The Specificity of Adaptation to Real-Time Formant Shifting

Ewen MacDonald[1], Elizabeth Pile[1], Hilmi Dajani[2], and Kevin Munhall[1]
[1] Queen's University, [2] University of Ottawa
E-mail: ewen.macdonald@queensu.ca

Abstract

In this study, two experiments were conducted to investigate the specificity of adaptation to real-time formant shifting. During the experiments, talkers were adapted to altered auditory feedback for one vowel (trained vowel) and received unaltered feedback for a different vowel (untrained vowel). In the first experiment, production of the untrained vowel was measured while the talker was in the process of adapting to the altered feedback for the trained vowel. In the second experiment, production of the untrained vowel was measured after talkers had adapted to the altered feedback for the trained vowel. In both experiments, talkers spontaneously modified production of the trained vowel in response to the altered auditory feedback. In the first experiment, talkers slightly altered production of the untrained vowel while the trained vowel was adapting to the altered feedback. In the second experiment, production of the untrained vowel was not altered after talkers had completely adapted to the altered feedback of the trained vowel. These results suggest that the degree of generalization depends on the conditions of adaptation and on the information available about the acoustic environment.

1. Introduction

Previous experiments have shown that talkers will spontaneously compensate for perturbations to the auditory feedback of their voice [1,2,4,5]. In these studies, the formants of a vowel were shifted using a real-time signal processing system. When talkers said a vowel, they heard themselves saying a different vowel. The talkers usually spontaneously compensated for this perturbation by shifting the frequency of their formants in the opposite direction in frequency to the perturbation. This compensation persisted for a short time after feedback was returned to normal, suggesting sensorimotor learning had taken place.

Previous research has investigated the specificity of this sensorimotor learning to test if adaptation to altered feedback of one vowel alters production of other vowels [2,5]. In these studies, the feedback for one vowel, the trained vowel, was perturbed. Over the course of the experiments, utterances of other vowels, the untrained vowels, were collected with feedback consisting of a loud masking noise. In both of these studies, transfer of learning to the untrained vowels was observed. However, these studies tested only a single, restrictive context for speech motor learning. In their paradigm, the talkers received formant-shifted feedback for one vowel yet they simultaneously produced a number of other vowels without feedback due to the masking noise. In this context, the extent of the feedback modification was ambiguous and subjects appeared to act as if a global change to the vowel space had occurred.

In the present study, we test different learning contexts by providing normal feedback while collecting utterances of untrained vowels. (While noise is used to mask bone-conducted feedback, it is presented at a much lower level than the auditory feedback presented over headphones.)

The extent and nature of generalization during learning depends on the availability of information that allows the learner to differentiate the specific and general 'lessons' to be gained. Two experiments were conducted to explore different contexts in which information about the perturbation is manipulated. In the first experiment, we investigated generalization while talkers were in the process of adapting to altered auditory feedback. Over the time-course of learning, talkers produced utterances of both the trained and untrained vowel while receiving perturbed and normal auditory feedback respectively. In the second experiment, we investigated generalization after talkers had already adapted to altered auditory feedback. Talkers received massed practice with the perturbed vowel but no exposure to the untrained vowel until after the time-course of learning. For both experiments, the same trained utterance ("head") and untrained utterance ("hid") were used.

2. Experiment 1: Specificity during learning

2.1. Participants

Twenty-two female talkers participated in this experiment. All spoke English as their first language, reported no history of auditory or speech impairments, and were screened to ensure audiometric thresholds were normal (< 25 dB HL over a range of 500 to 4000 Hz). The protocol for this study was approved by the institutional ethics review board and talkers provided informed consent.

2.2. Equipment

The equipment used was the same as that previously reported in Purcell and Munhall [4]. The talkers were recorded using a headset microphone (Shure WH20), amplified using a Tucker-Davis Technologies MA3 microphone amplifier and low-pass filtered at a cutoff frequency of 4500 Hz (Frequency Devices 901 filter). This signal was digitized at 10 kHz sampling rate. When altered auditory feedback was desired, the signal was filtered in real time to produce formant shifts using a National Instruments PXI-8176 controller. For both normal and altered auditory feedback, noise was added using a Madsen Midimate 622 audiometer and the voice signal and noise were presented to the subject using headphones (Sennheiser HD 265) at 85 and 50 dB SPL respectively.

The manipulation of auditory feedback was achieved by filtering the voice in real-time. Voicing was detected using a statistical amplitude threshold technique. Formants in the speech were determined using an iterative Burg algorithm [3]. The formant estimates were used to calculate the filter coefficients so that a pair of spectral zeroes was positioned at the location of the existing formant frequency and a pair of spectral poles was positioned at the desired frequency of the new formant. The formant frequency estimate and new filter coefficients were computed every 900 μs.

2.3. Estimating model order

Before conducting the experiment, 5 utterances of 7 vowels in an hVd context were collected from each talker ("heed", "hayed", "hid", "head", "had", "hawed", and "who'd"). These were collected to select the AR model order used by the real-time formant shifting system to estimate formant frequency.

The AR model order (which ranged in value from 8 to 12) was selected to achieve the most stable and smooth tracking of formants near the trained vowel ("head"). The heuristic used was based on minimum variance in formant frequency over a 25 ms segment midway through the vowel.

2.4. Procedure

There were two phases in the experiment. In the Baseline phase, 20 utterances of both "head" and "hid" were collected from each talker. During this phase the auditory feedback was unaltered. Talkers heard their own voice from the microphone played back over the headphones at 85 dB SPL. In the Perturbation phase, the talkers alternated saying "head" and then "hid". Forty pairs of utterances were collected. When the talkers said "hid", the auditory feedback was normal (i.e., the same feedback as used in the baseline phase). However, when the talkers said "head", a real-time formant-shifting system was used to alter the auditory feedback heard over the headphones. F1 was increased by 200 Hz and F2 was decreased by 250 Hz. Thus, when talkers said the word "head" they heard themselves saying the word "had".

2.5. Results and discussion

For each utterance, the vowel was segmented by hand. Offline estimates of the formant frequencies were calculated at multiple points by sliding an analysis window (25ms in length) ten speech samples (1ms) per estimate using a similar algorithm to that used in online shifting. For each trial, a single "steady-state" F1 value was determined by averaging 40% of the F1 estimates starting, from 40% of the way into the vowel to 80% of the way through the vowel.

The F1 estimates were then normalized for each vowel for each individual by subtracting their baseline. The baseline for each vowel was defined as the average of the last 15 utterances from the Baseline phase (i.e. utterances 6-20). The normalized results were then averaged across individuals and can be seen in Figure 1. The results for "head" and "hid" are in blue circles and red triangles respectively. The vertical dashed line separates the results of the Baseline phase (left of the line) from those of the Perturbation phase (right of the line).

From the results it is clear that the production of both the trained ("head") and, to a smaller degree, the untrained vowel ("hid") were affected

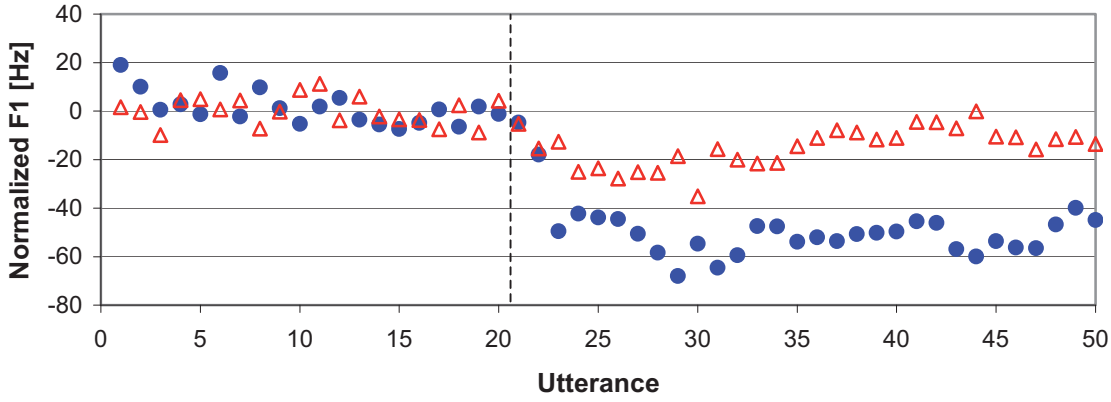

Figure 1. Average normalized F1 for utterances of the trained ("head", blue circles) and untrained ("hid", red triangles) vowels.

by the altered feedback provided during production of the trained vowel. However, as adaptation of the trained vowel reached steady-state (around utterance 35; 15 utterances into the perturbation phase), production of the untrained vowel appears to begin to return to baseline.

3. Experiment 2: Specificity after learning

3.1. Participants

Fifteen female talkers participated in this experiment. All spoke English as their first language, reported no history of auditory or speech impairments, and were screened to ensure audiometric thresholds were normal. The protocol for this study was approved by the institutional ethics review board and talkers provided informed consent.

3.2. Equipment

The equipment used was the same as in Experiment 1.

3.3. Estimating talker-specific parameters

Before conducting the experiment 5 utterances of 7 vowels in an hVd context were collected from each talker. These utterances were used to estimate the AR model order and the shift size.

The AR model order was selected using the same method as in Experiment 1.

Unlike Experiment 1, a different shift size was used for each talker. The individualized shift was determined from the difference in F1 and F2 between a talker's average production of the vowels in "head" and "had". On average, the shift size was 188 and -254 Hz for F1 and F2 respectively. This is similar to the 200 and –250 Hz shifts used in Experiment 1.

3.4. Procedure

There were three phases in the experiment. In the Baseline phase, 15 utterances of both "head" and "hid" were collected from each talker. During this phase the auditory feedback was unaltered. Talkers heard their own voice from the microphone played back over the headphones at 85 dB SPL. In the Perturbation phase, the talkers said "head" 40 times with formant shifted feedback. Thus, when talkers said the word "head" they heard themselves say "had". In the last phase, After Adaptation, talkers produced 40 utterances of the word "hid" with normal auditory feedback.

3.5. Results and discussion

As in Experiment 1, the vowels were segmented by hand and a "steady-state" F1 was estimated. These estimates were normalized for each individual.

The average normalized results for "head" and "hid" can be found in Figure 2. In this figure, the vertical dashed lines indicate the boundaries of the phases of the experiment. From the figure, it is clear that production of the trained vowel ("head") changed as talkers adapted to the altered feedback. Further, the adaptation was similar to that found in Experiment 1. However, the production of the untrained vowel ("hid") did not change between the Baseline and After Adaptation phases. Thus, the adaptation of the trained vowel was specific to

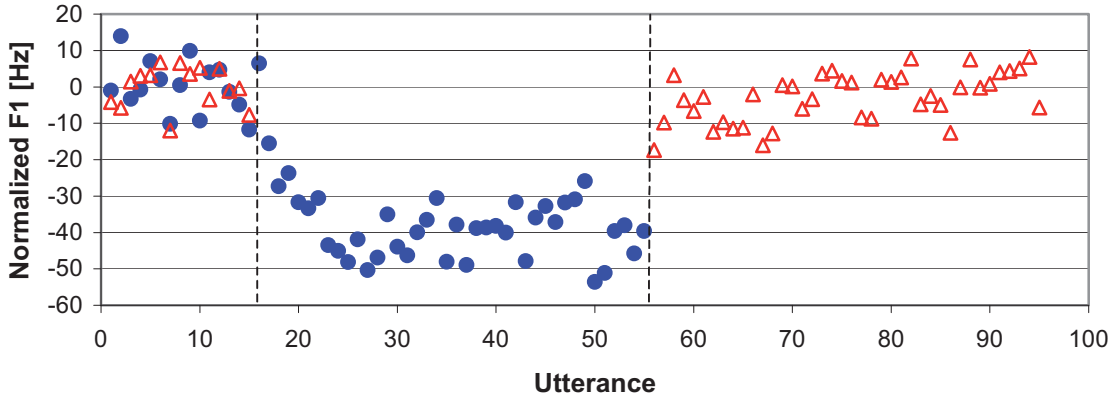

Figure 2. Average normalized F1 for utterances of the trained ("head", blue circles) and untrained ("hid", red triangles) vowels.

that vowel and did not generalize to the untrained vowel.

4. General discussion

The results of Experiment 1 showed that production of the untrained vowel was affected as the talker was adapting to the altered feedback during production of the trained vowel. As talkers' adaptation of the trained vowel reached steady state, their production of the untrained vowel began to return towards baseline. In Experiment 2, no change in production of the untrained vowel was observed. Thus, after a talker has adapted to the altered feedback of the trained vowel (i.e., production of the trained vowel has reached steady-state), the adaptation remains specific only to the trained vowel. Together, these results suggest that the specificity of adaptation varies over the time-course of adaptation.

A possible reason for the discrepancy between the present data and previous studies of generalization [2,5] may lie in the differences in auditory feedback used when collecting utterances of the untrained vowels. In the present study, talkers were always presented with unaltered feedback when saying untrained vowels. In the other studies, a loud masking noise feedback was presented so that talkers could not make use of auditory feedback to control production of the untrained vowel. Note, however, that using a loud masker assumes that the control system will not change its output in the absence of auditory feedback cues and this assumption warrants further examination.

References

[1] J. F. Houde, M. I. Jordan. Sensorimotor adaptation in speech Production. Science, 279(5354):1213-16, February 1998.

[2] J. F. Houde, J.F., *Sensorimotor Adaptation in Speech Production*. Doctoral thesis, Department of Brain and Cognitive Sciences, MIT, Cambridge, MA, 1997

[3] S. J. Orfandidis, *Optimum Signal Processing, An Introduction.* MacMillan, New York, 1988.

[4] J. F. Purcell, K. G. Munhall. Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, 120(2):966-77, August 2006.

[5] V. M. Villacorta, J. S. Perkell, F. H. Guenther. Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *Journal of the Acoustical Society of America*, 122(4):2306-19, October 2007.