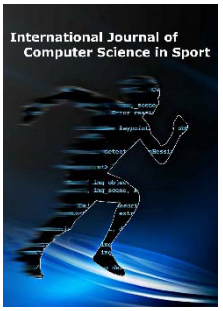
International Journal of Computer Science in Sport

Volume 16, Issue 3, 2017

Journal homepage: http://iacss.org/index.php?id=30

DOI: 10.1515/ijcss-2017-0012



Automated Feedback Selection for Robot-Assisted Training

Gerig, N.[1,2], Wolf, P.[1,2], Sigrist, R.[1,2], Riener, R.[1,2], Rauter, G.[1,2,3]

[1]Sensory-Motor Systems Lab, Swiss Federal Institute of Technology

[2]Spinal Cord Injury Center, University Hospital Balgrist, University of Zurich

[3]Department of Biomedical Engineering, University of Basel

Abstract

Robot-assisted training can be enhanced by using augmented feedback to support trainees during learning. Efficacy of augmented feedback is assumed to be dependent on the trainee's skill level and task characteristics. Thus, selecting the most efficient augmented feedback for individual subjects over the course of training is challenging.

We present a general concept to automate feedback selection based on predicted performance improvement. As proof of concept, we applied our concept to trunk-arm rowing. Using existing data, the assumption that improvement is skill level dependent was verified and a predictive linear mixed model was obtained. We used this model to automatically select feedback for new trainees. The observed improvements were used to adapt the prediction model to the individual subject. The prediction model did not over-fit and generalized to new subjects with this adaptation.

Mainly, feedback was selected that showed the highest baseline to retention learning in previous studies. By this replication of our former best results we demonstrate that a simple decision rule based on improvement prediction has the potential to reasonably select feedback, or to provide a comprehensible suggestion to a human supervisor. To our knowledge, this is the first time an automated feedback selection has been realized in motor learning.

KEYWORDS: VIRTUAL TRAINER, ROWING SIMULATOR, MOTOR LEARNING, LINEAR MIXED MODELS, STATISTICAL LEARNING

Introduction

Training robots complemented by virtual reality offer numerous possibilities to configure robot-assisted training, characterized by various exercise tasks, related task renderings, and by different types of augmented feedback. Feedback that is naturally present to a trainee performing a task physically must be recreated artificially when simulating the same task in virtual reality. This artificially recreated feedback is described as task renderings. In a rowing simulation for example, task renderings consist of a visually displayed landscape, oar-water interaction forces, and oar-water interaction sounds. Augmented feedback provides extrinsic information added along with the task renderings by providing additional cues, support or challenges to the trainee in order to enhance training.

Concurrent visual, auditory, or haptic feedback strategies have shown great potential in facilitating motor learning (Sigrist, Rauter, Riener, & Wolf, 2013). However, augmented feedback strategies should be designed with regards to the Guidance Hypothesis (Sigrist et al., 2013) and to avoid problems with Slacking (Reinkensmeyer, Akoner, Ferris, & Gordon, 2009). Their efficacy is dependent on the skill level of the subject and task characteristics (Marchal-Crespo et al., 2015; Sigrist et al., 2013). According to the Challenge Point Theory (Guadagnoli & Lee, 2004), optimal conditions for improvement are expected when the subject is challenged during training according to their skill level. Therefore, selecting augmented feedback to provide optimal conditions for improvement is challenging, especially if the skill level of the subject changes over the course of training.

Along with task characteristics, an individual subject's needs should be considered when selecting augmented feedback designs. Humans vary in their physical and cognitive abilities, preferences, and giftedness. For example, augmented visual feedback with little robotic support may be appropriate for a healthy subject, but performing the same task, an impaired or elderly subject with physical or cognitive deficits may require fewer visual details and more robotic support. Detailed visual feedback may be suitable for a subject that is used to virtual environments or computer games, but less detailed visual feedback and movement sonification might be beneficial for subjects used to playing musical instruments.

To overcome the challenges of selecting between different augmented feedback designs, the idea of automated feedback selection arose. Training or therapy robots may record large amounts of kinetic or kinematic data, which may serve as a basis for reasonably selecting feedback strategies. Instead of relying on the intuition and experience of a human operator, an automated feedback selection could rely purely on previous observations in the form of quantitative data. Such an automated selection process eventually increases training efficacy, reproducibility of successful training paradigms, and knowledge of the applied augmented feedback.

In terms of machine learning, selecting the most appropriate augmented feedback is a classical decision problem: Based on given observations, such as performance or participation features measured by the robot or entered by the operating trainer, one of a finite set of available augmented feedback designs is selected. Such decision making problems are commonly solved with classification approaches from supervised machine learning (Bishop, 2006), e.g. support vector machines or adaptive boosting. However, for the presented work we wanted to investigate another concept, which predicts the expected change in the subject's performance for each augmented feedback.

Classification approaches from supervised machine learning do often result in quantifications that are not understandable for typical users or training supervisors. In contrast, when selecting types of feedback based on predictions of expected improvement, these predicted

improvements for the different feedback options are a quantitative reasoning that can be interpreted. For many training scenarios the measurable states might not be sufficient to allow reasonable feedback selection and, therefore, a human supervisor may still be required. A human supervisor who understands the meaning of the performance metrics being used will also understand the meaning of the predicted improvements. Such a supervisor could weigh the predicted improvements against unmeasured relevant contextual information, e.g. motivational aspects, abnormal joint constraints, or impairments. Therefore, we considered a prediction-based concept to be more generally applicable and therefore more likely to result in greater user acceptance.

To select the most appropriate augmented feedback at a given point of time, there is no widely accepted ground truth and therefore no available labeled data for supervised classification. Labeling data with the help of expert human trainers is time consuming and suffers intra- and inter-rater variability. For a predictive approach, labeled data can be collected trivially whereby the labels are simply the observed changes in the performance metrics, dependent on the conducted feedback training. Also, based on their prediction accuracy, the resulting prediction models can be evaluated in the absence of a known ground truth for an optimal selection of feedback designs.

Modeling a training goal or selection rules for an optimal feedback selection is challenging. Intermediate training goals might change over the course of training, or different training goals might be applicable for different subjects. Supervised classifiers would have to be completely retrained based on a differently labeled data set. Predictive models of subjects' improvements do allow more flexibility. Training goals or selection rules must be defined based on the predicted performance metrics, but the applied selection rule can be changed whenever desirable. Using the same predictive models, it is also possible to compare different selection rules to each other. For example, one could test strategies with different complexities ranging from greedy strategies maximizing short term benefit to modeling the entire training plan by using informed search techniques (Russell, Norvig, & Davis, 2010) or dynamic programming (Bertsekas, 2012).

Predictive performance metrics are of great interest in motor learning. Findings of prediction models that generalize to new subjects could help to improve feedback designs. Researchers often conclude in their data analyses that certain measures can be used to predict the degree of motor learning (Joiner & Smith, 2008; Wu, Miyamoto, Castro, Ölveczky, & Smith, 2014). However, we have not found any literature that creates predictive models of such a finding, and quantifies generalization to new subjects in new experiments.

Thus, we aimed to show, as proof of concept, that an automated feedback selection using improvement predictions could be realized. We wanted to formulate the concept generally and derive the predictive models in a structured way, such that this process could be applied similarly to different fields of robot-assisted training. Feasibility of our predictive concept was tested with healthy subjects for a rowing task in a custom made large tendon-based rowing simulator (Rauter, Zitzewitz, Duschau-Wicke, Vallery, & Riener, 2010). The goal for our proof of concept was to realize an automated feedback selection that performs meaningfully by using a suitable predictive model, which is observation-based, and generalizes to new subjects.

Methods

Apparatus

For this work, our custom-made rowing simulator (see Figure 1) was used as a robotic training device (Rauter et al., 2011). This simulator was based on a real but trimmed rowing boat,

which was placed in the middle of a Cave Automated Virtual Environment (CAVE). The CAVE consisted of three 4.4 m x 3.3 m screens, which were placed in front and to the sides of the boat. Three projectors (Projection Design F3+, Norway) were used to display a visual ocean scenario and augmented visual feedback on the screens, the feedback and scenario were developed in Unity (Unity Technologies, CA, USA) and controlled for a minimum rate of 30 frames per second. Oar-water interaction sounds and auditory augmented feedback were implemented using C++ with an update rate of approx. 30 Hz and were delivered through over-ear stereo headphones (with a standard frequency range: 14 Hz to 26 kHz). Water resistance and haptic augmented feedback were realized with a tendon-based parallel robot (Rauter et al., 2010), which was controlled by a Matlab/Simulink® (r2013b, MathWorks, MA, USA) model running on an xPC target at a fixed update rate of 1 kHz. For this study, the rowing simulator was set up for sweep rowing on portside, meaning the subject manipulated a single oar with both hands on the left side in direction of travel. For water rendering, a virtual rowing model was implemented (Rauter et al., 2010), which allowed the subject to accelerate and decelerate the boat in the virtual environment when interacting with the virtual water. Water rendering and simulator realism were confirmed by a study showing that training in the simulator fostered skill gains in real rowing on water to a comparable extent as real rowing training on water (Rauter et al., 2013).

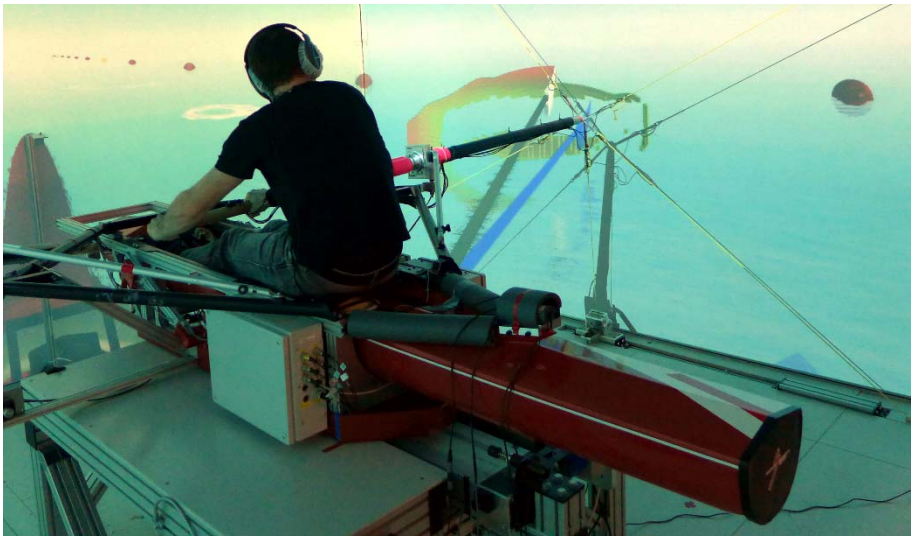


Figure 1. Apparatus with rowing scenario and visual augmented feedback.

Task

Subjects were asked to learn trunk-arm rowing, which is a common warm-up or training exercise in rowing, used to synchronize the rowers' movements with each other. A trunk-arm rowing stroke is a complex oar movement requiring coordination of trunk and arm movements to achieve a desired spatial and velocity profile. The trunk-arm rowing stroke consists of a drive and a recovery phase with the transitions between the two, i.e. the release from drive to recovery and the catch from recovery to drive. During drive, the completely immersed oar blade should be pulled through the water with a high velocity, in order to introduce the necessary energy to propel the boat. During recovery, the oar is moved back above the water with a slower, smooth and horizontal motion in order to prepare for the next catch. Reproducing a desired spatial and velocity profile is especially challenging due to oar-water interaction forces: the required forces vary depending on the relative velocity between the boat, oar, and water.

To train trunk-arm rowing, subjects were instructed to learn and repeat a given reference stroke

as accurately as possible with respect to both the spatial and velocity profile of this reference stroke. The reference stroke was prerecorded by a rowing expert at a typical training stroke rate of 24 strokes per minute. The recorded reference was post-processed, to give a smooth, cyclic trajectory with C^2-continuity and an exact duration of 2.5 s reflecting an exact stroke rate of 24 strokes/min. It was further rescaled to a movement range at the oar handle center of 0.67 m (44º) horizontally and 0.19 m (12.5º) vertically to ensure a reference movement that could be performed by subjects taller than the minimally requested 1.65 m (inclusion criteria). Oar blade rotation and leg movements were not included in the task, the oar blade was instructed to be kept in vertical orientation only and the rowing seat was fixed at a position, where the subject's legs were extended. This same recorded reference stroke (see Figure 2) was used in previous studies by the same authors (Rauter, Sigrist, Riener, & Wolf, 2015; Sigrist, Rauter, Marchal-Crespo, Riener, & Wolf, 2014).

Kinematic Evaluation

Kinematic data, i.e., the vertical and horizontal oar angles estimated from the measured rope lengths, were recorded at 100 Hz and analyzed in custom-written programs in Matlab® (MathWorks, MA, USA). Only test conditions were analyzed, errors during training were not evaluated. The measured data for these tests were segmented into rowing strokes at the point in time that corresponds to the minimal horizontal angle of the subject's movement. The first five strokes and the last stroke of each test condition were excluded from further analysis to avoid transition effects (subjects needed time to accelerate the boat to a steady state). Additionally, rowing strokes with a rate below 22 and above 26 strokes/min were excluded from analysis. The remaining rowing strokes were resampled to 250 data points and compared to the reference trajectory, which was also resampled to 250 data points. Rowing strokes were rated with two different dissimilarity metrics: spatial error and velocity magnitude error, i.e. speed error. We chose to evaluate these two metrics independently, since we did not know on how spatial accuracy should be weighed against temporal accuracy for this task.

Spatial error was defined as the overall spatial error in one rowing stroke obtained by using dynamic time warping (Giese & Poggio, 2000) with a zero weighting of the temporal shifts. Dynamic time warping assigns each sample of the subject's stroke to one sample of the reference by minimizing spatial and weighted temporal differences, under the constraint of time continuity (causal temporal order of samples). Spatial error was then calculated as the mean spatial difference between these assigned samples. The benefit of this metric compared to using a fixed temporal assignment (e.g. by the same index) is that small temporal shifts that have a large impact on spatial errors are not overrated (Vlachos, Hadjieleftheriou, Gunopulos, & Keogh, 2003). In other words, a reduction of a spatial error based on fixed temporal assignments could be the result of improvements in temporal accuracy only. In contrast, a reduction of a spatial error based on dynamic time warping may not only be the result of improvements in temporal accuracy.

The velocity magnitude error was assumed to be a sufficient metric for temporal accuracy, since for repetitive trunk-arm rowing, absolute temporal constraints are considered to be of low relevance. The velocity magnitude error was defined analogously to the spatial error with dynamic time warping, aligning the absolute value of the velocity between reference and user trajectories. Illustrative examples on how these alignments work for one good and one bad rowing strokes are displayed in Figure 2.

The subject's performance in a single test condition was evaluated with the average spatial error (ϵ_s) and the average velocity magnitude error (ϵ_v) of all valid strokes ($22 \leq stroke\ rate \leq 26$) in this test condition.

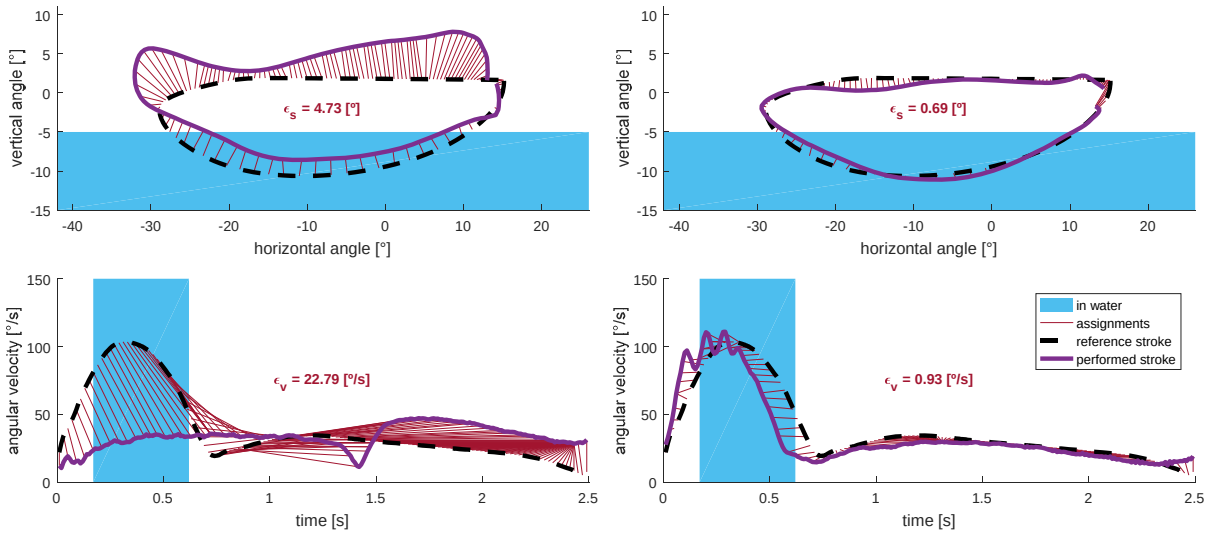


Figure 2. Illustrative rowing stroke examples of the applied kinematic evaluation with resulting single stroke errors. Bad examples of both spatial error (top) and velocity magnitude error (bottom) are shown on the left hand side, and good examples on the right side. For this illustration, only every second sample assignment was plotted.

Augmented Feedback Training

Subjects were provided with different augmented feedback designs during training. Augmented feedback was displayed in addition to the rowing simulation, i.e. the display of the virtual lake, oar immersion sounds and, where applicable, haptic water rendering. Seven different augmented feedback designs were used within this work. More technical details on these feedback designs are provided in our previous studies (Rauter et al., 2015; Sigrist et al., 2014).

The first augmented feedback, (i) *Visual*, consisted of a visual concurrent feedback (visible in Figure 1), where a virtual blue oar displayed the reference movement in addition to the subject's own oar. The blue reference oar was rendered with increasing transparency the closer the subject's oar was to the reference, fading out completely if the angular distance between subject's oar and the reference was less than 4º. Additionally, the subject's oar drew traces when the spatial deviation from the reference was greater than 3.6º vertically or 1.9º horizontally. These traces were drawn in green for low deviations and their color became increasingly reddish the more the subject deviated spatially from the reference path. The traces were faded out after 8 s to prevent the subject relying too strongly on these traces. Thus, all components of this visual feedback were designed such that the feedback faded out if the subject's performance was near the reference to avoid the subject relying on the feedback and becoming dependent on it.

The second augmented feedback, (ii) *AudioVisual*, consisted of feedback (i) plus sonification of the oar movement. The horizontal oar angle was mapped to pitch (from 54.5 Hz to 91.58 Hz) if the oar was outside the water. Normal purling sounds from the rowing simulation were played when the oar blade was immersed. The left headphone speaker was used to sonify the reference motion and the right headphone speaker was used to sonify the subject’s own oar blade. That way, subjects could synchronize their movement to the reference by minimizing sound differences between left and right ear.

The third augmented feedback, (iii) *PositionController*, consisted of a haptic guidance, where the robot fully guided the oar movement along the reference with a PD-Controller. The subject could just passively follow the movement by the robot, which is sometimes referred as full

haptic guidance. The position controller proportional gain ($p_{pos} = 6000$ Nm/rad) and the derivative gain ($d_{pos} = 170$ Nms/rad) ensured that the robot followed the reference independent of the subjects applied forces. *PositionController* was the only condition whereby water interaction forces were not haptically rendered.

The fourth augmented feedback, (iv) *PathController,* consisted of a different form of haptic guidance to that used in (iii), where the subject is only constrained spatially to stay inside a perceived tunnel around the reference path, but without any imposed velocity constraints. From a subject's perspective, the path controller created elastic forces pulling the oar back to the reference path, once the deviation was greater than a certain threshold, but if staying close to the reference only the rendered water interaction forces were perceived. Technically, the path controller forces were realized by the negative gradient of a conservative potential field, which is a passive control strategy that is safe for both user interaction and superposition with the haptic water rendering. In this augmented feedback, the conservative potential field, and therefore the perceived tunnel, was constant in all training sessions.

The fifth augmented feedback, (v) *AdaptivePathController,* consisted of a haptic feedback following the same principle as (iv), but here an additional assist-as-needed concept was implemented to decrease the haptic guidance in regions where the subject performed well, and increased the haptic guidance where the subject performed bad. Technically this was realized by scaling the potential field locally based on previous rowing strokes, i.e. increasing it in areas where the subject deviated a lot and decreasing the potential field where the subject was close to the reference path.

The sixth augmented feedback, (vi) *ReactivePathController,* consisted of another haptic feedback concept, which was also based on the *PathController* principle. If the subject deviated too much from the reference path but was approaching the reference path, they would receive the same supportive force that the *PathController* would provide. However, if the subject deviated too much from the reference path and was still moving further away, a reactive force would be applied instead. This reactive force was directed and scaled based on the current user velocity, e.g. it interrupted and stopped the user movement abruptly.

The seventh augmented feedback, (vii) *VisuoHaptic,* consisted of both the visual augmented feedback from *Visual* (i) and the reactive path controller from *ReactivePathController* (vi).

Experimental Protocol

The experimental protocol (Figure 3) was kept identical to our previous studies (Rauter et al., 2015; Sigrist et al., 2014) to facilitate comparisons between previous and new results.

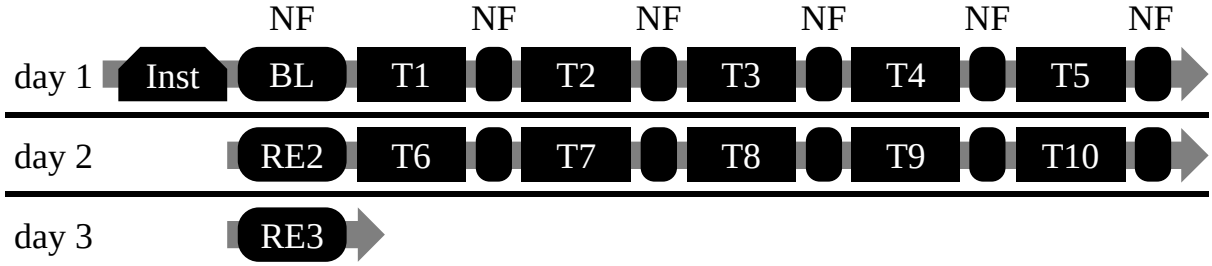


Figure 3. Experimental protocol, where NF denotes the no-feedback test conditions, Inst denotes the instruction phase, BL denotes the baseline test, T denotes the training conditions with augmented feedback, and RE2 and RE3 the retention tests on day 2 and 3 respectively.

The subjects were recorded on three consecutive days. On Day 1, they received general instructions and one investigator explained the handling and safety features of the simulator.

This general instruction was followed by a demonstration, where the investigator sat in the rowing simulator and demonstrated the movement to the subject. Thereby, the robot guided the movement with the *PositionController* for ~30 s. This primary instruction was performed to prepare subjects for interaction with the haptic guidance and to demonstrate the movement range and velocity the oar will be moved by the robot. Every subject was given a 180 s (≙ 72 reference strokes) practice session with the *PositionController* to obtain a basic idea of the reference movement before the baseline test. Subjects were asked to memorize spatial and temporal characteristics of the reference movement, since they would be requested to reproduce the reference movement in a subsequent baseline test. Subjects were not given any advice on how to profit from the robot-guided reference movement.

After instruction, a 180 s baseline test (BL) followed, where the subjects rowed without robotic guidance. Thus, the subjects had to perform the movement on their own, based on their memory of the instruction. In baseline and all subsequent no-feedback test conditions, the visual scenario, the purling sound of their oar in the right headphone, and the water interaction forces were rendered. The baseline test was followed by five training sessions which each consisted of a 180 s (≙ 72 reference strokes) training session with augmented feedback and a 60 s (≙ 24 reference strokes) no-feedback test condition. Which of the augmented feedback designs was provided during the trainings is dependent on the group and will be explained in detail later (see section *Feasibility Study*). Whenever a subject received a certain augmented feedback for the first time, the augmented feedback was described to the subject and a short familiarization period of maximum 60 s was additionally provided. A break of ~25 s was included between training with augmented feedback and no-feedback test conditions. On Day 2, a retention (RE2) test of 180 s was performed, followed by five training sessions according to the same procedure of Day 1. On Day 3, another retention test (RE3) of 180 s was performed.

During all no-feedback test conditions, the subjects were verbally instructed to increase or decrease the stroke rate if they left the range of 22 to 26 strokes/min in order to avoid effects on performance caused by a speed-accuracy trade-off. There were no verbal instructions on the stroke rate during the training conditions.

Automated Feedback Selection

We wanted to realize an automated selection of the available augmented feedback designs to train towards a desired trunk-arm movement. This automated selection should be based on improvement predictions to enable comprehensible reasoned selections, data labeling arising from quantitative analysis, and flexibility of applied selection rules. Additionally, we wanted to incorporate three general assumptions into our predictive models:

- The challenge point theory states that training exercises need to be matched to a subject's skill level (Guadagnoli & Lee, 2004) such that the subject can improve optimally. Therefore, we assume that when predicting improvement for a given augmented feedback, the current skill level of the subject plays an important role.
- The different augmented feedback designs cause different improvements. Selecting feedback based on predictions is only meaningful if there are different improvements observable for the available feedback designs.
- Individualization of the improvement prediction is important since there is a large inter-subject variability in humans when learning a motor task.

Skill level is an abstract term, and measuring the current skill level is challenging. In our trunk-arm rowing task, we interpreted skill as the grade to reproduce the reference rowing stroke by

oneself, in absence of any augmented feedback. Therefore, the best estimate of the current skill level was measured by our kinematic performance metrics average spatial error (ϵ_s) and average velocity magnitude error (ϵ_v) during no-feedback test conditions. Our experimental protocol (Figure 3) provided an alternating sequence of feedback trainings and no-feedback test conditions. This alternating sequence enabled us to use the performance achieved during last no-feedback test condition as an estimate of the current skill level.

We have come up with a prediction-based concept (illustrated in Figure 4), where the skill level measured during the last no-feedback test condition before the training ($\epsilon_s^{before}, \Delta\epsilon_v^{before}$) is used to predict the subject's improvements ($\Delta\hat{\epsilon}_s, \Delta\hat{\epsilon}_v$). over one 3 min training for each available augmented feedback. The augmented feedback for the next training could then be selected based on the predicted improvements ($\Delta\hat{\epsilon}_s, \Delta\hat{\epsilon}_v$).

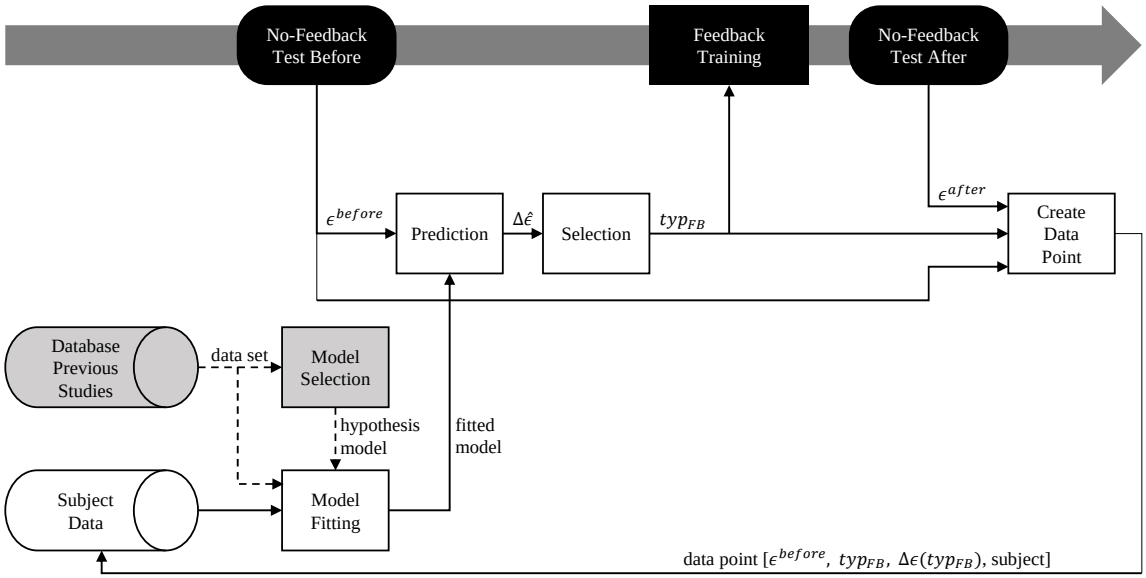


Figure 4. Automated Feedback Selection Concept.

Additionally, the subject's true improvements were measured by the change in our performance metrics from the no-feedback test condition before training to those recorded directly after the feedback training ($\Delta\epsilon = \epsilon^{before} - \epsilon^{after}$). A new data point consisting of the error before (ϵ^{before}), the performed feedback type (typ_{FB}), and the true observed improvements ($\Delta\epsilon_s, \Delta\epsilon_v$) could be added to a subject's database. This new data point could then be used to improve or individualize the prediction models that originated from data from previous studies.

Prediction

To realize the prediction based on the current skill level, we have chosen to use linear mixed models. For linear mixed models both theories and software implementations are available for model fitting, for objective model comparisons (e.g. likelihood ratio tests), and for conditional evaluation of the model predictions (West, Welch, & Galecki, 2006). Additionally, model comparison tests of linear mixed models can be interpreted as an explorative data analysis, which provide insights into the fitted data. Linear mixed models allowed us to include and test preliminary assumptions, i.e. the dependency of improvements on the current error level of the subject, the augmented feedback type, and a random factor for subject-specific individualization. The model finding procedure is described in the subsection *Model Selection*

following the description of our *Database Previous Studies*.

Selection

A greedy selection strategy was used to select between the augmented feedback designs for simplicity. The greedy strategy selected the augmented feedback with the highest sum of normalized, predicted improvements over the next three minute feedback training:

$$typ_{FB}^{*} = \text{argmax}_{typ_{FB}} \left(\frac{\Delta \hat{\epsilon}_s(typ_{FB})}{\tilde{\epsilon}_{sAV}} + \frac{\Delta \hat{\epsilon}_v(typ_{FB})}{\tilde{\epsilon}_{vAV}} \right),$$

where $\tilde{\epsilon}_{sAV} = 2.1$ °, $\tilde{\epsilon}_{vAV} = 3.4$ °/s were used to normalize the improvements by the average group errors from the retention test on day 3 (RE3) of the best group from our previous experiments: the *AudioVisual* group from (Sigrist et al., 2014).

Database Previous Studies

To develop our prediction models, we used data from our previous studies (Rauter et al., 2015; Sigrist et al., 2014). All seven different augmented feedback designs were tested in the previous studies. For each augmented feedback, a group of eight subjects was trained while receiving the same augmented feedback during all ten training conditions. In total, 560 training conditions of 3 minutes in duration were recorded. For each training condition one data point was obtained for both improvement in spatial error ($\Delta\epsilon_s$) and velocity magnitude error ($\Delta\epsilon_v$). Data points were obtained in the form of the error value ϵ^{before} from the test condition immediately before the training condition and the error reduction $\Delta\epsilon$. For both spatial error and velocity magnitude error 1120 data points were expected. However, due to some missing test condition data, there were 549 data points for the spatial error ϵ_s and 535 data points for the velocity error ϵ_v (1084 data points in total).

Model Selection

We started our linear mixed model selection with a base hypothesis model. We simplified our base hypothesis model until we achieved convergence in fitting our simplified hypothesis model to the data of our previous studies. The resulting hypothesis model was then compared to even more simplified, and to more complex models, by using likelihood ratio tests. With this procedure, we wanted to ensure that we were using an evidence based prediction model, which did not over-fit the data.

Our linear mixed model designs are reported using a form of the Wilkinson-Rogers notation, where the dependent variable is on the left side of the ~-symbol, the fixed and random factors on the right side of the ~-symbol, where random factors are denoted by being inside brackets (West et al., 2006). This notation does not only compactly represent the models, but can also directly be used for model specification in statistical software such as Matlab® (MathWorks, MA, USA) and R (R Development Core Team, 2008), which were both used for this work. For model selection, the R packages *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *RLRsim* (Scheipl, Greven, & Kuechenhoff, 2008) were used for model fitting and comparisons with likelihood ratio tests (R commands *lmer*, *anova*, *exactRLRT*). Matlab® was used, since it provides a linear mixed model implementation with conditional prediction capabilities (Matlab® commands *fitlme*, *predict*). Additionally, model comparisons with likelihood ratio tests were performed in Matlab® to confirm results from R (Matlab® command *compare*). Significance results from R and Matlab® were consistent, but the exact p-values differed slightly. Reported p-values are from the results in R, since there was no equivalent for the *RLRsim* package in Matlab®.

We used chi-square based likelihood ratio tests to compare the hypothesis model with the

simplified or extended models. These likelihood ratio tests correct for the different statistical degrees of freedom and provide the probability p_R, which indicates the probability that the increased likelihood of the more complex model is by chance. If p_R is larger than our chosen significance limit $\alpha = 0.05$, then the more complex model is not significantly better than the simplified model and should be rejected. For the special case, where a linear mixed model is compared to a model without random effect, *RLRsim* (Scheipl et al., 2008) was used instead of a chi-square based likelihood ratio test (with default settings using 10000 simulated values). The assumption of normally distributed residuals of the resulting hypothesis model was checked visually using a normal Q-Q plot (R commands *qqnorm*, *resid*).

The base hypothesis model $\Delta\epsilon \sim typ_\epsilon * typ_{FB} * \epsilon + (typ_\epsilon * typ_{FB} - 1|subject)$ represented the assumptions that for each error metric typ_ϵ and augmented feedback typ_{FB}, a different linear function based on the error value ($\epsilon = \epsilon^{before}$) before the training explains the error improvement $\Delta\epsilon$ over the 3 minute training. A positive $\Delta\epsilon$ was defined to be an improvement, i.e. a reduction of the error value. Additionally, an individualization term was modeled as a random factor for a given subject: the complete interaction between each error metric and augmented feedback ($typ_\epsilon * typ_{FB} - 1|subject$). The -1 denotes removal of a global intercept of this random factor. This individualization term can be interpreted as the subject's individual giftedness or non-giftedness to reduce a certain error with a certain augmented feedback.

For our prediction models, we restricted ourselves to training related information in this work. Therefore, we decided against increasing our model complexity by using subject demographics, e.g. age or gender. With this approach, we expected the resulting hypothesis model not to over-fit and to be reasonably specific regarding to our assumptions of skill, feedback, and individual subject dependences.

Creation of New Data Points and Model Fit Update

Each subject started with the same prediction model, i.e. the same model fit based on the same preliminary data. After recording the test condition that followed the training with the selected augmented feedback, the true error improvements ($\Delta\epsilon_s$ (typ^*_{FB}), $\Delta\epsilon_v$ (typ^*_{FB})) were calculated and used to update the model fit for this subject. To ensure that each subject had the same amount of data and comparable circumstances independent of the order of testing, the model fit was only updated with the subject's own data points. All data points were equally weighted for the fit. The maximum of 10 data points for each subject had relatively low weight against the minimum 535 preliminary data points and were assumed to have little impact on the nominal models. The subject's own data points were assumed to have most impact on the individualization random effect during conditional prediction. Therefore, individualization was assumed to arise from the subject random effect and not from a shift in the nominal models.

After completion of a no-feedback test, the next training configuration should be selected with minimum delay, however, there is a considerable amount of calculation necessary. First, the no-feedback test had to be evaluated, including rowing stroke segmentation and evaluation of the single strokes. Then, the model fit had to be updated and the conditional prediction had to be calculated for each available training configuration to select the next training type. To minimize the delay between the no-feedback test and the next training, the evaluation of rowing strokes in the no-feedback tests was implemented in real-time. For this real-time evaluation, the individual rowing strokes and reference were down-sampled to 101 data points instead of 250. Down-sampling to 101 instead of previously 250 data points did not critically affect the performance metrics.

Feasibility Study

To check if our concept fulfilled our goal, we performed a feasibility study with new subjects. The feasibility study should show that our predictive models generalize to new subjects and that feedback selection was reasonable.

Generalization of our predictive models was investigated by comparing the magnitude of prediction residuals for new independent data points to those fitted by the model. Additionally, to investigate if online updating of the prediction models is necessary, the magnitude of prediction residuals was compared between the used updated models and the constant models.

To ensure that our automated feedback selection performed in a reasonable fashion, we checked which feedback designs were selected, and compared the learning results from a group of subjects trained with automated feedback selection against our previously best preforming group, and a no-feedback control group.

Groups

The feasibility study was held in a parallel design with 8 healthy volunteer subjects per group according to the same experimental protocol as in our previous studies (Rauter et al., 2015; Sigrist et al., 2014). The best performing group in our previous experiments was the *AudioVisual* group (Sigrist et al., 2014), which received 10 trainings with the *AudioVisual* training configuration. This *AudioVisual* group will be reported for comparison.

The test group in this study was called *Predicted* group. The subjects of the *Predicted* group received augmented feedback according to our automated feedback selection.

In addition, to the new *Predicted* group, a control *NoFeedback* group was recorded. The *NoFeedback* group only received information on the desired movement during the instruction phase. Subjects in the *NoFeedback* group performed free trainings instead of trainings with augmented feedback, e.g. 3 min training time with the rowing scenario and water rendering. This group was tested to prove that our desired movement pattern was not simple enough for the subjects to remember completely from the instructions and to simply improve by getting used to the robot and the rendered environment.

Subjects

A total of 16 subjects (6 females, mean age 26.3 years, SD 3.5 years) were recruited, mainly from the university (students). The subjects were healthy, had normal hearing and normal or corrected-to-normal vision. All subjects had no prior experience with the task and confirmed to be non-rowers and perform at least half an hour of sport per week.

The subjects were pair-wise randomly assigned to one of two groups consisting of eight subjects while matching gender. A coin toss assigned the first subject of each pair, and the next subject of the same gender was assigned to the other group. The subjects were either assigned to the *Predicted* group (3 females, 5 males, 24-34 years, mean age 28 years, SD 3.1 years) or to the *NoFeedback* group (3 females, 5 males, 20-30 years, mean age 24.6 years, SD 3.2 years). Written informed consent was obtained from all subjects. The methods were approved by the ethics commission of the Federal Institute of Technology, Zurich, Switzerland (ETH Zurich, Ethics Commission, EK 2014-N-21). The study was carried out in accordance with the approved guidelines.

Statistical Evaluation

Statistical evaluation was performed in Matlab®. Within each test condition the two metrics spatial error and velocity error were averaged over all valid strokes. These mean values (ϵ_s and

ϵ_v) per test condition for each subject were used for the statistical analysis.

Baseline and retention tests were investigated using repeated-measures ANOVA for interaction between group (*AudioVisual, Predicted, NoFeedback*) and test (BL, RE2, RE3) and between group effects. Changes within each group were assessed using follow-up repeated measures ANOVA. Violations of sphericity were tested for with Mauchly's test for sphericity, but were not found to have occurred. Multiple comparisons were corrected using post-hoc Bonferroni tests.

Significant differences for a specific test condition (e.g. baseline test on Day 1) between the groups were assessed with One-way ANOVA. Multiple comparisons were corrected for with a Tukey-Kramer post-hoc test.

One-way ANOVA was also used to test whether the groups differed in the learning rate, i.e. the error reduction from one test condition to the next normalized by the error at the first of those two test conditions. Three different learning rates were calculated and tested for group differences: from baseline test on Day 1 to retention test on Day 2 (BL to RE2 normalized with BL), from baseline test on Day 1 to retention test on Day 3 (BL to RE3 normalized with BL), and from retention test on Day 2 to retention test on Day 3 (RE2 to RE3 normalized with RE2).

Results

Model Selection

The model selection results are summarized in Figure 5, starting from middle left with the base hypothesis model. Due to the limitation in our preliminary data that each subject received one type of augmented feedback only, fitting our base hypothesis model was not possible. In R, we did not achieve model convergence, probably due to aliasing between the fixed and the random effect of the typ_{FB} interaction. In Matlab® the model could not be fit because their implementation requires a complete rank dataset. Therefore, we simplified our base hypothesis model to the hypothesis model (Figure 5, middle), which could be supported by our data. This hypothesis model excluded the type of augmented feedback in the random effect. In other words, the modeled subject-specific giftedness only captured how a subject is gifted at decreasing a certain error. But the modeled subject-specific giftedness could not capture how gifted a subject is in using a certain augmented feedback to decrease a certain error.

According to the likelihood ratio tests, the hypothesis model explained the data significantly better than the simplified models (Figure 5, top row). A model without the discrimination of the error metric in the random effect showed a lower likelihood for the same complexity (Figure 5, middle row, right). Finally, models that included the global training number or the training day number were significantly better at explaining the data than the hypothesis model (Figure 5, bottom row). However, we have chosen to disregard these models. The available data was only from a two day protocol and we decided to prefer slightly worse predictions instead of adding a linear dependence based on two time points. Therefore, the hypothesis model (Figure 5, middle) was used as the prediction model for the feasibility study.

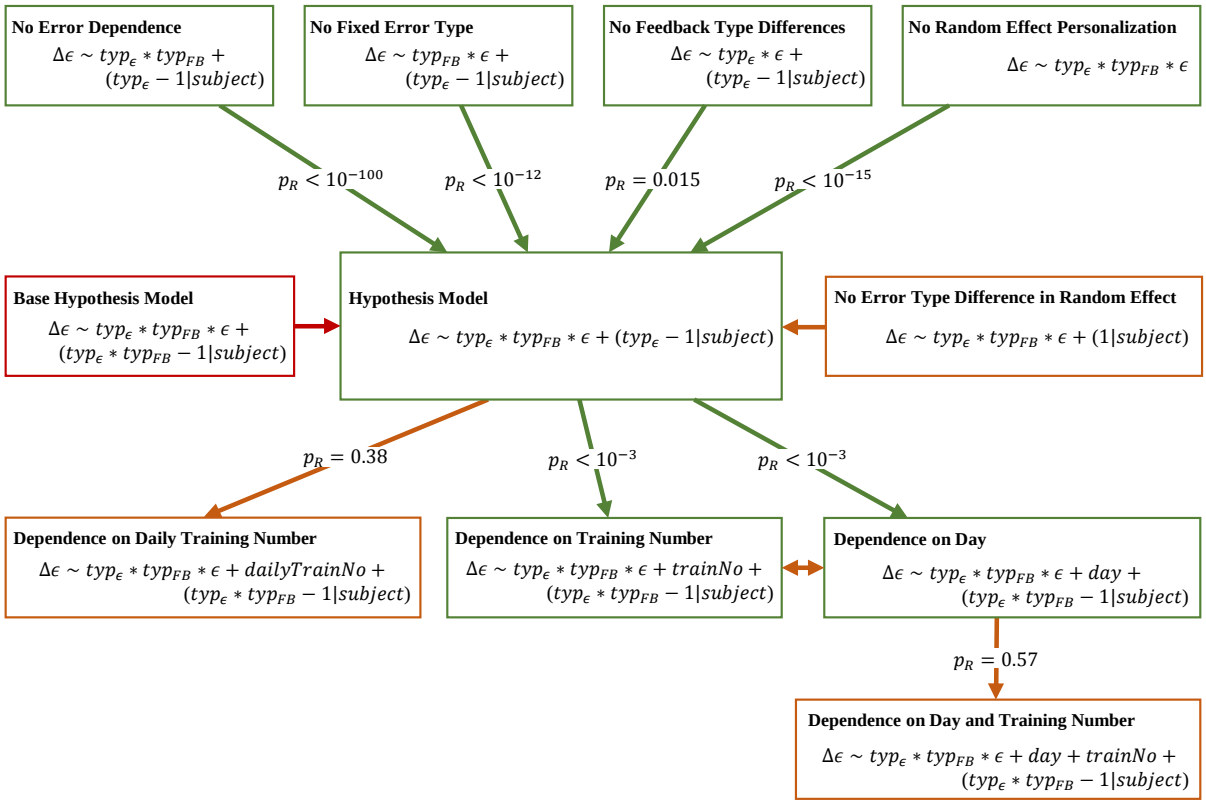


Figure 5. Model Selection: Arrows indicate direction of increasing model complexity, two-headed arrows indicate same amount of statistical degrees of freedom. The imprinted p_R-Value indicates the probability that the more complex model explains the data better than the less complex model only by chance (e.g. from likelihood ratio test). The arrow colors indicate if this probability is significant ($p < 0.05$) in green, non-significant in orange or not available due to missing model convergence in red. A green box indicates that the respective model is explaining the data significantly better than its ancestor. An orange box indicates that it is either not significantly better than its ancestor or worse in explaining the data compared to a model of equal complexity. A red box means that the model fitting failed to converge, either due to aliasing or limitations in the available data.

Feasibility Study

Selected Augmented Feedback Designs

In total, the *Predicted* group received *AudioVisual* feedback in 65 out of 80 trainings and *Visual* feedback for the 15 remaining trainings. Other possible augmented feedback designs were not selected.

Baseline to the Retention Tests

For spatial error, repeated-measures ANOVA tests revealed the significant main effect of tests, ($F_{(2,42)} = 3.67$, $p = .0340$) and the interaction between test and group ($F_{(4,42)} = 4.68$, $p = .0033$). Also for velocity error, main effect of tests ($F_{(4,42)} = 38.92$, $p = 2.7 \cdot 10^{-10}$) and interaction between group and test ($F_{(4,42)} = 7.59$, $p = .0001$) were significant.

The within group follow-up repeated-measures ANOVA (Table 1) showed that only the *AudioVisual* comparison group differed significantly from baseline to retention test on day 2 and from baseline to retention test on day 3 for the spatial error (Figure 6, top). For the velocity error, the *AudioVisual* group differed again significantly from baseline to retention test on day 2 and from baseline to retention test on day 3. But additionally, the *Predicted* group showed a trend in difference from their baseline to the retention test on day 2 and a significant difference

from their baseline to the retention test on day 3 (Figure 6, bottom).

Table 1. Results from baseline to retention, within group follow-up differences from the Bonferroni post-hoc that are significant or close to significance.

Spatial Error	**Test1**	**Test2**	**Difference**	**StdErr**	**p-Value**
AudioVisual	BL	RE2	1.32	0.22	.0016
AudioVisual	BL	RE3	1.62	0.31	.0038
Velocity Error	**Test1**	**Test2**	**Difference**	**StdErr**	**p-Value**
AudioVisual	BL	RE2	5.50	0.72	$3.7 \cdot 10^{-4}$
AudioVisual	BL	RE3	5.80	0.78	$4.3 \cdot 10^{-4}$
Predicted	BL	RE2	2.97	1.04	.0721
Predicted	BL	RE3	3.64	0.70	.0036

One-way ANOVA between the groups at fixed tests did not show any significant differences at baseline (Table 2). For the spatial error (Figure 6, top), the only differences were found in the retention test on day 3, the NoFeedback group was significantly different to both the AudioVisual group and the NoFeedback group. For the velocity error (Figure 6, bottom), the AudioVisual group differed significantly from the Predicted and the NoFeedback group at the retention tests on both day 2 and day 3.

Table 2. Between group differences at fixed tests from one-way ANOVA, corrected for multiple comparisons.

Spatial Error	**Group1**	**Group2**	**Lower**	**Mean**	**Higher**	**p-Value**
RE3	*AudioVisual*	*NoFeedback*	-3.27	-2.00	-0.73	.0019
RE3	*Predicted*	*NoFeedback*	-2.99	-1.72	-0.45	.0069
Velocity Error	**Group1**	**Group2**	**Lower**	**Mean**	**Higher**	**p-Value**
RE2	*AudioVisual*	*Predicted*	-7.70	-4.46	-1.22	.0062
RE2	*AudioVisual*	*NoFeedback*	-7.67	-4.43	-1.19	.0065
RE3	*AudioVisual*	*Predicted*	-7.03	-4.08	-1.14	.0058
RE3	*AudioVisual*	*NoFeedback*	-8.36	-5.41	-2.47	.0004

The between group differences in learning rate (Table 3) from baseline to the retention test on day 3 showed significant differences between the *NoFeedback* group and both the *AudioVisual* group and the *Predicted* group. Additionally, for the velocity error the learning rate from baseline to the retention test on day 3 of the *AudioVisual* group differed significantly to the learning rate of the *Predicted* group.

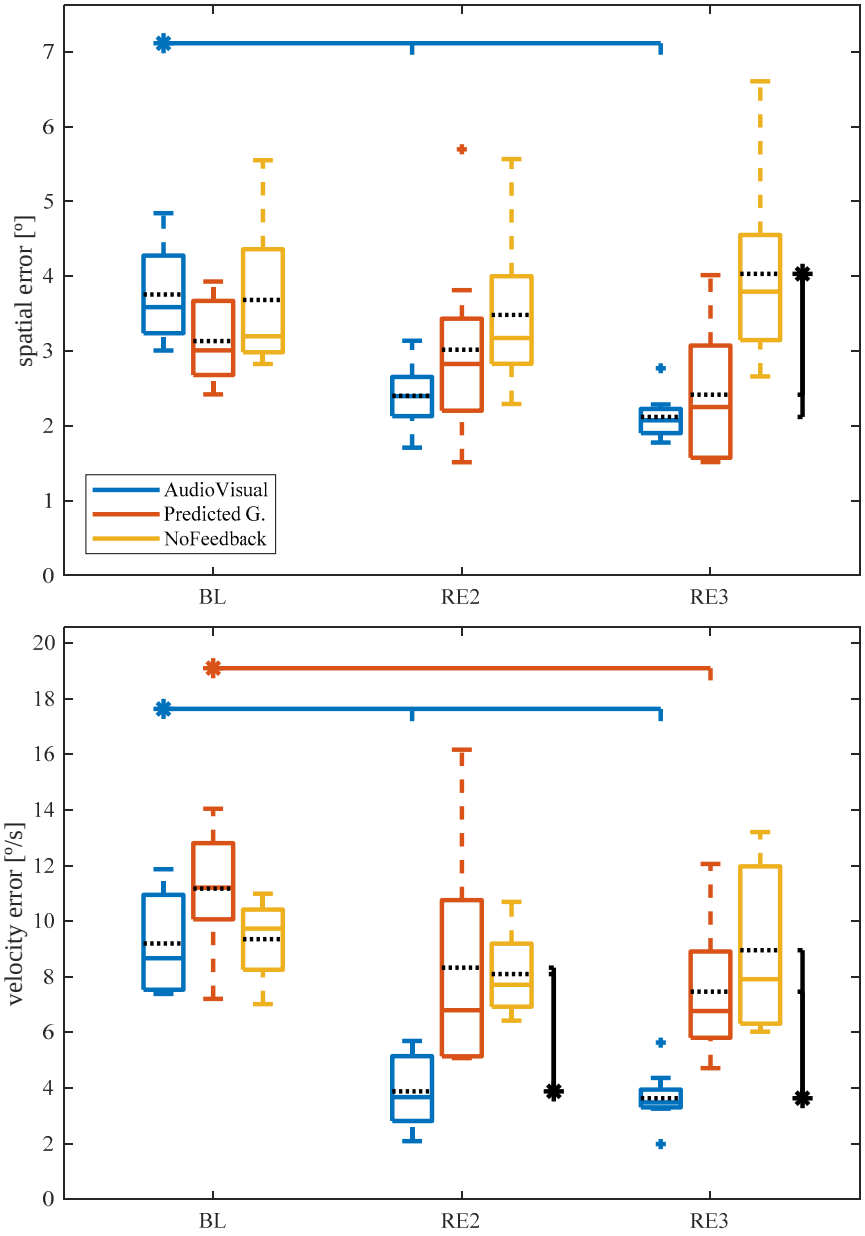


Figure 6. Results of the *AudioVisual* group, the *Predicted* group, and the *NoFeedback* group for the 3-minute test conditions. The group means are additionally plotted as black dashed lines into the box plot. Vertical bars in group colors denote differences between the test conditions in the within groups follow-up repeated measures ANOVA: the condition marked with the group-colored asterisk * is significantly different to the conditions with a small bar. The black vertical bars denote group differences at a fixed test condition from one-way ANOVA. A group mean marked with a black asterisk is significantly different to the group means at the same test condition marked with a small horizontal black bar.

Table 3. Between group differences in learning rate from one-way ANOVA, corrected for multiple comparisons.

Spatial Error Learning Rate	**Group1**	**Group2**	**Lower**	**Mean**	**Higher**	**p-Value**
from BL to RE2	*AudioVisual*	*NoFeedback*	-0.80	-0.38	0.04	.0808
from BL to RE3	*AudioVisual*	*NoFeedback*	-1.02	-0.62	-0.22	.0025
from BL to RE3	*Predicted*	*NoFeedback*	-0.84	-0.44	-0.03	.0332
from RE2 to RE3	*Predicted*	*NoFeedback*	-0.70	-0.33	0.04	.0850
Velocity Error Learning Rate	**Group1**	**Group2**	**Lower**	**Mean**	**Higher**	**p-Value**
from BL to RE2	*AudioVisual*	*Predicted*	-0.54	-0.30	-0.06	.0146
from BL to RE2	*AudioVisual*	*NoFeedback*	-0.71	-0.47	-0.23	.0002
from BL to RE3	*AudioVisual*	*Predicted*	-0.50	-0.28	-0.05	.0133
from BL to RE3	*AudioVisual*	*NoFeedback*	-0.77	-0.54	-0.32	$1.4 \cdot 10^{-5}$
from BL to RE3	*Predicted*	*NoFeedback*	-0.49	-0.26	-0.04	.0193

Prediction Residuals

For the spatial error, the prediction residuals (Figure 7, top) only show minor differences between the updating and the constant model. For the velocity error, the prediction residuals (Figure 7, bottom) of the constant model are larger than the prediction residuals for the updating model, except for the first training. For both metrics, the prediction residuals reach similar levels to the apriori data fit from the 8^{th} training on.

Discussion

In this work, a general concept of an automated feedback selection was elaborated that selects augmented feedback based on improvement predictions. The elaborated concept provides flexibility for a variety of different robot-assisted trainings, training goals, and performance metrics. To our knowledge, this is the first generally applicable virtual trainer concept, that closes the loop between the evaluation of subject performance and reasoned selection of different available augmented feedback designs.

The findings of our model selection will be discussed first in subsection *Model Selection and Implications* to provide insights into the existing data. Then, to elaborate whether the feedback selection was meaningful, we start to discuss the occurrences of the selected feedback designs to facilitate the interpretation of the results in learning from baseline to retention of the feasibility study in the section *Selected Augmented Feedback Designs*. Then, subsection *Baseline to the Retention Tests* will compare the progress of our *Predicted* group to the previously best performing group and to our *NoFeedback* group, to answer the question of whether our feedback selection was meaningful. Finally, the generalization of our prediction models to new subjects will be discussed in subsection *Prediction Residuals*.

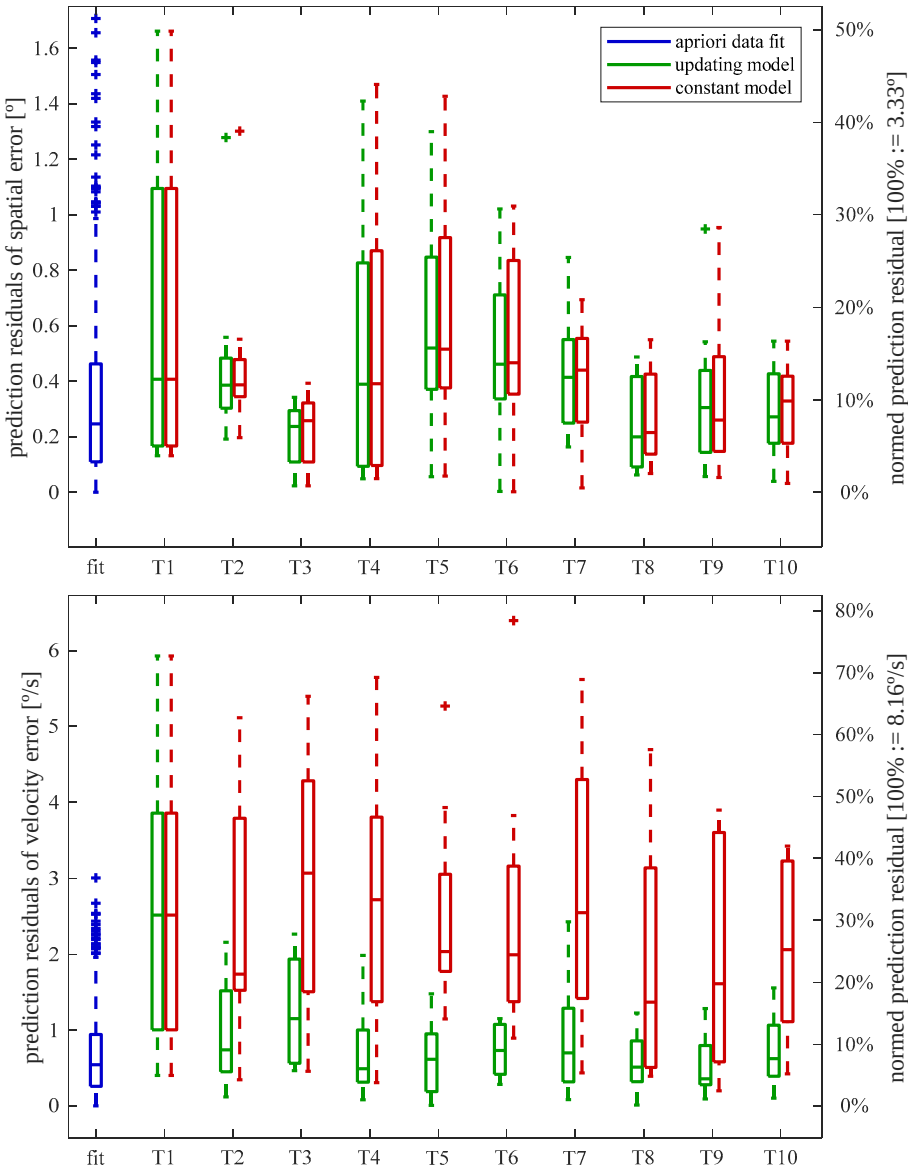


Figure 7. Prediction accuracy comparison between the used updating model and a constant model. The prediction accuracy is illustrated with the development of the prediction residuals over the 10 training conditions for the 8 subjects in the *Predicted* group. For reference, the blue box plot in the column labeled with *fit* shows the residuals from fitting the linear mixed models to the apriori data points, 549 data points for the spatial error and 535 data points for the velocity error were available. The green boxes show the prediction residuals of the used updating model. The red boxes show the prediction residuals from constant models that were only based on the apriori data. The left axis indicates the absolute value of the prediction residual in the original unit. The right axis indicates the residuals as percentage normed by the span between maximum and minimum error improvement observed during the validation study. For the spatial error metric, the maximum and minimum improvement observed were 2.14° and -1.19°. For the velocity error metric, the maximum and minimum improvement observed were 6.13°/s and -2.03°/s.

Model Selection and Implications

Using linear mixed models as a base for our prediction models comes with advantages but also limitations. Advantages compared to nonlinear modeling approaches are the robust and efficiently fitting algorithms, comprehensible form and the possibility to check if model assumptions are supported by preliminary data. However, the linearity of the models may be problematic: Firstly, when put to practical use, the model will extrapolate to areas where no

previous data has been collected, which can lead to a very poor performance. Secondly, the linear models will cross the zero line at some point and eventually predict negative improvement. While negative improvements or deterioration of performance with further training might occur, predicted deterioration should be interpreted very carefully. If training efficacy is flooring, predicted deteriorations might just be artifacts of the linear form. However, for our application, the predictions were used to select between augmented feedback designs. If the predicted improvements are negative, the predicted value is inconsequential since this feedback mode should not be chosen. If predicted improvements are negative for all available feedback designs, the exact values would also be of little importance, since a selection between ineffective results would have no relevance. Therefore, we were confident that the benefits of linear mixed models would outweigh the drawbacks.

Another limitation of our modeling and fitting process is that we did not include known boundaries or limitations in our noise modeling, e.g. our errors can only be positive and improvement is bound by the current error magnitude. That means assuming normality of residuals on the improvement will result in tails that violate those limitations. However, using another noise model would result in a less efficient model fitting and additional complexity to our method. To check if this simplification would be critical, the assumption of normally distributed residuals was checked by visual inspection. Single data points at the end of the tails deviated from the normal line, but were not considered critical. Those data points were not further investigated, because a violation of normality is assumed to be unimportant for the parameters of multilevel regression models (Gelman & Hill, 2007). Additionally, our sample size was sufficiently large (1084 data points) that we could assume the central value theorem would ensure that our fitting residuals approximate normality in the areas of interest.

Our model selection procedure using linear mixed models could also be understood through an explorative data analysis. However, we limited our discussion of the test results between the hypothesis model and the simplified or extended models (Figure 5) to the implications for the model predictions.

No Error Dependence: The test between the hypothesis model and the model that is independent of the error value showed that the expected improvement is dependent on the current skill level. This result was not surprising since we considered an error metric and error improvement: the error is bounded to be larger than zero, the improvement is linearly bounded as well, i.e. the improvement cannot be higher than the current error value. Therefore, it is expected to find a correlation between error improvement and error value before the training for each non-detrimental augmented feedback. However, this result highlights the importance of taking the current skill level into account when predicting improvements.

Our simplification of using errors as a direct measure for skill level could limit the predictions of our model. Errors reflect an observable performance rather than an objective skill level. The relation between observable performance and related skill levels have been modeled in a more sophisticated manner in the field of cognitive skill learning, e.g. for a programming problem (Huang, Guerra-Hollstein, & Brusilovsky, 2016). However, a transfer from their concept to motor learning faces additional challenges: We have no binary performance outcomes and for the problem of learning a movement and related skills in motor learning provide a less clear structure than those of solving a programming problem. If a similar model was developed for motor learning, improvement modeling for individual feedback trainings based on a more sophisticated skill level model could provide better predictions.

No Fixed Error Type: This test checked if our two error metrics, spatial error ϵ_s and velocity magnitude error ϵ_v were different enough, such that it mattered if we predicted spatial improvement or velocity improvement. In our case, including knowledge of the error type

explained the data significantly better than the model without error type. In other words, the spatial error ϵ_s and the velocity magnitude error ϵ_v were only weakly correlated in our data. If such a test would fail, the informativeness of the chosen error metrics to measure different aspects of movement performance should be questioned.

No Feedback Type Difference: This test checked if the different augmented feedback designs caused significantly different improvements. In our data, the observed improvements differed significantly based on the different augmented feedback designs. If no significant differences were observed dependent on which augmented feedback had been used, the different augmented feedback designs might be too similar in their effect on the chosen error metrics. In such a case, a selection of individual augmented feedback designs could not be justified based on these error metrics.

No Random Effect: This test basically showed that we had intra-subject correlation, meaning that subject-specific information was observable in the data. Therefore, a part in the model that adapts to the individual subject was expected to improve the predictive capabilities of the model. We would assume that having a term that adapts its prediction to each specific subject would be of great importance. Many different factors of the human subject and behavior are unknown and eventually not measurable by a robot. An individualization of the prediction models to observations from this subject seems to be the only solution to compensate at least partially for such effects. However, the limitation in our previous data that each subject was only trained with one type of augmented feedback, resulted in a model with a random effect that explained only a global giftedness or non-giftedness to reduce either spatial or velocity magnitude error. We expected that not considering subject giftedness for each available augmented feedback would be a critical limitation for studies where longer protocols or less strict inclusion criteria were applied. E.g. our hypothesis model would fail to learn that a blind subject cannot learn with visual augmented feedback, but would assume the subject to be generally too ungifted to improve.

Dependence on Daily Training Number: The effect of including the increasing training numbers on one day was not significant. A small negative effect was observed, which could be interpreted as increasing subject exhaustion or loss of focus. However, since this effect was not significant, we assume that our healthy subjects were not critically exhausted or had lost focus within 5 consecutive trainings in one day.

Dependence on Training Number: The observed significant dependence on the training number showed that there was small positive effect of increasing training numbers. One explanation could be that the subjects could profit more from later trainings, because they were more experienced with the augmented feedback designs. If such an experience with a specific augmented feedback would be crucial for its efficiency, our concept of automatically switching between different feedback might be detrimental. However, this effect was not visible when looking at the training number within one day, therefore we raised the question if this observation was just based on an effect of the day.

Dependence on Day: The observed significant dependence on the training day showed that subjects improved more on the second day of training – once the improvement was corrected for the other dependencies. We have decided to ignore this effect and use the hypothesis model without dependence on day for our feasibility study, mainly because we were not confident regarding how this effect would generalize to a longer training protocol. The other model dependencies on current error value, augmented feedback type, and the individual subject seem intuitively more general and independent of the training protocol used to obtain the preliminary data. Therefore, we decided to test the hypothesis model without this effect. We preferred showing a working principle based on slightly worse performing prediction models,

than one that might be too specific by relying on our two day protocol. Nevertheless, we realized that the model including the dependence on the day reached a very similar likelihood to the model with the dependence on the training number. We assumed that the previously observed dependence on training number was therefore based on this dependence on day only. To make sure that no valuable significant information was hidden in the training number except for the day effect, we tested with a model including day and training number.

Dependence on Day and Training Number: The model including both day and training number did not explain the data significantly better than the model only including the day. If there was a significant positive effect of training number left, that could have been a hint that subjects did get increased benefit when receiving the same augmented feedback multiple times, e.g. due to increased understanding of the feedback. However, this was not observed in our data. Therefore, we were confident that with the given short instructions our subjects could profit from the augmented feedback from the first time they experienced it. Thus, we did not expect severe negative consequences if an automated selection often switches between feedback designs.

All tested ancestor models provided significantly less model evidence, probabilities were below our chosen significance level of $p < 0.05$. Therefore, we considered our model complexity justified (not expected to over-fit). The model of the same statistical complexity without error type discrimination in the random effect was fit and discarded because it had a slightly lower likelihood. In summary, the results of our model selection procedure confirmed our three assumptions: dependency on current skill level, dependency on the feedback type, and the necessity of individualization due to the intra-subject correlation.

Selected Augmented Feedback Designs

Only the *AudioVisual* and the *Visual* feedback, which share the same visual augmented feedback, were selected. When trained exclusively with either *AudioVisual* or *Visual* augmented feedback no group differences in learning from baseline to retention were observed in a prior study (Sigrist et al., 2014). Therefore, these two augmented feedback designs were very similar and differences in their selection may be based purely on chance. One of our basic assumptions was that different feedback designs were better suited at different error levels, measured by spatial and velocity magnitude error. However, this basic assumption was not fulfilled between those two feedback designs. With the strong similarity of the two chosen feedback designs in mind, the occurrences seem to indicate that in our two day protocol the same augmented feedback may have been optimal with respect to our selection strategy.

A possible explanation for these two augmented feedback designs to be selected only is that our greedy selection strategy on the error weighting based on the previous best group performances. This greedy selection did not enable planning or evaluation of different reasonable training strategies, e.g. selecting an augmented feedback that specifically helps to reduce spatial error (e.g. *PathController*) and focusing on reducing velocity error in later trainings could only happen if a subject had a very high spatial error compared to the velocity error. However, our subjects tended to have higher velocity errors, but more comparable spatial errors to the previous *AudioVisual* group (Figure 6, bottom). Therefore, feedback designs that did not provide a velocity reference (*PathController*, *AdaptivePathController*, *ReactivePathController*) were unlikely to be selected through our selection strategy.

Planning over all remaining trainings instead of a greedy strategy selecting the next training could reduce this limitation of not switching between feedback designs that focus on spatial or velocity error. Such planning can be realized using informed search techniques, e.g. Breadth-First Search (Russell et al., 2010). However, planning over multiple trainings would also cause

prediction errors to add up. Therefore, if such a strategy would still be beneficial needs to be investigated in a future iteration of this study.

Another possible explanation as to why only these two augmented feedback designs have been selected is that they were the only two that did not change the task dynamics. All the other augmented feedback designs contained a haptic concurrent feedback, which altered the forces that the subject had to exert in order to reproduce the reference stroke. This may be an indication that for learning to reproduce our rowing movement, an augmented feedback that does not alter task dynamics may be superior. A major challenge in learning to reproduce our rowing movement is the slightly changing task dynamics, which are caused by slight variations in the relative velocity between boat and water. Haptic concurrent feedback might interfere with learning of how to compensate for these slight variations in required task dynamics.

In summary, the selected augmented feedback designs indicated that with our task protocol and feedback designs we observed the trivial case, where all subjects progressed in a way that one feedback training was superior to the others for their observed error levels. For different error levels, different feedbacks could have been selected. From our previous data we were not expecting the subjects to progress in this manner. Despite this unexpected behavior, our general concept resulted in a meaningful selection: Selecting the feedback based on predictions for single 3 minute trainings with a greedy strategy resulted in the same gold standard that showed the highest baseline to retention learning in previous studies. We do not claim that this finding generalizes to situations with other augmented feedback designs or other trained tasks. However, we demonstrated that a very simple decision rule may be sufficient for reasonable feedback selection, if the selection rule is based on improvement predictions.

Baseline to the Retention Tests

Since the *Predicted* group only received the *AudioVisual* and *Visual* feedback, a similar learning to the formerly tested *AudioVisual* group (Sigrist et al., 2014) could have been expected. However, some differences in behavior were found, which will be discussed in detail for the two different errors.

Spatial error was not significantly decreased over test conditions in the *Predicted* group (Figure 6, top). In contrast, spatial error was significantly decreased over test conditions in the *AudioVisual* group. However, no group differences between *Predicted* and *AudioVisual* were found either for the error values at any test condition or in learning rates. The missing significant decrease over test conditions in the *Predicted* group may be based on the lower baseline mean error or the higher observed variance in our *Predicted* group. Therefore, the observed tendency of a decrease in spatial error was in line with our expectations.

Velocity error was significantly decreased from BL to RE3 in the *Predicted* group. However, the absolute errors of the *Predicted* group were significantly higher than those of the *AudioVisual* group and the learning rates were found to be significantly lower. This basically means that both groups did improve, but the new *Predicted* group improved significantly less than the formerly measured *AudioVisual* group. The only difference in the group protocols was that the *Predicted* group received some *Visual* feedback trainings without sonification. Since the formerly measured *AudioVisual* and *Visual* groups (Sigrist et al., 2014) did not show such differences, we assume that the differences are caused by a systematic change between the two different studies, in which these groups were recorded. Even if the same protocol, the same task and the same prototype robot were used, small systematic changes can never be prevented completely. In our case there were minor updates in the robot hardware, namely exchanging the oarlocks to allow for a higher pretension in robot control and required sensor recalibration. These slight changes in our robot seem to have affected the task characteristics enough to be

measurable in the outcome, i.e. we assume that reproducing the velocity profile of the reference movement accurately had become more difficult.

The *NoFeedback* control group matched our expectation of not showing any learning from baseline to the retention tests.

Prediction Residuals

Spatial Error Reduction

The prediction residuals of spatial error improvements reached similar levels to the model fitting residuals (Figure 7, top). Therefore, for the spatial error improvements, the model generalized well and did not overfit our a priori data.

The spatial predictions are similar to the a priori fit residuals for most of the trainings. The median prediction residual for the 8th and subsequent training conditions remained lower than 10% and the upper bound of the 95% confidence interval remained below 17% (Figure 7, top). However, only small differences in prediction accuracy were observed between the updating and the constant model and therefore an updating model may be an unnecessary effort for predicting the spatial error improvements. In other words, the knowledge from having observed the *AudioVisual* group in the previous study was reliable enough to predict the improvements in spatial error of the *Predicted* group.

The low prediction errors might have resulted from a very low spatial error decrease in general. The *Predicted* group did not significantly decrease their spatial error from baseline to retention. On average, the *Predicted* group decreased their spatial error by 0.68º, the informativeness of a median prediction residual of around 0.33º is therefore questionable.

The prediction residuals seem noticeably higher for the training conditions 1, 4, 5, or 6 than for the other training conditions (Figure 7, top). Intuitively it seems clear that in the beginning, there should be higher prediction residuals, since the models could not individualize the predictions to the subject. However, the missing individualization in the beginning does not explain the observed rise in residuals in training 4, 5, or 6. Training 4 and 5 correspond to the last two trainings on the first day. We can only speculate the reasons for the worse predictions at these trainings. Some subjects might lose attention after receiving the almost same augmented feedback training for the fourth time in a row. Another possible reason could be that some subjects were less motivated or focused during their 4th and 5th test condition. The 6th training corresponds to the first training on the second day. A possible reason for the increase in prediction residuals may be that subjects came from a variety of different previous occupations and showed differences in readapting to the task and the simulator.

Velocity Error Reduction

For the velocity error improvements (Figure 7, bottom), the prediction residuals of the updating model reached similar levels to the model fitting residuals as well. The prediction residual median and percentiles of the a priori data fit seem smaller than those of the updating model. This difference between priori data fit and updating model seem larger for the prediction residuals of velocity error than for those of spatial error. However, only 8 data points were used to compute medians and percentiles of the updating model per training, whereas for the a priori data fit 535 data points were used. The 95% confidence of the updating model residuals was observed inside the span of the a priori residuals. Therefore, also for the velocity error improvements we conclude that our model generalized well and did not over fit our apriori data.

For the updating model, the median prediction residuals for the 8th and subsequent trainings

remained lower than 8% and the upper bound of the 95% confidence interval stayed below 13% (Figure 7, bottom). But, the constant model performed worse than the updating model for the velocity error. One reason that the constant model performed worse could be that the subject-specific differences in giftedness or learning ability had a large impact on how subjects learned to reproduce the velocity profile. The a priori fit residuals could reach comparably low levels, since the data for the fit included 10 data points per subject. The 10 data points per subject were directly used to fit the individualization random effect. However, for new subjects this optimal offset was unknown, and the average of the existing subjects was taken as a best guess. If the subject-specific differences were large, then taking the average could be a bad guess for a new subject and would result in poor performance. In contrast to the constant model, the updating model could improve the estimate of this individualization random effect with each new subject data point.

The probably dominant reason for the larger prediction residuals of the constant model is that subjects in the *Predicted* group did improve significantly less than the *AudioVisual* group from our previous studies (see discussion in section *Baseline to the Retention Tests*). However, this *AudioVisual* group data was the only basis within the constant model for predicting improvements with the *AudioVisual* training configuration.

On average, the *Predicted* group decreased their velocity error for 3.64º/s from baseline to retention on day 3, the prediction residual of around 0.65º/s at the last three training conditions seems reasonable. Therefore, the prediction of velocity error improvement seems reasonable while significant learning effect between baseline and retention on day 3 (Figure 7, bottom) are present. The model predictions seem comparable for all but the first training.

Our feasibility study is an example that even when using the same protocol, the same training configurations, and robot, even minor changes lead to observable effects in the outcome metrics. This seems to be a strong indication that an updating model, i.e. statistical learning, is necessary for predicting outcomes or selecting of robotic assisted trainings. The updating model used could correct this offset beginning at the 2nd training with only one data point from the subject. This very fast correction arises from the individualization random effect, i.e. the model just takes the new subject for being less gifted in reducing its velocity error. With a longer protocol, where more data would be added to the model than the 10 data points from the individual subject, the observed offset in velocity error improvements would more and more be corrected by the fixed effect of the velocity error. Simplified, the model would adapt to new observations in this context similar to a multi-rate model (e.g. the one used by (Joiner & Smith, 2008) to model human learning) with a fast learning part, i.e. individualization random effect, and a slow learning part, i.e. error type effect.

Conclusion

We introduced a general concept of an automated selection of augmented feedback based on predictions of the human subjects' performance improvements. The introduced concept is very general in the sense that it is does not require application specific knowledge. The concept is independent from knowledge specific to the task, the performance metrics required to quantify improvement, the available augmented feedback, or the robotic hardware. The automated feedback selection concept only requires data in the form of previous observations and can learn from new observations to cope with changing conditions or to individualize to new subjects.

We successfully implemented the automated feedback selection on a robot-assisted trunk-arm rowing training study. The improvement of kinematic performance metrics in robot-assisted

trunk-arm rowing was dependent on the individual subject and current level of performance. Linear mixed models that were updated with new observations generalized well, reaching prediction errors similar to the model fit residuals. Model fit updates were necessary to correct small changes in the robotic rowing simulator that affected subjects' improvements.

To our knowledge, this is the first time an automated feedback selection has been realized in motor learning. With our feasibility study, we could successfully demonstrate that a reasonable feedback selection can be realized based only on evaluated kinematic data. Furthermore, to simplify transfer to other applications, our prediction-based concept can be applied as a comprehensible feedback suggestion for human supervisors instead of only automated selection.

Acknowledgments

We want to thank Sarah Grimm and the team of the ETH statistical consulting for their valuable input and discussions. Special thanks go to Michael Herold-Nadig for his support on technical and safety issues. We want to thank Linda Seward for her support in revising and proofreading the manuscript. We also want to thank Marco Bader for his technical contributions to the simulator. Further, we want to thank Mark van Raai for the realization of the visual rowing scenario and visual feedback. We want to thank Samantha Fox for her help in conducting the measurements, and all the volunteering subjects for participating.

This work was supported by the SNF-Grant "Acceleration of complex motor learning by skill level-dependent feedback design and automatic selection", CR23I2_152817., and CRRP "Neuro-Rehab" of the University of Zurich.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. Retrieved from http://CRAN.R-project.org/package=lme4

Bertsekas, D. P. (2012). Dynamic Programming and Optimal Control. Athena Scientific optimization and computation series. Athena Scientific. Retrieved from https://books.google.ie/books?id=H-PSMwEACAAJ

Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Gelman, A., & Hill, J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research. Cambridge University Press. Retrieved from https://books.google.ch/books?id=lV3DIdV0F9AC

Giese, M. A., & Poggio, T. (2000). Morphable models for the analysis and synthesis of complex motion patterns. *International Journal of Computer Vision*, *38*(1), 59–73.

Guadagnoli, M. A., & Lee, T. D. (2004). Challenge point: a framework for conceptualizing the effects of various practice conditions in motor learning. *Journal of Motor Behavior*, *36*(2), 212–224. doi:10.3200/JMBR.36.2.212-224

Huang, Y., Guerra-Hollstein, J. D., & Brusilovsky, P. (2016). Modeling Skill Combination Patterns for Deeper Knowledge Tracing. UMAP (Extended Proceedings).

Joiner, W. M., & Smith, M. A. (2008). Long-term retention explained by a model of short-term learning in the adaptive control of reaching. *Journal of neurophysiology*, *100*(5), 2948–2955.

Marchal-Crespo, L., Wolf, P., Gerig, N., Rauter, G., Jaeger, L., Vallery, H., & Riener, R. (2015). The role of skill level and motor task characteristics on the effectiveness of

robotic training: first results. Rehabilitation Robotics (ICORR), 2015 IEEE International Conference on (pp. 151–156). IEEE.

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Rauter, G., Sigrist, R., Koch, C., Crivelli, F., Raai, M. van, Riener, R., & Wolf, P. (2013). Transfer of Complex Skill Learning from Virtual to Real Rowing. *PLoS ONE*, *8*(12), 1–18. doi:10.1371/journal.pone.0082145

Rauter, G., Sigrist, R., Marchal-Crespo, L., Vallery, H., Riener, R., & Wolf, P. (2011). Assistance or challenge? Filling a gap in user-cooperative control. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3068–3073). San Francisco, California. doi:10.1109/IROS.2011.6094832

Rauter, G., Sigrist, R., Riener, R., & Wolf, P. (2015). Learning of temporal and spatial movement aspects: A comparison of four types of haptic control and concurrent visual feedback. *IEEE Transactions on Haptics*, *8*(4), 421-433.

Rauter, G., Zitzewitz, J. von, Duschau-Wicke, A., Vallery, H., & Riener, R. (2010). A tendon based parallel robot applied to motor learning in sports. 3rd IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob), 2010 (pp. 82–87). Tokyo, Japan. doi:10.1109/BIOROB.2010.5627788

Reinkensmeyer, D. J., Akoner, O., Ferris, D. P., & Gordon, K. E. (2009). Slacking by the human motor system: computational models and implications for robotic orthoses. Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE (pp. 2129–2132). IEEE.

Russell, S., Norvig, P., & Davis, E. (2010). *Artificial intelligence: a modern approach.* Upper Saddle River, NJ: Prentice Hall.

Scheipl, F., Greven, S., & Kuechenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis*, *52*(7), 3283–3299.

Sigrist, R., Rauter, G., Marchal-Crespo, L., Riener, R., & Wolf, P. (2015). Sonification and haptic feedback in addition to visual feedback enhances complex motor task learning. *Experimental brain research*, *233*, 909–925.

Sigrist, R., Rauter, G., Riener, R., & Wolf, P. (2013). Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review. *Psychonomic Bulletin & Review, 20*(1), 21–53. doi:http://dx.doi.org/10.3758/s13423-012-0333-8

Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., & Keogh, E. (2003). Indexing multi-dimensional time-series with support for multiple distance measures. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '03 (pp. 216–225). Washington, D.C.: ACM. doi:10.1145/956750.956777

West, B. T., Welch, K. B., & Galecki, A. T. (2006). Linear Mixed Models: A Practical Guide Using Statistical Software. CRC Press. Retrieved from http://books.google.com.au/books?id=LSJ__7lDSdgC

Wu, H. G., Miyamoto, Y. R., Castro, L. N. G., Ölveczky, B. P., & Smith, M. A. (2014). Temporal structure of motor variability is dynamically regulated and predicts motor learning ability. *Nature neuroscience, 17*(2), 312-321.