Precise Feedback Control Underlies Sensorimotor Learning in Speech

Chris Vaughn¹ and Sazzad M. Nasir²

¹Georgia State University, Atlanta, GA
²Northwestern University, Evanston, IL

Corresponding Author: Sazzad M. Nasir
Department of Communication Sciences and Disorders
Northwestern University
2240 Campus Drive
Evanston, IL 60208
Phone: 847 491 2450
Email: s-nasir@northwestern.edu

Copyright © 2014 by the American Physiological Society.
ABSTRACT

Acquiring the skill of speaking in another language, or for that matter a child's learning to talk, does not follow a single recipe. People learn by variable amounts. A major component of speech learnability seems to be sensing precise feedback errors to correct subsequent utterances that help maintain speech goals. We have tested this idea in a speech motor learning paradigm under altered auditory feedback, in which subjects repeated a word while their auditory feedback was changed online. Subjects learned the task to variable degrees, with some simply failing to learn. We assessed feedback contribution by computing one-lag covariance between formant trajectories of the current feedback and the following utterance that was found to be a significant predictor of learning. Our findings rely upon a novel use of information rich formant trajectories in evaluating speech motor learning, and argue for their relevance in auditory speech goals of vowel sounds.

INTRODUCTION

What are the learnability factors of speech? As children learn to talk many factors, such as their ability to process speech sounds, or how well they can repeat what they heard, may contribute to their learning progression. While a great deal is known about language learnability (Gierut 2007; Perfors et al. 2011; Clark and Lappin 2013) the sensorimotor basis of speech learnability is not well understood. An important component in speech learning seems to be the detection of errors received through sensory feedback, and use that information to rectify mistakes in utterances produced subsequently for maintaining speech goals (Perkell 2012; Guenther 2006; Hickok et al. 2011). We provide evidence that a better learner is more able to correct for the errors received through feedback, and thus shows better control in the use of feedback in updating subsequent production. This process helps speakers in maintaining their speech goals. A direct demonstration of this intuitive concept is lacking in the context of speech learning.
We tested the idea that precise feedback control contributes to speech motor learning under a paradigm (Houde and Jordan 1998; Jones and Munhall 2005; Purcell and Munhall 2006a; Purcell and Munhall 2006b; Rochet-Capellan and Ostry 2011; Villacorta et al. 2007; Lametti et al. 2012) of altered auditory feedback. In the speech motor task the auditory feedback of a target vowel sound was altered in real time as subjects repeated the word “Head.” Over the course of training they changed their production in order to compensate for the perturbation that helps maintain their speech goals. Subjects learned the task by variable amounts and a portion of them failed to compensate for the perturbation. We expect that individual differences in learnability partly lie in their use of current feedback in updating subsequent utterances trials that can be measured by one-lag covariance (van Beers 2009, 2012) between vowel formant trajectories of the current feedback and the subsequent production. The reason that formant trajectory is an important factor in speech learnability is that it carries detailed information about the evolution of a vowel sound that provides more precise representation of its auditory goals. It is found that over the course of training learners show a progressive decrease in one-lag covariance, which means that their production becomes less and less similar to the feedback received, thus, reaching their speech goals. The non-learners, on the other hand, do not compensate for the perturbation, produce utterances that remain very similar to the feedback received and, consequently, they follow the perturbation and fail to reach their intended goals. Their one-lag covariance does not change significantly over the course of learning. The one-lag covariance between formant trajectories, thus, constitutes a significant factor of speech learnability.

Although it may seem surprising that formant trajectory (Cai et al. 2011; Weismer et al. 1992) informs about the speech motor learnability, it speaks to the complex landscape of speech goals (Lametti et al. 2012). It is now firmly established that speech has somatosensory goals
(Tremblay et al. 2003; Nasir and Ostry 2008; Guenther and Vladusich 2012; Perkell 2012), as during talking we take care to correct for deviations in the movement on the scale of few millimeters. It has been shown that subjects carry multiple speech goals (Lametti et al. 2012) spread over both somatosensory and auditory domains. As movement trajectories contribute to somatosensory goals, it is therefore expected that their auditory counterparts, formant trajectories, would constitute auditory goals of the vowel sounds. More precisely, detailed information contained in the trajectories is used to form stable speech goals upon repetition of the same speech sounds. Hence, trial-by-trial variations in formant dynamics carry information on the reliance of auditory feedback in vowel production. The inclusion of formant dynamics also better aligns auditory goals with their perceptual counterparts, since perception is mediated by formant transitions, information of which can only be gleaned from formant trajectories (Lindbolm and Studdert-Kennedy 1967; Liberman et al, 1956; Strange 1989). Based on our findings we anticipate that the studies of formant dynamics will provide constraints for better understanding of the neural bases of speech production. Using formant trajectories it is possible to map the variability landscape in vowel production that could then be used to assess how sensory feedback contributes to the ongoing control of speech sounds and the role it plays in motor speech disorders.

MATERIALS AND METHODS

**Subjects:** 31 native English speakers (26 females) participated in the experiments (29.5 ± 3.05 yrs). The Northwestern University Institution Review Board approved the experimental protocol. Subjects reported normal speech and hearing and gave informed consent before participating. All subjects were naive to the experimental manipulation upon initial recruitment.

**Experimental setup and task:** Subjects were seated in front of a computer monitor in a sound attenuating booth during testing. Subjects wore headphones (Shure SRH-840) and spoke into a
unidirectional microphone (Sennheiser). Subjects were instructed to speak the word displayed
on the monitor at a conversational loudness level that was monitored by a digital level meter.
Each word prompt lasted 1.5 s and the inter-prompt interval was approximately 2.5 s.

The auditory adaptation experiment had three phases that were run in blocks of 12 trials. The
first was a 6-block baseline phase, the second a 15-block training phase, and the third a 5-block
after-effect phase. In the baseline and after-effect phases subjects received normal auditory
feedback while in the training phase auditory feedback was perturbed by shifting the first two
formant frequencies of the target vowel /æ/ in the utterance ‘HEAD’ towards /I/. The estimated
formant shifts were obtained during a screening phase prior to the experiment when the vowel
space was mapped out, and on average the first and the second formant frequencies were
shifted by -264.6 Hz and 193.0 Hz respectively. In the initial screening participants were
presented words that span the entire vowel space. We used seven different vowels with each
vowel sound presented 6 times in random order. This screening was used to generate the base
formant frequencies for each vowel and choose the best linear predictive coding (LPC) model
order that ranged from 8 to 12. We assessed formant information during baseline, at the first
30% and the last 30% of training utterances constituting respectively training start and training
end phases, as well as the first 30% of the after-effect phase.

Auditory perturbations: We altered vowel formant frequencies in real time during speech
production following the methods of Purcell and Munhall (2006a, 2006b). The LabView real time
language implemented in the National Instruments PXI system can estimate the formant
frequencies using the Burg algorithm and update the LPC parameters of the speech signal
10000 times per second. The participant’s voice was recorded at 10 kHz to obtain offline
estimates of the formant frequencies. The intensity of the feedback signal played back to
participants was adjusted to 80 dB in order to minimize bone conducted unaltered auditory feedback.

**Acoustical analysis and assessing adaptation:** The first and second formant frequencies were extracted for each spoken word using PRAAT and customized Matlab routines. For each subject the formant frequencies were normalized relative to their baseline average. To assess speech motor learning we focused only on the first formant frequency since auditory feedback perturbation predominantly affected the vowel height.

In order to quantify adaptation we wanted to make sure that the formant frequency at the end of training phase was sufficiently different from those at the baseline phase. This was done on a per subject basis using a one way ANOVA followed by Tukey's HSD posthoc test according to which 16 subjects altered their production such that the average formant frequency at the end training was significantly higher ($p < 0.01$) than the baseline.

The amount of learning was quantified by taking the average difference between the formant frequencies of the end of training and baseline phases that was then normalized by the baseline average. The learners therefore showed a positive amount of learning while for the non-learners it was mostly negative.

It should be noted that learners and non-learners received similar amount of auditory shifts ($t(29) = 1.22; p > 0.5$) and, hence, the observed differences between the two groups can’t be attributed to the differences in feedback perturbation received.

**Formant trajectory analysis:** For each vowel utterance we extracted the first formant trajectory that spanned the voiced segment by using PRAAT. We then time-normalized all the trajectories
by resampling them at 200 sample points. These time-normalized trajectories were next
smoothed through convoluting with a 5 point Hanning window and repositioned by subtracting
the subject’s average baseline formant frequency. For each utterance we extracted two formant
trajectories, one for production and the other from the feedback.

One-lag cross-covariance estimates were generated for each utterance by computing
covariance between the formant trajectory during the production of the $n$-th trial ($P_n$) and the
feedback trajectory at the previous ($n-1$)-th trial ($F_{n-1}$). The maximum of absolute covariance was
used for further analysis.

Additionally cross-correlation coefficients were generated between the two time series
consisting of the first formant frequencies from the production trials and those from the
preceding feedback trials.

Statistical analysis: A split-plot ANOVA followed by Tukey’s HSD posthoc tests were used to
assess significant statistical differences for each covariance measure by doing a comparison
between learners and non-learners, and among the experimental phases (baseline, early and
late training, and after-effect). The same split-plot test was used to assess the differences in
terms of learning between learners and non-learners.

RESULTS
We have tested the idea that detailed feedback information as conveyed by formant trajectory is
a significant determinant of learnability in speech motor learning. The experiment began with a
baseline phase in which subjects repeated the target word “Head” under normal auditory
feedback. Figure 1A depicts the real time perturbation of auditory feedback of the target vowel
sound. It shows that during auditory perturbation the first formant trajectory (cyan) is shifted
downwards (red). Figure 1B shows the mean first formant frequencies from a representative subject obtained by averaging over the utterances in a block. During the training phase, following the baseline (blue), auditory feedback (cyan) was altered in real time by shifting downwards the formant frequencies that altered the target vowel /æ/ more towards /I/. Subjects who learned the task compensated for the auditory perturbation by progressively shifting the first formant frequency upwards (red) so that the auditory feedback received at the end of the task became similar to /æ/. An after-effect phase (black) followed the training phase, in which auditory feedback returned to normal. To assess learning we focused on baseline, early and end training and after-effect phases of the experiment (indicated by gray bar in Figure 1B) and performed ANOVA followed by posthoc tests (see Methods) on a per-subject basis. Subjects learned the speech motor task by variable amounts including some subjects who didn’t learn the task at all. The learners showed significant compensation to offset the auditory perturbation and the non-learners, on the other hand, either failed to offset or simply followed the perturbation. The learners, thus, eventually reach their intended speech goals, while the non-learners do not. Figure 1C shows first formant frequency (±SEM) normalized relative to the baseline, and averaged across subjects, for both learners (N = 16) and non-learners (N = 15). In the baseline phase formant production is similar between the groups (blue: learner; cyan: non-learner). In the training phase learners clearly showed compensation to the auditory perturbation by shifting their production upward (red), while the non-learners did not show any such compensation, and their production (magenta) seem to follow the downward auditory perturbation. During the after-effect phase, in the presence of normal auditory feedback, production gradually approached the baseline level for both groups, but they differed in the slope of their approach (black: learner; gray: non-learner). A split-plot ANOVA revealed significant differences between the two groups (F(1, 87) = 69, p < 0.0001).

In order to understand how feedback is used in rectifying production errors that help maintain
speech goals we next took into account information about entire formant dynamics as incorporated in their trajectories. For each utterance the time-normalized formant trajectory spanning the voiced part of the segment was obtained first, as time normalization takes into account the differences in vowel duration (see Methods). Figures 2A and 2B respectively show two example traces of formant trajectories at different experimental phases from a representative learner and a non-learner. The traces were chosen from the sample traces towards the end of each experimental phase from a typical learner and a non-learner. Notice that for the learner there occurs relative to baseline (blue) a progressive upward shift of the formant trajectories as auditory perturbation is introduced (red: early training; cyan: late training), while the trajectory approaches baseline once again in after-effect (gray). On the other hand the formant trajectories for the non-learner do not show any consistent shift over the course of training. Figure 2C shows averaged formant trajectories (± SEM) for both groups at different experimental phases, demonstrating substantial differences between learners (blue) and non-learners (gray).

To examine the role of feedback in learning, we computed one-lag covariance between the formant trajectories associated with the current feedback and the subsequent production trials. Figure 3A displays probability distributions of one-lag covariance over all the trials for learners and non-learners at different experimental phases. Notice that under normal auditory feedback, during the baseline and after-effect phases, the distributions are strikingly similar between the two groups, and in after-effect both distributions are skewed to the right. In contrast, in the presence of auditory feedback alteration the two groups are highly dissimilar; with the learners the distribution is skewed to the left, while for the non-learners it is to the right. This presumably highlights differences in the feedback control, and hence error compensation, between the two groups.
Figure 3B shows one-lag covariance averaged across subjects for different experimental phases for both groups. Consistent with the probability distributions, both groups have similar covariance values for the baseline and the after-effect when subjects receive normal auditory feedback. Over the course of training however, during which feedback is altered, the covariance progressively decreases for the learners; this illustrates that the current production of a speech utterance becomes increasingly dissimilar to the feedback received in the previous trial. A split-plot ANOVA between the groups and the experimental phases revealed significant main effects (Group: F(1, 116) = 20.53, p < 0.00001, Phase: F(3,116) = 67.79, p<0.00001) with an interaction (F(3, 116) = 17.04, p<0.00001). Further post-hoc analyses (Tukey’s HSD) found significant group differences (p < 0.01) at the beginning and the end of the training, while no differences were observed between the baseline and the after-effect. Moreover, for the learners the one-lag covariance over the course of training differed significantly from the baseline and the after-effect (p < 0.01). For the non-learners no consistent patterns in significant differences were observed except that the early training phase differed from the baseline.

To evaluate whether one-lag covariance is a predictor of learning we correlated between the amount of learning and the covariance (Figure 3C) and found a significant negative correlation between the two (r = - 0.83; p < 0.0001). Hence, the better the learner is, the higher the dissimilarity is that builds up over the course of training between the current feedback and the subsequent production.

Finally, we carried out additional one-lag correlation analyses using point estimators such as the average of the formant trajectory for each spoken utterance. Figure 4A shows a modified one-lag cross correlation computed for all experimental phases and between the two groups (van Beers 2009, 2012). Statistical analyses didn’t reveal any differences between the groups (F(1, 116) = 0.04, p > 0.8). This finding further underscores the utility of formant trajectories in
revealing differences in feedback control that lead to speech motor learnability under altered auditory feedback. We also computed one-lag cross-correlation (Figure 4B) that failed to find any significant differences between the two groups (F(1,116) =0.36, p > 0.5). Thus the role of trial-by-trial feedback error in speech learning seems only to be seen when one takes into account the entire dynamics of formant evolution.

DISCUSSION

In summary, the ability to effectively use feedback in modifying subsequent production differed significantly between the learners and the non-learners of the speech motor learning task under altered auditory feedback. Over the course of learning, the learners demonstrated a progressive decrease in trial-by-trial one-lag covariance between the formant trajectories of the auditory feedback received and the following speech utterance. On the other hand, the non-learners simply maintained a high degree of similarity between auditory feedback and the subsequent speech production during the baseline and throughout the training period. The one-lag covariance between feedback and subsequent formant production was also found to be a significant predictor of learning: over the course of learning better learners showed a larger decrease in the covariance while the non-learners showed an increase.

We also analyzed the correlation between feedback and subsequent production using average formant frequency for each trial, rather than the entire trajectory, and failed to detect any differences between learners and non-learners. Additionally, using the average formant frequencies we computed one-lag correlations, which have been successful in understanding motor tasks such as a dart throw, but were of little use in distinguishing learners from the non-learners in the speech motor task under study (van Beers 2009, 2012). The differences in feedback control between the two groups appear only to be picked up by taking formant trajectories into account.
We have used time-normalization in order to take into account vowel duration differences across the utterances. Our findings, however, are not an artifact of time normalization. We obtained qualitatively similar results, namely that learners and non-learners differed significantly in the way they control their auditory feedback, just by using formant trajectories without any time-normalization.

Our analyses focused only on the first formant trajectories, since in the learning task it was the vowel height that was the primary target for perturbation. The first formant frequency was perturbed in proportion by about 50%, compared to less than 10% in the second, and therefore the role of the second formant frequency in the motor task is expected to be minimal.

It is rather remarkable that little attention has been paid to learning variability in speech motor tasks. As in any motor task, speech is rather variable across and within the talkers, but what contributes to their variability? Under the assumption that speech motor learning largely entails trial-by-trial manipulation of feedback error, it is quite plausible that online feedback control in its ability to use current feedback to rectify subsequent production error constitutes a major ingredient of learning variability in speech. This intuitive hypothesis on the sensorimotor basis of speech learning variability is confirmed by our findings. It is interesting to note that learners, by using information rich formant trajectories, progressively decorrelate between feedback received and subsequent production in order to compensate for the auditory perturbation introduced. Our findings thus demonstrate the use of feedback in modifying production as a primary determinant of speech motor learning.

What do formant trajectories tell us about the speech goals? Over the last several years it has been convincingly demonstrated that speech goals are not only auditory but also somatosensory in nature (Tremblay et al. 2003; Nasir and Ostry 2008; Perkell 1992). Speakers
care about making fairly precise speech movements, which underpin their somatosensory goals in speech. It is similarly likely that auditory speech goals are constituted from detailed acoustical information over time, just as somatosensory speech goals are made of the movement trajectory of speech articulators. As for the vowel sounds, the formant trajectories would then amount to more detailed auditory speech goals. During talking the nervous system has access to fine-grain information about both the articulatory movements made and the ensuing acoustical dynamics. Consequently, as we produce an utterance we care not only about making accurate speech movements, but also that what we hear is correct over its entire course of production.

A great deal is known about the role of formant trajectories in speech perception. Formant transitions carry information that aid in the perception of the vowel sounds (Strange 1989; Weismer et al. 1988). An analogous role of formant trajectories in speech production is underexplored. Our analyses on feedback control will help us understand the role of the trajectories in speech, and will add to the growing body of literature on speech motor learning.

Although feedback control seems to underlie observed differences in speech motor learning, it could be just one among other exogenous and endogenous factors that affect one’s ability to learn a speech motor task. It has been shown that linguistic experiences or task instructions influence outcomes in speech motor learning (Mitsuya et al. 2013; Munhall et al. 2009). Similarly, attention or task fatigue could also contribute to speech learnability. It will be worthwhile to undertake a systematic approach to identify factors that contribute to the differences in speech motor learning. This might allow us to understand better speech development or help devise novel therapeutic targets for motor speech disorders.

It could be argued that what is being reported here are actually differences in speech motor
adaptation as opposed to learning. In similar altered feedback paradigms of motor learning involving human arm movement there are evidences for generalization, retention and interference suggesting that the object of study is learning (Shadmehr and Mussa-Ivaldi 1994; Mattar and Ostry, 2007; Brashers-Krug et al. 1996). Moreover, evidences of after-effect and context specificity observed in speech motor learning argue strongly in favor of learning as opposed to adaptation (Rochet-Capellan et al. 2012).

REFERENCES


Figure 1. Subjects learn the speech motor task by variable amounts. A. Real-time perturbation of the first formant frequency of the target vowel sound in “Head.” The formant trajectory is
shown during baseline (cyan) and when it is shifted downwards (red). B. First formant frequency from a representative subject at different experimental phases. Over the course of training under the downward auditory feedback perturbation (cyan) the subject alters his production by changing the first formant frequency upward (red) relative to the baseline (blue). During the after-effect (black) production approaches the baseline once feedback returns to normal level. C. Normalized first formant frequency averaged across subjects. It can be seen that subjects learn the task by different amount. While both learners and non-learners start at the similar baseline level (blue vs. cyan) learners oppose the perturbation by shifting production upward (red) and the non-learners seem to follow the perturbation with a downward shift in their production (magenta). Both groups similarly have different after-effect profiles (black vs. gray). The line thickness represents SEM.

Figure 2. Formant trajectory is sensitive to learning differences. A. Example traces of normalized formant trajectories from a learner at different phases of the motor task. Trajectories start to differ with the onset of auditory perturbation (red) relative to baseline (blue), and shifts upwards at the end of training (cyan), before becoming similar to baseline again during after-effect phase (gray). B. Non-learners do not show any systematic trend in the relative shift of their trajectories at various experimental phases. C. Formant trajectories averaged across subjects reveal that learners (blue) shift their production upwards during training, while for the non-learners they remain relatively unchanged. The line thickness represents SEM.

Figure 3. One-lag covariance computed using formant trajectories is a predictor of learning. A. Probability distribution of one-lag covariance for learners and non-learners computed by taking all the trials at different experimental phases. A. Both groups (black vs. gray) start out with similar distributions during the baseline, while for the learners it shifts to the left during the training phase and skews to the right for both groups during the after-effect. Both learners
(black) and non-learners (gray) have overlapping distribution during normal auditory feedback. B. Learners show progressive decrease in covariance over the course of learning. While both groups (black vs. gray) have similar covariance during baseline, at the beginning of training learners show a decrease relative to the baseline that becomes more pronounced at the end of training. The non-learners on the other hand don’t show any such systematic pattern. C. The amount of learning correlates well with one-lag covariance; the worse a non-learner is the higher the covariance is and vice versa.

Figure 4. One-lag correlations using point estimators such as the average formant frequency don’t pick up learner and non-learner differences. A. One-lag correlations were computed using current feedback and next production following a modified method of computing cross-correlation (van Beers 2009, 2012). There are no differences between the two groups (black vs. gray) at various experimental phases. B. One-lag cross correlations were computed using average formant trajectory of current feedback and next production. Again no-differences are found between the two groups.
A

B

C

Learner

Non-learner

Normalized time

Frequency (Hz)

Normalized time

Frequency (Hz)

Normalized time

Baseline

Training start

Training end

After effect

Baseline

Training start

Training end

After effect
A

Baseline | Training | After effect

Probability

Covariance

Learner
Non-learner

B

Baseline | Training start | Training end | After effect

Covariance

C

Baseline | Training | After effect

Learning
A

Correlation
Baseline Training start Training end After effect

B

Learner Non-learner
Baseline Training start Training end After effect