Actions are organized around goals or intentions. In speech production, there has been no agreement on how best to discuss speech goals. However, the auditory feedback perturbation methodology provides a window into the nature of speech goals. To the extent that subjects are sensitive to variation in an acoustic attribute, this attribute must be part of the controlled intention of articulation. In this presentation, we will review a series of studies that speak to this issue. In one study, we examined how intentionality of speech production influences compensatory formant production by instructing subjects to use a cognitive strategy in order to make the feedback sound consistent with the intended vowel. In other studies, we have explored the specificity of vowel formant compensation by comparing cross-language differences. The results indicate that speech goals are 1) very specific, defined by a phonemic category and its relationship with neighboring categories, and 2) multivariate. We will discuss these results by contrasting compensatory behaviors in reaching and limb movements to those observed in speech studies. The presence of a system of categories in speech may result in differences in the way speech goals are represented.
INTRODUCTION

The speech production process begins with speaker’s intention to communicate. In most accounts the process of articulations includes a phase where a phonological representation in the mind is transformed into physical form (i.e., articulatory gestures and consequent sounds). The transformation of such categorical mental sound representations to movements is fundamentally different from other motor behaviors, such as reaching, whose goals are defined in the environment (e.g., specified in a visual plane) that are usually not categorical. These differences in the nature of motoric targets might be reflected in how people control behaviors.

One way to examine how motoric goals are defined and achieved is to see how erroneous behaviors are corrected, using a real-time perturbation paradigm. In both visuomotor and auditory speech perturbations, subjects generally compensate by moving opposite to the direction of the perturbations. Recently, Mitsuya et al. (2011) have shown that speech vowel compensations may be unique in that they are produced with respect to the vowel category and its local neighbors.

In the present study, we replicate an experiment carried out with visuomotor adaptation. Mazzoni and Krakauer (2006) reported that when subjects are given a strategic target to cancel the perturbation all at once, they were able to aim at the given target initially, but they slowly began to overshoot the correction. Taylor and Ivry (2012) observed that this overshoot did not persist; however, they too reported the same behavior shortly after the subjects started aiming at the strategy target. Here, we test whether the use of an explicit cognitive strategy to overcome the perturbation of formant frequencies will show a similar pattern to that observed for reaching.

Mazzoni and Krakauer (2006) suggested that the observed overshoot in a reaching experiment might be due to the motor system trying to resolve the difference between the predicted versus observed trajectories of movements. Given that speech goals seem to be represented differently as a system of targets, then explicit strategies using different vowel categories to overcome the perturbation may result in different patterns of compensation.

METHODS

Participants

Nineteen female students of Queen’s University participated in the current experiment. The use of one gender was to reduce the differences in formant structure across participants. The average age was 19.6 (ranging from 18-21 years), and all of them learned English as their first language. Each participant was tested in a single session. No participants reported speech or language impairments and all had normal audiometric hearing thresholds over a range of 500-4000 Hz.

Equipment

Equipment used in this experiment was the same as the reported in Munhall et al. (2009), MacDonald et al. (2010, 2011) and Mitsuya et al. (2011). Speakers were tested in a sound attenuated booth in front of a computer monitor with a headset microphone (Shure WH20) and headphones (Sennheiser HD 265). The microphone signal was amplified (Tucker-David Technologies MA 3 microphone amplifier), low-pass filtered with a cutoff frequency of 4.5Hz (Hrohn-Hite 3384 filter), digitized at 10 kHz and filtered in real-time to produce formant shifts (National Instruments PXI-8106 embedded controller). The manipulated speech signal was then amplified and mixed with speech noise (Madsen Midimate 622 audiometer). This signal was presented through the headphones that the speakers wore. The speech and noise were
presented at approximately 80 and 50 dBA SPL respectively.

**Acoustic processing**

Voicing detection was done using a statistical amplified-threshold technique, and the real-time formant shifting was done using an IIR filter. An iterative Burg algorithm (Orfanidis, 1988) estimated formant frequencies every 900 μs. Prior to the experimental data collection, a parameter, the model order to determine the number of coefficients used in the auto-regressive analysis was estimated by collecting seven English vowels /i, i, e, e, æ, o, o, u/ were presented in an /hVd/ context (“heed”, “hid”, “hayed”, “head”, “had”, “hawed”, “who’d”). These words were randomly presented on a computer screen in front of the speakers, and they were instructed to say the prompted word without gliding the tone, or pitch. These utterances were analyzed with model orders ranging from 8 to 12. For each speaker, the best model order was selected based on minimum variance in formant frequency over a 25 ms segment in the middle portion of the vowel (MacDonald *et al.*, 2010).

For offline formant analysis, an automated process estimated the vowel boundaries in each utterance, based on the harmonicity of the power spectrum. These estimates were then manually inspected and corrected if required.

**Procedure**

Speakers produced 100 utterances of the word “head (/hEd/)” with a visual prompt on the screen in front of them. The prompt lasted 2.5 s with the inter-trial interval of approximately 1.5 s. The 100 utterance-session consisted of three experimental phases. In the first phase, Baseline (utterances 1-20), speakers received normal feedback through the headphones (i.e., amplified and noise added but with no change in formant frequency). In the second phase, Perturbation (utterance 21-60), speakers received altered feedback in which F1 was increased by 200 Hz and F2 was decreased by 250 Hz. This perturbation made the feedback sound more like “had (/hæd/).” Immediately after the 23rd trial, the experiment was paused and the experimenter instructed the speaker to say “hid (/hɪd/)” to make the sound they heard through the headphones more consistent with the sound of the word “head.” Then the experiment was resumed. In the third phase, Return (utterances 61-100), the perturbation was removed abruptly and the feedback went back to normal.

**FIGURE 1:** Feedback shift applied to the first formant (dotted line) and second formant (solid line). The vertical dashed lines denote the boundaries of the three phases: Baseline, Perturbation, and Return.
RESULTS

The baseline average of F1 was calculated for each speaker, based on the last 15 utterances of the Baseline phase (i.e., utterances 6-20), then the raw F1 value in Hz was normalized by subtracting the speaker’s baseline average from F1 value of each utterance. Figure 1 shows the overall average of normalized Formants. As can be seen, immediately after the perturbation was introduced, speakers already started to adjust their formant production (utterance 22 and 23). When the instruction was given after the 23rd utterance, 17 speakers correctly followed the instruction and produced “hid” at utterance 24 when the experiment was resumed. The remaining 3 speakers started saying “hid” at utterance 25.

![Figure 2: Averaged normalized F1 (solid circles) and F2 (open cercles). The vertical dashed lines denote the boundaries of the three phases: Baseline, Perturbation, and Return.](image)

The question we were examining was whether speakers would change the production of strategy vowel. In order to verify this, the average magnitude of compensation was compared across three points in the experiment, 1) the Perturbation phase after the cognitive strategy was given (utterances 25-40), 2) the last part of the Perturbation phase (utterances 46-60), and 3) the last part of the Return phase (utterances 86-100). An analysis of variance (ANOVA) was conducted with the three time points as within-subject factors, and it was significant on both F1 ($F_{[2, 36]} = 12.34, p < 0.05$) and F2 ($F_{[2, 36]} = 15.17, p < 0.05$). This significance is solely due to the fact that speakers’ production differed between the Perturbation and Return phases because post hoc analyses revealed that there was no difference between the two points within the Perturbation phase (F1: $t_{[18]} = 1.15, p > 0.05$; F2: $t_{[18]} = 1.02, p > 0.05$).

The results of Mazzoni and Krakauer (2006) and Taylor and Ivry (2012) imply that the adaptation to the visual rotation was implicitly global. The introduction of a cognitive strategy to resolve the discrepancy still resulted in a perturbation and compensation situation. In our experiment, it is possible that the cognitive strategy of producing "hid" during the Perturbation phase might have been affected by the introduction of perturbation. In order to examine whether the vowel /i/ was produced differently from speaker’s resting state, we compared the formant values of /i/ produced in the Perturbation phase and those collected during the prescreening session. The analysis revealed that speakers’ production did not differ with both F1 ($t_{[18]} = 1.12, p > 0.05$) and F2 ($t_{[18]} = 1.53, p > 0.05$), indicating that the perturbation did not induce an implicit learning on the vowel /i/.

It is important to note that the reason why the group average formant values did not go back to the resting point in the Return phase was that some speakers continued to say “hid” until the end of the experiment, failing to make the feedback consistent with the sound “head”.

---

T. Mitsuya and K. Munhall

We separated these people from the speakers who switched back to saying “head” in the Return phase. This yielded 11 speakers who kept saying “hid” (Stay group), and 8 speakers who switched back (Switch group).

**Figure 3:** Normalized F1 (solid) and F2 (open) production averaged across speakers in Switch group (circles) and Stay group (diamond)

Clearly, some speakers were cognizant of the task making the auditory feedback consistent with a particular vowel by producing another vowel, while others ignored the auditory feedback altogether and just produced the cognitive target as a new target regardless of its relationship with the feedback target at all. All of these results indicate no overshoot or implicit global adaptation for perturbation. However, the difference in the two types of observed behavior might have interacted with the overshoot effect somehow, thus we separated the groups and compared the group average of the magnitude of compensation during the perturbation phase; however, the groups did not differ (F1: t(17)= .69, p > 0.05; F2: t(17)= 1.88, p > 0.05). Moreover, the Stay group’s formant values during Perturbation and Return phases did not differ F1: t(10)= -.87, p > 0.05; F2: t(10)= .87, p > 0.05), indicating there was no change 1) in the way the speakers were adapting to the perturbation with the strategy vowel regardless of whether they were attending to the feedback or just focusing on producing the strategy vowel and 2) in the production of the strategy vowel regardless of the introduction and removal of the perturbation.

**DISCUSSION**

The current study was set up to examine the difference between the compensatory behavior of auditory perturbations of formant production and the visuomotor adaptation when people were given a specific strategy to correct their behavior for the perturbation given (Mazzoni and Krakauer, 2006; Taylor and Ivry, 2012). Unlike the results reported in these visuomotor adaptation studies, we did not observe an overshoot due to the cognitive strategy target as if subjects were implicitly adapting to the perturbation regardless of the target. Instead, they were able to reliably produce the vowel without the influence of perturbation. These results imply that perturbation applied to a vowel does affect the production of another vowel around the auditory target, suggesting adaptation is not global. Mazzoni and Krakauer (2006) postulated that the motor system’s intolerance for two simultaneous targets is predicated on the assumption that the rotation of visual space is applied globally. But this does not seem to be the case with speech production goals and how they are represented in the F1/F2 acoustic space.

Resistance to perturbation while producing a vowel that is given as a cognitive target
indicates that 1) representation of vowels is more than acoustic attributes, at least the properties that were perturbed in the current study and 2) speakers’ intention to produce a phonological category plays an important role in the stable production of the category, rather than the controlling acoustic attributes independently.

ACKNOWLEDGMENTS

This research was supported by Natural Sciences and Engineering Research Council of Canada.

REFERENCES


