Movement Error Rate for Evaluation of Machine Learning Methods for sEMG-Based Hand Movement Classification

Arjan Gijsberts, Manfredo Atzori, Claudio Castellini, Henning Müller, and Barbara Caputo

***Abstract*—There has been increasing interest in applying learning algorithms to improve the dexterity of myoelectric prostheses. In this work, we present a large-scale benchmark evaluation on the second iteration of the publicly released NinaPro database, which contains surface electromyography data for 6 DOF force activations as well as for 40 discrete hand movements. The evaluation involves a modern kernel method and compares performance of three feature representations and three kernel functions. Both the force regression and movement classification problems can be learned successfully when using a nonlinear kernel function, while the exp-χ^2 kernel outperforms the more popular radial basis function kernel in all cases. Furthermore, combining surface electromyography and accelerometry in a multimodal classifier results in significant increases in accuracy as compared to when either modality is used individually. Since window-based classification accuracy should not be considered in isolation to estimate prosthetic controllability, we also provide results in terms of classification mistakes and prediction delay. To this extent, we propose the movement error rate as an alternative to the standard window-based accuracy. This error rate is insensitive to prediction delays and it allows us therefore to quantify mistakes and delays as independent performance characteristics. This type of analysis confirms that the inclusion of accelerometry is superior, as it results in fewer mistakes while at the same time reducing prediction delay.**

***Index Terms*—Electromyography, machine learning, prosthetics.**

I. Introduction

MACHINE learning is increasingly being employed in the research setting to improve myoelectric control of prostheses (see [1], [2] and references therein). Potential advantages of these methods over traditional approaches include an increased level of dexterity and a more intuitive form of control [3]. Furthermore, these learned models adapt to the specific signals provided to them, so that precise positioning of the electrodes is no longer essential to achieve acceptable performance [4].

The Non-Invasive Adaptive Prosthetics (NinaPro) project aims to support this stream of research by publicly releasing large-scale datasets of myoelectric data [5], [6]. In the present work, we perform a benchmark evaluation on the second version of the NinaPro database, which at the moment contains data collected from 40 intact subjects. The evaluation covers two distinct approaches to myoelectric control, namely force regression for 6 degrees of freedom (DOFs) (i.e., the four fingers and two axes for the thumb) and classification of 40 different hand movements. We employ a modern kernel-based learning algorithm and compare combinations of linear as well as nonlinear kernels with three different feature representations. Following recent promising results on the inclusion of accelerometry (ACC) as an auxiliary modality [7], [8], we also investigate the benefit of combining surface electromyography (sEMG) with accelerometry in a multimodal classifier.

Recent studies have found that the commonly used window-based classification accuracy is only weakly related to online controllability [9], [10]. It certain cases, methods with a lower overall classification accuracy actually performed better in terms of controllability [10], [11]. Hargrove *et al.* [10] therefore caution against using classification accuracy as the sole measure of performance, suggesting that besides the accuracy also the type of errors affects controllability [11]. Smith *et al.* [12] provided insight on this distinction by varying the window length of feature extraction. They found that longer window lengths led to increased classification accuracy as well as higher controller delays. Both these consequences have an opposite effect on controllability, indicating that error rates as well as delays are important offline indicators for controllability that should be considered jointly.

A shortcoming of using the standard window-based accuracy in this context is that it equally penalizes both misclassifications (e.g., false activations) as well as mistakes due to controller delay. This means that window-based accuracy is in fact partially dependent on controller delays, reducing the effectiveness of considering both measures as competing characteristics. We therefore propose the movement error rate (MER) as an alternative for window-based accuracy, inspired by the similar word and phoneme error rates commonly used in automated speech recognition [13]. This error rate measures the similarity of the true and predicted sequences of movements, rather than sequences of windows, and is therefore insensitive to delays in

Manuscript received July 31, 2013; revised December 17, 2013; accepted January 16, 2014. Date of publication January 29, 2014; date of current version July 03, 2014. This work was supported in part by the Swiss National Science Foundation Sinergia Project 132700 NinaPro (http://www.idiap.ch/project/ninapro).

A. Gijsberts and B. Caputo are with the Institute de Recherche Idiap, Martigny, Switzerland.

M. Atzori and H. Müller are with the Information Systems Institute, University of Applied Sciences Western Switzerland (HES-SO Valais), Sierre, Switzerland.

C. Castellini is with the Robotics and Mechatronics Center of the DLR, German Aerospace Center, Weßling, Germany.

Digital Object Identifier 10.1109/TNSRE.2014.2303394

the predictions. This allows classification mistakes and prediction delays to be quantified as independent performance characteristics. We use this joint performance characterization to further establish the improvement of the multimodal classifier over the sEMG-only classifier.

The remainder of this paper is organized as follows. The data acquisition, exercises, and data postprocessing steps are described in Section II. Subsequently, the experimental setup for the benchmark evaluation is detailed in Section III, which contains a description of the considered feature representations and learning method, as well as the definition of the MER. Both the regression and classification benchmark results are presented in Section IV, which additionally contains further analysis in terms of the MER and delay tradeoff. A discussion of the results as well as pointers to future work are given in Section V, while the paper is concluded in Section VI.

II. Database

The database used in this work is the second version of the database released within the Ninapro project [5], which aims to support the scientific community working on sEMG-based hand prostheses by publicly releasing large-scale databases [6]. The two database versions share a common acquisition procedure, in which myoelectric activity is recorded while subjects perform multiple repetitions of a large set of hand movements. Practical experience with the first version and feedback from amputated subjects have led to a number of improvements; for instance, the number of repetitions for each movement has been reduced from 10 to 6 to limit fatigue and cognitive load on amputated subjects. In addition, the use of a different type of electrodes allows recording raw myoelectric signals, while for the single digit movements we now record forces at the fingertips rather than hand kinematics. The motivation for this latter modification is to support research both on discrete movement classification as well as proportional control of the individual fingers. In the following, the acquisition setup and protocol as well as low-level postprocessing are described in more detail.

A. Acquisition Setup and Protocol

The primary component in the acquisition setup is a Delsys™ Trigno Wireless System®, which consists of a base station and multiple wireless sEMG electrodes. These electrodes are equipped with a self-contained rechargeable battery and they allow an operative range of 40 m. Myoelectric signals are sampled at a rate of 2 kHz with a baseline noise of less than 750 nV RMS. An advantage of these specific electrodes is that they also integrate a 3-axes accelerometer sampled at 148 Hz. The base station receives the sEMG and accelerometry streams over a proprietary wireless communication protocol and relays these via a standard USB connection to the laptop responsible for data acquisition.

There is debate in the scientific literature about the optimal placement strategy for sEMG electrodes. Some prefer to carefully position the electrodes with respect to the muscular anatomy of the forearm [14], while others have reported success when combining dense sampling with machine learning techniques [15]. Hargrove *et al.* [4] found that machine learning based methods are insensitive to nominal electrode placement, provided that the same locations are used for training and testing. In our acquisition setup, 12 electrodes were attached to the subject's arm following a hybrid of both strategies (see Fig. 1). The first eight electrodes were placed around the forearm to obtain a dense sampling of the muscles located at the proximal part of the forearm. Their exact position was determined by placing the first electrode on the forearm in exact sagittal correspondence to the radiohumeral joint. The remaining seven electrodes were placed equidistant in the same sagittal plane around the forearm. This plane was the one most proximal to the biceps while keeping the forearm perpendicular to the upper arm. Four additional electrodes were instead targeted at specific muscles, which were identified by palpation while the subject was repeatedly contracting the muscle. Electrodes 9 and 10 were placed on the main activity spots of the *extensor digitorum communis* and the *flexor digitorum superficialis*, while electrodes 11 and 12 were placed on the biceps and triceps. These muscles were selected based on their importance for motor control of the hand and forearm, and since these muscles are still available in the majority of transradial amputees. Prior to attaching the electrodes with adhesive tape, the skin of the subject was carefully cleaned with isopropyl alcohol. To prevent displacement or even detachment during the experimental procedure, the electrodes were subsequently secured using a latex-free self-adhesive bandage, as seen in Fig. 2.

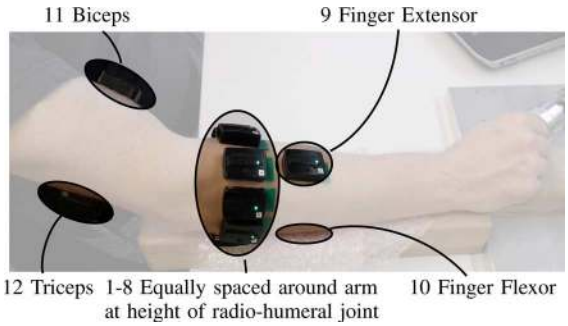

Fig. 1. Placement of the 12 electrodes on the arm. Electrode on the finger flexor is occluded by the arm and therefore not visible in this image.

Fig. 2. Acquisition setup for the (a) discrete movement and (b) force exercises.

During the acquisitions, subjects were seated at a desk resting their arm comfortably on the desktop. A laptop in front of the subject provided visual stimuli while at the same time acquiring data from all measurement devices. Subjects were asked their consent prior to the experiment and to fill in a brief questionnaire concerning personal data, such as age, gender, height, weight, laterality, and self-reported health status. The

TABLE I
DESCRIPTION OF 40 MOVEMENTS

	#	Description
Hand and wrist movements	1	Thumb up
	2	Extension of index and middle finger while flexing others (cf. "V-sign")
	3	Flexion of ring and little finger while extending others
	4	Thumb opposing base of little finger
	5	Abduction of the fingers
	6	Fingers flexed together in fist
	7	Pointing index
	8	Adduction of extended fingers
	9-10	Wrist supination and pronation (rotation axis through the middle finger)
	11-12	Wrist supination and pronation (rotation axis through the little finger)
	13-14	Wrist flexion and extension
	15-16	Wrist radial and ulnar deviation
	17	Wrist extension with closed hand
Grasps and functional movements	18-19	Large and small diameter grasp
	20-21	Fixed hook and index finger extension grasp
	22	Medium wrap
	23-24	Ring and prismatic four fingers grasp
	25-26	Stick and writing tripod grasp
	27-29	Power, three finger, and precision sphere grasp
	30-32	Tripod, prismatic, and tip pinch grasp
	33-35	Quadpod, lateral, and parallel extension grasp
	36	Extension type grasp
	37	Power disk grasp
	38	Open a bottle with a tripod grasp
	39	Turn a screw
	40	Cut something

acquisition procedure was approved by the Commission Cantonale Valaisanne d'Étique Médicale under identifier CCVEM 010/11 in the canton of Valais, Switzerland.

B. Exercises 1 and 2: Discrete Movements

In the first two exercises, subjects were instructed to perform a large set of hand movements, which were demonstrated by means of a video on the acquisition laptop. In this manner, they performed six consecutive repetitions of the 40 movements described in Table I, where each repetition lasted around 5 s. To ensure a consistent start and end position, repetitions were alternated with a rest posture lasting approximately 3 s. The set of movements was selected from the hand taxonomy, robotics, and rehabilitation literature (see [5] for more information), with the aim of covering the majority of hand postures encountered in daily activities. Furthermore, the sequence of movements was not randomized as to encourage repetitive, almost unconscious movements.

To avoid muscle fatigue, the 40 movements were split over two exercises. The first exercise covered 17 hand and wrist movements and lasted around 23 min, while the second exercise took 31 min and consisted of the remaining 23 grasps and functional movements. Both exercises were separated by approximately 5 min of rest, even though no subject reported fatigue at the end of either exercise. Prior to starting the acquisition, each subject was introduced to the experimental procedure by means of a short training sequence.

TABLE II
DESCRIPTION OF NINE FORCE PATTERNS

#	Description
1	Flexion of the little finger
2	Flexion of the ring finger
3	Flexion of the middle finger
4	Flexion of the index finger
5	Abduction of the thumb
6	Flexion of the thumb
7	Flexion of the index and little finger
8	Flexion of the ring and middle finger
9	Flexion of the index finger and the thumb

C. Exercise 3: Finger Forces

In the third and final exercise, subjects were required to produce a set of nine force patterns (see Table II) by pressing with one or more digits of their dominant hand. The activations involved six DOF, namely flexion of the five digits as well as abduction of the thumb. An initial calibration phase was performed to establish the rest and maximal voluntary contraction (MVC) force levels for all DOFs. The actual exercise required subjects to match the force levels indicated by bar stimuli (i.e., one for each DOF) on the laptop screen. This stimulus followed a bell-shaped curve reaching up to 80% of the MVC force level established during calibration. Although subjects did not receive feedback of their own forces during the acquisition, each of the patterns in Table II was preceded by a brief training phase with visual feedback that allowed them to adjust to 80% MVC.

For this exercise, the previously described setup was extended with a Finger-Force Linear Sensor (FFLS) [16]. This device measures flexion and extension forces of the four parallel fingers using a linear single-axis strain gage force sensor, while flexion and extension as well as abduction and adduction forces of the thumb are measured using a similar dual-axis sensor. These sensors are characterized by high signal repeatability, minimal drift over time, almost perfect linearity, and virtually nonexistent hysteresis (both parameters deviate no more than 0.3%). Each force sensor was connected to a dedicated amplifier, whose outputs were subsequently acquired at 100 Hz using a National Instruments DAQ card (NI-DAQ PCMCIA 6024E, 12-bit resolution).

As seen in Fig. 2(b), the sensors were placed according to the anatomy of a hand on a solid base, which allowed repositioning of the sensors to accommodate different hand sizes. A wooden support was placed in front of the FFLS to support the wrist and forearm, while a wooden block shaped to fit the palm was placed under the subject's hand to promote a stable hand configuration and to avoid wrist flexion as well as forearm pronation or supination during pressing. The four fingers were attached to the sensors using Velcro hook-and-loop straps with minimal slack, to ensure accurate force readings in positive and negative directions. Similarly, custom made gypsum casts in varying sizes were used for backlash-free attachment of the subject's thumb to the dual-axis sensor. The subject's forearm was not constrained other than resting on the wooden supports, as to increase comfort and encourage natural movements by allowing some freedom of movement. Subjects were however instructed

to only activate the indicated digits and to refrain from flexing the wrist.

D. Subjects

A total of 40 intact subjects participated in the data acquisition, consisting of 28 men and 12 women, 34 right-handed and 6 left-handed subjects. The age, weight, and height averages and standard deviations are, respectively, 29.9 ± 3.9 y, 70.9 ± 14.2 kg, and 172.8 ± 10.4 cm. All self-reported properties are available in anonymous form as part of the database.

E. Postprocessing

Each sample from each device was assigned a high-resolution timestamp at the moment of acquisition in a reference time based on the CPU's invariant timestamp counter. These timestamps were used during postprocessing to synchronize the data streams. More specifically, all streams were supersampled to the 2 kHz sampling rate of the sEMG stream using linear interpolation (real-valued streams) or nearest-neighbor interpolation (discrete streams). Prior to synchronization, the sEMG signals were cleaned from 50 Hz (and its harmonics) power-line interference using a Hampel filter [17].

A difficulty with the described acquisition procedure is that the movements performed by the subjects in the first two exercises may not match perfectly with the video stimulus. On several occasions, a subject would start the actual movement slightly after the start of the video and finish the movement either in advance or with some delay. This misalignment between the stimulus and the actual movement can be attributed to human reaction times as well as our explicit instruction to perform natural movements rather than exactly copying the kinematics of the video stimulus. The resulting erroneous movement labels have been corrected using an offline generalized likelihood ratio approach [18], which realigns the movement boundaries by maximizing the likelihood of a rest-movement-rest sequence.

III. Experimental Setup

We employ the control scheme proposed by Englehart and Hudgins [3], which consists of preprocessing the signals, segmenting them in windows, subsequently extracting features from the windows, and finally classification or regression based on the extracted features. These phases will be detailed in the following sections. A nearly identical setup has been used for both the force regression and movement classification benchmarks.

A. Preprocessing, Windowing, and Data Split

All channels were standardized to have a zero mean and unit standard deviation, based on statistics calculated solely on data from the training set. After this scaling, the signals were segmented using a sliding window with a length of 400 ms (800 samples). Although this window length is larger than in related work, preliminary experiments indicated that longer windows resulted in higher accuracy (see also [12], [18]). The increment of the sliding window was set to 10 ms (20 samples).

The data for each subject was split into training and test sets based on repetitions: the second and fifth repetition for each movement were used for testing, while the training set contained the remaining four repetitions. To ensure computational feasibility, the training and hyperparameter optimization sets were reduced in size by subsampling at regular intervals of 10 and 40 windows for classification (i.e., a window increment of 100 and 400 ms), and subsampling at intervals of 2 and 8 for the regression benchmark (i.e., a window increment of 20 and 80 ms). This configuration resulted in roughly 15 000 training samples in both settings.

B. Features

Selecting an appropriate feature representation is one of the most important determinants for regression and classification accuracy. To minimize the chance of reporting suboptimal performance in our benchmark, we select three popular types of features for sEMG data based on their diversity and their excellent results in earlier studies [18], [8]. For accelerometry signals, on the other hand, we follow the suggestion by Fougner *et al.* [7] and use the mean values within a processing window as features.

1) Root Mean Square: Perhaps the most commonly used feature representation for sEMG is the root mean square (RMS) of the signal. A compelling argument for this feature type is that (under ideal conditions) there is a quasi- or curvilinear relationship between the RMS value and the force exerted by a muscle [19]. Furthermore, the RMS of a signal is easily implemented in digital as well as analog systems.

2) sEMG Histogram: The second feature type is the sEMG Histogram (HIST) [20], which computes a histogram within the analysis window given a predefined number of bins. The HIST feature has demonstrated excellent performance for sEMG-based movement classification [20], [18]. Instead of fixing the lower and upper thresholds based on the extrema of the signal, we exploited the fact that the signals were standardized and set the thresholds to three standard deviations. In addition, extremal bins captured the outliers on each side, so that the effective bin edges were $[-\infty, -3, \ldots, +3, \infty]$. The total number of bins was fixed at 20.

3) Marginal Discrete Wavelet Transform: A more advanced representation that has recently gained popularity is the discrete wavelet transform (DWT). This transformation decomposes the signal in terms of a basis function (i.e., the wavelet) at different levels of resolution, resulting in a high-dimensional frequency-time representation. Lucas *et al.* [21] have demonstrated that it is sufficient for sEMG-based classification to preserve only the marginals at each level of the decomposition, thereby drastically reducing the dimensionality of the feature representation. Henceforth, this variant will be referred to as marginal discrete wavelet transform (mDWT). Although a variety of wavelet functions have been used in the context of sEMG [22], preliminary experiments on our data revealed that the 7th order Daubechies wavelet performed slightly better than others in a small pool of candidate functions. The marginal coefficients up to the third level obtained with this wavelet function have thus been used in the experimental validation.

4) Mean Value: Following related work [7], [8], the mean value (MEAN) within the processing window after interpolation is used as feature for the ACC modality. The dense place-

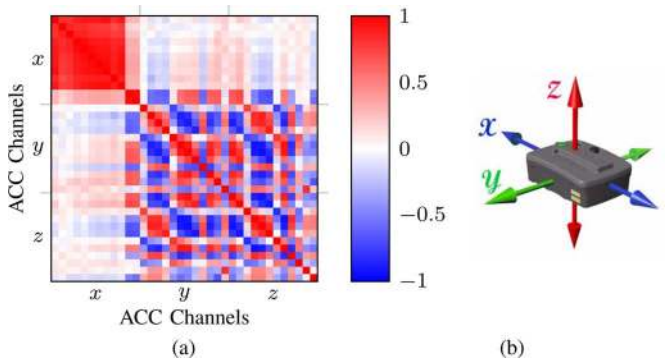

Fig. 3. (a) Correlation matrix of the 12×3 accelerometry channels (at original 148 Hz sampling rate) and for reference (b) orientation of the axes in the electrodes (figure taken from the Delsys Trigno Wireless System manual). This image is best viewed in color.

ment and regular orientation of the electrodes causes many of the ACC channels to be highly correlated (see Fig. 3). However, this redundancy was not been found to effect overall performance and all 36 channels (12 electrodes $\times$ 3 axes) have thus been used.

C. Learning Method

As a learning method we employed the Kernel Regularized Least Squares (KRLS) algorithm [23]. This kernel method is similar to the well-known support vector machine [24] in terms of formulation as well as practical performance [23], [25], but it offers multiple advantages in the context of our study. First, it can be applied in near identical form for both regression[1] as well as classification tasks. Furthermore, training KRLS consists of solving a linear system of equations, allowing multiple output dimensions to be learned simultaneously at negligible additional cost. This results in a considerable reduction of computational requirements, since our regression problem involves estimating forces for 6 DOF, while the multiclass classification problem is reduced to 41 binary classification problems (i.e., 40 movements and rest) using the well-known one-versus-all reduction.

1) Kernels: KRLS (and many other algorithms) can be used on nonlinear problems by employing so-called kernel functions, which implicitly map the data into a high or even infinite dimensional feature space. The *defacto* standard kernel function is the radial basis function (RBF)

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2) \text{ for } \gamma > 0$$

which has demonstrated excellent performance in a large variety of application domains. Nonetheless, some studies suggest that the exp-χ^2 kernel

$$k(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\gamma \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{x_i + y_i}\right) \text{ for } \gamma > 0$$

may be more appropriate for histogram-like feature representations. This is of interest in our setting, since all considered sEMG features produce non-negative representations. Finally,

[1]The algorithm is also known as Kernel Ridge Regression [25].

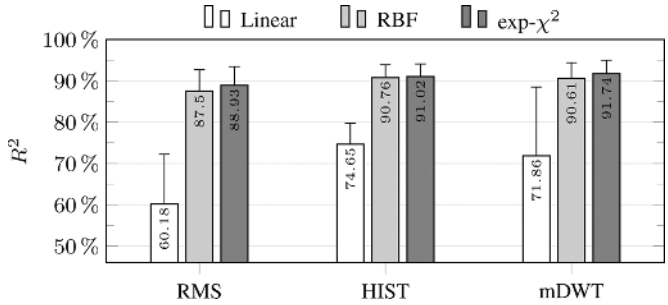

Fig. 4. Coefficient of determination R^2 for different types of sEMG features and kernels, averaged over the 6 DOFs and the 40 subjects. Error bars indicate unit standard deviation.

we also include the canonical linear kernel $k(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$ to establish whether nonlinearity is in fact required.

2) Combining Multiple Cues: Combination of multiple cues (i.e., either features or modalities) can be implemented by concatenating the individual feature vectors or by integrating the predictions of an ensemble of cue-specific classifiers. In kernel-based methods, however, it is more appropriate to combine cue-specific kernel functions, since this corresponds to concatenation in the implicit feature spaces induced by the respective kernels. Here we consider a linear combination of C cue-specific kernels

$$k(\boldsymbol{x}, \boldsymbol{y}) = w_1 k_1(\boldsymbol{x}_1, \boldsymbol{y}_1) + \cdots + w_C k_C(\boldsymbol{x}_C, \boldsymbol{y}_C)$$

where w_c for $1 \leq c \leq C$ weights the contribution of each kernel.

3) Hyperparameter Optimization: The KRLS algorithm requires setting a regularization parameter λ, which balances the tradeoff between under- and overfitting. This parameter was tuned together with the kernel parameter γ and the cue weights w_c (when applicable) using four-fold cross validation, where each of the folds corresponds to one of the four training repetitions. This particular splitting of the folds ensures that the distributional differences among repetitions were taken into account when optimizing the hyperparameters. To increase the likelihood of finding a (nearly) optimal configuration, parameters were selected using a dense grid search with $\lambda \in \{2^{-16}, 2^{-15}, \ldots, 2^3\}$, $\gamma \in \{2^{-20}, 2^{-19}, \ldots, 2^3\}$, and $w_c \in \{0.0, 0.1, \ldots, 1.0\}$ such that $\sum_{c=1}^{C} w_c = 1$.

D. Movement Error Rate

A problem with the window-based accuracy is that it does not distinguish between "true" mistakes (e.g., confusion between movements) and errors due to prediction delays. To address this shortcoming, we propose the MER as an alternative error measure. This error rate is motivated by and similarly defined as the so-called word or phoneme error rates in the field of speech recognition (e.g., [13]). Algorithm 1 describes the procedure to compute the MER. The first step is to erase adjacent duplicates in both the sequence of true labels as well as the predictions. Subsequently, the difference between the true and predicted sequence of movements is measured using the normalized Levenshtein distance [26], which counts the minimum number of

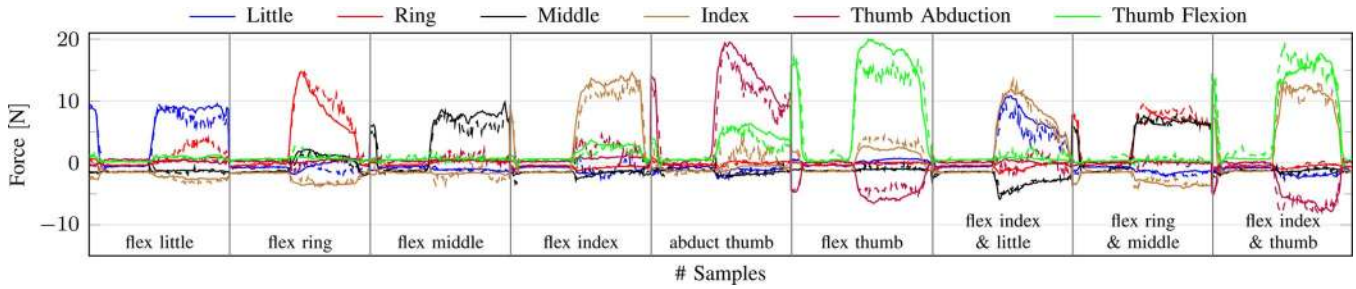

Fig. 5. Extract of the true (solid) and predicted (dashed) forces for the fifth repetition of all nine force patterns for the first subject in case of the mDWT+exp-χ^2 regressor. All signals have been subsampled by a factor of 5 to improve visualization. This figure is best viewed in color.

insertions, deletions, and substitutions to change one sequence into the other. Removing the adjacent duplicates has the effect that movements rather than windows become the atoms and that movement start and duration become irrelevant. This allows prediction delays (i.e., the first correct prediction after a label change) to be used as an orthogonal (nonredundant) performance measure.

Algorithm 1 Movement Error Rate

Require: true labels $\boldsymbol{y} = [y_1, \ldots, y_m]$, predictions $\hat{\boldsymbol{y}} = [\hat{y}_1, \ldots, \hat{y}_m]$

1: $\boldsymbol{z} \leftarrow$ EraseAdjacentDuplicates $(\boldsymbol{y})$
2: $\hat{\boldsymbol{z}} \leftarrow$ EraseAdjacentDuplicates $(\hat{\boldsymbol{y}})$
3: MER $\leftarrow$ Levenshtein $(\boldsymbol{z}, \hat{\boldsymbol{z}})$/Length $(\boldsymbol{z})$
4: **return** MER

IV. Benchmark

In this section, we establish benchmark results for both the force regression as well as movement classification tasks, to determine the feasibility of the tasks and to guide future experiments on the NinaPro dataset. Specifically, it is of interest to compare performance of the feature representations and kernel functions. We also quantify the performance gain when including the accelerometer modality in a multimodal classifier, as compared to an sEMG-only classification strategy. This section is concluded with a characterization of the classification results in terms of MER and prediction delay.

A. Force Regression

The accuracy on the force regression task, measured in terms of coefficient of determination R^2, is presented in Fig. 4. For each combination of kernel and feature type we report the average performance over all 40 subjects and the standard deviation. All nonlinear regressors achieve an acceptable performance irrespective of the feature type, which exceeds 90% for both HIST and mDWT feature types. A sign test comparing the individual R^2 scores of the mDWT/exp-χ^2 regressor with those of the other combinations reveals that its improvement is statistically significant ($p \leq 7‰$), although the absolute differences among the four best performing combinations are relatively small. A more intuitive understanding of the performance

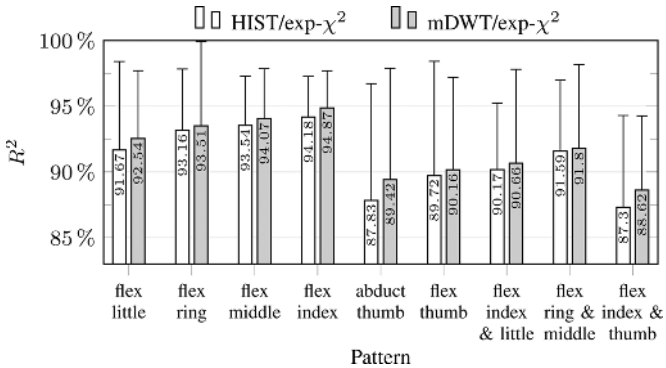

Fig. 6. Average coefficient of determination R^2 per pattern (see Table II) for both HIST and mDWT feature types with the exp-χ^2 kernel. Error bars indicate unit standard deviation.

is provided in Fig. 5, which shows an extract of the true and predicted forces for one repetition of all nine patterns. The forces are generally predicted rather well, though the residuals seem dependent on the magnitude of the force. Interestingly, the regressor even learned the involuntary negative forces most likely caused by synergistic or compensatory mechanisms [27].

The low performance of the linear kernel indicates that the capacity of linear models is not sufficient to capture the relationship between sEMG signals and forces at the fingertips. The lack of capacity seems confirmed by the fact that the HIST feature demonstrates higher performance than the other features, since the higher feature dimensionality of the former effectively increases the capacity of the linear regressor (e.g., in terms of VC dimensionality [28]). Regardless, the much higher capacity provided by the nonlinear kernels seems a necessity to obtain acceptable performance.

To investigate the relative difficulty of the nine force patterns from Table II, we report the individual performance per pattern for the two best performing feature-kernel combinations in Fig. 6. Patterns involving the individuated activation of the four fingers (patterns 1–4) are all characterized by high performance, while patterns involving the thumb or simultaneous activation of multiple digits show considerably worse performance. This difficulty of predicting thumb activations was observed as well by Kõiva *et al.* [29]. A likely explanation for this phenomenon is that in our acquisition no sEMG activity is recorded from the majority of thumb muscles. These muscles are located either in the hand proper or in the distal part of the forearm and they would therefore not be available in most amputees.

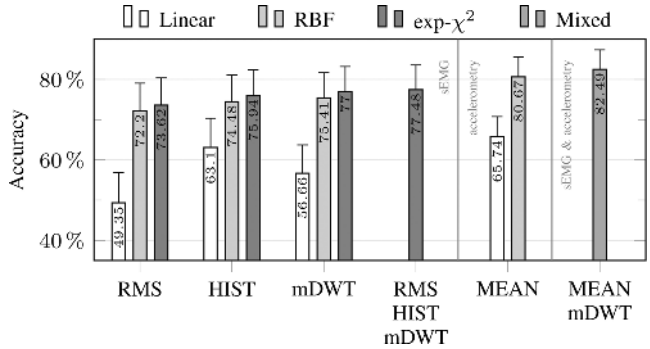

Fig. 7. Average classification accuracy over the 40 subjects when using (from left to right) the three sEMG based features individually with three different kernels and when combining these with the exp-χ^2 kernel, the MEAN feature of the ACC modality with two kernels, and when combining the sEMG and ACC modalities with the best performing kernels. Error bars indicate unit standard deviation.

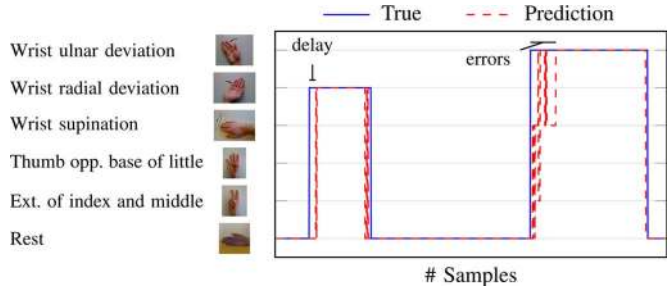

Fig. 8. Extract of the true labels and predictions taken from the first subject containing a repetition of a wrist radial deviation and a wrist ulnar deviation. Predictions were produced by the multimodal sEMG + ACC classifier.

B. Movement Classification

Results on the 41-class (40 movements and the rest posture) classification problem are shown in Fig. 7, which reports the average accuracy over the 40 subjects. Concentrating on the individual sEMG features (the three left-most groups in Fig. 7), we observe that also on this task the linear kernel performs significantly worse than the nonlinear kernels ($p \ll 1‰$). The higher dimensional HIST representation performs again better than mDWT features with the linear kernel, confirming that linear classifiers lack the capacity to learn the relation between sEMG and movement classes as well as finger forces. Among the two nonlinear kernels, we note a performance increase of around 1.5% for the exp-χ^2 kernel, regardless of the feature type. Thanks to the large number of subjects in our study, this difference can be shown to be statistically significant ($p \ll 1‰$). This confirms earlier indications that the exp-χ^2 kernel performs particularly well with non-negative feature representations, and that blindly choosing the "standard" RBF kernel can lead to suboptimal performance.

A common strategy to further increase performance is to combine several feature types in a multi-cue classifier (see Section III-C2). Contrary to the results reported by Gijsberts and Caputo [8], we observe a small increase in accuracy when combining the three sEMG features with exp-χ^2 kernels (see Fig. 7). This is because the contributions of each cue (i.e., the weights w_c) were tuned during hyperparameter optimization, while Gijsberts and Caputo [8] kept these weights fixed at C^{-1}. Though the increase in performance is significant ($p \leq 2.3‰$), it is arguable whether an advantage of less than 0.5% justifies the increased computational cost.

1) Including Accelerometry: Interestingly, the accuracy of the MEAN features over the accelerometry modality with RBF kernel is almost 81% and thus significantly higher than any of the considered sEMG features ($p \ll 1‰$). This confirms earlier observations that accelerometry is highly informative for hand movement classification. Furthermore, the right-most bar in Fig. 7 shows that integrating the best performing individual sEMG and ACC feature-kernel combinations in a multimodal classifier achieves a significant improvement of almost 2% over the ACC-only classifier ($p \ll 1‰$) and more than 5% over the sEMG-only classifier ($p \ll 1‰$). This proves empirically that the ACC and sEMG modalities are to be considered complementary, as has been suggested previously [7], [8].

C. Movement Error Rate versus Delay

A limitation of the window-based accuracy is that it does not distinguish between different types of mistakes made by the classifier. Consider for instance the extract in Fig. 8, where the subject was instructed to perform wrist radial and ulnar deviations alternated with the rest posture. The onsets of both movements are characterized by different problems: in the first case, the classifier suffers a loss due to a prolonged prediction of the rest posture after the start of the wrist radial deviation (i.e., a delay); in contrast, in the second case the classifier mistakes the wrist ulnar deviation for a variety of other movements (i.e., errors). The window-based accuracy fails to differentiate between both cases, as it assigns equal loss to delays and "real" mistakes.

Both types of mistakes can be quantified independently using respectively the MER and prediction delay. Furthermore, note that many of the errors in the example in Fig. 8 could have been avoided by temporal smoothing of the predictions during post-processing. Fig. 9(a) shows the effect of varying the window size k of a sliding majority vote on both the MER as well as the prediction delay. Increasing the amount of smoothing lowers the MER at the cost of larger prediction delay, and vice versa. It follows that the MER and prediction delay are competing characteristics that can be regulated using the smoothing parameter k. The standard window-based accuracy cannot capture this tradeoff, since a reduction of errors due to temporal smoothing would be offset by the increasing loss due to prediction delay.

Fig. 9(a) also gives a more complete insight into the synergy between the sEMG and accelerometry modalities. For a given delay, the multimodal classifier attains a lower MER than either unimodal classifier and, similarly, for a given MER it has a lower prediction delay. Particularly interesting is that the multimodal classifier achieves a considerably lower minimal prediction delay (i.e., below 300 ms) than either unimodal variant, demonstrating that the integration of modalities is instrumental in reducing errors as well as prediction delays.

In Section III-A, we mentioned selecting a relatively large window length of 400 ms based on the accuracy obtained in preliminary experiments. Fig. 9(b) shows the effect of varying the window length in terms of the MER and prediction delay. Without temporal smoothing (i.e., $k = 1$), we observe that a

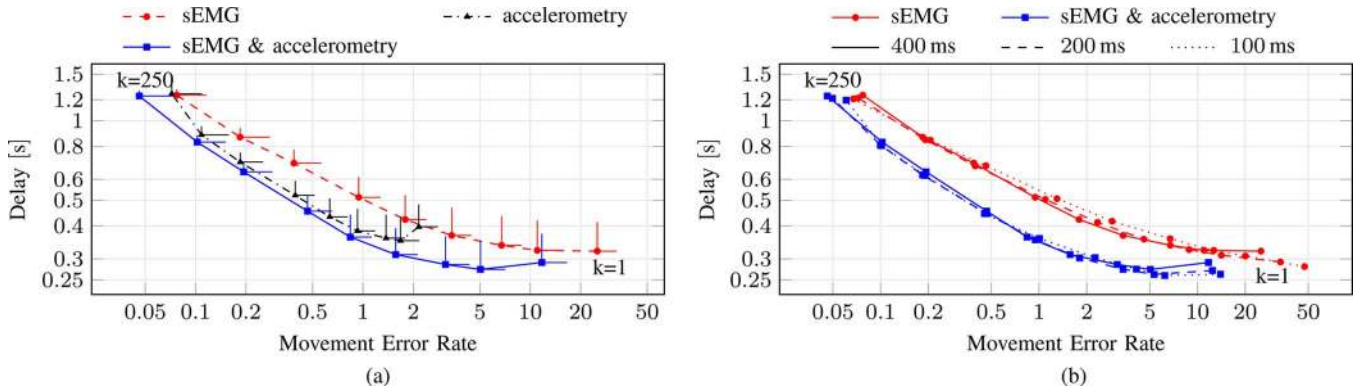

Fig. 9. Tradeoff between MER and prediction delay when varying the smoothing parameter $k \in \{1, 3, 5, 11, 25, 50, 100, 150, 250\}$ for the (a) sEMG-only mDWT/exp-χ^2 classifier, the ACC-only MEAN/RBF classifier, and the multimodal $\mathrm{mDWT} + \mathrm{MEAN}$ classifier, and (b) when varying the (feature) window length. Marks indicate the mean as averaged over all movements and subjects, while the error bar indicates a unit standard deviation in either axis (omitted in the right panel for clarity). (a) Unimodal versus multimodal. (b) Varying window length.

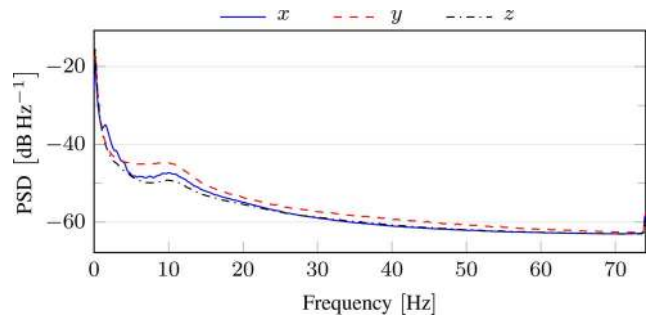

Fig. 10. Power spectral density of the accelerometer channels (at original 148 Hz sampling rate) estimated using Welch's method and grouped by Cartesian axis.

larger window length decreases the MER at the cost of higher prediction delay. When predictions are smoothed, however, the curves for different window lengths become nearly identical. This is not surprising, since the total length of historical data increases linearly with k, eventually dominating the length of the analysis window. Choosing the correct amount of smoothing seems therefore more crucial at attaining an optimal tradeoff between delay and prediction errors than varying the window length of the features.

V. Discussion

A. Benefit of Accelerometry

The high performance of the ACC-only and sEMG + ACC classifiers in Fig. 9(a) confirms that accelerometry is useful for movement classification. Earlier studies have shown that mechanomyography (MMG) measured using accelerometers can indeed be used for prosthetic control (e.g., [30]). In our setting, however, the performance cannot be attributed solely to measuring muscle activations with MMG. The power spectral density (PSD) in Fig. 10 reveals that the accelerometers captured primarily the gravitational field (near 0 Hz) and upper limb movement (approximately 0 to 6 Hz [31]), and to a lesser extent MMG (around 10 Hz [32]). Although the "motion-artefacts" caused by upper limb movements are often regarded as undesirable in the context of MMG [32], our results confirm earlier findings that these signals are instead useful for prosthetic control [7], [8]. Furthermore, Fougner *et al.* [7] have shown that measuring the gravitational field allows counteracting the so-called "limb position effect", which refers to deterioration of myoelectric control performance depending on the position or orientation of the arm.

B. Balancing Error Rate and Delay

It is evident from Fig. 9 that the tradeoff between MER and prediction delay can be adjusted by temporal smoothing or by varying the analysis window length. An important question that follows is which tradeoff results in optimal controllability, and whether this tradeoff is subject or task specific. Hargrove *et al.* [10] have suggested that false activations of the limb are more costly than those that can cause a pause in motion, implying that it would be preferable to reduce MER while maintaining delays within an acceptable range. Estimates for acceptable levels of delay (i.e., before controllability degrades drastically) range from 50 ms [33] to 300 ms [3].

The delays we found in our evaluation seem comparatively high with respect to these suggestions. One of the reasons is that we measured prediction delay as the first *correct* prediction after a label change, which is likely to be larger than pure controller delay in the presence of mistakes. This definition is identical to the motion-selection time used by Li *et al.* [34], who in their experiments found this time to be around 200 ms in amputated as well as intact arms. Furthermore, prediction delays are strongly dependent on the correctness of the (desired) movement boundaries. In our acquisition procedure, functional and grasp movements were performed on real objects and required an initial reaching movement. These reaching movements cause sEMG activity and were thus included as part of the movement, although the correct label during these transitory movements is obviously ambiguous. It is therefore plausible that the prediction delays in Fig. 9 are overestimated.

C. Practical Use of Movement Error Rate

There has been a recent shift from offline evaluation of myoelectric control systems to real-time closed-loop evaluation (e.g., [35]). Though undeniably preferable, real-time evaluation is often not practically feasible when comparing a large number of approaches. The motivation for joint characterization of the

MER and prediction delay is to allow an offline evaluation that has been demonstrated to be correlated with online controllability [12]. Consider for instance the curves for the sEMG-only and the multimodal sEMG+ACC classifiers in Fig. 9; the multimodal classifier demonstrates lower MER as well as lower prediction delays. This is a strong indicator that this classifier would also perform better in terms of controllability. There may be potential use for the MER in online scenarios as well. An advantage over window-based accuracy is that it does not require knowledge of the exact start and end time of the desired movements. Instead, a mere ordered sequence of desired movements is sufficient to compute the MER.

D. Future Directions

The planned future work will concentrate on two distinct directions. First, the benchmark evaluation will be confirmed on data from actual amputees. This would allow us to quantify to which degree results on intact subjects translate to amputees. Integration of accelerometry seems particularly useful, since lower arm dynamics may be less affected by the amputation than myoelectric signals. Second, the proposed analysis in terms of MER and prediction delays depends strongly on the correlation of related quantities (i.e., classification accuracy and controller delay) with online controllability [12]. Online control experiments are necessary to further establish this correlation, ideally in a user-centric scenario in which an amputee performs daily-life tasks using a real prosthesis.

Most results discussed in this paper focus on a movement classification setting, as opposed to the regression setting native to the proportional control approach. Also for proportional control there are indications that offline performance is at most weakly correlated with online control performance [36]. However, that work investigated whether the R^2 measure was correlated with a number of online performance indices. Whether there are other offline performance measures (e.g., correlation coefficient, prediction delay) that give reliable estimates of online proportional control performance is therefore still an open question.

VI. Conclusion

This paper presented a benchmark evaluation on the second revision of the publicly available NinaPro database. The evaluation involved two distinct approaches to myoelectric control, namely predicting forces at the four fingers and two axes of the thumb, as well as movement classification of 40 different hand movements in 40 intact subjects. The benchmark results indicate that a nonlinear kernel method can reach acceptable levels of performance on either problem. The exp-χ^2 kernel, which has not been commonly used in the present context, demonstrates higher classification accuracy than the popular RBF kernel for all considered (non-negative) feature representations in the regression as well classification settings. With respect to movement classification, accelerometry and sEMG were found to be complementary modalities and significant gains were achieved when both are combined in a multimodal classifier.

Recent studies have found that the commonly used window-based accuracy is only weakly related to online controllability, partially because it cannot distinguish between confusion between movement classes and prediction delays. We addressed this shortcoming by proposing the movement error rate, which measures the similarity of the actual and the predicted sequence of movements instead of windows. This metric is insensitive to prediction delays and therefore allows errors and delays to be quantified as two independent and competing characteristics. The balance between the error rate and delays can be regulated by means of temporal smoothing or by altering the analysis window length. Furthermore, this form of analysis confirmed the benefit of integrating accelerometry, as the multimodal classifier reduced both errors as well as prediction delay as compared to the sEMG-only classifier.

Acknowledgment

The authors would like to thank all subjects for their participation in the data acquisitions and the anonymous reviewers for their insightful comments.

References

[1] B. Peerdeman, D. Boere, H. Witteveen, R. H. Veld, H. Hermens, S. Stramigioli, H. Rietman, P. Veltink, and S. Misra, "Myoelectric forearm prostheses: State of the art from a user-centered perspective," *J. Rehab. Res. Devel.*, vol. 48, no. 6, pp. 719–738, 2011.

[2] A. Fougner, Ø. Stavdahl, P. J. Kyberd, Y. G. Losier, and P. A. Parker, "Control of upper limb prostheses: Terminology and proportional myoelectric control—A review," *IEEE Trans. Neural Systems Rehab. Eng.*, vol. 20, no. 5, pp. 663–677, Sep. 2012.

[3] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, vol. 50, no. 7, pp. 848–854, Jul. 2003.

[4] L. Hargrove, K. Englehart, and B. Hudgins, "A training strategy to reduce classification degradation due to electrode displacements in pattern recognition based myoelectric control," *Biomed. Signal Processing Contr.*, vol. 3, no. 2, pp. 175–180, 2008.

[5] M. Atzori, A. Gijsberts, S. Heynen, A.-G. M. Hager, O. Deriaz, P. van der Smagt, C. Castellini, B. Caputo, and H. Müller, "Building the Ninapro database: A resource for the biorobotics community," in *Proc. IEEE Int. Conf. Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 1258–1265.

[6] [Online]. Available: [Online]. Available: http://ninapro.hevs.ch

[7] A. Fougner, E. Scheme, A. D. C. Chan, K. Englehart, and Ø. Stavdahl, "A multi-modal approach for hand motion classification using surface EMG and accelerometers," in *Proc. Annu. Int. Conf. IEEE Eng. Medicine and Biology Society (EMBC)*, 2011, pp. 4247–4250.

[8] A. Gijsberts and B. Caputo, "Exploiting accelerometers to improve movement classification for prosthetics," in *Proc. IEEE Int. Conf. Rehab. Robotics (ICORR)*, 2013.

[9] B. Lock, K. Englehart, and B. Hudgins, "Real-time myoelectric control in a virtual environment to relate usability vs. accuracy," in *Proc. Myoelectric Symp.*, 2005.

[10] L. Hargrove, Y. Losier, B. Lock, K. Englehart, and B. Hudgins, "A real-time pattern recognition based myoelectric control usability study implemented in a virtual environment," in *Proc. Annu. Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBS)*, 2007, pp. 4842–4845.

[11] L. Hargrove, E. Scheme, K. Englehart, and B. Hudgins, "Multiple binary classifications via linear discriminant analysis for improved controllability of a powered prosthesis," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 18, no. 1, pp. 49–57, Jan. 2010.

[12] L. Smith, L. Hargrove, B. Lock, and T. Kuiken, "Determining the optimal window length for pattern recognition-based myoelectric control: Balancing the competing effects of classification error and controller delay," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 19, no. 2, pp. 186–192, Mar. 2011.

[13] J. Keshet and S. Bengio, *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*. New York, NY, USA: Wiley, 2008.

[14] C. J. D. Luca, "The use of surface electromyography in biomechanics," *J. Appl. Biomechanics*, vol. 13, pp. 135–163, 1997.

[15] C. Castellini, E. Gruppioni, A. Davalli, and G. Sandini, "Fine detection of grasp force and posture by amputees via surface electromyography," *J. Physiol. (Paris)*, vol. 103, no. 3–5, pp. 255–262, 2009.

[16] R. Kõiva, B. Hilsenbeck, and C. Castellini, "FFLS: An accurate linear device for measuring synergistic finger contractions," in *Proc. Annu. Int. Conf. IEEE Eng. Medicine Biology Soc. (EMBC)*, 2012, pp. 531–534.

[17] D. P. Allen, "A frequency domain hampel filter for blind rejection of sinusoidal interference from electromyograms," *J. Neurosci. Methods*, vol. 177, no. 2, pp. 303–310, 2009.

[18] I. Kuzborskij, A. Gijsberts, and B. Caputo, "On the challenge of classifying 52 hand movements from surface electromyography," in *Proc. Annu. Int. Conf. IEEE Eng. Medicine Biology Soc. (EMBC)*, 2012, pp. 4931–4937.

[19] E. Criswell, *Cram's Introduction to Surface Electromyography*. London, U.K.: Jones & Bartlett, 2010.

[20] M. Zardoshti-Kermani, B. C. Wheeler, K. Badie, and R. M. Hashemi, "EMG feature evaluation for movement control of upper extremity prostheses," *IEEE Trans. Rehab. Eng.*, vol. 3, no. 4, pp. 324–333, Jul. 1995.

[21] M.-F. Lucas, A. Gaufriau, S. Pascual, C. Doncarli, and D. Farina, "Multichannel surface EMG classification using support vector machines and signal-based wavelet optimization," *Biomed. Signal Processing Contr.*, vol. 3, no. 2, pp. 169–174, Apr. 2008.

[22] J. Rafiee, M. A. Rafiee, N. Prause, and M. P. Schoen, "Wavelet basis functions in biomedical signal processing," *Expert Systems With Applications*, vol. 38, no. 5, pp. 6190–6201, May 2011.

[23] R. Rifkin, G. Yeo, and T. Poggio, "Regularized least squares classification," in *Advances in Learning Theory: Methods, Model and Applications*. Amsterdam, The Netherlands: IOS Press, 2003, vol. 190, pp. 131–154.

[24] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[25] C. Saunders, A. Gammerman, and V. Vovk, "Ridge regression learning algorithm in dual variables," in *ICML'98: Proc. Fifteenth Int. Conf. Machine Learning*, 1998, pp. 515–521.

[26] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[27] J. R. Martin, M. L. Latash, and V. M. Zatsiorsky, "Interaction of finger enslaving and error compensation in multiple finger force production," *Experimental Brain Res.*, vol. 192, no. 2, pp. 293–298, 2009.

[28] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1995.

[29] R. Kõiva, B. Hilsenbeck, and C. Castellini, "Evaluating subsampling strategies for SEMG-based prediction of voluntary muscle contractions," in *Proc. Int. Conf. Rehab. Robotics (ICORR)*, 2013.

[30] J. Silva, W. Heim, and T. Chau, "MMG-based classification of muscle activity for prosthesis control," in *Proc. 26th Annu. Int. Conf. IEEE Eng. in Medicine Biology Soc. (IEMBS)*, 2004, vol. 1, pp. 968–971.

[31] S. B. Thies, P. Tresadern, L. Kenney, D. Howard, J. Y. Goulermas, C. Smith, and J. Rigby, "Comparison of linear accelerations from three measurement systems during "reach & grasp"," *Med. Eng. Phys.*, vol. 29, no. 9, pp. 967–972, Nov. 2007.

[32] A. O. Posatskiy and T.-H. Chau, "The effects of motion artifact on mechanomyography: A comparative study of microphones and accelerometers," *J. Electromyography Kinesiol.*, vol. 22, no. 2, pp. 320–324, 2012.

[33] D. S. Childress and R. F. Weir, "Control of limb prostheses," *Atlas of Limb Prosthetics*, vol. 2, pp. 175–198, 2004.

[34] G. Li, A. E. Schultz, and T. A. Kuiken, "Quantifying pattern recognition-based myoelectric control of multifunctional transradial prostheses," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. 18, no. 2, pp. 185–192, 2010.

[35] A. M. Simon, L. J. Hargrove, B. A. Lock, and T. A. Kuiken, "Target achievement control test: Evaluating real-time myoelectric pattern-recognition control of multifunctional upper-limb prostheses," *J. Rehab. Res. Devel.*, vol. 48, no. 6, pp. 619–627, 2011.

[36] N. Jiang, I. Vujaklija, H. Rehbaum, B. Graimann, and D. Farina, "Is accurate mapping of EMG signals on kinematics needed for precise online myoelectric control," *IEEE Trans. Neural Syst. Rehab. Eng.*, vol. PP, no. 99, pp. 1–1, 2013.