Clinical Neurophysiology xxx (2009) xxx–xxx

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Automated identification of ERP peaks through Dynamic Time Warping: An application to developmental dyslexia

Sara Assecondi [a,*], A.M. Bianchi [b], H. Hallez [a], S. Staelens [a], S. Casarotto [b,c], I. Lemahieu [a], G.A. Chiarenza [d]

[a] *Ghent University, Department of Electronics and Information Systems, MEDISIP-IBBT-IbiTech, De Pintelaan 185, B-9000 Ghent, Belgium*
[b] *Politecnico di Milano, Department of Biomedical Engineering, IIT Unit, Milan, Italy*
[c] *Università degli Studi di Milano, Department of Clinical Sciences “L. Sacco”, Milan, Italy*
[d] *Azienda Ospedaliera G. Salvini, Rho Hospital, Department of Child and Adolescent Neuropsychiatry, Rho, Italy*

ARTICLE INFO

Article history:
Accepted 20 June 2009
Available online xxxx

Keywords:
Event-related potentials
Automated identification
Peak-picking
Dynamic Time Warping
Developmental dyslexia

ABSTRACT

Objective: This article proposes a method to automatically identify and label event-related potential (ERP) components with high accuracy and precision.
Methods: We present a framework, referred to as peak-picking Dynamic Time Warping (ppDTW), where *a priori* knowledge about the ERPs under investigation is used to define a reference signal. We developed a combination of peak-picking and Dynamic Time Warping (DTW) that makes the temporal intervals for peak-picking adaptive on the basis of the morphology of the data. We tested the procedure on experimental data recorded from a control group and from children diagnosed with developmental dyslexia.
Results: We compared our results with the traditional peak-picking. We demonstrated that our method achieves better performance than peak-picking, with an overall precision, recall and *F*-score of 93%, 86% and 89%, respectively, versus 93%, 80% and 85% achieved by peak-picking.
Conclusion: We showed that our hybrid method outperforms peak-picking, when dealing with data involving several peaks of interest.
Significance: The proposed method can reliably identify and label ERP components in challenging event-related recordings, thus assisting the clinician in an objective assessment of amplitudes and latencies of peaks of clinical interest.

1. Introduction

Cognitive brain potentials are very useful and fascinating because they allow us to explore in a non-invasive way the higher cognitive functions that determine the development of human behaviour and thoughts. The study of such phenomena is currently limited by several factors, such as the presence of artefacts on the recordings, the low signal-to-noise ratio and the marked inter- and intra-individual variability of the potentials. These troubles are intrinsic to biological systems and above all to the ones involved in cognitive processes. However, they can be faced by applying proper mathematical approaches and models.

Event-Related Potentials (ERPs) are characterised by peaks and troughs, the amplitudes and latencies of which may represent quantitative measures useful to track functional neuronal mechanisms. Absence of some peaks or changes in amplitude and latency that are significant with respect to a control population, or a control condition, may be a signature of the presence of a specific group or dysfunction of the brain.

The manual quantification of event-related potentials continues to be the usual clinical practice and constitutes the gold standard reference for evaluating the efficacy of automatic methods. It is surely hard to replace the competency of an experimenter trained by years of clinical practice with a mathematical model. However, the identification of peaks and troughs is often doubtful because the marked variability of cognitive middle and late latency components can produce different views even in skilled experimenters. Therefore, the employment of mathematical approaches is important not only to reduce the analysis time but also to make the analysis results unequivocal and more reliable.

The techniques proposed in the literature to automatically score ERPs can be grouped into two categories: methods that assume a linear latency jitter between ERPs from different subjects (all the peaks have the same latency variability), and methods that assume a non-linear latency jitter (allowing different level of variability depending on the latency of the peak). The problem can be reduced to finding the optimal alignment between two time series. In the first case, that is when only a linear shift is allowed, the jitter

[*] Corresponding author. Tel.: +32 9 332 43 26.
E-mail address: sara.assecondi@gmail.com (S. Assecondi).

1388-2457/$36.00 © 2009 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.clinph.2009.06.023

between time series can be estimated using matched filters (Woody, 1967) or maximum-likelihood approaches (Jaśkowski and Verleger, 1999; Pham et al., 1987) but the inter-subject variability, reflected in local contractions or extensions of the time axis, is not taken into account. In the second case, that is when the mapping is assumed to be non-linear, other techniques have been developed. In peak-picking (Gratton et al., 1989; Derbyshire et al., 1967), a positive (or negative) peak is identified as the global maximum (or minimum) in a pre-defined time interval. The main assumption of this method is that each time range must contain only one peak. However, the time ranges are usually chosen on the basis of a normal population and they may not be able to deal with the large inter-subject variability of cognitive ERPs.

Techniques for non-linear time warping (Casarotto et al., 2005; Wang et al., 2001; Gupta et al., 1996; Picton et al., 1988) have been developed to find the optimal alignment between curves. Dynamic Time Warping (DTW) is a technique originally developed in the field of speech processing (Sakoe and Chiba, 1978): it allows a non-linear mapping of the time axis of two series, according to morphological characteristics, rather than to time latencies in the signals. The two time series are mapped onto a common time axis, giving a time correspondence between samples of the reference and samples of the signal. If meaningful peaks are known on the reference, they are, therefore, automatically identified on the signal as well.

It has already been proven that time warping approaches are superior to methods assuming a linear jitter between subjects and that DTW is superior to peak-picking (Wang et al., 2001; Jaśkowski and Verleger, 1999) for the detection of prominent components, such as the P300. However, to the authors' knowledge, only Casarotto et al. (2005) have used DTW for automatic and simultaneous identification and labelling of several peaks. The main limitation of the work of Casarotto et al. (2005) is the use of an average of several control subjects as a reference. In this case, as a consequence of the physiological variability of the waveforms, some features may not be visible on the reference, and peaks may be systematically missed.

The aim of this article is to propose a method to automatically measure and label ERP components, that, on the one hand, uses *a priori* knowledge of the ERP under investigation in the DTW constraints as well as in the computation of the reference signal; on the other hand, it integrates two previously introduced independent approaches, achieving high accuracy and precision.

We developed a framework, referred to as ppDTW, based on the integration of DTW and peak-picking that makes the temporal interval for peak-picking adaptive on the basis of the morphology of the data. We tested the procedure on experimental data recorded from normal children and children diagnosed with developmental dyslexia, a neuro-behavioral disorder characterised by a specific reading disability (World Health Organization, 2007). These recordings are particularly challenging because of the number of peaks on the ERP and because of the inter-subject variability, even more accentuated when dealing with children and with cognitive functions. We compared our results to the classical peak-picking method, in terms of their ability to discriminate between several different peaks, elicited by the same stimulus. We show that our hybrid method outperforms peak-picking, when dealing with data involving several peaks.

2. Methods

2.1. Experimental data

In this article, two samples are considered, namely a control group of normal children, and children diagnosed with developmental dyslexia. Data are summarised in Table 1. Each subject performed three different reading tasks: Letter Presentation (LP) and Symbol Presentation (SP), consisting in passive observation of letters or symbols displayed on a screen, respectively, and Letter Recognition (LRE), consisting in reading aloud single letters. For a detailed description of the experimental protocol, we refer the reader to Casarotto et al. (2004). All subjects were previously informed of the experimental procedure and written consent was obtained from parents and children. The entire experimental procedure was approved by the ethic committee of the hospital (Azienda Ospedaliera G. Salvini, Rho Hospital, Rho, Italy). The control group is a subset of the normal population used to derive *a priori* information while the dyslexic group is an independent set.

Associated with the cognitive stages elicited by the stimulus, a total of 11 peaks are identified on the ERP (Chiarenza and Casarotto, 2004), as shown in Fig. 1. According to their latencies, these peaks can be grouped in the early, middle and late latency components. The associated variability is shown in Fig. 2, panel C.

The EEG was recorded through 10 leads located according to the modified 10–20 international system. The ElectroOculoGram (EOG) for the evaluation and removal of ocular movements and blinks was also recorded. The signals were bandpass filtered between 0.02 Hz and 30 Hz and sampled at 250 Hz. The ocular and blinking artifacts were removed by means of principal component analysis, as described in Casarotto et al. (2004). The EEG was then segmented into epochs of 4 s (2 s pre- and 2 s post-stimulus), averaged to obtain the average ERP for the single subject. At least 51 trials were averaged for each subject. For further analysis, the digitised averaged signals were subsampled at 62.5 Hz.

2.2. Peak-picking

In the peak-picking method (Gratton et al., 1989), a positive (or negative) peak is identified as the global maximum (or minimum) in a pre-defined time interval. The main assumption for this method is that each time range can contain only one peak. The time ranges are usually chosen by an expert, on the basis of personal experience and knowledge of the phenomenon under investigation, thus being highly subjective and invariant throughout subjects. This last assumption may be violated by the intrinsic inter-subject variability of cognitive responses. The labelling of the peaks is achieved by assigning a specific label to each fragment. In this implementation, time intervals are defined as twice the mean latency variability for the considered peak, across channels and across tasks, as measured in a control sample, as shown in Fig. 2, panel B, lower part and panel C.

2.3. Dynamic Time Warping

Time warping (Cohen, 1986; Sakoe and Chiba, 1978) is a procedure in which the time scales of two signals are stretched or

Table 1
Data description. For each task (Letter Presentation (LP), Symbol Presentation (SP), Letter Recognition Externally Paced (LRE)) and for each group (control, dyslexic) the number of subjects in each group, the mean number of trials averaged for each subject (with the standard deviation) and the mean age for the subjects in the group (with the standard deviation) are reported.

Group	Number of subjects	Age (years)	Task	Number of trials	
		$\mu \pm \sigma$		$\mu \pm \sigma$	min
Control	26	9.6 ± 0.7	LP	92 ± 18	69
			SP	96 ± 19	68
			LRE	114 ± 32	76
Dyslexic	23	9.5 ± 0.7	LP	104 ± 30	71
			SP	103 ± 26	68
			LRE	124 ± 35	72

Please cite this article in press as: Assecondi S et al. Automated identification of ERP peaks through Dynamic Time Warping: An application to developmental dyslexia. Clin Neurophysiol (2009), doi:10.1016/j.clinph.2009.06.023

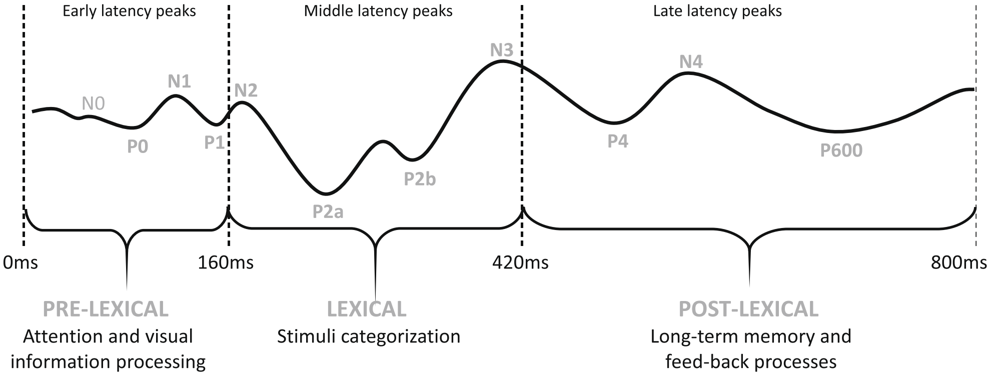


Fig. 1. Example of a reading-related ERP: physiologically meaningful peaks are divided according to their latency and the stage of the cognitive process they are related to (Chiarenza and Casarotto, 2004).

shrunk in order to reduce distortions due to normal morphological differences in the time bases of each waveform.

Let us consider two digitised time series: a reference pattern $\boldsymbol{t}(i)$, $i = 1, \ldots, I$ and a sample pattern $\boldsymbol{s}(j)$, $j = 1, \ldots, J$, where i and j are the digitised time bases of $\boldsymbol{t}$ and $\boldsymbol{s}$, respectively. The aim of time warping is to find an optimal mapping of the two time axes i and j onto a common time axis k, in such a way that a given distance measure is minimised.

The warping function F is defined as a sequence of points $\boldsymbol{c}(k) = (i(k), j(k))$ in the (ij) plane, as shown in Fig. 3, panel C, right. Each point of F gives the matching of the point $i(k)$ on the time axis of $\boldsymbol{t}(i)$ and the point $j(k)$ on the time axis of $\boldsymbol{s}(j)$ and represents the optimal mapping of the two time axes. If there was a linear relation between the sequences, their mapping function would be $i(k) = \frac{I}{J} j(k)$. However, F is usually not linear and some constraints must be imposed in order to avoid meaningless paths.

2.3.1. Constraints

First of all, in the case of ERPs, the first and last points of the reference $\boldsymbol{t}(i)$ and the sample $\boldsymbol{s}(j)$ must coincide. Secondly, the two functions $i(k)$ and $j(k)$ must be continuous and monotonically increasing from point $(1,1)$ to point (I,J), as the time axis is.

The *p-parameter* is a local constraint that defines the minimum (maximum) amount of allowed expansion (1/compression) of the time axis, thus constraining the possible path of F. The *p-parameter* is defined as the ratio between the number of steps in the diagonal direction and the number of steps in horizontal or vertical direction. Fig. 3, panel B, shows the possible local paths when $p = 1$.

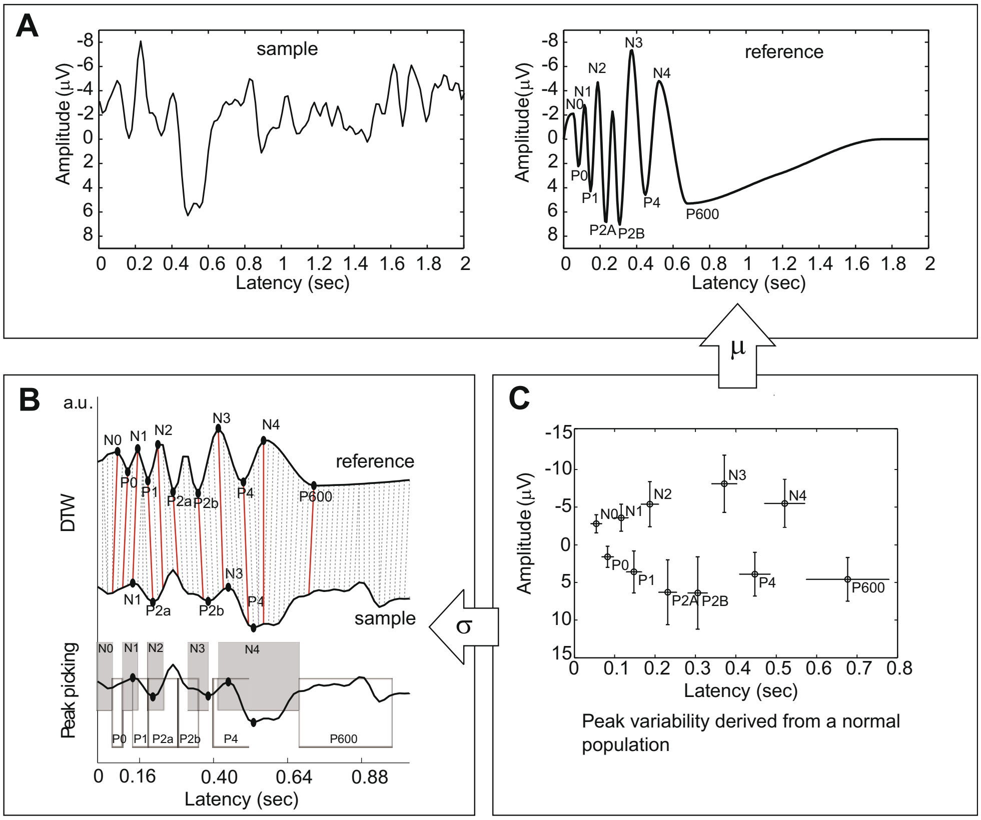


Fig. 2. Flowchart describing the method. (A) The reference signal and a sample signal. (B) The sample and the reference are aligned by means of DTW, then peak-picking is applied in interval adaptively shifted in time, according to the morphology of the data. (C) *A priori* information (mean and standard deviation of amplitudes and latencies of peaks, as measured in a normal population) is used to obtain a reference signal and to define parameters of ppDTW (i.e., warping window and interval for peak-picking).

Please cite this article in press as: Assecondi S et al. Automated identification of ERP peaks through Dynamic Time Warping: An application to developmental dyslexia. Clin Neurophysiol (2009), doi:10.1016/j.clinph.2009.06.023

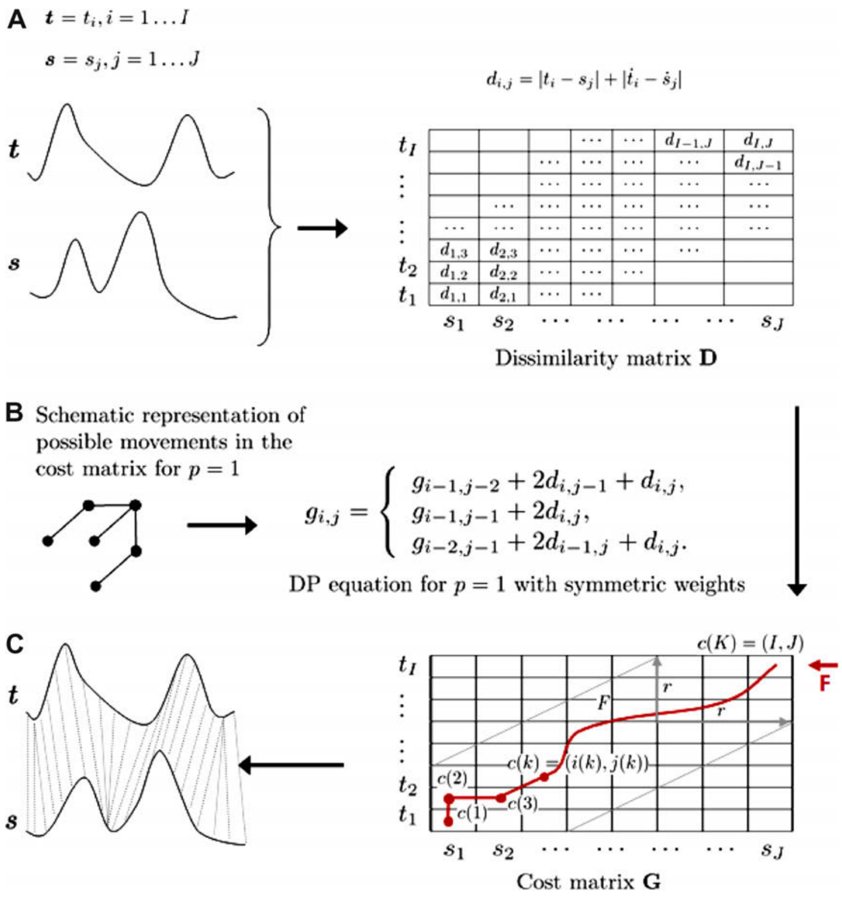

A
$t = t_i, i = 1 \ldots I$
$s = s_j, j = 1 \ldots J$
t
s
$d_{i,j} = |t_i - s_j| + |\dot{t}_i - \dot{s}_j|$
t_I
t_2
t_1
s_1 s_2 ⋯ s_J
$d_{1,1}$ $d_{1,2}$ $d_{1,3}$ $d_{2,1}$ $d_{2,2}$ $d_{2,3}$ $d_{I-1,J}$ $d_{I,J}$ $d_{I,J-1}$
Dissimilarity matrix D
B
Schematic representation of possible movements in the cost matrix for $p = 1$
$g_{i,j} = \begin{cases} g_{i-1,j-2} + 2d_{i,j-1} + d_{i,j}, \\ g_{i-1,j-1} + 2d_{i,j}, \\ g_{i-2,j-1} + 2d_{i-1,j} + d_{i,j}. \end{cases}$
DP equation for $p = 1$ with symmetric weights
C
t
s
$c(K) = (I, J)$
F
r
r
$c(k) = (i(k), j(k))$
$c(2)$
$c(1)$
$c(3)$
t_I
t_2
t_1
s_1 s_2 ⋯ s_J
Cost matrix G

Fig. 3. Dynamic Programming Algorithm: (A) Dissimilarity matrix D. (B) Left: paths allowed by a p-value of 1; right: DP equation used to compute the Cost matrix G. (C) Inside the warping window, limited by r, the warping function F (red line) is found, giving the optimal mapping of the two time axes i and j. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

The *r-parameter* is a global constraint that limits the possible displacements of the warping function in the (i,j) plane, with respect to the diagonal, due to the fact that some points in the (i,j) plane are too far apart to possibly generate a meaningful warping (i.e., points at the beginning of the reference and points at the end of the sample). A simple warping window, shown in Fig. 3, panel C, is defined by a constant r, called window length, as follows (Sakoe and Chiba, 1978):

$$|i(k) - j(k)| \leqslant r. \tag{1}$$

2.3.2. Distance metric

As mentioned above, the optimum warping function minimises a certain distance measure between time series. The distance measure in DTW is very important and must reflect the properties common to the features one wants to match, such as amplitude and shape (i.e., derivative). A possible distance metric, used in this work, accounting for amplitude and slope of the signal, is defined as follows:

$$d(i,j) = |t(i) - s(j)| + |\dot{t}(i) - \dot{s}(j)|, \tag{2}$$

where the first term is the absolute value of the difference between the amplitude of $\boldsymbol{t}$ and $\boldsymbol{s}$ and the second term is the absolute value of the difference between the first derivative of $\boldsymbol{t}$ and $\boldsymbol{s}$.

2.3.3. Dynamic programming

Once all the parameters (distance metric, *p-parameter*, and *r-parameter*) are defined, the optimal path has to be found. Sakoe and Chiba (1978) proposed Dynamic Programming (Bellman and Dreyfus, 1962) as an efficient way of solving the problem of path finding.

DTW is described in Fig. 3. First, a *Distance Matrix* $\mathbf{D} = [d(i,j)]$ is calculated, according to Section 2.3.2. Then, from $\mathbf{D}$ a *Cost Matrix* $\mathbf{G} = [g(i,j)]$ is derived as follows:

$$g(i,j) = \min_k \sum_{k=1}^{K'} d(i(k), j(k)) \cdot w(k), \tag{3}$$

where $g(i,j)$ is the minimum cost necessary to align the first i points of the reference with the first j points of the sample, K' is the length of the path associated to the minimum cost, $w(k)$ are weighting coefficients to avoid bias towards the diagonal.

The warping function F is found by searching in $\mathbf{G}$ for the minimum cost path that joins the upper right corner (point (I,J)) with the lower left corner (point $(1,1)$) and it also is the path associated to the minimum distance.

2.4. Algorithm

A flowchart describing the proposed method is shown in Fig. 2. The procedure is applied channel-wise and involves the following steps:

1. A reference signal is calculated as the interpolation of mean amplitudes and latencies, across channels and across tasks, of those peaks one wants to identify on the signals, as derived from a normal population. The reference is shown in Fig. 2, panel A, right.

2. The single-channel ERP is aligned to the reference, by means of DTW. The r parameter, that defines the width of the warping window, is chosen as the minimum inter-peak latency between two consecutive negative (or positive) peaks in normal cognitive ERPs, that is, 100 ms (Luck, 2005). The p is empirically chosen equal to 1, as in Casarotto et al. (2005). The warping is shown in Fig. 2, panel B, upper part.
3. The alignment of the reference with the signal produces a temporal correspondence between samples on the reference and samples on the signal. Therefore, physiologically relevant peaks identified on the reference are automatically identified on the signal as well, as shown by the red lines in Fig. 2, panel B.
4. In order to refine the identification, an *a posteriori* peak-picking is applied. For each peak identified on the signal by the warping, a symmetric temporal window around the peak, of the same width as the searching windows used in traditional peak-picking, as shown in Fig. 2, panel B, lower part, is defined and a search for maxima (or minima) is performed in that interval. If a maximum (or a minimum) is actually present in the temporal window, the point is marked as a positive (or negative) peak, otherwise the peak is considered missing.
5. The procedure is then repeated for the other ERP channels.

2.5. Evaluation

In order to quantify and evaluate the performance of the aforementioned methods, the proposed ppDTW and the traditional peak-picking are compared. The automatic scoring on experimental data are compared with the scoring of an expert clinician (G.A.C.) and the following quantities are defined (Makhoul et al., 1999):

- correct identification (C): a peak is identified at the same latency by the expert and by the method;
- substitution (S): a peak is identified at a different latency by the expert and by the method;
- deletion (D): a peak is identified by the expert but not by the method;
- insertion (I): a peak is identified by the method but not by the expert;
- total number of peaks in the reference (N); and
- total number of measured peaks (M).

From the above quantities, precision (P), recall (R) and F-score (F) are calculated as follows:

$$\begin{aligned}
\mathrm{P} &= \frac{C}{C+S+I} = \frac{C}{M} \\
\mathrm{R} &= \frac{C}{C+S+D} = \frac{C}{N} \qquad (4) \\
\mathrm{F} &= \left[\frac{1}{2P}+\frac{1}{2R}\right]^{(-1)} = \frac{2PR}{P+R}
\end{aligned}$$

where $N = C+S+D$ and $M = C+S+I$. Precision depends on the insertions or false positives, while recall depends on the deletions, or false negatives. The F-score is the weighted harmonic mean of precision and recall and it represents the global performance when the same importance is given to both precision and recall. The F-score is related to all the possible identification errors (insertions, deletions and substitutions). Precision, recall and F-score have values between 0 and 1.

Using SPSS 15.0 (SPSS Inc., Chicago, IL, USA), we performed an analysis of variance (ANOVA) investigating the main and interaction effects of Method, Task, Latency and Group on precision, recall and F-score. For the sake of clarity and to maintain the focus on the performance of the two methods, in the following, we only show and discuss results significant in at least one of the performance measures. We applied a ($2 \times 2 \times 3 \times 3$) ANOVA with 3 repeated-measurement factors (Method (pp, ppDTW), Task (LP, SP, LRE) and Latency Range (early, middle and late latency components) and 1 between-group factor (normal and dyslexic)). The Greenhouse–Geisser correction was applied to the Task and Latency Range effects. Significant factors were then tested by means of a paired two-tailed Student's t test and the absolute t-values are reported. Differences are labelled as significant, very significant or highly significant, when $p < 0.05$, $p < 0.01$ and $p < 0.001$, respectively.

Since ERP peaks are considered clinically relevant only when their peak-to-peak amplitude exceeds the background noise, we calculate precision and recall only for peaks exceeding the mean peak-to-peak amplitude in the baseline plus one standard deviation.

3. Results

In this section, we present the performance of the two methods in terms of precision, recall and F-score, as achieved when the automated scoring is compared to the scoring of an expert clinician (gold standard).

Fig. 4 shows an example of the performance on one ERP channel for different tasks and different pathologies. The black arrows mark disagreement between the automatic scoring and the visual scoring of an expert clinician. We see that, even with noisy time series, ppDTW is able to identify peaks that are missed by peak-picking.

The significant interactions and main effects are reported in Table 2 while the marginal means and follow-up analysis by means of paired Student's t-tests are reported in Table 3.

Main effects: The main effects for precision, recall and F-score are shown in Fig. 5. We found a main significant effect of Method and Task in recall and F-score. A main significant effect of Latency was found in precision, recall and F-score. No main effect was found for Group for any of the measures.

- *Method:* Methods differed for recall and F-score, ppDTW being better than peak-picking, while they did not differ for precision.
- *Task:* Recall and F-score were worse for the LP and the LRE tasks than for the SP task and did not differ between the LP and LRE tasks. The task factor did not affect the precision.
- *Latency:* Precision, recall and F-score were worse for middle and late latencies than for early latencies. Precision was worse for the late compared to the middle latencies. Recall was worse for the middle than for the late latencies. F-score did not differ between middle and late latencies.

Interactions: We found a significant interaction between Method and Latency for all the measures, between Task and Latency for recall and F-score and between Method, Latency and Group for precision.

- *Method* × *Latency:* The interaction of Method and Latency is shown in Fig. 6. Table 4 reports the marginal means and the t-statistic, when pairwise t-tests were performed on the significant interaction factors. Considering precision, methods differed for early latencies only, with ppDTW being worse than peak-picking. Considering recall and F-score, though ppDTW being better than peak-picking for all latencies (given a significant main effect of Method), these effects were more evident at the middle latencies.

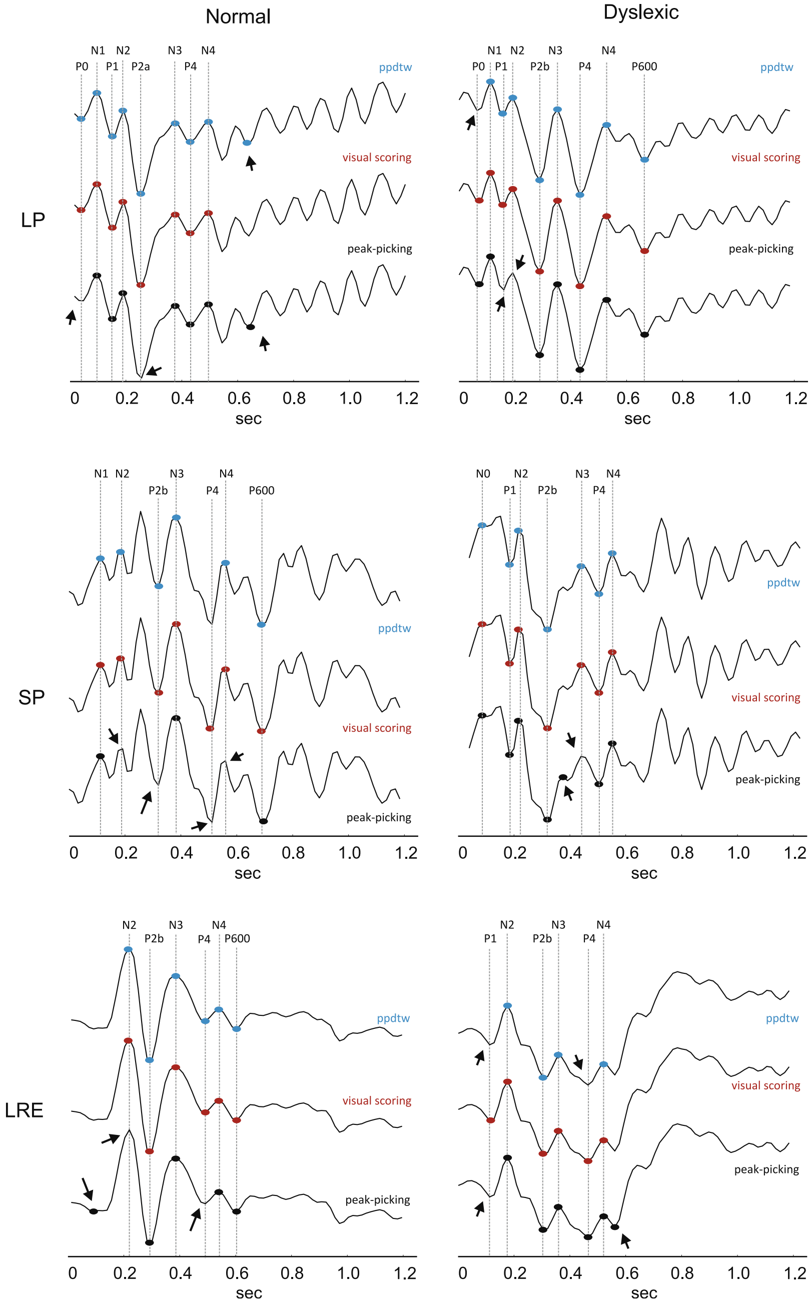


Fig. 4. Performance on one ERP channel for different tasks and different pathologies. The black arrows mark disagreement between the automatic scoring and the visual scoring of an expert clinician.

- *Task × Latency:* Recall at the middle latencies was worse for the LP task than for the SP and LRE tasks and did not differ between the SP and the LRE tasks. At the late latencies, recall was worse for the LRE compared to the SP and LP tasks. *F*-score at the middle latencies was worse for the LP than for the SP and the LRE tasks and did not differ between the SP and LRE tasks. At the late latencies, the *F*-score was worse for the LRE task than for the SP task and did not differ between the SP and LP tasks nor between the LP and LRE tasks.
- *Method × Latency × Group:* With a follow-up analysis, reported in Table 5, we found in the normal Group only a main significant effect of Latency, precision being worse for late latencies than for early and middle latency and not differing between early and middle latencies. In the dyslexic Group, we found a significant main effect of Latency, with precision being worse for late and middle latencies than for early latencies and for late than for middle latencies. We also found a significant interaction of Method and Latency, with ppDTW being worse than peak-picking for early latencies and better for late latencies. The methods did not differ at middle latencies.

4. Discussion

The development of automated methods for the identification of ERP components is a complex and delicate problem. Because of the inter-subject variability, an experienced experimenter is required to identify peaks in a consistent and homogeneous way across subjects. The inter-subject variability is determined by multiple factors. Firstly, after an external stimulus, certain components

Table 2
Main effects and interactions significant in at least one of the measures. For precision, recall and *F*-score the results of the ANOVA are reported (*F* = *F*-value; df(a,b) = degrees of freedom of the sum of square due to factor (a) and due to error (b); *p*-value). For factors marked by * the Greenhouse–Geisser correction was applied. The significant *p*-values are reported in bold.

Measure	Factor	F	df(a,b)	*p*-Value
Precision	Group	1.34	1,47	0.254
	Method	0.25	1,47	0.623
	Latency*	99.13	2,94	**0.000**
	Task*	1.89	2,94	0.158
	Method × Latency*	3.94	2,94	**0.033**
	Task × Latency*	1.50	4,188	0.220
	Method × Latency* × Group	5.36	2,94	**0.011**
Recall	Group	1.52	1,47	0.223
	Method	121.83	1,47	**0.000**
	Latency*	58.01	2,94	**0.000**
	Task*	4.35	2,94	**0.016**
	Method × Latency*	63.03	2,94	**0.000**
	Task × Latency*	3.36	4,188	**0.017**
	Method × Latency* × Group	0.04	2,94	0.953
F-score	Group	2.03	1,47	0.161
	Method	108.34	1,47	**0.000**
	Latency*	62.67	2,94	**0.000**
	Task*	4.93	2,94	**0.010**
	Method × Latency*	45.29	2,94	**0.000**
	Task × Latency*	3.20	4,188	**0.025**
	Method × Latency* × Group	1.18	2,94	0.308

Table 3
Pairwise comparison of the marginal means for the main effects. For precision, recall and *F*-score the results of the pairwise *t*-test are reported (*t* = *t*-value; df = degrees of freedom). The significant *p*-values are reported in bold.

		Method			
Measure	Pairs	Pairs mean (%)	–	–	–
Precision	pp–ppDTW	93–93	–	–	–
Recall	pp–ppDTW	80–86	–	–	–
F-score	pp–ppDTW	85–89	–	–	–
		Task			
Measure	Pairs	Pairs mean (%)	*t*	df	*p*-Value
Precision	SP–LP	94–92	1.56	48	0.125
	SP–LRE	94–92	1.88	48	0.066
	LP–LRE	92–92	0.03	48	0.977
	SP–LP	85–82	2.74	48	**0.009**
Recall	SP–LRE	85–82	2.38	48	**0.022**
	LP–LRE	82–82	0.35	48	0.727
	SP–LP	88–86	2.60	48	**0.012**
F-score	SP–LRE	88–86	2.72	48	**0.009**
	LP–LRE	86–86	0.31	48	0.757
		Latency			
Measure	Pairs	Pairs mean (%)	*t*	df	*p*-Value
Precision	Early–middle	97–96	2.98	48	**0.005**
	Early–late	97–86	10.81	48	**0.000**
	Middle–late	96–86	10.02	48	**0.000**
Recall	Early–middle	91–76	13.47	48	**0.000**
	Early–late	91–82	6.55	48	**0.000**
	Middle–late	76–82	3.61	48	**0.001**
F-score	Early–middle	93–84	13.00	48	**0.000**
	Early–late	93–83	10.00	48	**0.000**
	Middle–late	84–83	0.87	48	0.390

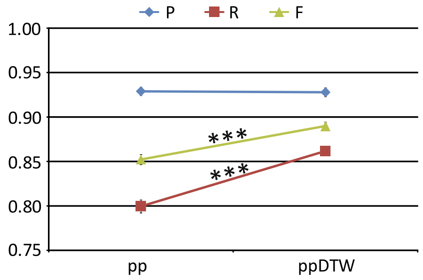
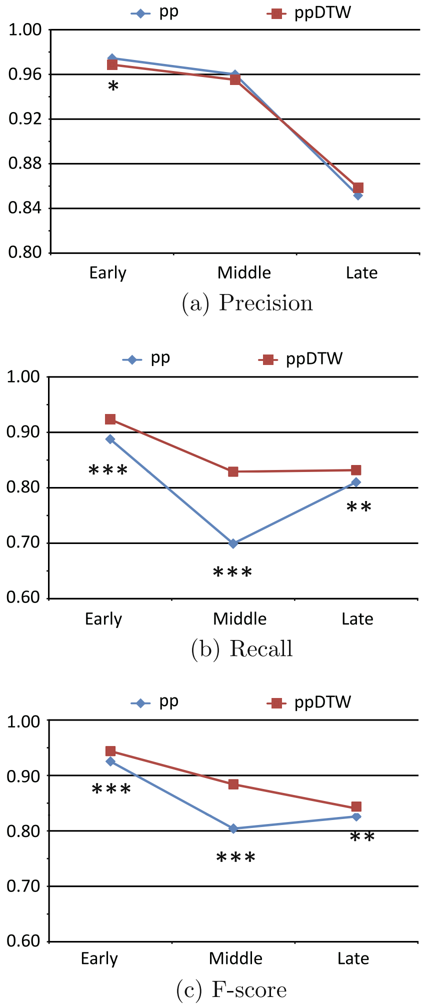


Fig. 5. Comparison of Precision, Recall and *F*-score for method. Significant differences are marked by stars ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$).

may not be elicited and their amplitudes, latencies and scalp topographies may vary depending on the task used. This is especially relevant with cognitive tasks. Secondly, the psychological state of the subject during the experiment and the physiological individual differences are additional sources of variability. This variability is even more pronounced in children, where the maturation of the brain and of the psychological functions plays an important role in determining polarity, amplitude and latency of cognitive ERP components. Lastly, the inter-subject variability is further increased by the comparison of healthy and pathological subjects, especially in children. Any procedure to automatically identify ERP peaks must, therefore, take all these factors of variability into account and translate in a mathematical model the experience of years of ERP analysis and scoring.

In this article, a method for automated identification and labelling of ERP components is proposed and compared against the more traditional peak-picking (Gratton et al., 1989; Derbyshire et al., 1967). The performance of both methods is evaluated in



(a) Precision

(b) Recall

(c) F-score

Fig. 6. Effect of the interaction between Method and Latency on the performance of ppDTW and peak-picking. Significant differences are marked by stars ($^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$). (a) Precision; (b) recall; and (c) *F*-score.

Table 4
Pairwise comparison of the marginal means for the interaction effects, significant in at least one of the measures. For precision, recall and *F*-score the results of the pairwise *t*-test are reported (*t* = *t*-value; df = degrees of freedom). The significant *p*-values are reported in bold.

	Method × Latency				
Measure	Pairs	Pairs mean (%)	*t*	df	*p*-Value
Precision	(pp,early)–(ppDTW,early)	98–97	2.58	146	**0.011**
	(pp,middle)–(ppDTW,middle)	96–96	1.42	146	0.158
	(pp,late)–(ppDTW,late)	85–86	1.60	146	0.111
Recall	(pp,early)–(ppDTW,early)	89–92	5.81	146	**0.000**
	(pp,middle)–(ppDTW,middle)	70–83	17.79	146	**0.000**
	(pp,late)–(ppDTW,late)	81–83	3.33	146	**0.001**
F-score	(pp,early)–(ppDTW,early)	93–94	4.97	146	**0.000**
	(pp,middle)–(ppDTW,middle)	80–88	15.01	146	**0.000**
	(pp,late)–(ppDTW,late)	83–84	2.95	146	**0.004**
	Task × Latency				
Measure	Pairs	Pairs mean (%)	*t*	df	*p*-Value
Recall	(SP,early)–(LP,early)	91–90	0.55	97	0.584
	(SP,early)–(LRE,early)	91–91	0.30	97	0.762
	(LP,early)–(LRE,early)	90–91	0.98	97	0.328
	(SP,middle)–(LP,middle)	78–74	4.35	97	**0.000**
	(SP,middle)–(LRE,middle)	78–77	1.20	97	0.234
	(LP,middle)–(LRE,middle)	74–77	2.85	97	**0.005**
	(SP,late)–(LP,late)	85–82	1.59	97	0.114
	(SP,late)–(LRE,late)	85–79	3.59	97	**0.001**
	(LP,late)–(LRE,late)	82–79	1.77	97	0.079
F-score	(SP,early)–(LP,early)	94–93	1.35	97	0.180
	(SP,early)–(LRE,early)	94–94	0.57	97	0.569
	(LP,early)–(LRE,early)	93–94	0.93	97	0.355
	(SP,middle)–(LP,middle)	86–82	3.74	97	**0.000**
	(SP,middle)–(LRE,middle)	86–85	0.81	97	0.419
	(LP,middle)–(LRE,middle)	82–85	2.88	97	**0.005**
	(SP,late)–(LP,late)	86–83	1.84	97	0.070
	(SP,late)–(LRE,late)	86–81	3.77	97	**0.000**
	(LP,late)–(LRE,late)	83–81	1.75	97	0.083

terms of precision and recall, when the automatic scoring is compared to the scoring of an expert clinician, considered as gold standard. The novelty of the proposed method, when compared to previously introduced approaches based on DTW, resides, on the one hand, in the use of *a priori* knowledge of the ERP under investigation in the DTW constraints as well as in the computation of the reference signal. On the other hand, ppDTW integrates the two previously introduced independent approaches: DTW and peak-picking. In our framework, DTW is followed by an *a posteriori* search for maxima and minima. This search is equivalent to applying peak-picking in intervals that are adaptively shifted in time, as determined by the warping with a reference signal, on the basis of the morphology of the data. This characteristic overcomes the main limitation of peak-picking, that is, the fact that the searching interval are pre-defined for all subjects and, therefore, the difficulty of coping with the inter-subject variability of ERPs. The good performance obtained in dyslexic children, when the template is computed from the control group, and the lack of a significant main effect for the factor Group show that *a priori* knowledge is only a little constraint for the method.

As stated above, the alignment of the reference signal with the ERP provides a temporal correspondence between samples of the reference and samples of the ERP. As a consequence, features identified on the reference are automatically identified and labelled on the ERP. Therefore, it is important that the reference presents all the features we are interested in. However, when a reference is defined as the average across subjects, as in Casarotto et al. (2005), some peaks may not be visible on the reference. This is due to the large inter-subject variability, which is characteristic for cognitive ERPs. In this article, we propose an alternative reference signal: mean amplitudes and latencies for those peaks that we want to identify on the signals are derived as the mean amplitude and latency across channels and across tasks, from a normal population. These points are then interpolated, thus obtaining a reference curve that represents the average peaks. The use of such a reference signal avoids the possibility that some peaks are systematically missed because they are not present on the reference. It is worth mentioning that a template can be derived from the literature data, from a proper database, from theoretical models, etc., thus making the method even more general.

The amplitudes and latencies of peaks in the normal population were measured by identifying a time window for each peak and subsequently applying peak-picking. The time window for a given peak was determined by superimposing the single-subject averages and by taking into account the earliest and the latest peaks. Then, the measures obtained by peak-picking were visually inspected by an expert and, if it was the case, corrected. This kind of "informed"

Table 5
Interaction of Method × Latency × Group for precision. The results of the follow-up ANOVA and pairwise *t*-test are reported (*F* = *F*-value; df(a, b) = degrees of freedom of the sum of square due to factor (a) and due to error (b); *p*-value, *t* = *t*-value; df = degrees of freedom). For factors marked by * the Greenhouse–Geisser correction was applied. The significant *p*-values are reported in bold.

				Normal group				
Factor	*F*	df(a,b)	*p*-Value	Pairs	Pairs mean (%)	*t*	df	*p*-Value
Method	4.23	1,25	0.050	–	–	–	–	–
				Early–middle	96–96	0.40	25	0.691
Latency*	54.29	2,50	**0.000**	Early–late	96–84	7.93	25	**0.000**
				Middle–late	96–84	7.80	25	**0.000**
Method × Latency*	0.28	2,50	0.706	–	–	–	–	–
				Dyslexic group				
Factor	*F*	df(a,b)	*p*-Value	Pairs	Pairs mean (%)	*t*	df	*p*-Value
Method	0.67	1,22	0.421	–	–	–	–	–
				Early–middle	98–96	3.43	22	**0.012**
Latency*	49.85	2,44	**0.000**	Early–late	98–87	7.33	22	**0.000**
				Middle–late	96–87	7.17	22	**0.000**
				(pp,early)–(ppDTW,early)	98–97	3.36	22	**0.003**
Method × Latency*	13.05	2,44	**0.000**	(pp,middle)–(ppDTW,middle)	96–95	1.12	22	0.273
				(pp,late)–(ppDTW,late)	86–88	3.65	22	**0.001**

search allowed us to greatly reduce the time necessary to identify the amplitudes and latencies and to only correct the misdetections. The measures identified in this way were then used to construct the template for ppDTW. Therefore, the method is valuable even when dealing with new original paradigms, for which literature data are lacking but some components are expected, because of the intrinsic design and structure of the experiment.

Similarly, the definition of the *r* parameter is based on values measured in a normal population. The use of *a priori* information may guide the user during the choice of the parameters, though small changes of the value of *r* do not significantly affect the performance of the method.

We assume that the template contains all the peaks that could possibly arise on the ERP. However, someone may also be interested in identifying a subset of peaks, representative for the main morphology of the ERP. In this case ppDTW would still be able to cope with the inter-subject variability, better than peak-picking, since ppDTW would still be able to exploit the morphology of the signal.

The performance of automated methods for the identification of ERP components is strongly affected by the inter-subject variability of the latencies of meaningful peaks. This effect is clearly shown by the results of the ANOVA test performed, where a main effect of Task and Latency was found. Better performance is achieved at early latencies, characterised by less variability. For the same reason, better results are obtained in the SP task, in which late, more variable, components are less prominent than in the LP or LRE tasks.

The ANOVA study also showed that recall achieved by ppDTW is always significantly higher than by peak-picking, meaning that ppDTW is less likely than peak-picking to miss a peak. This is due to the fact that the searching intervals for ppDTW are shifted in time, according to the morphology of the data, and thus not fixed as in peak-picking. The overall absence of significant differences in precision is due to the fact that the *a posteriori* search for peaks, that is very efficient, is equivalent in both peak-picking and ppDTW.

Only in early latency peaks the precision achieved by peak-picking is significantly higher than in ppDTW. This is due to the fact that early latency peaks have a small latency variability, thus being less prone to jitter due to inter-subject variability. In peak-picking, the searching intervals are defined *a priori*: once they are carefully chosen, peak-picking is likely to give very good performance with peaks with small latency variability. However, some flexibility, that is a trade-off between very high performance at the early latencies or good performance at both the early and the late latencies must be given to ppDTW in order to deal at the same time with peaks with small and large latency variability.

5. Conclusion

We proved that taking into account the morphology of the data, prior to peak-picking significantly improves the performance of the automated detection method. We demonstrated that the proposed method is successfully applied to ERPs recorded in both normal and dyslexic children, for different tasks and when applied to different latency ranges. We show that our hybrid method achieves the best performance.

Acknowledgments

The authors thank Jonne Oldenburg and Ruth Seurinck (Department of Experimental Psychology, Ghent University, Ghent, Belgium) for their contribution to the statistical analysis.

The authors thank the two anonymous referees for their useful comments and suggestions.

References

Bellman R, Dreyfus S. Applied dynamic programming. Princeton (NJ): Princeton University Press; 1962.

Casarotto S, Bianchi A, Cerutti S, Chiarenza G. Principal component analysis for reduction of ocular artefacts in event-related potentials of normal and dyslexic children. Clin Neurophysiol 2004;115(3):609–19.

Casarotto S, Bianchi A, Cerutti S, Chiarenza G. Dynamic time warping in the analysis of event-related potentials. EMB Mag 2005;24(1):68–77.

Chiarenza GA, Casarotto S. Imparare a leggere: i meccanismi psicofisiologici. Quad acp 2004;11(5):212–5.

Cohen A. Biomedical signal processing: compression and automatic recognition, vol. 2; 1986 [Chapter: Time warping].

Derbyshire A, Driessen G, Palmer C. Technical advances in the analysis of single acoustically evoked potentials. Electroenceph Clin Neurophysiol 1967;22:476–81.

Gratton G, Kramer A, Coles G, Donchin E. Simulation studies of latency measures of components of the event-related brain potential. Psychophisiology 1989;26(2):233–48.

Gupta L, Molfese D, Tammana R, Simos P. Nonlinear alignment and averaging for estimating the evoked potential. IEEE Trans Biomed Eng 1996;43(4):348–56.

Jaśkowski P, Verleger R. Amplitudes and latencies of single-trial ERP's estimated by a maximum-likelihood method. IEEE Tran Biomed Eng 1999;46(8):987–93.

Luck S. An introduction to the event-related potential technique. Cambridge (MA): MIT Press; 2005.

Makhoul J, Kubala F, Schwartz R, Weischedel R. Performance measures for information extraction. In: Proceedings of DARPA broadcast news workshop; 1999. p. 249–52.

Pham DN, Möcks J, Kohler W, Gasser T. Variable latencies of noisy signals: estimation and testing in brain potential data. Biometrika 1987;74(3):525–33.

Picton T, Hunt M, Mowrey R, Rodriguez R, Maru J. Evaluation of brain-stem auditory evoked potentials using dynamic time warping. Electroenceph Clin Neurophysiol 1988;71(3):212–25.

Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process 1978;ASSP-26(1):43–9.

Wang K, Begleiter H, Porjesz B. Warp-averaging event-related potentials. Clin Neurophysiol 2001;112:1917–24.

Woody C. Characterization of an adaptive filter for the analysis of variable latency neuroelectrical signals. Med Biol Eng 1967;5:539–53.

World Health Organization. International statistical classification of diseases and related health problems, 10th ed.; 2007.

Please cite this article in press as: Assecondi S et al. Automated identification of ERP peaks through Dynamic Time Warping: An application to developmental dyslexia. Clin Neurophysiol (2009), doi:10.1016/j.clinph.2009.06.023