Exploring the relationship between EMG feature space characteristics and control performance in machine learning myoelectric control

Franzke, A W; Kristoffersen, M B; Jayaram, V; Van Der Sluis, C K; Murgia, A; Bongers, R M

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2021

Exploring the relationship between EMG feature space characteristics and control performance in machine learning myoelectric control.

A.W. Franzke, M.B. Kristoffersen, V. Jayaram, C.K. van der Sluis, A. Murgia, *Member, IEEE,* R.M. Bongers

Abstract— In myoelectric machine learning (ML) based control, it has been demonstrated that control performance usually increases with training, but it remains largely unknown which underlying factors govern these improvements. It has been suggested that the increase in performance originates from changes in characteristics of the Electromyography (EMG) patterns, such as separability or repeatability. However, the relation between these EMG metrics and control performance has hardly been studied. We assessed the relation between three common EMG feature space metrics (separability, variability and repeatability) in 20 able bodied participants who learned ML myoelectric control in a virtual task over 15 training blocks on 5 days. We assessed the change in offline and real-time performance, as well as the change of each EMG metric over the training. Subsequently, we assessed the relation between individual EMG metrics and offline and real-time performance via correlation analysis. Last, we tried to predict real-time performance from all EMG metrics via L2-regularized linear regression. Results showed that real-time performance improved with training, but there was no change in offline performance or in any of the EMG metrics. Furthermore, we only found a very low correlation between separability and real-time performance and no correlation between any other EMG metric and real-time performance. Finally, real-time performance could not be successfully predicted from all EMG metrics employing L2-regularized linear regression. We concluded that the three EMG metrics and real-time performance appear to be unrelated.

Index Terms — Electromyography, Machine Learning, Pattern Analysis, Prosthetics, Training

I. INTRODUCTION

Machine learning (ML) based control strategies have been suggested as an advanced control paradigm for myoelectric upper limb prostheses which could make prosthesis operation more intuitive and faster compared to conventional, direct control (DC) [1]–[7]. Nonetheless, recent studies suggest that user training and guidance by a coach is needed to learn ML prosthesis control [8]–[11].

It is yet not well understood which underlying factors and mechanisms influence an individual's ability to control a ML-based myoelectric device. Such knowledge might help improve and individualize training of prosthetic users. It seems widely accepted that an individual's ability to generate consistent and distinguishable patterns of electromyographic (EMG) signals is necessary for successful ML-based myoelectric control [10], [12]–[14]. However, the relation between such EMG signal characteristics and prosthesis control ability is hardly studied [12], [15].

In ML-based myoelectric control, most algorithms are trained with EMG features that are extracted from the time-windowed raw EMG signals. The ML algorithm tries to distinguish the different movement classes in the EMG feature space and it has been shown that some characteristics of the EMG feature space, depicted by metrics which estimate EMG pattern qualities such as separability, are strongly related to the ML algorithm's offline performance [15].

It has also been argued that such feature space characteristics reflect the user's ability to control a myoelectric device and that changes in control ability might be related to changes in those characteristics [12], [14], [16]. Undeniably, EMG feature space characteristics might be a promising and suitable target in rehabilitation training since they are easy to define, quantify and monitor. Therefore, understanding the link between the EMG feature space and ML myoelectric control performance might offer theoretical support for designing training methods which address device controllability in rehabilitation training.

However, whereas strong relations have been found between feature space characteristics and offline performance, inconclusive results have been found with regard to the relation between characteristics of the feature space and users' control ability in real time evaluations, i.e. with the user being in the control loop [12], [14]–[16]. This distinction between offline and real-time (i.e., online) performance is important as it has been shown that offline performance is a poor predictor of real time control ability [17]. Moreover, only few studies attempted to quantify the relation between EMG feature space characteristics and real time control ability [12], [15], [18].

The results of these studies on able-bodied persons showed a significant correlation between real-time performance and an

T

EMG feature space metric which estimated the separability between movement classes in the feature space [12], [15]. This separability index (SI) was moreover found to be significantly higher in subjects who had experience with ML myoelectric control, compared to naïve subjects [12]. However, the correlation between the separability index and real-time performance was only moderate in both studies (r = 0.53 and r = 0.54, respectively) and much lower than the correlation between the separability index and offline performance (r = 0.94) [15]. Remarkably, both studies showed that several participants achieved high real time performance despite a low separability index. Moreover, in another recent work it was found that between-class separability changed after training, but the authors found no correlation between separability and real-time performance [18]. The same study found a significant, but low correlation between a metric for EMG pattern variability and real-time performance [18].

In summary, there is some support for a link between EMG feature space metrics and real-time performance, but the results so far remain inconclusive and the correlations which have been found were only low to moderate. Moreover, no attempt has been made so far to investigate the relation between a combination of feature space metrics and real-time performance, although it has been shown that changes in EMG feature space characteristics can greatly differ between individuals, thus the relevant feature space metric combinations might vary considerably between persons [14].

The goal of this study was to establish whether a relation exists between the EMG feature space metrics and offline as well as real-time performance in individuals who learn ML myoelectric control during repeated training sessions. We focused on the EMG feature space metrics which were suggested in the literature, i.e. metrics related to separability, variability and repeatability of the movement classes in the feature space. We first examined whether any of the performance measures and the EMG feature space metrics changed during the training. Subsequently, we assessed the relation between the performance measures and each of the three EMG metrics. Finally, we analyzed whether real-time performance could be predicted from all EMG feature space metrics via L2-regularized linear regression.

We hypothesized that both, offline and real-time performance, would increase during the training. Furthermore, we hypothesized that separability would increase and variability as well as repeatability would decrease as a response to the training. We also hypothesized that strong correlations would be found between offline performance and separability, but only a low to moderate correlation would be found between any of the EMG metrics and real-time performance. In contrast, we hypothesized that real-time performance could be predicted with the help of machine learning methods when taking all EMG feature space metrics into account.

II. Methods

A. Ethical approval

The local ethics committee approved the study (ECB/2017.01.12_1). All participants were informed about content, procedure and goal of the study. All subjects gave written consent prior to the start of the experiment.

B. Subjects

Twenty able-bodied university students (mean age in years: 22($\pm$2.8), 11 females) were recruited. Handedness was assessed by completing the handedness questionnaire of the Edinburgh inventory [19], [20]. Exclusion criteria were any neurological pathologies or musculoskeletal complaints interfering with study outcomes.

C. Myoelectric machine learning system

Eight commercially available double differential electrodes were used (13E200=50 AC, Otto Bock Healthcare Products GmbH, Vienna, Austria), which pre-amplified and band-pass filtered the EMG signals. The electrodes were placed at equal distances around the thickest part of the forearm. EMG data were sampled at 1000 Hz and streamed to a laptop computer via Bluetooth connection. Software provided by Otto Bock Healthcare Products GmbH (Vienna, Austria) was used to record EMG data, to train a classifier and to run a match-prompt test. EMG data were divided into sliding time-windows of 128 ms with a 32 ms overlap. Four time-domain features (mean absolute value, waveform length, slope sign changes and zero crossings) were extracted, yielding a 32-dimensional feature space [21]. A linear-discriminant-analysis (LDA) classifier was used to classify each feature vector as one of seven movement classes (wrist supination, wrist pronation, wrist flexion, wrist extension, hand open, fine pinch grip, and lateral thumb grip) and a "rest" class [2], [21]. Moreover, a proportionality estimator was employed to facilitate proportional control. Details of the system are described in Amsuess et al. [5]. The overall system delay was 150 ms, which is considered within the optimal range for myoelectric control [22], [23].

D. Study Protocol

The experiment was conducted over five consecutive days. Each day, subjects completed a session of 45-60 minutes which consisted of three blocks. Hence, each participant completed 15 blocks in total. Each block consisted of one system training and a following match prompt test. The ML system was retrained with one data set of movements only, thus data from prior blocks or prior days were not used during system training. This procedure was chosen to enable us to study the precise relation between one set of EMG system training data and the corresponding performance in a match-prompt test (See D.2: Match-prompt test). Before the start of the experiment, all subjects received a general introduction to myoelectric ML control by one of the experimenters. It was explained what myoelectric signals are and how pattern recognition in myoelectric control works. Furthermore, the concepts of pattern separability, variability, and repeatability were explained.

1) System Training

Subjects were seated on a comfortable chair with their arms resting on an armrest, yielding an elbow flexion angle of approximately 90 degrees. EMG data were recorded from the participant's non-dominant arm. Moreover, wrist and fingers were splinted with a medical brace to avoid the effects of joint movements on the EMG data [24]. Before the start of each block, the RMS of each electrode was briefly visually inspected in an 8-axis plot (Top left panel in Figure 1) to ensure that there

was no noise or "dead channel" e.g. due to bad electrode-skin contact. The experimenter furthermore monitored the RMS of all electrodes during each block to ensure that there were no abnormalities in the signals. Each system training started with a calibration where participants were instructed to perform each movement at 100% maximum voluntary contraction (MVC) while avoiding arm trembling or muscle pain. Subsequently, participants were asked to repeat each movement at 30%, 60% and 90% proportion of their MVC. Thus, in total, the participants performed 21 movements (7 different movements at 3 different contraction strength levels each), and three repetitions without any muscle contractions (rest). Continuous feedback about their EMG pattern was provided by an 8-axis plot which displayed the root mean square of each electrode on one axis (Figure 1, top-left panel). Continuous feedback about the contraction strength was provided through a red dot, which moved up and down proportionally to the achieved percentage of their MVC. The red dot moved over time from left to right in front of a blue trapezoid shape. Participants were asked to try to follow the blue line by adjusting their contraction strength. (Figure 1, top-right panel). After data collection, the LDA classifier was trained. The offline performance was calculated from the system training data only (see E: performance metrics).
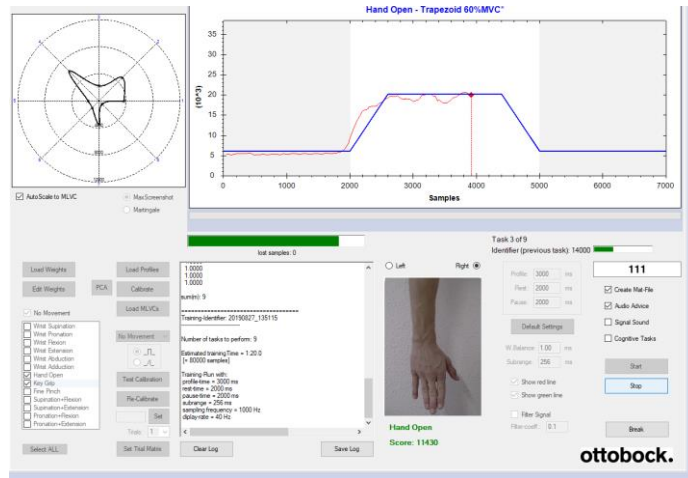

Figure 1: Recording of training data. The eight axes in the top left figure correspond to the eight electrodes' root mean square (where the center of the figure represents zero activity and the outer boundary represents 100% of their MVC). The blue line in the right top shows the target contraction strength, whereas the red dot shows the participants' achieved contraction strength (indicated as percentage of their MVC). The picture in the lower middle shows the target movement.

2) Match Prompt Test

After each system training, participants completed a match prompt test (Figure 2), which was similar to the motion test by Kuiken et al. [25]. The participants were asked to perform one out of the seven different movements, at one out of the three different contraction strength levels (yielding in total 21 different cues in each match prompt test). The prompted movement was presented as a picture on the screen, the expected contraction strength was presented as marks on a vertical bar. The movement which was predicted by the LDA classifier was continuously presented as a picture. The

contraction strength was continuously displayed in the form of a vertical bar (Figure 2).

Participants were given three seconds to perform the corresponding movement. Importantly, there was a two-second pause after each prompt, which allowed the participants to study the prompted movement. Moreover, the participants were told that the three-second timer didn't start to run as long as the "rest" class was classified. This meant that additionally to the two-second pause, the participants could take their time to focus on the prompted movement and decide when they would start their attempt to perform the corresponding contraction. The next cue was presented when the three-second timer had passed or when participants managed to continuously hold the correct movement and the prompted contraction strength level for two seconds. The contraction strength was considered correct if the estimated level was within a defined margin of the prompted level. Those margins were 15%, 20% and 30% for the prompted levels of 30% MVC, 60% MVC and 90% MVC, respectively. The real-time performance was calculated based on the match-prompt data only (see E: performance metrics).
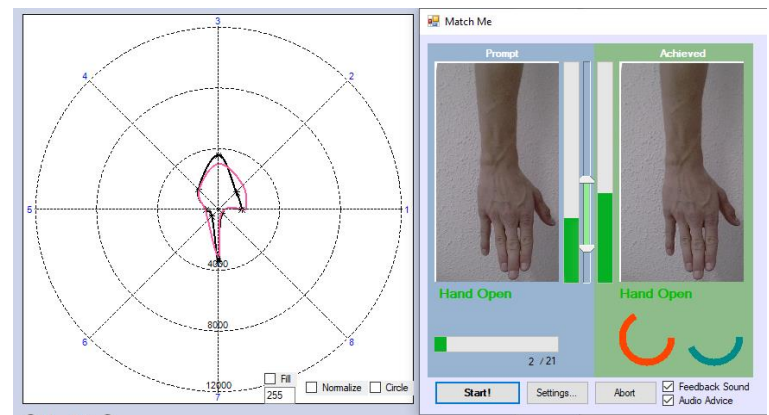

Figure 2: Match prompt test. Left panel: The eight axes correspond to the eight electrodes' root mean square (where the center of the figure represents zero activity and the outer boundary represents 100% of the participant's MVC). The black line continuously displayed the current EMG activation pattern, whereas the colored line displayed the mean EMG pattern of the movement which was classified.
Right panel: The prompted movement is shown in the left picture, whereas the classified movement is shown in the right picture. The prompted contraction strength level is shown in the left green bar, whereas the achieved contraction strength is indicated by the right green bar. The contraction strength tolerance level is indicated by the two white blocks in between both green bars. The growing red circle displayed the time which was available to complete the current attempt, the growing blue circle indicated the duration of how long the correct movement and the correct contraction strength were maintained.

3) Feedback After Match Prompt Test

After each match prompt test, the participants received feedback about their EMG patterns. This feedback was based on visual inspection of the EMG patterns in an 8-axis plot (similar to the left panel in Figure 2) and three different EMG metrics. These metrics were calculated from the EMG data of the most recent system-training and match prompt test. They reflected the following three characteristics of EMG data: (1) Separability, i.e. the distance between two different classes in the feature space, (2) variability, i.e. the variance around the class mean, and (3) repeatability, i.e. the similarity between a class in the training data and its repetition in the match prompt

test [12]. Separability and variability were calculated within the training data set, whereas the repeatability was calculated within training and test data. The test data only contained those parts of the EMG data where the three-second timer was running, i.e. where the participants were attempting to perform a prompted movement. The definitions of these metrics are described in *F) Feature space metrics*. The participants were told which class showed the lowest separability, which class showed the highest variability, and which class showed the highest repeatability. This procedure was chosen so that the participants could focus on the most problematic classes, and to not overwhelm the participants by showing three metrics for all of the seven movement classes. The participants were then asked to focus in the next block on these specific movement classes to make them less variable ("try to hold the contraction in a more stable manner for this movement") and more repeatable ("for this movement, try to repeat the same contraction in the match prompt test, compared to the training phase"). Moreover, they received guidance on how to alter one movement to make it more separable from a conflicting movement based on suggestions in the literature [13], [14]. In order to illustrate a simplified version of the feature space, the participants were shown the root-mean-square of the raw EMG for each class, displayed in an 8-axis plot (similar to the left panel in Figure 2). For each class, the average of the three different contraction strength level was shown as a static line, and each class was depicted in a different color. The participants then had up to two minutes to inspect how their EMG patterns corresponded to changes in the contraction.

E. Performance Metrics

1) Offline performance

The software which was used during the system training and the match prompt test did not offer an offline-analysis framework. Therefore, the offline analysis of EMG training data was performed in BioPatRec, an open-source EMG pattern recognition software suite [26]. The features were calculated from consecutive time windows of 128 ms, with a 32 ms overlap. The feature vectors of different contraction strength levels belonging to the same movement class were then added to form one movement class. Subsequently, the feature vectors were randomly assigned to a training, validation and testing set. This procedure was repeated 10 times as a cross-validation to remove any bias of the randomly assigned feature vectors [15], [26].

Offline accuracy (OffAcc) was then defined as the ratio of correct predictions over total predictions:

$$OffAcc = \frac{correct\ predictions}{total\ predictions}$$

2) Real-time performance

The real-time performance was based on the EMG data from the match prompt test and defined as the online accuracy (OnAcc). It was calculated as the percentage of time-windows where the predicted movement matched the prompted movement.

$$OnAcc = \frac{correct\ predictions\ (in\ match\ prompt\ test)}{total\ predictions\ (in\ match\ prompt\ test)}$$

F. Feature Space Metrics

For all feature space metrics, data recorded from different contraction strength levels belonging to one movement were added to form one movement class. Moreover, for the feedback provided after each match-prompt test, the feature space metrics were calculated for every class individually to identify the movement which showed the worst result in each metric. However, for the data analysis, the feature space metrics were always averaged over all seven movement classes.

1) Separability Index

The Separability Index (SI) was suggested to reflect distances between different movement classes in the EMG feature space [12]. It was calculated as the average distance of all movement classes to their most conflicting neighbor. In the initial definition only the covariance of one movement class was taken into account [12]. We used an adapted version of the separability index where the covariance of both compared movement classes is considered [15]. The adapted SI was then defined as:

$$SI\_adapted = \frac{1}{7} \sum_{j=1}^{7} \left(\frac{1}{2} \sqrt{\left(\mu_j - \mu_{Cj} \right)^T S^{-1} \left(\mu_j - \mu_{Cj} \right)} \right)$$

Where μ_j is the centroid of class j, μ_{Cj} is the centroid of the most conflicting class (with respect to class *j*), and *S* is defined as:

$$S = \frac{S_j + S_{Cj}}{2}$$

Where S_j is the covariance of class j, and S_{Cj} is the covariance of the most conflicting class (with respect to class *j*).

2) Mean-semi-principal axis (MSA)

The MSA was proposed as a measure for intra-class variability [12]. It considers the feature vectors of each class as a cluster in the shape of a hyper-ellipsoid. The size of the cluster is approximated through singular value decomposition of the feature vector and subsequently calculating the geometric mean of the singular values.

$$MSA = \frac{1}{7} \sum_{j=1}^{7} \left(\prod_{k=1}^{n} a_{jk} \right)^{\frac{1}{n}}$$

Where a_{jk} is the *k*th of *n* singular values of class *j* and *n* is equal to the number of dimensions (i.e. in this study, *n* was equal to 32, as EMG was measured with eight electrodes and four time-domain features were calculated).

3) Repeatability Index

The repeatability index measures how well the participants reproduced EMG patterns in the match prompt test (test data),

compared to the EMG pattern they produced during the system training procedure (training data) [12]. It was calculated as half the Mahalanobis distance between the feature vector centroid of a movement class in the training data and the feature vector centroid of the same movement class in the test data, averaged over all seven movement classes. Similarly to the adapted SI, we used an adapted version of the RI, where the covariance of both classes is taken into account:

$$RI\_adapted = \frac{1}{7} \sum_{j=1}^{7} \left(\frac{1}{2} \sqrt{\left(\mu_{TRj} - \mu_{TEj} \right)^T S^{-1} \left(\mu_{TRj} - \mu_{TEj} \right)} \right)$$

Where μ_{TRj} is the centroid of class j in the training data, μ_{TEj} is the centroid of class j in the test data, and S is defined as:

$$S = \frac{S_{TRj} + S_{TEj}}{2}$$

Where S_{TRj} is the covariance of class j in the training data and S_{TEj} is the covariance of class j in the test data.

G. Statistics and analysis

During the experiment it was discovered that for the first eight participants, the feedback given after each match-prompt test was wrongly calculated due to a coding error. The SI_adapted, RI_adapted, and MSA were calculated based on one feature only (mean absolute value) instead of the whole feature set, as we intended. This was then corrected so the remaining participants received feedback based on the entire feature set. However, before conducting any further analysis of the data, we first assured the validity of the data by assessing if the differences in feature selection had any effect on the real-time performance improvements over time (see next paragraph). The corresponding methods and results are described in the appendix at the end of the paper. As we found no effect of the feature selection difference on real-time performance, we performed the main data analyses based on all participants. For all the analyses reported the significance level was set to 0.05.

1) Change in performance outcomes and EMG feature space metrics

To test whether any of the performance outcomes and the EMG feature space metrics changed over the 15 blocks, we performed planned comparisons (linear contrast) in five separate one-way repeated measure ANOVAs on each performance outcome and each EMG metric, with block (block 1, block 2,… block 15) as within-subject factor. If Mauchly's test indicated that the sphericity assumption was violated, the Greenhouse-Geisser correction was used to estimate the degrees of freedom. Effect sizes were calculated using generalized eta-squared statistics [27]. We were primarily interested in the planned linear contrast, but to provide as much information as possible, we also report the significant omnibus effect.

2) Partial correlation Between Performance Outcomes and EMG Feature Space Metrics

We assessed the strength of the relation between each EMG feature space metric and each performance outcome by calculating the Pearson's product-moment correlation coefficient for each pair of performance outcome and EMG feature space metric.

3) Predicting real-time performance from all EMG feature space metrics

To test whether real-time performance could be predicted from the combination of all EMG feature space metrics, we attempted to fit a linear function via L2-regularized linear regression that predicted the real time performance using the three EMG feature space metrics generated from test and training data. Results were generated using 10-fold cross validation, with the fitting of the regularization parameter done by an inner loop with 3-fold cross-validation. To see whether we were successful at predicting real-time performance, we calculated the correlation of the predicted performances with the real performances. Significance was calculated via permutation-based null distribution testing with 100 permutations: the true correlation coefficient was compared against correlation coefficients calculated based on data where the true labels were shuffled, and the p-value reflects what percentage of shuffled data correlation coefficients were larger than the true one.

III. RESULTS

Twenty participants completed all 15 blocks, for a total of 300 blocks (i.e. system trainings and match-prompt tests) being performed. Data of four blocks of different participants were not saved correctly and one block was removed from the analysis, as the participant achieved 12.4% online accuracy and we believe that such low real-time performance was due to a lack of focus or that the participant mixed up some of the movements. Therefore, in total 295 blocks were analyzed.

Moreover, due to an issue with the automatic naming of the EMG files from the motion test, these files were often overwritten and therefore not stored permanently, so the repeatability index (RI, adapted) could only be computed for 115 blocks. The missing files were randomly distributed over participants and blocks.

A. Change in performance outcomes and EMG feature space metrics

The repeated measure ANOVA for the offline performance showed a significant effect for block (F(14, 266) = 2.277; p = .006; $\eta_G^2 = 0.107$) but the planned linear contrast of blocks showed no significant difference. No significant effects were found in the separate repeated measure ANOVAs for any of the three EMG feature space metrics. The repeated measure ANOVA for real-time performance revealed a significant effect of block (F(14, 266) = 11.256; p < .0001; $\eta_G^2 = 0.372$). The planned linear contrast of blocks showed a significant difference (F(1, 19) = 50.437; p < .0001; $\eta_G^2 = 0.726$). Figure 3 shows data of all performance outcomes and EMG feature space metrics for all blocks.

B. Partial correlation between the EMG feature space metrics and offline/real-time performance

A significant correlation between the SI (adapted) and offline accuracy was found (r = 0.72, p < 0.0001, 95% CI [0.71, 0.78]). No significant correlation was found between the other two EMG feature space metrics and offline accuracy (Figure 4, left panels).

There was a low but significant correlation between the SI (adapted) and real-time performance (r = 0.27, p < 0.0001, 95% CI [0.10, 0.26]). No significant correlation was found between the other two EMG feature space metrics and real-time performance (Figure 4, right panels).

C. Predicting real-time performance from all EMG feature space metrics

The predicted real-time performance showed only poor correlations to the true online accuracy for the movements wrist-extension (r = 0.13) and fine-pinch (r = 0.2). No significant correlations were found between the predicted real-time performance and the true real-time performance for all other movements. Figure 5 shows the correlations between predicted and true real-time performance for each individual movement (panel 1 to panel 7), and for the average of all seven movements (panel 8), respectively.

IV. DISCUSSION

Our aim was to establish whether a relationship existed between three common EMG feature space metrics and offline as well as real-time performance in myoelectric ML based control. Whereas a strong correlation was found between offline performance and separability, none of the three EMG feature space metrics showed a strong correlation with real-time performance. Moreover, real-time performance could not be predicted from the three EMG feature space metrics with the help of machine learning methods. Our findings suggest that the EMG feature space characteristics which were estimated with the three metrics SI (adapted), RI (adapted) and MSA do not appear to accurately reflect the gains in real-time performance in ML myoelectric control.

A. Improvements in real-time performance vs. no changes in offline performance or EMG feature space metrics

Our analyses showed that real-time performance (depicted by the online accuracy) increased over the 15 blocks, which indicates that the participants' skill in controlling the output of the myoelectric ML system improved. It should be noted that all participants were naïve to myoelectric control and the match prompt test. Therefore, the increase in online accuracy indicates that the participants got more familiar with the task and found ways to improve their performance. However, neither the offline-performance, nor any of the EMG feature space metrics showed a significant change over the 15 blocks. This finding suggests that offline performance appears to be unrelated to real-time performance [17]. Moreover, this finding is in agreement with [12], where it was found that none of the three

metrics showed a significant change over two days of training myoelectric ML control.

However, our results differ from two other studies, where changes of the EMG feature space were studied during repeated training [16], [18]. He et al. [16] found that the repeatability of movement classes decreases over training, whereas the separability between movement classes did not show a significant change. The participants in the study of He et al. received no feedback about their EMG, which might explain the different findings with regard to our study. This interpretation is supported by Kristoffersen et al. [18], where groups who received different types of feedback showed different changes in EMG characteristics. In agreement with the literature, the current study shows that real-time performance increases with training, but it appears that changes in EMG feature space characteristics as a response to training are strongly dependent on the type of feedback. The absence of change in any of the EMG metrics in our study is particularly surprising because the participants were specifically coached to improve EMG pattern separability, variability and repeatability. This might imply that explicitly providing feedback about these EMG metrics is not effectively making individuals change their EMG patterns. Moreover, the absence of change in EMG metrics in parallel with a significant change in real-time performance implies that these EMG metrics and real-time performance seem to be unrelated.

B. Correlation between EMG feature space metrics and real-time performance

We found that neither the repeatability, nor the within-class variance showed a significant correlation to real-time performance. Furthermore, in contrast to Bunderson et al. [12] and Nilsson et al. [15], we only found a low correlation between separability and real-time performance, despite an identical separability definition in our study and that of Nilsson et al. It is noteworthy however that Nilsson et al. used time-to-completion as real-time performance metric, which is different from online accuracy, which was used in our study and Bunderson et al. Nonetheless, the two above studies and our study appear to show a similar pattern with regard to the separability, that is: whereas the overall probability of achieving high real-time performance seems to increase with higher separability, the latter does not guarantee high performance and there is large variance in the correlation. Furthermore, high real-time performance is possible despite relatively low separability. It appears that none of the EMG metrics studied here have high predictive power with regard to real-time performance. A possible explanation for the absence of any strong correlation between the EMG metrics and real-time performance might be that rather than a straightforward relation, there is a complex interplay between the EMG feature space characteristics which determines real-time performance.

In order to test this hypothesis, we tried to predict the real-time performance from all three EMG feature space metrics via L2 regularized linear regression. However, we found that the real-time performance could not be successfully predicted, as we found no, or only poor correlations between the predicted and the true real-time performance. Altogether, these results
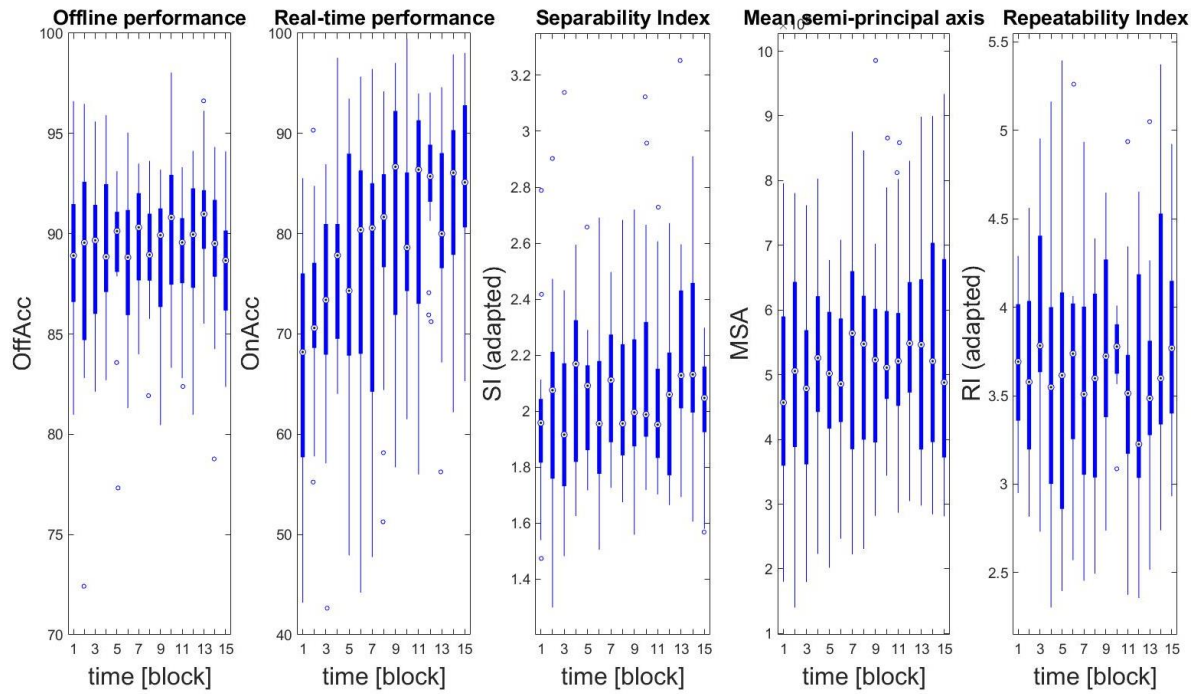
Figure 3: Performance outcomes and EMG feature space metrics over all blocks. The edges of the blue boxes indicate the 25th and the 75th percentiles, respectively. The white circles in the blue boxes indicate the median. The blue whiskers indicate the most extreme datapoints (which are not considered outliers), whereas the blue circles indicate outliers.

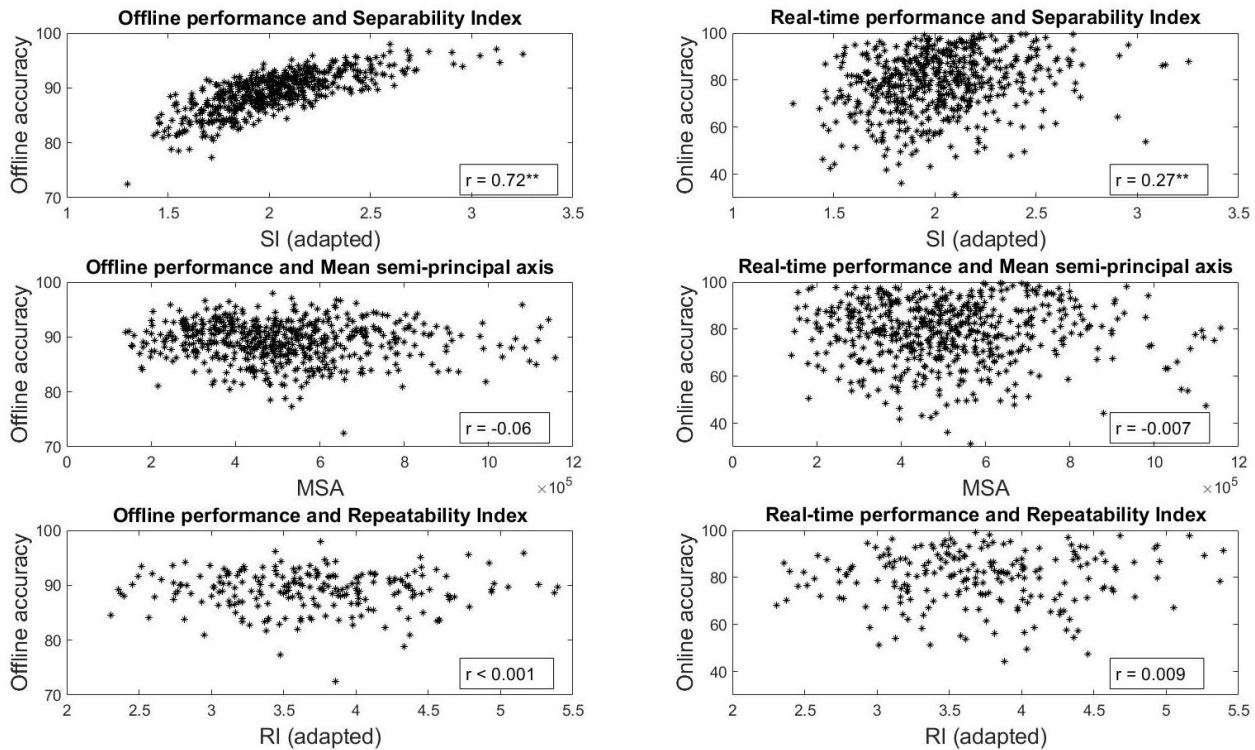Figure 4: Correlations between the three feature space metrics and offline accuracy are shown in the three plots on the left side. Correlations between the three feature space metrics and online accuracy are shown in the three figures on the right side. Significance of correlations is indicated by asterisks: ** $p < 0.01$.
Abbreviations: SI: separability index. MSA: mean-semi-principal axis. RI: repeatability index.
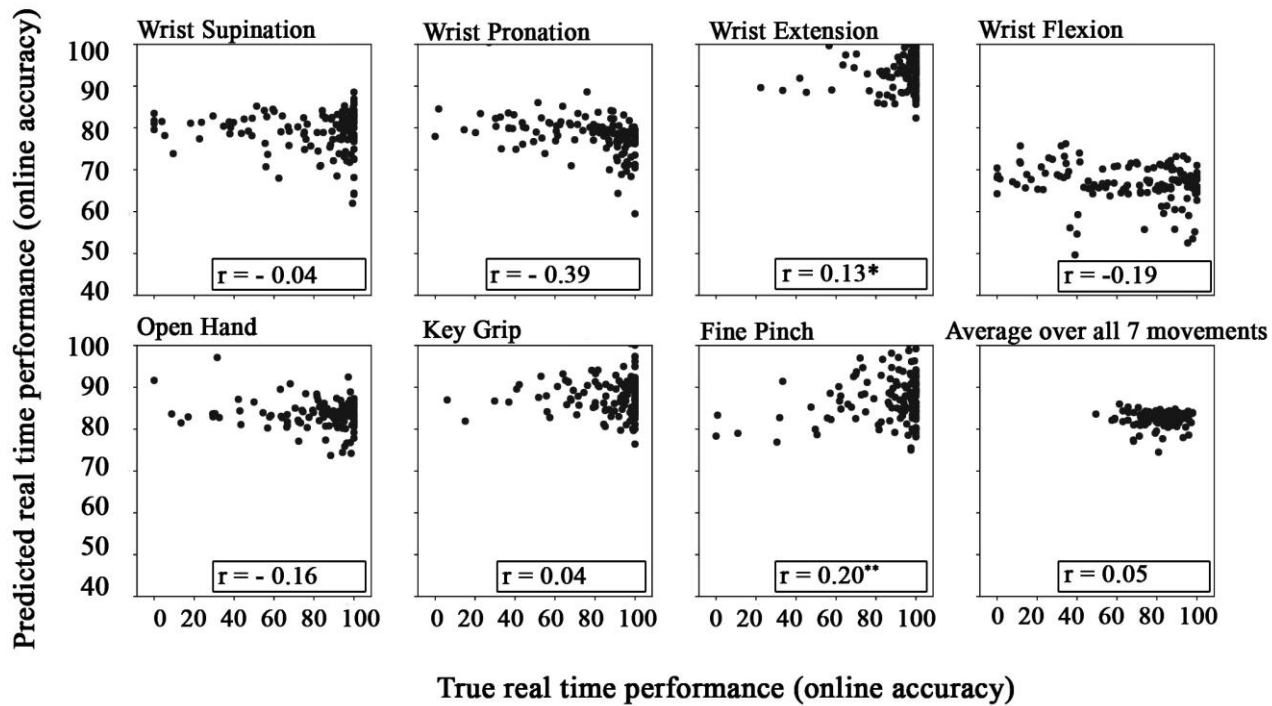
Figure 5: Correlation between predicted online accuracy and true online accuracy for each individual movement class and for the average over all seven movement classes, respectively. Significance of correlations is indicated by asterisks: *p < 0.05, ** p < 0.01.

show that none of the EMG feature space metrics alone, or in combination, appear to be predictive with regard to real-time performance.

C. Absence of strong correlations between EMG feature space metrics and real-time performance

Both our study and the study of Nilsson et al. [15] showed strong correlations between separability and offline performance, but only low to moderate correlations between separability and real-time performance. It appears that as soon as the user is in the control loop and receives feedback about the output of the machine learning algorithm, the EMG patterns are affected and the relation between EMG characteristics and performance changes. Indeed, Nilsson et al. noted a shift in distributions between the time when data was recorded and the time when the match-prompt test was executed [15]. Similarly, we found that the repeatability is on average nearly twice as high as the separability (see panel 3 and 5 in Figure 3), which means that the distance between a movement class in the training data and the same class in the match-prompt test data is considerably larger than the distance between a movement class in the training data and its most conflicting neighbor class. Nilsson et al. attributed this shift in distributions to the time between the recording of the training data and the execution of the match-prompt test. However, this time gap was very short in our experiment (usually less than two minutes). Moreover, the experiment was carried out in a highly controlled lab condition, where able-bodied participants were comfortably seated and had their arm in a relaxed position. Therefore, we believe that it is the feedback during the match-prompt test and the participants' continuous response and adaptation to that feedback which caused the drift in distributions and therefore a

relatively large repeatability. This interpretation is supported by the results of He et al. [16], who found that the repeatability index significantly decreased (i.e. improved) over time when participants received no feedback. Such a distribution drift might also explain why the EMG feature space metrics such as the separability and the variability of movement classes in the training data are of little predictive power with regard to real-time performance.

D. Implications

It appears that the metrics which have proven useful in predicting a machine learning algorithm's offline performance (e.g. separability) [15], are of little use in predicting the real-time performance of myoelectric ML based control. Therefore, these metrics might be ill-suited to assess an individual's skill in controlling a myoelectric ML device, even though it is commonly assumed that, for successful ML myoelectric control, EMG patterns must be separable, consistent and repeatable. However, it is important to note that the overall offline accuracies in this study were relatively high from the start (between 85% and 95%). Such high offline accuracies have been found to generally yield well controllable systems [28]. Therefore, the lack of predictive power of the chosen EMG metrics might be specific for individuals and control systems who achieve such high offline accuracies. Also, the finding that the participants' improvements in real-time control were not governed by changes in the chosen EMG metrics does not imply that changes in these metrics cannot lead to improvements in real-time control. Therefore, the field of ML myoelectric control might benefit from further studies which explore the changes in EMG characteristics as a function of training and investigate the relation between EMG

characteristics and real-time performance. Such knowledge could prove crucial in improving the training paradigms for ML myoelectric control.

Likewise, it would be important to deepen the knowledge of how EMG patterns are affected by feedback [18]. Many studies about ML myoelectric control were conducted on offline data without the user in the loop, but it remains questionable how well offline scenarios predict real-time control performance. There is evidence for a correlation between offline accuracy and completion rate in able-bodied who perform the Target Achievement Test [23], [28]–[30], whereas other studies conclude that offline analyses are poorly connected to real-time performance [15], [17], [31]. One of the reasons for inconclusive results and a potential lack of correlation might be the response of individuals to feedback and therefore a change in the EMG activity during real-time control, causing a drift in the feature space distributions [15].

Finally, we also encourage researchers to explore new routes in the training paradigms. Our results suggest that providing explicit feedback about the EMG feature space characteristics and qualitative feedback about the EMG is not very effective in causing a corresponding change in the EMG. It might be that the improvements in real-time performance which we observed are merely the result of repeatedly performing the task, rather than the result of our specific coaching. This interpretation is supported by studies which show that performance can increase during training without feedback [16], [18]. Therefore, it might be fruitful to explore serious-games or other training approaches which can facilitate implicit learning strategies.

E. Limitations

The results might be specific for individuals with intact limbs and not directly transferrable to individuals with amputation, since the altered anatomy and physiology of affected limbs might have an effect on the characteristics of EMG [32]. Therefore, we cannot rule out that the absence of a correlation between the EMG feature space metrics and the real-time performance reported here, would be different in individuals with amputation.

This study used a virtual match-prompt test to assess real-time performance which is different from actual prosthetic use, the ultimate goal of rehabilitation training [10], [31]. However, we chose the match prompt task to avoid the influence of confounding factors on performance, such as skin/electrode shift or the effects of upper limb kinematics on the EMG. Finally, we used a classic (LDA) pattern recognition system as this has become the golden standard in the field and is the most widely used type in commercial ML myoelectric control. However, it might be that the relation between EMG feature space metrics and real-time performance is strongly dependent on the type of machine learning algorithm, so the results might not be directly transferrable to fundamentally different algorithms, e.g. regression-based or neural-network techniques.

APPENDIX

To test whether the difference in feature selection had an effect on real-time performance, we carried out the following analysis: The participants were divided into two groups; one

group which had received feedback based on one feature only (group: "one-feature-feedback") and a group which had received feedback based on the entire feature set (group "all-features-feedback"). Then, a mixed-model ANOVA was performed with block (block 1, block 2, … block 15) as within-subject factor and feedback (one-feature-feedback vs. all-features-feedback) as between-subject factor on real-time performance (online accuracy). If Mauchly's test indicated that the sphericity assumption was violated, the Greenhouse-Geisser correction was used to estimate the degrees of freedom. Effect sizes were calculated using generalized eta-squared statistics [27]. If the feature selection difference in the feedback would have had an effect on the performance, we expected the main effect of feedback as well as the interaction effect to be significant.

The mixed-model ANOVA results showed a significant main effect for block on the real-time performance: $F(14, 252) = 10.251; p < 0.0001; \eta_G^2 = 0.363$. No significant effect of group on real-time performance was found: $F(1, 18) = 3.824; p = 0.66; \eta_G^2 = 0.175$ and no significant interaction effect between block and group was found: $F(14, 252) = 0.555; p = 0.898; \eta_G^2 = 0.30$. Based on these results we concluded that the real-time performance was not influenced by the differences in feature selection. Most importantly, they did not appear to have an impact on performance improvements over time.

V. REFERENCES

[1] F. R. Finley and R. W. Wirta, "Myocoder-computer study of electromyographic patterns.," *Arch. Phys. Med. Rehabil.*, vol. 48, no. 1, pp. 20–24, Jan. 1967.

[2] 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, 2003.

[3] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *J. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 643–660, 2011.

[4] N. Jiang, H. Rehbaum, I. Vujaklija, B. Graimann, and D. Farina, "Intuitive, online, simultaneous, and proportional myoelectric control over two degrees-of-freedom in upper limb amputees," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 3, pp. 501–510, 2014.

[5] S. Amsuess, P. Goebel, B. Graimann, and D. Farina, "A multi-class proportional myocontrol algorithm for upper limb prosthesis control: Validation in real-life scenarios on amputees," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 23, no. 5, pp. 827–836, 2015.

[6] S. Amsuess *et al.*, "Context-dependent upper limb prosthesis control for natural and robust use," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 7, pp. 744–753, 2016.

[7] D. Farina and O. Aszmann, "Bionic limbs: clinical reality and academic promises.," *Sci. Transl. Med.*, vol. 6, no. 257, p. 257ps12, Oct. 2014.

[8] T. A. Kuiken, L. A. Miller, K. Turner, and L. J.

Hargrove, "A Comparison of Pattern Recognition Control and Direct Control of a Multiple Degree-of-Freedom Transradial Prosthesis," *IEEE J. Transl. Eng. Heal. Med.*, vol. 4, no. c, 2016.

[9]   L. J. Hargrove, L. A. Miller, K. Turner, and T. A. Kuiken, "Myoelectric Pattern Recognition Outperforms Direct Control for Transhumeral Amputees with Targeted Muscle Reinnervation: A Randomized Clinical Trial," *Sci. Rep.*, vol. 7, no. 1, pp. 1–9, 2017.

[10]   L. Resnik, H. Huang, A. Winslow, D. L. Crouch, F. Zhang, and N. Wolk, "Evaluation of EMG pattern recognition for upper limb prosthesis control: a case study in comparison with direct myoelectric control," *J. Neuroeng. Rehabil.*, vol. 15, no. 1, p. 23, 2018.

[11]   F. Cordella *et al.*, "Literature review on needs of upper limb prosthesis users," *Front. Neurosci.*, vol. 10, no. MAY, pp. 1–14, 2016.

[12]   N. E. Bunderson and T. A. Kuiken, "Quantification of feature space changes with experience during electromyogram pattern recognition control," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 20, no. 3, pp. 239–246, 2012.

[13]   M. A. Powell and N. V Thakor, "A Training Strategy for Learning Pattern Recognition Control for Myoelectric Prostheses.," *J. Prosthet. Orthot.*, vol. 25, no. 1, pp. 30–41, 2013.

[14]   M. A. Powell, R. R. Kaliki, and N. V. Thakor, "User training for pattern recognition-based myoelectric prostheses: Improving phantom limb movement consistency and distinguishability," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 22, no. 3, pp. 522–532, 2014.

[15]   N. Nilsson and M. Ortiz-Catalan, "Estimates of Classification Complexity for Myoelectric Pattern Recognition," *Proc. - Int. Conf. Pattern Recognit.*, pp. 2682–2687, 2017.

[16]   J. He, D. Zhang, N. Jiang, X. Sheng, D. Farina, and X. Zhu, "User adaptation in long-term, open-loop myoelectric training: implications for EMG pattern recognition in prosthesis control," *J. Neural Eng.*, vol. 12, no. 4, p. 046005, 2015.

[17]   M. Ortiz-Catalan, F. Rouhani, R. Branemark, and B. Hakansson, "Offline accuracy: A potentially misleading metric in myoelectric pattern recognition for prosthetic control," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, vol. 2015-Novem, pp. 1140–1143, 2015.

[18]   M. B. Kristoffersen, A. W. Franzke, C. K. van der Sluis, A. Murgia, and R. M. Bongers, "The Effect of Feedback During Training Sessions on Learning Pattern-Recognition-Based Prosthesis Control.," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 10, pp. 2087–2096, Oct. 2019.

[19]   M. Cohen, "Brain Mapping, Handedness Questionnaire," 2008. [Online]. Available: http://www.brainmapping.org/shared/Edinburgh.php. [Accessed: 20-Jul-2005].

[20]   R. C. Oldfield, "The assessment and analysis of handedness: the Edinburgh inventory.,"

Neuropsychologia, vol. 9, no. 1, pp. 97–113, Mar. 1971.

[21]   B. Hudgins, P. Parker, and R. N. Scott, "A New Strategy for Multifunction Myoelectric Control," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 1, pp. 82–94, 1993.

[22]   T. R. Farrell and R. F. Weir, "The optimal controller delay for myoelectric prostheses," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 1, pp. 111–118, 2007.

[23]   L. H. Smith, L. J. Hargrove, B. A. Lock, and T. A. 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. Rehabil. Eng.*, vol. 19, no. 2, pp. 186–192, 2011.

[24]   M. B. Kristoffersen, A. W. Franzke, C. K. Van Der Sluis, R. M. Bongers, and A. Murgia, "Should Hands Be Restricted When Measuring Able-Bodied Participants To Evaluate Machine Learning Controlled Prosthetic Hands?," *IEEE Trans. Neural Syst. Rehabil. Eng.*, 2020.

[25]   T. A. Kuiken *et al.*, "Targeted muscle reinnervation for real-time myoelectric control of multifunction artificial arms.," *JAMA*, vol. 301, no. 6, pp. 619–628, Feb. 2009.

[26]   M. Ortiz-Catalan, R. Brånemark, and B. Håkansson, "BioPatRec: A modular research platform for the control of artificial limbs based on pattern recognition algorithms.," *Source Code Biol. Med.*, vol. 8, no. 1, p. 11, 2013.

[27]   R. Bakeman, "Recommended Effect Size Statistic," *Behav. Res. Methods*, vol. 37, no. 3, pp. 379–384, 2005.

[28]   A. J. Young, L. J. Hargrove, and T. A. Kuiken, "The effects of electrode size and orientation on the sensitivity of myoelectric pattern recognition systems to electrode shift," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 9, pp. 2537–2544, 2011.

[29]   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. Rehabil. Res. Dev.*, vol. 48, no. 6, pp. 619–628, 2011.

[30]   B. Lv, X. Sheng, D. Hao, and X. Zhu, "Relationship between Offline and Online Metrics in Myoelectric Pattern Recognition Control Based on Target Achievement Control Test," *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, pp. 6595–6598, 2019.

[31]   I. Vujaklija *et al.*, "Translating research on myoelectric control into clinics-are the performance assessment methods adequate?," *Front. Neurorobot.*, vol. 11, no. FEB, pp. 1–7, 2017.

[32]   E. Campbell, A. Phinyomark, A. H. Al-Timemy, R. N. Khushaba, G. Petri, and E. Scheme, "Differences in EMG Feature Space between Able-Bodied and Amputee Subjects for Myoelectric Control," *Int. IEEE/EMBS Conf. Neural Eng. NER*, vol. 2019-March, pp. 33–36, 2019.