MOTOR LEARNING OF ARTICULATOR TRAJECTORIES IN THE PRODUCTION OF NOVEL UTTERANCES

Mark Tiede[a,b], *Christine Mooshammer*[a], *Louis Goldstein*[a,c],
Stefanie Shattuck-Hufnagel[b] & *Joseph Perkell*[b]

[a]Haskins Laboratories, USA; [b]Speech Comm. Group, MIT R.L.E., USA;
[c]Department of Linguistics, U.S.C., USA
tiede@haskins.yale.edu

ABSTRACT

In this study EMA has been used to observe speech articulator movements during successive productions of a variety of novel polysyllabic nonsense words conforming to English phonotactics. Analysis of sensor trajectories comparing initial and final repetitions shows in general reduction of overall duration and distance travelled, lower variability, and fewer acceleration peaks. Comparison of consonant closure timings delimited using velocity extrema suggests that as fluency increases, the overlap between adjacent consonantal gestures also increases.

Keywords: speech motor control, speech production, motor learning, EMA, STI, FDA

1. INTRODUCTION

Motor learning is the process of mastering goal-directed patterns of movement to improve performance, whose general characteristics include increasing movement speed and consistency while simultaneously improving performance accuracy [9]. In the context of the skilled movements associated with speech, little is known about the optimization process through which speakers attain fluency with a novel utterance. Previous work assessing motor learning on a single nonce word (/θɹeɪm•po•fɹa•mo•dɪs/ “thraimpoframodis”) has reported that kinematic duration and variability are reduced as a function of practice, while overall production accuracy increases [7].

This work extends the scope of that study, by examining nine additional utterances, and by investigating new approaches for assessing the time course and extent of learned fluency. The target utterances are novel, in that both the words themselves and their syllabic constituents are not words of English. Because the syllables adhere to English phonotactics they are however *potential* words (or compounds) of English. This distinction ensures that the learning task consists of adapting existing patterns of articulatory movement available to fluent English speakers, as opposed to developing entirely new patterns (e.g., an English monolingual confronting Polish “Szczebrzeszyn”).

2. METHODS

2.1. Participants

Four female and three male young adult native speakers of American English with normal hearing and no apparent speech deficits were recruited and paid to participate in this study.

2.2. Materials

Participants were asked to produce the 10 nonsense words shown in Table 1. These ranged from four to six syllables in length and all syllables were constructed to conform with English phonotactics.

Table 1: Stimulus words used. The complexity measure is the sum over each word of 1 point/syllable, 1 point/coda, and 1 point for each cluster.

Syllables	Complexity	Stimulus Word
5	9	thraimpoframodis
4	11	blertdoibtradisp
5	12	krubdrathraimtrobeel
5	12	proomfreckpoyfrokosp
5	13	mabeprotvaspreedrep
5	14	praksteebdrongspovasp
6	14	borkspasprumlankperwoo
5	15	splonktretsfavepodasp
6	15	splumprofresproompadoip
6	16	splampredfrothaspodisp

Within a trial a given word was shown to the participant on a computer screen. Each of the ten words was presented once, in randomized order, within a block. Each block was repeated from 8 to 10 times within the experiment, interspersed with additional unrelated material.

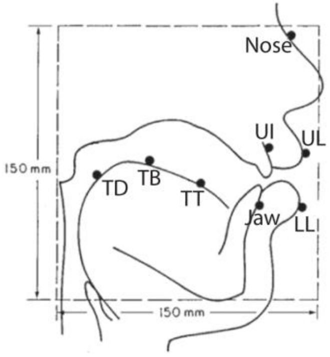
All participants produced at least one production of the target word per trial. Four participants were instructed to produce each word two or three times sequentially within each trial;

however, only the first production from each trial is analyzed here. Two participants received prompting for suggested pronunciation in the form of an audio recording played once at the start of each trial; all others were instructed to infer pronunciation from their knowledge of English orthography.

2.3. Recordings

An electromagnetic articulometer (EMA; Carstens AG500) was used to transduce the time-varying locations of sensors attached to each participant's speech articulators at a 200 Hz sampling rate (see Figure 1). Concurrently recorded audio was sampled at 16 kHz. Movement data were corrected for head motion and aligned to each participant's occlusal plane.

Figure 1: EMA sensor placement. Two additional head motion reference sensors (not shown) were located on the mastoid processes.



2.4. Data analysis

Data analysis was performed within Matlab (The MathWorks) using custom developed procedures. Individual productions were segmented using consistent articulatory landmarks associated with the first and last phones of each utterance. For example, minimum lip aperture (Euclidean distance between UL and LL) was used to delimit instances of "blertdoibtradisp." Two criteria were used to exclude productions if necessary: excessive duration (greater than 2 std. deviations from the median for that participant/utterance), and gross mispronunciation (more than 2 deviations from that participant's preferred pronunciation of the word, judged by listening and ignoring stress).

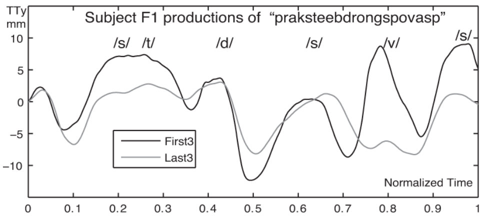
2.4.1. Variability measures

Production variability was assessed using two methods applied to the vertical component of movement for the tongue dorsum (TD), tongue tip (TT), Jaw and lower lip (LL) sensors.

In the first of these each signal was amplitude normalized by subtracting its mean and dividing by its standard deviation, then linearly time normalized to a standard number of samples. These samples were then binned into 50 groups (2% intervals), and standard deviations were computed across the normalized samples within each bin. The Spatio-Temporal Index (STI) is the sum of these 50 standard deviations [8]. STI compared across different groups (early vs. late repetitions) provides a metric for observing changes in articulatory fluency: as fluency increases, articulatory trajectories are expected to converge on a stable pattern, and overall variability (as reflected by the STI measure) should decrease.

The second approach used nonlinear time warping. The algorithm evolved a set of strictly increasing and smoothness-constrained transformations of time (the warping functions) such that the distance of each time-warped (normalized) signal from the average across all normalized signals of a group was minimized [3]. The individual warping functions provide the phase difference of each signal from the average, and the amplitude difference is obtained between corresponding samples of each time-warped signal and the mean signal (see Figure 2). From these separate STI-like measures for phase and amplitude were computed by summing across binned standard deviations at 2% intervals as above. To distinguish these from the linear STI measure these will be referred to as the FDA (Functional Data Analysis; [6]) measures of variability. Again, as indices of variability, both are expected to decrease as articulatory fluency improves.

Figure 2: Mean TTy for first/last three repetitions aligned through time warping on significant events. Note loss of unnecessary /v/ gesture after practice.



2.4.2. Economy of effort measures

Additional measures were evaluated directly on the un-normalized sensor trajectories, including utterance duration, distance travelled (sensor path integral), and the number of signal acceleration peaks. In these measures, the expectation is that learning reflected in increased fluency is driven by a general ‘economy of effort’ principle [5], and thus each is expected to decrease with increased fluency.

2.4.3. Measures of relative phasing

The effect of practice on licensed overlap between constriction gestures formed with non-competing articulators was assessed through relative phasing, by labeling maximum constriction points for /b/ and /t/ within the /d/$_1$:/d/$_2$ context in productions of “blert**d**oi**bt**ra**d**isp” using velocity thresholding criteria (cf. [2]). Improved fluency is expected to lead to a decrease in relative phasing.

3. RESULTS

Reported measures were computed over the mean of the first three and last three productions, after pruning for excessive duration and dysfluency trials. Words are listed in order of complexity.

3.1. Variability

Variability measures combine results computed on the vertical component of movement for TD, TT, Jaw, and LL sensors. 5 of 7 subjects and 9 of 10 words show decreased STI after practice (Fig. 3).

Figure 3: STI collapsed over words (left), and over subjects (right).

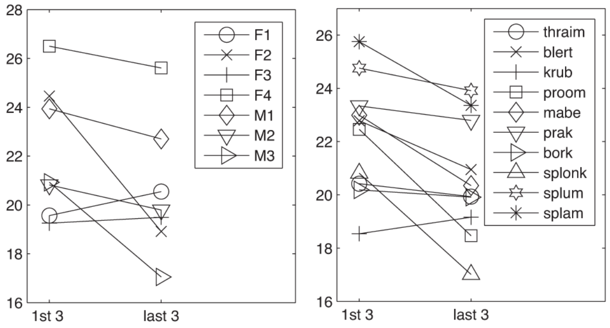


Figure 4: FDA amplitude variability collapsed over words (left), and over subjects (right).

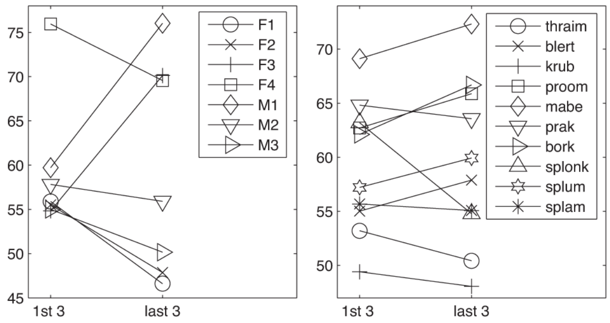
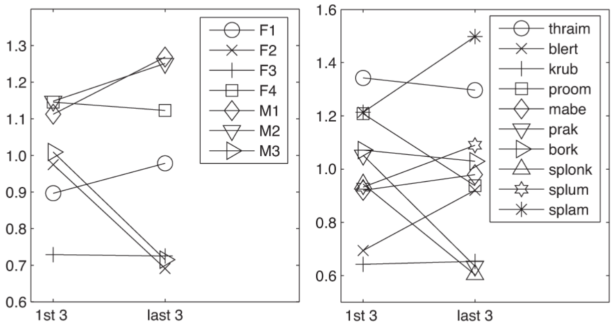


Figure 5: FDA phase variability collapsed over words (left), and over subjects (right).



FDA amplitudes show 5 of 7 subjects and 5 of 10 words with decreased amplitude variability after practice (Fig. 4). Of the two subjects that show an increase in variability, M1 received an initial audio prompt. FDA phasing measures (Fig. 5) show 4 of 6 subjects and 5 of 9 words with decreased variability after practice. Subject F3 and word “krubdrathraimtrobeel” show effectively no difference.

3.2. Economy of effort

EOE measures have been converted to z-scores computed by subject over repetitions to facilitate comparison, averaged across the first/last 3 trials.

Utterance durations (Fig. 6) show unambiguous effects of increased fluency, as 6 of 7 subjects and all 10 words had decreased production times.

Figure 6: Normalized utterance durations collapsed over words (left), and over subjects (right).

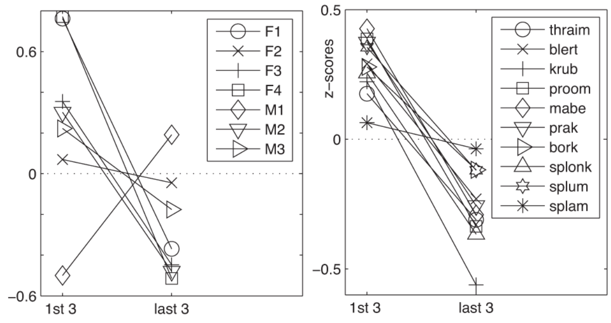
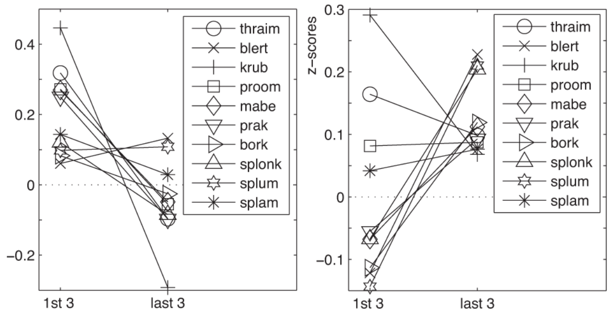
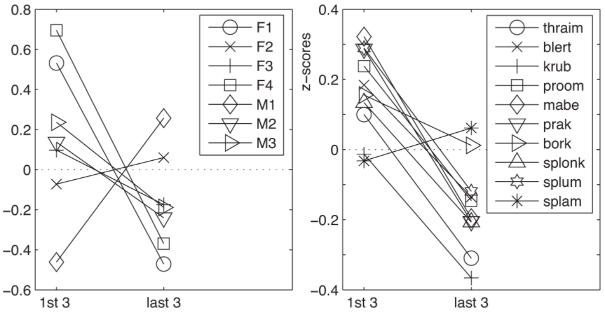


Figure 7: Normalized distance (left) and average sensor velocity (right) collapsed over subjects.



The remaining EOE measures combine results from sensor trajectories for TD, TT, Jaw, and LL, collapsed over subjects. With two exceptions, overall articulator distance travelled declined with practice, and average sensor velocity increased (Fig. 7). The number of acceleration peaks (Fig. 8) uniformly decreased, with the exception of the most complex utterance ("splampredfrothaspo-disp"). Collapsed across words, all but two subjects showed declines in acceleration peaks; of these M1 received audio prompting.

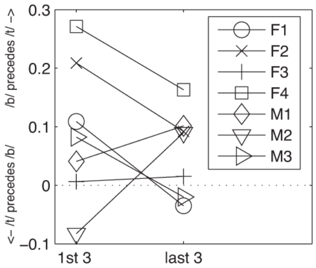
Figure 8: Normalized number of acceleration peaks collapsed over words (left), and over subjects (right).



3.3. Relative phasing

Gestural overlap was assessed over instances of "blert**d**oi**bt**ra**d**isp" by converting /b/ and /t/ maximum constriction offsets to percentages of /d/$_1$:/d/$_2$ context duration and subtracting. This relative phasing measure decreased with practice for most subjects; two subjects with initial inverted /tb/ order corrected this with practice.

Figure 9: /b/:/t/ phasing in "blert**d**oi**bt**ra**d**isp" as percentage differences (/t/-/b/) of enclosing /d/$_1$:/d/$_2$ context duration averaged over the first / last three repetitions.



4. DISCUSSION

In general the results obtained confirm general expectations for an unsupervised motor learning task: repeated productions of the respective utterances lead to greater efficiency and accuracy in the observed patterns of articulator movements, manifested particularly clearly in reduced duration and the number of acceleration peaks (an index of movement smoothness [4]). Increased overlap expressed as the relative timing between constriction gestures, known to occur with changes in rate and style [1], is here shown to be a function of improved fluency as well. Some effects of utterance complexity were observed, with the more complicated sequences incompletely optimized given the number of repetitions available. Finally, the attempt to facilitate learning by presenting a suggested pronunciation to two of the participants as a pre-production audio prompt was a failure: F3 did not achieve fluency more quickly than those who did not receive the prompt, and M1 showed indications that such prompting inhibited optimization (particularly as reflected by increased FDA amplitude variability and number of acceleration peaks). Speakers apparently prefer to find their own paths to fluency when mastering novel utterances.

5. ACKNOWLEDGMENTS

This work was supported by NIH research grant R01-DC008780.

6. REFERENCES

[1] Browman, C., Goldstein, L. 1990. Tiers in articulatory phonology, with some implications for casual speech. In Kingston, J., Beckman, M. (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press, 341-376.

[2] Byrd, D., Lee, S. Campos-Astorkiza, R. 2008. Phrase boundary effects on the temporal kinematics of sequential tongue tip consonants. *J. Acoust. Soc. Am.* 123, 4456-4465.

[3] Lucero, J., Munhall, K., Gracco, V., Ramsay, J. 1997. On the registration of time and the patterning of speech movements. *J. Speech Lang. Hear. Res*. 40, 1111-1117.

[4] Nelson, W.L. 1983. Physical principles for economies of skilled movements. *Biol. Cybern*. 46, 135-147.

[5] Perkell, J.S., Zandipour M., Matthies, M.L., Lane, H. 2002. Economy of effort in different speaking conditions. I. A preliminary study of intersubject differences and modeling issues. *J. Acoust. Soc. Am*. 112, 1627-1641.

[6] Ramsay, J.O., Munhall K.G., Gracco, V.L., Ostry, D.J. 1996. Functional data analyses of lip motion. *J. Acoust. Soc. Am*. 99, 3718-3727.

[7] Schulz, G.M., Stein, L., Micallef, R. 2001. Speech motor learning: preliminary data. *Clin. Ling. Phon*. 15, 157-161.

[8] Smith, A., Goffman, L., Zelaznik, H.N., Ying, G., McGillem, C. 1995. Spatiotemporal stability and patterning of speech movement sequences. *Exp. Brain Res*. 104, 493-501.

[9] Wolpert, D.M., Ghahramani, Z., Flanagan, J.R. 2001. Perspectives and problems in motor learning. *Trends Cogn. Sci*. 5, 487-494.