components.

The fifth variation ($DBA_{DEV}$) is similar to $DBA_{LEN}$ in that it uses a modified DTW to select the sequence most similar to the ERP. However, this variation uses a slightly more complicated measure of similarity. At each step in the path through the scoring matrix, the integer distance from the diagonal is computed. This is most easily done by taking the magnitude of the difference between the x and y coordinates of the current position in the matrix, $|x - y|$. This quantity is cumulatively added to the value calculated during the previous steps in the path, integrating the diagonal distance along its path through the matrix. The reasoning behind this variation of the algorithm is that multiple brief deviations from the diagonal, corresponding to temporal shifts of small time periods such as a single wave component, may be more optimal than a chronic deviation from the diagonal in which a large chunk of the sequence is shifted temporally. Repeated departures and returns to the diagonal would increase path length, causing $DBA_{LEN}$ to favor the latter, while $DBA_{DEV}$ would favor the former due to the decreased duration of deviations from the diagonal.

## Analysis

To quantify the performance of each algorithm in finding a consensus in the EEG data, magnitude and latency of the N250 peak were measured on both the single-trial EEG sequences and the consensus sequences separately. Each of the 28 subjects was analyzed individually. The single-trial peak values were averaged across trials for each subject (92 or 93 trials each) to provide a measurement of the prototypical N250 peak latency and magnitude for that subject. These prototypical values were then subtracted from the values measured from each consensus sequence, to provide a measurement of error for each method.

N250 magnitude and latency were detected using a local minimum search in the time window of 215 to 295ms post-stimulus. This is a relatively wide window compared to the 232 to 280ms used by Scott et al. (2008), as it is being applied to raw EEG data which has a much higher latency variability than ERP waveforms (Luck, 2005). Peak locality was tested using sample windows of 24ms (3 data points) on either side of the peak to reduce incorrect feature

detection (Luck, 2005); if no local minimum was detected under these criteria, the search was repeated using iteratively smaller locality-testing windows until a minimum was found.

For both magnitude and latency of N250, the average error of each consensus method was evaluated across subjects using two definitions of error: mean absolute error (MAE) and root mean square error (RMSE). MAE acts as a precise measurement of the average error, while RMSE is affected more strongly by outliers, providing insight into the distribution of the errors (Willmott and Masuura, 2005). For each error measurement, the analogous deviation measurement (mean amplitude deviation (MAD) and standard deviation (STD), respectively) was performed on the original EEG dataset on a per-subject basis; then averaged across subjects to give a sense of scale to the consensus sequences' error values. MAD and STD represent the average deviation from the mean of any randomly selected single-trial EEG sequence; as such, if the consensus sequences have a higher error than the single-trial deviation, they are poor estimators of that value.

**Results**

The error measurements of each method, as well as the average single-trial EEG deviations, are depicted in *Table 1*; the statistical significances of the error differences, relative to those of ERP, are shown in *Table 2*.

Error measurements of consensus sequences for N250 features

| | Peak Latency MAD | Peak Latency STD | Peak Magnitude MAD | Peak Magnitude STD |
|---|---|---|---|---|
| EEG (avg) | 0.0163 | 0.0218 | 3.4019 | 5.7433 |
| **Method** | **Peak Latency MAE** | **Peak Latency RMSE** | **Peak Magnitude MAE** | **Peak Magnitude RMSE** |
| ERP | 0.0198 | 0.0229 | 4.5298 | 6.7204 |
| DBA_0 | 0.0184 | 0.021 | 3.2459 | 5.1139 |
| DBA_erp | 0.0185 | 0.0213 | 2.6488 | 4.5949 |
| DBA_dtw | 0.0223 | 0.0257 | 2.7243 | 4.6546 |
| DBA_len | 0.0165 | 0.0198 | 2.4776 | 4.3659 |
| DBA_dev | 0.0182 | 0.0216 | 2.5327 | 4.1489 |

***Table 1*: Error measurements (mean amplitude error and root mean square error) for the various methods tested. Errors have been marked in light pink if they are higher than the corresponding deviation measured in the single-trial EEG sequences, which are depicted at the top of the table. The lowest error of each measurement has been marked in green. Two DBA errors that exceeded corresponding ERP errors have been marked in dark pink.**

All of the tested DBA variations performed better than ERP in measurements of N250 magnitude error. The improvement was statistically significant ($p < 0.05$) for all methods except $DBA_0$, the unmodified version of the technique. Notably, the p-value for $DBA_0$ was still fairly low ($p=0.07$), and may prove to be significant with a larger sample size.

Significance of errors relative to ERP

| Method | Peak Latency | Peak Magnitude |
|---|---|---|
| DBA_0 | 0.5443 | 0.0754 |
| DBA_erp | 0.6463 | 0.002 |
| DBA_dtw | 0.4559 | 0.0021 |
| DBA_len | 0.2413 | 2.84E-04 |
| DBA_dev | 0.5069 | 7.59E-04 |

***Table 2*: Significance of the error improvements relative to ERP (see *Table 2*). P-values of a two-sided Wilcoxon rank sum test are shown, with the null hypothesis that each DBA variation's errors come from a distribution with median identical to that of the ERP errors. P-values less than 0.05 are considered statistically significant and marked in green.**

DBA also showed improvements relative to ERP in measurements of N250 latency error, with the exception of $DBA_{DTW}$. However, these improvements were not statistically significant ($p > 0.05$); further experimentation with a larger number of subjects is required to confirm the existence or non-existence of these latency improvements, and they are likely to be small if they do exist. Interestingly, comparison of the latency errors with the latency deviations of the single-trial EEG sequences suggests that *all* tested methods were poor estimators of N250 latency, regardless of any improvements over ERP. MAE measurements of latency error were higher than the corresponding MAD values, suggesting that randomly-selected single-trial EEG sequences could act as better estimators of N250 peak latency than the consensus sequences produced by any of the tested methods. However, with the exceptions of ERP and $DBA_{DTW}$, the algorithms' RMSE latency errors were still lower than the corresponding average single-trial deviation, suggesting that $DBA_0$, $DBA_{ERP}$, $DBA_{LEN}$, and $DBA_{DEV}$ were in fact less prone to latency outliers than single-trial EEGs.

The inaccuracy observed in N250 peak latencies may have been an artifact introduced by the measurement method. The peak-searching algorithm inherently restricts the range of measured latencies, by searching only within a predefined temporal window. This might have prevented some latency values from being correctly measured; in particular, if the peak-searching window tends to exclude early values and late values unequally, it could introduce a bias into the measured average peak latency, increasing the errors measured from any consensus sequence that doesn't share this bias.

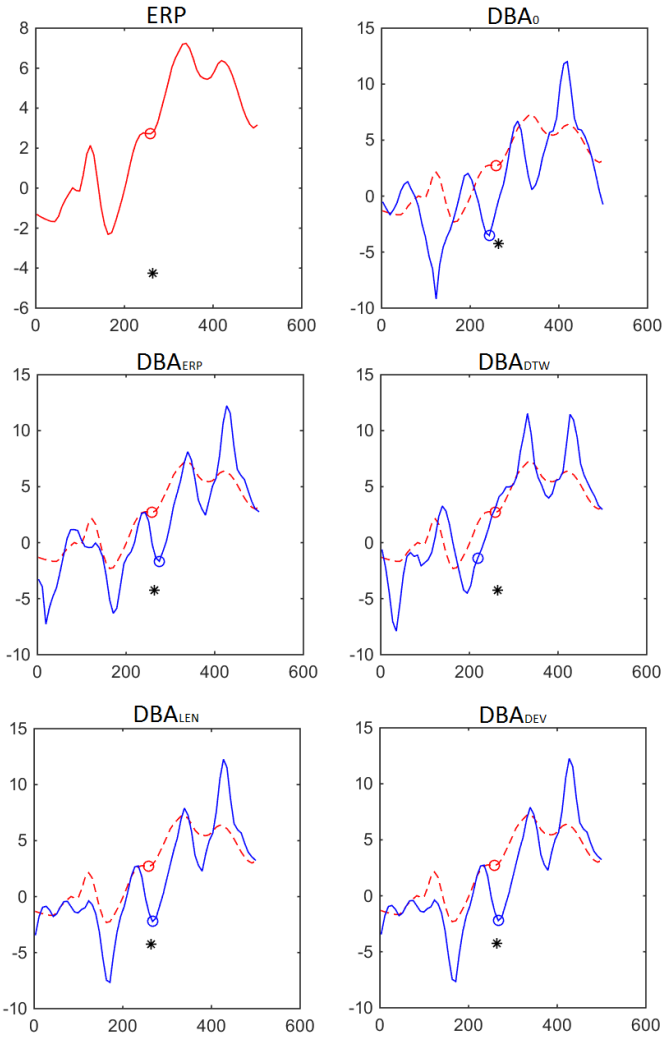


***Figure 2*: Waveforms of the various consensus sequences for a representative single subject. Consensus sequences (blue) are overlaid on the ERP sequence for reference (dashed red). N250 peaks are labeled with circles of the corresponding color. The average N250 peak of the EEG data is indicated with a black asterisk. The top-left plot displays the ERP only.**

Qualitatively, all methods of DBA alignment produced visually similar waveforms; the consensus sequences produced by each method for a single subject, including demarcations for each of the measured N250 peaks, are illustrated in *Figure 2.* This single subject was selected for its representation of consistent trends across subjects and its reflection of the results shown in *Table 1*. Each of the DBA methods produced a sequence consisting of a series of oscillations whose peaks and troughs corresponded respectively to the positive and negative components of the ERP waveform. The most notable deviations from this common pattern are the shift of the early components in $DBA_{0,}$ which resulted in the N170 occurring at approximately the same location as the P100 of the ERP waveform, and a notable lack of an N250 component in the $DBA_{DTW}$ sequence. $DBA_{DTW}$ often failed to produce a local minimum at N250 at all, causing the peak searching algorithm to fail. This may account for the particularly large latency errors observed in that version of the algorithm. Notably, in all DBA variations other than $DBA_{DTW}$, the N250 appears as a deep trough, with a size comparable to that of the N170, which is typically the most prominent negative feature in ERP waveforms (Luck, 2005).

**Discussion**

The results support the hypothesis that DBA can be used to calculate more accurate EEG consensus sequences than ERP, at least insofar as the measures used in this study are concerned. Peak amplitude errors were very large in ERP, and DBA provided a statistically significant improvement; peak latency errors also showed improvements, but they were not statistically significant, so further research with larger data sets will be required to confirm their existence or lack thereof.

It was also confirmed that initial sequence selection can play a significant role in the final outcome of DBA averaging, contrary to the suggestion of the original authors (Petitjean et al., 2011).

In particular, $DBA_{LEN}$ outperformed all other tested algorithms in mean amplitude error of peak latency, root mean square error of peak latency, and mean amplitude error of peak amplitude. Therefore, measuring path length when aligning each input sequence with the ERP seems to be the best approach for selecting an initial sequence when working with EEG data. Yet, $DBA_{LEN}$ was outperformed by $DBA_{DEV}$ in terms of the root mean square error of peak amplitude, indicating that a more complicated measure (the integral of deviation from the diagonal in the warping path) may provide sequences with more consistent amplitudes, and in certain cases this may produce better results than $DBA_{LEN}$.

Arbitrary initial sequences and initial sequences selected by their standard DTW score against the ERP have been shown to produce visually apparent deviations from the standard waveform outlined by ERP and confirmed by $DBA_{LEN}$ and $DBA_{DEV}$, and may not be appropriate choices for future work involving EEG analysis via DBA. Increased numerical errors relative to $DBA_{LEN}$ and $DBA_{DEV}$ confirm this visually-guided conclusion. $DBA_{ERP}$, which directly uses the ERP as its starting sequence, did not perform as well as $DBA_{LEN}$ or $DBA_{DEV}$, but it did not perform particularly poorly in any situation either.

Shortcomings of the automated local peak searching algorithm used for analysis may have confounded results, and future research should test these findings with more robust peak-finding methods, perhaps including manual peak identification. Additional component features such as onset-latency may prove useful for quantifying consensus sequence accuracy as well,

though measurements such as component area should be avoided, since they are often conserved even after a component has been flattened by ERP averaging (Luck, 2005). Results may also vary from component to component, so future research should apply these methods to more than just the N250 component.

On a physiological note, the results of this study seem to suggest that the N250 component is much larger than typically seen in ERP, appearing as large as the N170 in most of the DBA consensus sequences. This may be a special case, since the nature of this data set gives it an especially pronounced N250 component, but it might be speculated that a large N250 component exists in many other data sets in which it was not previously observed, due to cancellation during the ERP averaging process.

In conclusion, DTW barycenter averaging might prove to be a very useful technique in future EEG analysis. Modifications to the initial sequence selection process outlined in this article, as well as utilization of the noise-reducing modification proposed by Kotas et al. (2015), make consensus sequences calculable from EEG data, and these consensus sequences provide quantifiable improvements over current EEG analysis techniques. If these results are confirmed by future studies, variations of DBA might provide cognitive scientists with a more accurate conception of the prototypical electrophysiological response to various external events. Furthermore, the preservation of components that are otherwise canceled during ERP analysis may allow researchers to uncover significant effects that were previously obscured. Undoubtedly, the ability to calculate accurate consensus sequences of EEG data would be of huge benefit to future electrophysiological studies.

**References**

Aarabi, A., Kazemi, K., Grebe, R., Moghaddam, H. A., & Wallois, F. (2009). Detection of EEG transients in neonates and older children using a system based on dynamic time-warping template matching and spatial dipole clustering. NeuroImage, 48(1), 50-62.

Assecondi, S., Bianchi, A., Hallez, H., Staelens, S., Casarotto, S., Lemahieu, I., & Chiarenza, G. (2009). Automated identification of ERP peaks through Dynamic Time Warping: An application to developmental dyslexia. Clinical Neurophysiology, 120(10), 1819-1827.

Casarotto, S., Bianchi, A., Cerutti, S., & Chiarenza, G. (2005). Dynamic time warping in the analysis of event-related potentials. IEEE Eng. Med. Biol. Mag. IEEE Engineering in Medicine and Biology Magazine, 24(1), 68-77.

Curran, T. (N.d.). [EEG analysis of trained vs. untrained visual classification]. Unpublished data, Department of Psychology and Neuroscience, University of Colorado at Boulder.

Gupta, L., Molfese, D., Tammana, R., & Simos, P. (1996). Nonlinear alignment and averaging for estimating the evoked potential. IEEE Transactions on Biomedical Engineering IEEE Trans. Biomed. Eng., 43(4), 348-356.

Huang, H., & Jansen, B. (1985). EEG waveform analysis by means of dynamic time-warping. International Journal of Bio-Medical Computing, 17(2), 135-144.

Keogh, E. J., & Pazzani, M. J. (2001). Derivative Dynamic Time Warping. Proceedings of the 2001 SIAM International Conference on Data Mining, 1-11.

Kotas, M., Leski, J. M., & Moroń, T. (2015). Dynamic Time Warping Based on Modified Alignment Costs for Evoked Potentials Averaging. Advances in Intelligent Systems and Computing Man–Machine Interactions 4, 305-314.

Luck, S. (2005). An introduction to the event-related potential technique (2nd ed.). Cambridge, Mass.: MIT Press.

Myers, C. S., & Rabiner, L. R. (1981). A Comparative Study of Several Dynamic Time-Warping Algorithms for Connected-Word Recognition. Bell System Technical Journal, 60(7), 1389-1409.

Petitjean, F., Ketterlin, A., & Gançarski, P. (2011). A global averaging method for dynamic time warping, with applications to clustering. Pattern Recognition, 44(3), 678-693.

Petitjean, F., & Gançarski, P. (2012). Summarizing a set of time series by averaging: From Steiner sequence to compact multiple alignment. Theoretical Computer Science, 414(1), 76-91.

Picton, T., Hunt, M., Mowrey, R., Rodriguez, R., & Maru, J. (1988). Evaluation of brain-stem auditory evoked potentials using dynamic time warping. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 71(3), 212-225.

Sakoe, H., & Chiba, S. (1990). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. Readings in Speech Recognition, 159-165.

Scott, L. S., Tanaka, J. W., Sheinberg, D. L., & Curran, T. (2008). The role of category learning in the acquisition and retention of perceptual expertise: A behavioral and neurophysiological study. Brain Research, 1210, 204-215.

Wang, K., Begleiter, H., & Porjesz, B. (2001). Warp-averaging event-related potentials. Clinical Neurophysiology, 112(10), 1917-1924.

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research Clim. Res., 30, 79-82.