Robust Common Spatial Patterns based on Bhattacharyya Distance and Gamma Divergence

Stephanie Brandl[1], Klaus-Robert Müller[1,2], *Member, IEEE*, and Wojciech Samek[3], *Member, IEEE*
[1]Department of Machine Learning, Berlin Institute of Technology, Berlin, Germany
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea
[3]Machine Learning Group, Fraunhofer Heinrich Hertz Institute, Berlin, Germany
s_brandl@mailbox.tu-berlin.de, klaus-robert.mueller@tu-berlin.de, wojciech.samek@hhi.fraunhofer.de

***Abstract*— The computation of task-related spatial filters is a prerequisite for a successful application of motor imagery-based Brain-Computer Interfaces (BCI). However, in the presence of artifacts, e.g., resulting from eye movements or muscular activity, standard methods such as Common Spatial Patterns (CSP) perform poorly. Recently, a divergence-based spatial filter computation framework has been proposed which enables significantly more robust computation with respect to artifacts by using Beta divergence. In this paper we integrate two additional divergence measures, namely Bhattacharyya distance and Gamma divergence, into the divergence-based CSP framework and evaluate their robustness using simulations and data set IVa from BCI Competition III.**

I. Introduction

Brain-Computer Interfacing (BCI) [1] [2] serves as a non-muscular communication system which detects brain signals from the electroencephalogram (EEG) and translates them into control commands for a computer device. People affected by diseases such as amyotrophic lateral sclerosis (ALS), brainstem stroke, multiple sclerosis or muscular dystrophies, and especially those who are completely locked-in, could benefit from such technology, since it provides a way to communicate without any muscular control. A popular mental strategy is so-called motor imagery, which provides a system based on imagined movements that cause significant and detectable changes in EEG. The main challenge in constructing a BCI device is extracting relevant features from a high-dimensional EEG to translate brain signals efficiently into control signals [3], [4], [5].

Common Spatial Patterns (CSP) is a well established feature-extraction algorithm in motor imagery-based BCIs [6] [7] that detects synchronization and desynchronization processes and uses them to compute spatial filters which make it possible to discriminate between different imagined movements. Since the original version of CSP is sensitive to nonstationarities, more robust CSP versions have been proposed over the last few years [8] [9] [10] [11] [12]. Considering end users, BCI efficiency still needs to be enhanced so that it can work in an out-of-lab context. One important issue is the robustness of the system. It has recently been shown that CSP can be embedded in a divergence framework using Beta divergence, which is robust to outliers [13][14]. Divergence functions [15] are popular measures of discrepancy between probability distributions. Integrating them into machine learning methods is not entirely new, as different algorithms have already been transformed using divergence functions [16] [17] [18]. In this paper, we integrate Bhattacharyya distance and Gamma divergence into the divergence-based CSP (divCSP) framework and evaluate their robustness.

Detailed derivations and an implementation of the proposed algorithms are available at `www.divergence-methods.org`.

*This work was supported by the by the Federal Ministry of Education and Research (BMBF) under the project Adaptive BCI (FKZ 01GQ1115), by the DFG, by the World Class University Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology, under Grant R31-10008 and by the Brain Korea 21 Plus Program through the National Research Foundation of Korea funded by the Ministry of Education.

II. Divergence-based framework for CSP

Spatial filtering is a common way to detect discriminative features by enhancing the signal-to-noise ratio in motor imagery. As we have mentioned, a well-established spatial filter algorithm is Common Spatial Patterns (CSP). Motor imagery usually comes along with synchronization and desynchronization effects (ERS/ERD) in mu and beta rhythms over the sensorimotor cortex during and after imagined movements [19]. CSP detects those effects by maximizing the variance of class 1 while minimizing the variance of class 2 and vice versa. This problem can be solved by a generalized eigenvalue problem

$$\Sigma_1 w_i = \lambda_1 \Sigma_2 w_i$$

where Σ_1 and $\Sigma_2 \in \mathbb{R}^{\mathrm{DxD}}$ are the average covariance matrices of class 1 and 2, w_i are the spatial filters and λ_i the corresponding eigenvalues. The obtained spatial filters $W = [w_1, w_2, \ldots, w_D]$ then have to be sorted according to their contributing discriminative qualities [14]. In [13] it was shown that the subspace spanned by these spatial filters is equal to the span of filters that maximize the symmetric Kullback-Leibler (KL) divergence between the probability distribution of both classes:

$$\mathrm{span}(W) = \mathrm{span}(V^*)$$
$$V^* = \underset{V}{\mathrm{argmax}}\, \tilde{D}_{kl}\left(\mathcal{N}\left(0, V^\top \Sigma_1 V\right) || \mathcal{N}\left(0, V^\top \Sigma_2 V\right)\right)$$

TABLE I: Overview of the objective functions and properties of CSP divergence methods, with $\bar{\Sigma}_c^i = V^T \Sigma_c^i V$ and $\bar{\Sigma}_c$ denoting the projected covariance matrix of trial i and class c and the class average. Note that $\rho = \frac{1}{\beta}\sqrt{\frac{1}{(2\pi)^{\beta d}(\beta+1)^d}}$.

Method	**Objective** $\sigma(V)$	**robust**	**parameter**
CSP	$\frac{1}{2}\operatorname{tr}((\bar{\Sigma}_1)^{-1}(\bar{\Sigma}_2)) + \frac{1}{2}\operatorname{tr}((\bar{\Sigma}_2)^{-1}(\bar{\Sigma}_1)) - d$	no	no
Beta	$\rho \sum_{i=1}^{n} (\lvert\bar{\Sigma}_1^i\rvert^{-\frac{\beta}{2}} + \lvert\bar{\Sigma}_2^i\rvert^{-\frac{\beta}{2}} - (\beta+1)^{\frac{d}{2}} (\lvert\bar{\Sigma}_2^i\rvert^{\frac{1-\beta}{2}} \lvert\beta\bar{\Sigma}_1^i + \bar{\Sigma}_2^i\rvert^{-\frac{1}{2}} + \lvert\bar{\Sigma}_1^i\rvert^{\frac{1-\beta}{2}} \lvert\beta\bar{\Sigma}_2^i + \bar{\Sigma}_1^i\rvert^{-\frac{1}{2}}))$	yes	β
Bha	$\frac{1}{2}\sum_{i=1}^{n} (\ln(\lvert\bar{\Sigma}_1^i + \bar{\Sigma}_2^i\rvert) - \frac{1}{2}\ln\lvert\bar{\Sigma}_1^i\rvert - \frac{1}{2}\ln\lvert\bar{\Sigma}_2^i\rvert - d\ln(2))$	yes	no
Gamma	$\frac{1}{4\gamma}\sum_{i=1}^{n} (\frac{1}{2}\ln(\lvert\gamma\bar{\Sigma}_1^i + \bar{\Sigma}_2^i\rvert) + \frac{1}{2}\ln(\lvert\bar{\Sigma}_1^i + \gamma\bar{\Sigma}_2^i\rvert) - \ln\lvert\bar{\Sigma}_1^i\rvert - \ln\lvert\bar{\Sigma}_2^i\rvert - d\ln(2))$	yes	γ

where V^* maximizes the symmetric KL divergence between the probability distribution of both classes and $\mathcal{N}(\mu, \Sigma)$ denotes the Gaussian distribution, with mean μ and covariance Σ. Symmetric KL divergence between two continuous probability distributions, $p(x)$ and $q(x)$, is defined as

$$\tilde{D}_{kl} = \int p(x) \log \frac{p(x)}{q(x)} dx + \int q(x) \log \frac{q(x)}{p(x)} dx$$

Since KL divergence is sensitive to outliers, we investigated robust divergence measures to enhance classification accuracy.

III. Robustness and robust discrepancy measures

A. Robustness Property

The goal of a robust CSP algorithm is to reliably compute task-related spatial filters, V, even when data is heavily contaminated. In the divergence framework, robustness can be achieved by decomposing the divergence between the average class distributions into the sum of trialwise divergences and limiting the influence of single (potentially outlier) terms (see [13]). This changes the objective function into

$$V^* = \underset{V}{\operatorname{argmax}} \sum_i \tilde{D}_{kl} \left(\mathcal{N}\left(0, V^\top \Sigma_1^i V\right) \,||\, \mathcal{N}\left(0, V^\top \Sigma_2^i V\right) \right)$$

where Σ_c^i refers to trialwise covariance matrices of class c. Note that this approach assumes a balanced number of trials for each class and also requires robust divergence measures. We call a divergence function, D, *robust* if the inclusion of a single additional outlier trial (e.g. $\Sigma_1^{i^*}$) does not significantly increase the value of the objective function. Mathematically, this translates into stating that the ratio α between the objective functions (with and without the outlier trial) is close to 1, that is

$$\alpha = \frac{\max_V \left\{ \sigma(V) + D\left(\mathcal{N}(0, V^\top \Sigma_1^{i^*} V) || \, \mathcal{N}(0, V^\top \Sigma_2^i V) \right) \right\}}{\max_V \sigma(V)}$$

where $\sigma(V)$ is the corresponding objective function (see Table I). If this ratio is very large (e.g., because it grows linearly or exponentially with $||V^\top \Sigma_1^{i^*} V||$), then the outlier trial has a significant influence on the spatial filter computation. In other words, the algorithm focuses on the outlier trial instead of considering the majority of the 'clean trials'.

B. Beta divergence

Beta divergence has been considered to be a robust replacement in the divCSP algorithm [13]. Beta divergence [20] between two continuous probability distributions is defined (for $\beta > 0$) as

$$D_\beta(p(x) \,||\, q(x)) = \frac{1}{\beta} \int (q^\beta(x) - p^\beta(x)) p(x) dx - \frac{1}{\beta+1} \int (q^{\beta+1}(x) - p^{\beta+1}(x)) dx.$$

The objective function of the Beta divergence CSP algorithm (β-divCSP) is defined in the second row of Table I, where $\tilde{D}_\beta$ denotes the symmetric Beta divergence and Σ_1^i and Σ_2^i are trial-wise covariance matrices of classes 1 and 2, respectively. In the following, we will show that Beta divergence is not the only well-suited robust distance measure in the divCSP framework.

C. Bhattacharyya distance

Bhattacharyya distance, a divergence-type measure between two populations [21], is widely used in various fields, such as computer vision [22], multiclass classification [23] and bayesian classification [24]. The key advantages in using the Bhattacharyya distance in the divCSP framework are its easy evaluation and lack of parameters, which often make calculations significantly more complicated. The Bhattacharyya distance between two continuous probability distributions $p(x), q(x)$ is defined as

$$D_{bha}(p(x) \,||\, q(x)) = -\ln\left(\int \sqrt{p(x)q(x)} dx \right).$$

Incorporating Bhattacharyya distance into the divCSP (bhadivCSP) framework of [14] leads to the objective function defined in the third row of Table I.

D. Gamma divergence

Gamma divergence, first defined in 2008 in [25], is considered to be a *super robust*[1] divergence measure. It has been used in robust parameter estimation [25], robust blind source separation [26] and clustering algorithms [27]. Symmetric

[1]The phenomenon of the breakdown point of an estimator being larger than 50% is termed super robustness.

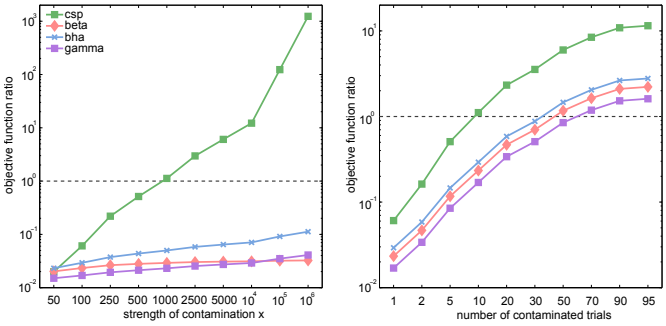

Fig. 1: *Left*: The impact of a single outlier on the objective function ratio. If the ratio exceeds 1, then the method will not extract the task-related source. *Right*: The objective function ratio after contaminating various numbers of trials.

Gamma divergence between two continuous probability distributions, $p(x) = p,\ \ q(x) = q$, is defined (for $\gamma > 0$) as

$$\tilde{D}_\gamma(p \mid\mid q) = \frac{1}{2}(d_\gamma(p,q) - d_\gamma(p,p) + d_\gamma(q,p) - d_\gamma(q,q))$$

where

$$d_\gamma(p,q) = -\frac{1}{\gamma}\log\left(\int p\, q^\gamma dx\right) + \frac{1}{1+\gamma}\log\left(\int q^{1+\gamma} dx\right).$$

The objective function of the Gamma divergence CSP (γ-divCSP) algorithm is defined in the last row of Table I.

IV. Evaluation

A. Simulations

In the first simulation experiment, we visualized the effect of a single outlier on the presented divergence functions. Therefore, we chose a two-dimensional example of covariance matrices, where Σ_1 is constant over 100 trials and Σ_2 is constant over 99 trials and includes one contaminated outlier trial (Σ_2^{100}):

$$\Sigma_1^{1:100} = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma_2^{1:99} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \Sigma_2^{100} = \begin{pmatrix} 1 & 0 \\ 0 & x. \end{pmatrix}$$

The correct discriminative source is the first one that would be detected by spatial filter $w_1 = (1,0)^T$. Depending on x, a non-robust method would be influenced by this outlier such that $w_2 = (0,1)^T$ would be preferred over w_1, meaning, the method would detect the wrong source. In the left panel of Figure 1, we visualize the ratio of the objective functions when applying filters w_2 and w_1: $\frac{\sigma(w_2)}{\sigma(w_1)}$. One can see that the standard CSP algorithm is not robust, because the ratio grows very quickly and becomes larger than 1 (i.e., CSP quickly prefers w_2 over w_1). Beta divergence and the two robust discrepancy measures proposed in this paper successfully reduce the influence of the outlier and prefer the correct filter, w_1, even for very large outlier values of x.

In a second experiment, we evaluated the robustness of the presented divergences when the number of outlier trials increases. We used the same example as before, with an outlier value (x) of 100. We added an increasing number of outliers to the data and computed the ratio of the objective functions. The right panel of Figure 1 depicts the results. Here, we again observe a significant increase in robustness when using the robust discrepancy measures. The standard CSP algorithm broke down (ratio of objective functions exceeds 1) when more than 10 outliers are contained in the data, whereas the γ-divCSP method withstood up to 70 contaminated outlier trials in this example. Note that we have fixed the parameters β and γ for β-divCSP and γ-divCSP respectively to 0.2 in the presented simulation experiments.

B. BCI Competition Dataset

TABLE II: Mean classification accuracies for four CSP methods

	aa	al	av	aw	ay
CSP	66.07	96.43	58.16	**88.84**	80.95
Beta	**74.11**	96.43	**70.41**	77.23	80.95
Bha	72.32	96.43	**70.41**	63.84	80.95
Gamma	**74.11**	96.43	69.39	83.04	80.95

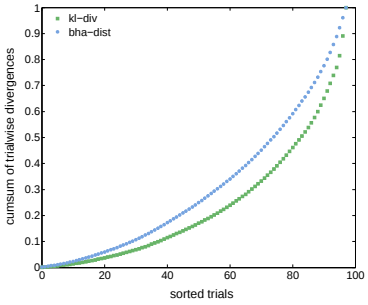

Fig. 2: Trialwise KL divergences and Bhattacharyya distances of bha-divCSP projected data of participant *av*. The divergences have been sorted, normed and accumulated.

The described methods were applied to data set IVa [28] from BCI Competition III [29], which contains EEG recordings from five healthy participants performing right hand and foot motor imagery without getting any feedback. The target class was visualized by letters appearing behind a fixation cross and a randomly moving object. The EEG signal was recorded from 118 Ag/AgCl electrodes, band-pass filtered between 0.05 and 200 Hz and downsampled to 100 Hz. We manually selected 68 electrodes, mainly those covering the motor cortex, and divided the data into a training and a testing set according to BCI Competition III. Parameters for Gamma and Beta divergence were determined individually for each participant by a 2-fold cross validation using the following values: [0,0.0001,0.001,0.01,0.05,0.1,0.15,0.2,0.25,0.5,0.75,1,1.5,2,5], where 0 is equivalent to original CSP and $\gamma = 1$ in γ-divCSP is equivalent to bha-divCSP.

Classification accuracies can be seen in Table II. Note, that for *al* and *ay* parameter selection yielded 0, which means

that, at least during cross validation, CSP outperformed all three divCSP methods.

It seems that the recordings of participants *aa* and *av* were especially affected by artifacts, because all robust divCSP methods have improved classification accuracy over the standard CSP baseline. Due to the trade-off between robustness and efficiency, not all users appear to benefit from using robust divergence-based spatial filtering. Another effect which certainly seems to play a role is the lack of data in some experiments. Participants *aw* and *ay* only did 56 (26 foot, 30 right hand) and 28 (10 foot, 18 right) training trials, respectively, where due to symmetry (see section III-A) even only 52 and 20 were used. In the end they do not seem to have benefited from applying robust divergence-based CSP.

In Figure 2 we seek to explain the reason for the performance increase of user *av* by depicting the impact of single trials on the sum in the objective function of the divCSP algorithms (see Table I). We assume that, for Bhattacharyya distance, outlier trials of greater distance have smaller impacts on the total sum than for KL divergence. This effect of downweighting becomes obvious when cumulating the sorted and normed trialwise distances. Bhattacharyya distance is closer to a linear function than KL divergence, which means that those trials having greatest distance have a major impact on the sum of KL divergences, such that outliers highly influence the result.

V. Conclusion

The ability to control a BCI in the real world represents a major challenge wherefore robustness plays an important role. Since brain recordings are highly sensitive and differ from brain to brain, it is difficult to balance the trade-off between accuracy and robustness. Simulations and the application to real data have shown that the proposed discrepancy measures work significantly better than original CSP in case of heavy contamination. Selecting the optimal parameter highly complicates calculations, especially when only few training data is available (e.g. participant *aw*). Future work could focus on a preselection regarding different CSP methods and enhancing the parameter selection.

References

[1] G. Dornhege, J. del R. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, Eds., *Toward Brain-Computer Interfacing*. Cambridge, MA: MIT Press, 2007.

[2] B. Graimann, B. Z. Allison, and G. Pfurtscheller, *Brain-computer interfaces: Revolutionizing human-computer interaction*. Springer, 2010.

[3] R. Tomioka and K.-R. Müller, "A regularized discriminative framework for EEG analysis with application to brain-computer interface," *NeuroImage*, vol. 49, no. 1, pp. 415–432, 2009.

[4] B. Blankertz, S. Lemm, M. S. Treder, S. Haufe, and K.-R. Müller, "Single-trial analysis and classification of ERP components – a tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[5] S. Lemm, B. Blankertz, T. Dickhaus, and K.-R. Müller, "Introduction to machine learning for brain imaging," *NeuroImage*, vol. 56, no. 2, pp. 387–399, 2011.

[6] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing Spatial filters for Robust EEG Single-Trial Analysis," *IEEE Signal Proc. Magazine*, vol. 25, no. 1, pp. 41–56, 2008.

[7] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial eeg during imagined hand movement," *IEEE Trans. Rehab. Eng.*, vol. 8, no. 4, pp. 441–446, 1998.

[8] W. Samek, C. Vidaurre, K.-R. Müller, and M. Kawanabe, "Stationary common spatial patterns for brain-computer interfacing," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026013, 2012.

[9] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain-computer interface," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 4, pp. 610–619, 2013.

[10] M. Kawanabe, W. Samek, K.-R. Müller, and C. Vidaurre, "Robust common spatial filters with a maxmin approach," *Neural Computation*, vol. 26, no. 2, pp. 1–28, 2014.

[11] D. Devlaminck, B. Wyns, M. Grosse-Wentrup, G. Otte, and P. Santens, "Multi-subject learning for common spatial patterns in motor-imagery bci," *Computational Intelligence and Neuroscience*, vol. 2011, no. 217987, pp. 1–9, 2011.

[12] H. Kang and S. Choi, "Bayesian multi-task learning for common spatial patterns," in *Int. Workshop on Pattern Recognition in NeuroImaging (PRNI)*, 2011, pp. 61–64.

[13] W. Samek, D. Blythe, K.-R. Müller, and M. Kawanabe, "Robust spatial filtering with beta divergence," in *Advances in Neural Information Processing Systems 26*, 2013, pp. 1007–1015.

[14] W. Samek, M. Kawanabe, and K.-R. Müller, "Divergence-based framework for common spatial patterns algorithms," *IEEE Reviews in Biomedical Engineering*, vol. 7, pp. 50–72, 2014.

[15] S. Amari and A. Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.

[16] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[17] M. Kawanabe, W. Samek, P. von Bünau, and F. Meinecke, "An information geometrical view of stationary subspace analysis," in *Artificial Neural Networks and Machine Learning - ICANN 2011*, ser. LNCS. Springer, 2011, vol. 6792, pp. 397–404.

[18] M. Mihoko and S. Eguchi, "Robust blind source separation by beta divergence," *Neural Computation*, vol. 14, no. 8, pp. 1859–1886, 2002.

[19] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain–computer interfaces for communication and control," *Clinical neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.

[20] S. Eguchi and Y. Kano, "Robustifying maximum likelihood estimation," *Tokyo Institute of Statistical Mathematics, Tokyo, Japan, Tech. Rep*, 2001.

[21] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *The Indian Journal Of Statistics*, vol. 7, no. 4, pp. 401 – 406, 1946.

[22] F. J. Aherne, N. A. Thacker, and P. I. Rockett, "The bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998.

[23] E. Choi and C. Lee, "Feature extraction based on the bhattacharyya distance," *Pattern Recognition*, vol. 36, no. 8, pp. 1703–1709, 2003.

[24] C. C. Reyes-Aldasoro and A. Bhalerao, "The bhattacharyya space for feature selection and its application to texture segmentation," *Pattern Recognition*, vol. 39, no. 5, pp. 812–826, 2006.

[25] H. Fujisawa and S. Eguchi, "Robust parameter estimation with a small bias against heavy contamination," *Journal of Multivariate Analysis*, vol. 99, no. 9, pp. 2053–2081, 2008.

[26] P.-W. Chen, H. Hung, O. Komori, S.-Y. Huang, and S. Eguchi, "Robust Independent Component Analysis via Minimum Divergence Estimation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 4, pp. 614–624, 2013.

[27] A. Notsu, O. Komori, and S. Eguchi, "Spontaneous clustering via minimum gamma-divergence," *Neural computation*, vol. 26, no. 2, pp. 421–448, 2014.

[28] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in noninvasive eeg single-trial classifications by feature combination and multiclass paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993 –1002, 2004.

[29] B. Blankertz, K.-R. Müller, D. Krusienski, G. Schalk, J. Wolpaw, A. Schlögl, G. Pfurtscheller, J. del R. Millán, M. Schröder, and N. Birbaumer, "The bci competition iii: validating alternative approaches to actual bci problems," *IEEE Trans. on Neural Syst. and Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, 2006.