Current Biology, Vol. 15, 1768–1772, October 11, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.cub.2005.08.063

Remapping Auditory-Motor Representations in Voice Production

Jeffery A. Jones[1,*] **and K.G. Munhall**[2]
[1]Department of Psychology
Wilfrid Laurier University
Waterloo, Ontario N2L 3C5
Canada
[2]Departments of Psychology and Otolaryngology
Queen's University
Kingston, Ontario K7L 3N6
Canada

Summary

Evidence regarding visually guided limb movements suggests that the motor system learns and maintains neural maps between motor commands and sensory feedback [1–3]. Such systems are hypothesized to be used in a feed-forward control strategy that permits precision and stability without the delays of direct feedback control [4]. Human vocalizations involve precise control over vocal and respiratory muscles. However, little is known about the sensorimotor representations underlying speech production. Here, we manipulated the heard fundamental frequency of the voice during speech to demonstrate learning of auditory-motor maps. Mandarin speakers repeatedly produced words with specific pitch patterns (tone categories). On each successive utterance, the frequency of their auditory feedback was increased by 1/100 of a semitone until they heard their feedback one full semitone above their true pitch. Subjects automatically compensated for these changes by lowering their vocal pitch. When feedback was unexpectedly returned to normal, speakers significantly increased the pitch of their productions beyond their initial baseline frequency. This adaptation was found to generalize to the production of another tone category. However, results indicate that a more robust adaptation was produced for the tone that was spoken during feedback alteration. The immediate aftereffects suggest a global remapping of the auditory-motor relationship after an extremely brief training period. However, this learning does not represent a complete transformation of the mapping; rather, it is in part target dependent.

Results and Discussion

Several lines of research suggest that movements, including speech, are planned and then "supervised" by systems that monitor and compare internally and externally generated feedback [5, 6]. Evidence from a number of arm-movement studies that involved visuomotor [1, 2, 7–9] and dynamic perturbations [3, 7, 10–12] indicates that the motor system learns and maintains neural maps of the relationships among the musculature, environment, motor commands, and sensory feedback. The nervous system may use these "internal models" to predict movement outcome and provide internal feedback to the planning and control systems. These internal feedback loops effectively avoid the delays associated with sole reliance on peripheral feedback [4]. One of the fundamental questions concerning internal models is the degree to which learning extends beyond the specific training conditions. When visual or force feedback is perturbed during arm movement, the motor system quickly learns to adjust, and this learning generalizes to other visuospatial [1, 2, 8] and force conditions [9, 11].

In this study, we applied a novel extension of this feedback-perturbation paradigm to speech motor control. Specifically, we manipulated the heard fundamental frequency of the voice during speech to demonstrate the learning of new auditory-motor relationships in Mandarin speakers. Mandarin is a tone language in which the meaning of a word is dependent on the pitch of the utterance. There are four primary tones associated with each monosyllabic morpheme; these tones in Mandarin represent pitch targets achieved by individual speakers [13]. In the experiment, subjects were asked to produce two of the four standard Mandarin tones, tone 1 and tone 2. Figure 1 shows the pitch contours typically observed when a Mandarin speaker produces the word 'ma' inflected with these two tones.

The Queen's University Ethics Committee approved all experimental procedures, and the subjects gave informed consent before participating. The subjects (nine women) participated in two experimental sessions that took place on different days in order to reduce the possibility of vocal fatigue. Sessions took place in a double-walled soundproof booth. Subjects were seated in front of a computer monitor and pronounced the word 'ma' (depicted with traditional Chinese script) presented on the screen. Subjects pressed a mouse button to initiate successive trials. They were instructed to produce the word as similarly as possible on each trial but were not informed that their feedback would be manipulated in any way.

Each of the sessions consisted of three distinct phases (see Figure 2). These phases occurred without notice or interruption for the subjects. In the first phase of one session, subjects produced 10 productions of tone 1 while receiving normal auditory feedback through earphones. The mean pitch value for these initial utterances was taken as the subjects' baseline pitch for the session. This "baseline" phase was followed immediately by the "training" phase, in which subjects produced tone 1 120 times. The pitch of the subjects' auditory feedback in this phase was increased by 1/100 of a semitone for each successive utterance until the feedback received was one semitone above the subjects' true vocal pitch. This increase was followed by 20 trials in which the one-semitone difference was maintained. After the training phase, subjects produced 20 more utterances while receiving normal auditory feedback in a "test" phase. Any differences observed

*Correspondence: jjones@wlu.ca

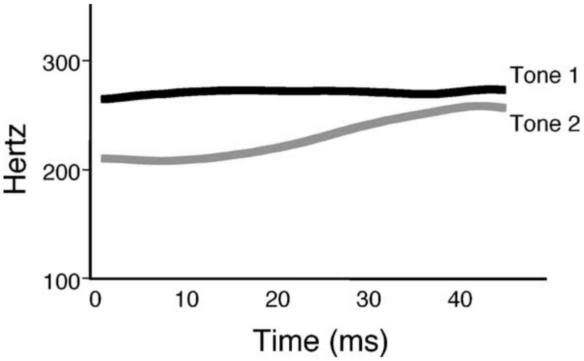

Figure 1. An example of a Typical Fundamental-Frequency Contour for a Speaker's Production of the Word "Ma" Said as Tone 1 and Tone 2

The different tone contours specify the meaning of the morpheme. For example, "ma" means "mother" when produced as tone 1 and "hemp" when produced as tone 2. Note that the tones will vary in absolute pitch because of individual pitch production ranges and coarticulatory constraints.

between the mean pitch values for the "baseline" and "test" phases were indications of an aftereffect caused by exposure to the altered auditory feedback.

To test whether adaptation to one tone generalized to productions of another tone category, we had the same speakers participate in a second session. In this session, the "training" phase also involved the speakers producing tone 1 while receiving auditory feedback that gradually increased in frequency. However, during the "baseline" phase speakers produced the word 'ma' as tone 2. Speakers were again asked to produce tone 2 during the "test" phase when feedback was returned to normal. To test for aftereffects, the mean pitch values for tone 2 utterances produced during this "test" phase were compared to the tone 2 utterances produced in the "baseline" phase.

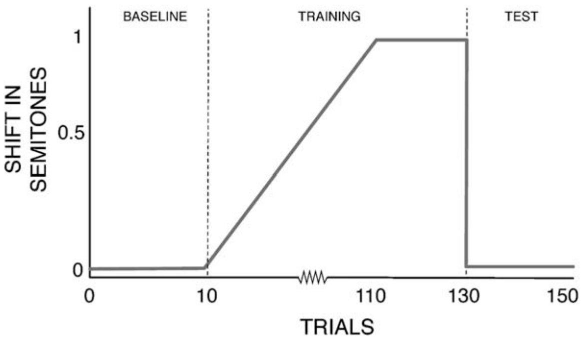

Figure 2. A Schematic Depicting the Altered-Frequency Feedback Paradigm Used in the Experiment

The first ten trials were used to acclimate the subject to the task as well as to measure the speakers' normal baseline fundamental frequency for production of the tones. During the training phase, speakers heard their fundamental frequency shifted up 1/100 of semitone on each trial until, after 100 trials, they heard their voice one full semitone above the fundamental frequency they were actually producing. They heard their voice shifted one semitone for a total of 20 trials before they finally heard an unaltered version of their productions during the test phase.

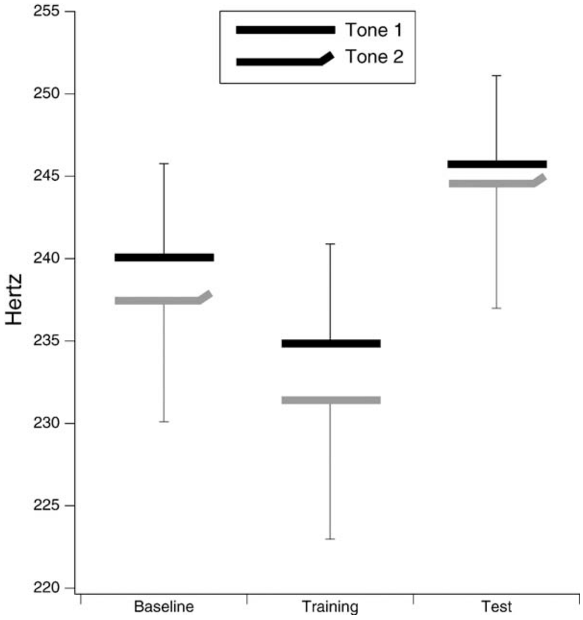

Figure 3. The Average Fundamental Frequency for Tone 1 Productions, and the Average Maximum Frequency of Tone 2 Productions, for the Baseline, Training and Test Phases

Stylized icons indicate the two tone categories, and shading is used to show the two test sessions. The session in which speakers produced tone 1 during the baseline and test phases is in black. The session in which speakers produced tone 2 during the baseline and test phases is in gray. The error bars show the standard errors of the means.

All utterances were recorded with a headset microphone fixed approximately 7 cm from the mouth. Microphone signals were amplified and filtered with a 9 kHz cut-off. The pitch of speech signals was then shifted with a sound-effects generator (Eventide Harmonizer H3000-D/SX). The processing required for the pitch shifting was very close to real time and introduced only a slight delay of around 4 ms. The pitch-shifted speech was then mixed with pink noise and a multi-speaker babble in an effort to decrease the amount of natural acoustic feedback the subjects received regarding their true vocal pitch. The signal and noise were transmitted through earphones into subjects' ear canals. The masking noise was presented at approximately 75 dB SPL, but the amplitude of the speech signal depended on each subject's speaking level. Each session was recorded with a sampling frequency of 48 kHz on a digital audiotape. The data were then low-pass filtered (with a 5 kHz cutoff) and digitized with a sampling rate of 11.025 kHz. The fundamental frequency of the utterance produced in each trial was determined with an autocorrelation algorithm.

Figure 3 shows the average pitch values of the final five productions of tone 1 in the baseline and training phases as well as the average pitch values of the first five productions of tone 1 made in the test phase. Because tone 2 is dynamic, in that it starts low and ends at a higher frequency, we used the average maximum pitch value as our measure (see Figure 3). An ANOVA revealed no significant difference between tone 1 and tone 2 during the three phases of the experiment ($F < 1$). However, as one can see in the figure, there was a

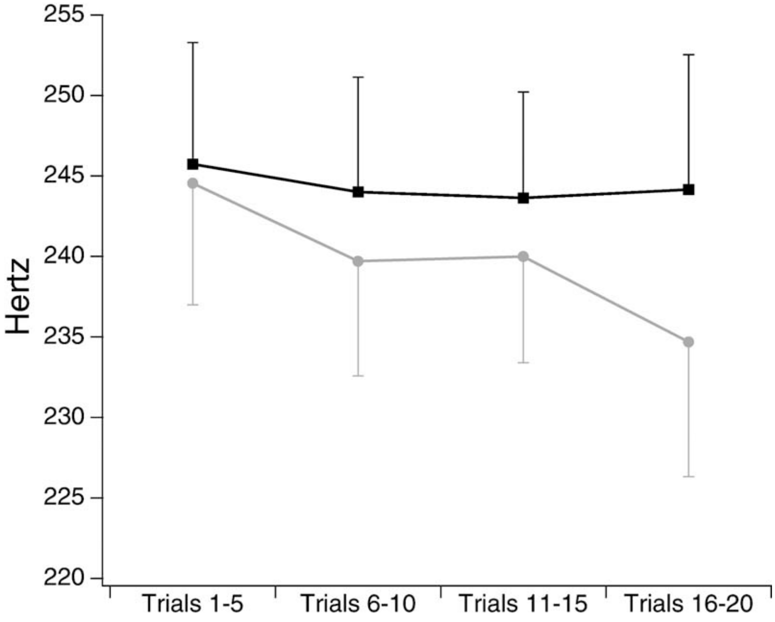

Figure 4. The Average Fundamental Frequency for Tone 1 and the Average Maximum Frequency of Tone 2 in Blocks of Five Trials during the Test Phase

The session in which speakers produced tone 1 is in black. The session in which speakers produced tone 2 is in gray. The error bars show the standard errors of the means.

significant effect of condition [F (2, 16) = 21.3; $p < .001$]. When speakers heard the pitch of their voice shifted up during the training phase, their productions of tone 1 were lower than the tone 1 and tone 2 productions made during the baseline phase (Student-Newman-Keuls test; $p < .01$). This finding is consistent with previous observations that show that shifting feedback regarding voice fundamental frequency leads to compensations in the opposite direction of the shift [14–17]. In addition to these compensations, aftereffects resulted from exposure to the altered auditory feedback. The average frequency of both tone 1 and tone 2 productions was higher after the feedback that speakers received unexpectedly returned to normal (test phase) in comparison to the average observed during the baseline phase (Student-Newman-Keuls test; $p < .01$; see Figure 3). The aftereffect observed for tone 1 replicates our previous work [16, 18]; this study clearly demonstrates that this adaptation generalizes to another, untrained tone. However, the durability of the aftereffects differed across tone 1 and tone 2. When we compared the fundamental frequency for the entire test phase, we found that the effect persisted when speakers were asked to produce tone 1 for the 20 utterances we recorded immediately after training, but the pitch of tone 2 productions decreased over this same time period toward the subjects' normal baseline frequency (see Figure 4). The interaction pattern of responses in the test condition did not quite reach significance [F (3,24) = 2.9, $p > .05$]; however, a trend analysis revealed a strong linear decline in the frequency of tone 2 productions [F (1, 8) = 949; $p < .001$] that did not exist for tone 1 productions [F (1, 8) = 1.9; $p > .05$]. Post-hoc analysis showed that the average of the final five tone 2 productions was significantly lower than the other utterances produced during the test period (Student-Newman-Keuls test; $p < .05$).

Our results suggest that the nervous system uses internal models when planning and controlling the pitch of the voice. Vocal fundamental frequency is a complex motor output determined jointly by the air pressure below the vocal folds, activation level in a network of intrinsic and extrinsic laryngeal muscles, and biomechanical forces that occur in the laryngeal tissue as a result of postural and articulatory adjustments throughout the head and neck area [19]. The active motor control of vocal pitch must accommodate these many influences, and part of this control involves a systematic mapping between produced vocal pitch and the laryngeal motor system. The learning demonstrated in our study indicates that this mapping is calibrated continuously through auditory feedback.

The immediate aftereffects shown for the two Mandarin tones suggest a global remapping of vocal pitch space during the training period. The adaptation observed for the trained tone was extremely robust and, surprisingly, lasted for the duration of the test phase. Future work should address the persistence of these adaptations to acoustic-motor mismatches. Studies investigating the control of arm movements suggest that this type of learning may survive months without intervening exposure [20]. On the other hand, the differential rate of decay of the tone 1 and 2 aftereffects indicates that the learning is not a complete transformation of the mapping but is in part more local and target dependent. A similar blend of local and global adaptation has been observed in other sensorimotor learning paradigms. In studies in which subjects are trained to produce arm movements while in a novel dynamic environment, motor learning generalizes to movements scaled temporally or spatially [7]. However, when the perturbations involve spatial translations or rotations, generalization decays as the distance from the original training position increases [1, 11].

We know surprisingly little about the neural mechanisms involved in sensorimotor control of speech. Certainly, the speech motor-control system in children must somehow adapt to gradual changes in the shape and size of their vocal tract due to growth. Adults too experience changes in their vocal tract (loss of teeth, wearing dental appliances). Speakers must modify their previously learned articulatory-acoustic relationships in order to produce perceptually adequate speech sounds. Several recent neurophysiological studies on fish [21], bats [22], songbirds [23], and monkeys [24–26] support the idea that internal feedback mechanisms underlie the control of vocalization. Brain imaging and stimulation studies of human vocalization also suggest feedforward control of speech. For example, a number of magneto-encephalographic studies show that responses from the auditory cortex are both delayed and damped when subjects hear their own speech productions [27–29]. Significantly smaller magnetic-field recordings from the auditory cortex are observed when subjects speak as opposed to when they hear taped versions of their speech [27]. In addition, multiunit recordings made from patients who were speaking while undergoing craniotomies for the treatment of epilepsy have shown both excitatory and inhibitory events in the lateral temporal cortex [30].

Imaging techniques that take advantage of blood flow (fMRI and PET) have yielded further evidence of internal feedback. Wise et al. [31] used PET and demonstrated a reduction in activation during speech production in a periauditory region. However, other studies have shown increased activation in auditory cortical regions in response to talkers hearing their own voice during speech [32–34]. These increases may be indicative of centers dedicated to providing auditory feedback to the motor system or centers that are functional in processes that inhibit other auditory regions.

The mixed pattern of generalization we observed reflects the multidimensional nature of the sensorimotor learning problem and the structure of the underlying internal models being used to solve this problem [7]. In the case of tone production, the nervous system must balance the continuous demands of motor control with the constraints of discrete linguistic “targets.” Our results imply that separate representations are responsible for the production of individual tone categories and that the apparent global generalization might be accounted for by locally weighted learning of these tone targets [35].

Acknowledgments

This work was funded by Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Institute on Deafness and Other Communication Disorders (NIDCD). We thank Adrienne Steer for her help in analyzing portions of the data.

Received: July 14, 2005
Revised: August 18, 2005
Accepted: August 19, 2005
Published: October 11, 2005

References

1. Ghahramani, Z., Wolpert, D.M., and Jordan, M.I. (1996). Generalization to local remappings of the visuomotor coordinate transformation. J. Neurosci. *16*, 7085–7096.
2. Imamizu, H., Uno, Y., and Kawato, M. (1995). Internal representations of the motor apparatus: Implications from generalization in visuomotor learning. J. Exp. Psychol. Hum. Percept. Perform. *21*, 1174–1198.
3. Thoroughman, K.A., and Shadmehr, R. (1999). Electromyographic correlates of learning an internal model of reaching movements. J. Neurosci. *19*, 8573–8588.
4. Wolpert, D.M., and Miall, R.C. (1996). Forward models for physiological motor control. Neural Netw. *9*, 1265–1279.
5. Desmurget, M., and Grafton, S. (2000). Forward modeling allows feedback control for fast reaching movements. Trends Cogn. Sci. *4*, 423–431.
6. Guenther, F.H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. Psychol. Rev. *102*, 594–621.
7. Goodbody, S.J., and Wolpert, D.M. (1999). The effect of visuomotor displacements on arm movement paths. Exp. Brain Res. *127*, 213–223.
8. Field, D.P., Shipley, T.F., and Cunningham, D.W. (1999). Prism adaptation to dynamic events. Percept. Psychophys. *61*, 161–176.
9. Vetter, P., Goodbody, S.J., and Wolpert, D.M. (1999). Evidence for an eye-centered spherical representation of the visuomotor map. J. Neurophysiol. *81*, 935–939.
10. Shadmehr, R., and Holcomb, H.H. (1997). Neural correlates of motor memory consolidation. Science *277*, 821–825.
11. Shadmehr, R., and Moussavi, Z.M. (2000). Spatial generalization from learning dynamics of reaching movements. J. Neurosci. *20*, 7807–7815.
12. Shadmehr, R., and Mussa-Ivaldi, F.A. (1994). Adaptive representation of dynamics during learning of a motor task. J. Neurosci. *14*, 3208–3224.
13. Yip, M. (1995). Tone in east Asian languages. In The Handbook of Phonological Theory, J.A. Goldsmith, ed. (Cambridge, MA: Blackwell).
14. Kawahara, H. (1995). Hearing voice: Transformed auditory feedback effects on voice pitch control. In Proceedings of the 1995 International Joint Conference on Artificial Intelligence Workshop on Computational Auditory Scene Analysis. (Montreal, Canada: IJCAI), pp. 143–148.
15. Burnett, T.A., Freedland, M.B., Larson, C.R., and Hain, T.C. (1998). Voice F0 responses to manipulations in pitch feedback. J. Acoust. Soc. Am. *103*, 3153–3161.
16. Jones, J.A., and Munhall, K.G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. J. Acoust. Soc. Am. *108*, 1246–1251.
17. Donath, T.M., Natke, U., and Kalveram, K.T. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. J. Acoust. Soc. Am. *111*, 357–366.
18. Jones, J.A., and Munhall, K.G. (2002). The role of auditory feedback during phonation: Studies of Mandarin tone production. J. Phonetics *30*, 303–320.
19. Titze, I.R. (1994). Principles of Voice Production (Englewood Cliffs: Prentice-Hall).
20. Shadmehr, R., and Brashers-Krug, T. (1997). Functional stages in the formation of human long-term motor memory. J. Neurosci. *17*, 409–419.
21. Weeg, M.S., Land, B.R., and Bass, A.H. (2005). Vocal pathways modulate efferent neurons to the inner ear and lateral line. J. Neurosci. *25*, 5967–5974.
22. Smotherman, M., Zhang, S., and Metzner, W. (2003). A neural basis for auditory feedback control of vocal pitch. J. Neurosci. *23*, 1464–1477.
23. Solis, M.M., Brainard, M.S., Hessler, N.A., and Doupe, A.J. (2000). Song selectivity and sensorimotor signals in vocal learning and production. Proc. Natl. Acad. Sci. USA *97*, 11836–11842.
24. Muller-Preuss, P., and Ploog, D. (1981). Inhibition of auditory cortical neurons during phonation. Brain Res. *215*, 61–76.
25. Eliades, S.J., and Wang, X. (2003). Sensory-motor interaction in the primate auditory cortex during self-initiated vocalizations. J. Neurophysiol. *89*, 2194–2207.
26. Eliades, S.J., and Wang, X. (2005). Dynamics of auditory-vocal interaction in monkey auditory cortex. Cereb. Cortex *15*, 1510–1523.

27. Houde, J.F., Nagarajan, S., and Merzenich, M. (2000). Modulation of auditory cortex during speech production: An MEG study. In Proceedings of the Fifth Seminar on Speech Production: Models and Data. (Munich, Germany: Unversitat Munchen), pp. 249–252.
28. Numminen, J., and Curio, G. (1999). Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. Neurosci. Lett. *272*, 29–32.
29. Numminen, J., Salmelin, R., and Hari, R. (1999). Subject's own speech reduces reactivity of the human auditory cortex. Neurosci. Lett. *265*, 119–122.
30. Creutzfeldt, O., Ojemann, G., and Lettich, E. (1989). Neuronal activity in the human lateral temporal lobe. II. Responses to the subjects own voice. Exp. Brain Res. *77*, 476–489.
31. Wise, R.J., Greene, J., Buchel, C., and Scott, S.K. (1999). Brain regions involved in articulation. Lancet *353*, 1057–1061.
32. Price, C.J., Wise, R.J., Warburton, E.A., Moore, C.J., Howard, D., Patterson, K., Frackowiak, R.S., and Friston, K.J. (1996). Hearing and saying. The functional neuro-anatomy of auditory word processing. Brain *119*, 919–931.
33. McGuire, P.K., Silbersweig, D.A., and Frith, C.D. (1996). Functional neuroanatomy of verbal self-monitoring. Brain *119*, 907–917.
34. Paus, T., Perry, D.W., Zatorre, R.J., Worsley, K.J., and Evans, A.C. (1996). Modulation of cerebral blood flow in the human auditory cortex during speech: Role of motor-to-sensory discharges. Eur. J. Neurosci. *8*, 2236–2246.
35. Atkeson, C.G., Moore, A.W., and Schaal, S. (1997). Locally weighted learning for control. Artif. Intell. Rev. *11*, 75–113.