35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Long Term Stability of Surface EMG Pattern Classification for Prosthetic Control*

Sebastian Amsüss[1], Liliana P. Paredes[2], Nina Rudigkeit[3],
Bernhard Graimann[2], Michael J. Herrmann[3] and Dario Farina[1]

***Abstract*— Long-term functioning of a hand prosthesis is crucial for its acceptance by patients with upper limb deficit. In this study the reliability over days of the performance of pattern classification approaches based on surface electromyography (sEMG) signal for the control of upper limb prostheses was investigated. Recordings of sEMG from the forearm muscles were obtained across five consecutive days from five healthy subjects. It was demonstrated that the classification performance decreased monotonically on average by 4.1% per day. It was also found that the accumulated error was confined to three of the eight movement classes investigated. This contribution gives insight on the long term behavior of pattern classification, which is crucial for commercial viability.**

I. INTRODUCTION

Pattern recognition of surface electromyographic (sEMG) signals is a promising approach for facilitating intuitive use of multifunctional upper limb prostheses. However, despite the research efforts of several groups (e.g., [1]-[4]), no commercial product is available to date. One reason for this failure is the lack of robustness of the proposed methods in real-life settings. Unintended prosthesis movements are indeed rated as highly frustrating [5] and the reliability of a prosthesis is a crucial factor for prosthesis acceptance [6]. Therefore, studies have been conducted to investigate the deteriorating effects of electrode shift with respect to the muscles [7], [8], varying arm postures [9], [10], and the changes in muscle activities due to fatigue [11], [12]. The reliability of classification across different days of use has also been addressed by [13], although for only one subject. Additionally a system that allows for retraining of the prosthesis whenever the user feels the necessity for it has been recently proposed [14], but a general understanding of the long-term dependence on classification accuracy is still needed. Therefore, in this study we aimed at quantifying changes in EMG classification accuracy as a function of time in the range of five days.

*This work has been supported by the European Commission via the Industrial Academia Partnerships and Pathways (IAPP), Grant No.251555 (AMYO) and was conducted within the Bernstein Focus Neurotechnology (BFNT) Göttingen

[1]S.Amsüss and D.Farina are with Dept. of Neurorehabilitation Engineering, Georg August University 37073 Göttingen, Germany sebastian.amsuess@bccn.uni-goettingen.de dario.farina@bccn.uni-goettingen.de

[2]L.P.Paredes, B.Graimann are with the Otto Bock Healthcare GmbH, 37115 Duderstadt, Germany liliana.paredes@ottobock.de bernhard.graimann@ottobock.de

[3]N.Rudigkeit is with Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany nina@nld.ds.mpg.de

[3]M.J.Herrmann is with Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany and with School of Informatics, The University of Edinburgh, EH89AB UK mherrmann@inf.ed.ac.uk

II. MATERIALS AND METHODS

A. Subjects

Five (two experienced and three naive) healthy subjects (age 25.8±0.8 years, 3 males and 2 females), were recruited to participate. The naive subjects were introduced to the topic and the systems used in a 15-min introduction and test session before the beginning of the main experiment. The study was approved by the local ethics committee and the subjects signed an informed consent prior to their participation.

B. Data collection

The data were collected using eight dry bipolar electrodes (Otto Bock HealthCare Products GmbH, Vienna, Austria, 13E200=50AC). The skin area was treated with abrasive gel, cleaned with alcohol and with a small amount of conductive gel to minimize the electrode-skin impedance. The electrodes were placed equidistantly around the circumference of the right forearm (all subjects were right-handed), 6.5-7.5 cm (depending on the subject's arm length) distal to the olecranon, using a custom-made mounting device. The exact locations of the electrodes were marked using a skin friendly, sweat and water resistant pen and renewed every day for accurate repositioning of the electrodes. The raw signals were amplified to a range of 0-4.5 V in the bandwidth 20-450 Hz with the inclusion of a 50 Hz notch filter. The processed signals were then sampled at 1kHz with a 10bit A/D converter and transferred to a personal computer via Bluetooth by the Axonmaster (Otto Bock HealthCare Products GmbH, Vienna, Austria), where they were recorded using a custom application.

C. Experimental protocol

The subjects were seated in front of a computer monitor. The maximum long term voluntary contraction (MLVC), defined as the maximum contraction that the subject was able to hold over a period of approximately 20s, was determined for each subject and movement. During data recording, trapezoidal profiles (t_{rise}=1s, $t_{plateau}$=3s, t_{fall}=1s) were displayed at three force levels: 30%MLVC, 60%MLVC and 90%MLVC. The sum of root mean square (RMS) values of the eight sEMG signals was calculated and displayed to the subjects as biofeedback and they were asked to trace the given profiles to the best of their abilities by activating their forearm muscles corresponding to movements of the following eight classes: wrist supination (WS), wrist pronation (WP), wrist flexion (WF), wrist extension (WE), hand opening (HO), key grip (KG), fine pinch (FP), and no movement (NM). The classes were indicated to the subjects using images with text captions and audio advice. For this study, only the static parts of the signals ($t_{plateau}$) were used. The mean tracking error was calculated as the average of the mean square errors (MSE) between the given profile trapezoids and the actual contractions. Every movement/force combination was repeated 15 times, resulting in a total of 1080s (8 classes x 3 force levels x 15 repetitions x 3 s) of data per subject per day. These sessions were repeated on five consecutive days.

D. Signal Processing

For classification, four widely used time domain features were used [2], [3], [12], [14], calculated in intervals of 128 ms and a frame increment of 50 ms: RMS, zero crossings (ZC), slope sign changes (SSC) and waveform length (WL). Classification was performed offline by applying linear discriminant analysis (LDA). Assuming multivariate Gaussian feature distribution and homoscedastic covariances for each class, LDA constitutes the optimal Bayesian classifier [15]. After this rule, decide among C target labels for class i s.t. the conditional probability $P(\cdot|\cdot)$ for $\boldsymbol{x}$ is maximized:

$$\frac{P(\boldsymbol{x}|i)\cdot P(i)}{\sum P(\boldsymbol{x})} > \frac{P(\boldsymbol{x}|j)\cdot P(j)}{\sum P(\boldsymbol{x})}, \quad \forall i \neq j \in \{1,\ldots,C\} \tag{1}$$

Assuming Gaussian distribution for the probability density function $P(\boldsymbol{x}|k)$

$$P(\boldsymbol{x}|k) = \frac{1}{(2\pi)^{\frac{n}{2}}\cdot|\Sigma_k|^{\frac{1}{2}}} exp\left(-\frac{d_k}{2}\right) \tag{2}$$

where $d_k = (\boldsymbol{x}-\boldsymbol{\mu}_k)\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T$ with $\boldsymbol{x},\boldsymbol{\mu}_k \in \mathbb{R}^n$, $\Sigma_k \in \mathbb{R}^{nxn}$ and k is a class label. Then combining (1) and (2) the decision function results in

$$\underset{i}{\text{argmax}}\left(\log P(i) - \frac{1}{2}|\Sigma_i| - \frac{d_i}{2}\right). \tag{3}$$

Further assuming homoscedastic covariances (Σ_i=Σ_j=Σ), (3) can further be simplified to the decision function $f(\boldsymbol{x},i)$

$$f(\boldsymbol{x},i) = \underbrace{\log P(i) - \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i}_{Cg_i} + \boldsymbol{x}^T\underbrace{\Sigma^{-1}\boldsymbol{\mu}_i}_{Wg_i} \tag{4}$$

and decide for class i s.t. $f(\boldsymbol{x},i) > f(\boldsymbol{x},j)\ \forall i \neq j \in C$. Cg_i and Wg_i can be calculated during the training phase of the classifier for each class and combined, so that classification of an unknown vector $\boldsymbol{x}$ can be achieved in only one matrix multiplication and is therefore a very efficient method. The classification accuracy (ratio between correct classifications and total classifications) was calculated within and between days. For the within days analysis, a five-fold cross-validation scheme was applied and the average classification accuracies were used. For the between days evaluations, the entire data set of one day was used for the classifier training and the entire data set of another day was tested. Every possible combination of train- and test days was employed. All results are presented as mean ± standard deviation.

III. RESULTS

The classification accuracy within days per subject was on average 97.9%±0.8. The results for a representative subject are shown in Fig. 1, indicating the highest classification accuracies within days and a decrease in accuracy when the test data and the training data were not obtained in the same day. It was further investigated whether a training effect occurred within five subsequent days of data recordings in tracking the given force profiles. As shown in Fig. 2, this effect was globally not observed. Moreover, there was no difference in this performance between experienced subjects (Subjects 1 and 5) and naive subjects.

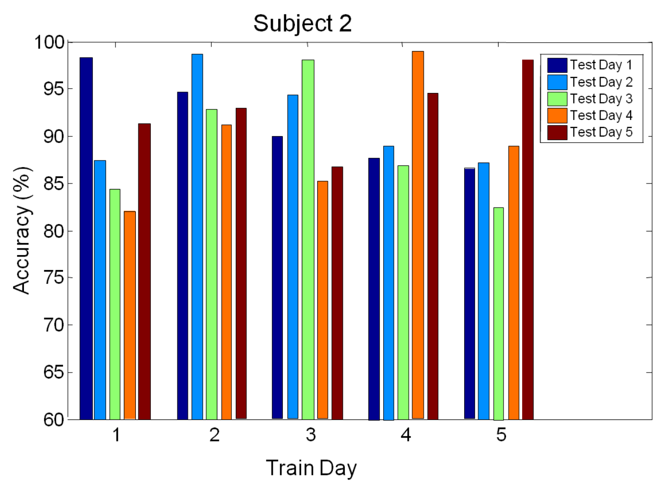

Fig. 1. Classification accuracies for a representative subject are high when classifier training - and test data are obtained from the same day (98.42%±0.4 for this subject) and tend to drop monotonically with increasing separation between training and test day. This trend is observable for all days, except day 5, on which a re-ascent was found in four of the five subjects.

In Fig. 3 the average classification accuracies are displayed, indicating that the accuracy tended to decrease from the maxima of within day accuracies monotonically as a function of the number of days separating the training and test set. On average, the classification accuracy dropped by 4.17% per day between training- and test days. Interestingly, in four of the five subjects the classification accuracies of the fifth day were consistently higher than the average for other between day assessments. This can be observed representatively in Fig. 1 for one subject and is also clear in Fig. 3, where an average increase in accuracy can be observed for the last day for all training days. In fact, in Fig. 3 it can be observed that when classifying data that were recorded before the data used for the classifier training,

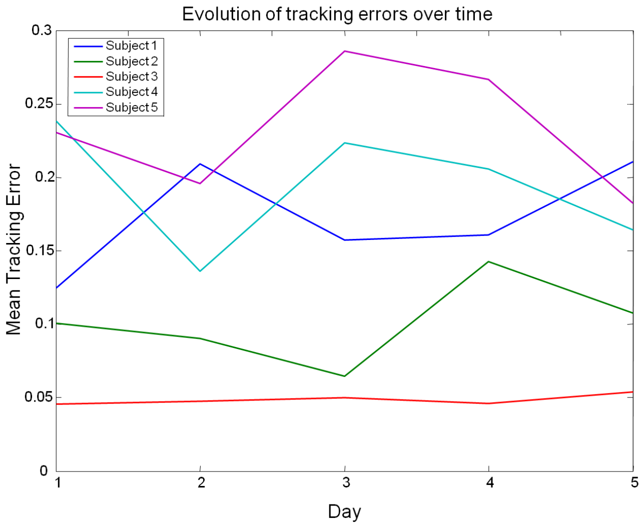

Fig. 2. Evolution of tracking errors per subject. No significant increase or decrease in the subjects' abilities of following the demanded force profiles could be observed within five subsequent days of data recording.

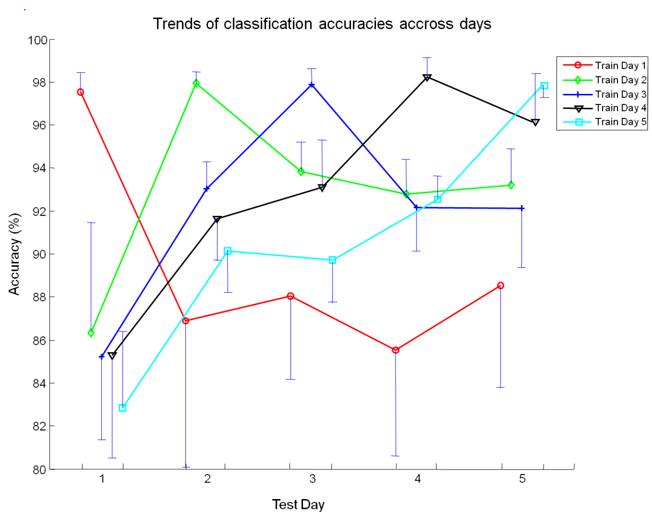

Fig. 3. Trends of classification accuracies, averaged over subjects (n=5). The classification accuracy decreases when the interval between train and test day increases. Error bars indicate one standard deviation and are drawn only in one direction for clarity.

there is a monotonic decrease in classification accuracy. In order to investigate the origin of classification errors, a histogram of the misclassifications averaged over all days and subjects was created from the respective test confusion matrices. Significant rates of misclassifications were defined as 5%. In Fig. 4 these classification error ratios (10 out of 64 were found to be significant) are plotted in descending order together with their corresponding standard deviations. It was found that confusions between classes HO and FP contributed the two largest portions of classification errors. As the standard deviations of these particular pairs of classes were relatively high, this type of error did not occur consistently across days and subjects. Class HO was present in three of the four highest classification errors, indicating that this movement is more difficult to classify consistently over days. Misclassifications of class WF had comparatively lower standard deviations, suggesting that the difficulty of classifying this movement was consistent for all subjects and days. Class WS was present in six of the lower seven significant misclassification ranks, also with low standard deviations, again demonstrating that classification of this movement was difficult consistently across subjects and days. In fact, 76.5% of all misclassifications involved classes WS, HO, or FP.

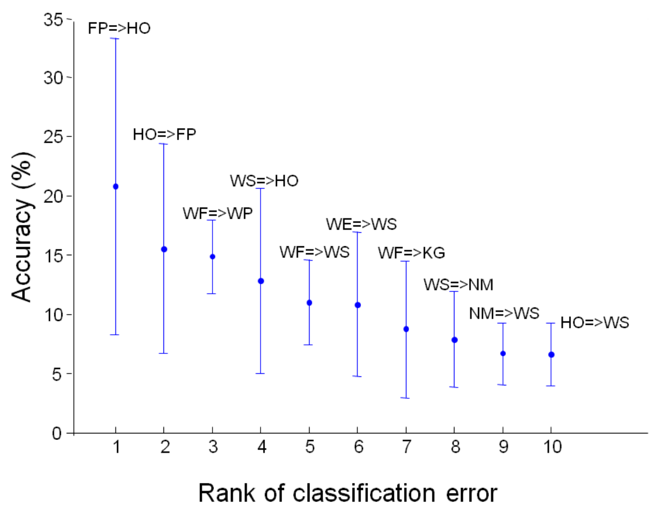

Fig. 4. Histogram of the largest classification errors, ranked in descending order, 1 standard deviation. The labels next to the bars indicate the actual and the calculated classes (e.g. FP=>HO means data from class FP were wrongly labeled as data from class HO).

IV. DISCUSSION

In the present study the relation between time and EMG classification accuracy was investigated. The results demonstrated that classification accuracy decreases as a function of time. In [7], [17] it was shown that classification accuracy determines the controllability of a prosthesis.Therefore, a functional online test would likely also reveal the same decreasing trend of task completions. Hence, the dependence on time of the classification accuracy is very important for the long term usability of a prosthesis. However, the reason of the gradual increase in classification error is not yet known. Not surprisingly, we demonstrate that the highest classification accuracy is achieved when training- and test data stem from the same day and session, using 5-fold cross validation. But in subsequent days, one may rather have expected random variations due to placement errors of the electrodes, stochastic changes in electrode-skin impedance from day to day and performance/concentration of the subject in performing the requested movements. Electrode placement was controlled very carefully with the skin markings. Yet it is possible that the relative position of the skin and the underlying muscles was slightly different each day due to compression exerted to keep the electrodes in place. In order to minimize variations in the electrode skin impedance, the skin was treated prior to recordings each day in the same manner, as described in Section II B. All of these uncertainties are of stochastic nature, therefore other, time dependent factors are likely involved and responsible for the monotonic decrease of classification accuracy. This may involve the motivation of the subjects. As presented, classification accuracy tended to be higher than average on the last day, which may support this assumption (motivation of last day in study). Another possible reason is that the repeatability of each movement improves over time, as described in [6], which compared class separabilities between novice and experienced subjects in a very similar setup as used in this study over the course of two days. An indicator for this may be the fact that when classifying data back in time a monotonic decrease in classification accuracy can be observed in Fig. 3, however especially data from the last day seem to be very well classified. A possible explanation would be that after a certain amount of training, subjects were able to perform the requested movements in a more repeatable manner. Therefore, the feature clouds for each movement would become more concise across days. As a consequence, such data can be very well classified but on the other hand do not yield a generalizing classifier. Another possible reason for this saturation is that after five days the increase in classification error converges. Moreover no correlation between the error in tracking the force profiles and classification accuracy was observed. Therefore, the individual correlations of tracking performance and classification accuracies for all movements have to be analyzed in detail and their contribution to the overall classification accuracy, which is out of the scope of this study. The average within days accuracy was 97.9% and decreased on average 4.2% per day during the investigated period of five days. It was also shown that misclassifications of three classes (WS, HO, FP) accounted for 76.5% of the total averaged misclassifications. These results indicate that with the signal processing methods and classification approach applied in this study, a decrease of functionality has to be expected within the range of a few days and that certain sub groups of classes require more attention than other movements. Possibly including data from several days into the classifier training might help to improve the stability of prosthesis control, but this has not been investigated in this study.

Finally, it is worth noting that the current study is limited in the number of subjects and days investigated and in the constraint of analysis of able-bodied subjects only. Five days can be considered long term when compared to most other studies, dealing with intra-session analysis, but is still relatively short compared to a real prosthesis usage of several years. The trends found in this study should be confirmed with a significantly longer period of investigation. The presented results must be considered in the context of these limitations. A greater subject group, including subjects with amputations, should also be employed for further extending our conclusions.

ACKNOWLEDGMENTS

The authors would like to extend their gratitude to Jakob L. Dideriksen for the fruitful discussions and helpful comments and suggestions for the manuscript. Furthermore we would like to thank David Hofmann for his help in recruiting subjects for this study.

REFERENCES

[1] C. Toledo, L. Leija, R. Muoz, A. Vera, A. Ramrez, Upper Limb Prostheses for Amputations Above Elbow: A Review. 2009 Pan American Health Care Exchanges Conference and Workshop Proceedings

[2] E. Scheme, K. Englehart, Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use. The Journal of Rehabilitation Research and Development, 48(6), pp. 643-660, 2011.

[3] M. Zecca,S. Micera, M.C. Carrozza, P. Dario, Control of multifunctional prosthetic hands by processing the electromyographic signal. Critical reviews in biomedical engineering, 30(4-6), pp. 459-85, 2002.

[4] B. Peerdeman, D. Boere, H. Witteveen, R. Huis in't Veld, H. Hermens, S. Stramigioli, H. Rietman et al., Myoelectric forearm prostheses: State of the art from a user-centered perspective. The Journal of Rehabilitation Research and Development, 48(6), pp. 719-738, 2011.

[5] L. J. Hargrove, E. J. Scheme, K. B. Englehart, B. S. Hudgins, Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis. IEEE transactions on neural systems and rehabilitation engineering, 18(1), pp. 49-57, 2010.

[6] N. E. Bunderson, T. Kuiken, Quantification of feature space changes with experience during electromyogram pattern recognition control. IEEE transactions on neural systems and rehabilitation engineering: 20(3), pp. 239-46, 2012.

[7] A.J. Young, L. J. Hargrove, T. Kuiken, The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift. IEEE transactions on bio-medical engineering, 58(9), pp. 2537-44, 2011.

[8] L. Mesin, R. Merletti, A. Rainoldi, Surface EMG: the issue of electrode location. Journal of electromyography and kinesiology, 19(5), pp. 719-26, 2009.

[9] N. Jiang, S. Muceli, B. Graimann, D. Farina, Effect of arm position on the prediction of kinematics from EMG in amputees. Medical & biological engineering & computing.

[10] Y. Geng, P. Zhou, G. Li, Toward attenuating the impact of arm positions on electromyography pattern-recognition based motion classification in transradial amputees. Journal of neuroengineering and rehabilitation, 9(74), 2012.

[11] T. Vukova, M. Vydevska-Chichova, N. Radicheva, Fatigue-induced changes in muscle fiber action potentials estimated by wavelet analysis. J. of electromyography and kinesiology, 18(3), pp. 397-409, 2008.

[12] B. Wan, L. Xu, Y. Ren, L. Wang, S. Qiu, X. Liu, et al., Study on Fatigue Feature from Forearm SEMG Signal Based on Wavelet Analysis, Proceedings of the 2010 IEEE International Conference on Robotics and Biomimetics, 1229-1232, 2010

[13] P. Kaufmann, K. Englehart, M. Platzner, Fluctuating emg signals: Investigating long-term effects of pattern matching algorithms, Proceedings of the Annual International Conference of the IEEE EMBS. 6537 - 6360, 2010.

[14] L. C. Chicoine, A. M. Simon,L. J. Hargrove, Prosthesis-Guided Training of Pattern Recognition-Controlled Myoelectric Prosthesis, Proceedings of the 34th Annual International Conference of the IEEE EMBS, 2012

[15] S. Theodoridis, K. Koutroumbas, Pattern Recognition. Elsevier: Burlington, MA 01803, USA (ISBN 978-1-59749-272-0), 4^{th} Edt., pp.24ff, 2009

[16] D. Tkach, H. Huang, T. Kuiken, Study of stability of time-domain features for electromyographic pattern recognition. Journal of neuroengineering and rehabilitation, 7(21) pp 1-13, 2010

[17] L.H. Smith, L.J. Hargrove, B. Lock, T. Kuiken, Determining the optimal window length for pattern recognition-based myoelectric control: balancing the competing effects of classification error and controller delay. IEEE transactions on neural systems and rehabilitation engineering: 19(2), pp. 186-92, 2011.