STRUCTURE OF FORCE AS A PREDICTOR OF ORAL MOTOR LEARNING IN HEALTHY YOUNGER AND OLDER ADULTS

BY

CHRISTINA R. BRONSON-LOWE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Speech and Hearing Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Associate Professor Julie Hengst, Chair
Associate Professor Jacob Sosnoff, Director of Research
Associate Professor Raksha A. Mudar
Assistant Professor Georgia Malandraki, Purdue University

Abstract

This study examined the relationship of lingual and labial force structure to learning of oral motor continuous fine force tasks using a pursuit tracking paradigm. It investigated how error and the temporal and frequency structure of force during baseline performance predicted oral motor learning in healthy younger and older adults, drawing on dynamical systems (Bernstein, 1967, as cited in Newell et al., 2003), bidirectional complexity change (Vaillancourt & Newell, 2002) and optimal variability theories (Stergiou, Harbourne, & Cavanaugh, 2006) to explain interacting effects of age and task demand.

Right-handed younger (18-28 years of age, *N* = 20) and older (71-79 years of age, *N* = 21) adults participated in 2 days' practice matching constant, 0.75-Hz sinusoidal, and complex periodic (hereafter "multicosine") visual targets by pursing the lips or elevating the tongue to exert submaximal force whose magnitude controlled the height of a visual trace. Targets were centered at 15% of maximal voluntary force (MVF) determined individually per participant and effector. Over the two days, participants practiced matching each target a total of 35 times with each effector. On the third day, learning was assessed in retention trials (unmodified tasks) and transfer trials (multiple task characteristics individually, systematically modified; only transfer to 10% and 20% MVF target force levels is reported here).

Measures of force structure (approximate entropy, ApEn; fuzzy measure entropy, FuzzyMEn; proportion of power, PoP, in 0-1 Hz, 1-2 Hz and 2-3 Hz bands) and error (normalized root mean square error, NRMSE) at baseline (day 1, first trial of each effector x task condition) were related to measurements of reduction in error vs. baseline in retention and transfer trials on the third day ($\Delta_{final}NRMSE_{ret}$, $\Delta_{final}NRMSE_{trn}$), using primarily linear mixed effects modeling. Results are presented organized by hypotheses within specific aims. Because each hypothesis was assessed using multiple measures, only a selection of the results is covered here for brevity.

Specific aim 1. Assess applicability of previous findings on effects of age and task to oral effectors.

Hypothesis 1a. Older adults' force structure will differ task-dependently from younger adults' (lower entropy and a greater proportion of low-frequency power when the task demands high entropy and reduced low-frequency power, and vice versa).

At baseline, task and age group interacted (ApEn: $F(2, 205) = 9.555$; FuzzyMEn: $F(2, 205) = 9.515$; both $p < 0.0005$). Follow-up analysis showed that only younger adults altered entropy across task, (ApEn: $F(2, 100) = 17.173$; FuzzyMEn, $F(2, 100) = 20.492$, both $p < 0.0005$). Younger adults produced

higher-entropy force than older adults only on the constant task (ApEn: $F(1, 41) = 9.407$, $p = 0.004$; FuzzyMEn, $F(1, 41) = 10.297$, $p = 0.003$).

In retention trials, younger adults' entropy was higher than older adults' in the *lip x constant force* condition (ApEn: $t(39) = -4.339$, $p = 0.002$; FuzzyMEn: $t(39) = -4.295$; $p = 0.001$) and lower in the *tongue x sine* condition (ApEn: $t(39) = 4.741$; FuzzyMEn: $t(39) = 4.191$; both $p = 0.001$). These effects suggest younger adults adapted structure of output to task demand more closely than the older adults.

Hypothesis 1b. Adaptability (immediate): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing trial 2 to trial 1 on day 1 within each effector x task combination.

There was no significant effect of age or age-task interaction with this minimal practice.

Hypothesis 1c. Adaptability (after practice): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing day 1 trial 1 to day 3 retention trial 1 within each effector x task combination.

Entropy change with practice depended upon an interaction of age group and task (ApEn: $F(2, 205) = 5.890$, $p = 0.003$; FuzzyMEn: $F(2, 205) = 4.950$, $p = 0.008$). For the multicosine task, younger adults did not change entropy with practice, while older adults increased it (ApEn: $t(67.503) = 2.675$, $p = 0.014$; FuzzyMEn: $F(1, 41) = 4.559$, $p = 0.039$, NS). For the sine task, younger adults decreased entropy with practice, while older adults increased it (ApEn: $t(75.08) = 4.308$, $p = 0.001$; FuzzyMEn: $F(1, 41) = 9.657$, $p = 0.003$).

Hypothesis 1d. Older adults' reduction in error vs. baseline on retention and transfer trials after two days' practice will be less than younger adults'.

On retention trials, age group interacted significantly with both effector ($F(1, 205) = 5.702$, $p = 0.018$) and task ($F(2, 205) = 3.871$, $p = 0.022$). Younger adults showed greater reduction in NRMSE than older adults only with the tongue ($F(1, 41) = 7.38$, $p = 0.009$) and on the sine task ($F(1, 41) = 11.288$, $p = 0.002$).

Hypothesis 1e. Structure of force will differ by task. The constant task will elicit the highest entropy, lowest proportion of low-frequency power, and greatest proportion of higher-frequency power. The sine task will elicit the lowest entropy, greatest proportion of low-frequency power, and lowest proportion of higher-frequency power. The multicosine task will be intermediate.

At baseline, both younger and older adults responded to the increased high-frequency content of the multicosine target compared to the sine target by decreasing power in the 0-1 Hz band and

increasing it in the 1-2 Hz band (all $p \leq 0.002$; see Table 20). On retention trials, entropy had increased with practice to a greater degree for the constant task than for both variable tasks (ApEn, $F(2, 205) = 34.918$; FuzzyMEn, $F(2, 205) = 32.121$; main effects and pairwise comparisons of constant to sine and multicosine, all $p < 0.0005$).

Specific aim 2. Assess differences in motor variability between oral effectors.

Hypothesis 2a. The tongue will produce less complex force than the lip (lower-entropy, greater dominance of low-frequency power).

Hypothesis 2b. The effects of age group and effector on entropy will interact.

At baseline, effector and age group interacted (ApEn: $F(1, 205) = 10.806$, $p < 0.0005$; FuzzyMEn, $F(1, 205) = 9.769$, $p = 0.002$). For older adults only, entropy was higher for the tongue (ApEn: $F(1, 105) = 23.591$; FuzzyMEn, $F(1, 105) = 20.794$; both $p < 0.0005$). On retention trials, older adults still produced higher-entropy force with the tongue (ApEn: $F(1, 105) = 22.708$; FuzzyMEn, $F(1, 105) = 22.767$; both $p < 0.0005$). Younger adults' force production on retention trials showed higher entropy with the lip for the constant task (which demands high-entropy force; ApEn: $F(1, 20) = 11.950$, $p = 0.002$; FuzzyMEn: $F(1, 20) = 10.768$) and slightly lower entropy with the lip for the more structured variable tasks (significant only for multicosine, ApEn: $F(1, 20) = 10.821$, $p = 0.004$; FuzzyMEn: $F(1, 20) = 11.775$, $p = 0.003$), suggesting that adaptation to task demand with practice may have been better with the lip.

Specific aim 3. Assess utility of baseline performance measures in predicting de novo learning of fine-force pursuit tracking tasks in oral effectors.

Hypothesis 3. (a) Error at baseline ($NRMSE_{initial}$) and a measure of temporal structure, (b) higher maximal force entropy (maxApEn or maxFuzzyMEn) at baseline or (c) greater adaptability of entropy at baseline, will predict reduction in error vs. baseline on retention and transfer trials ($\Delta_{final}NRMSE_{ret}$, $\Delta_{final}NRMSE_{trn}$) in pursuit tracking tasks after controlling for age group, effector and task.

$NRMSE_{initial}$ was a significant predictor of reduced error compared to baseline for both retention and transfer trials in all pairings with the various entropy-based predictors (all $p < 0.0005$). Parameter estimates ranged from -0.89 to -0.91, suggesting that poor initial performance functioned as a marker of greater room for improvement.

Maximum entropy also significantly predicted $\Delta_{final}NRMSE_{ret}$ (maxApEn model: $F(1, 245.948) = 7.005$, $p = 0.009$; maxFuzzyMEn model: $F(1, 245.823) = 5.414$, $p = 0.021$). 1-unit increases in maxApEn and maxFuzzyMEn were estimated to predict changes in $\Delta_{final}NRMSE_{ret}$ of 0.32 and 0.14 respectively (worsening of performance), after controlling for age group, task, effector and $NRMSE_{initial}$.

The predictive value of initial change in entropy varied by task for retention trials ($\Delta_{initial}$ApEn: $F(2, 236.302) = 3.514$, $p = 0.031$; $\Delta_{initial}$FuzzyMEn: $F(2, 236.860) = 5.229$, $p = 0.006$). For the constant task, which demands force output of high entropy, higher entropy on trial 2 than trial 1 on day 1 predicted a decrease in $\Delta_{final}NRMSE_{ret}$ (i.e. a greater reduction in error by day 3). For the sine target, which requires force output of low entropy, higher entropy on trial 2 than trial 1 on day 1 predicted an increase in $\Delta_{final}NRMSE_{ret}$ (i.e. a lesser reduction in error by day 3).

Initial change in entropy significantly predicted transfer of learning to tasks with a higher target force level in both models ($\Delta_{initial}$ ApEn model: $F(1, 232.077) = 11.853$; $\Delta_{initial}$ FuzzyMEn model: $F(1, 234.461) = 10.437$, both $p = 0.001$). 1-unit increases in $\Delta_{initial}$ ApEn and $\Delta_{initial}$ FuzzyMEn were estimated to predict changes in $\Delta_{final}NRMSE_{trn}$ of -0.17 [-0.27, -0.07] and -0.09 [-0.15, -0.04] respectively (improved transfer), after controlling for age group, task, effector and $NRMSE_{initial}$.

These and other results from this study suggest that (Aim 1) task-dependent effects on force structure and the bidirectional complexity hypothesis of healthy aging developed in non-oral systems can be applied to fine-force control in pursuit tracking tasks using the lip and tongue; (Aim 2) oral effectors' structure of force differs, influenced by age, and (Aim 3) baseline behavioral measures can predict learning (measured as reduction in error) after two days' practice.

Initial adaptability of entropy predicted better performance on retention trials if the direction of change was in line with task demand, and worse performance if the direction of change was counter to task demand. This effect comports with the idea of variability in early learning as an exploration of task space (Dhawale, Smith, & Ölveczky, 2017; Stergiou, Harbourne, & Cavanaugh, 2006; Wu, Miyamoto, Gonzalez Castro, Ölveczky, & Smith, 2014) and therefore an active support of learning, rather than a hindrance to be suppressed. Optimal variability in a learning context suggests the ability to shift temporal/frequency structure of force output in the direction demanded by a goal or task. The reduction in adaptability of force structure seen in the older adult participants may play a role in changes in learning with aging.

This prediction can be made from a small enough data set to have potential clinical applicability. Older adults remain robustly able to learn and to adjust complexity of oral force output, though with limitations most consistent with the loss of adaptability hypothesis.

To three of my many role models:

Dr. Audrey Holland, for her curiosity, engagement and compassion

Rachael Faith Brown Gray, MSW, and Jeanette Bronson, for bravery, resilience, humor and grace in the face of uncooperative bodies

Acknowledgements

I thank Dr. Julie Hengst (Chair), Dr. Jake Sosnoff (Director of Research), Dr. Raksha Mudar and Dr. Georgia Malandraki (committee members) for their support and feedback.

I also appreciate the contributions of multiple other people to this work. I thank Dr. Chengyu Liu (Newcastle University, United Kingdom) and Dr. Peng Li (Brigham and Women's Hospital, Harvard University) for discussion of Dr. Liu's fuzzy measure entropy algorithm and for sharing Matlab code for its calculation. The LabVIEW code used in this work was developed from it.

For assistance with the fuzzy measure entropy algorithm I thank Dr. Mark Hasegawa-Johnson (Department of Electrical and Computer Engineering, Beckman Institute, UIUC).

For all their statistical advice I thank Dr. Sa Shen (Research Biostatistician, Center on Health, Aging and Disability, UIUC), Dr. Isabella Ghement (Ghement Statistical Consulting Company Ltd.) and the personnel at Laerd Statistics (Lund Research).

For review of infection control procedures, statistical assistance and manuscript review, unflagging support and encouragement, and extensive co-parenting and household management I thank Dr. Daniel Bronson-Lowe, CIC, FAPIC (Baxter International).

For manuscript review and a lifetime of teaching I thank Theresa G. Lowe, M.A., Special Projects Coordinator, Yuma School District One.

For assistance with the amendment review process I thank Lea Ann Carson, UIUC Institutional Review Board.

For instruction in LabView I thank Zach Jones and Andrew Watchom, National Instruments. I also thank Julia Petrella, Applications Engineer (National Instruments) for technical support.

For assistance in recruiting participants I thank Dr. David Gooler; Dr. Suma Devanga; Cathy Miller; Ben Franklin; Dr. Art Kramer; Anya Knecht; the Champaign-Urbana Lions Club; Dr. Pamela A. Hadley; Dr. Jeffrey Woods; Suzanne Nanney; Daniel Lee; Anthony Early, Mahomet Senior Citizens Group.

For departmental and Graduate College support I thank Dr. Karen Iler Kirk, Dr. Pamela A. Hadley, Dr. Cynthia J. Johnson, Dr. Laura Segebart DeThorne and Dr. Anne T. Kopera.

For nurturing and teaching our daughter so that I could focus on this work, I thank Jefferyann Winfield and the staffs of The Caring Place and Montessori Habitat School.

Finally, I thank my daughter, Rachael Alejandra Bronson-Lowe, for her encouragement, patience, and many acts of kind, thoughtful support while I finished my puzzle.

Table of Contents

Introduction

A Clinical Challenge: Prediction of Therapy Benefit

The question of potential to benefit from therapy arises in many contexts. Patients and their families need to know prognoses and to judge estimated effort, cost and risk of side effects against the potential for benefit. Answering the overall question for a particular patient requires asking two further questions: what does it mean for this patient to benefit from therapy, and what can predict this type of benefit?

What does it mean to benefit from therapy? Bain and Dollaghan (1991) proposed a definition of clinically significant change as "a change in client performance that (a) can be shown to result from treatment rather than from maturation or other uncontrolled factors, (b) can be shown to be real, rather than random, and (c) can be shown to be important, rather than trivial." For a clinically significant change to represent benefit, the change must be an improvement, that is, a change in a desired direction from a measurement taken at baseline to a later measurement.

Further description of benefit can be situated within the International Classification of Functioning, Disability and Health (ICF), a biopsychosocial model adopted by the World Health Organization in 2001. This model attempts to merge the medical and social models of health and disability, acknowledging that biological and psychological factors at the individual level, as well as interpersonal and societal factors, combine in complex ways to produce the constructs of health and disability. The model allows description at various levels (body functions and structures, activities and participation), any of which can interact with environmental and personal contextual factors (World Health Organization, 2002; Vaillancourt, Larsson, & Newell, 2003).

For instance, consider a person with dysphagia, who may have disability at any level of the ICF. Embarrassment from choking or drooling episodes may restrict participation in meals with family. Inability to eat safely by mouth is an activity limitation. A tumor blocking epiglottic inversion, or lingual weakness leading to poor bolus formation and propulsion, is a disruption of bodily structure or function. Environmental factors such as availability of altered-consistency food and liquid or non-oral feeding methods, postural supports, feeding assistance or safety cues interact with personal factors such as cognitive status, willingness to consume altered-consistency diets, and emotional importance attached to mealtime interactions with loved ones to determine safety, participation and quality of life. Other SLP-treated disorders have analogous multilevel effects.

Therapeutic benefit can be assessed at any level. For example, a person reliant on non-oral feeding for primary nutrition may benefit from therapy training family members to safely provide small bites or sips to permit meal participation; from therapy focusing on altering diet/liquid consistencies to permit safe oral intake (activity level); or from therapy addressing airway protection during the swallow (body function level).

Judgment of therapeutic benefit can also be affected by timing. When a disorder is the result of a condition expected to improve with time, showing that change resulted from therapy can be challenging. Bain and Dollaghan (1991) note the need to rule out maturation in younger clients. For speech-language pathologists (SLPs) in the acute setting, establishing the additional effect of therapy above and beyond the effects of spontaneous recovery is notoriously difficult (Marshall, 1997), though some evidence suggests benefit to early intervention for dysphagia (Carnaby, Hankey, & Pizzi, 2006; Geeganage, Beavan, Ellender, & Bath, 2012; Momosaki et al., 2015; Takahata, Tsutsumi, Baba, Nagata, & Yonekura, 2011). In addition, the time elapsed from baseline to later measurement affects the judgment of importance. Within-session improvement may fulfill Bain and Dollaghan's first two conditions; the goal of therapy, however, is to create a more lasting change that generalizes, or transfers, to functional situations in the patient's everyday life.

Though the time periods for "acute" intervention and analysis of its benefit differ, most cover on the order of weeks to months. Marshall (1997) reviewed literature on aphasia treatment in the "early post-onset period" covering 1-3 months post-stroke. The early-intervention dysphagia literature cited above covers interventions ranging from near-immediate (within 24 hours of hospital admission: Takahata et al., 2011) to "acute or subacute…within six months" (Geeganage et al, 2012). All assess the effects of an entire course of treatment. However, particularly in the acute hospital setting where medical status and fatigue may fluctuate significantly over very short time periods, patients and clinicians may need to make on-the-spot judgments about the potential benefit of a single therapy session vs. the cost in effort (for the patient) and time (for the clinician).

Assuming the patient has already been diagnosed with a treatable disorder and deemed to be a treatment candidate, the question here is not "should this patient be offered treatment?" but "can this patient benefit from the planned treatment *right now* (vs. later today or tomorrow, or an alternate treatment)?" When the primary barrier to participation is an acute medical status change, the decision is simple: defer treatment and reassess readiness at the appropriate later time. However, commonly the decision is less clear. Conditions such as fatigue, hypotension or hypoxia may occur in acutely ill patients

and may affect learning and memory (Du, Romano, Aloyo, & Harvey, 1995; Mizunoya, Oyaizu, Hirayama, & Fushiki, 2017; Qaid et al., 2017) or patients' estimates of available vs. required effort, even when mild enough not to simply preclude therapy participation. Their effects on patients' ability to benefit from therapy are unknown. Evidence from healthy young adults practicing unfamiliar motor tasks under conditions of exercise-induced fatigue is mixed (Alderman, 1965; Carron, 1969; Carron & Ferchuk, 1971; Cochran, 1975; Godwin & Schmidt, 1971; Schmidt, 1969) and does not necessarily generalize to older or clinical populations. Should the clinician encourage the patient to participate in therapy, even if performance will likely be relatively poor and the patient feels the required effort is high? Alternatively, should the clinician defer therapy – allowing the patient to rest, but risking a missed therapy day – or switch treatment strategies? What kind of evidence can improve this clinical decision?

Predictors of benefit may be highly dependent on the type of benefit sought. Because of its relevance to treatment of multiple disorders at the body function level, this work focused on the prediction of benefit to motor control: that is, voluntary movement and/or force production with control of timing, sequence, coordination, placement and/or level of force. Examples include voluntary prolongation of the opening of the upper esophageal sphincter during the pharyngeal swallow by prolonging laryngeal movement (Kahrilas, Logemann, Krugler, & Flanagan, 1991) and coordination of the swallow with phase of respiration (Martin-Harris et al., 2015). Behavioral therapy at this level involves learning. Depending on therapeutic need, a patient may be asked to relearn a task, modify performance of a previously learned task, or learn a new task.

What is motor learning? Motor learning is an increase in neural capacity to control an effector system (biological or artificial) to produce a desired output. By analogy to Bain and Dollaghan's (1991) definition of clinically significant change, it results from practice (mental and/or physical) producing that output, rather than from recovery, maturation, or change in the effector system (though these can all coexist with or spur motor learning), and it can be shown to be real, rather than random.

Under this definition, becoming able to play the piano, speak, use a prosthesis after limb amputation, drive a car, write, or type all count as motor learning. Learning the identity of an output (e.g. a phone number or a song) is not motor learning, but becoming able to dial that number with little to no visual attention, or play the song, is. Recovering the ability to draw due to healing of a broken wrist is not motor learning, though learning a new drawing method to compensate for the injury is. Lifting a greater weight due to muscular hypertrophy, or drawing a thicker line by using a thicker pencil or cursor, changes the effector system but is not motor learning; becoming able to coordinate a lifting

effort with respiration and posture-stabilizing muscles' activations, or varying line width using a stylet on a pressure-sensitive tablet, is.

Increased capacity to control cannot be measured directly. Proxy measures include evaluation of changes in behavioral performance and neural activation or microstructure (e.g. cortical maps, synaptic organization and white matter changes (Fields, 2015; Hosp & Luft, 2011). Behavioral performance change is often measured in terms of accuracy (how closely output produced matches the desired output, when this can be known), time to complete task, or an efficiency measure before vs. after practice. Learning is assumed to have taken place when practice is followed by increased accuracy or efficiency or decreased time, or when these measures are unaffected in the face of task alterations such as increased difficulty.

What predicts single-session therapeutic benefit to motor learning? One potential reasonable expectation of benefit from a single session is that improvement in performance of a task targeted in therapy persist until the next therapy session and generalize to closely related therapeutic tasks. Thus an intermediate refinement of the question is: what information (low-cost in patient effort, clinician time and equipment cost to obtain) can predict single-session improvement in task performance which persists for at least one day and transfers to other closely related tasks? An answer may improve short-term prognostication and could guide choice of treatment timing and strategy.

Many previously evaluated predictors of learning are either not readily or cheaply available or are not practical for evaluation of in-the-moment readiness to learn, e.g. profiles of prefrontal and striatal dopaminergic genes (Frank, Doll, Oas-Terpstra, & Moreno, 2009), white and gray matter microstructure and neural activation (Della-Maggiore, Scholz, Johansen-Berg, & Paus, 2009; Kincses et al., 2008; Tomassini et al., 2011).

Short-time-frame motor learning, within a single experimental session, has been predicted from neural activity. Wu, Srinivasan, Kaur, & Cramer (2014) trained healthy right-handed young adults (8/17 women, mean age 22.1 ± 3.0 years) in a pursuit rotor task, for which participants were asked to keep a cursor on a target oscillating along an arc at 50% of the participant's maximum movement speed. Task performance was measured as percent time that the cursor overlapped the target by greater than 50%. Two practice blocks each consisted of four twenty-second trials; three test blocks before, between, and after the practice blocks consisted of a single eighty-second trial each. Learning was quantified as absolute change in task performance from the first to the last test block. Electroencephalography (EEG) data were collected at rest prior to any task performance and during each test block, including average

absolute power at each electrode and β coherence (20-30 Hz) between each pair of electrodes (used to estimate functional connectivity). β coherence with left M1 (contralateral primary motor area for the hand) at rest and during the first test block significantly predicted learning.

Özdenizci et al. (2017) trained healthy right-handed young adults (7/21 women, mean age 23.8 ± 3.1 years) in a rapid-reaching task with adaptation to a velocity-dependent force field. Participants moved a manipulandum, which resisted in a velocity-dependent fashion, from a central resting position to one of four randomly selected, equally spaced targets. Task performance was measured as the area between the observed movement trajectory to target position and the ideal straight line between those points. Improvement was quantified as the ratio of mean performance during trials 1-10 to mean performance during trials 30-40. EEG data were recorded for all trials and showed that beta activity in sensorimotor areas, fronto-parietal attention networks and parieto-occipital areas was predictive of improvement in task performance. No follow-up was done to assess longer-term retention or transfer of learning in either study.

While predictively valuable and focused on immediate readiness for learning, this type of data is available only in high-resource settings, and requires more trials than practical for a high-fatigue clinical patient. In the context of therapy focused on motor control, motor performance measures may serve as low-cost, easily obtainable predictors.

Stimulability (ability to produce target output with the assistance of a cue or model) has been recognized across a variety of populations as a marker of readiness to learn, dating from Milisen (1954, as cited in Long & Olswang, 1996). Much of this assessment has been conducted with children. Bain and Olswang (1995), for instance, found that for children 30-36 months of age with specific expressive language impairment and mean length of utterance from 1.0-1.3 morphemes, greater initial stimulability for production of two-word utterances predicted greater language development over nine weeks. Beeson, Rising and Volk (2003) , on the other hand, treated older adults with persistent severe aphasia at least two years post-stroke, using Copy and Recall Treatment (“repeated copying of target words in the presence of pictured stimuli, followed by recall trials in the form of written picture naming”) over 4-5 months. The only one of their eight participants to show no response to treatment was also the only one who showed no within-session mastery of targeted words (their definition of stimulability).

Stimulability may thus offer a rapid, low-cost predictor of therapeutic benefit. However, it has limitations for use in acute care settings, where people with dysphagia often make up a large portion of

those referred for SLP services. First, development of procedures for the dynamic assessment of stimulability within speech-language pathology has focused on speech and language tasks. This limitation could be overcome with development of a dysphagia-specific cueing hierarchy. More importantly, it may not be safe to assess stimulability for some target tasks with some patients, e.g. any task involving intake of food or liquid by mouth in a patient known to have high risk of aspiration sequelae. Any predictive task and measure should be safely useable with this population.

One promising possibility is the measurement of initial accuracy and variability in performance of pursuit tracking tasks. This combination offers a way to rapidly collect data on motor learning in control-focused tasks, which at low force should be minimally taxing.

Pursuit Tracking Task

The pursuit tracking task allows examination of motor accuracy, variability, and learning. It elicits continuous behavior[1] as opposed to the discrete movements of a ballistic aiming or stop consonant production task. Versions of this task have been used for at least sixty years (Chernikoff, Birmingham, & Taylor, 1955; Craik, 1947; Noble, Fitts, & Warren, 1955); the modern version detailed here applies to all studies labeled "pursuit tracking" in this work unless otherwise noted. Participants are seated in front of a monitor on which a target pattern is displayed. A second visual trace on the same monitor moves rightward; participants control its height by pressing on a load cell with greater force to move the trace upwards, or reduced force to allow the trace to descend. Participants are asked to make the trace match the target pattern. Target force levels are usually determined with respect to participants' maximal voluntary contractions (MVC) and are normally well below maximum force. In all such studies described in this document unless otherwise noted, participants used the dominant side for non-oral effectors; for studies investigating learning, participants were given knowledge-of-results type feedback (e.g. absolute or root mean square error) after each trial. Time series of force output are analyzed, with varying short lengths of the initial data cropped to avoid including initial approach to and stabilization around the target. In some studies 1-2 final seconds of data are also cropped to avoid anticipatory reduction of effort.

[1] "Continuous behavior" refers to continuous isometric force in most studies described here, but other variants exist. Craik (1947) discusses turning a handle with constant angular velocity or acceleration, and Franks, Wilberg and Fishburne (1982) asked participants to pivot a lever to track one-dimensional sinusoidal variation.

This paradigm presents several benefits. Because the target is known, deviation from it (accuracy and variability) can be quantified; because it is controlled by the experimenter, target shape, duration and force level can be systematically manipulated. Because many data points per second can be collected, within-trial variability can be assessed, but trials can remain short enough to permit multiple repetitions for assessment of inter-trial variability. Because the idea is simple and targets are set at participant-specific levels, participants of a wide range of ages and strengths can attempt the task. Its main disadvantage is that it is not representative of most functional tasks.

Motor Learning and Initial Accuracy

Initial accuracy could theoretically predict learning in two contrasting ways. Higher initial accuracy could indicate task aptitude, in which case it should predict greater learning. Under this reading, accuracy is related to stimulability: rather than fixing a criterion for acceptable performance and assessing the level of support required to achieve it, one fixes a level of support (here, a visible target and standard instruction) and assesses the resulting performance. Alternately, initial accuracy could indicate room to improve, in which case lower initial accuracy would predict greater learning. This interpretation can only be valid where performance improvement is possible for the population evaluated – i.e. where room to improve exists. Results supporting this interpretation for relatively easy tasks or for healthy populations may not generalize to difficult tasks or clinical populations.

Limited data support the second interpretation. Barbado Murillo, Caballero Sánchez, Moreside, Vera-García, and Moreno (2017) predicted within-session learning during an error-based task (standing balance on an unstable surface). Learning was measured as performance improvement: the difference from baseline to post-practice in average distance of the center of pressure from its mean position. It was significantly correlated with initial performance, with the direction of correlation suggesting that worse initial performance indicated greater room for improvement. Wu et al. (2014, reviewed above) also evaluated initial accuracy in their pursuit rotor task (percent time on target at test 1) as a predictor of within-session motor learning (percent improvement in time on target from test 1 to test 3). It was not significant, perhaps because their measure focused on time meeting an arbitrary criterion of accuracy rather than distance from target. This work used an accuracy measure quantifying distance from target. Because participants were healthy adults and the tasks were expected to be learnable, initial accuracy was expected to function as a measure of room for improvement rather than aptitude.

Motor Learning and Variability

Skilled motor performance requires reliably producing accurate, well-timed responses to varying internal and external stimuli in service of diverse goals from a range of starting conditions. When an action is performed multiple times, even if the goal and starting conditions are held constant, there will be differences between trials in timing, magnitude and/or duration of the action, even if all fall within acceptable limits on each trial; or, if an action is prolonged (e.g. maintenance of a posture or a level of force), moment-to-moment performance will vary around the goal. Accounting for this variability has been an important goal of multiple theories of motor control and learning.

Information-theoretic perspective. One influential view of motor control, often called the information-processing or information-theoretic perspective, can be traced to Shannon's (1948) work developing a general theory of communication (independent of medium or message). He posited a model in which a message is transmitted from a source to a destination, with stochastic noise modifying the signal. The noise is modeled as white Gaussian noise, which requires independence between successive values and equal representation of all frequencies (Pierce, 1980). Applied to motor control, the message is a motor command, sent from the motor cortex (source) to an effector (destination). Mean performance is assumed to indicate the intended motor command, while variability around the mean is assumed to result from noise. This view suggests that variability marks poor skill, is to be avoided, and will decrease as a task is learned. Operationally, measures of the magnitude of variability, such as standard deviation (SD), coefficient of variation (CV) or spatiotemporal index (STI)[2] (Smith, Goffman, Zelaznik, Ying, & McGillem, 1995), are predicted to decrease with practice; or the assumed signal can be compared to the assumed noise as a signal-to-noise ratio (SNR, mean / SD), which would be expected to increase with practice.

The predicted changes in magnitude of variability with learning have been seen in both oral and manual motor control literature. For example, Grigos (2009) examined articulator movement variability across productions in a longitudinal study of six typically developing children (initial mean age 20 months) acquiring a voicing contrast (/p/ vs. /b/). Children began the study, not having acquired the voicing contrast, with voice onset time (VOT) for /p/ < 20 ms. They were seen every three weeks until

[2] CV = SD / mean (the reciprocal of the signal-to-noise ratio described above). STI is the sum of standard deviations calculated at 2% intervals over repetitions of time- and amplitude-normalized trajectories. Lower values indicate lesser variability.

they acquired the contrast (12-21 weeks), meaning that they produced VOT for /p/ > 25 ms, plus a perceptible voicing contrast on at least 90% of occurrences across two consecutive sessions three week apart. In each session, coordinates of reflective markers on the lips and jaw (relative to a forehead marker) were captured during videotaping of elicited productions of /papa/ and /baba/ in play scenarios and digitized. Articulator movements associated with production of /p/ and /b/ were identified from time-aligned acoustic recordings, refined algorithmically based on zero-crossings of the jaw velocity signal. STI was calculated to describe variability across productions. For both lips and the jaw, STI decreased from pre-acquisition to acquisition, but only for /papa/, not for /baba/. Voice onset time increased across sessions as STI decreased, with four of six participants showing a significant negative correlation between the two, again for /papa/ only. The author interpreted these findings as evidence for learning of articulatory strategy, rather than maturation, which she had hypothesized would affect both voiced and voiceless production.

Reduced magnitude of variability with learning has been found in a nonspeech oral motor task in healthy adults as well. Testa, Rolando, and Roatta (2011) asked seventeen participants (9 women; mean age 28.4 ± 6.67 years) to complete a pursuit tracking task using unilateral jaw-closing force on an intermolar sensor to match five-second constant targets at 10%, 20%, 30%, 50% and 70% maximal voluntary contraction (MVC). The set of target force levels was repeated three times per side of the mouth, alternating sides, on each of two consecutive days. Performance was assessed using mean distance (MD: mean absolute difference between target and force produced; MD > 0), offset error (OE: difference between average force and target; OE > 0 indicates overshoot, OE < 0 undershoot) and coefficient of variation (CV). Findings included increased accuracy (reduced MD and OE) and decreased magnitude of variability (CV) from day 1 to day 2. The authors interpreted these findings as evidence of a learning effect, though they cautioned that given the limited practice, learning may not have reached a plateau. For the purpose of this work, it is notable that even such a limited training regimen produced a decrease in variability.

Further examples abound (Deutsch & Newell, 2004; Mukherjee, Koutakis, Siu, Fayad, & Stergiou, 2013; Newell et al., 2003; Sosnoff & Voudrie, 2009). However, while these investigations do support the prediction of the information-theoretic perspective that magnitude of variability should decrease or SNR increase with practice, they also showcase an important limitation. As noted by Deutsch and Newell (2004), mean and standard deviation, because their calculation does not account for data order, cannot address the time or frequency structure of the data and therefore cannot test the claim that motor

variability merely represents noise. Multiple studies, therefore, have incorporated measures of time and frequency structure[3] into their data analyses. Where the structure of motor time series data has been found not to fit the characteristics posited within the information-theoretic perspective (successive values independent, all frequencies equally represented), then its claim that motor variability represents noise has not been supported. Rather, when values show some dependence on previous values, then a causal process or multiple interacting processes occurring across time can be hypothesized to exist, and these can be associated with characteristic prominent frequencies in the power spectrum. Thus while the magnitude of variability changes with practice as predicted by the information-theoretic perspective, the structure of variability, rather than representing noise, provides a rich information source regarding motor control processes. This approach is the dynamical systems framework.

Dynamical systems framework. This framework descends from Bernstein's concept of motor learning as mastery of redundant degrees of freedom (Latash, Scholz, & Schöner, 2007; Bernstein, 1967, as cited in Newell et al., 2003). "Degrees of freedom" may refer to either the physical system at any level of analysis (e.g. joints, muscles, motor units, neurons) or the active or dynamical degrees of freedom (dimension) describing a behavior in state space.[4] Consider moving an effector to a target location, such as the tip of the index finger to a letter on a keyboard or the tip of the tongue to the alveolar ridge. In state space, the many possible paths to the target are specified in three dimensions, but at any level of motor control, far more than three elements act to produce the desired outcome.

A goal in state space – a desired configuration of a system – is called an attractor. If the goal is a steady state, such as constant isometric force production, the attractor in state space is a fixed point with zero dimension (Vaillancourt & Newell, 2002). Higher-dimension attractors can also exist. Sinusoidally varying isometric force, for instance, is an example of a limit cycle oscillator with one

---

[3] Each such measurement used by a study included in this review has a brief description in Appendix A. Measures used in this study are covered in greater detail.

[4] "State space" describes the possible states of a system and contains one dimension per descriptor. Time-varying system configuration is represented as a path through state space. To describe the relevant state (here, physical location) of the index fingertip in the main-text example, the state space needs three dimensions corresponding to the three physical spatial dimensions. If the relevant state were "force produced perpendicular to the surface of a pressure transducer," only one dimension would be needed. These state spaces describe *outcomes*. Approximate entropy, correlation dimension and other related measures (Appendix A) use outcome signals to estimate a reconstruction of a state space describing the process(es) producing the outcome.

dimension. The problem of motor learning can be viewed as learning to keep the system near the attractor in state space.

This does not imply that the dimension of motor output must approach the dimension of the task attractor. Newell and Vaillancourt (2001) and Vaillancourt and Newell (2002) proposed that a complex control system attempting to maintain output near a steady state must dampen the effects of any high-amplitude or dominant rhythms which, unopposed, would cause the system's state to oscillate outside the boundaries of acceptability. This damping is accomplished by multiple other, offset rhythms; their summed output maintains the system within a threshold deviation from the fixed goal state. The presence of multiple rhythms in system output drives its dimension higher, rather than closer to the dimension of the fixed-point attractor (zero). When the attractor is a limit cycle oscillator, on the other hand, the dimension of the attractor (one) is higher than that of a fixed-point attractor, but the dimension of ideal motor output is expected to be lower than that of ideal motor output for a steady state, because a dominant rhythm matching that of the attractor should emerge, while other rhythms should be suppressed. This argument is not specific to a particular effector nor even to voluntary motor control, and thus should apply to oral motor as well as to manual motor control.

This viewpoint does not contradict the information-theoretic perspective's prediction that *magnitude* of variability should decrease with practice, because successful emphasis or damping of rhythms maintaining a system near an attractor should reduce variability and error around the target state (Newell et al., 2003; Slifkin & Newell, 1999; Slifkin & Newell, 2000; Slifkin, Vaillancourt, & Newell, 2000). However, it does make two predictions conflicting with the information-theoretic perspective: 1) the temporal or frequency structure of variability should differ from random noise, reflecting coordination of multiple influences on system behavior, and 2) changes should be expected in the *structure* of variability with practice as system coordination alters to more closely approximate output to a task goal, with the direction of these changes dependent upon the dimension of the task attractor.

Several pursuit tracking studies supporting these predictions across the lifespan, each with healthy participants using the dominant hand, are described below. Each study's analyses included time- and frequency-domain measures of force structure: approximate entropy (ApEn) and proportion of power (PoP).

Focusing on children, Deutsch and Newell (2004) investigated the effects of age and practice on force variability to determine whether reductions in magnitude of motor variability seen in older children and young adults were due to reduction in sensorimotor system noise. Three groups of

participants (20 with mean age 6.4 ± 0.29, 18 with mean age 10.6 ± 0.26, 20 with mean age 20.7 ± 1.38; handedness unspecified) were asked to use a pinch grip (thumb and index finger against oppositely oriented load cells with output combined) to match a constant target in the pursuit tracking task. Within each age group, half of participants' target force levels were set at 5% MVC and half at 25% MVC. Participants performed fifteen fifteen-second trials on each of five consecutive days. Their force output was recorded at 50 Hz, with the initial five seconds cropped, leaving time series of length 10 s x 50/s = 500 for analysis. All analyses were completed on the index finger and thumb individually and on their combined signal; as results were similar, only those for the combined signal were reported. PoP analysis included three frequency bands (0-4, 4-8 and 8-12 Hz). Signal-to-noise ratio (SNR) was calculated as mean / standard deviation.

SNR was found to increase as a function of practice. Practice-associated increase was greater for the 25% MVC target level regardless of age group, and greater for the young adult group than the children's groups regardless of target force level. ApEn increased with both age and practice, though the effect of practice on ApEn became less as age increased. (No comparisons between specific age groups were reported.) Within frequency bands, practice was associated with declining PoP in the 0-4 Hz range, but no change within the higher bands. Comparison across frequency bands showed that PoP was greater in the 0-4 Hz band than in the higher bands across all age groups, target force levels and days of practice. PoP did not differ between the two higher-frequency bands on the first day of practice, but was higher from 4-8 Hz than from 8-12 Hz on remaining days.

The authors interpreted these findings as demonstrating that for 6-, 10- and 20-year-olds, practice matching a constant force target led to a less regular (less predictable) force structure with a broader frequency profile; chronological or developmental age was responsible for less of the variance in performance than practice. That is, the structure of variability in participants' force production was meaningful, not random, and it changed with practice in a way expected from the dimension of the task attractor (a fixed, zero-dimension point in state space).

Newell et al. (2003) recruited young to middle-aged adults to investigate changes in structure of force output depending upon practice and the dimension of the task attractor. Separate experiments were run for constant vs. sine targets using the dominant index finger in the pursuit tracking task. Twelve adults (ages 24-50) practiced matching a constant target, and twelve (ages 22-25) a 1-Hz sine target; this frequency was chosen based on previous work finding it to be most participants' dominant output frequency in continuous isometric force production (Slifkin & Newell, 1999; Slifkin et al., 2000).

Within each task group, participants were assigned randomly to either 10% or 40% MVC target force levels. Participants performed twenty-five twenty-five-second trials on each of six consecutive days. On the sixth day, both visual traces (target and participant-controlled) and error feedback were suppressed. Force output was recorded at 50 Hz, with the initial five seconds cropped, leaving time series of length 20 s x 50/s = 1000 for analysis. In addition to root mean square error (RMSE) and ApEn, correlation dimension was calculated for a subset of the trials. Proportion of power (PoP) analyses split the 0-5 Hz range into 17 equal bins (0.2928 Hz).

For the adults matching the constant force target, ApEn increased over the five days of practice with the target visible, regardless of target force level, then decreased on the sixth day with removal of visual information. Modal frequency (unspecified frequency with the greatest power) did not change with practice, but the level of power at that frequency decreased, and range in which 95% of the power within the 0-5 Hz band was concentrated increased from a mean of 2.25 Hz on the first day to 3.10 Hz on the fifth: that is, power spectrum became more broadband. Correlation dimension increased as a function of practice and was higher in the 10% MVC target force level group than the 40% MVC group, but decreased on withdrawal of visual information.

For the adults matching the sine force target, ApEn decreased over the five days of practice with the target visible, regardless of target force level, then decreased further on the sixth day with removal of visual information. Modal frequency (0.99 Hz vs. the 1-Hz target) did not change with practice, but the level of power at that frequency increased, and range in which 95% of the power within the 0-5 Hz band was concentrated decreased from a mean of 2.11 Hz on the first day to 1.11 Hz on the fifth: that is, power spectrum became more narrowband. Correlation dimension decreased as a function of practice regardless of target force level group, but increased on withdrawal of visual information.

The authors interpreted these findings as support for a practice-associated and crucially task-dependent change in number of dynamical degrees of freedom. In their view, practice-associated changes in coordination of system components results from both task goals and physical and other (e.g. informational or energetic) constraints on action.

In their study of younger and older adults, Sosnoff & Voudrie (2009) examined the influences of age, task, and practice on the temporal structure of isometric force variability to determine whether older adults' lesser adaptability to task constraints could be attributed to lack of task familiarity or was characteristic of aging. Right-handed younger and older adults (36 each, mean ages 22.9 ± 3.4 and 72.1 ± 4.5 years respectively) were asked to use right index finger flexion onto a load cell to match constant

and sinusoidal targets. Targets were centered at 15% MVC, with the sinusoid fluctuating by ±5% MVC at 1 Hz. Participants performed five twenty-second trials for each target (order counterbalanced) on each of five consecutive days. Their force output was recorded at 140 Hz, with the initial seven and final one seconds cropped, leaving time series of length 12 s x 140/s = 1680 for analysis. PoP analysis was completed separately for three frequency bands (0-4, 4-8 and 8-12 Hz). In addition to ApEn and PoP, Sosnoff and Voudrie evaluated adaptability across task (ApEn difference score for individual constant vs. sinusoidal trials). Though the authors' focus was on temporal structure of force, they also calculated coefficient of variation (CV = standard deviation / mean) for each time series. Three-way mixed model ANOVA (age group x day x target) was used to analyze dependent variables' mean values over the five trials for each target x day condition.

For both tasks, they found decreasing CV associated with practice, with no further significant change following the third day of practice. The effects of practice and age group on force structure varied by task. For the constant task, practice was associated with increased ApEn, decreased PoP in the 0-4 Hz band, and increased PoP in the higher bands; younger adult participants had higher ApEn than older adult participants. For the sine task, practice was associated with decreased ApEn, increased PoP in the 0-4 Hz band, and decreased PoP in the higher bands; younger adult participants had lower ApEn than older adult participants. ApEn difference score increased with practice and was greater for younger adults than older adults throughout practice.

The results are consistent with those of Newell et al. (2003), in that a fixed-point task attractor elicited less regular output and reduced pre-eminence of a single frequency band after practice, while practice matching a limit cycle attractor was associated with the opposite findings. While the reduction in magnitude of variability with practice could be explained by either the information-theoretic or dynamical systems perspective, only the latter can explain the systematic practice-associated, task-dependent changes in structure of variability. [5]

The dynamical systems framework has also been applied to the assessment of motor learning in a clinical population. Mukherjee, Koutakis, Siu, Fayad and Stergiou (2013) examined stroke survivors' adaptation to a variable force field in an aiming task to determine how variability of hand movement changed as participants learned the task. Twelve premorbidly right-handed participants (four with right-

[5] This study is discussed further in 'Generalization to healthy aging.'

sided weakness; two women; 1 hemorrhagic, 11 ischemic strokes, all unilateral; mean age 62.92 ± 8.07 years; mean time post-onset 18.58 ± 12.47 months) were randomly assigned to experimental and control groups. Both groups used their affected hand grasping a robotic manipulandum to control a cursor, moving it from a central starting position to one of eight equiangularly spaced peripheral targets. Targets were ordered in counterclockwise sequence over cycles of 8 trials. After eighty practice trials, participants completed forty trials with no alteration from the baseline task, then 240 trials each on that day and the next with an additional force exerted through the manipulandum, perpendicular to hand velocity. Participants in the experimental group (unbeknownst to them) had augmented visual feedback of error: while cursor position along the line from start to target was unchanged, cursor distance from that ideal-performance line was doubled. After the force-altered trials on the second day, participants completed a further forty washout trials with no additional force (same as baseline task). One week later, participants completed forty more unaltered-task trials to test transfer of skill to a non-dynamic environment, followed by forty trials in the practiced dynamic-force environment to test retention.

Positional time series data were recorded as deviation perpendicular to the line between starting position and target; SD and ApEn of these time series were determined. Mixed ANOVAs were used to detect significant effects of participant group and trial type on the time series measures. There were no main effects of participant group or interactions between group and trial type. ApEn increased significantly with practice in the dynamic-force environment, then decreased during washout trials. On dynamic-force retention trials, ApEn remained higher than during the early trials of day 1, but this increase did not transfer to unaltered-task trials without the additional dynamic force. SD significantly increased from baseline when the dynamic force was introduced, decreased with practice within the dynamic-force environment, then increased again on washout trials. These practice-related changes in SD did not persist until retention trials, and neither SD nor ApEn showed transfer to reaching in the absence of the variable force field.

Because changes in structure of variability were maintained until retention trials, while changes in magnitude of variability were not, the authors concluded that structure of variability was a more sensitive indicator of motor training than magnitude of variability, for this task and these participants.

In sum, the dynamical systems framework has been used productively to analyze manual motor learning in healthy participants across a wide range of ages, in a clinical population, and using both isometric continuous force and aiming tasks. Results suggest that structure of motor variability is not

consistent with a "noise" interpretation, but rather contains relevant information on how body systems coordinate to learn to produce desired motor output.

Optimal variability as support for motor learning. Newell et al. (2003) suggested an additional idea contrasting with the information-theoretic perspective on variability as a sign of poor performance. They describe initial changes in structure of motor output as indicative of attempts to match task demands by varying coordination strategies within the space of available system components, their possible couplings, and various constraints (efficiency, sensory input, cost-benefit ratio of failure to success, etc.). That is, initial structural variability may actually support learning if it indicates improved exploration of the task. In this reading it should relate to stimulability, as greater unaided exploration of a task space should predispose a learner towards taking advantage of support to either expand the task space or exploit previously discovered best strategies within the known task space.

Wu, Miyamoto, Gonzalez Castro, Ölveczky, and Smith (2014) tested whether early variability supported motor learning. They asked participants to trace trajectories shown on a screen, with the view of their active hand occluded. Baseline trials were completed without feedback, following which participants were trained on trajectories for which they were given knowledge-of-results feedback after each trial. Participants whose task-relevant variability during baseline attempts was higher than the mean showed more rapid performance improvement over hundreds of trials than those whose baseline task-relevant variability was below the mean.

Stergiou, Harbourne, and Cavanaugh (2006) discuss potential learning-supportive characteristics of variability. They describe temporal structure of time series in terms of complexity vs. predictability. A simple series such as $f(t) = c$ or $f(t) = sin(t)$ has both low entropy (high predictability) and low complexity. A series of random numbers has high entropy (low predictability), while remaining low-complexity, because it is generated by a simple random process rather than multiple processes coordinating over varying timescales. A high-complexity series may appear to be random, but actually contain underlying deterministic structure (with or without a random component), yielding intermediate predictability. See Figure 1, adapted from Stergiou et al.'s Figure 2. Stergiou et al. suggest that high-complexity (specifically, chaotic) output structure is ideal for physiologic systems in general and specifically for motor control, indicating capacity to flexibly adapt to stress, perturbation, and changing goals.

Chaos arises from deterministic systems with nonlinear responses to perturbation (Bassingthwaighte, Liebovitch, & West, 1994; Faure & Korn, 2001). These systems are "sensitively dependent upon initial conditions," as state-space trajectories from similar states may rapidly diverge.

Thus despite their deterministic nature, their output is difficult to predict. It is bounded but aperiodic and so may appear random or noisy (Pincus, 2001). Such a system is capable of periods of stable output, but small perturbations can grow rapidly. The concept of sensitive dependence upon initial conditions is important for biological systems in general and motor control in particular because relevant internal and external factors (e.g. tongue position and muscle activation level) are in constant flux.

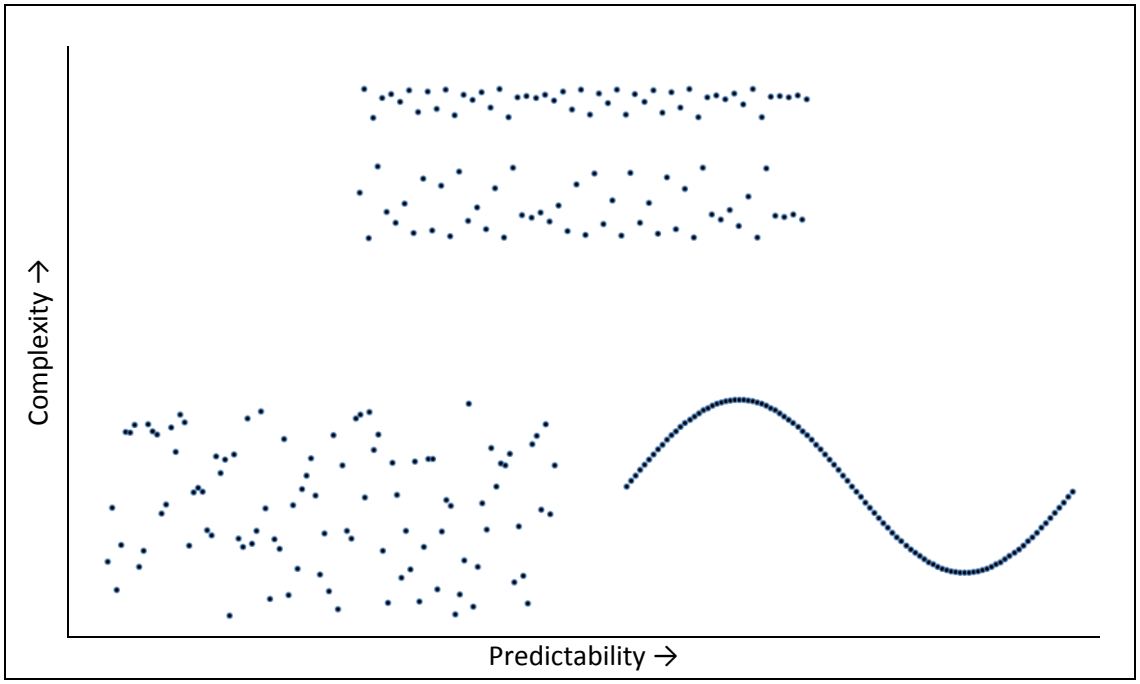

Figure 1. Time series complexity vs. predictability. Adapted from Figure 2 of Stergiou, Harbourne and Cavanaugh (2006). Lower left series: random. Lower right series: sine. Upper central series: logistic equation [$x_{i+1} = kx_i(1 - x_i)$, $k = 3.6$, $x_0 = 0.5$].

Temporal and frequency structure of time series data is quantified using nonlinear measurement tools, including those described in Appendix A. The studies reviewed in the previous section showed that these nonlinear analyses can fruitfully characterize structure of motor performance data in the context of learning, but why should they suggest an optimal variability?

A highly regular, ordered (low-entropy) motor output signal (when not required by the task or goal) suggests inappropriate domination by a single control process. This type of control may allow only limited flexibility to respond to a perturbation or to scale output to a goal, as only one component's frequency, amplitude and phase can be adjusted. At the other extreme, a highly irregular, disordered or random signal (high entropy) suggests that no ordered control process has significant influence. A high-

entropy motor output signal, therefore, may also be associated with poor motor performance because no controlling process acts to guide the system towards a goal or in response to a perturbation. Chaotic and other complex processes have intermediate levels of order or entropy, suggesting multiple control processes coordinated across timescales. Melby (2002) showed that self-adjusting, dynamical systems adapt their control parameters to function "at the edge of chaos," where a small perturbation suffices to shift between periodic and chaotic output. This adaptation represents an effective control strategy, as a chaotic system can be controlled more rapidly, finely and efficiently than a linear system (Bassingthwaighte et al., 1994).

Rapid, fine, efficient control could support adaptability to stress, perturbation or differing task demands (Stergiou et al., 2006; Stergiou & Decker, 2011). Stergiou et al. thus suggest that healthy biological systems have optimal variability characterized by complex, chaotic structure; deviation from this ideal leads to overly rigid behavior in the case of too-predictable structure, or unstable behavior in the case of noisy, too-unpredictable structure, and thus to loss of health. They propose that healthy development and learning are also characterized by this complex, chaotic structure of variability supporting adaptability to perturbation and differing task demands. Clinical populations with altered structure of motor variability might then display motor behavior that is "rigid, inflexible, and highly predictable (i.e., stereotypical), or alternatively, random, unfocused, and unpredictable" (Stergiou et al., 2006, p. 121). Stergiou and colleagues proposed an intervention strategy: nurture development of optimal variability by teaching a variety of strategies to promote active engagement of the individual within their environment.

They illustrated their proposed therapeutic and measurement approach with two case studies focusing on postural motor control. In the first, two one-year-old boys with right spastic hemiplegic cerebral palsy were treated for gross motor delays. One boy (JK) was given physical therapy with a focus on increased variability of movement. The other (LM) was given a home exercise program "of a more static nature" (p. 125). Center-of-pressure time series during sitting were quantified using linear and nonlinear measures. After two months of treatment, both boys achieved independent sitting and reaching for objects; only JK also developed multiple other posture-change skills and limited right hand use. One month post-intervention, JK's gross motor development continued to be greater than LM's, with multiple postural changes and early ambulation. The authors attributed JK's greater progress to increased complexity of postural control seen in the nonlinear measurements.

The second case study assessed standing postural control in an 18-year-old male athlete recovering post-concussion, for whom pre-concussion baseline data were available. Center-of-pressure time series during standing were assessed using linear and nonlinear measures (Equilibrium Score and ApEn respectively) at the beginning of soccer season and daily for four days after the concussion. While Equilibrium Score returned to baseline level by the fourth day post-concussion, ApEn remained below baseline and in fact diminished further on the fourth day. The authors suggested that nonlinear measures may detect subtle motor impairments not visible using linear measures.

Optimal variability may support non-motor learning as well, as typically developing children could be distinguished from children with dyslexia using nonlinear measures of variability. Wijnants, Hasselman, Cox, Bosman, and Van Orden (2012) assessed single-word reading in 15 dyslexic children (ages 7-8) and 15 "reading-age"-matched children (ages 6-7). Five hundred fifty one-syllable (2-8 letter) words were presented individually on a computer screen, which the children were asked to read as quickly and accurately as possible. Response times were recorded with millisecond precision; inter-stimulus interval was 500 ms. Ordered reaction times series were analyzed using linear (mean, SD) and nonlinear measures (PoP, standardized dispersion analysis, detrended fluctuation analysis and recurrence quantification analysis). Results showed that, in addition to slower and more variable response times (higher mean and SD), dyslexic children's frequency structure of response times was closer to white noise and showed lower recurrence rates and higher fractal dimension, while non-dyslexic children showed 1/f scaling (indicative of complexity; see Appendix A: Frequency Domain Measures: Spectral Slope). Though the study task did not itself involve learning, the participants were selected to vary in their ability to learn in the domain tested. The authors argued that their results should be interpreted in line with similar results from physical and physiological systems, in which 1/f scaling suggests interrelated task-specific control processes coordinating over multiple timescales and the emergence of successful behavior from the interactive dynamics of the entire system rather than any component in isolation.

In accord with Stergiou and colleagues' work (Stergiou, Harbourne, & Cavanaugh, 2006; Stergiou & Decker, 2011), the association of task-optimal complexity with improved current performance may extend to indicating physiological capacity or readiness of the motor system to change its performance. Flexible coordination of motor control processes operating across different time scales may enhance both exploration of an unfamiliar task space, via generation of varying responses, and ability to shape

fluctuations in performance towards a task goal. Changes in temporal and frequency structure co-occurring with learning and recovery may thus have predictive value.

Baseline force structure as a predictor of learning. Prediction of later change from nonlinear measures of system function has been previously observed in physiological systems. For instance, Fleisher, Pincus, and Rosenbaum (1993) collected heartbeat interval data pre- and post-operatively for 23 high-risk adult patients undergoing elective surgical procedures. Nine (age 76 ± 9) had post-operative ventricular dysfunction, while 14 (age 69 ± 10) did not. While mean pre-operative ApEn of the heartbeat interval time series did not differ between groups, lowest post-op ApEn (collected before adverse events) was lower in the group with ventricular dysfunction. Two consecutive post-op values of ApEn < 0.7, 8-10 hours apart, had a sensitivity of 88% and specificity of 79% for association with ventricular dysfunction (positive predictive value of 73%, negative predictive value of 92%[6]). The authors suggested that post-operative measurement of heartrate ApEn could provide an early indicator of upcoming ventricular dysfunction. Similarly, (Kluge et al., 1988) found that infants who later died of sudden infant death syndrome (SIDS) differed from control infants in the extent of coordination between cardiac and respiratory activity (extent of respiratory sinus arrhythmia).

If nonlinear measures can detect subtle indicators of disease (with expected future deterioration) or current difficulty learning, they may be able to predict learning as well. There is limited evidence in support of this idea. Barbado Murillo et al. (2017; methods review earlier) performed detrended fluctuation analysis, a measure of the structure of motor variability, on center-of-pressure time series collected during a standing balance task on an unstable surface. They found that participants with relatively unstructured motor variability (interpreted as greater ability to adjust posture) improved performance within a practice session more than those whose motor variability was highly structured. This investigation extends that work to oral motor control, to healthy older adults, to tasks varying by structure of ideal performance, and to retention and transfer trials one day after cessation of practice. If baseline force structure is predictive of later learned performance, such information could be used to improve prognostic estimates and/or guide (re)habilitative strategies for disordered speech or swallow.

[6] Calculated from data in the paper; the authors provided only sensitivity and specificity.

Measurement of baseline force structure. What measure of baseline force structure is most appropriate for prediction of learning? This investigation evaluated two possibilities: (1) maximal entropy, and (2) change in entropy from the first to the second attempt at each task.

Maximal entropy during initial task performance. Humans produce force to match a target via coordination of multiple neural oscillators (central drives operating over varying timescales) whose contributions add dynamical degrees of freedom, i.e. dimension, to the force signal (Newell et al., 2003). A higher-dimensional signal may indicate greater independence among the oscillators, while lower dimension may indicate that the oscillators are more tightly coupled. Lower maximal dimensionality may indicate lesser ability to decouple oscillators to vary force along independent dimensions, possibly limiting flexibility to alter performance. In turn, one would then predict lesser motor learning, because more ideal couplings of force control processes may lie outside the available solution space.

A caveat should be made explicit: task alterations such as changed target force level, isometric vs. non-isometric or discrete vs. continuous force may add, subtract or modulate the potential contribution of force control processes. Since most functional task goals require absolute rather than relative force or timing, individual physiological variability will also dictate limits. Thus a baseline measure may be limited to predicting learning in tasks which are similar in some respect such as target force level, effector or muscle contraction type.

How should this hypothesis be operationalized? Though the discussion has been in terms of force dimension, this investigation did not use the correlation dimension measurement, because it was developed for use with completely deterministic, noiseless systems with unlimited amounts of data and was not intended for statistical uses (Pincus, 1995). Pincus recommended approximate entropy due to its agreement with dimension changes for low-dimensional systems and its applicability to noisy, finite-length data sets. This investigation used approximate entropy as well as one of its descendants, fuzzy measure entropy, determining a maximum for each effector x task condition. These measures are unlikely to accurately represent the true maximum of which a participant is capable and are better viewed as an estimate over a small sample of initial attempts at unfamiliar tasks.

Adaptability of force structure to disparate goals. If the goal of force production (in state-space terms) is to match the trajectory of force output to the task attractor, then the ability to produce particularly high-dimension output may be less important than the ability to adjust the temporal/frequency structure of the output to match the task attractor within the space of possible states. Immediate within-task adaptability, or the effectiveness of search strategy within the state space,

may then be a better predictor of eventual skill. For instance, if the task demands very regular (low-entropy) output, and one's initial attempt produces force of moderate entropy, lower- entropy force on the second attempt suggests adaptation in a direction congruent with task demand, while higher-entropy force on the second attempt suggests a maladaptive change.

Predicting learning outcomes from an immediate adaptability measure rather than single-trial entropy presents two advantages. First, comparison of successive trials suggests how much a participant can learn from the minimum possible increment of practice and feedback, potentially making a stronger predictor than a static measure. Second, it may offer improved clinical applicability: if a patient cannot perform the goal task at all at baseline, one can still construct a rougher prediction by using a proxy task similar in relevant respects to the true goal task, but within the patient's capabilities. If the proxy task were selected to require control processes similar to those required by the goal task, it could function as an effective alternate predictor.

Various measures of adaptability are possible. The ideal measure would account for not only absolute change in entropy (e.g. Sosnoff & Voudrie, 2009), but also for the relative change compared to both the participant's maximum possible entropy and the entropy of the target. However, given the difficulty of measuring maximum possible entropy, and the known differences between entropy of a target trace and entropy of ideal human performance following that trace, absolute difference in entropy was used here for simplicity.

Generalization to oral effectors. It has been suggested that due to the many differences in structure and function among oral effectors and between oral and non-oral effectors, findings may not generalize from literature evaluating non-oral effectors, or from one oral effector to another (Kent, 2004). For instance, while lateral difference is an important feature of motor control for much of the body, it has a more limited role in oral motor control.[7] However, while differences in effector control or composition could lead to differences in the *level of variability* (magnitude or structure) between oral and non-oral or among oral effectors, the finding of *task-dependent changing variability with practice* should hold because it relies on the idea that change in behavioral output arises from emerging coordination between multiple components of a control system, regardless of the identities of those

[7] It has been observed as chewing side preference, correlated with hand, foot, ear and eye side preference (Arslan, Ínal, Demir, Ölmez, & Karaduman, 2017; Nissan, Gross, Shifman, Tzadok, & Assif, 2004). Since most speech and swallow tasks entail grossly symmetric oral function, oral motor laterality was not a focus of this investigation.

components. Likewise, the idea that immediate change in structure of force output may predict future learning is not effector-specific. The physical characteristics of effectors are unlikely to be the primary driver of the kinds of changes seen in the learning studies reviewed thus far, because in each the structure of motor output changed with relatively brief, low-force training less than that ordinarily recommended to drive processes such as muscle hypertrophy (American College of Sports Medicine, 2009). As lips and tongue both contribute to normal speech and swallow functions, and their control may be jointly or individually impaired in clinical populations, both effectors were included in this work.

Support for the principle of reduced magnitude of variability with oral motor learning was reviewed previously (Grigos, 2009; Testa et al., 2011). A few studies have compared pursuit tracking in oral and non-oral effectors, finding more variable performance in the former (Bronson-Lowe, Loucks, Ofori, & Sosnoff, 2013; Loucks, Ofori, & Sosnoff, 2012; Ofori, Loucks, & Sosnoff, 2012; van Steenberghe, Bonte, Schols, Jacobs, & Schotte, 1991), but very little evidence addresses intra-oral comparisons or how temporal structure of oral force changes with task or with learning.

McHenry et al. (1999) compared maximal strength, magnitude of variability (as “steadiness”), and other measures of force generation in the upper lip, lower lip, tongue and jaw in 10-member groups of healthy women aged 20-39 (mean 31.1 ± 5.1 years), 40-59 (mean 48.8 ± 6.9 years), 60-79 (mean 66.8 ± 4.5 years) and 80-100 (mean 85.7 ± 4.69 years). Steadiness was measured in a pursuit tracking task with constant targets matched over ten five-second trials per effector. It was calculated as criteria percentage (CP), measuring the percentage of time during which force was within 10% of the target after first attainment of 90% of the target force. The analyzed level of target force was 0.5 N (an absolute rather than relative target, in contrast to most recent pursuit tracking studies). ANOVA was performed for each effector separately and found the effect of age group on CP not significant. Effector differences and age x effector interaction were not tested.

Visual inspection of their Figure 3 suggests the possibility of an age x effector interaction, as steadiness appears greater for the lips than for the tongue and jaw in the three younger groups and this difference appears smaller in the oldest group. Comparison of these results to other pursuit tracking tasks reviewed in this work is difficult, however, because the use of an absolute force target across age

groups and effectors means that the target varied widely in percentage of maximal force.[8] Differences between average strength of lower lip and tongue appear relatively small across the three younger age groups, meaning that the observed differences in steadiness vs. the absolute target would likely be preserved in a comparison of steadiness vs. the relative target. On the other hand, the lower lip appears stronger than the upper, meaning that it may have a greater advantage in steadiness vs. a relative target than is apparent compared to the absolute target. The upper lip's strength appears consistently lower than that of the tongue, which would tend to reduce its apparent advantage in steadiness if compared to the relative target. See "Declining strength in oral and manual effectors."

Holtrop, Loucks, Sosnoff, and Sutton (2014) examined submaximal force control of index finger flexion, lip pucker and anterior tongue elevation in healthy younger (12 of 18 female, mean age 22.6 ± 2.0 years) vs. older (8 of 14 female, mean age 67 ± 4.5 years) adults, handedness unspecified, in a pursuit tracking task using a constant target. Targets were presented at 10% and 20% MVC for 25 seconds. Participants were asked to complete three trials per effector at each force level during a single session. Force output was recorded at 100 Hz, with the initial five seconds cropped, leaving time series of length 20 s x 100/s = 2000 for analysis. Force variability was analyzed for both its magnitude (coefficient of variation, CV) and structure (approximate entropy, ApEn).

Findings included significant interactions of age group and effector for both measures. ApEn decreased significantly for the finger and lip, but not for the tongue, in older vs. younger adults. CV was higher in older adults for each effector. These age-related changes occurred despite the absence of significant difference in MVC between the younger and older adults. The authors did not report the significance of effector differences within age group, although it appears from their Table 1 that participants' force had lower ApEn and higher CV for the tongue than the lips in both age groups (across force levels, and with unknown significance). That is, force output was more predictable or regular and had greater magnitude of variability for the tongue. The effector differences were attributed to differing physical composition and habitual task demands of oral vs. non-oral effectors; because the focus was on age differences, the discussion did not touch on intra-oral comparisons.

---

[8] Based on their Figure 4, 0.5 N appears to range from approximately one-twenty-fourth (4%) of MVF for the lower lip of the average participant in the second-oldest group to approximately one-sixth (17%) of MVF for the upper lip of the average participant in the oldest group.

This work adds to this literature by comparing oral effectors' force variability in both steady and varying continuous submaximal force tasks and examining the interaction of effector and baseline force structure in prediction of learning. Based on interpretation of data in Holtrop et al. (2014) and McHenry et al. (1999), the tongue is expected to produce lower-entropy force than the lip, at least for the constant task.

Generalization to healthy aging. The first use of the dynamical systems framework to explain changes in healthy aging was the loss of complexity hypothesis (Lipsitz & Goldberger, 1992). Lipsitz and Goldberger suggested that aging entailed decreasingly complex dynamics in all physiologic systems, due to loss of or detrimental changes in either the systems' components or their nonlinear coupling, leading to reduced adaptation to stress. They reviewed data on neuroendocrine and cardiovascular function in aging (e.g. reduced entropy of heart rate and blood pressure variability) to argue for their theory, but did not apply it to voluntary motor tasks.

Reviewing data on motor control (locomotion and bimanual finger movement) and endocrine system function, Vaillancourt and Newell (2002) argued for a modified, bidirectional complexity hypothesis of both aging and disease. That is, what is lost is not complexity itself but rather the capacity to adjust output force structure to task demands. A task demanding constant output, for example, can be said to have a fixed-point (zero-dimension) attractor in state space. To produce such output, dominant (high-amplitude) rhythms must be dampened by the addition of multiple other rhythmic components. Since the damping is not perfect, actual output is highly complex and entropy measurements will be relatively high. On the other hand, a task demanding oscillatory output has a one-dimensional limit cycle attractor. To produce this output, the target rhythm must be enhanced and competing rhythms suppressed, which will produce relatively simple, low-entropy force structure (though of higher complexity than the attractor itself, since the competing rhythms will not be perfectly suppressed). If older adults are less able to adapt the structure of their force output to task demand, they should be less successful at cancelling out high-amplitude rhythms, yielding lower-entropy force for a constant target than younger adults, and less successful at emphasizing a single dominant rhythm, yielding higher-entropy oscillatory force than younger adults.

This prediction has been supported in multiple studies, of which three examples using the pursuit tracking task in healthy adults are discussed below. These studies examined a variety of isometric index finger force contours at low to moderate forces.

Vaillancourt and Newell (2003) investigated the effects of task demand on age-related changes in force complexity. They examined time and frequency structure of force output in pursuit tracking tasks using constant and 1-Hz sine targets by adults in three age groups each evenly split by gender: young (mean age 22.1± 1 years), old (mean age 67 ± 2 years), and older-old (mean age 82 ± 5 years). Each target was presented in two consecutive trials at 5%, 10%, 20%, and 40% MVC for 25 seconds (force and target order randomized) during a single session. Force output was recorded at 100 Hz, with the initial five seconds cropped, leaving time series of length 20 s x 100/s = 2000 for analysis. Analyses of variability structure included approximate entropy (ApEn), detrended fluctuation analysis (DFA), spectral slope and degrees of freedom (SS and SDF), and proportion of power (PoP) in three frequency bands (0-4, 4-8, 8-12 Hz).

Results showed that advancing age was associated with ApEn, SS, and SDF decreasing for the constant task and increasing for the sine task, and with the DFA scaling index α increasing for the constant task and decreasing for the sine task. PoP analysis, run separately for each task, showed an age group by frequency bin interaction due to age group differences in the 0-4 Hz band: while all groups' peak power occurred at a similar frequency (~1 Hz), power increased significantly at that frequency with each difference in age (older-old > old > young).

The authors interpreted their findings as showing that with advancing age, force became less complex for the constant task and more so for the sine task, consistent with the idea that older adults have more difficulty either increasing or decreasing the dimension of their force output to meet task demands. They pointed out that the loss of complexity hypothesis may still hold over longer timescales, but at the scale at which external environmental or task demands operate, the bidirectional loss of adaptability of force complexity is more relevant. Finally, the concentration of both power and age-related change in the 0-4 Hz frequency bin suggests that the observed effects of age were due to changes in sensorimotor processing rather than physiological tremor.

Sosnoff, Vaillancourt, and Newell (2004) examined electromyographic (EMG) frequency structure during pursuit tracking tasks, to evaluate effects of age and task demand on EMG oscillations at multiple time scales, possibly indicating changes in central firing synchronizing the collective activity of pools of motor units. Healthy young (6/15 women, mean age 24.9 ± 3.8 years), old (7/15 women, 67.8 ± 4.1 years), and older-old (8/15 women, 79.7 ± 4.2 years) adults were asked to use dominant index finger abduction to match 1-, 2-, 3-, and 4-Hz sinusoidal targets each centered at 5% and 25% MVC with a range of ±5% MVC. Participants completed three twenty-five-second trials per force x frequency

condition during a single session, with order of force and frequency randomized. Force output was recorded at 100 Hz, with the initial four and final one seconds cropped, leaving time series of length 20 s x 100/s = 2000 for root mean square error (RMSE) and modal force frequency calculations. Intramuscular EMG during trials was recorded from the active finger's first dorsal interosseous muscle and processed to enhance low-frequency spectral power. Modal frequency (frequency with greatest power density in a spectrum) was determined for both force output and EMG, used to assess how modal force frequency related to both target frequencies (1-4 Hz) and modal EMG frequency. Proportion of EMG power (PoP_{EMG}) was examined in four unequal frequency bands (0-5, 5-15, 15-35 and 35-50 Hz, based on the modal analysis and previous work). Coherence analysis (maximal coherence and, when coherence was significant over five adjacent frequency bins, phase) between force output and EMG signals was used to assess coupling between the two. Mixed model ANOVA (age group x target frequency x target force level) was performed for each dependent variable. K-means cluster analysis was used to evaluate the relationship of EMG and force modal frequencies separately for each force modal frequency[9] and age group.

Findings included an age x target frequency interaction in the force output data: young adult participants produced force with a modal frequency near the target frequency for all four targets, while modal frequency was significantly lower for the older-old adults for the 3-Hz target and both old and older-old adults for the 4-Hz target. Young adult participants' RMSE was lower than both older age groups' RMSE only for the 1-Hz target. Descriptive/visual analysis of a sample young adult participant's normalized EMG spectra by task suggested that the 1-Hz target elicited a broad spectrum with several dominant peaks, while all higher-frequency targets elicited a single dominant peak near the target frequency. This observation was supported by the k-means analysis, which sorted young adult participants' modal EMG frequencies into three clusters at the lowest modal force frequency (0.78 Hz), but only single clusters (all with mean values < 5) for all higher modal force frequencies. In both older age groups, the presence of multiple clusters of modal EMG frequencies persisted at higher modal force frequencies. Consistent with the k-means analysis, PoP_{EMG} analysis found that younger adult participants

[9] Possible force modal frequency values occurred in increments of 0.78 Hz rather than perfectly overlapping the target frequencies due to the bin size attained in the Fourier transformation of the time series to the frequency domain. This analysis thus examined the relation of modal EMG frequency to the modal frequency of force actually produced rather than to target frequency.

had greater relative power in the 0-5 Hz band than the older age groups, but only for the 3- and 4-Hz targets. Across age groups, relative power increased in the 5-15 Hz band with increasing target frequency. In the 15-35 Hz band, the young group had less relative power than both older groups for the 3-Hz target at both force levels and the 4-Hz target at the higher force level. In the 35-50 Hz band, young adult participants had less relative power than both older groups for the 3- and 4-Hz targets. EMG/force coherence was greater for young adult participants than for both older groups, with the difference more pronounced at the higher target force level, with the exception of no age effect for the 4-Hz target at the lower force level. The phase relationship between EMG and force signals was negative, meaning force oscillations followed EMG oscillations. Age group effects on phase were significant only in two target force x target frequency conditions: the older-old participants had a smaller phase difference than the young participants at 5% MVC for the 1-Hz target, but a greater phase difference at 25% MVC for the 2-Hz target.

The presence of an age effect on EMG-force coherence for almost all target force x frequency conditions, coupled with the absence of an age effect on EMG-force phase difference for almost all conditions, was interpreted as suggesting that the age-related changes seen in motor performance were due to central factors (stronger coupling between task-specific oscillatory drive and motor unit pools) rather than peripheral changes in duration of transition from electrical signal to mechanical force. Results were interpreted as support for age- and task-mediated changes in EMG frequency structure such that older adult participants were less able to shift EMG power across frequency bands to meet task demands (e.g. emphasizing a dominant target frequency to synchronize motor unit output for a relatively smooth, low-entropy force oscillation, vs. desynchronizing motor units' output by more broadly distributing power across frequencies to produce complex, high-entropy force for constant-target tasks), suggesting impaired coordination of excitation and inhibition of multiple neural oscillators.

Recall that Sosnoff and Voudrie (2009) examined whether older adults' relatively reduced adaptability of force structure to task demands could be modified with practice. Briefly, young and old adults were asked to use index finger abduction to match constant and 1-Hz sine targets over five days' practice. Though both age groups demonstrated reduced magnitude of variability with practice and evolution of force structure in the direction demanded by the task (ApEn increased for the constant task and decreased for the sine task), the old adults changed ApEn within a more restricted range than the younger adults and showed greater magnitude of variability throughout the study. The authors interpreted their findings as support for the hypothesis that loss of adaptability (rather than loss of

complexity) is characteristic of aging and is not due to lack of familiarity with a task. They speculated that both the changes in the force structure observed with practice and the loss of adaptability seen with aging could be due to changes in motor unit synchrony.

Neither complexity change hypothesis has been directly tested in the oral motor literature. However, there is indirect evidence of altered force structure for older adults in oral effectors. Marzullo et al. (2010) contrasted normalized EMG lip power during perturbed and nonperturbed word repetitions. While both young and old adults produced high-frequency EMG power, only young adults were able to increase power from 30-50 Hz to reduce lip displacement during perturbation trials. Older adults were unable to make this shift and consequently showed greater lip displacement following perturbation. These results can be interpreted as a decrease in adaptability with aging.

If immediate within-task adaptability of force temporal/frequency structure predicts motor learning, but older adults are less able to adapt force structure to task demand, the strength of the predictive relationship may vary by age. Consequently, both younger and older adults were included in this investigation to assess this possibility.

The role of strength. It has been suggested that the relative magnitude and the regularity of variability may be inversely related to muscle strength (Hamilton, Jones, & Wolpert, 2004; Sosnoff & Newell, 2006b) and consequently that changes in variability seen with aging may be explained by declining strength. The mechanisms potentially responsible are not clear and may include various physiological changes or the inherently greater difficulty of producing low relative magnitude of variability at extremely low absolute levels of force (Sosnoff, Valantine, & Newell, 2006). The latter factor would disproportionately affect weaker participants, for whom a given relative force target equates to a lower absolute force. Other work has found age-related changes in force variability in the absence of significant declines in maximal strength (Bronson-Lowe et al., 2013; Sosnoff & Voudrie, 2009; Vaillancourt et al., 2003).

If strength does have a role in older adults' changing variability of force, its effects may differ across effectors. Age-related declines in strength are well established for the tongue (Adams et al., 2013; Crow & Ship, 1996; Nicosia et al., 2000), but findings for other oral effectors have been less consistent (Clark & Solomon, 2012; Fogel & Stranc, 1984; McHenry, Minton, Hartley, Calhoun, & Barlow, 1999; Wohlert & Smith, 1998). The two studies to measure strength of both lip and tongue were focused on strength changes in aging and did not compare the effectors to each other (Clark & Solomon, 2012; McHenry et al., 1999), but some suggestions can be made based on their reported data.

Briefly, Clark and Solomon (2012) evaluated tongue strength (anterior and posterior dorsum elevation, protrusion, and lateralization) and lip and cheek compression strength in healthy young (43/68 women, mean age 22.9 ± 3.5 years), middle-aged (25/60 women, 44.7 ± 8.8 years) and old (15/43 women, 70.8 ± 7.1 years) adults. Age and effector interacted significantly: strength was decreased in the older group compared to the middle group for all four tongue measures and compared to the younger group for protrusion and lateralization, but there was no effect of age group on lip compression strength. Effectors were not directly compared, but per their Table 4, lip compression strength was lower than all four measures of tongue strength across all age groups. If strength has the proposed role in variability, tongue force would have lower relative magnitude of variability and higher entropy (opposite to the untested differences reported in Holtrop et al., 2014; their lip task involved puckering rather than compression, but this should only have exacerbated strength differences).

McHenry et al. (1999; see "Generalization to oral effectors") compared maximal strength and other measures of force generation in the upper lip, lower lip, tongue and jaw in 10-member groups of healthy women aged 20-39 (mean age 31.1 ± 5.1 years), 40-59 (48.8 ± 6.9 years), 60-79 (66.8 ± 4.5 years) and 80-100 (85.7 ± 4.69 years). The upper lip and tongue showed trends for decreasing strength with age ($p < 0.107$ and $p < 0.045$ respectively, both nonsignificant after Bonferroni correction). Effector differences were not tested. Visual inspection of their Figure 4 shows somewhat lower strength for the lips and tongue in the oldest group, lower strength for the upper lip than for the tongue across age groups, and the possibility of an age x effector interaction due to tongue strength declining more than upper lip strength in the oldest group.

Recall that McHenry et al. (1999) also measured variability (as "steadiness") during a very short constant-target pursuit tracking task, with the target set at an absolute level of 0.5 N for all age groups and effectors regardless of maximal voluntary force (MVF). Extrapolating from Figures 3 and 4 to estimate differences in steadiness with the target expressed relative to MVF (see "Generalization to oral effectors"), the lower lip may have greater steadiness/lower magnitude of variability than both the tongue and the upper lip, despite having grossly similar strength to the former. If correct, this raises the possibility that strength has a stronger influence on variability when effector composition is similar (upper vs. lower lip) than when it differs (lip vs. tongue).

In sum, variability differences have been observed across age groups, and appear to exist in reported data on different oral effectors but have not been statistically evaluated. Oral effector strength differences also appear, untested, in reported data, but have an unclear relationship to variability

differences. Given the lack of clarity and the possibility of an age x effector interaction in strength, the interaction term was included in analyses in this work.

Since the role of strength loss in age-related variability change was not the primary focus of this work, participants were recruited only into older and younger adult groups, rather than further stratifying the older adults into groups who would and would not be expected to show strength loss. Because sex differences in strength in the oral effectors were not observed to change direction with advancing age in the single study to evaluate the question (Clark & Solomon, 2012), sex was not included as a factor in this investigation.

Generalization to functional tasks. The pursuit tracking studies reviewed previously have used constant and sinusoidal force targets to elicit force whose structure can be clearly differentiated by task. These targets were also used in this work for the same reason; however, a visual pursuit tracking task with either contour does not resemble any common oral motor task. Visual feedback is seldom relevant for either speech or swallow, and neither requires sustained performance of either constant or perfectly periodic force. To partially address the second of these limitations, a third contour was chosen to balance the constraints of improved similarity to speech with the goal of eliciting force whose structure would be clearly differentiable by task.

Eliciting force more complex (higher-entropy) than that produced for a sine target required a target with multiple component frequencies. Plausible learnability and the need to elicit force less complex (lower-entropy) than that produced in the effort to match a constant target suggested the number of additional frequency components beyond the fundamental should be small. A model of speech motor control that views speech articulation as movement at a fast timescale (consonant gestures) superimposed on movement at a slower timescale (vowels) suggested that the target's structure should entail reduced amplitude for the higher frequencies (Öhman, 1966)[10].

A waveform developed by Stanley and Franks (1990; hereafter, "multicosine") fit these constraints. It consisted of a fundamental dominant frequency with two superimposed, lower-amplitude

---

[10] Öhman's model was developed using VCV sequences in which all C were stop consonants. He used vowel formant transitions to argue that intervocalic coarticulation took place across the intervening consonant and thus "stop-consonant gestures are actually superimposed on a context-dependent vowel substrate that is present during all of the consonantal gesture" (p. 165). The multicosine waveform was chosen to mimic only this basic idea of multiple-timescale superimposition, not to act as a realistic model of speech. Using a more complex, realistic model would have risked reduced differentiation of force complexity from that elicited by the constant target.

multiples of the fundamental (parameterization details in Methods). The entropy of the target itself was higher than that of either the constant or sine targets, but the expected entropy of force output during attempts to match it was intermediate between the others, as the presence of multiple component frequencies was expected to drive entropy higher than that seen for a sine target, while the preservation of regular, predictable structure and the need to emphasize a small number of dominant frequencies was expected to keep entropy lower than that seen for a constant target.

Within-session practice vs. retained and transferred learning: commentary on methods. The studies of practice reviewed earlier shared two important characteristics which may have biased change in their participants towards temporary effects only. A 100% knowledge-of-results feedback schedule (Sosnoff & Voudrie, 2009; Deutsch & Newell, 2004; Newell et al., 2003) may support improved performance during practice sessions, but suppress learning as assessed in retention and transfer trials without knowledge of results (Salmoni, Schmidt, & Walter, 1984). Blocked practice, in which a single skill is practiced repeatedly (Deutsch & Newell, 2004; Newell et al., 2003; Sosnoff & Voudrie, 2009; Wijnants et al., 2009), may enhance performance during acquisition, but depress learning outcomes compared to random practice of several tasks (contextual interference, Hall & Magill, 1995; Magill & Hall, 1990).

Because of its clinical importance, this investigation assessed learning as retention and transfer the day after practice, rather than performance change during practice. Tasks were structured to maximize retention and transfer even at the expense of performance during practice sessions. Learning enhancement techniques included random vs. blocked practice, reduced frequency of knowledge of results after initial task familiarization, and integrated variable priority training, in which participants complete the whole task on each trial but emphasize different aspects of performance (force amplitude vs. timing) on different trials (Fabiani et al., 1989; Gopher, Weil, & Siegel, 1989). This training approach has been found to hasten learning and improve task mastery (Boot et al., 2010; Prakash et al., 2012).

Specific Aims and Hypotheses

This investigation collected original data for an investigation with three primary aims related to motor learning and variability in healthy younger and older adults. See Table 1.

Table 1

Specific Aims and Hypotheses

Specific aims	Hypotheses
1. Assess applicability of previous findings on effects of age and task to oral effectors.	1a.* Older adults' force structure will differ task-dependently from younger adults' (lower entropy and a greater proportion of low-frequency power when the task demands high entropy and reduced low-frequency power, and vice versa). 1b. Adaptability (immediate): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing trial 2 to trial 1 on day 1 within each effector x task combination. 1c. Adaptability (after practice): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing day 1 trial 1 to day 3 retention trial 1 within each effector x task combination. 1d. Older adults will show less reduction in error relative to baseline on retention and transfer trials after two days' practice than younger adults. 1e.* Structure of force will differ by task. The constant task will elicit the highest entropy, lowest proportion of low-frequency power, and greatest proportion of higher-frequency power. The sine task will elicit the lowest entropy, greatest proportion of low-frequency power, and lowest proportion of higher-frequency power. The multicosine task will be intermediate.
2. Assess differences in motor variability between oral effectors.	2a.* The tongue will produce less complex force than the lip (lower-entropy, greater dominance of low-frequency power). 2b.* The effects of age group and effector on entropy will interact.
3. Assess utility of baseline performance measures in predicting *de novo* learning of fine-force pursuit tracking tasks in oral effectors.	3. (a) Error and a measure of temporal structure, (b) higher maximal force entropy or (c) greater adaptability of entropy at baseline, will predict retention and transfer in pursuit tracking tasks after controlling for age group, effector and task.

* These hypothesis are expected to hold before and after practice.

Aim 3 is the main interest, but aims 1 and 2 will be addressed first in Results to build the case for the importance of age group, effector, and task in this data set prior to their use as predictors for aim 3.

Methods

Participants

Younger and older right-handed adults were recruited. Younger adults (10/20 women) ranged in age from 18-28 years (M ± SD, 22.887 ± 2.843 years). Older adults (10/21 women) ranged in age from 71-79 years (M ± SD, 75.322 ± 2.691 years). Younger adults had 16.75 ± 2.45 years of education; older adults, 15.67 ± 2.89 years. This difference was not significant (two-tailed independent samples *t-test* with equal variances assumed, $t(39) = -1.293$, $p = 0.204$). Other criteria included:

1. No reported history of neurological, psychiatric or speech-language disorders or motor impairments. Investigator (certified speech-language pathologist with six years' experience with geriatric clinical populations) observed for signs of speech or other motor disorders during conversation, and rejected one potential participant on this basis (see Results: Missing Data), but no formal speech, motor or cranial nerve screening was performed.
2. No signs, symptoms or diagnosis of temporomandibular joint disorder in the past 5 years
3. No use of anti-depressant or anticholinergic drugs (e.g. antihistamines, tricyclic antidepressants, antipsychotics, antiemetics), anticonvulsants, or antianxiety drugs, any of which may affect neural function
4. No use of dentures
5. Central and lateral incisors present and functional
6. Functional vision (natural or corrected) and cognitive-communication skills to understand the chart displays used for performance of the experimental task
7. For the older adult participants, a score of at least 27 on the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), or at least 26 if the potential participant had no more than an eighth-grade education (O'Bryant et al., 2008; Kukull et al., 1994; Bravo & Hébert, 1997).
8. No oral pain, infection or lesion (screened prior to inclusion and daily before testing)

All provided informed consent. Experimental procedures were approved by the Institutional Review Board at the University of Illinois prior to participant recruitment.

Handedness preference (direction and strength) was measured using the handedness subscale of the Lateral Preference Inventory (Coren, 1993). To be included, participants had to score at least 1, i.e. prefer to use their right hand for at least one more activity than their left hand. Handedness criteria

were included in this oral motor control study due to handedness-group differences on multiple visual, auditory and motor tasks (Gabbard, 1998) and to control for any potential subtle oral laterality (Arslan et al., 2017; Nissan et al., 2004).

Functional (corrected or uncorrected) vision and chart understanding was assessed with a custom screen: During the informed consent process, the potential participant was shown a simulated task (replay of previously collected data), and the chart display was explained including identification of target and data signals and relationship of pressure on the transducer to behavior of the data line on the chart. After a short delay with other screens serving as distractions, the participant was shown two simulated tasks and asked to identify in each case the target line, data line, a point at which the person providing the data had needed to press more gently to match the target, and a point at which the person had needed to press harder. Each question was worth one point, with 7 out of 8 required to pass. The only question permitted to be missed was the final one, 'needed to press harder' in the constant case, because the visible difference between target and data was very small. Appendix B includes screenshots of the tested cases. Use of glasses, contacts or hearing aids was noted daily, but the functional vision screen was completed only once.

Instrumentation

Stimulus presentation. Stimuli were presented on a Lenovo P500 laptop anti-glare screen, 15.6"/39.6 cm diagonal (16:9 aspect ratio: 1366 x 768 pixels), ~45 cm from the participants' eyes; chair height, screen height and screen angle were adjusted for comfort and best view. Participants were encouraged to wear glasses if they normally did so for reading or computer work.

Targets were presented as a thick red line (R255 G000 B010) on a black background, with the participant's force represented as a thinner blue line (R051 G153 B255) that remained visible when it crossed the target line. The plot area of the display was 1246 pixels wide by 207 pixels high, yielding a horizontal gain of approximately 78 pixels/sec for a 16-second task. Vertical scale of the chart ranged from 0% to 30% of a participant's effector-specific maximal voluntary force (MVF), yielding a vertical gain of 0.145%MVF/pixel. Controlling vertical gain in this way rather than Newtons/pixel allowed for visual angle to remain constant across tasks with differing dynamic ranges and across effectors and participants with differing maximal strengths.

Signal collection. Separate transducers, shown in Appendix C, were used to collect lip and tongue force signals. The transducers' sensitivity was under 0.01 Newton.

An adjustable chin rest was used to minimize extraneous movement. The lip force transducer was fixed using a nonslip-mounted vise on the table. Its two horizontal prongs were placed inside the commissures of the lips, to the depth of the prongs' rounded divots. Pressure was exerted medially against the prongs by pursing the lips. The tongue force transducer was held bimanually with elbows resting on the table, stabilized within the mouth by indentations for the upper and lower teeth to minimize jaw motion. Pressure was exerted by elevating the anterior tongue against the lower surface of the curved metal tab protruding from the transducer's end.

Transducer signals for all conditions, sampled at 100 Hz, were routed through a bridge amplifier (Biocommunication Electronics LLC, Madison, WI). The signals were sent to an NI-USB-9234 signal acquisition A/D board and then to custom Labview programs to provide target signals and feedback (National Instruments, Austin, TX).

Any trial interrupted by a perturbation (cough, jostling of chair/table/transducer setup, etc.) was restarted; that is, all recorded and analyzed trials were free of perturbation.

Calibration. Each transducer was calibrated using known test masses (0 – 450 g in increments of 50 g, not all tested during every calibration; brass scale weights, Ohaus, Parsippany, NJ). Calibration before every experimental session and occasionally on non-experimental days began with testing at 0 g to check for and correct baseline drift. Multiple nonzero masses were also tested before the majority of experimental sessions (60% for the lip transducer, 65% for the tongue transducer) and occasionally on non-experimental days. Baseline was also retested and corrected to zero immediately before each participant began the day's trials, and at the beginning of each set of five trials.

Table 2 lists criteria for accepting a calibration trial, developed empirically based on early calibration sessions prior to experimental data collection. Two criteria were chosen: absolute value of slope close to zero to ensure values were not drifting appreciably during a trial, and mean close to zero when test mass was 0 g, to ensure baseline was close to a known correct value. Figure 2 and Table 3 show the actual frequency distribution of unique masses tested per session. For both transducers, 86% of multi-mass testing sessions had at least 4 test masses.

Table 2

Required Criteria for an Acceptable Calibration Trial and Metrics of Actual Trials

	Required	Actual (M ± SD, # trials)	
		Lip	Tongue
|Linear slope|	< 10^{-4} volts/sec	6.0 x 10^{-7} ± 1.2 x 10^{-6} volts/sec (1079 trials)	4.5 x 10^{-6} ± 7.2 x 10^{-6} volts/sec (947 trials)
Mean for trials at 0g	≤ 0.003 volts	-0.00025 ± 0.00096 volts (346 trials)	-0.00013 ± 0.0015 volts (224 trials)
Visual review	no obvious perturbation	achieved for all accepted trials	

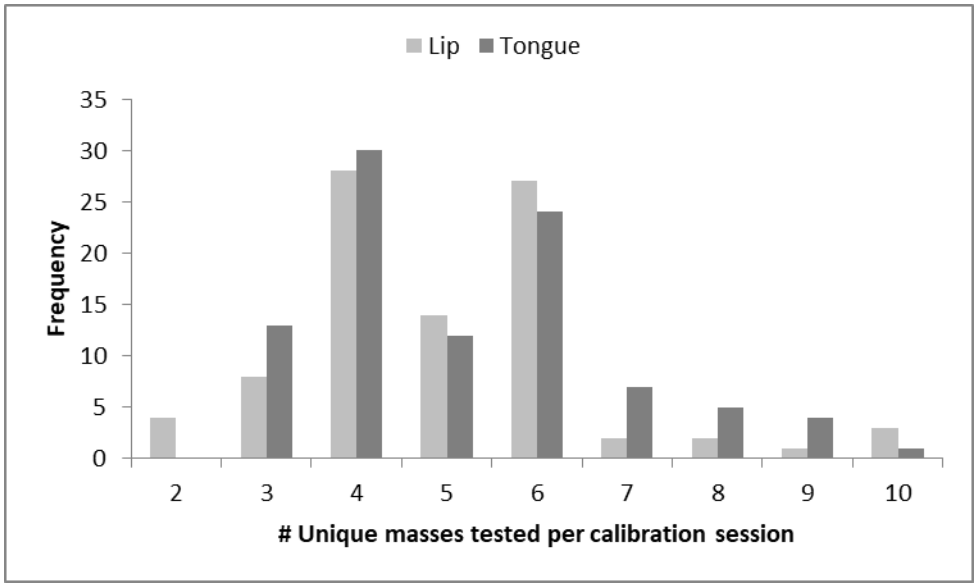

Figure 2. Frequency distribution of unique masses tested per calibration session.

Table 3

Frequency Distribution of Unique Masses Tested per Session

	Lip		Tongue	
Unique masses tested per calibration session	Frequency (N, %)	Cumulative frequency (% ≥)	Frequency (N, %)	Cumulative frequency (% ≥)
2	4 (4.5%)	100%	0 (0%)	100%
3	8 (9.0%)	95%	13 (14%)	100%
4	28 (31%)	87%	30 (31%)	86%
5	14 (16%)	55%	12 (13%)	55%
6	27 (30%)	39%	24 (25%)	43%
7	2 (2.2%)	9.0%	7 (7.3%)	18%
8	2 (2.2%)	6.7%	5 (5.2%)	10%
9	1 (1.1%)	4.5%	4 (4.2%)	5.2%
10	3 (3.4%)	3.4%	1 (1.0%)	1.0%

Linear regression of signal vs. test mass using all acceptable trials was performed to develop conversion equations for transducer data from volts to Newtons. Proportion of variance accounted for by a linear fit to the data was high for both transducers: $r^2_{lip} = 0.81$, $r^2_{tongue} = 0.94$.

Infection prevention. Following every experimental session, transducers were soaked for eight minutes in Revital-Ox RESERT XL (Accelerated Hydrogen Peroxide) High Level Disinfectant (Steris, Mentor, OH), then washed for one minute in tap water per manufacturer directions. VERIFY Chemical Monitoring Strips for Resert Solutions (Steris) were used once per experimental day to verify activity of the disinfectant. New disinfectant was used every 21 days (per manufacturer guidelines) or when a test strip indicated reduced activity, whichever came first. Transducers air-dried completely prior to next use. Self-adhesive textured plastic wrap was used to cover the portions of the transducer contacting oral mucosa, replaced for every session. Computer and desk were cleaned with alcohol wipes before and after each session.

Tasks

Each task was done using the lips and the tongue separately. All were demonstrated using previously recorded data to familiarize the participant with the visual display, target contours and feedback function. (See Appendix C.) Instructions were designed to elicit an external focus of attention for improved performance and retention (Freedman, Maas, Caligiuri, Wulf, & Robin, 2007; Wulf, Höß, & Prinz, 1998; Kal, van der Kamp, & Houdijk, 2013).

Maximal voluntary force. Participants were instructed to produce the highest resultant force (Barlow & Burton, 1990) or maximal voluntary force (MVF) possible by elevating the tongue or pursing the lips (Barlow & Muller, 1991; Loucks et al., 2010). Three MVF trials were conducted with each effector, using visual feedback and verbal encouragement. Maximum force reached on each trial was marked on subsequent trials by a horizontal target line and participants were encouraged to exceed their previous performance. To avoid fatigue, participants were told they did not need to maintain the maximal force until trial completion, but could relax once they had made their best effort within a trial. See Figure 3.

Kamen (Kamen, 1983) found that maximal force developed within less than three seconds, but his participants were college-aged men and oral effectors were not tested. To allow for possibly slower force development, each MVF trial lasted six seconds, with sixty seconds between trials. The highest value of the three trials was used, so long as the lowest value was at least 90% of the highest; else, a fourth and final trial was completed and the highest value taken. This value was used to calculate target force levels for experimental tasks.

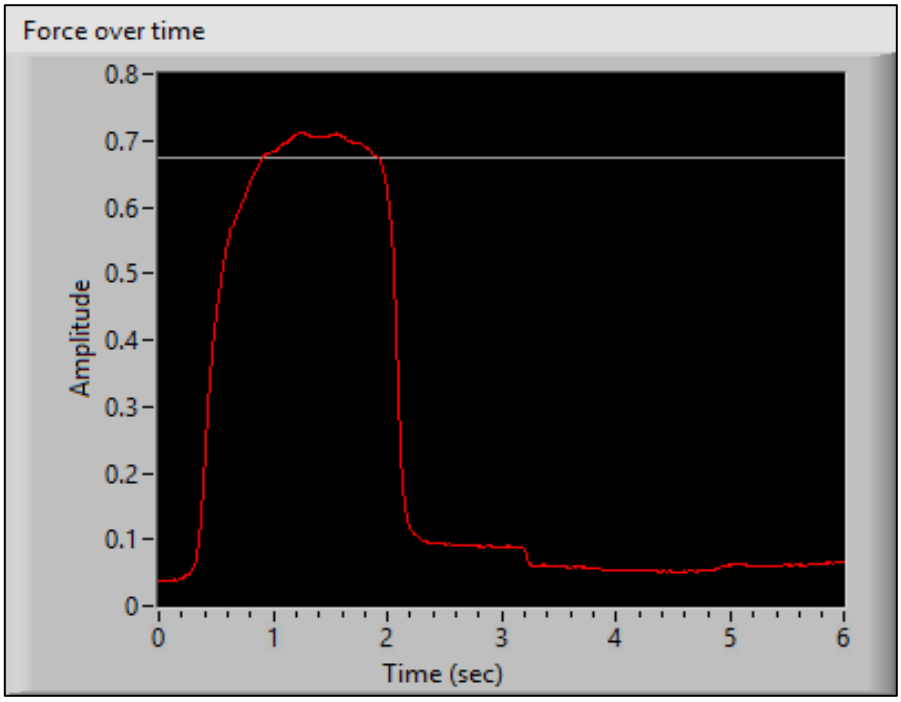

Figure 3. Maximal voluntary force visual feedback. Red line represents current attempt, updating in real time. Horizontal gray line represents maximum achieved on previous attempt.

Fine-force stimuli. Participants were asked to match three force contours: constant force, a sine wave, and a multicomponent cosine wave. Stimuli were centered at 15% of the effector's maximal voluntary force, ranging ±5% for the variable contours. This force level was chosen to avoid fatigue and to be approximately comparable to forces used for speech, which have been estimated to be 20% or less of maximal voluntary force for the lips (Goldberg, 2000; Muller, Milenkovic, & MacLeod, 1985). Swallow pressures have been estimated at 46-68% of maximal tongue pressure (Youmans, Youmans, & Stierwalt, 2009); no comparable level of force was used in this work due to the possibility of fatigue.

All tasks were performed with visual feedback. Both the target waveform and the participant's force were visible, to assess learning of the ability to adapt force complexity to visually specified task demand rather than the creation of an internal representation of a particular force profile. Trial length was chosen by balancing required N for nonlinear analyses with practical constraints; see Appendix D.

Multicomponent cosine (multicosine) waveform design. The multicosine stimulus is based upon the waveform used by Stanley and Franks (Stanley & Franks, 1990):

$$y(t) = \frac{A}{2} + Ccos(\omega t) + \frac{C}{2}cos(2\omega t) + \frac{C}{4}cos(4\omega t), \quad (1)$$

where $\omega = 2\pi f$, A/2 is used to adjust mean force and C to adjust amplitude. Let C = 1 such that the relative amplitudes are 1, 0.5 and 0.25 as frequency progressively doubles. On a log power/log-frequency graph (power proportional to amplitude squared), the slope of the line through those points is -2. (This is related to a signal with spectral slope β = -2, but no intermediate frequencies are included. It is called *three-point slope* here to differentiate it from the usual spectral slope.)

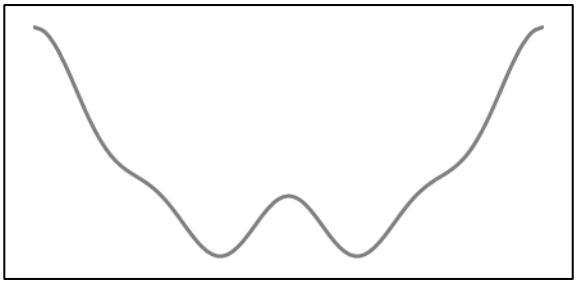

Figure 4. Multicosine waveform.

Frequency components for variable tasks. Frequency components met two constraints. First, overall difficulty: tasks had to be easy enough to permit immediate performance with an expectation of improvement upon two days' practice for both age groups, while challenging enough to allow room for improvement and variation across participants. The chosen fundamental frequency (0.75 Hz) is within a range that healthy young adults have learned to produce in a visuomotor tracking task with the lip with

as little as two trials (Moon, Zebrowski, Robin, & Folkins, 1993). The highest frequency in the multicosine stimulus, 3 Hz, is consistent with the speech rate of approximately three to four syllables/second reported for young and older adults (Ramig, 1983; Goozée, Stephenson, Murdoch, Darnell, & LaPointe, 2005), on the low end to allow for task unfamiliarity and older adults' previously reported difficulty producing higher-frequency force fluctuations (Sosnoff et al., 2004).

Second, targets were to be differentiated and ordered in complexity, both conceptually and per entropy measures, to increase from constant to sine to multicosine. The conceptually simplest target (constant force) has zero entropy per either measure. Two strategies were considered for the variable tasks: matching the multicosine task's highest frequency to the sine task's fundamental frequency, vs. matching their lowest (fundamental) frequencies. Table 4 compares the resulting targets. The match-lowest strategy differentiated the targets and ordered them as desired. The match-highest strategy did neither, because its multicosine signal's F0 value was enough lower than the sine's F0 to reduce its entropy to less than the sine signal's and to near zero. Consequently, the match-lowest target signals (right-hand column) were used.

Table 4

Target Series' Fuzzy Measure Entropy, Using N = 1100, Sampling Rate F_s = 100 Hz, m = 2, r = 0.2

	Strategy	
	Match highest: sine F0 = 1* Hz	Match lowest: sine F0 = 0.75† Hz
Signal	multicosine = 0.25, 0.5, 1 Hz	multicosine = 0.75, 1.5, 3 Hz
Constant	0	0
Sine	0.249	0.163
Multicosine	0.069	0.317

* Not 0.75 Hz: if multicosine's highest frequency were 0.75 Hz, its lowest-frequency component would be 0.1875 Hz, with no integer number of repeats in an integer number of seconds until 3 periods at 16 seconds, vs. 1 repeat in 4 seconds for 0.25 Hz.
† Not 1 Hz: the multicosine signal's highest-frequency component then would have been 4 Hz, crowding the limit of physiological feasibility even for younger adults.

Experiment Structure

Sessions occurred once per day over three continuous days, at a consistent time of day for each participant. Participants practiced each force contour fifteen times per effector on day 1 (3 tasks x 15 trials x 2 effectors = 90 trials total) and twenty times per effector on day 2 (3 x 20 x 2 = 120 trials total). Similar studies lasting five days asked participants to perform totals of 75 trials (Deutsch & Newell,

2004), 125 trials (Newell et al., 2003), discounting the sixth day with no visual traces) and 25 trials per task (Sosnoff & Voudrie, 2009), and these practice amounts were all sufficient to create performance change. This design achieved 35 trials per task x effector combination over two practice days, with retention and transfer assessed on the third. Since the eventual clinical goal is prediction of learning from a single session, a shorter time scale than five days' practice was preferred, but one day of practice (15 trials per task x effector combination) was judged potentially insufficient.

Because pilot data suggested rapid change over the first few trials, participants were allowed to see demonstrations of the tasks during the informed consent process but were not allowed to practice them until participating in the experiment. Figure 5 summarizes the experimental structure described in more detail in the following sections. For this investigation, only data from the initial trials on day 1 and initial trials of 'retention' and 'transfer to altered target force level' on day 3 were analyzed.

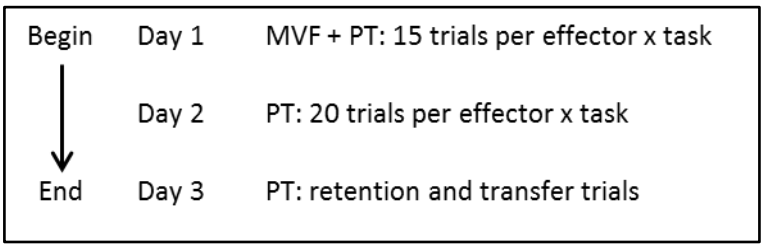

Figure 5. Summary experimental structure schematic. MVF = maximal voluntary force. PT = pursuit tracking.

Day 1. See Figure 6 for Day 1 experiment structure. Following screening and informed consent, maximal voluntary force was assessed for each effector, and all days' tasks were scaled to those MVF values. Order of effectors tested was determined randomly for the first participant in each age x sex subgroup, then alternated within those groups. Participants rested for one minute after each MVF trial.

Practice of the pursuit tracking tasks was structured to maximize familiarization with the force contours. Following MVF measurement and rest, participants completed three task blocks per effector (ordered as above), each block comprising five trials of each task. Tasks within a block were ordered to increase in complexity of the target signal: constant, sine, multicosine. All blocks for one effector were completed sequentially. Participants rested for at least ten seconds between each trial to reduce history effects on force output (Herzog, 2004) and to avoid fatigue.

For the variable force tasks, participants were guided to focus on matching the *amplitude* of force changes during the first block with each effector (five trials per task), their *timing* during the second block, and both amplitude and timing during the third block (integrated variable priority training;

see instructional scripts in Appendix C). The requested focus of the current trial was always visible on the monitor below the target/data plot. Any change in focus at the beginning of a five-trial set was announced in a text dialogue box requiring acknowledgment before data collection continued. The experimenter also verbally cued the change in focus when the text dialogue box appeared.

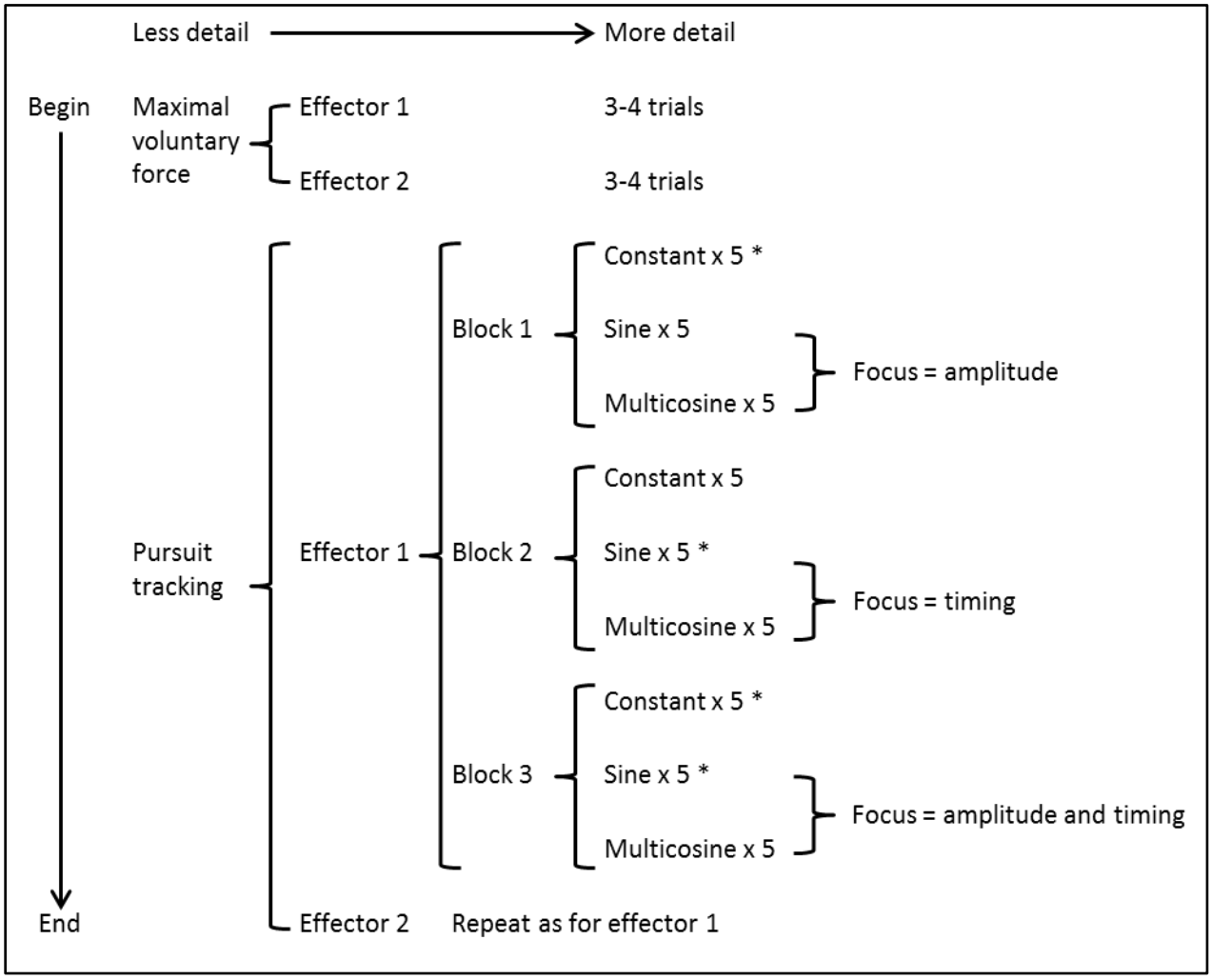

Figure 6. Day 1 experiment structure. Task order is top to bottom, with detail increasing to the right. Maximal voluntary force and pursuit tracking tasks were ordered by effector; order of effectors alternated within age x sex subgroups. For pursuit tracking tasks, within each effector, each of three blocks consisted of one five-trial set per task, with sets in a fixed order by task. The only possible focus for the constant task is "match amplitude," because the target does not vary over time. For variable tasks, focus varied by block. Each 5-trial set marked with * began with cues to attend to the new focus.

Knowledge of results was given after each trial (a graph of normalized root mean square error, NRMSE – see Figure 7 and the Measures section below) and each task (average NRMSE for the 5 trials).

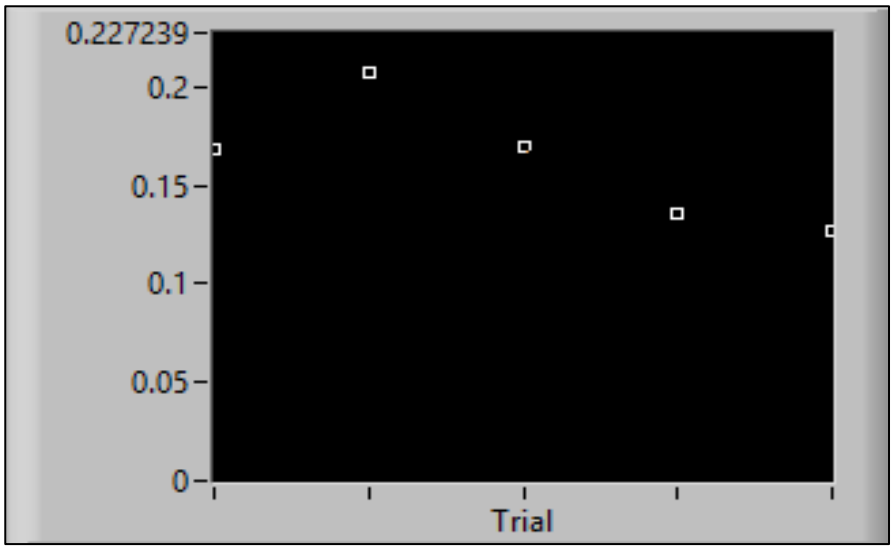

Figure 7. Screenshot of knowledge-of-results feedback graph shown to participants. See text for description of use.

These graphs were shown and explained to participants prior to the experiment and reviewed upon first appearance within the experiment. The investigator answered questions and verbally confirmed participant understanding of the feedback. Participants were told dot height represented "about how far away from the target you were overall during this trial, so the lower the dot, the closer you stayed to the target." Within each set of five trials of a task x effector combination, one new dot appeared after completion of each trial, and dots from previous trials remained visible; participants were encouraged to compare performance to previous trials in the set and to improve performance on the next trial in the set. The feedback graph was only visible after completion of a trial.

The investigator intermittently gave verbal knowledge-of-performance feedback after a trial (e.g. "you overshot," "your line went up too late"), in line with the current focus. A trial was restarted if invalid, e.g. interrupted by a cough, but had to be accepted or rejected before the investigator or participant could see knowledge-of-results feedback and was not redone based on that feedback.

Day 2. Day 2 was structured to maximize learning, using random practice with lower frequency of knowledge of results. The most effective frequency of this feedback may depend on task complexity (Schmidt, Young, Swinnen, & Shapiro, 1989) and random vs. blocked practice condition (Del Rey &

Shewokis, 1993). Since frequency of feedback is not a focus of this work, a consistent absolute rate of feedback after every 5 trials was used, which may be optimal (Schmidt, Lange, & Young, 1990).

Participants completed four task blocks per effector, each still comprising five trials per task (Figure 8). Effector x task x single-focus (amplitude or timing, 1 block each) combinations were randomly ordered, followed by effector x task x dual-focus combinations (2 blocks) randomly ordered. Participants saw knowledge-of-results feedback at the end of each 5-trial combination (graph showing all five trials' NRMSE values and average NRMSE across trials). The investigator gave no feedback until the completion of a 5-trial set and then focused on eliciting the participant's judgment of their own performance.

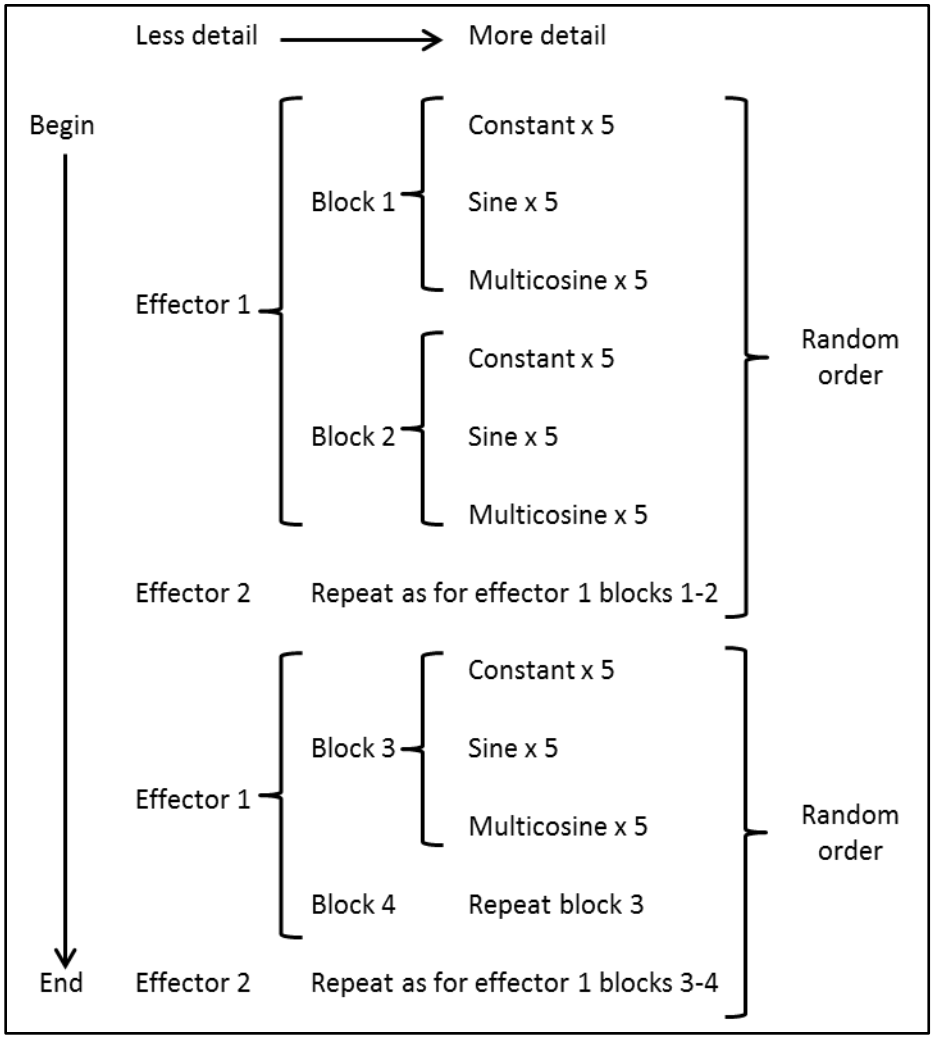

Figure 8. Day 2 experiment structure. Task order is top to bottom, with detail increasing to the right. Within each effector, each of four blocks comprised one five-trial set per task. For variable tasks, focus varied by block. Task order was randomized across effector x block (focus) x task combinations, done separately for blocks 1 & 2 (1 amplitude, 1 timing) vs. blocks 3 & 4 (both amplitude AND timing).

Day 3. Retention and transfer trials took place on Day 3 (Figure 9). Participants were instructed to stay as close as possible to the target (i.e. match both amplitude and timing) for all trials. Five retention trials of each task were completed first, with effector x task combinations randomly ordered. Transfer trials assessed transfer to slightly decreased then slightly increased target force levels: 10% and 20% MVF (two trials each), vs. the practiced value of 15%. Effector x task combinations were again randomly ordered, though within each combination the order of target force levels was not.

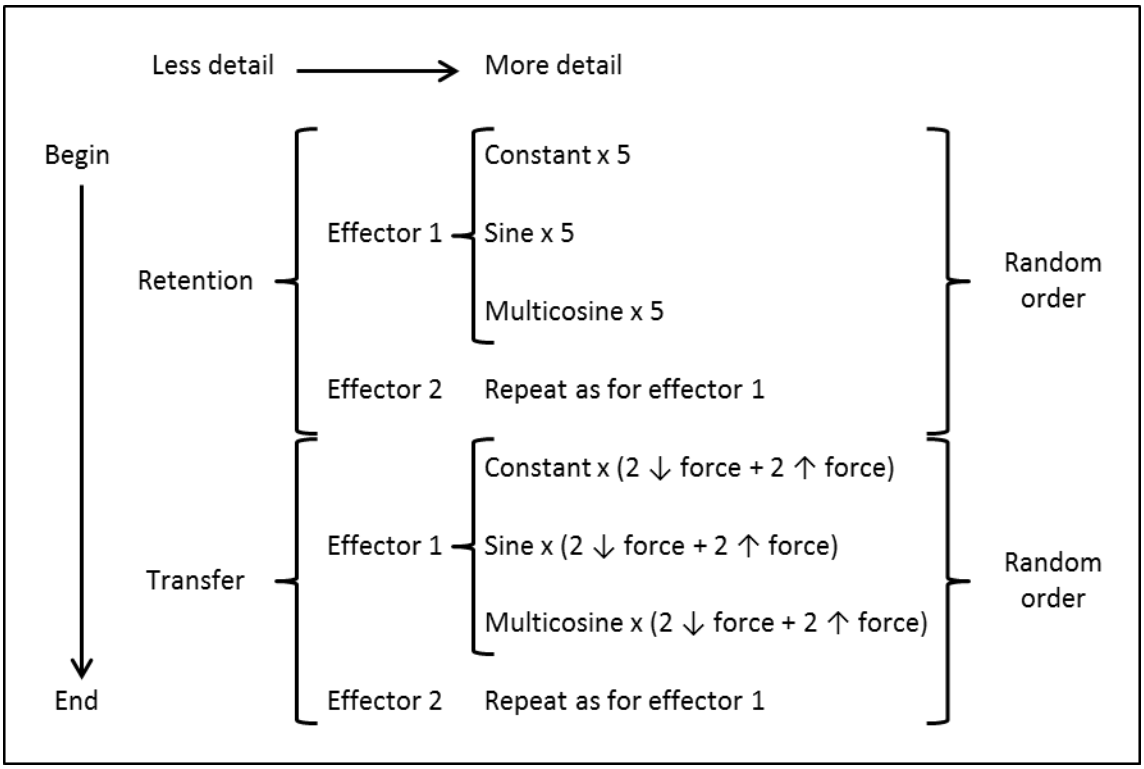

Figure 9. Day 3 experiment structure. Retention trials were completed first, followed by transfer trials. Within each effector, each task was tested in sets of consecutive trials: 5 for retention, 4 for transfer (2 each of lower and higher target force levels). Participants were requested to focus on keeping their line as close to the target as possible (the instructions from the "amplitude and timing" focus condition) for all tasks. Within retention and transfer trials separately, task order was randomized across effector x task combinations.

Measures

These measures are used for all analyses reported in Results.

Accuracy. Root mean square error has been used to index accuracy. Its appropriateness for pursuit tracking tasks is well established (Deutsch & Newell, 2004; Franks, Wilberg, & Fishburne, 1982; Newell et al., 2003). Studies using variable targets for pursuit tracking tasks have calculated RMSE as

$$\boldsymbol{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\boldsymbol{f_i} - \boldsymbol{T_i})^2}{\boldsymbol{N-1}}} \quad (2)$$

where f_i and T_i refer respectively to the force produced and the target force at time *i* (Sosnoff et al., 2004; Sosnoff & Voudrie, 2009). Because participants targeted different absolute levels of force (percentage of participant-x-effector-specific maximal voluntary force) and some tasks' target force varied over time, RMSE was normalized by moment-to-moment target force level:

$$\boldsymbol{NRMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\frac{\boldsymbol{f_i} - \boldsymbol{T_i}}{\boldsymbol{T_i}}\right)^2}{\boldsymbol{N-1}}} \quad (3)$$

The formula for RMSE is nearly identical to the calculation for sample standard deviation, while that for NRMSE simplifies to RMSE/T for a constant target (nearly identical to the calculation for coefficient of variation). The difference, and the reason they function as measures of accuracy rather than variability, is in the comparison to the target (prescriptive) rather than the mean (descriptive).

"Baseline NRMSE" refers to the NRMSE of the first trial per effector of each task the participant completed (day 1, block 1, trial 1). "Retention" and "transfer" NRMSE refer to the NRMSE of the first day-3 trial per effector x task combination under the named condition (retention: unaltered task; transfer: altered target force level). NRMSE was also used to provide feedback to participants.

Temporal structure of variability. Temporal structure was assessed using approximate entropy (Pincus, Gladstone, & Ehrenkranz, 1991) and fuzzy measure entropy (Liu et al., 2013); further information on each is available in Appendix A. ApEn was used for comparability to previous literature and FuzzyMEn as a potential improvement. The latter offered two particular advantages for this work. Firstly, since work on temporal structure of oral motor output is less common than investigations of other muscles, it is unknown whether the standard value of *r* (0.2 times the standard deviation of the time series) is the most appropriate, and thus a function less affected by choice of *r* was preferable. More importantly, two of the tasks required nonstationary force to be exerted. Processes causing small rapid fluctuations superimposed on the larger voluntary force changes, e.g. normal physiological tremor, may create similar vector shapes at different levels of force (Liu et al., 2013). These are best captured by an algorithm examining both local and global similarity.

ApEn and FuzzyMEn were calculated using custom LabVIEW functions written and optimized for computational time by the investigator based on MATLAB code for FuzzyMEn provided by Dr. Peng Li. Parameter choice and testing of their performance are described in Appendix D.

"Baseline" ApEn and FuzzyMEn refer to values for the first trial per effector of each task the participant completed (day 1, block 1, trial 1). "Retention" refers to the first of five retention trials per effector x task combination on day 3. "Transfer" refers to the first transfer trial in each condition (higher target force level vs. lower target force level) per effector x task combination on day 3.

Frequency structure of variability. A power spectral density function was determined using Welch's method: the signal was divided into eight segments, a Gaussian window was applied to each to reduce spectral leakage, and the periodograms of each segment were averaged. Using segment length = 256 with a series length of N = 1100 yielded 52.7% overlap of successive segments. FFT size of 512, with a 100 Hz sampling rate, yielded a frequency bin width of 0.195 Hz.

Proportion of power (PoP) was calculated as power within a frequency band divided by power in the entire frequency range (up to 50 Hz - the Nyquist frequency, half the sampling rate). Originally, 0-4 Hz, 4-8 Hz, and 8-12 Hz bands were used, following previous work (Deutsch & Newell, 2004; Sosnoff & Voudrie, 2009)[11]. This setup was revised to 1-HZ-wide[12] bands from 0 to 4 (see Results: Data Quality: Reframing of spectral analysis measures). Recall that sine and multicosine tasks targeted fundamental frequencies of 0.75 Hz. Ideal performance of either would yield the largest proportion of power in the 0-1 Hz band; the multicosine target also had successively smaller components at 1.5 and 3 Hz.

Testing of this methodology with signals of known spectral peaks and slope is described in Appendix E.

Software Development

Extensive custom routines were written in LabVIEW. A main dashboard was created to calibrate the transducers, run task demonstrations and start and stop the experiment.

Calibration functions included real-time visual data display during trials, systematic generation of file names and internal file components, data logging, and parameterized flagging of problematic trials for visual review and 'use/do not use' decision logging. (See Calibration.)

---

[11] These authors offered different though not mutually exclusive reasons for choosing these frequency bands. Deutsch and Newell (2004) stated that "earlier experiments showed these to be the main frequency ranges influenced differentially as a function of age" (p. 325, no further details provided). Sosnoff and Voudrie (2009) cite earlier work suggesting that the 0-4 Hz band is associated with sensorimotor processing, for instance use of feedback, while physiological tremor contributes to power in the 8-12 Hz band.

[12] 100/512 = 0.1953125. Technically all frequency bands discussed here are multiples of this number (0-0.9765625 Hz, 0.9765625 – 1.953125 Hz etc.). Rounding in category descriptors is used for convenience.

Task demonstrations displayed the specified task target, then overlaid it with previously recorded data, mimicking the timing of an actual data collection trial. Feedback was provided for up to five demonstration trials per task (see Figure 7). This capability was used during the informed consent process to teach the participant what to expect and how to interpret it.

Starting the experiment initiated routines performing multiple tasks: completing the initial and daily screens; cuing the participant to ask any questions and confirming consent (required daily before data collection but documented only on day 1); generating the daily task list (ordered and randomized as described above); presenting all tasks, with effector, task and focus cued at the beginning of each block of five trials; recording both raw and cropped time series of the transducers' data under systematically generated filenames; calculating all trial-level analyses (NRMSE, ApEn, etc.) and storing them with the trial data; displaying feedback and progress through the task list; and counting down the specified rest periods between trials. In the event of a program interruption, the Resume Experiment function retrieved the participant's MVF and the previously generated task list, and prompted identification of the last successfully recorded trial to enable continuation from the appropriate point. The experiment could be stopped at any point, or a trial could be restarted without stopping and restarting the experiment. If a trial was restarted, the rejected trial data could optionally be logged with an explanatory comment (e.g. "participant coughed").

All data and associated calculations were logged using TDMS file format. This format contains internal *group* and *channel* structure, as well as file-, group- and channel-level properties, allowing each trial run to be labeled with all relevant characteristics and all calculations to be stored with the data upon which they were based.

The current version of the software is capable only of repeating the present work. However, it was structured with the intent of being upgradeable to eventually permit collection of data from any compatible transducer, using any specified repeating target pattern with investigator-controlled numbers of trials, trial order randomization, etc. for flexibility in design of related experiments.

Pre-Analysis Data Transformation

Data for the maximal voluntary force task were transformed from volts to Newtons using the effector-specific equations developed during calibration. All pursuit task data were transformed to a percentage of the used effector's MVF and detrended using a least-squares linear fit. Normalized root mean square error was calculated prior to detrending. These calculations were performed in LabVIEW.

Statistical Analysis

All analyses were performed using SPSS (IBM, Version 24). Only significant results are reported, unless a nonsignificant result directly addresses one of the research questions.

Modeling tools. Linear mixed effects (LME) modeling with random subject intercepts, maximum likelihood estimation and type III sums of squares was the primary analysis tool. This method allows repeated measurement of the same participant under different conditions (here, combinations of effector and task) and includes in the model within-participant similarity of responses across differing conditions. When data did not support fitting a random subjects intercept, mixed-model[13] repeated measures analysis of variance (RMANOVA) was used instead, permitting inclusion of fixed effects only.

Specific models' construction is described in the following subsections. In general, for each measure, the postulated model was fitted to the corresponding data. Any interaction terms not reaching statistical significance were discarded and the simplified model rerun until (i) no significant interactions remained in the model or (ii) only significant interactions remained in the model. For case (i), the significance of the main effects was evaluated and only the significant main effects were interpreted. For case (ii), interactions were further investigated by testing the significance of the corresponding simple effects[14]. Main effects were reported only if the factor in question did not participate in any significant interactions; if it did, only its simple effects were reported. Simple simple main effects of age group within effector x task conditions were tested using independent-samples *t-test*s with bootstrapping. Bonferroni adjustment of the significance criterion was used to control familywise error, meaning that at each level of each analysis path, the chance of falsely rejecting the null hypothesis is 5%. Appendix F describes the criteria used for each potential term evaluated.

One-way analyses of change over time: occurrence of learning. To check that learning occurred for both age groups under all conditions, separate analyses were performed for each age group x effector x task combination. Normalized root mean square error was compared at two time points: baseline (day 1, block 1, trial 1) and either retention or transfer (day 3, trial 1 of each type of task). Linear mixed effects models with random subjects intercepts and a main effect of time were fit.

---

[13] In "linear mixed effects," "mixed" refers to the inclusion of both fixed and random effects. In "mixed-model RMANOVA," "mixed" refers to the inclusion of both between- and within-subjects factors.

[14] A simple effect means the effect of categorical factor A on the dependent variable for specific levels of categorical factor B with which it significantly interacts, e.g. "the effect of task on ΔNRMSE for older adults."

Two-way analysis: maximal voluntary force. MVF was assessed using a two-way linear mixed effects model with one between-subjects factor (age group) and one within-subjects factor (effector), each with two levels. The initial model included the main effects, the two-way interaction between them, and a random subjects intercept.

Three-way analyses: accuracy and entropy. Analyses of accuracy (NRMSE) and entropy (ApEn, FuzzyMEn) characterized initial performance, final performance, initial adaptability and change between initial and final performance using LME or RMANOVA to delineate the effects of age group, effector and task on specific measures defined in each Results subsection. These models used one between-subjects factor (age group, two levels: younger vs older adults) and two within-subjects factors (effector, two levels: lip vs tongue; task, three levels: constant, sine, multicosine). Initial models were full-factorial, i.e. included all main effects and interactions, as well as random subjects intercepts (LME) where possible.

Four-way analyses: proportion of power in spectral frequency bands. Spectral analyses (proportion of power) were completed using four-way mixed-model RMANOVA, with the factors listed above plus frequency band, a within-subjects categorical factor with three levels: 0-1 Hz, 1-2 Hz, and 2-3 Hz. Only these three were used to avoid unstable modeling results due to multicollinearity (because for each trial, the sum of all the bands is 1). The highest-frequency band was omitted because it contained relatively little energy. These models were fit using the main effects of each factor, the two-way interactions of frequency band with each other factor, and the three-way interactions of age group and frequency band with (separately) task and effector. Only terms including frequency band are interpreted: varying effector, task, age group or a combination can vary the distribution of proportional power across bands, measured by including the frequency band term, but cannot vary the sum across bands (always 1). These analyses should be interpreted with caution given the proportional nature of the data, which violates the assumption of an unbounded dependent variable.

Three-way analyses: prediction of change. To test the hypotheses that specific continuous quantitative predictor variables (each described in the Results section for its model) would predict learning, linear mixed effects models were used. The dependent variable for all of these models was the change in normalized root mean square error from initial performance to either retention trials or transfer trials to higher target force level: day-3 retention or transfer trial 1 minus day 1, block 1, trial 1 for each effector x task combination ($\Delta_{final}NRMSE_{ret}$ or $\Delta_{final}NRMSE_{trn}$). Age group, effector and task were included as categorical factors along with two continuous quantitative predictor variables per model: normalized root mean square error on day 1, block 1, trial 1 ($NRMSE_{initial}$) and one entropy measure.

Every model also included the two-way interactions of each categorical factor with the quantitative entropy predictor under consideration in that model. Given sample size, no other interaction terms were considered, limiting focus to the entropy predictor and its interactions with the categorical factors. A significant main effect of a categorical factor was interpreted only if the categorical factor did not interact significantly with the entropy predictor.

Testing of model assumptions and response to violations. Normality of fitted-model residuals' and predicted random effects' distributions were assessed using the Shapiro-Wilk statistic, recommended for groups of N < 50; violations were accepted if skewness fell between -2 and 2. When skewness fell outside this range, removal of extreme values or data transformation was used to attempt remediation and the analysis was repeated on the altered data.

Square root and base-10 logarithmic transformation were used for moderately and strongly skewed data respectively, after reflect transformation[15] if skew was negative. These transformations compress the data range, with greater compression as data value increases, to reduce skewness of the distribution; compression is more pronounced with the logarithmic than with the square root transformation. Square root and logarithm transformations can only be used on positive data and only compress its range for certain values ($x > 0.25$ for square root and $x > 1$ for logarithm); below these inflection points, they instead expand the range. See Figure 10. When data needed to be transformed but fell within this problematic range, 1 was added to the data prior to transformation (de Smith, 2015).

Results of model assumption testing are reported, and transformation or removal of extremes was used, only when assumptions were violated.

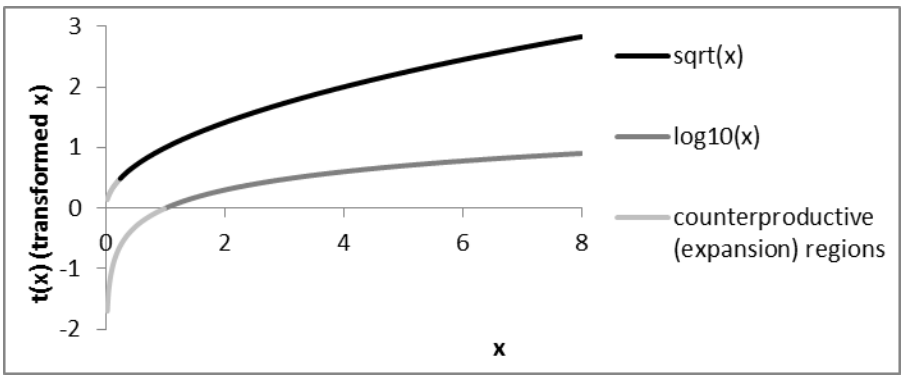

Figure 10. Transformations' shapes.

[15] Reflect transformation subtracts the data from one greater than its maximum value, making all values ≥ 1 and reversing their magnitude order. The other transformations do not alter data order.

Results

Organization and Reporting Conventions

For aims 1 and 2, hypotheses not dependent on practice were evaluated at both initial and final performance, followed by hypotheses regarding change with practice.

In the Data Quality section, where the spread of the actual data sample is relevant, figures are reported as mean ± standard deviation. Elsewhere, a probabilistic statement about the probable true mean is more appropriate, and figures are reported as mean ± standard error.

Consistent graphing conventions are used throughout for column plots: age group is differentiated by column fill (filled for older adults, unfilled for younger adults) and effector by shade (dark gray for tongue, medium gray for lip). Error bars represent ± 1 standard error. Unless otherwise noted, all graphs display original (not transformed) data. Significance is indicated on the graphs using a single asterisk, with thin horizontal lines indicating grouping of only the single elements directly below tick marks in the line (e.g. “all columns for ‘older adult, lip’ across task”) and thick horizontal lines for grouping all elements directly below the line (e.g. “all age group x effector combinations within a task”).

Temporal structure of force was measured by approximate entropy (ApEn) and fuzzy measure entropy (FuzzyMEn), both with parameters $m = 2$, $r = 0.2$, $N = 1100$. Because the pattern of significant results was the same or very similar for each, the two measures are presented together.

Most statistical test results are presented within the text. When there is a significant three-way interaction with the attendant follow-up analyses for simple two-way interactions, simple simple main effects, and pairwise comparisons, statistical results are presented in a table formatted to visually clarify the structure of the analysis, and the accompanying text summarizes and interprets the table without repeating the numbers.

Missing Data

Eleven participants who passed the telephone screen, scheduled and showed up for day 1 of the experiment were not included: eight for reasons specific to them and three due to equipment issues. All participants were paid for time completed, regardless of whether their data could be used.

Participant-based. Three participants provided full or nearly full data that could not be used. One older man admitted at the end of the final experimental session that his stated age and the birthdate on his driver’s license were incorrect; he was actually one year too young to participate.

One older woman developed mild soreness at the commissure of the lips while using the lip transducer on day 3; one prong had rotated and was chafing. There was no visible skin tear or

inflammation. She did not complete the remaining lip trials but was able to use the tongue transducer without discomfort. There were no other significant adverse events.

One older woman who completed all three days initially appeared to be an outlier with unusually poor performance, but it was established after experiment completion that her lip MVF value had been far too low, which led to target forces much lower (and therefore more difficult to control) than the intended 15% MVF target. Her performance immediately improved to comparable to her age group peers with the correct MVF value, but as she was no longer naïve to the task, she could not redo the experiment.

Five other participants did not complete the experiment. One withdrew after day 1 due to a scheduling conflict, and four withdrew early in day 1. Of these, one had spasmodic torticollis with a notable head tremor, undisclosed during recruitment screening; two had oral pain due to ill-fitting dentures or same-day dental work; one could not find a comfortable posture due to arthritis pain in her back and shoulders, which readjustment and additional padding of the chair did not resolve.

Instrument-based. For one younger man, the transducers' behavior was unstable, with frequent baseline drifts and jumps; he attempted the tasks enough times during the unsuccessful initial session that rescheduling was not appropriate.

The lip transducer experienced one episode of failure to transmit data, with abrupt cessation near the end of one participant's day-2 session. The transducer was returned to the manufacturer for repair. (Calibration values after repair were similar to those seen before failure.) That participant and another scheduled on the same days, both older men, provided complete tongue data but were excluded from analyses.

To screen for data-capture problems not detected during experimental sessions, all task data trials for both transducers were checked for runs of more than 10 consecutive values of exactly 0 (equivalent to 0.1 seconds at the sampling rate of 100 Hz), and for runs of more than 10 unchanging values. Either type of run could have indicated a problem within the transducer or transmission failure along the transducer-amplifier-computer path. Trials were also checked for runs of more than 500 values (5 seconds) less than 5% task-specific MVF, as a general flag to trigger visual review. The only trials excluded on these bases were those described above.

Data Quality

Recall that any trial during which performance was perturbed (e.g. by a cough) was restarted. Such trials were not formally tracked but are estimated to have occurred less often than once per

participant, i.e. less than 1% of trials. All trials were reviewed and accepted by the investigator during the experiment immediately upon completion, prior to either the experimenter or the participant seeing NRMSE feedback. Acceptance required only absence of evidence of perturbation or signal loss and was not based on level of success tracking the target.

The possible effect of transducer noise was assessed by comparing mean transducer signal in zero-mass calibration trials to the lowest target (in volts). Across all tasks, the lowest relative target was 5% MVF (during sine and multicosine transfer trials altering target force level to 10% ± 5% MVF). Across all tasks and participants, the lowest absolute target was 5% MVF for the participants with the lowest maximal voluntary forces per effector, giving the equation

$$maximum\ relative\ noise\ amplitude = 100\% * \frac{(zero\ mass\ calibration\ trial\ mean)_{effector}}{0.05 * MVF_{minimum\ across\ participants\ by\ effector}} \quad (4)$$

Zero-mass calibration trial means were -0.00025 ± 0.00096 volts for the lip, -0.00013 ± 0.0015 volts for the tongue (from Table 2). One of the older women had the lowest MVF for the lip, at 3.06 N (0.528 volts); an older man had the lowest MVF for the tongue, at 6.74 N (1.67 volts). Relative amplitudes of transducer noise at 5% MVF were 0.946% and 0.155% respectively. Thus even for the lowest target forces across all participants, transducer noise would have affected an otherwise perfect match to target by less than 1%.

Reframing of spectral analysis measures. Brief exploratory analysis of proportion of power figures showed that power was overwhelmingly concentrated in the 0-4 Hz band: across all analyzed trials, mean ± standard error proportion of power in this band was 0.97 ± 0.00062. Seventy percent of trials had values above the mean, and 95% had values above 0.90; see Figure 11. There was correspondingly little energy above 4 Hz. This suggested that any changes in frequency structure of participants' force output took place within the 0-4 Hz range – plausible given the influence of aging and visuomotor processes on power from 0-1 Hz (Baweja, Kennedy, Vu, Vaillancourt, & Christou, 2010; Fox et al., 2013) and the specific task demands in this experiment (variable-force targets contained frequencies of 0.75, 1.5 and 3 Hz).

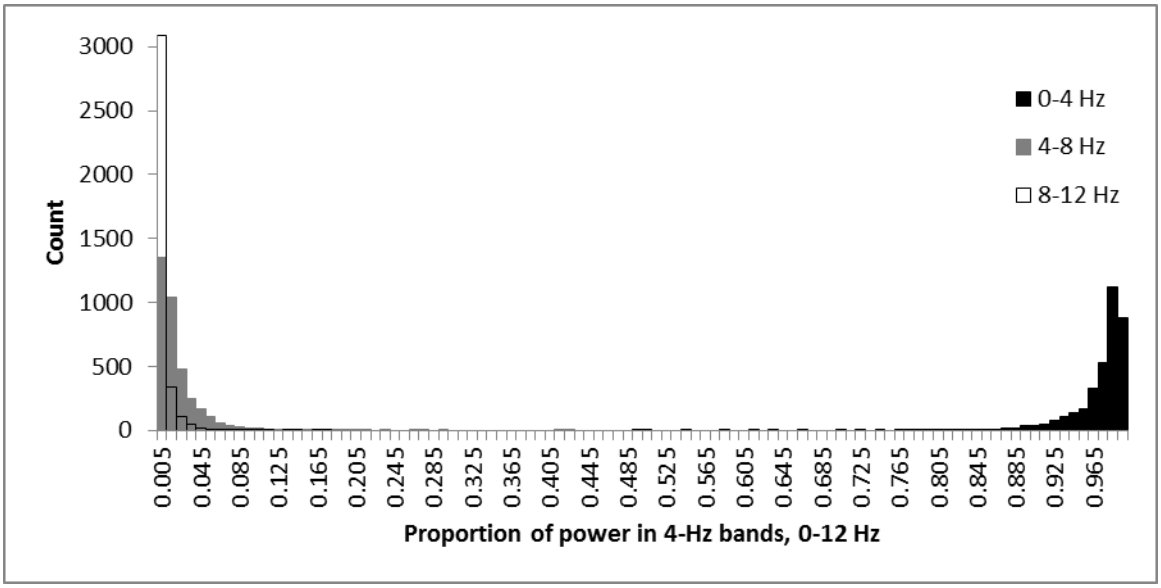

Figure 11. Proportion of power distributions using original frequency bands. The 8-12 Hz band columns are transparent, not white-filled, for visibility vs. the 4-8Hz distribution.

Consequently, frequencies above 4 Hz were not analyzed. To make statistical discrimination among samples more likely by matching measurement scale to scale of expected change, the 0-4 Hz band was split into approximately 1-Hz wide bands from 0-1, 1-2, 2-3 and 3-4 Hz (Figure 12). Each was composed of the sum of five 0.195-Hz-wide bins.

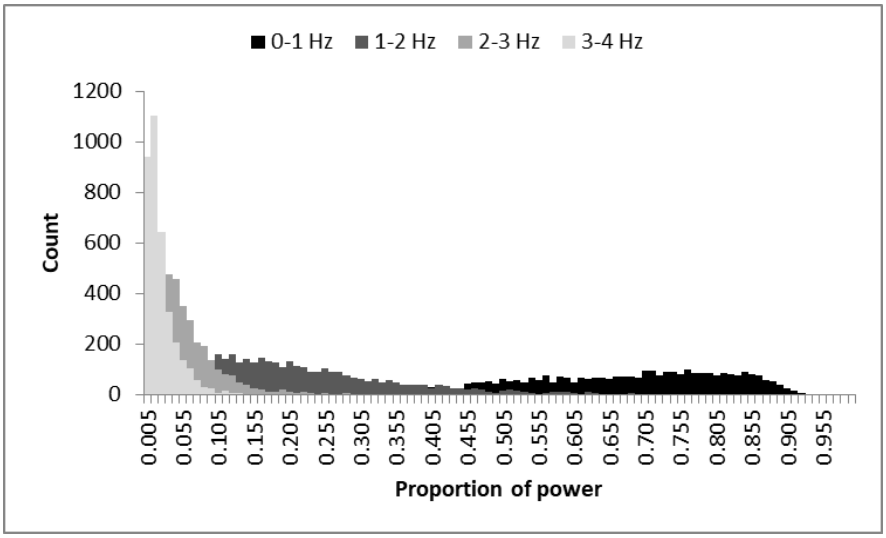

Figure 12. Proportion of power distribution using narrower frequency bands.

Maximal Voluntary Force

This analysis used linear mixed effect modeling to evaluate maximal voluntary force in Newtons (MVF) to determine how it was affected by age group, effector and their interaction. The interaction did

not reach significance in the full model ($F(1, 41) = 1.427$, $p = 0.239$) and was dropped for the simplified, main-effects-only model.

MVF did not vary by age group ($F(1, 41) = 0.389$, $p = 0.536$) but was greater for tongue than lip by 6.29 N ($F(1, 41) = 64.362$, $p < 0.0005$, 95% CI 4.71 – 7.87 N). See Figure 13.

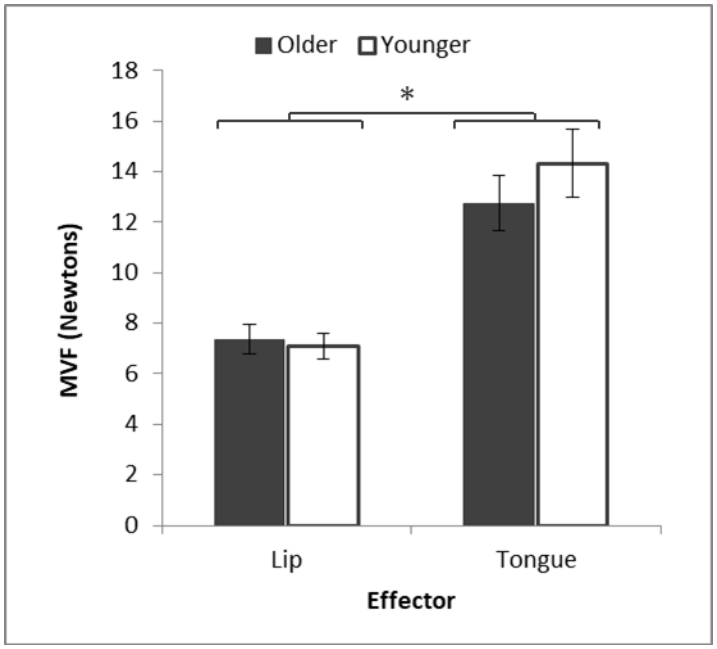

Figure 13. Maximal voluntary force(MVF) by age group and effector (Newtons, M ± SE).
* Significant at $p < 0.0005$.

Pursuit Tracking: Representative Examples

A representative was chosen from each age group by summing participants' absolute differences from the age group mean NRMSE on each effector x task combination. In each age group, the participant with the lowest sum was taken to be overall nearest group-mean performance. The participants thus chosen were a younger man, age 24, and an older man, age 73.

Initial performance. Figures 14-19 and Tables 5-10 show participant force traces on each task's first trial on day 1 and their associated measures. Only the analyzed portion of the trace is shown. Proportions of power do not sum to 1.0 due to rounding and exclusion of power above 4 Hz.

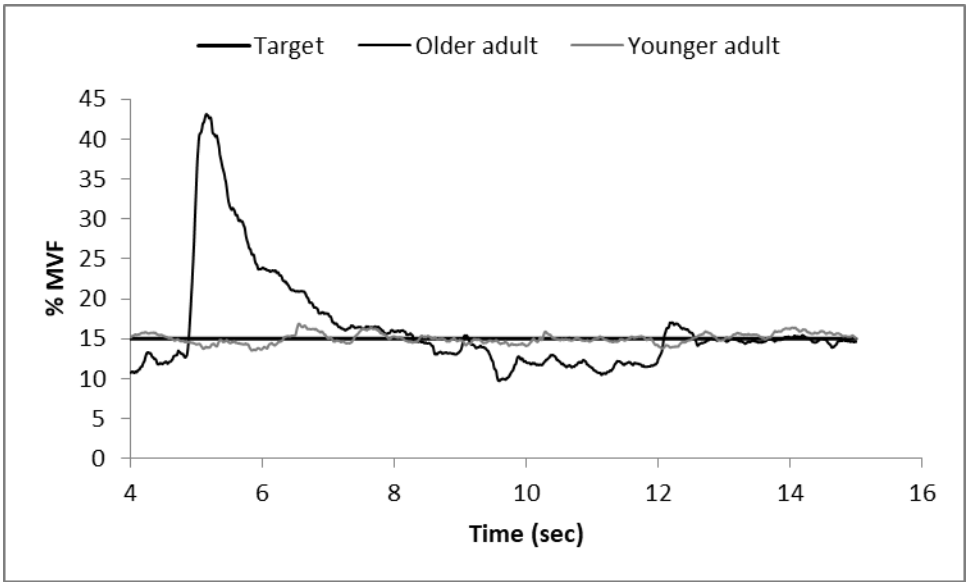

Figure 14. Representative participants' initial force trace for constant task using lip: day 1, trial 1.

Table 5

Representative Participants' Initial Performance Measures for Constant Task Using Lip

Measure	Older adult	Younger adult
NRMSE	0.442	0.043
ApEn	0.089	0.439
FuzzyMEn	0.073	0.637
PoP 0-1 Hz	0.900	0.748
PoP 1-2 Hz	0.085	0.171
PoP 2-3 Hz	0.010	0.026
PoP 3-4 Hz	0.003	0.017

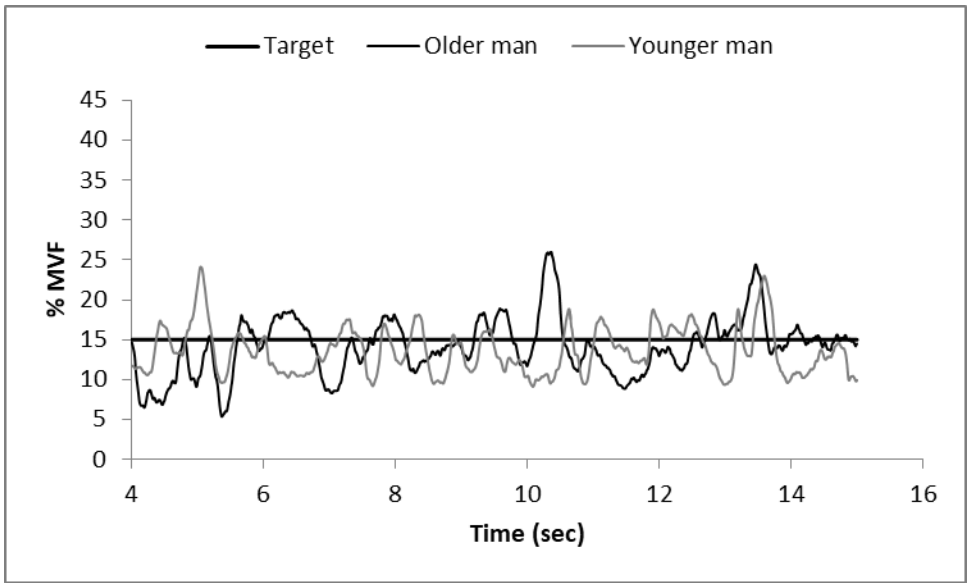

Figure 15. Representative participants' initial force trace for constant task using tongue: day 1, trial 1.

Table 6

Representative Participants' Initial Performance Measures for Constant Task Using Tongue

Measure	Older adult	Younger adult
NRMSE	0.248	0.218
ApEn	0.350	0.409
FuzzyMEn	0.455	0.532
PoP 0-1 Hz	0.617	0.465
PoP 1-2 Hz	0.297	0.366
PoP 2-3 Hz	0.064	0.128
PoP 3-4 Hz	0.012	0.020

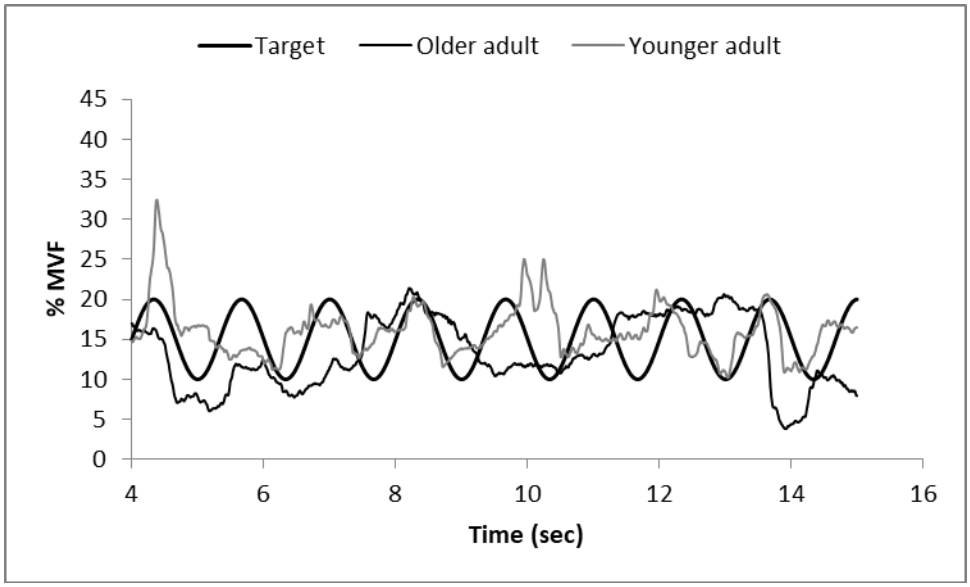

Figure 16. Representative participants' initial force trace for sine task using lip: day 1, trial 1.

Table 7

Representative Participants' Initial Performance Measures for Sine Task Using Lip

Measure	Older adult	Younger adult
NRMSE	0.404	0.319
ApEn	0.129	0.297
FuzzyMEn	0.131	0.361
PoP 0-1 Hz	0.944	0.782
PoP 1-2 Hz	0.041	0.121
PoP 2-3 Hz	0.008	0.048
PoP 3-4 Hz	0.002	0.023

Figure 17. Representative participants' initial force trace for sine task using tongue: day 1, trial 1.

Table 8

Representative Participants' Initial Performance Measures for Sine Task Using Tongue

Measure	Older adult	Younger adult
NRMSE	0.368	0.343
ApEn	0.333	0.379
FuzzyMEn	0.385	0.500
PoP 0-1 Hz	0.626	0.525
PoP 1-2 Hz	0.268	0.363
PoP 2-3 Hz	0.080	0.080
PoP 3-4 Hz	0.014	0.019

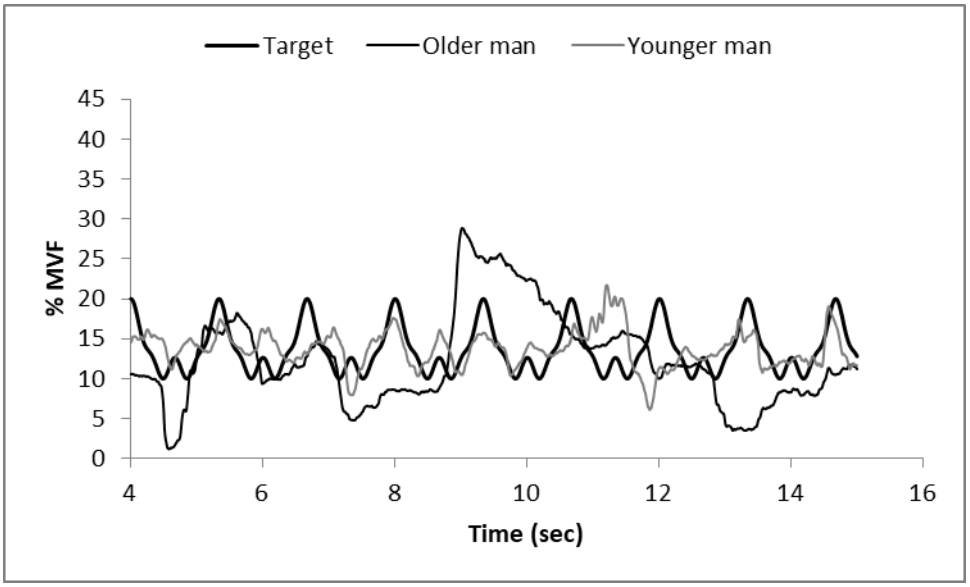

Figure 18. Representative participants' initial force trace for multicosine task using lip: day 1, trial 1.

Table 9

Representative Participants' Initial Performance Measures for Multicosine Task Using Lip

Measure	Older adult	Younger adult
NRMSE	0.483	0.265
ApEn	0.116	0.369
FuzzyMEn	0.119	0.487
PoP 0-1 Hz	0.929	0.628
PoP 1-2 Hz	0.058	0.288
PoP 2-3 Hz	0.008	0.042
PoP 3-4 Hz	0.002	0.017

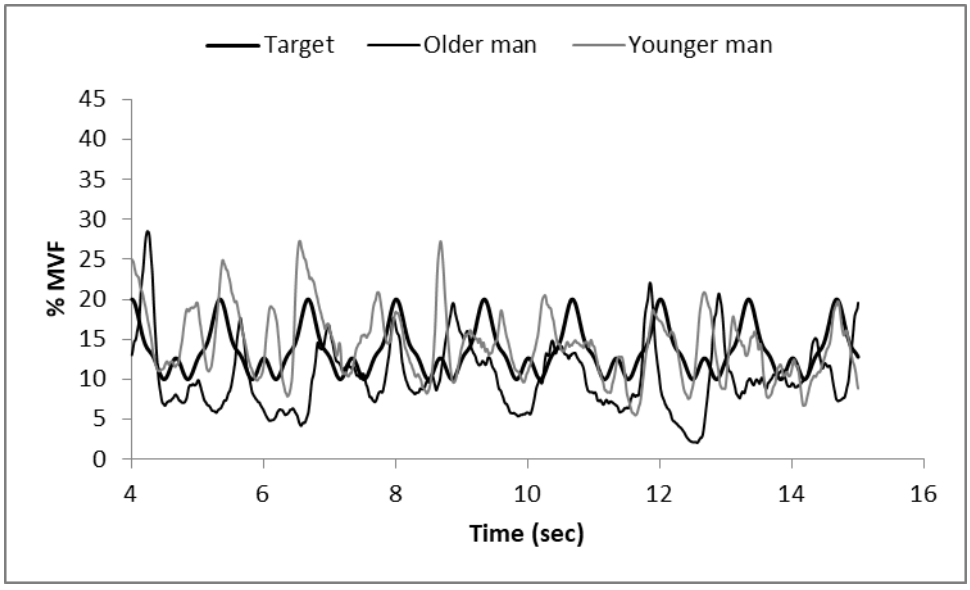

Figure 19. Representative participants' initial force trace for multicosine task using tongue: day 1, trial 1.

Table 10

Representative Participants' Initial Performance Measures for Multicosine Task Using Tongue

Measure	Older adult	Younger adult
NRMSE	0.418	0.315
ApEn	0.304	0.420
FuzzyMEn	0.382	0.580
PoP 0-1 Hz	0.670	0.308
PoP 1-2 Hz	0.250	0.424
PoP 2-3 Hz	0.053	0.209
PoP 3-4 Hz	0.017	0.041

Final performance. Figures 20-25 show force traces from the same participants showcased previously, to the same scale. Traces are taken from each participant's first retention trial on day 3. Time scale reflects cropping of first four seconds and last one second of each trial. Proportions of power do not sum to 1.0 due to rounding and exclusion of power above 4 Hz. The accompanying tables show measures based on those trials.

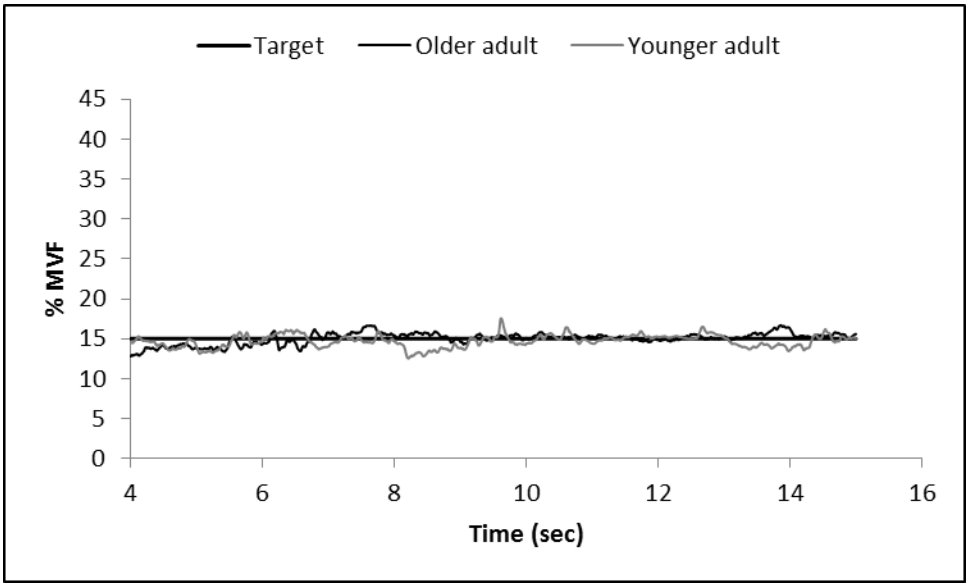

Figure 20. Representative participants' final force trace for constant task using lip: day 3, retention trial 1.

Table 11

Representative Participants' Final Performance Measures for Constant Task Using Lip

Measure	Older adult	Younger adult
NRMSE	0.048	0.055
ApEn	0.569	0.528
FuzzyMEn	0.854	0.777
PoP 0-1 Hz	0.630	0.662
PoP 1-2 Hz	0.133	0.186
PoP 2-3 Hz	0.060	0.044
PoP 3-4 Hz	0.103	0.036

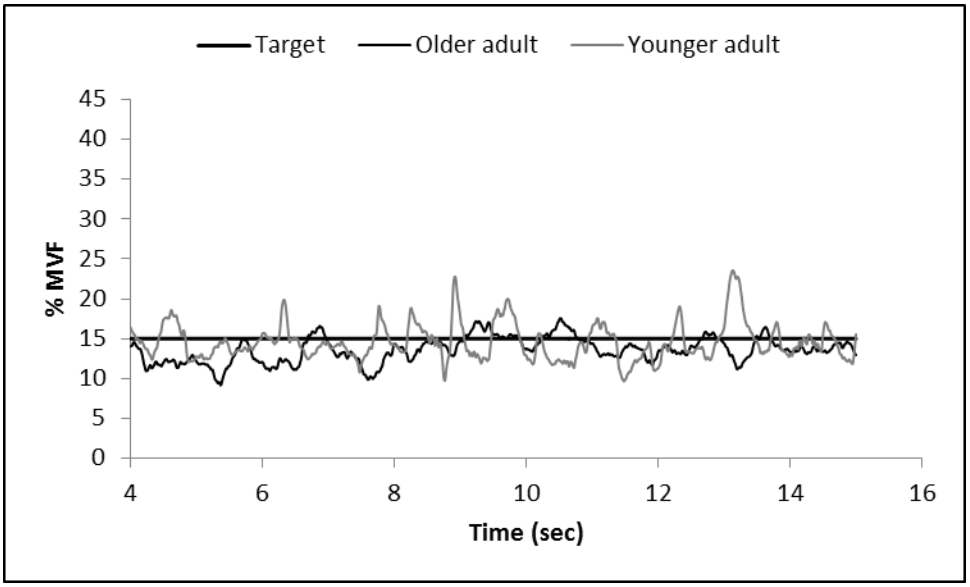

Figure 21. Representative participants' final force trace for constant task using tongue: day 3, retention trial 1.

Table 12

Representative Participants' Final Performance Measures for Constant Task Using Tongue

Measure	Older adult	Younger adult
NRMSE	0.141	0.157
ApEn	0.414	0.476
FuzzyMEn	0.580	0.657
PoP 0-1 Hz	0.671	0.358
PoP 1-2 Hz	0.269	0.386
PoP 2-3 Hz	0.024	0.103
PoP 3-4 Hz	0.016	0.083

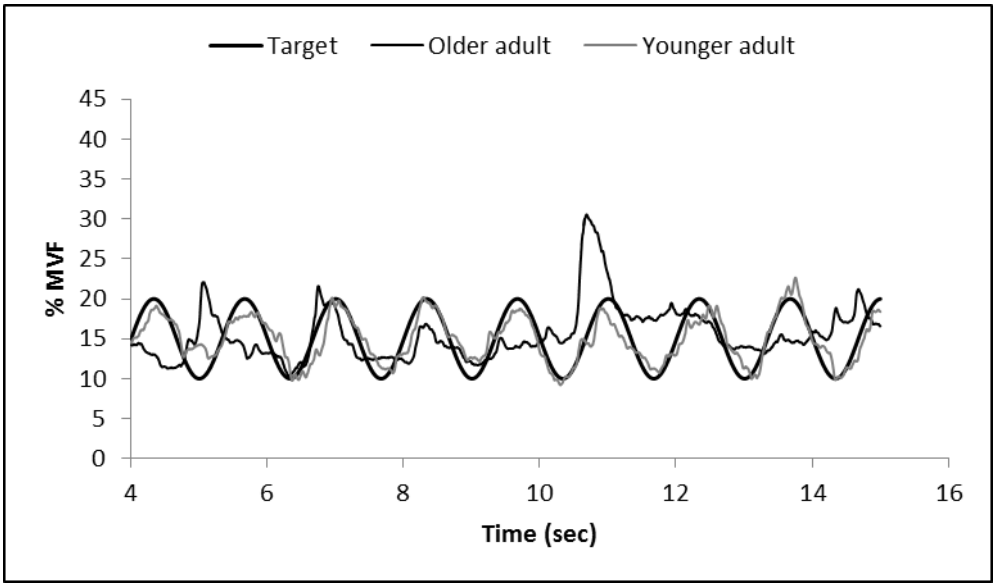

Figure 22. Representative participants' final force trace for sine task using lip: day 3, retention trial 1.

Table 13

Representative Participants' Final Performance Measures for Sine Task Using Lip

Measure	Older adult	Younger adult
NRMSE	0.328	0.128
ApEn	0.180	0.317
FuzzyMEn	0.215	0.410
PoP 0-1 Hz	0.788	0.875
PoP 1-2 Hz	0.164	0.085
PoP 2-3 Hz	0.020	0.019
PoP 3-4 Hz	0.015	0.007

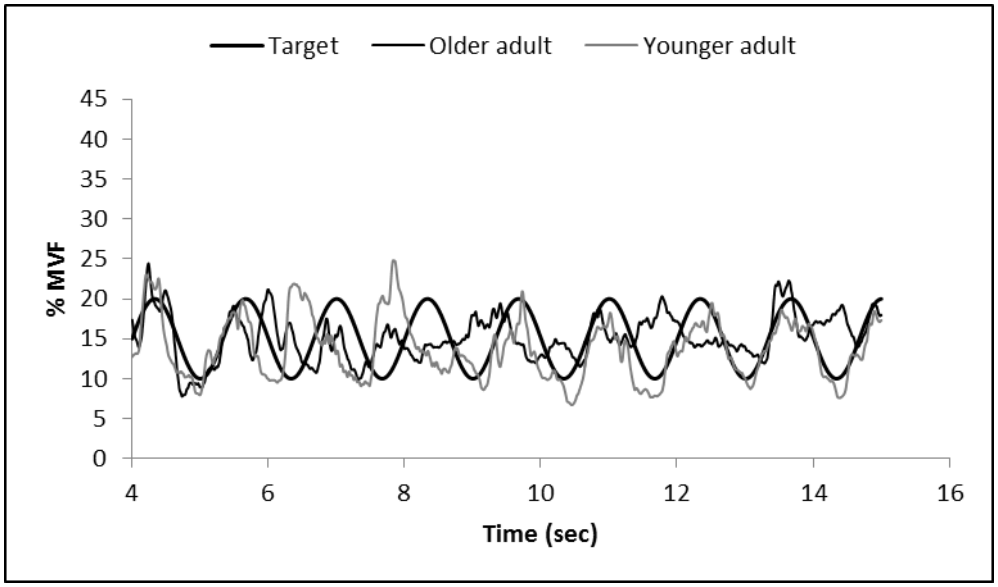

Figure 23. Representative participants' final force trace for sine task using tongue: day 3, retention trial 1.

Table 14

Representative Participants' Final Performance Measures for Sine Task Using Tongue

Measure	Older adult	Younger adult
NRMSE	0.312	0.308
ApEn	0.406	0.373
FuzzyMEn	0.539	0.477
PoP 0-1 Hz	0.617	0.745
PoP 1-2 Hz	0.247	0.180
PoP 2-3 Hz	0.073	0.023
PoP 3-4 Hz	0.032	0.024

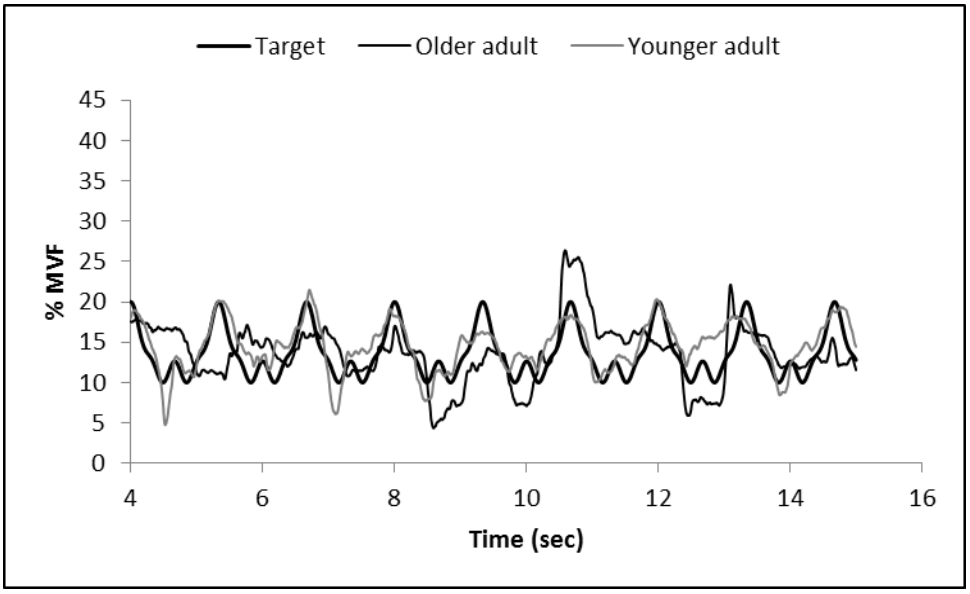

Figure 24. Representative participants' final force trace for multicosine task using lip: day 3, retention trial 1.

Table 15

Representative Participants' Final Performance Measures for Multicosine Task Using Lip

Measure	Older adult	Younger adult
NRMSE	0.293	0.186
ApEn	0.220	0.313
FuzzyMEn	0.260	0.412
PoP 0-1 Hz	0.851	0.760
PoP 1-2 Hz	0.076	0.182
PoP 2-3 Hz	0.044	0.032
PoP 3-4 Hz	0.013	0.012

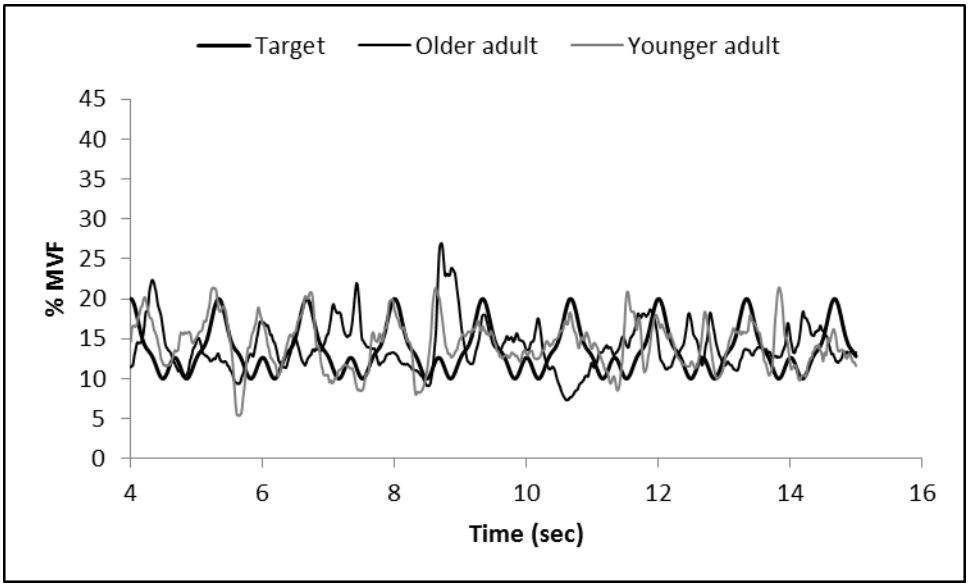

Figure 25. Representative participants' final force trace for multicosine task using tongue: day 3, retention trial 1.

Table 16

Representative Participants' Final Performance Measures for Multicosine Task Using Tongue

Measure	Older adult	Younger adult
NRMSE	0.382	0.257
ApEn	0.348	0.477
FuzzyMEn	0.474	0.651
PoP 0-1 Hz	0.548	0.362
PoP 1-2 Hz	0.306	0.449
PoP 2-3 Hz	0.063	0.118
PoP 3-4 Hz	0.048	0.034

Occurrence of learning. Because prediction of learning was the most important aim of the work, it was necessary to show that the amount of practice provided was sufficient to elicit learning. These linear mixed model analyses evaluated normalized root mean square error (NRMSE) at two time points: baseline (day 1, block 1, trial 1) and trial 1 of retention or transfer tasks on day 3. The purpose was to determine whether learning occurred for both age groups under each effector x task condition after two days' practice. See Statistical Analysis for details. No graphs are presented for these analyses, which focused only on main effect of time across two points.

Retention. For retention trials, a reduction in NRMSE from baseline indicated improvement in performance on the practiced tasks. When models were fit to the original data, those marked in the table could not be adequately fit due to model assumption violations. Data transformations led to successful fitting for only one of these. Instead, for each, participants with extreme values were removed and the analyses re-run on non-transformed data. All alternate analyses were consistent in significance and direction of difference with results of the original-data analyses reported here.

Table 17

Effect of Time on Normalized Root Mean Square Error for Each Combination of Age Group, Effector and Task, Comparing Day 1, Trial 1 with Day 3, Retention Trial 1

Age group	Effector	Task	Test statistic	Significance
Older adult	lip	constant	$F(1, 21) = 8.932$	$p = 0.007$*
		sine	$F(1, 21) = 5.750$	$p = 0.026$*
		multicosine†	$F(1, 21) = 4.773$	$p = 0.040$*
	tongue	constant	$F(1, 21) = 15.931$	$p = 0.001$*
		sine	$F(1, 21) = 33.806$	$p < 0.0005$*
		multicosine	$F(1, 21) = 3.431$	$p = 0.078$
Younger adult	lip	constant†	$F(1, 20) = 10.439$	$p = 0.004$*
		sine	$F(1, 20) = 43.093$	$p < 0.0005$*
		multicosine†	$F(1, 20) = 17.505$	$p < 0.0005$*
	tongue	constant	$F(1, 20) = 38.687$	$p < 0.0005$*
		sine	$F(1, 20) = 73.817$	$p < 0.0005$*
		multicosine	$F(1, 20) = 41.433$	$p < 0.0005$*

Note. All significant differences were in the direction of reduction in error on day 3.
† Alternate analyses were completed (see text); all figures are from original analyses.

Participants demonstrated retention of previously practiced skill after a one-day delay regardless of age group, task or effector, with the single exception of older adults performing the multicosine task with their tongues, for which the effect of time did not reach significance. See Table 17.

This pattern is consistent with comments by the participants, who felt this task combination to be particularly difficult.

Transfer. For transfer trials, lower NRMSE on day 3 than day 1 indicated application of learned skill to a related task. Models were fit separately for the two transfer conditions (reduced and increased target force level) because of the expected difference in task difficulty.

Transfer to lower target force. When linear mixed effects models were fit to the original data, those starred in the table could not be adequately fit due to model assumption violations. Removal of participants with extreme values led to adequate model fit for three models. Paired t-tests with bootstrapping were used for the other two. Where alternate analyses agreed with the original model, the original model's test statistics are reported. In the two models for which the alternate analysis failed to find the significant effect detected by the original model, the alternate analysis is reported.

Table 18

Effect of Time on Normalized Root Mean Square Error for Each Combination of Age Group, Effector and Task, Comparing Day 1, Trial 1 with Day 3, Transfer Trial 1 when Target Force Level was Reduced

Age group	Effector	Task	Test statistic	Significance	Direction of change
Older adult	lip	constant	$F(1, 21) = 3.379$	$p = 0.080$	
		sine	$F(1, 21) = 9.431$	$p = 0.006$*	↑
		multicosine†	$F(1, 21) = 0.322$	$p = 0.576$	
	tongue	constant	$F(1, 21) = 10.814$	$p = 0.004$*	↓
		sine	$F(1, 21) = 0.104$	$p = 0.750$	
		multicosine†	$F(1, 20) = 3.628$	$p = 0.071$	
Younger adult	lip	constant†	$F(1, 18) = 2.742$	$p = 0.115$	
		sine†	$F(1, 40) = 1.504$	$p = 0.227$	
		multicosine†	$F(1, 40) = 6.499$	$p = 0.015$*	↑
	tongue	constant	$F(1, 20) = 16.940$	$p = 0.001$*	↓
		sine	$F(1, 20) = 25.418$	$p < 0.0005$*	↓
		multicosine	$F(1, 20) = 3.515$	$p = 0.075$	

† Alternate analyses were completed (see text).

NRMSE on transfer trials with lower target force level was lower than baseline error for only three conditions, all using the tongue: older and younger adults tracking a constant target and younger adults tracking sine targets. NRMSE was significantly *higher* than baseline for two conditions, both using the lip: older adults tracking a sine target and younger adults tracking a multicosine target. For the other

conditions, no significant difference was found. Participants demonstrated inconsistent transfer of skill to familiar target patterns at lower force levels.

Transfer to higher target force. When linear mixed effects models were fit to the original data, several (marked in Table 19) could not be adequately fit due to model assumption violations. Removal of participants with extreme values led to adequate model fit for four models; a paired-samples t-test with bootstrapping was used for the last. All results of alternate analyses were consistent in significance and direction of difference with results of the original-data analyses reported here. Participants demonstrated consistent transfer of skill to familiar target patterns at higher force levels regardless of age group, task or effector.

Table 19

Effect of Time on Normalized Root Mean Square Error for Each Combination of Age Group, Effector and Task, Comparing Day 1, Trial 1 with Day 3, Transfer Trial 1 when Target Force Level was Increased

Age group	Effector	Task	Test statistic	Significance	Direction of change
Older adult	lip	constant	$F(1, 21) = 12.211$	$p = 0.002$*	↓
		sine†	$F(1, 21) = 40.483$	$p < 0.005$*	↓
		multicosine*	$F(1, 21) = 16.922$	$p < 0.0005$*	↓
	tongue	constant†	$F(1, 21) = 34.426$	$p < 0.0005$*	↓
		sine	$F(1, 21) = 50.200$	$p < 0.0005$*	↓
		multicosine	$F(1, 21) = 33.351$	$p < 0.0005$*	↓
Younger adult	lip	constant†	$F(1, 40) = 4.691$	$p = 0.036$*	↓
		sine	$F(1, 20) = 63.749$	$p < 0.0005$*	↓
		multicosine†	$F(1, 40) = 44.227$	$p < 0.0005$*	↓
	tongue	constant	$F(1, 20) = 30.903$	$p < 0.0005$*	↓
		sine	$F(1, 20) = 95.933$	$p < 0.0005$*	↓
		multicosine	$F(1, 20) = 61.812$	$p < 0.0005$*	↓

† Alternate analyses were completed (see text).

Specific Aims 1 and 2: Non-Practice-Related Hypotheses

These hypotheses were expected to hold both before and after practice.

Hypothesis 1a. Older adults' force structure will differ task-dependently from younger adults' (lower entropy and a greater proportion of low-frequency power when the task demands high entropy and reduced low-frequency power, and vice versa).

Hypothesis 2a. The tongue will produce less complex force than the lip (lower-entropy, greater dominance of low-frequency power).

Hypothesis 2b. The effects of age group and effector on entropy will interact.

Initial task performance. All measurements analyzed in this section are taken from the first trial of each condition (effector x task) on day 1. This single trial per condition was chosen for analysis, rather than a mean of the five trials per initial block of each condition, to capture performance when the task was as nearly novel to the participant as possible (since change in performance even within the first block was expected); to provide the baseline measurement against which immediate adaptability (trial 2 – trial 1) could be compared; and because if these results are to be applicable to clinical populations particularly in acute care, significance must be detectable based on very few trials in light of clinical participants' expected decreased stamina.

Temporal structure. These linear mixed effects analyses separately evaluated approximate entropy (ApEn) and fuzzy measure entropy (FuzzyMEn), for which lower values indicate more regular, predictable temporal structure of force. The purpose was to determine how age group, effector and task affect force temporal structure during initial attempts at unfamiliar tasks. See Figure 26, Figure 27, and Statistical Analysis for model details.

Task and age group interacted (ApEn: $F(2, 205) = 9.555$; FuzzyMEn: $F(2, 205) = 9.515$; both $p < 0.0005$). Follow-up analysis showed that only younger adults altered entropy across task, (ApEn: $F(2, 100) = 17.173$; FuzzyMEn, $F(2, 100) = 20.492$, both $p < 0.0005$). They showed significantly reduced entropy for the sine task vs. both others (vs. constant, both measures $p < 0.0005$; vs. multicosine, $p = 0.002$ for ApEn, $p = 0.005$ for FuzzyMEn) and perhaps higher entropy for the constant task vs. multicosine ($p = 0.006$ for ApEn, $p = 0.067$ (NS) for FuzzyMEn). Younger adults produced higher-entropy force than older adults only on the constant task (ApEn: $F(1, 41) = 9.407$, $p = 0.004$; FuzzyMEn, $F(1, 41) = 10.297$, $p = 0.003$).

Effector and age group interacted (ApEn: $F(1, 205) = 10.806$, $p < 0.0005$; FuzzyMEn, $F(1, 205) = 9.769$, $p = 0.002$). Follow-up analysis showed entropy higher for the tongue for older adults only (ApEn: $F(1, 105) = 23.591$; FuzzyMEn, $F(1, 105) = 20.794$; both $p < 0.0005$). For the lip only, younger adults had higher entropy (ApEn: $F(1, 41) = 7.212$, $p = 0.010$; FuzzyMEn, $F(1, 41) = 7.562$, $p = 0.009$).

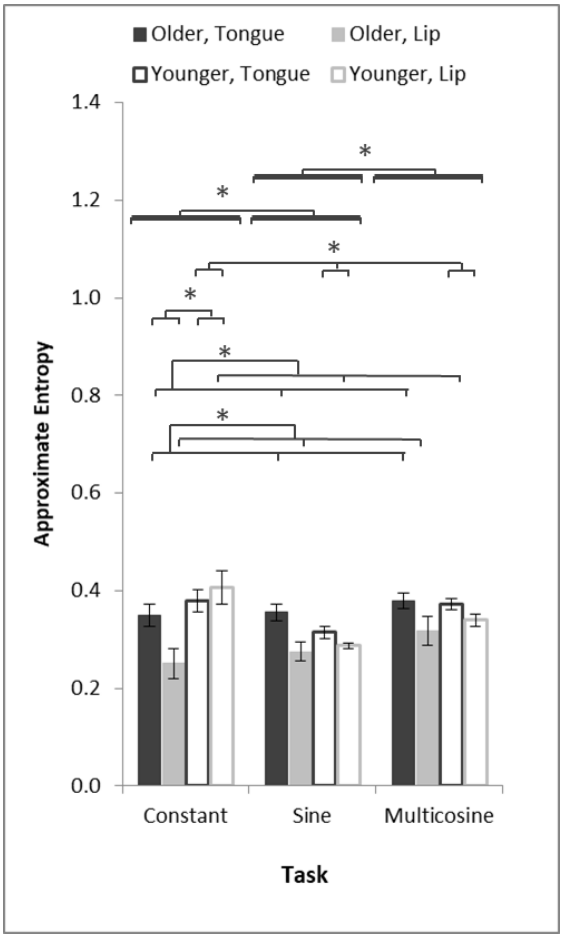

Figure 26. Initial complexity by age group, effector and task (ApEn, m = 2, r = 0.2, N = 1100; M ± SE): day 1, trial 1.
* Significant at $p < 0.01$.

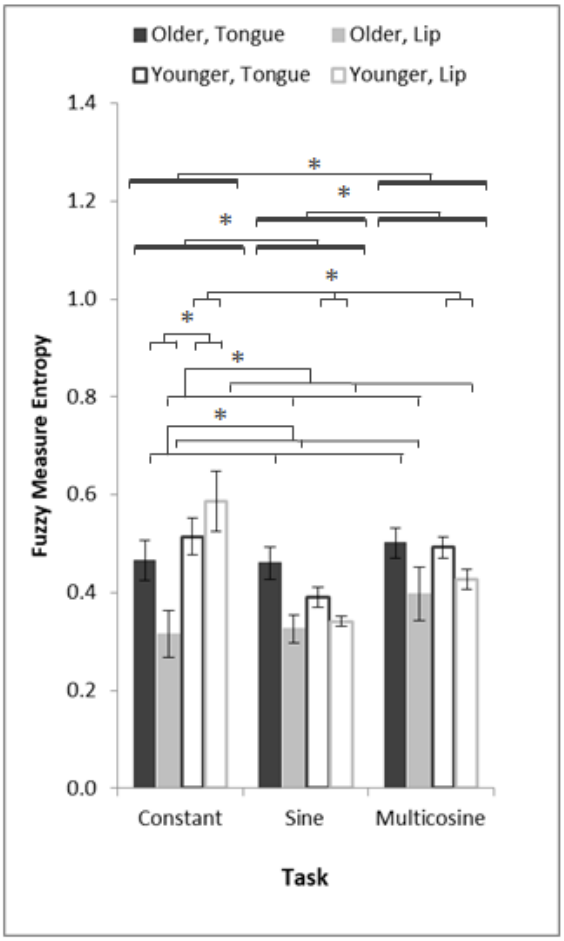

Figure 27. Initial complexity by age group, effector and task (FuzzyMEn, m = 2, r = 0.2, N = 1100; M ± SE): day 1, trial 1.
* Significant at $p < 0.01$.

For best comparability to the results of (Holtrop et al., 2014), this analysis was repeated using only the data from the constant task. Results were the same. Effector and age group again interacted (ApEn: $F(1, 41) = 6.625$, $p < 0.014$; FuzzyMEn, $F(1, 41) = 6.828$, $p = 0.012$). For older adults only, entropy was significantly higher for the tongue (ApEn: $F(1, 21) = 7.282$, $p = 0.013$; FuzzyMEn, $F(1, 21) = 6.227$, $p = 0.021$). For the lip only, younger adults showed higher entropy (ApEn: $t(39) = -3.344$, $p = 0.003$; FuzzyMEn, $t(39) = -3.528$, $p = 0.001$).

Frequency structure. This analysis evaluated proportion of power (PoP) in 1-Hz-wide frequency bands from 0 to 3 Hz; measurement values range from 0 to 1 with a higher value indicating a greater proportion of the total power in the specified band. The purpose of this analysis was to describe how

age group, effector and task affected the proportional distribution of power across the different frequency bands during participants' initial attempts at the tasks on day 1. Recall that the sine and multicosine tasks' targeted fundamental frequencies are 0.75 Hz; the multicosine target has additional, successively smaller components at 1.5 and 3 Hz. See Statistical Analysis for model details and cautions.

When the four-way model was run on all 41 participants, the distribution of studentized residuals for one effector x task x frequency band condition had skew > 2, due to a single participant's outlying value. When this participant (an older man) was excluded and the model re-run, all cells' studentized residual distributions had skew within acceptable limits, and the pattern of significant results did not differ from the original analysis reported here.

Due to the large number of comparisons necessitated for follow-up of multiple significant interactions, test statistics are reported in Table 20 (following Figure 30). Accompanying text summarizes and interprets table results, following the order of presentation in the table, without repeating the numbers.

Within every combination of age group, task and effector, differences between frequency bands followed the same pattern: greatest proportion of power in the 0-1 Hz band, followed by the 1-2 Hz band, followed by the 2-3 Hz band, with every pairwise comparison significant at $p < 0.0005$ (omitted from the table). See all figures in this section; significance is marked for this contrast only in Figure 28.

Across effector and within each task, younger and older adults' frequency band profiles did not significantly differ; see Figure 28 for an example. Across effectors, both younger and older adults responded to the increased high-frequency content of the multicosine target compared to the sine target by decreasing power in the 0-1 Hz band and increasing it in the 1-2 Hz band. Only the younger adults were able to decrease power in the 0-1 Hz band for the constant target compared to the sine target, congruent with the difference in task demand. For all of these effects, see Figure 29. Across tasks and age groups, the lip produced greater power in the 0-1 Hz band than the tongue, while the tongue produced greater power than the lip in the higher bands (Figure 30).

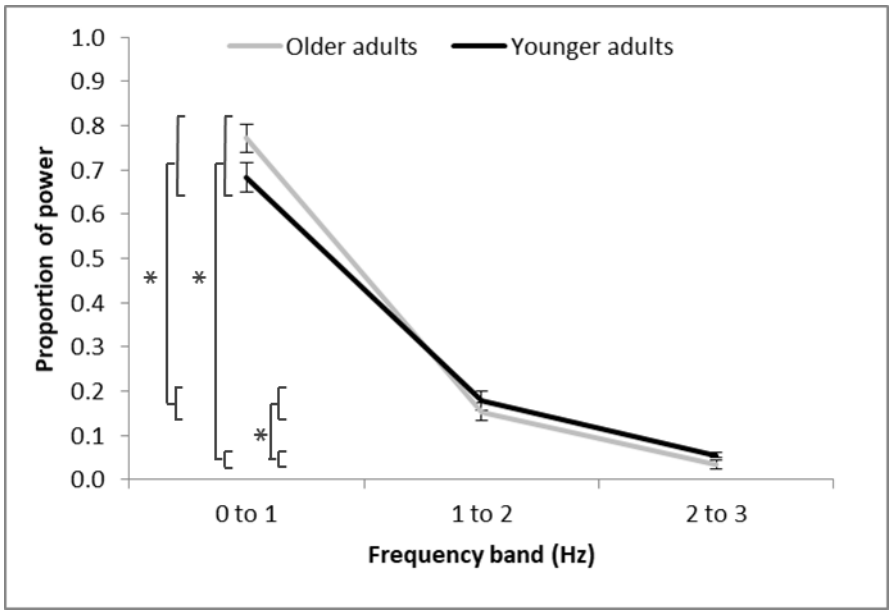

Figure 28. Initial proportion of power by frequency band and age group for the constant task performed with the lip: day 1, trial 1.
* Significant at $p < 0.0005$.

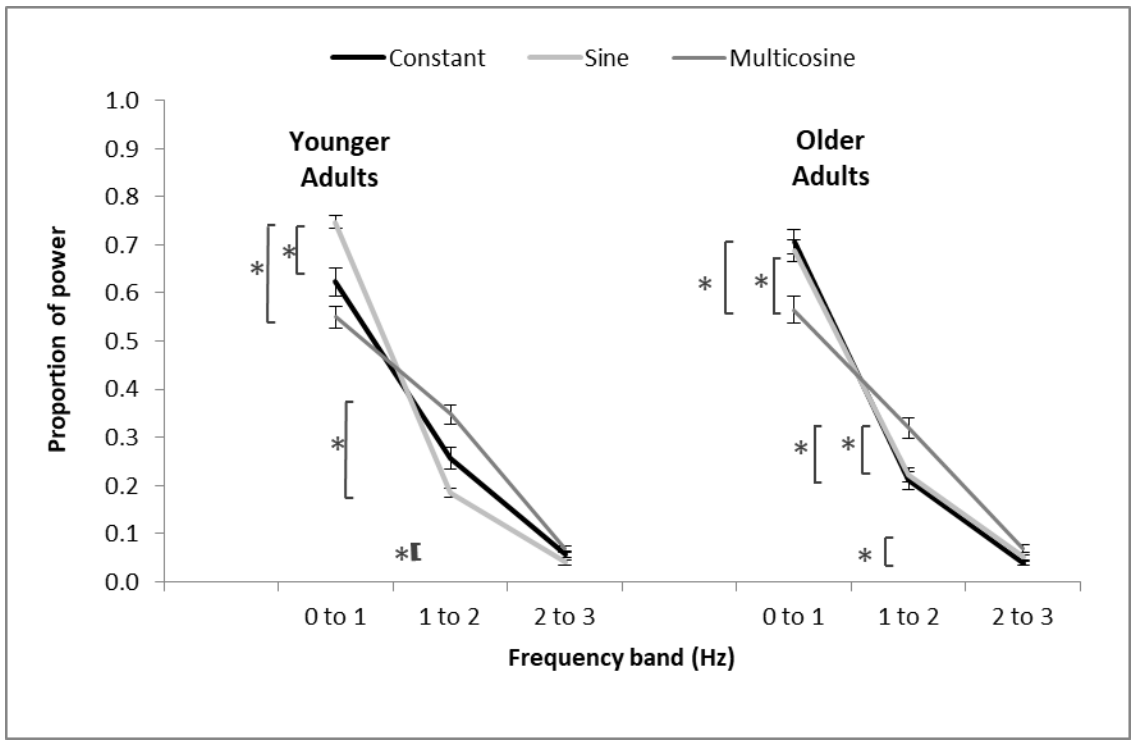

Figure 29. Initial proportion of power by frequency band, task and age group: day 1, trial 1.
* Significant at $p < 0.01$.

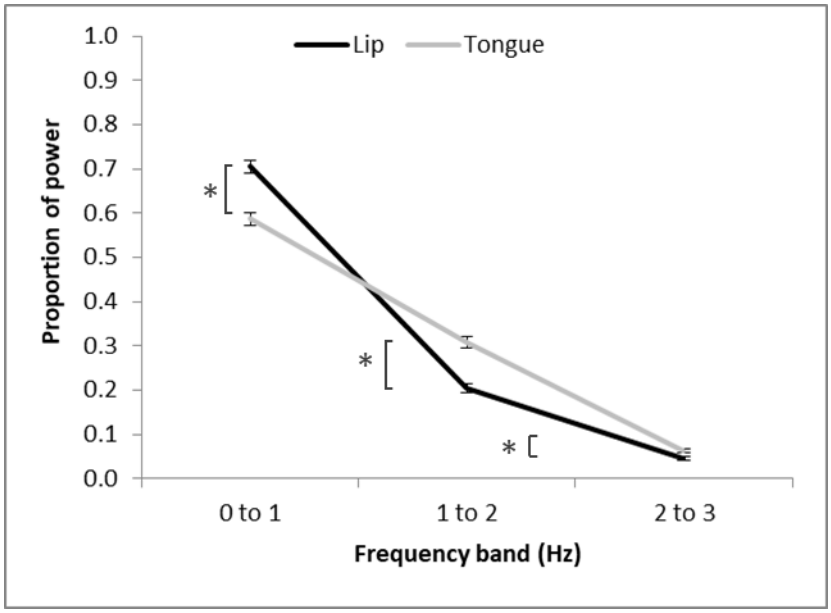

Figure 30. Initial proportion of power by frequency band and effector, across task and age group: day 1, trial 1.
* Significant at $p < 0.01$.

Table 20

Proportion of Power by Age Group, Effector, Task and Frequency Band on Day 1, Block 1, Trial 1

Term evaluated and *level of evaluation*			Test statistic[a]	Significance[b]
2-way interaction (effector x frequency band) [b]			$F(1.141, 44.509) = 41.112$	$p < 0.0005$*
Simple main effect of effector	*band: 0-1 Hz* [c] (L > T)		$F(1, 39) = 37.718$	$p < 0.0005$*
	band: 1-2 Hz [c] (T > L)		$F(1, 39) = 50.454$	$p < 0.0005$*
	band: 2-3 Hz [c] (T > L)		$F(1, 39) = 8.139$	$p = 0.007$*
Simple main effect of frequency band	*effector = L* [d]		$F(1.123, 43.779) = 686.898$	$p < 0.0005$*
	effector = T [d]		$F(1.150, 44.832) = 350.429$	$p < 0.0005$*
3-way interaction: age group x task x frequency band [b]			$F(1.999, 77.953) = 4.813$	$p = 0.011$*
Simple 2-way interaction: age group x frequency band	*task = C* [c]		$F(1.063, 41.448) = 3.988$	$p = 0.050$
	task = S [c]		$F(1.111, 43.334) = 5.041$	$p = 0.027$
	task = M [c]		$F(1.190, 46.428) = 0.436$	$p = 0.546$
Simple 2-way interaction: task x frequency band	*age group = OA* [d]		$F(2.092, 41.835) = 13.213$	$p < 0.0005$*
	simple simple main effect of freq. band	*task = C* [e]	$F(1.049, 20.974) = 247.061$	$p < 0.0005$*
		task = S [e]	$F(1.077, 21.547) = 308.021$	$p < 0.0005$*
		task = M [e]	$F(1.136, 22.727) = 94.754$	$p < 0.0005$*
	simple simple main effect of task	*band: 0-1 Hz* [e]	$F(2, 40) = 13.834$	$p < 0.0005$*
		C = S [f]		$p = 1.000$
		C > M [f]		$p < 0.0005$*
		S > M [f]		$p = 0.002$*

Table 20 (cont.)

Simple 2-way interaction: task x frequency band (cont.)	age group = OA (cont.)			
	simple simple main effect of task (cont.)	*band: 1-2 Hz* [e]	$F(2, 40) = 12.318$	$p < 0.0005$*
		C = S [f]		$p = 1.000$
		C < M [f]		$p = 0.003$*
		S < M [f]		$p = 0.001$*
		band: 2-3 Hz [e]	$F(2, 40) = 9.698$	$p < 0.0005$*
		C = S [f]		$p = 0.165$
		C < M [f]		$p = 0.001$*
		S = M [f]		$p = 0.103$
	age group = YA [d]		$F(1.825, 34.680) = 22.212$	$p < 0.0005$*
	simple simple main effect of freq. band	*task = C* [e]	$F(1.072, 20.376) = 122.158$	$p < 0.0005$*
		task = S [e]	$F(1.200, 22.795) = 1015.582$	$p < 0.0005$*
		task = M [e]	$F(1.247, 23.701) = 120.920$	$p < 0.0005$*
	simple simple main effect of task	*band: 0-1 Hz* [e]	$F(2, 38) = 23.603$	$p < 0.0005$*
		C < S [f]		$p < 0.0005$*
		C = M [f]		$p = 0.145$
		S > M [f]		$p < 0.0005$*
		band: 1-2 Hz [e]	$F(2, 38) = 21.477$	$p < 0.0005$*
		C = S [f]		$p = 0.005$
		C = M [f]		$p = 0.021$
		S < M [f]		$p < 0.0005$*
		band: 2-3 Hz [e]	$F(2, 38) = 7.389$	$p = 0.002$*
		C = S [f]		$p = 0.081$
		C = M [f]		$p = 0.377$
		S = M [f]		$p = 0.006$

Note. Age groups: older adults (OA), younger adults (YA). Effectors: lip (L), tongue (T). Tasks: constant (C), sine (S), multicosine (M). Formatting shows the structure of follow-up analyses: terms are evaluated within italicized levels to the right of the term; if a term is significant at a given level, the next-simplest terms are listed below it with further indentation.

[a] For interaction and main effects, F-statistic. For pairwise comparisons, mean difference ± standard error, [95% confidence interval] – significant if 95% CI does not include 0.
[b] For interactions in the full model, the standard for significance is $p < 0.05$. All follow-up analyses' p criteria were Bonferroni-corrected; see analysis-specific footnotes for the values used and Appendix F for their derivation. [c] Significance criterion is $p < 0.017$. [d] Significance criterion is $p < 0.025$.
[e] Significance criterion is $p < 0.008$. [f] Significance criterion is $p < 0.003$. * Meets the criterion for statistical significance.

Final task performance. This section examines variables' values for the first retention trial in each effector x task condition on day 3 – the final performance of the unmodified practiced tasks.

Temporal structure. These analyses separately evaluated approximate entropy (ApEn) and fuzzy measure entropy (FuzzyMEn), for which lower values indicate more regular, predictable temporal structure of force. The purpose was to determine how age group, effector and task affect force entropy on the first day-3 retention trial in each effector x task condition after two days' practice of unfamiliar tasks. See Figure 31 and Statistical Analysis for model details. Due to the large number of comparisons necessitated for follow-up of a significant three-way interaction, test statistics are reported in Table 21. Accompanying text summarizes and interprets table results, following the order of presentation in the table, without repeating the numbers.

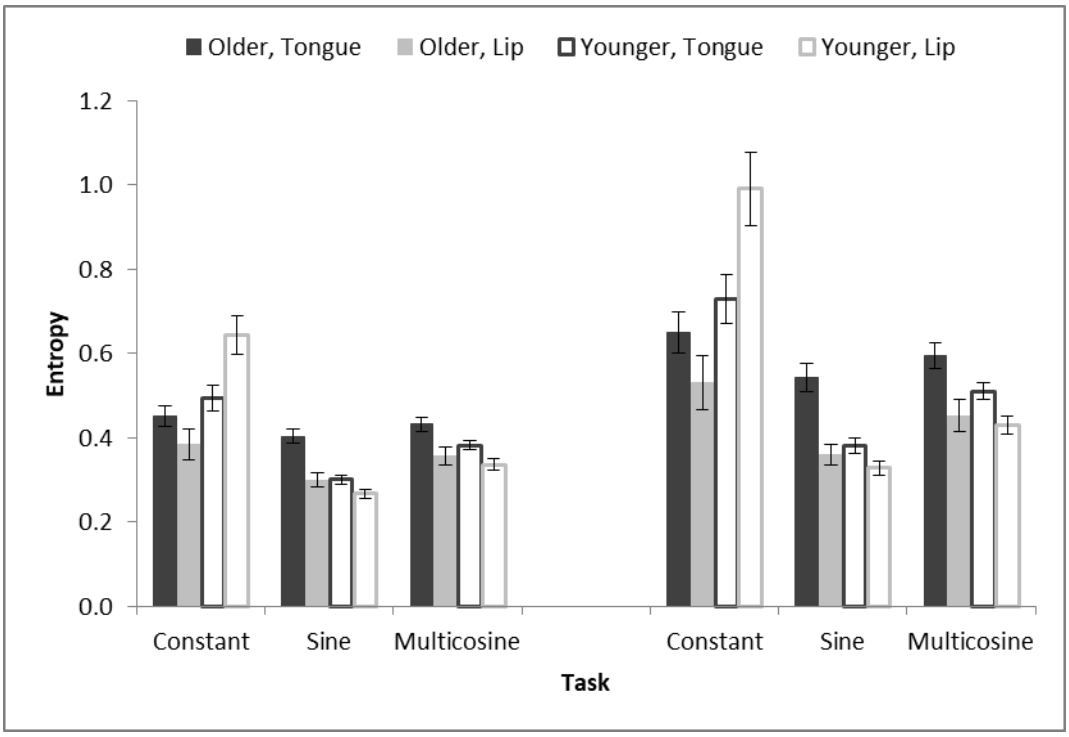

Figure 31. Final complexity by age group, effector & task (ApEn on the left and FuzzyMEn on the right, each m = 2, r = 0.2, N = 1100; M ± SE): day 3, first retention trial. Significance not indicated; see text and tables.

Averaged across task, older adults produced higher-entropy force with the tongue than the lip. Younger adults' force production showed higher entropy with the lip for the constant task (which demands high-entropy force) and slightly lower entropy with the lip for the more structured variable tasks (significant only for multicosine), suggesting that adaptation to task demand with practice may have been better with the lip.

All participants had force entropy ordered by task (constant > multicosine > sine), but for older adults only the constant > sine difference was significant. For younger adults, the constant > multicosine

and constant > sine differences were significant for both effectors, while the multicosine > sine difference was significant only for the tongue.

Younger adults' entropy was higher than older adults' in the *lip x constant force* condition and lower in the *tongue x sine* condition; in both cases, the contrast suggests younger adults adapted structure of output to task demand more closely than the older adults.

Table 21

Approximate Entropy (AE) and Fuzzy Measure Entropy (FZ) on Day 3 Retention Trial 1

Term evaluated and *level of evaluation*				Test statistic[a]	Significance[b]
3-way interaction (age group x effector x task) [b]			AE	$F(2, 205) = 5.053$	$p = 0.007$*
			FZ	$F(2, 205) = 4.537$	$p = 0.012$*
Simple 2-way interaction: effector x task	*age group = OA* [c]		AE, FZ both NS for interaction; simple main effects significant		
	simple main effect: effector [c] (T > L)		AE	$F(1, 105) = 22.708$	$p < 0.0005$*
			FZ	$F(1, 105) = 22.767$	$p < 0.0005$*
	simple main effect: task [c]		AE	$F(2, 105) = 5.059$	$p = 0.008$*
			FZ	$F(2, 105) = 6.777$	$p = 0.002$*
		C > S [d]	AE		$p = 0.007$*
			FZ		$p = 0.001$*
	age group = YA [c]		AE	$F(2, 100) = 12.185$	$p < 0.0005$*
			FZ	$F(2, 100) = 10.979$	$p < 0.0005$*
	simple simple main effect of effector	*task = C* [d] (L > T)	AE	$F(1, 20) = 11.950$	$p = 0.002$*
			FZ	$F(1, 20) = 10.768$	$p = 0.004$*
		task = M [d] (T > L)	AE	$F(1, 20) = 10.821$	$p = 0.004$*
			FZ	$F(1, 20) = 11.775$	$p = 0.003$*
	simple simple main effect of task	*effector = L* [c]	AE	$F(2, 40) = 60.379$	$p < 0.0005$*
			FZ	$F(2, 40) = 57.790$	$p < 0.0005$*
		C > S [e]	AE		$p < 0.0005$*
			FZ		$p < 0.0005$*
		C > M [e]	AE		$p < 0.0005$*
			FZ		$p < 0.0005$*
		effector = T [c]	AE	$F(2, 40) = 35.382$	$p < 0.0005$*
			FZ	$F(2, 40) = 29.771$	$p < 0.0005$*
		C > S [e]	AE		$p < 0.0005$*
			FZ		$p < 0.0005$*
		C > M [e]	AE		$p < 0.0005$*
			FZ		$p < 0.0005$*
		S < M [e]	AE		$p = 0.003$*
			FZ		$p = 0.023$

Table 21 (cont.)

Simple 2-way interaction: age group x task	*effector = L* [c]		AE	$F(2, 82) = 23.792$	$p < 0.0005$*
			FZ	$F(2, 82) = 21.761$	$p < 0.0005$*
	simple simple main effect of age group [e]	*task = C* (YA > OA)	AE	$t(39) = -4.339$	$p = 0.002$*
			FZ	$t(39) = -4.295$	$p = 0.001$*
		task = M	AE	$t(39) = 0.793$	$p = 0.424$
			FZ	$t(39) = 0.503$	$p = 0.624$
	effector = T [c]		AE	$F(2, 82) = 10.995$	$p < 0.0005$*
			FZ	$F(2, 82) = 8.330$	$p = 0.001$*
	simple simple main effect of age group [e]	*task = S* (OA > YA)	AE	$t(39) = 4.741$	$p = 0.001$*
			FZ	$t(39) = 4.191$	$p = 0.001$*

Note. Nonsignificant results are included only when AE and FZ results disagree. Age groups: older adults (OA), younger adults (YA). Effectors: lip (L), tongue (T). Tasks: constant (C), sine (S), multicosine (M). [a] For interactions and most main effects, F-statistic. For simple simple main effects of age group, independent-samples *t-test* with bootstrapping. For pairwise comparisons, SPSS provides a significance criterion but no test statistic. [b] For the three-way interaction, standard for significance is $p < 0.05$. All follow-up analyses' p criteria were Bonferroni-corrected; see analysis-specific footnotes for the values used and Appendix F for their derivation. [c] Significance criterion is $p < 0.025$. [d] Significance criterion is $p < 0.017$. [e] Significance criterion is $p < 0.008$. * Meets the criterion for statistical significance.

Frequency structure. This analysis evaluated proportion of power (PoP) in 1-Hz-wide frequency bands from 0 to 3 Hz; values range from 0 to 1 with a higher value indicating a greater proportion of the total power in the specified band. The purpose of this analysis was to describe how age group, effector and task affected the proportional distribution of power across the different frequency bands during day-3 retention trials (first trial in each effector x task condition) after two days' practice of unfamiliar tasks. Recall that the sine and multicosine tasks' targeted fundamental frequencies are 0.75 Hz; the multicosine target has additional, successively smaller components at 1.5 and 3 Hz.

When the four-way model was run on all 41 participants, the distributions of studentized residuals for three effector x task x frequency band conditions were excessively skewed. Because the direction of skew was not consistent, no transformation was appropriate. When participants with extreme residuals in those conditions (three older adults) were excluded and the model re-run, model fit was acceptable. The latter analysis is reported.

Due to the large number of comparisons necessitated for follow-up of multiple significant interactions, test statistics are reported in tables, one per significant three-way interaction. Accompanying text summarizes and interprets table results, following the order of presentation in the tables, without repeating the numbers.

Within every combination of age group, task and effector, differences between frequency bands followed the same pattern: greatest proportion of power in the 0-1 Hz band, followed by the 1-2 Hz band, followed by the 2-3 Hz band, with every pairwise comparison significant at $p < 0.0005$ (omitted from the table). See all figures in this section; significance for this effect is marked only on Figure 33.

The three-way interaction of age group, effector and frequency band was significant only when participants with extreme values were excluded (the sole difference between the models' results). For both age groups, proportion of power was greater in the lip than in the tongue from 0-1 Hz and greater in the tongue above 1 Hz. See Table 22 and Figure 32.

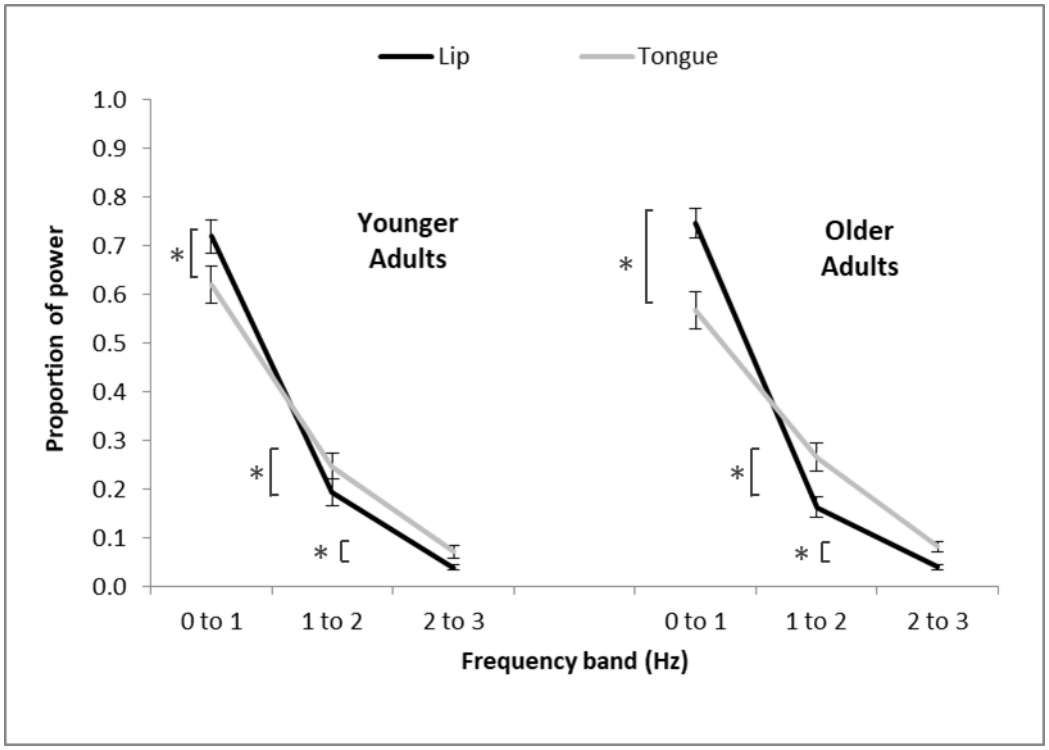

Figure 32. Proportion of power by age group and effector: day 3, first retention trial.
* Significant at $p \leq 0.005$.

Table 22

Proportion of Power by Age Group, Effector, Task and Frequency Band on Day 3, Retention trial 1, Part 1: Age Group x Effector x Frequency Band Interaction

Term evaluated and *level of evaluation*			Test statistic	Significance[a]
3-way interaction: age group x effector x frequency band [a]			$F(1.218, 43.839) = 4.561$	$p = 0.031$*
Simple 2-way interaction: age group x effector	*band: 0-1 Hz* [b]		$F(1, 36) = 5.326$	$p = 0.027$
	band: 1-2 Hz [b]		$F(1, 36) = 3.387$	$p = 0.074$
	band: 2-3 Hz [b]		$F(1, 36) = 1.096$	$p = 0.302$
Simple 2-way interaction: effector x frequency band	*age group = OA* [c]		$F(1.255, 21.335) = 35.097$	$p < 0.0005$*
	simple simple main effect of effector	*band: 0-1 Hz* [d]	$F(1, 17) = 38.516$ (L > T)	$p < 0.0005$*
		band: 1-2 Hz [d]	$F(1, 17) = 24.911$ (L < T)	$p < 0.0005$*
		band: 2-3 Hz [d]	$F(1, 17) = 25.818$ (L < T)	$p < 0.0005$*
	simple simple main effect of freq. band	*effector = L* [e]	$F(1.070, 18.187) = 482.900$	$p < 0.0005$*
		effector = T [e]	$F(1.316, 22.366) = 160.895$	$p < 0.0005$*
	age group = YA [c]		$F(1.147, 21.799) = 19.638$	$p < 0.0005$*
	simple simple main effect of effector	*band: 0-1 Hz* [d]	$F(1, 19) = 24.694$ (L > T)	$p < 0.0005$*
		band: 1-2 Hz [d]	$F(1, 19) = 9.978$ (L < T)	$p = 0.005$*
		band: 2-3 Hz [d]	$F(1, 19) = 34.090$ (L < T)	$p < 0.0005$*
	simple simple main effect of freq. band	*effector = L* [f]	$F(1.088, 20.664) = 651.685$	$p < 0.0005$*
		effector = T [f]	$F(1.375, 26.133) = 425.326$	$p < 0.0005$*
Simple 2-way interaction: age group x frequency band	*effector = L* [c]		$F(1.091, 39.273) = 1.797$	$p = 0.188$
	effector = T [c]		$F(1.339, 48.199) = 2.727$	$p < 0.094$

Note. Age groups: older adults (OA), younger adults (YA). Effectors: lip (L), tongue (T). Tasks: constant (C), sine (S), multicosine (M). Formatting shows the structure of follow-up analyses: terms are evaluated within italicized levels to the right of the term; if a term is significant at a given level, the next-simplest terms are listed below it with further indentation.

[a] For 3-way interactions, the standard for significance is $p < 0.05$. All follow-up analyses' p criteria were Bonferroni-corrected. See footnotes for the values used and Appendix F for their derivation.

[b] Significance criterion is $p < 0.017$. [c] Significance criterion is $p < 0.025$. [d] Significance criterion is $p < 0.008$. [e] Significance criterion is $p < 0.013$. * Meets the criterion for statistical significance.

The three-way interaction of effector, task and frequency band was significant (see Table 23 and Figures 33-35). For all tasks, proportion of power was greater in the lip than in the tongue from 0-1 Hz and greater in the tongue above 1 Hz.

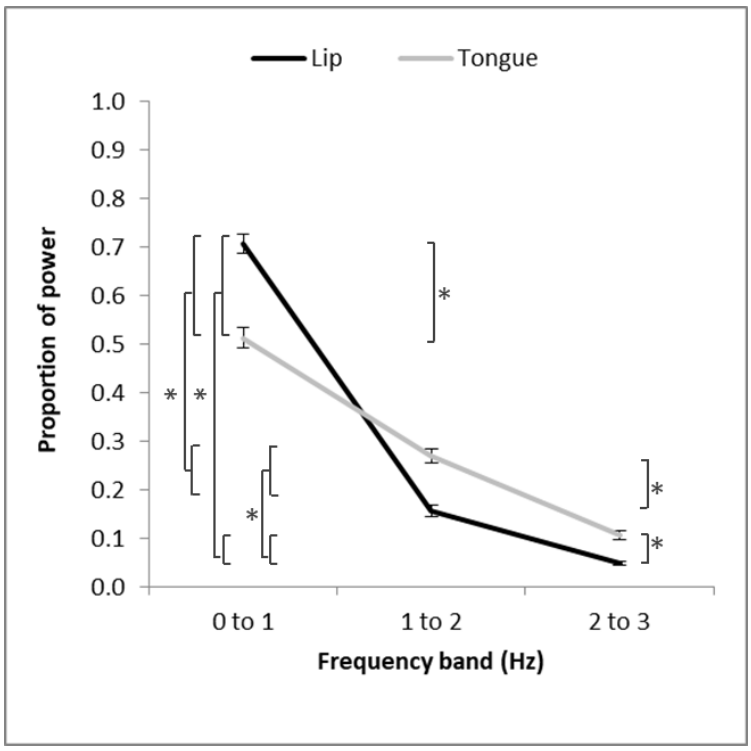

Figure 33. Proportion of power by frequency band and effector for the constant task: day 3, first retention trial.
* Significant at $p < 0.0005$.

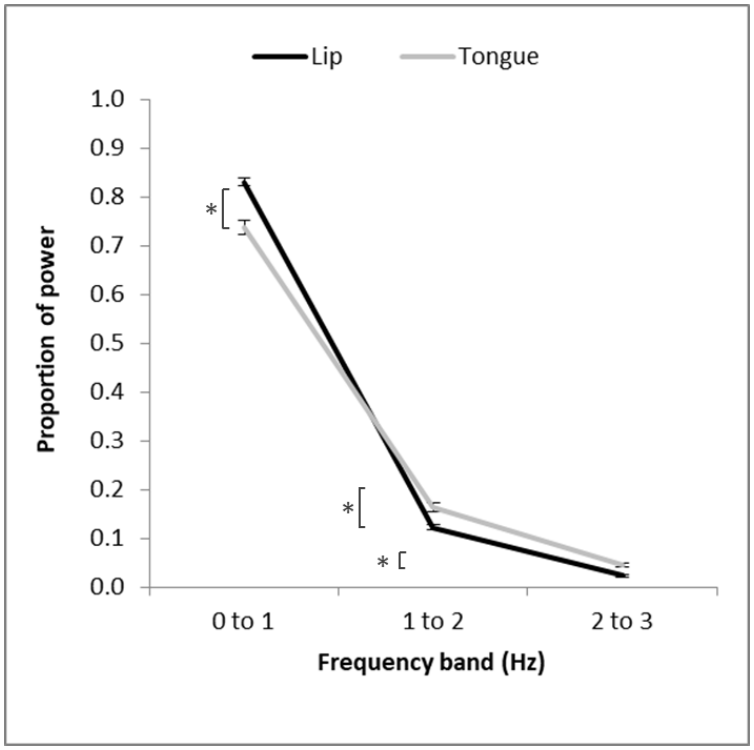

Figure 34. Proportion of power by frequency band and effector for the sine task: day 3, first retention trial. * Significant at $p < 0.0005$.

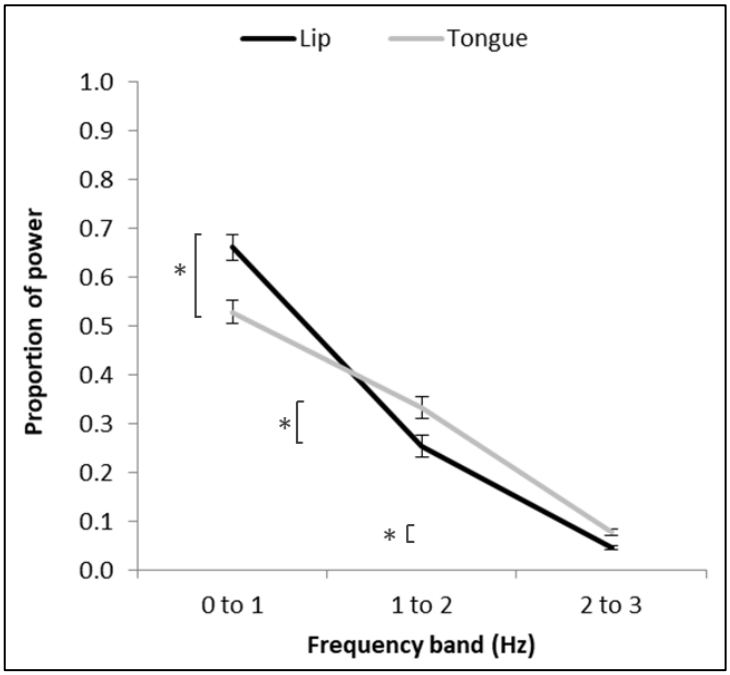

Figure 35. Proportion of power by frequency band and effector for the multicosine task: day 3, first retention trial. * Significant at $p \leq 0.002$.

Table 23

Proportion of Power by Age Group, Effector, Task and Frequency Band on Day 3, Retention trial 1, Part 2: *Effector x Task x Frequency Band Interaction*

Term evaluated and *level of evaluation*			Test statistic	Significance[a]
3-way interaction: effector x task x frequency band[b]			$F(2.3, 82.787) = 4.538$	$p = 0.010$*
Simple 2-way interaction: effector x task	*band: 0-1 Hz*[c]		$F(2, 72) = 4.054$	$p = 0.021$
	simple main effect of effector[c]		$F(1, 36) = 65.013$ (L > T)	$p < 0.0005$*
	simple main effect of task[c]		$F(1.511, 54.408) = 52.483$	$p < 0.0005$*
		C < S[d]		$p < 0.0005$*
		C = M[d]		$p = 1.000$
		S > M[d]		$p < 0.0005$*
	band: 1-2 Hz[c]		$F(2, 72) = 4.736$	$p = 0.012$*
	simple simple main effect of effector	*task = C*[e]	$F(1, 36) = 32.236$ (L < T)	$p < 0.0005$*
		task = S[e]	$F(1, 36) = 21.296$ (L < T)	$p < 0.0005$*
		task = M[e]	$F(1, 36) = 10.592$ (L < T)	$p = 0.002$*
	simple simple main effect of task	*effector = L*[f]	$F(1.492, 53.707) = 22.855$	$p < 0.0005$*
		C = S[d]		$p = 0.039$
		C < M[d]		$p < 0.0005$*
		S < M[d]		$p < 0.0005$*
		effector = T[f]	$F(1.462, 52.628) = 33.796$	$p < 0.0005$*
		C > S[d]		$p < 0.0005$*
		C = M[d]		$p = 0.053$
		S < M[d]		$p < 0.0005$*

Table 23 (cont.)

Simple 2-way interaction: effector x task (cont.)	*band: 2-3 Hz* [c]		*F*(2, 72) = 6.006	*p* = 0.004*
	simple simple main effect of effector	*task = C* [e]	*F*(1, 36) = 31.590 (L < T)	*p* < 0.0005*
		task = S [e]	*F*(1, 36) = 30.996 (L < T)	*p* < 0.0005*
		task = M [e]	*F*(1, 36) = 16.518 (L < T)	*p* < 0.0005*
	simple simple main effect of task	*effector = L* [f]	*F*(1.552, 55.889) = 19.867	*p* < 0.0005*
		C > S [d]		*p* < 0.0005*
		C = M [d]		*p* = 1.000
		S < M [d]		*p* < 0.0005*
		effector = T [f]	*F*(2, 72) = 19.801	*p* < 0.0005*
		C > S [d]		*p* < 0.0005*
		C = M [d]		*p* = 0.031
		S < M [d]		*p* = 0.001*
Simple 2-way interaction: task x frequency band	*effector = L* [g]		*F*(1.741, 62.680) = 22.918	*p* < 0.0005*
	simple simple main effect of freq. band	*task = C* [f]	*F*(1.158, 41.677) = 487.838	*p* < 0.0005*
		task = S [f]	*F*(1.068, 38.437) = 3463.659	*p* < 0.0005*
		task = M [f]	*F*(1.050, 37.811) = 169.110	*p* < 0.0005*
	effector = T [g]		*F*(2.037, 73.342) = 37.045	*p* < 0.0005*
	simple simple main effect of freq. band	*task = C* [f]	*F*(1.369, 49.296) = 119.087	*p* < 0.0005*
		task = S [f]	*F*(1.146, 41.252) = 1025.257	*p* < 0.0005*
		task = M [f]	*F*(1.185, 42.653) = 90.947	*p* < 0.0005*
Simple 2-way interaction: effector x frequency band	*task = C* [c]		*F*(1.188, 42.763) = 36.811	*p* < 0.0005*
	task = S [c]		*F*(1.121, 40.353) = 31.065	*p* < 0.0005*
	task = M [c]		*F*(1.207, 43.461) = 16.760	*p* < 0.0005*

Note. Age groups: older adults (OA), younger adults (YA). Effectors: lip (L), tongue (T). Tasks: constant (C), sine (S), multicosine (M). Formatting shows the structure of follow-up analyses: terms are evaluated within italicized levels to the right of the term; if a term is significant at a given level, the next-simplest terms are listed below it with further indentation.

[a] For interaction and main effects. For pairwise comparisons, SPSS provides a significance criterion but no test statistic. [b] For interactions in the full model, the standard for significance is $p < 0.05$. All follow-up analyses' *p* criteria were Bonferroni-corrected. See analysis-specific footnotes for the values used and Appendix F for their derivation. [c] Significance criterion is $p < 0.017$. [d] Significance criterion is $p < 0.003$. [e] Significance criterion is $p < 0.006$. [f] Significance criterion is $p < 0.008$. [g] Significance criterion is $p < 0.025$. * Meets the criterion for statistical significance.

The three-way interaction of age group, task and frequency band was significant (see Table 24 and Figures 36 and 37). For the sine task, younger adults' distribution of force matched the demands of the task more closely than older adults' across all frequency bands, with greater power than the older adults in the 0-1 Hz band and less power in the higher bands. In the 0-1 Hz band, younger and older adults had similar task differentiation: greater proportion of power for the sine task than for the other

tasks. In the 1-2 Hz band, older adults did not differentiate tasks, while younger adults had a clear separation of all three, with the greater proportion of power for the multicosine task, intermediate for the constant task, and least for the sine task, congruent with task demands. In the 2-3 Hz band, older adults had a greater proportion of power for the multicosine task than the sine task. Younger adults again differentiated all three tasks, with the greatest proportion of power for the constant task, intermediate for the multicosine task and least for the sine task, congruent with task demands.

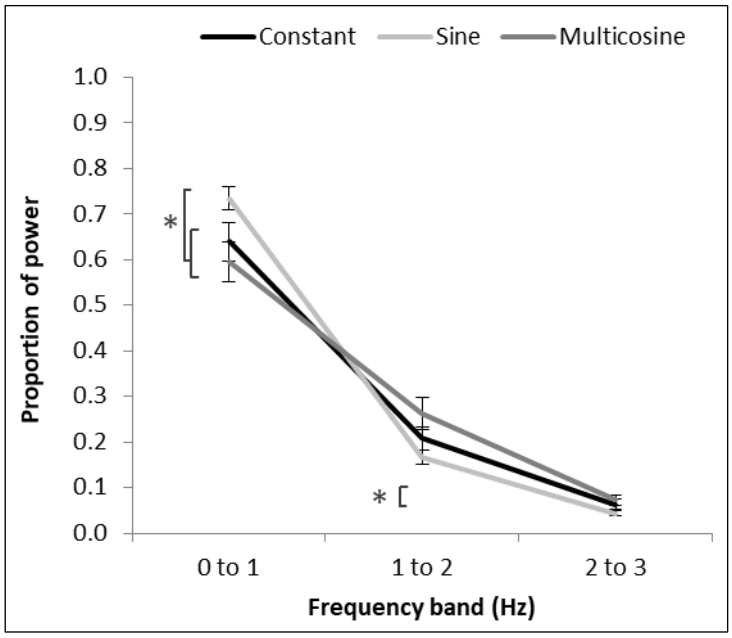

Figure 36. Proportion of power by frequency band and task for older adults: day 3, first retention trial. * Significant at $p < 0.002$.

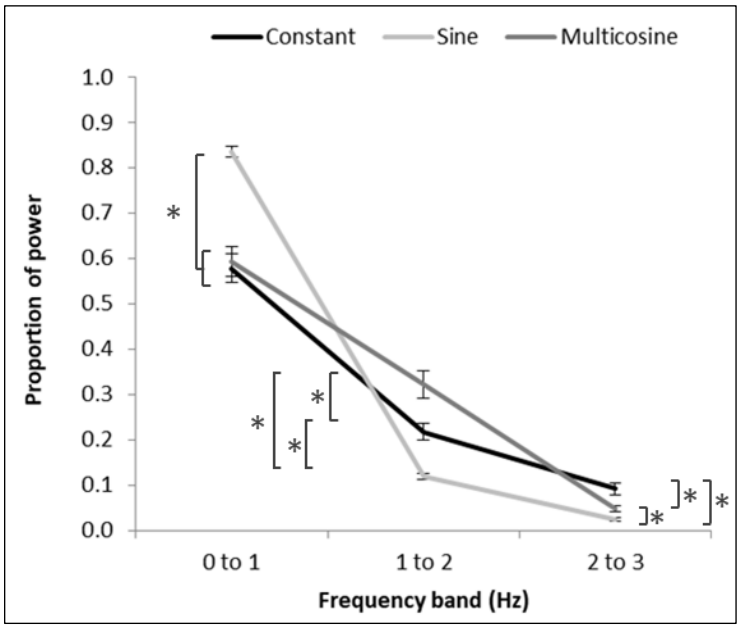

Figure 37. Proportion of power by frequency band and task for younger adults: day 3, first retention trial. * Significant at $p \leq 0.004$.

Table 24

Proportion of Power by Age Group, Effector, Task and Frequency Band on Day 3, Retention trial 1, Part 3: Age Group x Task x Frequency Band Interaction

Term evaluated and *level of evaluation*			Test statistic[a]	Significance[b]
3-way interaction: age group x task x frequency band [b]			$F(1.575, 56.690) = 7.107$	$p = 0.004$*
Simple 2-way interaction: age group x frequency band	*task = C* [c]		$F(1.437, 51.720) = 4.564$	$p = 0.025$
	task = S [c]		$F(1.134, 40.826) = 29.371$	$p < 0.0005$*
	task = M [c]		$F(1.072, 38.591) = 1.262$	$p = 0.272$
Simple 2-way interaction: age group x task	*band: 0-1 Hz* [c]		$F(1.511, 54.408) = 7.710$	$p = 0.003$*
	simple simple main effect of age group	*task = C* [d]	$F(1, 36) = 5.339$	$p = 0.027$
		task = S [d]	$F(1, 36) = 35.810$ (YA > OA)	$p < 0.0005$*
		task = M [d]	$F(1, 36) = 0.002$	$p = 0.963$
	band: 1-2 Hz [c]		$F(1.265, 45.526) = 5.462$	$p = 0.017$*
	simple simple main effect of age group	*task = C* [d]	$F(1, 36) = 0.309$	$p = 0.582$
		task = S [d]	$F(1, 36) = 16.119$ (OA > YA)	$p < 0.0005$*
		task = M [d]	$F(1, 36) = 0.2.595$	$p = 0.116$
	band: 2-3 Hz [c]		$F(2, 72) = 14.840$	$p < 0.0005$*
	simple simple main effect of age group	*task = C* [d]	$F(1, 36) = 7.533$	$p = 0.009$
		task = S [d]	$F(1, 36) = 22.135$ (OA > YA)	$p < 0.0005$*
		task = M [d]	$F(1, 36) = 9.476$ (OA > YA)	$p = 0.004$*

Table 24 (cont.)

Simple 2-way interaction: task x frequency band	*age group = OA* [e]		*F*(1.670, 28.382) = 9.567	*p* = 0.001*
	simple simple main effect of task	*band: 0-1 Hz* [f]	*F*(2, 34) = 11.003	*p* < 0.0005*
		C < S [g]		*p* = 0.002*
		C = M [g]		*p* = 0.655
		S > M [g]		*p* = 0.002*
		band: 1-2 Hz [f]	*F*(1.300, 22.092) = 7.156	*p* = 0.009
		band: 2-3 Hz [f]	*F*(2, 34) = 8.770	*p* = 0.001*
		C = S [g]		*p* = 0.020
		C = M [g]		*p* = 0.700
		S < M [g]		*p* = 0.002*
	simple simple main effect of freq. band	*task = C* [f]	*F*(1.200, 20.402) = 429.487	*p* < 0.0005*
		task = S [f]	*F*(1.139, 19.363) = 702.816	*p* < 0.0005*
		task = M [f]	*F*(1.081, 18.377) = 76.891	*p* < 0.0005*
	age group = YA [e]		*F*(1.408, 26.756) = 48.356	*p* < 0.0005*
	simple simple main effect of task	*band: 0-1 Hz* [f]	*F*(1.305, 24.802) = 52.273	*p* < 0.0005*
		C < S [g]		*p* < 0.0005*
		C = M [g]		*p* = 1.000
		S > M [g]		*p* < 0.0005*
		band: 1-2 Hz [f]	*F*(1.219, 23.154) = 42.963	*p* < 0.0005*
		C > S [g]		*p* < 0.0005*
		C < M [g]		*p* = 0.004*
		S < M [g]		*p* < 0.0005*
		band: 2-3 Hz [f]	*F*(2, 38) = 36.977	*p* < 0.0005*
		C > S [g]		*p* < 0.0005*
		C > M [g]		*p* < 0.0005*
		S < M [g]		*p* = 0.001*
	simple simple main effect of freq. band	task = C [f]	*F*(2, 38) = 238.628	*p* < 0.0005*
		task = S [f]	*F*(1.090, 20.712) = 5848.305	*p* < 0.0005*
		task = M [f]	*F*(1.061, 20.166) = 113.583	*p* < 0.0005*

Note. Age groups: older adults (OA), younger adults (YA). Effectors: lip (L), tongue (T). Tasks: constant (C), sine (S), multicosine (M). Formatting shows the structure of follow-up analyses: terms are evaluated within italicized levels to the right of the term; if a term is significant at a given level, the next-simplest terms are listed below it with further indentation.

[a] For interaction and main effects. For pairwise comparisons, SPSS provides a significance criterion but no test statistic. [b] For interactions in the full model, the standard for significance is $p < 0.05$. All follow-up analyses' *p* criteria were Bonferroni-corrected. See analysis-specific footnotes for the values used and Appendix F for their derivation. [c] Significance criterion is $p < 0.017$. [d] Significance criterion is $p < 0.006$. [e] Significance criterion is $p < 0.025$. [f] Significance criterion is $p < 0.008$. [g] Significance criterion is $p < 0.003$. * Meets the criterion for statistical significance.

Specific Aim 1: Practice-Related Hypotheses

These hypotheses related to changes expected with practice.

Hypothesis 1b. Adaptability (immediate): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing trial 2 to trial 1 on day 1 within each effector x task combination. All reported measurements were calculated as day 1: block 1: [trial 2 – trial 1] for each effector x task combination.

Temporal structure. These linear mixed effects models evaluated the effects of age group, effector and task on immediate change in entropy ($\Delta_{initial}$ApEn and $\Delta_{initial}$FuzzyMEn separately). A positive number indicates increase in entropy from trial 1 to trial 2, a negative number a decrease. See Statistical Analysis and Figure 38 for model details. For both measures, the full model had no statistically significant interaction terms; only two main effects reached significance.

Immediate change in entropy was greater for the lip than the tongue ($\Delta_{initial}$ApEn: $F(1, 205) = 6.133$, $p = 0.014$; $\Delta_{initial}$FuzzyMEn: $F(1, 205) = 7.436$, $p = 0.007$).

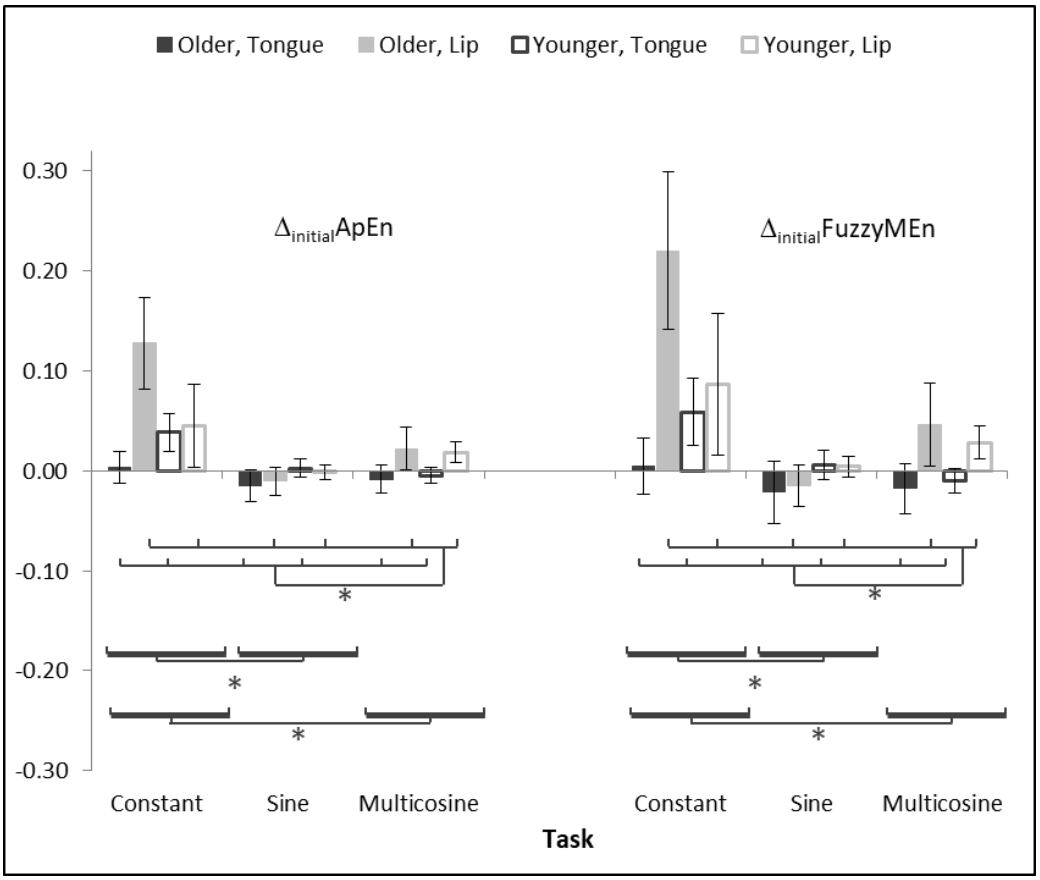

Figure 38. Immediate change (day 1, block 1, trial 2 - trial 1) in entropy by age group, effector and task (ApEn and FuzzyMEn, each m = 2, r = 0.2, N = 1100; M ± SE). * Significant at $p \leq 0.014$.

Immediate change in entropy differed by task ($\Delta_{initial}$ApEn: $F(2, 205) = 8.279$; $\Delta_{initial}$FuzzyMEn: $F(2, 205) = 8.030$; both $p < 0.0005$). Pairwise comparisons showed that immediate change was greater for

the constant task than for both other tasks (vs. sine: $\Delta_{initial}$ApEn, $p < 0.0005$; $\Delta_{initial}$FuzzyMEn, $p = 0.001$; vs. multicosine: $\Delta_{initial}$ApEn, $p = 0.008$; $\Delta_{initial}$FuzzyMEn, $p = 0.007$).

Proportion of power. The purpose of this analysis was to determine how age group, effector and task affect immediate changes in proportional distribution of power from 0-3 Hz from initial to second attempts at unfamiliar tasks. See Statistical Analysis for details. Change in proportion of power (PoP) was calculated for each of three frequency bands (0-1, 1-2, 2-3 Hz). A positive number indicates that PoP in the specified band increased from trial 1 to trial 2, a negative number that it decreased.

Mean change in PoP from trial 1 to trial 2 did not vary by frequency band, or by age group, effector, or task interacting with each frequency band (main effects and interactions all non-significant).

Hypothesis 1c. Adaptability (after practice): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing day 1 trial 1 to day 3 retention trial 1 within each effector x task combination.

Temporal structure. Change in complexity was assessed by subtracting entropy (ApEn or FuzzyMEn) on initial performance (day 1, block 1, trial 1 for each effector x task combination) from entropy on the first retention trial (Figure 39). A positive result indicated an increase in entropy.

There was equivocal evidence that entropy change with practice may depend upon an interaction of effector and task. For ApEn, the interaction was significant ($F(2, 205) = 4.307$, $p = 0.015$), but the simple main effect of effector did not reach significance within any task (closest for the constant task, $F(1, 41) = 4.696$, $p = 0.036$, with Bonferroni-adjusted criterion for significance 0.017). For FuzzyMEn, the interaction was significant in the full factorial model ($F(2, 205) = 3.061$, $p = 0.049$), but no longer so ($F(2, 205) = 2.951$, $p = 0.055$) when the non-significant three-way interaction was omitted.

Entropy change with practice depended upon an interaction of age group and task (ApEn: $F(2, 205) = 5.890$, $p = 0.003$; FuzzyMEn: $F(2, 205) = 4.950$, $p = 0.008$). ApEn task-specific submodels did not all support random subjects effects; independent-samples t-tests with bootstrapping and equal variances not assumed are reported instead. For the multicosine task, younger adults did not change entropy with practice, while older adults increased it (ApEn: $t(67.503) = 2.675$, $p = 0.014$; FuzzyMEn: $F(1, 41) = 4.559$, $p = 0.039$, NS). For the sine task, younger adults decreased entropy with practice, while older adults increased it (ApEn: $t(75.08) = 4.308$, $p = 0.001$; FuzzyMEn: $F(1, 41) = 9.657$, $p = 0.003$).

The simple main effects of task were significant, and in the same direction, for every level of effector and every level of age group; only the overall main effect and pairwise comparisons are reported. Entropy increased with practice to a greater degree for the constant task than for both

variable tasks (ApEn, $F(2, 205) = 34.918$; FuzzyMEn, $F(2, 205) = 32.121$; main effects and pairwise comparisons of constant to sine and multicosine, all $p < 0.0005$).

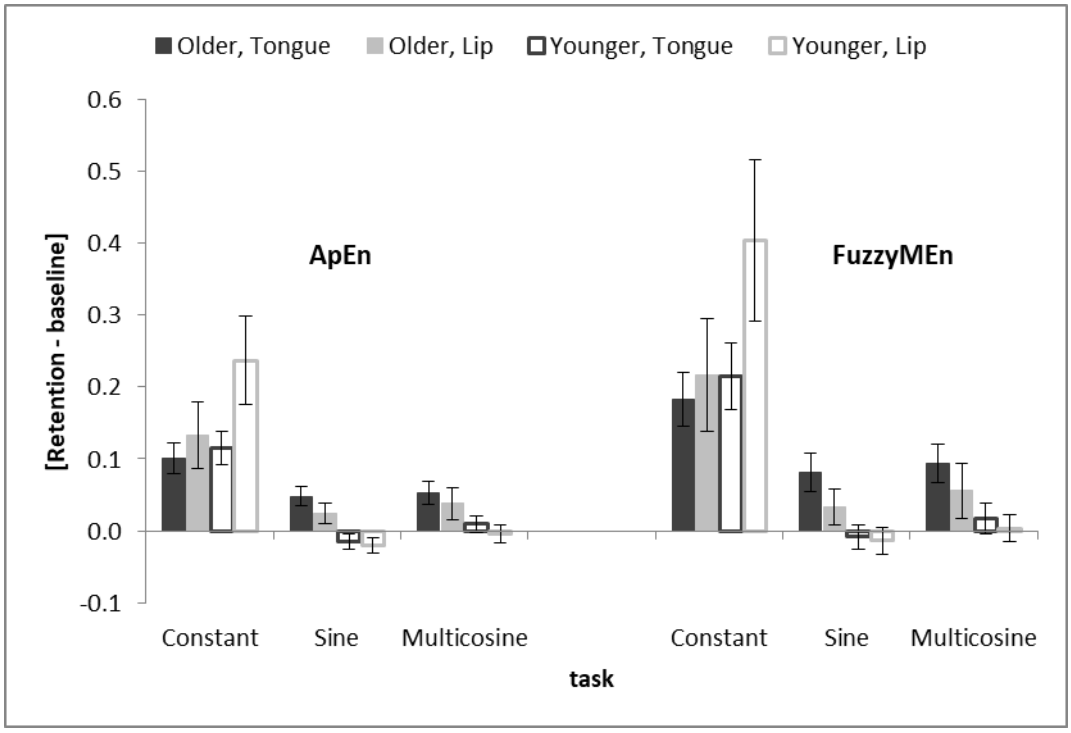

Figure 39. [Retention – baseline] approximate and fuzzy measure entropies (m = 2, r = 0.2, N = 1100) by age group, task & effector (M ± SE). Significance not indicated; see text and tables.

Frequency structure. The purpose of this analysis was to determine how spectral power from 0-3 Hz changed with practice, depending on effector used, task performed, and participant's age group. See Statistical Analysis for details. Change in proportion of power (PoP)was calculated for each of three frequency bands (0-1, 1-2, 2-3 Hz) by subtracting band-specific PoP on initial performance (day 1, block 1, trial 1 for each effector x task combination) from PoP on the first retention trial (Figure 40). A positive number indicates that PoP in the specified band increased, a negative number that it decreased.

Change in the distribution of spectral power across frequency bands with practice was significantly affected by task ($F(2.102, 81.978) = 10.681$, $p < 0.0005$). Follow-up analyses (see Table 25, following Figure 40) showed that the amount of change differed by frequency band for the constant and sine tasks. For the constant task, participants decreased PoP in the 0-1 Hz and 1-2 Hz bands and increased it in the 2-3 Hz band. See Figure 40. For the sine task, participants increased power in the 0-1 Hz band and decreased it in the 1-2 Hz band. The pattern for the multicosine task was similar to that for

the sine task, but because the increase at 0-1 Hz and decrease at 1-2 Hz were less extreme, the effect of frequency band did not reach significance.

Amount of change also differed by task within frequency band. In the 0-1 Hz band, PoP was decreased for the constant task and increased for the others, which did not differ from each other. In the 2-3 Hz band, PoP was increased for the constant task and unchanged for the others.

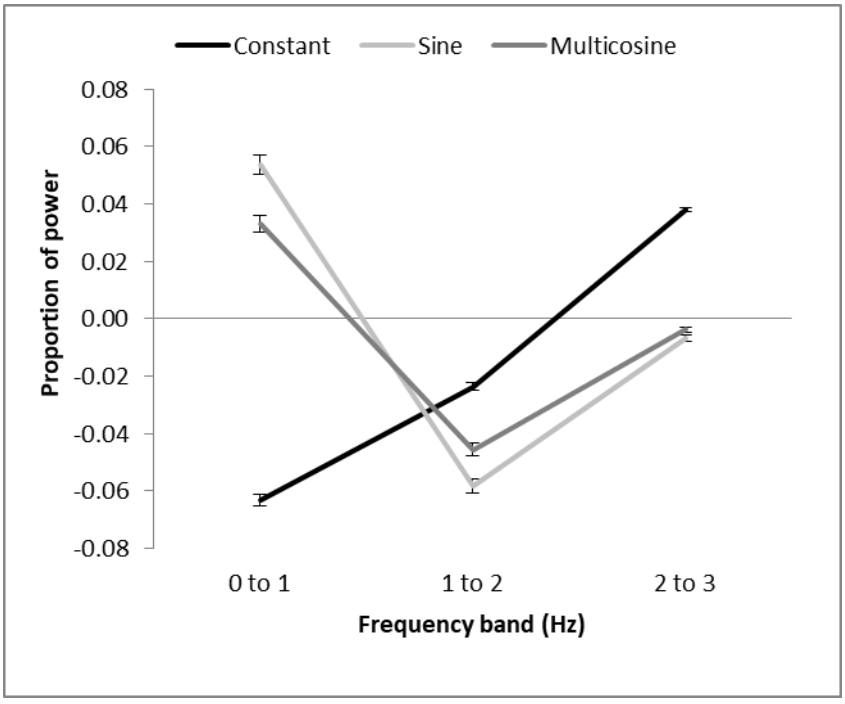

Figure 40. Change in proportion of power by frequency band and task from initial performance (day 1, trial 1 of each condition) to first retention trial (day 3). Significance not indicated; see text and tables.

Table 25

Change in Proportion of Power by Frequency Band from Day 1, Block 1, Trial 1 to Day 3, Retention Trial 1.

Term evaluated and *level of evaluation*		Test statistic [a]	Significance [b]
Simple main effect of frequency band	*task = C* [c]	*F*(1.313, 51.201) = 7.507	*p* = 0.005*
	0-1 = 1-2 Hz [d]		*p* = 0.760
	1-2 < 2-3 Hz [d]		*p* = 0.004*
	0-1 < 2-3 Hz [d]		*p* = 0.001*
	task = S [c]	*F*(1.231, 48.012) = 24.414	*p* < 0.0005*
	0-1 > 1-2 Hz [d]		*p* < 0.0005*
	1-2 < 2-3 Hz [d]		*p* < 0.0005*
	0-1 > 2-3 Hz [d]		*p* = 0.002*
	task = M [c]	*F*(1.136, 44.321) = 5.246	*p* = 0.023
Simple main effect of task	*band: 0-1 Hz* [c]	*F*(1.714, 66.838) = 14.054	*p* < 0.0005*
	C < S [d]		*p* < 0.0005*
	C < M [d]		*p* = 0.004*
	S = M [d]		*p* = 0.850
	band: 1-2 Hz [c]	*F*(1.710, 66.686) = 2.092	*p* = 0.138
	band: 2-3 Hz [c]	*F*(1.684, 65.677) = 23.073	*p* < 0.0005*
	C > S [d]		*p* < 0.0005*
	C > M [d]		*p* < 0.0005*
	S = M [d]		*p* = 1.0

Note. Tasks: constant (C), sine (S), multicosine (M). Formatting shows the structure of follow-up analyses: terms are evaluated within italicized levels to the right of the term; if a term is significant at a given level, the next-simplest terms are listed below it with further indentation.

[a] For interaction and main effects, F-statistic. For pairwise comparisons, SPSS provides a significance criterion but no test statistic. [b] For the task x frequency band interaction in the full model, the standard was $p < 0.05$. All follow-up analyses' *p* criteria were Bonferroni-corrected; see analysis-specific footnotes for the values used and Appendix F for their derivation. [c] Significance criterion is $p < 0.017$. [d] Significance criterion is $p < 0.008$. * Meets the criterion for statistical significance.

Hypothesis 1d. Older adults' reduction in error vs. baseline on retention and transfer trials after two days' practice will be less than younger adults'. This section characterizes initial accuracy, final accuracy, and learning by age group, effector and task.

Initial accuracy. This analysis used linear mixed effects modeling to evaluate normalized root mean square error (NRMSE), for which a lower value indicates more accurate performance, for the first trial of each effector x task combination on day 1. The purpose was to determine how age group, effector and task affect accuracy during initial attempts at unfamiliar tasks. See Figure 41 and Statistical Analysis for model details.

Model fitting to the original data yielded strongly positively skewed residuals (skew = 3.118). Residuals remained excessively skewed after data were square-root transformed (skew = 2.207). When a logarithmic transformation was used instead, both model residuals' and random effects' values of skewness were acceptable (1.649 and 1.013 respectively). Reported test statistics refer to the analysis of logarithm-transformed data, but the pattern of significance did not differ from the original analysis.

The only significant interaction in the final model occurred between age group and effector ($F(1, 205) = 9.459$, $p = 0.002$). Follow-up analysis of simple main effects showed that for younger adults only, NRMSE was greater for the tongue ($F(1, 100) = 21.991$, $p < 0.0005$), while for both effectors, older adults' NRMSE was greater than younger adults' (lip: $F(1, 41) = 22.010$, $p < 0.0005$; tongue: $F(1, 41) = 10.228$, $p = 0.003$).

Accuracy differed by task ($F(2, 205) = 31.182$, $p < 0.0005$). Pairwise comparison showed that NRMSE was lower on the constant task than the variable tasks (sine and multicosine, both $p < 0.0005$).

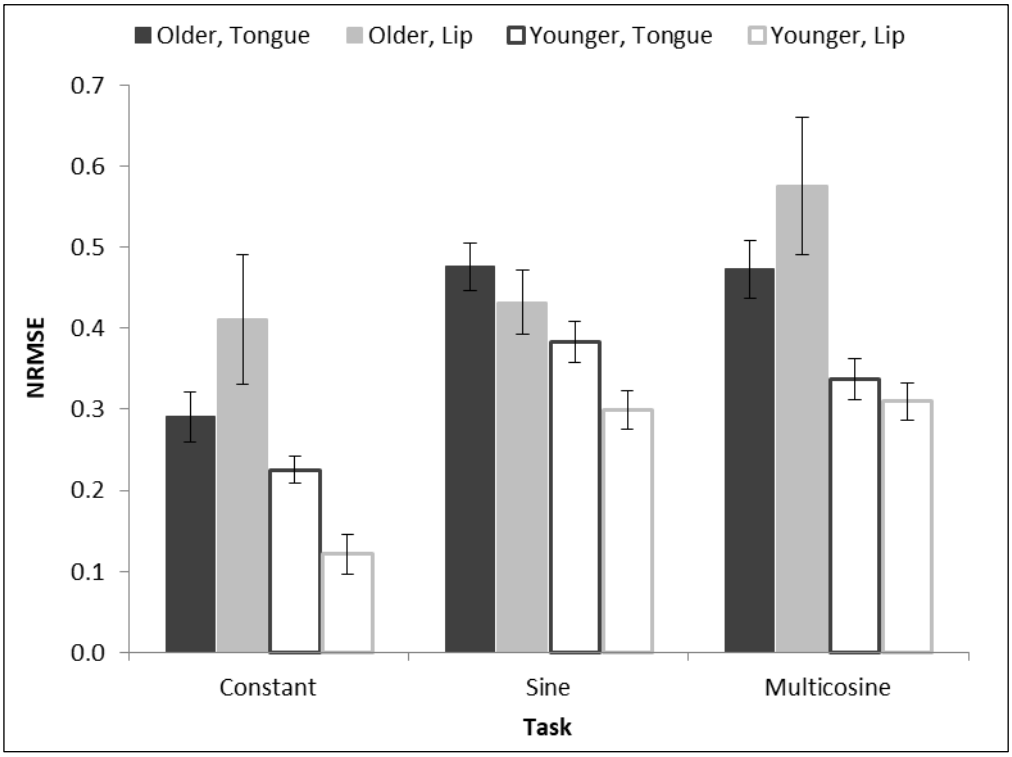

Figure 41. Initial performance normalized root mean square error (NRMSE) by age group, effector and task (M ± SE): day 1, trial 1. Significance not indicated; see text.

Final accuracy. This linear mixed model analysis evaluated normalized root mean square error (NRMSE), for which a lower value indicates more accurate performance. The purpose was to determine

how age group, effector and task affect accuracy on the first day-3 retention trial after two days' practice of unfamiliar tasks. See Figure 42 and Statistical Analysis for model details.

When the full model was fit to the original data, both level-1 and level-2 residuals had skew > 2. Only the former was corrected with a square-root transformation, while both were resolved with a logarithm transformation. Reported test statistics refer to the analysis of logarithm-transformed data. The pattern of significant results was the same as for the original data.

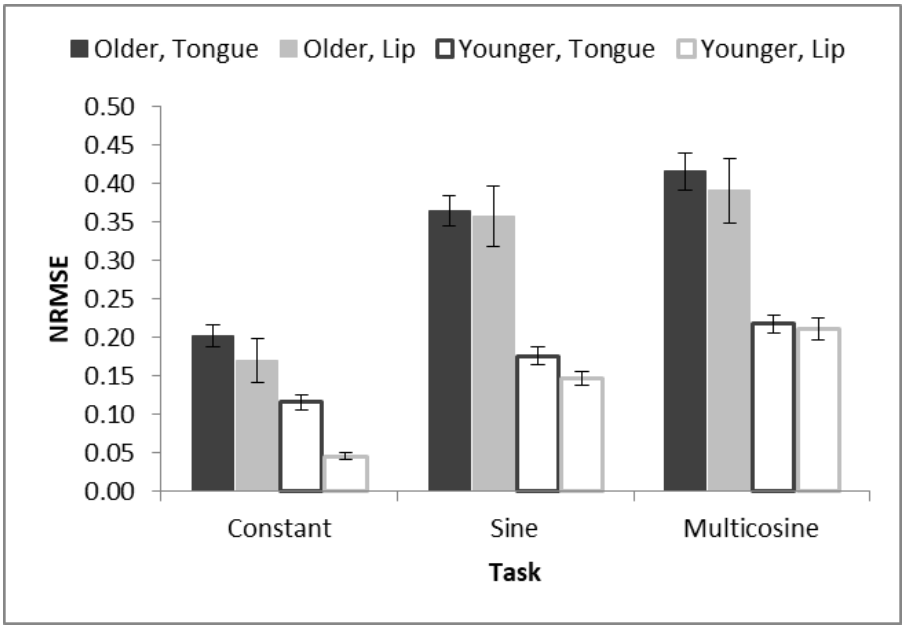

Figure 42. Normalized root mean square error (NRMSE) by age group, task & effector (M ± SE): day 3, first retention trial. Significance not indicated; see text.

The effects of age group and task interacted ($F(2,205) = 5.998$, $p = 0.003$). Follow-up analysis showed that older adults' NRMSE was higher than younger adults' for all tasks (constant: $F(1, 41) = 32.391$; sine: $F(1, 41) = 72.549$; multicosine, $F(1, 82) = 65.410$[16], all $p < 0.0005$).

Both age groups' NRMSE differed by task (older adults: $F(2, 105) = 53.352$; younger adults, $F(2, 100) = 94.339$; both $p < 0.0005$). Though participants in both age groups showed the same pattern, with NRMSE increasing from constant to sine to multicosine force targets, only the younger adults had each

[16] For task = multicosine only, model validity was questionable due to zero random subjects effects. A mixed-model RMANOVA analysis of age group and effector simple main effects for task = multicosine found the same age group difference, $F(1, 39) = 72.344$, $p < 0.0005$.

task significantly differentiated from both others (all $p < 0.0005$). For the older adults, the two variable tasks had higher NRMSE than constant force ($p < 0.0005$) but did not differ significantly from each other.

NRMSE was greater for the tongue than the lip ($F(1,205) = 11.180$, $p = 0.001$) by 0.012 ± 0.003.

Reduction in error following practice. The purpose of these analyses was to determine how age group, effector and task affect learning after two days' practice of unfamiliar tasks. Learning was assessed by subtracting NRMSE on initial performance (day 1, block 1, trial 1 for each effector x task combination) from the corresponding NRMSE on trial 1 of retention and transfer conditions. For retention trials, a negative number, i.e. reduction in error, indicated improvement in performance. For transfer trials, a negative number indicated application of learned skill to a related task. Linear mixed effects models were fit; see Statistical Analysis.

Retention trials. See Figure 43. When the full model was fit to the original data, residuals were negatively skewed. This issue was resolved with a *reflect + square root* transformation. The pattern of significance in the original-data model was the same as the transformed-data analysis reported here.

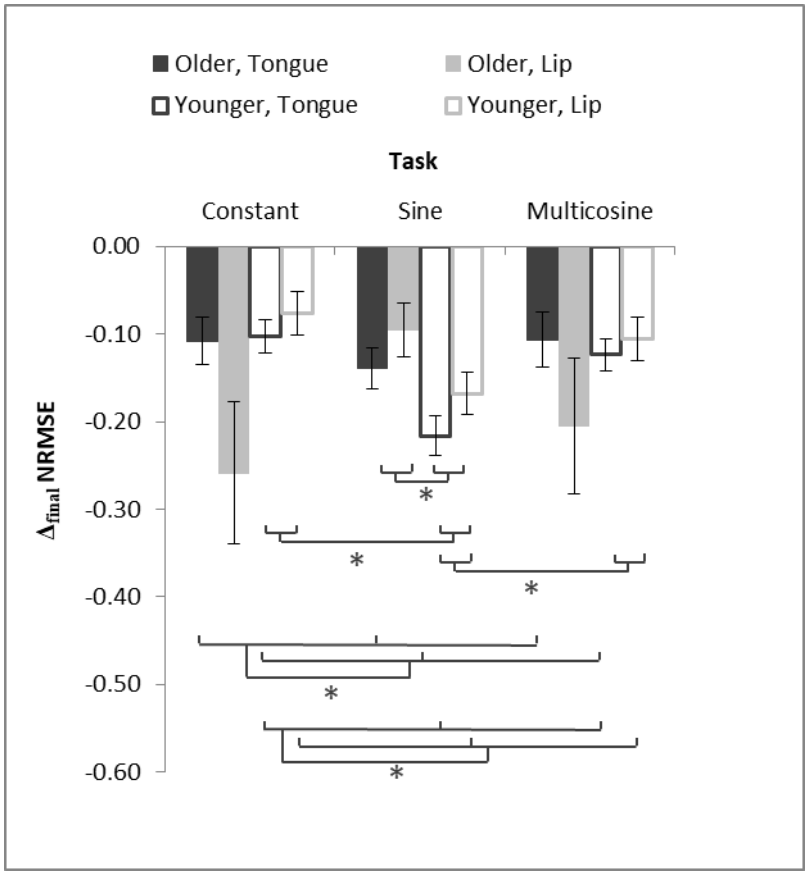

Figure 43. [Retention – baseline] normalized root mean square error (NRMSE) by age group, task & effector (M ± SE).
* Significant at $p \leq 0.024$.

Age group interacted significantly with both effector ($F(1, 205) = 5.702$, $p = 0.018$) and task ($F(2, 205) = 3.871$, $p = 0.022$).

Only younger adults' reduction in NRMSE differed by effector, with greater change for the tongue ($F(1, 100) = 5.276$, $p = 0.024$). Younger adults showed greater reduction in NRMSE than older adults only with the tongue ($F(1, 41) = 7.38$, $p = 0.009$).

Only younger adults' reduction in NRMSE differed by task ($F(2, 100) = 10.943$, $p < 0.0005$), with greater improvement on the sine task than both others ($p = 0.002$ vs. multicosine, $p < 0.0005$ vs. constant). Younger adults showed greater reduction in NRMSE than older adults only on the sine task ($F(1, 41) = 11.288$, $p = 0.002$).

Transfer trials. See Figure 44. Models were fit separately for the two transfer conditions (lower and higher target force) because of the difference in presence vs. absence of transfer seen previously.

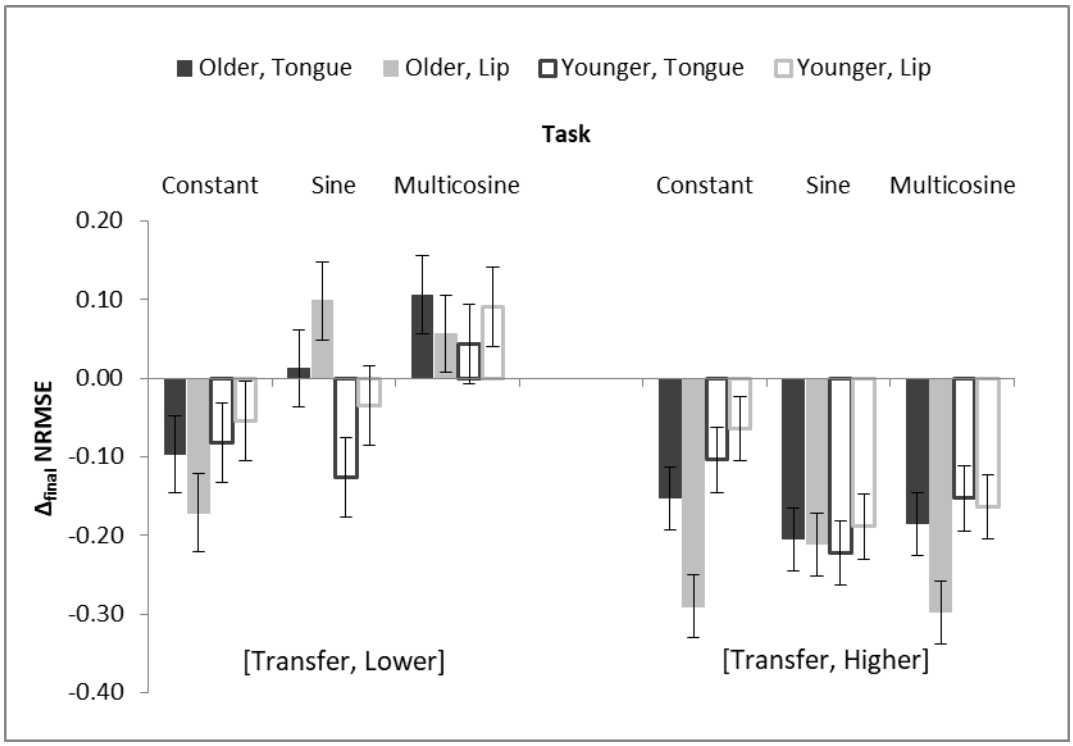

Figure 44. [Transfer – baseline] for transfer to lower and higher target force levels normalized root mean square error (NRMSE) by age group, task & effector (M ± SE). Significance not indicated; see text.

For transfer to lower target force, age group interacted with task ($F(2, 205) = 4.336$, $p = 0.014$). Both groups' transfer of skill depended on task (younger adults: ($F(2, 100) = 20.215$, $p < 0.0005$; older adults: ($F(2, 105) = 7.062$, $p = 0.001$). For younger adults, transfer of skill was greater for constant and

sine tasks than multicosine (both $p < 0.0005$). For older adults, transfer of skill was greater for the constant task than for both others ($p = 0.009$ vs. sine, $p = 0.002$ vs. multicosine). Within task, the effect of age group was significant only for the sine task ($F(1, 41) = 15.127$, $p < 0.0005$), for which the younger adults showed transfer of skill and the older adults did not.

For transfer to higher target force, when the full model was fit to the original data, residuals were negatively skewed. This issue was resolved with a reflect + square root transformation. With one exception noted below, the pattern of significance in the original-data model was the same as the transformed-data analysis reported here.

Age group interacted significantly with both effector ($F(1, 205) = 5.238$, $p = 0.023$) and task ($F(2, 205) = 3.114$, $p = 0.047$). The latter interaction was not significant in the original-data model ($F(2, 205) = 2.954$, $p = 0.054$) and is not marked on the graph, which shows original data. For the lip only, older adults showed a larger difference from baseline to higher-force transfer tasks than younger adults ($F(1, 41) = 7.806$, $p = 0.008$).

Younger adults' transfer of skill varied by task ($F(2, 100) = 15.933$, $p < 0.0005$): it was greater on both variable tasks than on the constant task (vs. sine, $p < 0.0005$; vs. multicosine, $p = 0.002$). For the constant task only, older adults showed a larger difference from baseline to higher-force transfer tasks than younger adults ($F(1, 41) = 7.078$, $p = 0.011$).

Specific Aim 3: Prediction of Learning

The purpose of all analyses in this section was to test the hypotheses that specific continuous quantitative predictor variables (each described in the subsection for its model) would predict learning; see Statistical Analysis for details. The dependent variable for all of these models was the change in normalized root mean square error from initial performance to either retention trials or trials of transfer to higher target force level: day-3 retention or transfer trial 1 minus day 1, block 1, trial 1 for each effector x task combination ($\Delta_{final}NRMSE_{ret}$ or $\Delta_{final}NRMSE_{trn}$). For $\Delta_{final}NRMSE_{ret}$, a more negative number indicated reduction in error from baseline (improvement in performance on the practiced tasks). For $\Delta_{final}NRMSE_{trn}$, lower normalized root mean square error on day 3 than day 1 indicated application of learned skill to a related task. These are the same outcomes analyzed in "Hypothesis 1d: Reduction in error following practice," which showed that age group, effector and task had significant influence. Thus, these factors were retained for this predictive modeling.

Hypotheses 3a and 3b. (a) Error and (b) higher maximal force entropy at baseline will predict retention and transfer in pursuit tracking tasks after controlling for age group, effector and task. The

primary continuous quantitative predictor of interest for these analyses was maximum entropy (maxApEn, maxFuzzyMEn) out of the five trials of the first block for each effector x task condition on day 1. The predictor of secondary interest was normalized root mean square error on day 1, block 1, trial 1 ($NRMSE_{initial}$). The choice to evaluate both a maximum entropy measure and $NRMSE_{initial}$ in the same model was made based on results of the opposite choice, explained below.

Recall that $NRMSE_{initial}$ has two potential interpretations as a predictor of reduced error after practice. If $NRMSE_{initial}$ is not merely an indicator of performance in the moment but also measures aptitude for or ability to learn the task, then higher (worse) $NRMSE_{initial}$ will predict higher $\Delta_{final}NRMSE_{ret}$ (also worse) and the parameter estimate will be positive. On the other hand, if higher initial performance error only provides a benchmark and does not indicate aptitude, then higher $NRMSE_{initial}$ allows greater room for improvement with practice, and the parameter estimate will be negative.

Retention. Initial models were fit using maximum entropy as the only quantitative predictor (along with categorical factors of age group, effector and task, and each one's interaction with maximum entropy). These models were unstable: random subjects effects were not supported,[17] and stepwise discarding of nonsignificant interaction terms with the highest *p*-values erased the significance of previously significant terms with each step, until a main-effects only model remained with no significant terms. Residuals in the initial model were strongly negatively skewed, and while a *reflect + logarithm* transformation mitigated this issue, it did not allow random subjects effects to be fit. Table 26 shows this instability for the model fitting process using maxApEn; the pattern was the same with maxFuzzyMEn. As the table shows, maximum entropy was often significant or nearly so, alone or in interaction with effector. However, given the model instability, these results are questionable.

Unlike analyses in previous sections, RMANOVA could not be used as an alternate analysis when random subjects effects were not supported in the LME models; it can fit a quantitative predictor for each subject, but not for each condition within subject. An alternate linear mixed effects modeling approach was therefore tried, including $NRMSE_{initial}$ as a second quantitative predictor rather than assessing its predictive utility separately.

---

[17] SPSS warnings: "The final Hessian matrix is not positive definite although all convergence criteria are satisfied…Validity of subsequent results cannot be ascertained"; "This covariance parameter [subject ID] is redundant. The test statistic and confidence interval cannot be computed."

Table 26

Changes in Significance of Quantitative Predictor and Interacting Terms during Unstable Stepwise Modeling Process

Model	Term	Test statistic	Significance
Model 1: main effects + all 2-way interactions with maxApEn	effector	$F(1, 246) = 3.078$	$p = 0.081$
	maxApEn	$F(1, 246) = 3.918$	$p = 0.049$*
	age group x maxApEn	$F(1, 246) = 0.714$	$p = 0.399$
	effector x maxApEn	$F(1, 246) = 4.877$	$p = 0.028$*
	task x maxApEn	$F(1, 246) = 1.284$	$p = 0.279$
↓			
Remove nonsignificant interaction terms.			
↓			
Model 2: main effects + 1 previously significant 2-way interaction	effector	$F(1, 246) = 2.301$	$p = 0.131$
	maxApEn	$F(1, 246) = 6.157$	$p = 0.014$*
	effector*maxApEn	$F(1, 246) = 3.348$	$p = 0.069$
↓			
Remove formerly significant interaction term.			
↓			
Model 3: main effects only	effector	$F(1, 246) = 1.364$	$p = 0.244$
	maxApEn	$F(1, 246) = 2.782$	$p = 0.097$
↓			
No significant terms remaining.			

Note. Age group and task main effect terms were included in all of these models but remained nonsignificant throughout and are not shown here.
* Significant at the $p < 0.05$ level.

$NRMSE_{initial}$ was a potentially important contributor to the model for two reasons. Firstly, controlling for initial performance could stabilize the model. Secondly, controlling for initial performance changes the tested hypothesis from 'After controlling for age group, effector and task, maximum entropy will predict reduction in error with practice' to 'After controlling for age group, effector, task, *and initial performance*, maximum entropy will predict reduction in error with practice.' This difference is important because the latter version makes the results more generalizable to other populations who might be expected to perform differently.

This alternate approach was partially successful in improving model quality. These models permitted fitting of random subjects effects, but model stability remained questionable: the initial models' residuals were excessively skewed, and both removal of two participants with extreme residuals and (separately) square root transformation led again to models unable to support random subjects effects. Thus the following results should be interpreted cautiously.

One borderline-significant interaction term was retained in the alternate models because of its consistent significance or near-significance across the variety of models tested: the interaction of effector with maximum entropy (maxApEn model: $F(1, 241.787) = 3.786$, $p = 0.053$; maxFuzzyMEn model: $F(1, 245.823) = 2.550$, $p = 0.112$). Controlling for this term as well as for age group, effector and task main effects, both quantitative predictors were significant.

$NRMSE_{initial}$ significantly predicted $\Delta_{final}NRMSE_{ret}$ (maxApEn model: $F(1, 242.978) = 797.01$; maxFuzzyMEn model: $F(1, 243.240) = 791.607$; both $p < 0.0005$), with a 1-unit increase (worsening) in $NRMSE_{initial}$ estimated to predict a change of -0.90 in $\Delta_{final}NRMSE_{ret}$ per both models (improvement in performance with practice) after controlling for age group, task, effector and maximum entropy.

Maximum entropy also significantly predicted $\Delta_{final}NRMSE_{ret}$ (maxApEn model: $F(1, 245.948) = 7.005$, $p = 0.009$; maxFuzzyMEn model: $F(1, 245.823) = 5.414$, $p = 0.021$). 1-unit increases in maxApEn and maxFuzzyMEn were estimated to predict changes in $\Delta_{final}NRMSE_{ret}$ of 0.32 and 0.14 respectively (smaller reduction in error), after controlling for age group, task, effector and $NRMSE_{initial}$.

Transfer. Initial linear mixed effects models were fit with main effects of age group, effector, task, $NRMSE_{initial}$, and maxApEn or maxFuzzyMEn. For both models, all interaction terms were nonsignificant, leaving only main effects in the final models. Both models also had excessively skewed random effects, which were resolved by the removal of a single older adult with extreme values. Pattern of significance was the same without this participant as in the original analysis reported here.

$NRMSE_{initial}$ significantly predicted change in error from baseline to transfer for higher target force levels in both models (maxApEn model: $F(1, 243.256) = 1181.922$; maxFuzzyMEn model: $F(1, 243.315) = 1184.732$, both $p < 0.0005$). A 1-unit increase (worsening) in $NRMSE_{initial}$ was estimated to predict a change of -0.91 in $\Delta_{final}NRMSE_{trn}$ per both models (improved transfer) after controlling for age group, task, effector and maximum entropy.

Maximum entropy did not significantly predict change in error from baseline to transfer for higher target force levels in either model (maxApEn model: $F(1, 240.019) = 0.008$, $p = 0.927$; maxFuzzyMEn model: $F(1, 240.638) = 0.043$, $p = 0.836$).

Hypotheses 3a and 3c. (a) Error and (c) greater adaptability of entropy at baseline will predict retention and transfer in pursuit tracking tasks after controlling for age group, effector, task and the other continuous quantitative predictor. The quantitative predictor of primary interest for these analyses was the difference in entropy from trial 1 to trial 2 of the first block for each effector x task

combination on day 1 ($\Delta_{initial}$ApEn or $\Delta_{initial}$FuzzyMEn). It was evaluated jointly with $NRMSE_{initial}$ because $NRMSE_{initial}$ was shown in the previous analyses to be a significant predictor.

Retention. After stepwise removal of nonsignificant interactions, the final models included main effects of the categorical factors and quantitative predictors, plus the interaction of task with $\Delta_{initial}$entropy. Models' residuals were moderately positively skewed. Rerunning the models sans the two participants with extreme values (the same two for each model) mitigated this weakness and produced the same pattern of results as in the original analyses reported here, with the exception noted below.

$NRMSE_{initial}$ remained significant in these models ($\Delta_{initial}$ApEn model, $F(1, 245.966) = 757.115$; $\Delta_{initial}$FuzzyMEn model, $F(1, 245.974) = 773.854$, both $p < 0.0005$), with 1-unit changes estimated to predict changes in $\Delta_{final}NRMSE_{ret}$ of -0.89 and -0.90 respectively, after controlling for age group, effector, task and $\Delta_{initial}$ApEn/$\Delta_{initial}$FuzzyMEn.

The effect of initial change in entropy varied by task ($\Delta_{initial}$ApEn: $F(2, 236.302) = 3.514$, $p = 0.031$; $\Delta_{initial}$FuzzyMEn: $F(2, 236.860) = 5.229$, $p = 0.006$). Follow-up analyses were affected by inclusion vs. exclusion of participants with extreme residuals. See Table 27. Initial change in entropy met the adjusted criterion for significance for each task under one circumstance: for the constant task, $\Delta_{initial}$ApEn was significant in the all-participants model; for the sine task, $\Delta_{initial}$FuzzyMEn was significant when participants with extreme residuals were excluded. The direction of the effect varied by task. For the constant task, which demands force output of high entropy, higher entropy on trial 2 than trial 1 on day 1 predicted a decrease in $\Delta_{final}NRMSE_{ret}$ (i.e. a greater reduction in error by day 3). For the sine target, which requires force output of low entropy, higher entropy on trial 2 than trial 1 on day 1 predicted an increase in $\Delta_{final}NRMSE_{ret}$ (i.e. a lesser reduction in error by day 3).

Table 27

Effect of Initial Change in Entropy on Change in Normalized Root Mean Square Error After Practice, Within Task, in Models Using All Participants vs. Excluding Two with Extreme Residuals

Task	Entropy measure	Model	Test statistic	Significance	Parameter estimate
Constant	$\Delta_{initial}$ ApEn	N = 41	$F(1, 80.703) = 6.877$	$p = 0.010$*	-0.138
		N = 39	$F(1, 78.000) = 5.540$	$p = 0.021$	-0.100
	$\Delta_{initial}$ FuzzyMEn	N = 41	$F(1, 81.482) = 5.715$	$p = 0.019$	-0.074
		N = 39	$F(1, 78.000) = 4.983$	$p = 0.028$	-0.055
Sine	$\Delta_{initial}$ ApEn	N = 41	$F(1, 77.164) = 2.105$	$p = 0.151$	0.247
		N = 39	$F(1, 77.982) = 3.411$	$p = 0.069$	0.251
	$\Delta_{initial}$ FuzzyMEn	N = 41	$F(1, 80.686) = 4.875$	$p = 0.030$	0.222
		N = 39	$F(1, 77.139) = 6.318$	$p = 0.014$*	0.197
Multicosine	$\Delta_{initial}$ ApEn	N = 41	$F(1, 82) = 0.710$	$p = 0.402$	0.167
		N = 39	$F(1, 68.048) = 0.005$	$p = 0.946$	-0.010
	$\Delta_{initial}$ FuzzyMEn	N = 41	$F(1, 82) = 1.349$	$p = 0.249$	0.126
		N = 39	$F(1, 68.399) = 0.192$	$p = 0.663$	0.036

Note. Neither initial entropy change measure was significant for the multicosine task in any model. Bonferroni-adjusted significance criterion is $p < 0.017$ (0.05 / 3 tasks).
* Significant at the $p < 0.017$ level.

Transfer. No interactions were significant; the final models included main effects of the categorical factors (age group, effector, task) and quantitative predictors ($NRMSE_{initial}$ and initial change in entropy). Both models also had excessively skewed random effects, which were resolved by the removal of a single older adult with extreme values. Pattern of significance was the same without this participant as in the original analysis reported here.

$NRMSE_{initial}$ significantly predicted change in error from baseline to transfer for higher target force levels in both models ($\Delta_{initial}$ ApEn model: $F(1, 243.505) = 1148.820$; $\Delta_{initial}$ FuzzyMEn model: $F(1, 243.303) = 1149.167$, both $p < 0.0005$). A 1-unit increase (worsening) in $NRMSE_{initial}$ was estimated to predict a change of -0.89 [-0.94, -0.84] in $\Delta_{final}NRMSE_{trn}$ per both models (improved transfer) after controlling for age group, task, effector and initial change in entropy.

Initial change in entropy significantly predicted transfer of learning to tasks with a higher target force level in both models ($\Delta_{initial}$ApEn model: $F(1, 232.077) = 11.853$; $\Delta_{initial}$FuzzyMEn model: $F(1, 234.461) = 10.437$, both $p = 0.001$). 1-unit increases in $\Delta_{initial}$ApEn and $\Delta_{initial}$FuzzyMEn were estimated to predict changes in $\Delta_{final}NRMSE_{trn}$ of -0.17 [-0.27, -0.07] and -0.09 [-0.15, -0.04] respectively (improved transfer), after controlling for age group, task, effector and $NRMSE_{initial}$.

Summary of Findings

Table 28

Summary of Significant Findings by Specific Aim and Hypothesis

Specific Aim 1. Assess applicability of previous findings on effects of age and task to oral effectors.

Hypothesis 1a. Older adults' force structure will differ task-dependently from younger adults' (lower entropy[a] and a greater proportion of low-frequency power when the task demands high entropy and reduced low-frequency power, and vice versa).

At baseline:

- ✓ Entropy, YA only: both measures, C > S, M > S; FZ only: C > M.[b]
- ✓ Entropy, task C only: YA > OA.
- ✓ PoP, 0-1 Hz band, YA only: C < S.[b]

In retention trials:

- ✓ Entropy: for OA, C > S; for YA, C > M and C > S for both effectors, M > S only for T.[b]
- ✓ Entropy: YA > OA in the "L x C" condition, YA < OA in the "T x S" condition.
- ✓ PoP, task S: YA > OA in the 0-1 Hz band, YA < OA in higher bands.
- ✓ PoP, 1-2 Hz band: OA did not differentiate tasks; YA differentiated all three, M > C > S.[b]
- ✓ PoP, 2-3 Hz band: OA had M > S. YA differentiated all three, C > M > S.[b]

Hypothesis 1b. Adaptability (immediate): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing trial 2 to trial 1 on day 1 within each effector x task combination.

- × Entropy: immediate change was affected by effector and task, but not by age group.
- × PoP: No tested factors had significant influence on immediate change.

Hypothesis 1c. Adaptability (after practice): Older adults will change structure of force to meet task demands less effectively than younger adults, comparing day 1 trial 1 to day 3 retention trial 1 within each effector x task combination.

- ✓ Entropy, task M: YA did not change with practice; OA increased.
- ✓ Entropy, task S: YA decreased with practice[b]; OA increased.

Hypothesis 1d. Older adults will show less reduction in error relative to baseline on retention and transfer trials after two days' practice than younger adults.

In retention trials:

- ✓ NRMSE reduction vs. baseline, T only: YA > OA.
- ✓ NRMSE reduction vs. baseline, task S only: YA > OA.

In transfer trials:

- ✓ NRMSE reduction vs. baseline, transfer to lower target force, task S only: YA > OA.
- ↩ NRMSE reduction vs. baseline, transfer to higher target force, L only: OA > YA.
- ↩ NRMSE reduction vs. baseline, transfer to higher target force, task C only: OA > YA.

Hypothesis 1e.[c] Structure of force and its change with practice will differ by task. The constant task will elicit the highest entropy, lowest proportion of low-frequency power, and greatest proportion of higher-frequency power. The sine task will elicit the lowest entropy, greatest proportion of low-frequency power, and lowest proportion of higher-frequency power. The multicosine task will be intermediate.

At baseline:

- ✓ PoP: in 0-1 Hz band, M < S; 1-2 Hz band, M > S

Table 28 (cont.)

Retention – baseline change: ✓ Entropy increase vs. baseline: C > S, C > M. ✓ PoP change vs. baseline, task C: 0-1 Hz and 1-2 Hz, ↓; 2-3 Hz, ↑ ✓ PoP change vs. baseline, task S: 0-1 Hz, ↑; 1-2 Hz, ↓
Specific Aim 2. Assess differences in motor variability between oral effectors. Hypothesis 2a. The tongue will produce less complex force than the lip (lower-entropy, greater dominance of low-frequency power). At baseline AND in retention trials: ⮌ PoP: L > T in the 0-1 Hz band; T > L in the higher bands. Hypothesis 2b. The effects of age group and effector on entropy will interact. At baseline: ✓ Entropy, OA only: T > L. ✓ Entropy, L only: YA > OA. In retention trials: ✓ Entropy, OA only: T > L. ✓ Entropy, YA only: task C, L > T; task M, T > L. ✓ Entropy, YA only: C > M and C > S for both effectors; M > S only for T. ✓ Entropy: YA > OA in the "L x C" condition, YA < OA in the "T x S" condition.
Specific Aim 3. Assess utility of baseline performance measures in predicting *de novo* learning of fine-force pursuit tracking tasks in oral effectors. Hypothesis 3. (a) Error and a measure of temporal structure, (b) higher maximal force entropy or (c) greater adaptability of entropy at baseline, will predict retention and transfer in pursuit tracking tasks after controlling for age group, effector and task. ✓ $NRMSE_{initial}$ predicted reduction in error on both retention and transfer trials in all models. In retention trials: ⮌ Higher baseline maximum entropy predicted smaller reduction in error. ✓ The effect of initial change in entropy varied by task: task C, $\Delta_{initial}$ApEn > 0 predicted greater reduction in error; task S, $\Delta_{initial}$FuzzyMEn > 0 predicted lesser reduction in error. In transfer trials: ✓ Immediate increase in entropy at baseline predicted improved transfer to higher target force.

Key:

✓ Supportive of the hypothesis.
× No difference found.
⮌ Difference was opposite the expected direction.
AE = approximate entropy
FZ = fuzzy measure entropy
PoP = proportion of power
NRMSE = normalized root mean square error
YA = younger adults
OA = older adults
L = lip
T = tongue
C = constant task
M = multicosine task
S = sine task

[a] "Entropy" refers to both entropy measures. If a finding was significant for only one, the measure abbreviation is used.

[b] Also supportive of Hypothesis 1e.

[c] Only main effects of task or interactions with effector are listed here. Interactions of age group and task are listed under other Aim 1 hypotheses.

Discussion

Specific Aim 1

Results from this study suggest that the bidirectional complexity hypothesis of healthy aging, which posits age-related loss of the ability to adapt structure of motor output to task demand rather than a unidirectional simplification of motor output (Vaillancourt & Newell, 2002), can be applied to fine-force control in pursuit tracking tasks using the lip and tongue.

Even during initial performance, younger adults responded to differing tasks' demands by producing force differentiable in entropy and frequency structure. Older adults, on the other hand, did not initially differentiate the entropy of their force output by task and did not reduce power in the lowest frequency band as needed for the constant target compared to the sine. Sosnoff and Voudrie's (2009) older participants appear to have performed similarly using index finger flexion: though the effect of task within age group was not reported, their Figure 3 suggests older adult participants' entropy did not initially differ between constant and sine tasks.

Following two days' practice, both older and younger adults differentiated force structure by task on retention trials, but remaining differences showed that, consistent with Sosnoff and Voudrie (2009), practice had not entirely compensated for age-related loss of adaptability of force structure.

The effects of task on force structure seen in this data set were as expected from the manual motor control literature (Newell & Vaillancourt, 2001; Vaillancourt & Newell, 2002). With a single exception (the older adults had increased entropy for the sine task after practice), all significant differences in entropy and proportion of power by task, whether assessing a single time point (baseline or retention trials) or the change between the two, supported the idea that compared to the other tasks, a constant target should elicit force with relatively high entropy and a greater proportion of higher-frequency power, the sine target should elicit the opposite pattern (low entropy, lowest-frequency force dominating to the greatest extent among the tasks), and performance should evolve towards higher entropy for constant tasks and lower entropy for sine tasks with practice.

The age x task interaction expected from the bidirectional loss of adaptability hypothesis evaluated in the manual motor control literature (Sosnoff & Newell, 2008; Sosnoff, Vaillancourt, & Newell, 2004; Sosnoff & Voudrie, 2009; Vaillancourt & Newell, 2002) partially held in this data set: the sine target, requiring low-entropy force dominated by a single frequency, elicited higher-entropy force in older adults than younger adults when using the tongue, while the constant target, requiring high-entropy force with a broader spectrum containing more high-frequency energy, elicited lower-entropy

force in the older adult than the younger adult participants using the lip. In both cases, the contrast suggests younger adults adapted structure of output to task demand more closely than the older adults. After practice, younger adults had maintained or decreased the entropy of their force for the variable tasks (in line with task demand), while older adults had increased entropy on these tasks (counter to task demand). The proportion of power analysis was similarly supportive, showing age x task interaction for every analyzed frequency band. The older adults showed greater power than the younger adults in the higher-frequency bands for the variable tasks and lesser power in the lowest band, consistent with the idea of reduced adaptability. The differential effect of effector could suggest that fine force control is better preserved in oral than manual effectors, in which case loss of adaptability should be significant in all oral effectors in participants older than the older adults in this data set. It may also suggest differential preservation of function within the oral motor system. Over all tasks in both age groups, the tongue had more power in the higher frequencies than the lip, and the lip had more power in the lower frequencies than the tongue; it is precisely in the tasks challenging these propensities (sine task for the tongue, constant task for the lip) where age differences were significant.

Though the significant differences predicted by the bidirectional complexity hypothesis did not appear under all conditions, the frequent interactions of age and task make this hypothesis a better fit than the two primary alternatives. The information-theoretic perspective (Shannon, 1948), based on the idea of variability as random sensorimotor noise increasing with age, would have predicted higher entropy and more broadband spectra (more high-frequency power) for older adults regardless of task. The loss of complexity hypothesis (Lipsitz & Goldberger, 1992) would have predicted lower entropy and more narrowband (dominant low-frequency) spectra for older adults, again regardless of task.

Though EMG data were not collected, the task-specific age group differences seen in frequency structure of force on retention trials are consistent with Sosnoff et al.'s (2004) proposal that elders' loss of adaptability of temporal structure of force is related to the impaired coordination of neural oscillators (motor neuron pools) whose contributions to central drive of the effectors occur on different time scales and thus add degrees of freedom (dimension) to the force signal.

Specific Aim 2

The existence of an effector difference was supported, but the hypothesized direction was not. Effector differences were found on both initial and final (retention) task performance. Older adults produced higher-entropy force with the tongue than the lip at both time points. Also at both times, across tasks and age groups, the lip produced greater power in the 0-1 Hz band than the tongue

(consistent with lower entropy), while the tongue produced greater power than the lip in the higher frequency bands (consistent with higher entropy). These results differ from those reported without statistical testing by Holtrop et al. (2014). Potential causes of difference include participant age (similar averages for the young adults, but older adults approximately eight years older on average in this study); differing target force levels (15% MVF vs. 10% and 20% MVF), the very different number of trials (three per effector per force level = 6 per effector during a single session, vs. 35 per effector per task = 105 per effector over two days), and most importantly, the inclusion of tasks requiring differing force structure vs. constant force only.

In retention trials, only the younger adult participants increased entropy of lip force for the constant task. The older adult participants were not able to make this adjustment in response to task demand. Holtrop et al. (2014) may have found a precursor to this effect of age in their somewhat younger old-adult participants: at 10% MVC (the lower and more difficult of their two target force levels), older adults had higher ApEn in the tongue than the lip – that is, in the most stressful condition, their participants performed more like the older adults in this data set. Further investigation can be done with the current data set, as the transfer trials used target force levels of 10% and 20% MVF. Though these trials took place after significant practice, vs. the unpracticed trials in Holtrop et al., the target force levels are the same and the comparison will be more direct. If older adults in this work continue to show higher entropy for the tongue at 10% and 20% MVF, the consistent difference would support an argument that even a sub-decade change in age for older adults may produce a differential effect of aging on the lip and tongue, with fine force control better preserved in the tongue at least for tasks requiring high-complexity output. Older adults were noted here to be less able to reduce entropy in the tongue for the sine task than younger adults, but no comparison can be made with results obtained by Holtrop et. al., who did not investigate variable tasks.

Immediate adaptability of entropy within task was greater for the lip across age groups. On final performance, effector differences in entropy by task suggested that for younger adults, adaptation of temporal structure of force to task demand with practice may have been better with the lip.

Age group and effector interacted significantly at both time points and in the change over time, in measures of both accuracy and force structure. While the effectors differed in strength, neither decreased significantly in strength with age, suggesting that the interactive effects of age and effector on force structure are more likely due to changing neural control with aging working on differently structured effectors than to declining strength, contra the suggestion of Sosnoff and Newell (2006).

Specific Aim 3

Three baseline behavioral measures each achieved success in predicting learning (measured as reduction in error) after two days' practice. Initial error was the strongest predictor. Consistent with Barbado Murillo et al. (2017), higher initial error represented greater room for improvement, suggesting that the measure functions as a performance benchmark rather than as a measure of aptitude or potential to learn the tasks. In participants with intact motor-learning capabilities, it is unsurprising that many participants were able to make this improvement with intense practice. In clinical populations whose disorders may directly bear on capacity for motor learning, the relation between current and eventual performance will likely be modified by other factors.

After controlling for age group, effector, task, and initial performance, greater baseline maximum entropy predicted smaller reduction in error (i.e. reduced learning) on retention trials only. The direction of the relationship between initial maximum entropy and performance change was contrary to the predicted relationship, which was hypothesized based on the ideas that (i) a greater maximum value of entropy indicated a system with fewer constrained degrees of freedom and thus greater potential adaptability to task demands with practice and (ii) maximal entropy values on the tasks (particularly the constant task) might be a reasonable proxy for this maximum value. However, while (ii) might be accurate after considerable practice (which pairs improved performance on the constant task with increased complexity), initial complexity values do not appear to indicate anything about a possible maximum. Two of the three tested tasks demand relatively low-complexity output. If high entropy on initial trials of these tasks is an indicator of poorer intra-trial adaptability to task, rather than a measure of a participant's maximal possible complexity, its prediction of smaller reduction in error with practice fits with the results from the initial adaptability analysis.

Initial adaptability of entropy predicted better performance on retention trials if the direction of change was in line with task demand, and worse performance if the direction of change was counter to task demand. This effect comports with the idea of variability in early learning as an exploration of task space (Dhawale, Smith, & Ölveczky, 2017; Stergiou, Harbourne, & Cavanaugh, 2006; Wu, Miyamoto, Gonzalez Castro, Ölveczky, & Smith, 2014) and therefore an active support of learning, rather than a hindrance to be suppressed. "Optimal" variability in a learning context suggests the ability to shift temporal/frequency structure of force output in the direction demanded by a goal or task. The reduction in adaptability of force structure seen in the older adult participants may play a role in changes in learning with aging.

Measure Choice: Approximate vs. Fuzzy Measure Entropy

Based on its more nuanced mathematics and previous improved performance vs. earlier entropy algorithms (including ApEn) in classifying heart failure vs. healthy control patients (Liu et al., 2013), FuzzyMEn was expected to perform better than ApEn in detecting subtle age group or effector differences. This expectation was not borne out; the measures performed nearly identically. As comparison of the measures was not the focus of this work, no analyses attempted to parse reasons for the difference in relative performance compared to previous literature. Parameter choice, frequency content of the analyzed signals, or their interaction may play a role. It may also be the case that in this data set, runs of data with lengths *m* and then *m* + 1 fell either well within or well outside of the tolerance *r* used to judge similarity, meaning that the additional nuance offered by FuzzyMEn for borderline cases had little effect.

Currently neither measure can be recommended over the other for analysis of data similar to that presented here. ApEn is slightly faster to calculate than FuzzyMEn, but the difference is of practical significance only when many calculations (thousands) are performed together.

Limitations

The most significant limitations of this work in terms of its applicability to predicting therapeutic benefit in clinical populations is that the participants were all healthy and the tasks required *de novo* learning, which is a less likely focus in therapy than relearning of lost skills. M1 damage or post-lesion reorganization, interference from previously learned motor skills, dysfunctional spontaneous reorganization or differences between healthy and damaged white matter microstructure and its response to learning may mean that mechanisms and patterns of relearning differ from *de novo* learning (Hosp & Luft, 2011; Johansen-Berg, Scholz, & Stagg, 2010). However, the importance of this limitation may differ between the predictors assessed. *Room to improve* (the supported interpretation of initial accuracy as a predictor) is likely to be specific to the population and task.

On the other hand, the arguments for entropy-based predictors relied on the idea of coordination between multiple components of a system – whatever those components might be and in whatever condition. Suppose healthy system H uses three components to perform a task, refining the coordinative relationships between them during the learning process. After an injury affecting H, substitute system S is used to perform the same task. S's components might be the same components and relationships or a subset thereof, some or all of which might be changed by the injury, or might incorporate some or all new components and relationships. But in any of those cases, the early efforts of

the components of S to establish, mend, or alter their relationships should predict how S *as it now exists* will be able to learn or relearn the task. In healthy adults or in clinical populations with chronic disorders, in which physical system components and links between them are not expected to change rapidly, entropy-based predictors could potentially predict change over a somewhat longer time frame than that examined here. In clinical populations in an acute recovery phase, the components of the system or their relationships may be changing enough as recovery takes place to limit the utility of predictions to short periods in which the relevant systems are in similar physical condition.

Related to the issue of *de novo* learning is the issue of task functionality or relevance. The visual targets used in this study were designed to differ maximally in complexity within the limits of plausible short-term learnability, to provide the best chance of eliciting differently complex output. In that sense task design was successful, but at the cost of similarity to any normal functional speech or swallowing task. Neither continuous fine force in a highly predictable pattern, nor the use of visual feedback to judge force output of one's own lips or tongue, closely resemble everyday experience. Thus generalizability to oral motor learning of functional tasks is questionable; it is not yet clear whether use of immediate adaptability to task demand to predict short-term therapy benefit is indicated. (It is unlikely to predict long-term benefit.)

No physiological or neurophysiological data (respiratory traces, electromyography etc.) were collected. Other than the observed differences in distribution of power, sources of complexity difference between effectors or age groups remain unknown. Some previous work on sustained force has asked participants to refrain from breathing during the task (Burnett, Laidlaw, & Enoka, 2000). Though this work did not ask participants to alter respiration, some commented that they held their breath or otherwise modified their breathing during trials.

Others commented that part of initial learning was deciding on "how to hold my tongue on this thing" (the transducer). Tongue posture and contact point with the transducer may have varied across participants and within participants across trials, although contact point was limited to anterior dorsal tongue surface. Consequently muscle activation patterns are also likely to have varied, potentially accounting for some portion of performance variability across and within participants. This early variability due to exploration of the task space may have contributed to participants' learning (Wu et al., 2014), but its extent, and therefore importance, cannot be assessed from the captured data.

The measure of performance accuracy (normalized root mean square error) is designed to assess how closely the participant matched the target from moment to moment and thus effectively

detects decreases in overshoot or undershoot (mismatch of force *amplitude* to target). However, it is less likely to detect improvement in *timing* of force amplitude change. A correctly shaped and level-matched pattern offset in time (phase) from a time-varying target may still score poorly. Several participants were observed to produce what both participant and investigator agreed were "better" target matches than their previous attempts without concomitant decrease in NRMSE, likely due to this issue. Reliance on NRMSE for feedback may consequently have biased participants towards focusing on matching the amplitude of the target, reducing their attention to matching temporal characteristics of the varying targets, and consequently reducing change in the complexity of their force output. (Changing only the amplitude of a signal does not alter its complexity.) Future work may need to incorporate a time-based performance measure in addition to NRMSE.

Current methodology relied on establishing a maximal voluntary force for each effector for consistent target scaling. Three to four trials per effector were used, with one-minute rests after each, which was adequate to avoid fatigue (per participant report) in this healthy adult population. In a clinical setting with high-fatigue patients, this regimen is both likely to exacerbate fatigue and unlikely to establish a value that is either consistent or representative of the patient's non-fatigued maximum. It may be more effective to allow the patient to set the target force level, for instance "press gently, at a level you feel you can comfortably control for ten seconds."

Future Directions

This study represented a first step in establishing the ability to predict very short-term therapy benefit from limited data on in-the-moment capacity for motor learning, using entropy measurements sensitive to the coordination of multiple control processes acting across different timescales. Further exploration of variability as a predictor of motor learning may support Stergiou et al.'s recommendation of variability as a therapy strategy, or suggest ways in which this strategy can be modified to better suit oral effectors and functional communication and swallowing tasks.

While the equipment used for the current study remains impractical for clinical use, study methods could potentially be adapted to more widely clinically available equipment such as the Iowa Oral Pressure Instrument (IOPI Medical LLC, Redmond, Washington).

Maximum entropy among an initial trial set, and change in entropy from the first to second trial, predicted small to moderate changes in amount of learning after two days' practice by healthy younger and older adults.

The predictive value of initial entropy adaptability may be improved by increasing the number of trials over which entropy change is assessed, though the amount of predictive improvement will have to be balanced against increased demand on participants (minimal for healthy adults but potentially prohibitive for high-fatigue clinical participants). Another possible route to improvement would be a change in the calculation of entropy adaptability to take into account the target complexity. An estimate of the true target in skilled human performance might involve the addition of a participant-specific neural noise estimate (based on signal from 30-50 Hz after correcting for equipment noise as in Sosnoff & Newell, 2011) to the mathematically generated target.

Effectiveness of entropy algorithms to detect and predict change may improve if their parameters can be systematically chosen to match expected data characteristics. Calculation of ApEn and FuzzyMEn with systematically varied parameters (*m* from 1 to 6 in increments of 1, *r* from 0.1 to 0.3 in increments of 0.05, *N* from 100 to 1100 in increments of 100) was performed for each trial across all three days of the experiment, but that analysis was outside the scope of this report. This work will be completed on the already-collected data from healthy adults and on signals of controlled content before application to clinical populations is attempted. While determining the appropriate values of *m* and *r* is most likely to contribute by improving sensitivity of measurement, *N* has particular practical relevance. Shorter trials would be preferable for clinical populations with significant fatigue, if adequate measurements can be obtained with a lower *N*.

Based on the investigator's observations during experimental sessions, participants appeared to demonstrate different patterns of learning over time and possibly per task: some seemed to make initially rapid then decelerating progress, while others showed slow but steady or initially slow but accelerating progress. Several appeared to experience a sudden improvement, particularly with the multicosine pattern, at varying times including as late as the last block of trials on the second day. Data and NRMSE/entropy measurements were recorded for all trials but have not yet been analyzed; a growth curve analysis might allow more fine-grained prediction of how and whether participants are likely to learn.

Several participants commented on metaphor-related learning strategies: "[the multicosine target] looks like a W, so I am drawing a W with my tongue" ; "I think of [the multicosine target] like a European ambulance sounds, you know, up and down like wooOOoo wooOOoo? I'm trying to do that." These participants felt that their mental visual or auditory images helped to improve their performance, although analysis of their claims in this data set is not possible, as not all participants reported on their

strategies. (After a few participants volunteered comments, the rest were asked about learning strategies.) It is possible that metaphorical expression of motor goals could be an effective therapy tool, particularly for effectors not ordinarily subject to visual observation, as there are links between metaphor, motor imagery or mental practice, and learning (Feltz & Landers, 1983; Mulder, 2007; Mulder, Zijlstra, Zijlstra, & Hochstenbach, 2004). (Note that the second commenter was trying to connect oral force production with sound, a much more normal sensorimotor match than the visual biofeedback used here.)

Conclusions

Entropy measurements of continuous fine force production, particularly change in entropy over time, show promise for predicting learning and support the concept of variability as an exploration of task space and a support of early learning.

This prediction can be made from a small enough data set to have potential clinical applicability. Older adults remain robustly able to learn and to adjust complexity of oral force output, though with limitations most consistent with the loss of adaptability hypothesis.

References

Adams, V., Mathisen, B., Baines, S., Lazarus, C. L., & Callister, R. (2013). A systematic review and meta-analysis of measurements of tongue and hand strength and endurance using the Iowa Oral Performance Instrument (IOPI). *Dysphagia, 28,* 350-369.

Alderman, R. B. (1965). Influence of local fatigue on speed and accuracy in motor learning. *Research Quarterly, 36,* 131-140.

American College of Sports Medicine (2009). American College of Sports Medicine position stand: Progression models in resistance training for healthy adults. *Medicine and Science in Sports and Exercise, 41,* 687-708.

Arslan, S. S., Ínal, Ö., Demir, N., Ölmez, M. S., & Karaduman, A. A. (2017). Chewing side preference is associated with hemispheric laterality in healthy adults. *Somatosensory and Motor Research, 34,* 92-95.

Bain, B. A. & Dollaghan, C. A. (1991). The notion of clinically significant change. *Language, Speech, and Hearing Services in Schools, 22,* 264-270.

Barbado Murillo, D., Caballero Sánchez, C., Moreside, J., Vera-García, F. J., & Moreno, F. (2017). Can the structure of motor variability predict learning rate? *Journal of Experimental Psychology: Human Perception and Performance, 43,* 596-607.

Barlow, S. M. & Burton, M. K. (1990). Ramp-and-hold force control in the upper and lower lips: developing new neuromotor assessment applications in traumatically brain injured adults. *Journal of Speech and Hearing Research, 33,* 660-675.

Barlow, S. M. & Muller, E. M. (1991). The relation between interangle span and in vivo resultant force in the perioral musculature. *Journal of Speech and Hearing Research, 34,* 252-259.

Bassingthwaighte, J. B., Liebovitch, L. S., & West, B. J. (1994). *Fractal physiology*. (vols. 2) Oxford: American Physiological Society.

Baweja, H. S., Kennedy, D. M., Vu, J., Vaillancourt, D. E., & Christou, E. A. (2010). Greater amount of visual feedback decreases force variability by reducing force oscillations from 0-1 and 3-7 Hz. *European Journal of Applied Physiology, 108,* 932-943.

Beeson, P. M., Rising, K., & Volk, J. (2003). Writing treatment for severe aphasia: Who benefits? *Journal of Speech, Language, and Hearing Research, 46,* 1038-1060.

Bernstein, N. A. (1967). *The co-ordination and regulation of movements*. Oxford: Pergamon Press.

Blackman, R. B. & Tukey, J. W. (1958). *The measurement of power spectra from the point of view of communications engineering*. New York: Dover.

Boot, W. R., Basak, C., Erickson, K. I., Neider, M., Simons, D. J., Fabiani, M. et al. (2010). Transfer of skill engendered by complex task training under conditions of variable priority. *Acta Psychologica, 135,* 349-357.

Bravo, G. & Hébert, R. (1997). Age- and education-specific reference values for the Mini-Mental and modified Mini-Mental State Examinations derived from a non-demented elderly population. *International Journal of Geriatric Psychiatry, 12,* 1008-1018.

Bronson-Lowe, C. R., Loucks, T. M. J., Ofori, E., & Sosnoff, J. J. (2013). Aging effects on sensorimotor integration: a comparison of effector systems and feedback modalities. *Journal of Motor Behavior, 45,* 217-230.

Burnett, R. A., Laidlaw, D. H., & Enoka, R. M. (2000). Coactivation of the antagonist muscle does not covary with steadiness in old adults. *Journal of Applied Physiology, 89,* 61-71.

Carnaby, G. M., Hankey, G. J., & Pizzi, J. (2006). Behavioural intervention for dysphagia in acute stroke: A randomised controlled trial. *Lancet Neurology, 5,* 31-37.

Carron, A. V. (1969). Physical fatigue and motor learning. *Research Quarterly, 40,* 682-686.

Carron, A. V. & Ferchuk, A. D. (1971). The effect of fatigue on learning and performance of a gross motor task. *Journal of Motor Behavior, 3,* 62-68.

Cavanaugh, J. T., Kelty-Stephen, D. G., & Stergiou, N. (2017). Multifractality, interactivity, and the adaptive capacity of the human movement system: a perspective for advancing the conceptual basis of neurologic physical therapy. *Journal of Neurologic Physical Therapy, 41,* 245-251.

Chen, W., Zhuang, J., Yu, W., & Wang, Z. (2009). Measuring complexity using FuzzyEn, ApEn, and SampEn. *Medical Engineering & Physics, 31,* 61-68.

Chen, X., Mohr, K., & Galea, J. M. (2017). Predicting explorative motor learning using decision-making and motor noise. *PLoS Computational Biology, 13,* e1005503.

Chernikoff, R., Birmingham, H. P., & Taylor, F. V. (1955). A comparison of pursuit and compensatory tracking under conditions of aiding and no aiding. *Journal of Experimental Psychology, 49,* 55-59.

Clark, H. M. & Solomon, N. P. (2012). Age and sex differences in orofacial strength. *Dysphagia, 27,* 2-9.

Cochran, B. J. (1975). Effect of physical fatigue on learning to perform a novel motor task [Abstract]. *Research Quarterly, 46,* 243-249.

Coren, S. (1993). The lateral preference inventory for measurement of handedness, footedness, eyedness, and earedness: Norms for young adults. *Bulletin of the Psychonomic Society, 31,* 1-3.

Corsini, G. & Saletti, R. (1988). A $1/f^{\gamma}$ power spectrum noise sequence generator. *IEEE Transactions on Instrumentation and Measurement, 37,* 615-619.

Craik, K. J. W. (1947). Theory of the human operator in control systems. I. The operator as an engineering system. *British Journal of Psychology: General Section, 38,* 56-61.

Crow, H. C. & Ship, J. A. (1996). Tongue strength and endurance in different aged individuals. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 51,* M247-M250.

Cuddington, K. & Yodzis, P. (1999). Black noise and population persistence. *Proceedings of the Royal Society of London Series B: Biological Sciences, 266,* 969-973.

de Smith, M. J. (2015). Statistical analysis handbook: a comprehensive handbook of statistical concepts, techniques and software tools. World Wide Web [Electronic version]. Available: http://www.statsref.com/HTML/index.html

Deffeyes, J. E., Harbourne, R. T., Stuberg, W. A., & Stergiou, N. (2011). Approximate entropy used to assess sitting postural sway of infants with developmental delay. *Infant Behavior and Development, 34,* 81-99.

Del Rey, P. & Shewokis, P. (1993). Appropriate summary KR for learning timing tasks under conditions of high and low contextual interference. *Acta Psychologica, 83,* 1-12.

Della-Maggiore, V., Scholz, J., Johansen-Berg, H., & Paus, T. (2009). The rate of visuomotor adaptation correlates with cerebellar white-matter microstructure. *Human Brain Mapping, 30,* 4048-4053.

Deutsch, K. M. & Newell, K. M. (2003). Deterministic and stochastic processes in children's isometric force variability. *Developmental Psychobiology, 43,* 335-345.

Deutsch, K. M. & Newell, K. M. (2004). Changes in the structure of children's isometric force variability with practice. *Journal of Experimental Child Psychology, 88,* 319-333.

Dhawale, A. K., Smith, M. A., & Ölveczky, B. P. (2017). The role of variability in motor learning. *Annual Review of Neuroscience, 40,* 479-498.

Du, W., Romano, A. G., Aloyo, V. J., & Harvey, J. A. (1995). Hypotensive stress retards associative learning in rabbits. *Neuroscience, 69,* 459-466.

Fabiani, M., Buckley, J., Gratton, G., Coles, M. G. H., Donchin, E., & Logie, R. (1989). The training of complex task performance. *Acta Psychologica, 71,* 259-299.

Faure, P. & Korn, H. (2001). Is there chaos in the brain? I. Concepts of nonlinear dynamics and methods of investigation. *Comptes Rendus de l'Académie des Sciences - Series III - Sciences de la Vie, 324,* 773-793.

Feltz, D. L. & Landers, D. M. (1983). The effects of mental practice on motor skill learning and performance: a meta-analysis. *Journal of Sport Psychology, 5,* 25-57.

Fields, R. D. (2015). A new mechanism of nervous system plasticity: activity-dependent myelination. *Nature Reviews Neuroscience, 16,* 756-767.

Fleisher, L. A., Pincus, S. M., & Rosenbaum, S. H. (1993). Approximate entropy of heart rate as a correlate of postoperative ventricular dysfunction. *Anesthesiology, 78,* 683-692.

Fogel, M. L. & Stranc, M. F. (1984). Lip function: a study of normal lip parameters. *British Journal of Plastic Surgery, 37,* 542-549.

Folstein, M., Folstein, S., & McHugh, P. R. (1975). Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Fox, E. J., Baweja, H. S., Kim, C., Kennedy, D. M., Vaillancourt, D. E., & Christou, E. A. (2013). Modulation of force below 1 Hz: age-associated differences and the effect of magnified visual feedback. *PLoS One, 8,* e55970.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12,* 1062-1068.

Franks, I. M., Wilberg, R. B., & Fishburne, G. J. (1982). Consistency and error in motor performance. *Human Movement Science, 1,* 123.

Freedman, S. E., Maas, E., Caligiuri, M. P., Wulf, G., & Robin, D. A. (2007). Internal versus external: Oral-motor performance as a function of attentional focus. *Journal of Speech, Language, and Hearing Research, 50,* 131-136.

Gabbard, C. (1998). Considering handedness in studies involving manual control. *Motor Control, 2,* 81-86.

Geeganage, C., Beavan, J., Ellender, S., & Bath, P. M. W. (2012). Interventions for dysphagia and nutritional support in acute and subacute stroke. *Cochrane Database of Systematic Reviews, 10,* CD000323.

Gilden, D. L. (2001). Cognitive emissions of 1/f noise. *Psychological Review, 108,* 33-56.

Gilden, D. L., Thornton, T. L., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science, 267,* 1837-1839.

Godwin, M. A. & Schmidt, R. A. (1971). Muscular fatigue and learning a discrete motor skill. *Research Quarterly, 42,* 374-382.

Goldberg, A. Z. (2000). *Force control of the lips during speech and non-speech tasks.* M.S. University of Wisconsin - Madison.

Goozée, J. V., Stephenson, D. K., Murdoch, B. E., Darnell, R. E., & LaPointe, L. (2005). Lingual kinematic strategies used to increase speech rate: comparison between younger and older adults. *Clinical Linguistics and Phonetics, 19,* 319-334.

Gopher, D., Weil, M., & Siegel, D. (1989). Practice under changing priorities: An approach to the training of complex skills. *Acta Psychologica, 71,* 147-177.

Grassberger, P. & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D: Nonlinear Phenomena, 9,* 189-208.

Grigos, M. I. (2009). Changes in articulator movement variability during phonemic development: a longitudinal study. *Journal of Speech, Language, and Hearing Research, 52,* 164-177.

Hall, K. G. & Magill, R. A. (1995). Variability of practice and contextual interference in motor skill learning. *Journal of Motor Behavior, 27,* 299-309.

Hamilton, A. F. d. C., Jones, K. E., & Wolpert, D. M. (2004). The scaling of motor noise with muscle strength and motor unit number in humans. *Experimental Brain Research, 157,* 417-430.

Hayes, M. H. (1996). *Statistical digital signal processing and modeling*. John Wiley & Sons.

Herzog, W. (2004). History dependence of skeletal muscle force production: Implications for movement control. *Human Movement Science, 23,* 591-604.

Holden, J. G. (2005). Gauging the fractal dimension of response times from cognitive tasks. In M.A.Riley & G. C. Van Orden (Eds.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (pp. 267-318). National Science Foundation.

Holtrop, J. L., Loucks, T. M. J., Sosnoff, J. J., & Sutton, B. P. (2014). Investigating age-related changes in fine motor control across different effectors and the impact of white matter integrity. *Neuroimage, 96,* 81-87.

Hosp, J. A. & Luft, A. R. (2011). Cortical plasticity during motor learning and recovery after ischemic stroke. *Neural Plasticity, 2011,* Article ID 871296, 9 pages.

Johansen-Berg, H., Scholz, J., & Stagg, C. J. (2010). Relevance of structural brain connectivity to learning and recovery from stroke. *Frontiers in Systems Neuroscience, 4,* 146.

Kahrilas, P. J., Logemann, J. A., Krugler, C., & Flanagan, E. (1991). Volitional augmentation of upper esophageal sphincter opening during swallowing. *American Journal of Physiology - Gastrointestinal and Liver Physiology, 260,* G450-G456.

Kal, E. C., van der Kamp, J., & Houdijk, H. (2013). External attentional focus enhances movement automatization: A comprehensive test of the constrained action hypothesis. *Human Movement Science, 32,* 527-539.

Kamen, G. (1983). The acquisition of maximal isometric plantar flexor strength: a force-time curve analysis. *Journal of Motor Behavior, 15,* 63-73.

Kasdin, N. J. (1995). Discrete simulation of colored noise and stochastic processes and $1/f^{\alpha}$ power law noise generation. *Proceedings of the IEEE, 83,* 802-827.

Kent, R. D. (2004). The uniqueness of speech among motor systems. *Clinical Linguistics and Phonetics, 18,* 495-505.

Kincses, Z. T., Johansen-Berg, H., Tomassini, V., Bosnell, R., Matthews, P. M., & Beckmann, C. F. (2008). Model-free characterization of brain functional networks for motor sequence learning using fMRI. *Neuroimage, 39,* 1950-1958.

Kluge, K. A., Harper, R. M., Schechtman, V. L., Wilson, A. J., Hoffman, H. J., & Southall, D. P. (1988). Spectral analysis assessment of respiratory sinus arrhythmia in normal infants and infants who subsequently died of sudden infant death syndrome. *Pediatric Research, 24,* 677-682.

Kukull, W. A., Larson, E. B., Teri, L., Bowen, J., McCormick, W., & Pfanschmidt, M. L. (1994). The Mini-Mental State Examination score and the clinical diagnosis of dementia. *Journal of Clinical Epidemiology, 47,* 1061-1067.

Latash, M. L., Scholz, J. P., & Schöner, G. (2007). Toward a new theory of motor synergies. *Motor Control, 11,* 276-308.

Lipsitz, L. A. (1995). Age-related changes in the "complexity" of cardiovascular dynamics: A potential marker of vulnerability to disease. *Chaos, 5,* 102-109.

Lipsitz, L. A. & Goldberger, A. L. (1992). Loss of 'complexity' and aging: potential applications of fractals and chaos theory to senescence. *Journal of the American Medical Association, 267,* 1806-1809.

Liu, C., Li, K., Zhao, L., Liu, F., Zheng, D., Liu, C. et al. (2013). Analysis of heart rate variability using fuzzy measure entropy. *Computers in Biology and Medicine, 43,* 100-108.

Long, S. H. & Olswang, L. B. (1996). Readiness and patterns of growth in children with SELI. *American Journal of Speech-Language Pathology, 5,* 79-85.

Loucks, T. M. J., Ofori, E., Grindrod, C. M., De Nil, L. F., & Sosnoff, J. J. (2010). Auditory motor integration in oral and manual effectors. *Journal of Motor Behavior, 42,* 233-239.

Loucks, T. M. J., Ofori, E., & Sosnoff, J. J. (2012). Force control under auditory feedback: effector differences and audiomotor memory. *Perceptual & Motor Skills, 114,* 915-935.

Magill, R. A. & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9,* 241-289.

Marshall, R. C. (1997). Aphasia treatment in the early postonset period: Managing our resources effectively. *American Journal of Speech-Language Pathology, 6,* 5-11.

Martin-Harris, B., McFarland, D. H., Hill, E. G., Strange, C. B., Focht, K. L., Wan, Z. et al. (2015). Respiratory-swallow training in patients with head and neck cancer. *Archives of Physical Medicine and Rehabilitation, 96,* 885-893.

Marzullo, A. C. d. M., Neto, O. P., Ballard, K. J., Robin, D. A., Chaitow, L., & Christou, E. A. (2010). Neural control of the lips differs for young and older adults following a perturbation. *Experimental Brain Research, 206,* 319-327.

McHenry, M. A., Minton, J. T., Hartley, L. L., Calhoun, K., & Barlow, S. S. (1999). Age-related changes in orofacial force generation in women. *Laryngoscope, 109,* 827-830.

Melby, P. C. (2002). *Adaptation to the edge of chaos and critical scaling in self-adjusting dynamical systems.* Ph.D. University of Illinois at Urbana-Champaign.

Miall, R. C., Weir, D. J., & Stein, J. F. (1985). Visuomotor tracking with delayed visual feedback. *Neuroscience, 16,* 511-520.

Milisen, R. L. (1954). A rationale for articulation disorders. In *The disorder of articulation: a systematic and clinical and experimental approach* (pp. 5-18).

Mizunoya, W., Oyaizu, S., Hirayama, A., & Fushiki, T. (2017). Effects of physical fatigue in mice on learning performance in a water maze. *Bioscience, Biotechnology and Biochemistry, 68,* 827-834.

Momosaki, R., Yasunaga, H., Matsui, H., Horiguchi, H., Fushimi, K., & Abo, M. (2015). Effect of dysphagia rehabilitation on oral intake in elderly patients with aspiration pneumonia. *Geriatrics & Gerontology International, 15,* 694-699.

Moon, J. B., Zebrowski, P., Robin, D. A., & Folkins, J. W. (1993). Visuomotor tracking ability of young adult speakers. *Journal of Speech and Hearing Research, 36,* 672-682.

Mukherjee, M., Koutakis, P., Siu, K.-C., Fayad, P. B., & Stergiou, N. (2013). Stroke survivors control the temporal structure of variability during reaching in dynamic environments. *Annals of Biomedical Engineering, 41,* 366-376.

Mulder, T. (2007). Motor imagery and action observation: cognitive tools for rehabilitation. *Journal of Neural Transmission, 114,* 1265-1278.

Mulder, T., Zijlstra, S., Zijlstra, W., & Hochstenbach, J. (2004). The role of motor imagery in learning a totally novel movement. *Experimental Brain Research, 154,* 211-217.

Muller, E. M., Milenkovic, P. H., & MacLeod, G. E. (1985). Perioral tissue mechanics during speech production. In J.Eisenfeld & C. DeLisi (Eds.), *Mathematics and computers in biomedical applications* (pp. 363-371). Amsterdam: Elsevier Science Publishers.

Newell, K. M., Broderick, M. P., Deutsch, K. M., & Slifkin, A. B. (2003). Task goals and change in dynamical degrees of freedom with motor learning. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 379-387.

Newell, K. M. & Vaillancourt, D. E. (2001). Dimensional change in motor learning. *Human Movement Science, 20,* 695-715.

Nicosia, M. A., Hind, J. A., Roecker, E. B., Carnes, M. L., Doyle, J., Dengel, G. A. et al. (2000). Age effects on the temporal evolution of isometric and swallowing pressure. *Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 55,* M634-M640.

Nissan, J., Gross, M. D., Shifman, A., Tzadok, L., & Assif, D. (2004). Chewing side preference as a type of hemispheric laterality. *Journal of Oral Rehabilitation, 31,* 412-416.

Noble, M., Fitts, P. M., & Warren, C. E. (1955). The frequency response of skilled subjects in a pursuit tracking task. *Journal of Experimental Psychology, 49,* 249-256.

O'Bryant, S. E., Humphreys, J. D., Smith, G. E., Ivnik, R. J., Graff-Radford, N. R., Petersen, R. C. et al. (2008). Detecting dementia with the Mini-Mental State Examination (MMSE) in highly educated individuals. *Archives of Neurology, 65,* 963-967.

Ofori, E., Loucks, T. M. J., & Sosnoff, J. J. (2012). Visuomotor and audiomotor processing in continuous force production of oral and manual effectors. *Journal of Motor Behavior, 44,* 87-96.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: spectrographic measurements. *Journal of the Acoustical Society of America, 39,* 151-168.

Özdenizci, O., Yalçin, M., Erdogan, A., Patoglu, V., Grosse-Wentrup, M., & Çetin, M. (2017). Electroencephalographic identifiers of motor adaptation learning. *Journal of Neural Engineering, 14,* 046027.

Peng, C. K., Havlin, S., Stanley, H. E., & Goldberger, A. L. (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos, 5,* 82-87.

Pierce, J. R. (1980). *An introduction to information theory: symbols, signals and noise*. (2nd, revised ed.) New York: Dover.

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the USA, 88,* 2297-2301.

Pincus, S. M. (1995). Approximate entropy (ApEn) as a complexity measure. *Chaos, 5,* 110-117.

Pincus, S. M. (1998). Approximate entropy (ApEn) as a regularity measure. In K.M.Newell & P. C. M. Molenaar (Eds.), *Applications of nonlinear dynamics to developmental process modeling* (pp. 243-268). Mahwah, NJ: Erlbaum.

Pincus, S. M. (2000). Irregularity and asynchrony in biologic network signals. *Methods in Enzymology, 321,* 149-182.

Pincus, S. M. (2001). Assessing serial irregularity and its implications for health. *Annals of the New York Academy of Science, 954,* 245-267.

Pincus, S. M., Gladstone, I. M., & Ehrenkranz, R. A. (1991). A regularity statistic for medical data analysis. *Journal of Clinical Monitoring, 7,* 335-345.

Prakash, R. S., De Leon, A. A., Mourany, L., Lee, H., Voss, M. W., Boot, W. R. et al. (2012). Examining neural correlates of skill acquisition in a complex videogame training program. *Frontiers in Human Neuroscience, 6,* 115.

Qaid, E. Y. A., Zakaria, R., Sulaiman, S. F., Mohd Yusof, N. A., Shafin, N., Othman, Z. et al. (2017). Insight into potential mechanisms of hypobaric hypoxia-induced learning and memory deficit - Lessons from rat studies [Epub ahead of print]. *Human & Experimental Toxicology,* 960327116689714.

Ramig, L. O. (1983). Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders, 16,* 217-226.

Richman, J. S. & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology - Heart and Circulatory Physiology, 278,* H2039-H2049.

Riley, M. A. & Van Orden, G. C. (2005). Tutorials in contemporary nonlinear methods. National Science Foundation.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: a review and critical reappraisal. *Psychological Bulletin, 95,* 355-386.

Schmidt, R. A. (1969). Performance and learning a gross motor skill under conditions of artificially-induced fatigue. *Research Quarterly, 40,* 185-190.

Schmidt, R. A., Lange, C., & Young, D. E. (1990). Optimizing summary knowledge of results for skill learning. *Human Movement Science, 9,* 325-348.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition, 15,* 352-359.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27,* 379-423, 623-656.

Slifkin, A. B. & Newell, K. M. (1999). Noise, information transmission, and force variability. *Journal of Experimental Psychology: Human Perception and Performance, 25,* 837-851.

Slifkin, A. B. & Newell, K. M. (2000). Variability and noise in continuous force production. *Journal of Motor Behavior, 32,* 141-150.

Slifkin, A. B., Vaillancourt, D. E., & Newell, K. M. (2000). Intermittency in the control of continuous force production. *Journal of Neurophysiology, 84,* 1708-1718.

Smith, A., Goffman, L., Zelaznik, H. N., Ying, G., & McGillem, C. (1995). Spatiotemporal stability and patterning of speech movement sequences. *Experimental Brain Research, 104,* 493-501.

Smith, A., Luschei, E. S., Denny, M., Wood, J. L., Hirano, M., & Badylak, S. F. (1993). Spectral analyses of activity of laryngeal and orofacial muscles in stutterers. *Journal of Neurology, Neurosurgery and Psychiatry, 56,* 1303-1311.

Sosnoff, J. J. & Newell, K. M. (2006a). Aging, visual intermittency, and variability in isometric force output. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 61,* 117-124.

Sosnoff, J. J. & Newell, K. M. (2006b). Are age-related increases in force variability due to decrements in strength? *Experimental Brain Research, 174,* 86-94.

Sosnoff, J. J. & Newell, K. M. (2008). Age-related loss of adaptability to fast time scales in motor variability. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 63,* 344-352.

Sosnoff, J. J. & Newell, K. M. (2011). Aging and motor variability: a test of the neural noise hypothesis. *Experimental Aging Research, 37,* 377-397.

Sosnoff, J. J., Vaillancourt, D. E., & Newell, K. M. (2004). Aging and rhythmical force output: loss of adaptive control of multiple neural oscillators. *Journal of Neurophysiology, 91,* 172-181.

Sosnoff, J. J., Valantine, A. D., & Newell, K. M. (2006). Independence between the amount and structure of variability at low force levels. *Neuroscience Letters, 392,* 165-169.

Sosnoff, J. J. & Voudrie, S. J. (2009). Practice and age-related loss of adaptability in sensorimotor performance. *Journal of Motor Behavior, 41,* 137-146.

Stanley, M. L. & Franks, I. M. (1990). Learning to organize the frequency components of a perceptual motor skill. *Human Movement Science, 9,* 291-323.

Stergiou, N. & Decker, L. M. (2011). Human movement variability, nonlinear dynamics, and pathology: is there a connection? *Human Movement Science, 30,* 869-888.

Stergiou, N., Harbourne, R. T., & Cavanaugh, J. T. (2006). Optimal movement variability: a new theoretical perspective for neurologic physical therapy. *Journal of Neurologic Physical Therapy, 30,* 120-129.

Takahata, H., Tsutsumi, K., Baba, H., Nagata, I., & Yonekura, M. (2011). Early intervention to promote oral feeding in patients with intracerebral hemorrhage: a retrospective cohort study. *BMC Neurology, 11*.

Takanokura, M. & Sakamoto, K. (2001). Physiological tremor of the upper limb segments. *European Journal of Applied Physiology, 85,* 214-225.

Testa, M., Rolando, M., & Roatta, S. (2011). Control of jaw-clenching forces in dentate subjects. *Journal of Orofacial Pain, 25,* 250-260.

Tomassini, V., Jbabdi, S., Kincses, Z. T., Bosnell, R., Douaud, G., Pozzilli, C. et al. (2011). Structural and functional bases for individual differences in motor learning. *Human Brain Mapping, 32,* 494-508.

Vaillancourt, D. E., Larsson, L., & Newell, K. M. (2003). Effects of aging on force variability, single motor unit discharge patterns, and the structure of 10, 20, and 40 Hz EMG activity. *Neurobiology of Aging, 24,* 25-35.

Vaillancourt, D. E. & Newell, K. M. (2002). Changing complexity in human behavior and physiology through aging and disease. *Neurobiology of Aging, 23,* 1-11.

Vaillancourt, D. E. & Newell, K. M. (2003). Aging and the time and frequency structure of force output variability. *Journal of Applied Physiology, 94,* 903-912.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132,* 331-350.

van Steenberghe, D., Bonte, B., Schols, H., Jacobs, R., & Schotte, A. (1991). The precision of motor control in human jaw and limb muscles during isometric contraction in the presence of visual feedback. *Archives of Oral Biology, 36,* 545-547.

Voss, R. F. & Clarke, J. (1975). '1/f noise' in music and speech. *Nature, 258,* 317-318.

Wijnants, M. L., Bosman, A. M. T., Hasselman, F., Cox, R. F. A., & Van Orden, G. C. (2009). 1/f scaling in movement time changes with practice in precision aiming. *Nonlinear Dynamics, Psychology, and Life Sciences, 13,* 79-98.

Wijnants, M. L., Hasselman, F., Cox, R. F. A., Bosman, A. M. T., & Van Orden, G. C. (2012). An interaction-dominant perspective on reading fluency and dyslexia. *Annals of Dyslexia, 62,* 100-119.

Wing, A., Daffertshofer, A., & Pressing, J. (2004). Multiple time scales in serial production of force: A tutorial on power spectral analysis of motor variability. *Human Movement Science, 23,* 569-590.

Wohlert, A. B. & Smith, A. (1998). Spatiotemporal stability of lip movements in older adult speakers. *Journal of Speech, Language, and Hearing Research, 41,* 41-50.

Wolf, A., Swift, J. B., Swinney, H. L., & Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena, 16,* 285-317.

World Health Organization (2002). Towards a common language for functioning, disability and health: ICF. World Wide Web [Electronic version]. Available: http://www.who.int/classifications/icf/icfbeginnersguide.pdf?ua=1

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

Wu, J., Srinivasan, R., Kaur, A., & Cramer, S. C. (2014). Resting-state cortical connectivity predicts motor skill acquisition. *Neuroimage, 91,* 84-90.

Wulf, G., Höß, M., & Prinz, W. (1998). Instructions for motor learning: differential effects of internal versus external focus of attention. *Journal of Motor Behavior, 30,* 169-179.

Yentes, J. M., Hunt, N., Schmid, K. K., Kaipust, J. P., McGrath, D., & Stergiou, N. (2013). The appropriate use of approximate entropy and sample entropy with short data sets. *Annals of Biomedical Engineering, 41,* 349-365.

Youmans, S. R., Youmans, G. L., & Stierwalt, J. A. G. (2009). Differences in tongue strength across age and gender: is there a diminished strength reserve? *Dysphagia, 24,* 57-65.

Appendix A: Common Measures of Time Series' Temporal and Frequency Structure

Consider two time series, shown below. The one on the left is entirely deterministic (rule-governed) and highly regular (predictable). The one on the right is a random reordering of the same values.

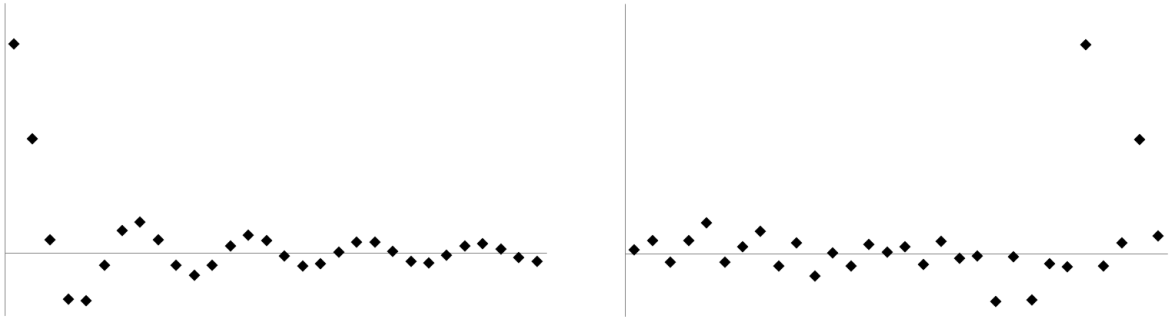

Figure 45. Time series: left, $f(t) = [\sin(t)]/t$; right, random reordering.

Any measure ignoring order of the values (e.g. mean, standard deviation) will yield the same result for both. To differentiate between them, a measure must account for their internal structure. This can be done in the time domain or, complementarily, the frequency domain. Many measurements of time series structure have been proposed (Bassingthwaighte et al., 1994; Faure & Korn, 2001), though their use in oral motor control research has been limited. Measures used primarily in non-oral motor control literature are briefly reviewed here.

Time Domain Measures

Entropy measures (approximate, sample, fuzzy and fuzzy measure entropy) are covered first as they were the most commonly used in the studies reviewed and two were used in this work. Other measures of temporal structure appear next in alphabetical order.

Entropy measures. Within a time series of length *N*, if two data runs of length *m* have similar values (within a tolerance *r*), approximate entropy ApEn(*m, r, N*) measures the likelihood that the runs will remain similar as the next data point is added to each run. ApEn can be described as measuring signal regularity, predictability or stability. Values close to zero indicate more regular, predictable output (e.g. a sine wave or other signal with power concentrated in a narrow frequency range) whereas values closer to 2 may represent a more irregular, unpredictable signal (e.g. white noise or other broadband signal; (Pincus, 1998). Chaotic systems have intermediate ApEn values (Deffeyes, Harbourne, Stuberg, & Stergiou, 2011). ApEn was developed for use with finite or noisy data in response to the limitations of measures intended for use with truly chaotic processes (Pincus, 1991; Pincus et al., 1991), such as the need for time-series length of 10^m to 30^m points (Wolf, Swift, Swinney, & Vastano, 1985) and the

inability to discriminate some noisy series (Pincus, 1998). Normalized ApEn, which calculates *r* as a fraction of the time series standard deviation, is preferred for clinical applications (Pincus, 2000; Slifkin & Newell, 1999) and has been used in studies with methods similar to those used here. Standard parameter values of $m = 2$ and $r = 0.2$ or 0.25 were used for studies described in this work unless otherwise noted. ApEn does not specify the number of degrees of a freedom in a system, but can suggest increase or decrease in the dimension of an attractor dynamic (Newell et al., 2003) and has been observed to be positively correlated with results of formal dimensional tests (Pincus, 1998).

ApEn is limited by sensitivity to all three input parameters (*m*, *r* and *N*) and a bias towards regularity (Yentes et al., 2013). Sample entropy (SampEn) simplifies the ApEn algorithm to reduce bias and dependence on *N* (Richman & Moorman, 2000) but is otherwise conceptually similar.

Fuzzy entropy (FuzzyEn) has four advantages over ApEn and SampEn (Chen, Zhuang, Yu, & Wang, 2009). It does not count vector self-matches and so lacks the bias towards regularity of ApEn. It is less sensitive to time series length. It judges similarity by vector shape rather than absolute coordinate and thus is less susceptible to nonstationarity. Finally, it classifies vectors' similarity using a fuzzy function to provide a gradual transition in similarity rating rather than the binary classification used by approximate and sample entropy. Thus, it is less sensitive to choice of *r*.

Fuzzy measure entropy (FuzzyMEn) refines FuzzyEn by measuring both local and global entropy, yielding finer discrimination of both constructed mathematical sequences and clinical vs. non-clinical populations (Liu et al., 2013).

Past investigations of motor control have typically used ApEn or SampEn to quantify temporal structure of an output signal such as steady force (Sosnoff & Newell, 2006a), postural sway (Deffeyes et al., 2011) or cyclic movements (Kal et al., 2013). FuzzyEn and FuzzyMEn have been used to examine EMG signals during manual motor tasks (Chen et al., 2009) and heart rate variability of normal and heart-failure patients (where both performed better than non-fuzzy entropy measures, and FuzzyMEn better than FuzzyEn; Liu et al., 2013). While time series lengths as short as $N = 75$ have been used in the calculation of ApEn (Pincus, 1998), $N = 200$ has elsewhere been suggested as a minimum (Yentes et al., 2013), and the motor control literature reviewed in the Introduction tended to have $N > 1000$. See Appendix D.

Correlation dimension (CD). CD (Grassberger & Procaccia, 1983) is an algorithm used to estimate fractal dimension, which can discriminate data produced from a deterministic process involving a few independent variables from randomly produced data. It characterizes purely deterministic

systems in detail, but discriminates poorly when data are noisy; this shortcoming motivated the development of ApEn (see (Bassingthwaighte et al., 1994) and (Pincus, 1998).

Detrended Fluctuation Analysis (DFA). DFA evaluates the presence of long-term autocorrelation within a time series with a scaling index, α. If the data are uncorrelated (random fluctuation), $\alpha = 0.5$; $\alpha > 0.5$ suggests persistence and $\alpha < 0.5$ suggests antipersistence (Barbado Murillo et al., 2017; Peng, Havlin, Stanley, & Goldberger, 1995).

Lyapunov Exponent (LE). LE measures sensitivity of a system's dependence upon its initial conditions, a key marker of chaotic behavior. A system is chaotic if one or more of its exponents, whose values relate to the speed with which the system loses predictability, are positive (Wolf et al., 1985).

Recurrence Quantification Analysis (RQA). RQA uses time-delayed copies of a single-variable time series as stand-ins for the dimensions of the multidimensional space needed to capture the dynamics of the system producing the time series (Riley & Van Orden, 2005). It can indicate the extent to which a system's behavior is random vs. deterministic, system stability, and the complexity of the system's attractor.

Standardized Dispersion Analysis (SDA). SDA measures the fractal dimension (FD) of a time series (the degree to which a line graph of the time series deviates from a one-dimensional straight line to occupy space in the second dimension of the graph). White noise has FD = 1.5. Correlated noise, suggesting coordination among system components, has $1 < FD < 1.5$ (Holden, 2005; Van Orden, Holden, & Turvey, 2003).

Frequency Domain Measures: Power Spectral Analyses

These measures characterize a time series in the frequency domain. Power spectral analysis (Gilden, Thornton, & Mallon, 1995; Wing, Daffertshofer, & Pressing, 2004) determines power across a range of frequencies. Controlling processes active at different timescales during force production contribute power at characteristic frequencies, e.g. around 2 Hz for visuomotor feedback (Miall, Weir, & Stein, 1985) or 8-12 Hz for upper limb physiological tremor (Takanokura & Sakamoto, 2001).

Spectral slope (SS). Spectral power can be modeled as a function of frequency, $P(f) = af^{\beta}$. SS is the slope of the log power/log frequency graph. White noise ($\beta = 0$, flat SS) reflects a random signal with no correlation between successive values (Gilden, 2001; Slifkin et al., 2000) and suggests either lack of control or relatively equal influence of contributing processes (Wing et al., 2004). Interactions between controlling processes regulate the relative amplitudes of their characteristic spectral peaks, leading to nonzero SS. Brown noise ($\beta = -2$) is characteristic of a random-walk system dominated by a few low-

frequency processes (Sosnoff & Newell, 2008; Gilden, 2001), while intermediate pink or 1/f noise ($\beta = -1$) results from limited interaction between multiple physiological control systems operating over a range of timescales (Sosnoff & Newell, 2008; Lipsitz, 1995). Rather than resulting from particular physical interactions, 1/f noise may be diagnostic of complex systems in general, representing a midpoint between extremes of order and randomness (Gilden et al., 1995; Gilden, 2001).

SS has two primary limitations. 1/f-like noise may be produced by both deterministic nonlinear and certain linear controlling processes, though there are distinguishing characteristics (Wing et al., 2004). Additionally, it is a summary measure. A more fine-grained picture of power allocation in force output is obtained through examination of proportion of power in each of multiple bandwidths.

Proportion of power (PoP). This measurement examines how signal power is divided among equal bandwidths to assess the relative contribution of low-, intermediate- and high-frequency processes to output. See Wing, Daffertshofer and Pressing (Wing et al., 2004) and Hayes (Hayes, 1996) for its calculation.

Investigations of motor control often use power spectral analysis to associate motor processes with characteristic spectral peaks or slopes (Gilden, 2001; Takanokura & Sakamoto, 2001; Voss & Clarke, 1975; Smith et al., 1993) or determine the effects of various factors such as age, feedback modality or task demand upon motor output (Sosnoff & Newell, 2008; Vaillancourt & Newell, 2003; Ofori et al., 2012). Power spectral analyses are limited by inconsistencies or bias in standard spectral estimation methods (e.g. fast Fourier transform, FFT) when data contain outliers or nonstationarities (Pincus, 2001).

Spectral degrees of freedom (SDF). SDF estimates number of frequency bins in a spectrum contributing power to the spectrum (Blackman & Tukey, 1958), calculated as the ratio of the squared sum of each bin's power estimate to the sum of squares of the bins' power estimates. A theoretical perfectly sharp peak (100% of spectral power in one bin, 0% in all other bins) would have SDF = 1, while a perfect white-noise spectrum (equal power in each bin) would have SDF equal to the number of bins. For colored-noise signals, SDF decreases as spectral slope of the signal becomes more steeply negative, i.e. as the output becomes dominated by a lower number of control processes.

Appendix B: Functional Vision and Cognition/Communication Screen

The screenshots below show the charts the participant is asked to interpret (sans arrows). Full-size, they occupy the entire width of the 39.6 cm (aspect ratio 16:9) monitor and are approximately 5 cm high.

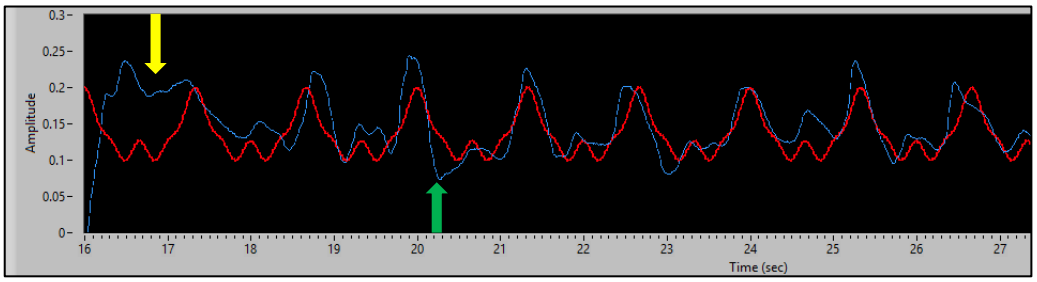

Figure 46. Multicosine simulated data used for vision screen.

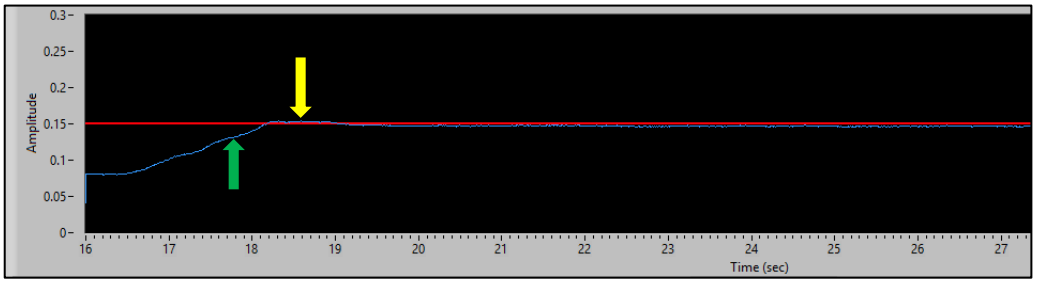

Figure 47. Constant simulated data used for vision screen.

Participant: these questions check your ability to clearly see the chart.

1. Which line is the target to match? [red line]
2. Which line shows how hard the person is pressing? [blue line]
3. When did they need to press more gently? [examples: yellow down arrows]
4. When did they need to press harder? [examples: green up arrows]

Appendix C: Transducer Images, Task Instruction Scripts and Screenshots

Transducer Images

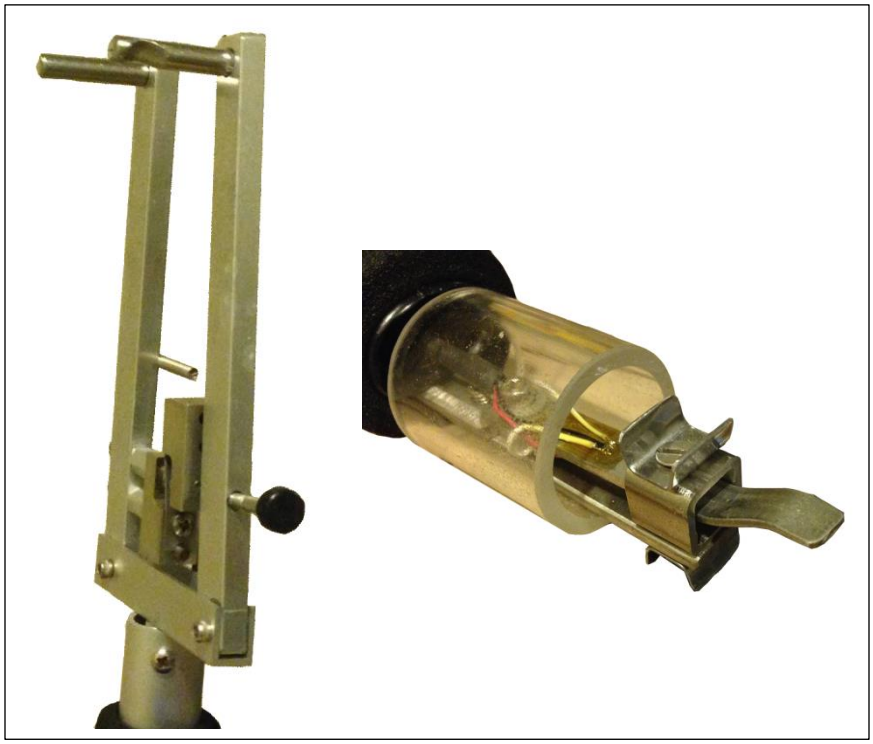

Figure 48. Left: lip transducer. Right: tongue transducer.

Maximal Voluntary Force Tasks Script

Now we will test the strength of your lips. Position these prongs just inside the corners of your mouth. Is the chair adjusted so that you are comfortable and the prongs are not pushing up or pulling down on your mouth? Keep your mouth closed and teeth together, but don't clench your jaw. You are going to purse your lips as hard as you can, while breathing out through your nose, like this. [Investigator demonstrates, not actually using transducer.] *Show me what you are going to do.* [Re-explain/re-demonstrate as needed.] *Once I hit start, do that again. When you do, the red line will go up farther the harder you purse. Make the line go up as high as you can. On the second and third tries, you'll see gray lines showing how high you've already gone, like in the demonstration. Try to get higher. You have six seconds, and you get to try three times. Do you have any questions? Say 'mm-hmm' when you're ready.*

Now we will test the strength of your tongue. Put this in your mouth, with your teeth in these grooves. Bite gently to keep the transducer in a stable position, but don't clench your jaw. You are going to push up against this part with your tongue as hard as you can, while breathing out through your nose,

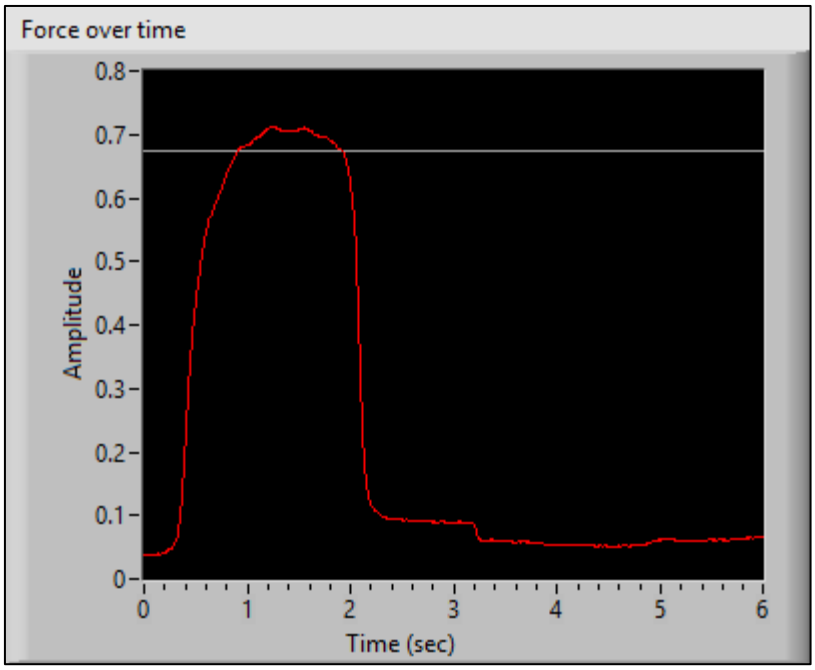

Figure 49. Maximal voluntary force visual feedback. Red line represents current attempt, updating in real time. Horizontal gray line represents maximum achieved on previous attempt.

like this. [Investigator demonstrates, not actually using transducer.] *Show me what you are going to do.* [Re-explain/re-demonstrate as needed.] *Some people have found this task to be uncomfortable. You do not have to press so hard that it hurts – if it hurts, stop.* [Remainder of script is the same as for the lip.]

Constant, Sine and Multicosine Tasks Script

For the rest of the tasks, you are going to match a line on a screen by pressing up with your tongue or pursing your lips. You will not have to press very hard. Keep your lips/tongue relaxed until I press 'start,' then try to make your line get as close to the target line as you can and stay with it. Sometimes the line will be straight. Sometimes it will be a simple up-and-down. Sometimes it will be a more complicated wave.

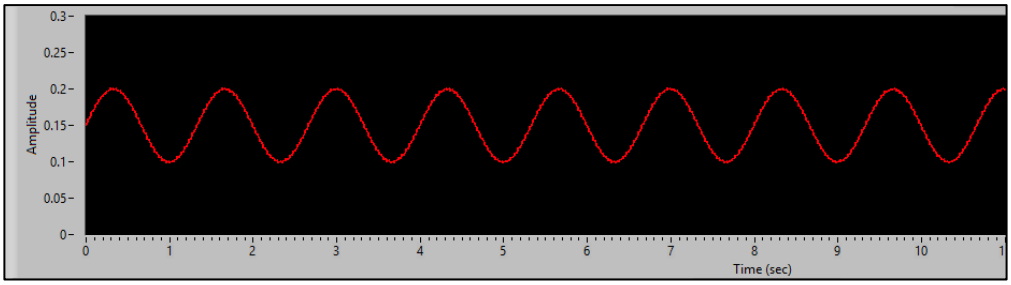

Figure 50. Sine task target. Constant and multicosine targets were shown in Appendix B.

For the variable tasks, participants were asked to vary their focus during different trial blocks.

Amplitude focus. *This time, try to match how high and low the line goes, even if your timing is a little off. So get your line this high and this low* [investigator points to maxima and minima]*, even if it happens a little before or after the target line is that high or that low.*

Timing focus (variable tasks only). *This time, try to match the timing of the target line. Get your line as high or as low as it's going to go at the same time as the target line, even if that means your line is a little too high or low. Make sure you don't let your line go all the way to the bottom of the screen.*

Amplitude and timing focus (variable tasks only). *This time, match both the timing of the target line and how high or low it is. Just stay as close to the line as you can.*

Appendix D: Fuzzy Measure Entropy and Approximate Entropy Parameters

Approximate entropy and its descendant algorithms are highly sensitive to their input parameters (*m*, *r* and *N*), and the ways in which they change with these parameters may depend upon the type of time series, e.g. periodic vs. logistic (Yentes et al., 2013; Liu et al., 2013). Yentes et al. recommended *m* = 2, *N* ≥ 200, and examination of several *r* values for ApEn (0.2 * time series standard deviation has been most often used). (Pincus, 1998) pointed out that ApEn becomes less reliable if *m* is "relatively large" or *r* is "too small," because matching runs become rare events as run length increases or match tolerance decreases. Parameter recommendations have not been established for FuzzyMEn, but similar logic should apply. Prior to data collection, the variable target time series used in this investigation were used to estimate the most appropriate parameter values for ApEn and FuzzyMEn. The constant target series was not used as its entropy was zero regardless of algorithm or parameters.

The test series were created in Matlab with F_0 = 0.75 Hz, sampling frequency F_s = 100 Hz to mimic the target time series used in data collection. Length was N = 1100 to match the length of cropped data series analyzed.

Table 29

Variable Target Series

Series	Equation form	Matlab code
Sample times	$t_i = i\Delta t = \frac{i}{F_s}$	t = [0:1099]/100;
Sine	$f_{sine}(t_i) = sin(2\pi F_0 t_i)$	s0_75 = [sin(t*pi*3/2)];
Multicosine	$f_{multicosine}(t_i) = cos(2\pi F_0 t_i) + 0.5cos(2\pi(2F_0)t_i) + 0.25cos(2\pi(4F_0)t_i)$	mc0_75 = [cos(t*pi*3/2) + 0.5*cos(t*pi*3) + 0.25*cos(t*pi*6)];

Note. This version was used for establishing ApEn and FuzzyMEn parameters. The targets used during data collection were created in LabView and had *N* = 1600 rather than *N* = 1100 to allow for cropping. LabView code is available from the author.

Liu et al. (2013) tested FuzzyMEn to show consistency and discrimination across a range of *r* values. That is, if FuzzyMEn(seriesA) > FuzzyMEn(seriesB) for one value of *r*, the algorithm is consistent if the inequality holds true as *r* varies. The algorithm discriminates well if, when tested on multiple series of known relative complexity, the values of the algorithm change consistently with changes in the series'

complexity. A similar procedure was used here. It was expected that the multicosine series would be rated more complex than the simple sine across a range of *m* and *r* values and that FuzzyMEn would show more consistent results than ApEn.

Choice of *m*

FuzzyMEn and ApEn values were evaluated at integer values of *m* from 1 to 6. Figures 51 and 52 show that FuzzyMEn discriminated the two series at all values of *m* and consistently ranked the multicosine series more complex than the sine series, while ApEn discriminated well only for *m* = 1 and *m* = 2. Consequently the conventional value of *m* = 2 was used in this study.

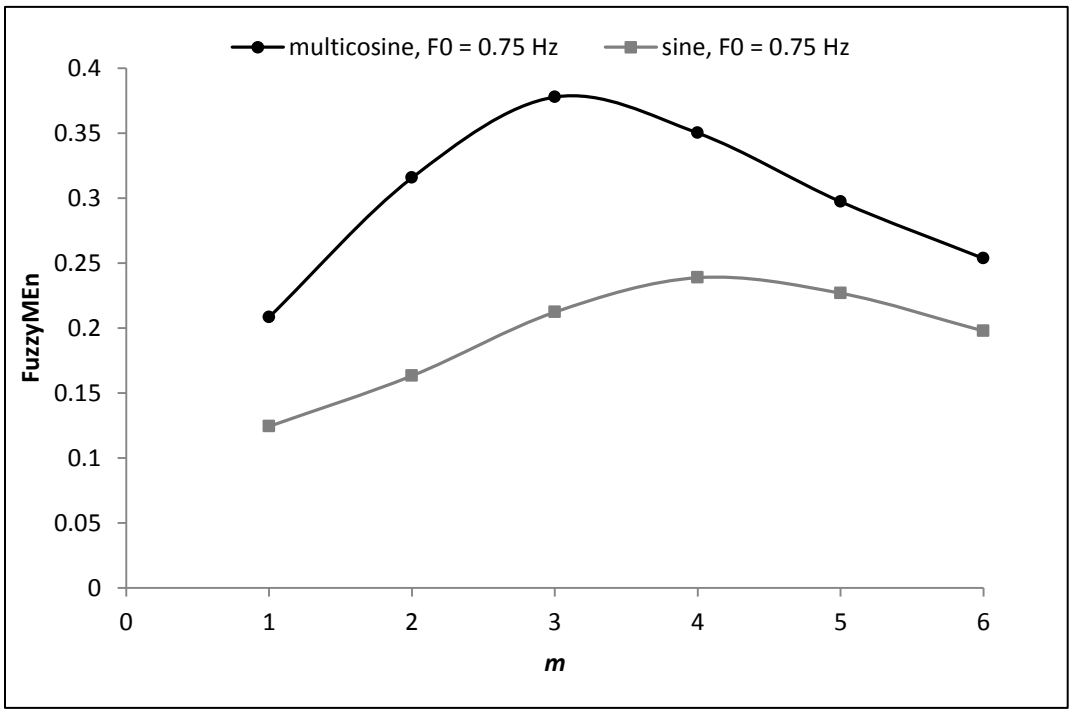

Figure 51. FuzzyMEn as a function of *m* for sine and multicosine series. For both series, *N* = 1100, *r* = 0.2, F_s = 100 Hz.

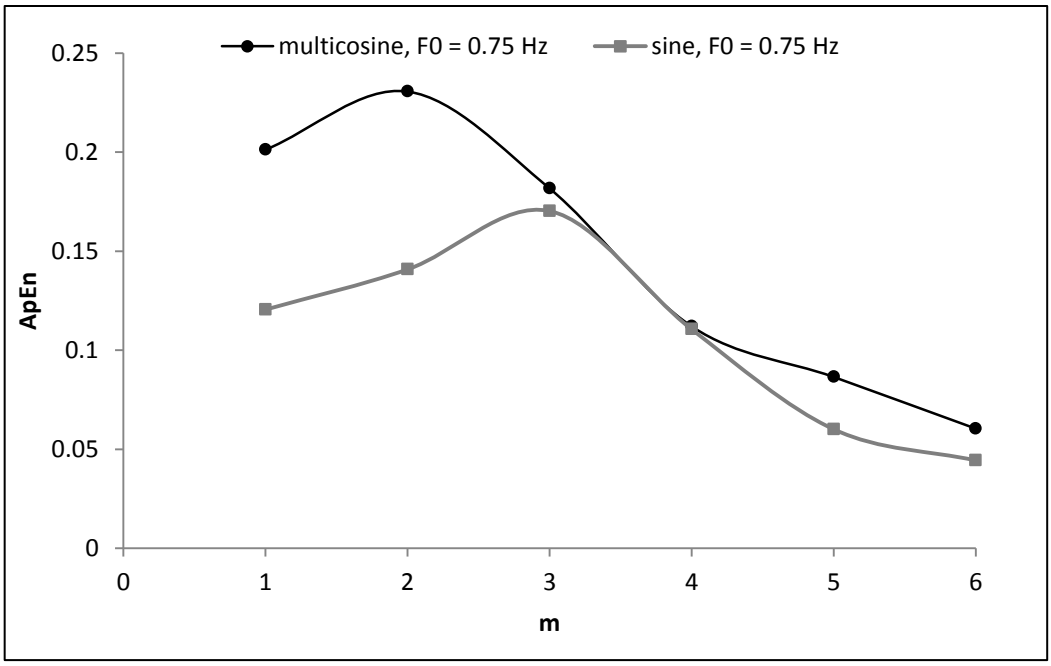

Figure 52. ApEn as a function of m for sine and multicosine series. For both series, N = 1100, r = 0.2, Fs = 100 Hz.

Choice of *r*

FuzzyMEn and ApEn values were calculated for each series with *r* (ApEn) and $r_L = r_F$ (FuzzyMEn) ranging from 0.01 to 1, incrementing by 0.01, $m = 2$, $N = 1100$ (Figures 53 and 54). The figures suggest that FuzzyMEn consistently ranks the complexity of the multicosine series above that of the simple sine series across *r*, while ApEn is consistent for these series only for $r \geq 0.11$. Both discriminate the series at the conventional value of $r = 0.2$, though FuzzyMEn changes more smoothly. This investigation therefore used $r = 0.2$ for comparability to previous literature.

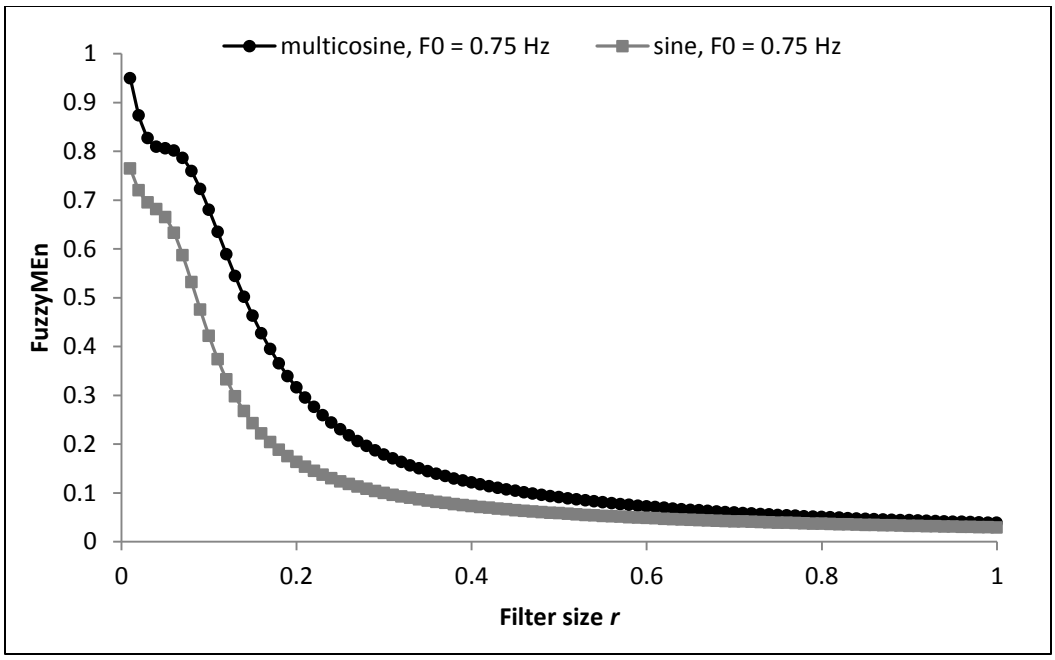

Figure 53. FuzzyMEn for tested target series as a function of filter size, 0.01 ≤ *r* ≤ 1.0. For all series, m = 2, N = 1100.

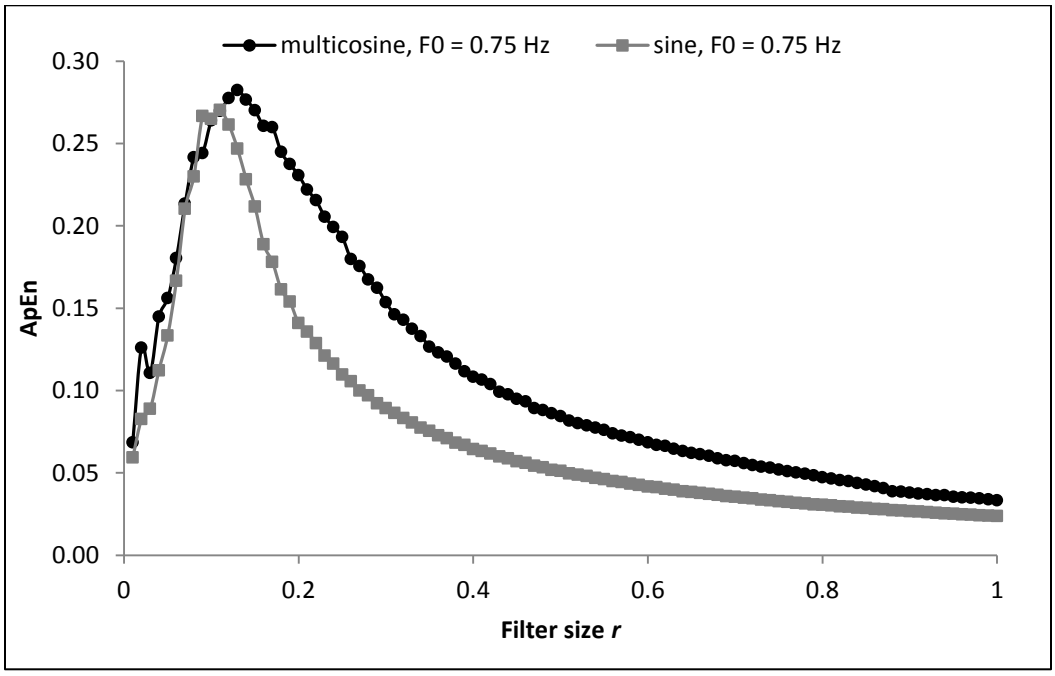

Figure 54. ApEn for tested target series as a function of filter size, 0.01 ≤ *r* ≤ 1.0. For all series, *m* = 2, *N* = 1100.

Choice of N

Given a fixed sampling rate (100 Hz), choice of *N* determines trial length. *N* was chosen to fit the following constraints:

1. Recommended $N \geq 200$ for nonlinear analyses (Yentes et al., 2013)
2. LabVIEW 2013 (the software used to display target and transducer signals and record data) has no type of graphical display inherently permitting a static pattern to be shown while another dynamic pattern develops. A sweep-refreshing graph was used to display the target signal only on its first run, then to refresh the target with the transducer signal visible on its second. The target signal and its refreshed version had to align, meaning trial length needed to accommodate an integer number of repeats of either variable-force pattern. Both had F0 = 0.75 Hz, meaning the lowest integer number of seconds with an integer number of periods was four (4 seconds = 3 periods). Thus trial length had to be a multiple of four seconds.
3. Following previous work (Bronson-Lowe et al., 2013) the first 4 and last 1 seconds were to be discarded to avoid ramp-up and end-anticipatory effects. Thus the multiple of four seconds chosen had to be large enough to allow for the subtraction of five seconds of data (N = 500). Eight seconds would leave N = 300 (2 x 4 sec x 100 samples/sec – 500 samples).
4. Trial length needed to balance providing a large enough N for entropy calculations with short enough length to avoid fatiguing participants and to permit a useful number of trials in a reasonable amount of time for participant convenience and cost.

Fuzzy measure entropy and approximate entropy of the variable target signals were evaluated for *N* from 50 to 1100 (corresponding to a trial length of 16 seconds – 5 seconds' cropped data) in increments of 25. The following figures show that both measures ranked the target series' complexity correctly and discriminated well once *N* reached 100. The algorithms rapidly converged on stable values for both series, though small fluctuations continued to occur at higher values of *N*.

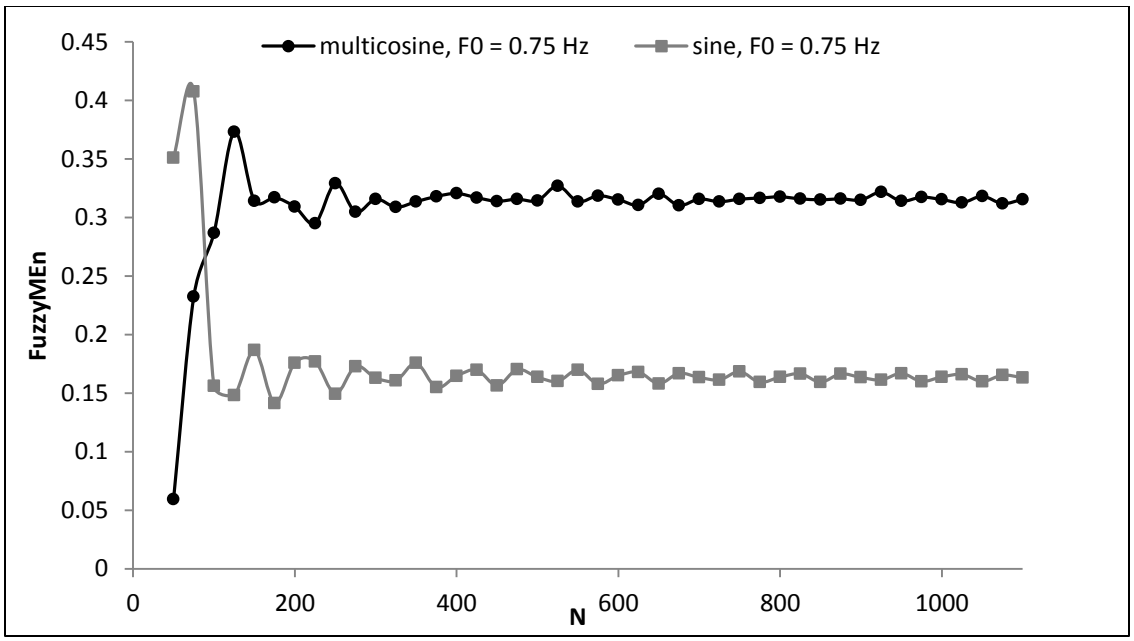

Figure 55. FuzzyMEn as a function of N for sine and multicosine series. For both series, m = 2, r = 0.2, F_s = 100 Hz.

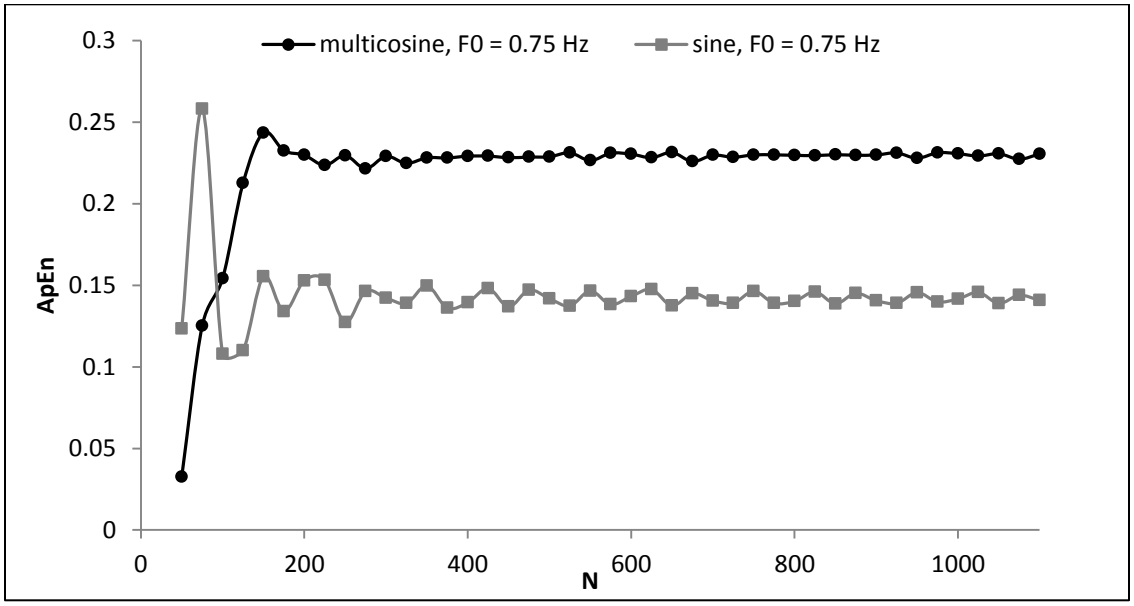

Figure 56. ApEn as a function of N for sine and multicosine series. For both series, m = 2, r = 0.2, F_s = 100 Hz.

To judge whether these fluctuations justified extending the series to yet higher N, fluctuation size was quantified by calculating the range of ApEn and FuzzyMEn values over the last five points (N = 1000 to N = 1100) as a percentage of their maximum. All were below 5%; see Table 30.

Table 30

Range of ApEn and FuzzyMEn Values over Five Values of N (1000 to 1100 in Increments of 25) as a Percentage of Their Maxima over Those Values of N.

Series	ApEn	FuzzyMEn
Sine	4.62%	3.58%
Multicosine	1.47%	1.96%

ApEn and FuzzyMEn values were next calculated out to *N* = 10,000 (*m* = 2, *r* = 0.2) for the sine series, since their variation by *N* was greater for that series. Little additional change was seen for either algorithm: ApEn(*N* = 1100) differed from ApEn(*N* = 10,000) by -0.11%, while FuzzyMEn(*N* = 1100) was 0.25% greater than FuzzyMEn(*N* = 10,000). The figure below shows how negligible the additional reduction in entropy fluctuation is over the range up to N = 10,000. The small benefit of increasing *N* was judged not to outweigh its costs. Trial length was not reduced, on the other hand, because of the likelihood that fluctuations in entropy based on series length would be greater for human fine-force data than for simple, highly regular mathematical functions.

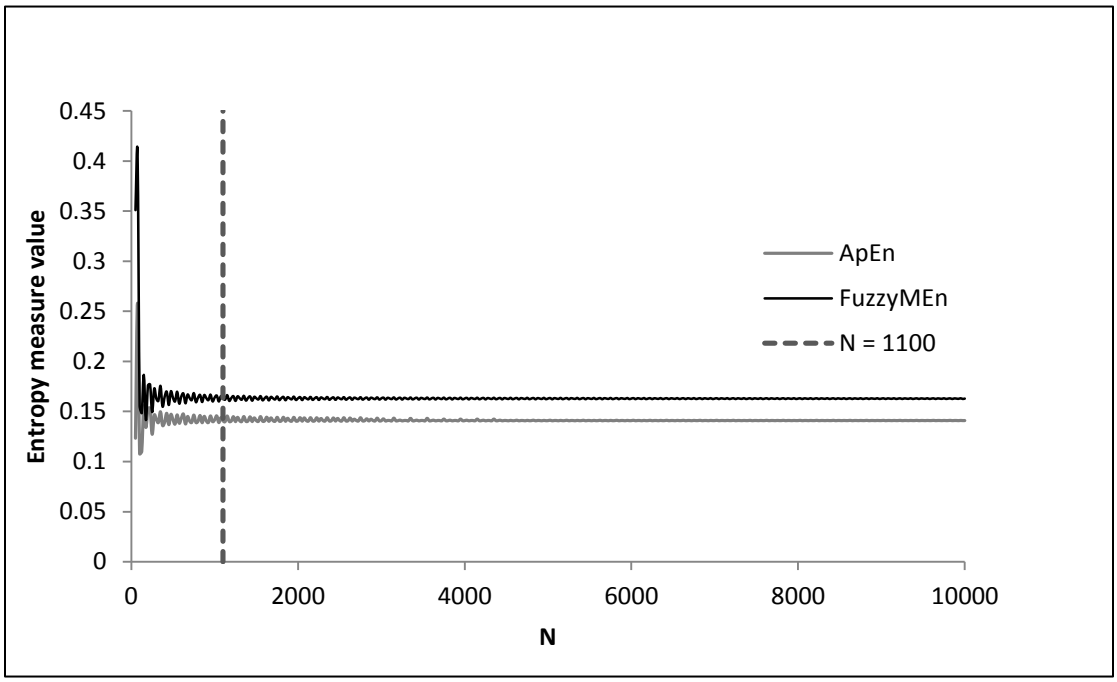

Figure 57. Approximate and fuzzy measure entropy by series length: sine, F0 = 0.75 Hz, Fs = 100 Hz, m = 2, r = 0.2.

Choice of Fuzzy Function

For fuzzy measure entropy, the fuzzy function used to classify vectors as close or distant was

$$f(d_{ij}, r) = e^{-\left(d_{ij}/r\right)^n} \quad (5)$$

as recommended by (Chen et al., 2009) and (Liu et al., 2013). d_{ij} describes distance between vectors *i* and *j* and *r* is a proportion of time series standard deviation. Values of n > 1 preferentially weight vector pairs with the lowest values of $\frac{d_{ij}}{r}$ (those with greatest similarity) while giving a smooth transition to more extreme values. See Figure 58. Following Liu et al. (2013), this work used $n_L = 3$, $n_F = 2$.

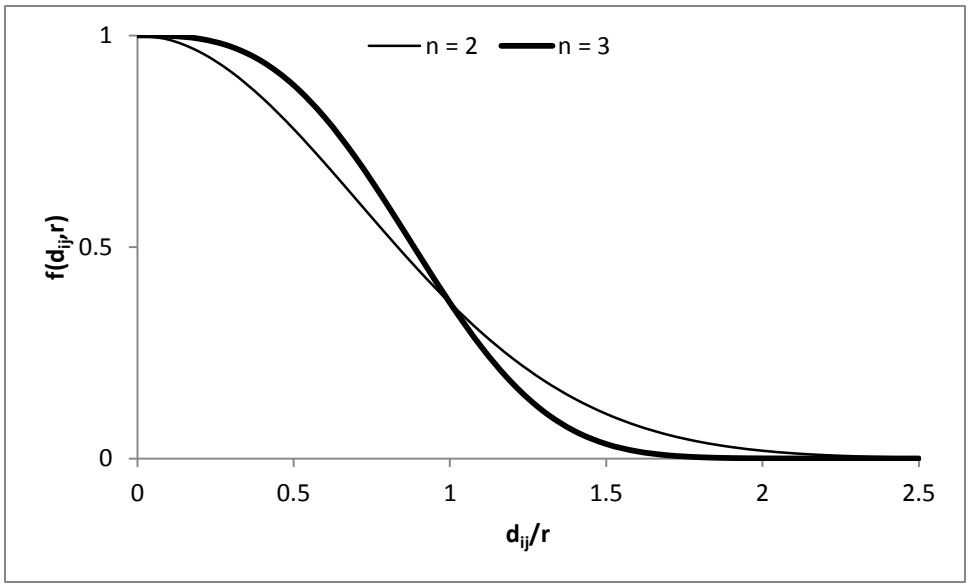

Figure 58. Fuzzy classification function $f(d_{ij}, r)$.

Appendix E: Spectral Analysis Testing

To test the function of the spectral analysis code and parameters, two types of signals were generated and analyzed using the parameters described in the power spectral analysis discussed in the Measures section.. All test signals had length = 1100, sampling rate 100 Hz to match experimental data.

First, colored noise signals were used to test accuracy of spectral slope calculation. Results are reported in Table 31. Least-squares linear fit of expected vs. mean calculated slopes yielded slope = 0.881(expected) - 0.0302, $r^2 = 0.981$.

Table 31

Testing Spectral Analysis Parameters: Signals with Approximately Known Spectral Slope

		Spectral slope	
Noise color	Generation method	Expected	Analyzed (mean ± std. error, 1000 trials)
White	Custom code[a]	0	-0.00101 ± 0.00159
	LabVIEW Uniform White Noise.vi, default parameters		-0.00207 ± 0.00166
	Spectral synthesis approximation[b]		0.00177 ± 0.00195
	Created in Sound Forge		0.000657 ± 0.00156
Pink	Filtering of white noise signal #1 above[c]	-1	-0.898 ± 0.00297
	Filtering of white noise signal #2 above[d]		-0.895 ± 0.00280
	Spectral synthesis approximation		-0.788 ± 0.00343
	Created in Sound Forge		-1.01 ± 0.00271
Brown	Running sum of white noise signal #1 above	-2	-1.87 ± 0.00647
	Filtering of white noise signal #2 above[c]		-1.85 ± 0.00691
	Spectral synthesis approximation		-1.96 ± 0.00449
	Created in Sound Forge		-1.95 ± 0.00554
Black	Spectral synthesis approximation	-3	-2.38 ± 0.0123[e]

[a] Samples from a uniform random distribution $0 \le x < 1$, with 0.5 subtracted so that mean = 0. [b] (Cuddington & Yodzis, 1999) [c] Infinite impulse response filter (Kasdin, 1995), appropriate for expected slopes $0 < x < 2$. [d] LabVIEW's Inverse f Filter.vi, filter specifications 0.1 – 50 Hz (Nyquist frequency), 10th order. This VI uses a zero-pole method of filter generation based on (Corsini & Saletti, 1988), appropriate for expected slopes $-2 < x < 2$. [e] Bimodal distribution with peaks around -2.2 and -3.2

Second, the target signals used for the variable force-matching tasks (sine and multicosine; see Fine-force stimuli in the Methods chapter) were analyzed to test accuracy of spectral peak detection. Results are reported in Table 32. Both target signals' spectra contained the expected number of peaks,

at the closest frequency bin boundaries to the targeted frequencies, with less than 4% divergence from the expected relative heights.

Table 32

Testing Spectral Analysis Parameters: Signals with Known Peaks

	Peaks expected		Peaks found	
Signal	Frequencies (Hz)	Heights (relative)	Frequencies (Hz)	Heights (relative)
Sine	0.75	1	0.781	1
Multicosine	0.75	1	0.781	1
	1.5	0.25	1.56	0.241
	3.0	0.0625	2.93	0.0607

Appendix F: Bonferroni-Adjusted Significance Criteria by Level of Analysis and Preceding Pattern of Significance

Analysis paths (full model significant finding → follow-up analysis):
A: 3-way interaction → simple 2-way interactions → simple simple main effects → pairwise comparisons
B: 2-way interaction → simple main effects → pairwise comparisons
C. Main effect → pairwise comparisons
Pairwise comparisons were used only for investigating main effects of task and frequency band (each 3 levels, compared in pairs). They were not necessary for factors with two levels.

Analysis paths by corresponding steps	Factor(s) tested	Bonferroni correction
A0. 3-way interaction	A. F_1 x F_2 x F_3	
B0. 2-way interaction	B. F_1 x F_2	n/a: $p < 0.05$
C0. Main effect	C. F_1	
A1. Simple 2-way interaction	A. F_2 x F_3 at each level of F_1	$0.05 / L_1$
B1. Simple main effect	B. F_2 at each level of F_1	$0.05 / L_1$
C1. Pairwise comparisons among main effect levels	C. pairs of levels of F_1	0.05 / (# pairs)
A2. Simple simple main effect	A. F_3 at each level of F_2, within each level of F_1 for which F_2 x F_3 was significant	$0.05 / (L_2 \times L_{1*})$
B2. Pairwise comparisons among simple main effect levels	B. pairs of levels of F_2, within each level of F_1 for which F_2 was significant	$0.05 / (\#\ pairs \times L_{1*})$
A3. Pairwise comparisons among simple simple main effect levels	A. pairs of levels of F_3, within each level of F_2 for which F_3 was significant, within each level of F_1 for which F_2 x F_3 was significant	$\frac{0.05}{\#\ pairs\ x\ L_{2*}\ x\ L_{1*}}$

Substitution rules:	Correction	Criterion
# pairs = 3	0.05 / (2 x 1 x 1)	$p < 0.025$
L_i = 2 or 3	0.05 / (2 x 1 x 2)	$p < 0.013$
L_{i*} = 1, 2 or 3	0.05 / (2 x 2 x 2)	$p < 0.006$
Arithmetic is shown for step 3. Ignore '1' terms to find step 1-2 equivalents.	0.05 / (3 x 1 x 1)	$p < 0.017$
	0.05 / (3 x 2 x 1)	$p < 0.008$
	0.05 / (3 x 2 x 2)	$p < 0.004$

Note. Analysis paths are shown first; then correction algebra is shown for corresponding steps of the various paths (step number indicates number of terms in correction denominator); arithmetic is collated last.

Abbreviations:
F_i = any factor
L_i = number of levels in F_i = number of tests performed, regardless of significance
L_{i*} = number of levels in F_i for which tested term was significant

Appendix G: Copies of IRB Documentation

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Office of the Vice Chancellor for Research

Office for the Protection of Research Subjects
528 East Green Street
Suite 203
Champaign, IL 61820

February 25, 2015

Torrey Loucks
Speech & Hearing Science
228 SHS Bldg
906 S Goodwin Ave
Champaign, IL 61820

RE: *Project MOV: Manual and oral variability*
IRB Protocol Number: 09284

Dear Dr. Loucks:

This letter authorizes the use of human subjects in your continuing project entitled *Project MOV: Manual and oral variability*. The University of Illinois at Urbana-Champaign Institutional Review Board (IRB) approved the protocol as described in your IRB-1 application, by expedited continuing review. The expiration date for this protocol, IRB number 09284, is 02/22/2016. The risk designation applied to your project is *no more than minimal risk*. Certification of approval is available upon request.

Copies of the attached date-stamped consent form(s) must be used in obtaining informed consent. If there is a need to revise or alter the consent form(s), please submit the revised form(s) for IRB review, approval, and date-stamping prior to use.

Under applicable regulations, no changes to procedures involving human subjects may be made without prior IRB review and approval. The regulations also require that you promptly notify the IRB of any problems involving human subjects, including unanticipated side effects, adverse reactions, and any injuries or complications that arise during the project.

If you have any questions about the IRB process, or if you need assistance at any time, please feel free to contact me at the OPRS office, or visit our Web site at http://www.irb.illinois.edu.

Sincerely,

Anita Balgopal, PhD
Director, Office for the Protection of Research Subjects

Attachment(s)

c: Jacob Sosnoff
Christina Bronson-Lowe
Linda West

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

Department of Kinesiology and Community Health

Speech and Hearing Science
901 S. Sixth St., Champaign, IL
217-244-6576 *office*
e-mail: tloucks@illinois.edu

Project title: Project MOV: Motor and Oral-motor Variability: Complexity and Motor Learning sub-study

Purpose of the study:

The main purpose of the present investigation is to examine the association between an individual's ability to move their mouth and their ability to move their fingers and arm. The sub-study's purpose is to examine how an individual's ability to control the lips and tongue predicts learning of a new lip/tongue task. This particular project is conducted under guidance of professor Torrey Loucks, in the Department of Speech and Hearing Science at the University of Illinois at Urbana-Champaign. If you agree to take part in this research, you will be asked to press your lips and tongue against a measuring device while being presented with a visual display on how hard you are pressing. You will need to accurately match the desired target force as indicated by the computer monitor. To assess the lip/tongue force a special measuring device will be placed in your mouth. You will be given feedback regarding your performance on all of the tasks. You will be asked to complete the test and surveys on three separate testing sessions. Overall the completion of each testing session will take about 90 minutes for the first two sessions and 60 minutes for the third. This research will provide information concerning the mechanisms contributing to the maintenance of force output. This study is being conducted by the University of Illinois Motor Control Laboratory and the Neurospeech Laboratory.

What you will do:

You will be asked to maintain force production at varying force levels using different visual feedback displays by pursing your lips or raising your tongue. For the oral motor force assessment a device will be placed in your mouth. You will be seated in a chair in front of a computer monitor. On the monitor you will see your force output and a target force level and varying types of visual feedback. You will be asked to minimize any deviations from the force target. A score indicating your performance will be provided at the end of every trial on the first day and every five trials on the second day. **We would also like to record your voice as you produce a single vowel sound such as 'a' in the word bought or 'i' in beat or 'u' in boot. You will be asked to produce these sounds**

using a high pitch and a low pitch. We would like to produce the sound for about 10 seconds or until you want to take another breath. Your voice will be recorded with a microphone. The voice recordings will take less than 5 minutes. Although there is no direct benefit by participating, your participation will add to the knowledge base concerning the mechanisms underlying the control of oral movement. You will receive $8.25/hour up to $35 or course credit for completing the study. If you chose to not complete the entire experiment you will be paid for the amount of time you have completed (not to exceed $35). Your parking expenses related to participation will also be paid for.

Your information and confidentiality:

Your identity will be protected. You will be assigned an ID number and any data collected will only be identified by that ID number. We will collect two types of information from you. One is your personal identification information, the other is the experimental data. Your personal information (name and contact information) will only be used to contact you during this study. This information will not be used in the data analysis, nor will it be released to others. Personal information collected will be kept separate in a locked cabinet. Your identity will be kept confidential to the extent required by law.

This research is basic research designed to help us understand the mechanisms underlying the control of force production. In any report stemming from this research only group data will be reported. No individual will be identified in any reports or presentations. The results of this study will provide information concerning the control processes responsible for the control of movement. It is anticipated that the results of this investigation will be disseminated to the lay and scientific community via presentations and scholarly publications. If you wish to have a copy of the results of this study please contact us.

Time requirements and risks:

The experiment typically consists of three testing sessions. The first two sessions include measuring how hard you can purse your lips or press upwards with your tongue, and learning and practicing new tasks requiring only gentle pressure with your lips and tongue. The third session tests how well you have learned the new tasks and whether you have gained skill in very similar tasks.

If we find that three sessions are insufficient for learning the tasks, we may ask if you would be willing to come back for up to two more sessions. You would be paid at the same rate as for the first two sessions, up to a total (across all your sessions) of up to $45 for four sessions or $55 for five. You will be asked each day whether you are willing to continue; consenting now does not mean that you are obligated to complete all sessions.

We do not anticipate any physical risks greater than normal risks associated with daily life. However, at anytime during data collection if you feel any risk or discomfort, you

may withdraw from testing. You will be given breaks throughout the data collection period as needed.

Refusal or withdrawal of participation:

Your participation is completely voluntary and you may withdraw at any time without penalty or loss of benefits to which you are otherwise entitled, and without influencing your relationship with the University of Illinois in any manner. If you choose to not complete the entire experiment you will be paid for the amount of time you have completed at the rate of $2.06/15 minutes (not to exceed $35, $45 or $55 depending on days of participation as outlined above).

If you have any questions concerning this study please feel free to contact Dr. Torrey Loucks (email: tloucks@illinois.edu or phone: 217-244-6576). You will be given a copy of the consent form upon completion of the study.

For any questions regarding the rights of a research subject, please contact the Institutional Review Board Office of the University of Illinois at Urbana-Champaign at (217) 333-2670 or irb@illinois.edu. Collect calls are welcome if you live outside of the local calling area.

I, _____, am 18 years of age or older and understand the above information and voluntarily consent to my participation.

_____ _____
Signature of Participant Date

_____ _____
Signature of Witness Date

Sincerely,

Torrey M. Loucks, PhD – Responsible Principal Investigator
Assistant Professor
Department of Speech and Hearing Science