Article

Language *and* Speech

Partial Compensation for Altered Auditory Feedback: A Tradeoff with Somatosensory Feedback?

Language *and* Speech
55(2) 295–308
© The Author(s) 2011
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/0023830911417802
las.sagepub.com
SAGE

Shira Katseff[1], John Houde[2] and Keith Johnson[1]
[1]University of California at Berkeley, USA
[2]University of California, San Francisco, USA

Abstract
Talkers are known to compensate only partially for experimentally-induced changes to their auditory feedback. In a typical experiment, talkers might hear their F1 feedback shifted higher (so that /ɛ/ sounds like /æ/, for example), and compensate by lowering F1 in their subsequent speech by about a quarter of that distance. Here, we sought to characterize and understand partial compensation by examining how talkers respond to each step on a staircase of increasing shifts in auditory feedback.

Subjects wore an apparatus which altered their real time auditory feedback. They were asked to repeat visually-presented *hVd* stimulus words while feedback was altered stepwise over the course of 360 trials. We used a novel analysis method to calculate each subject's compensation at each compensation step relative to their baseline. Results demonstrated that subjects compensated more for small feedback shifts than for larger shifts. We suggest that this pattern is consistent with vowel targets that incorporate auditory and somatosensory information, and a speech motor control system that is driven by differential weighting of auditory and somatosensory feedback.

Keywords
adaptation, altered auditory feedback, compensation, vowel production

Introduction

Given the many opportunities for errors in the formulation and execution of an utterance, running speech contains remarkably few mistakes. The speech motor control system accomplishes this in part by using incoming feedback to tune the details of planning and execution. A precise understanding of how feedback is used to control articulation is an important part of understanding how speech is planned and represented.

This paper focuses on the contributions of somatosensory and auditory feedback to the control of articulation. Somatosensory feedback reaches the central nervous system through mechanoreceptors on the surfaces of the vocal tract articulators and from stretch receptors in its muscles. These

Corresponding author:
Shira Katseff, 1203 Dwinelle Hall, University of California at Berkeley, Berkeley, CA 94720-2650, USA
Email: shira.katseff@canterbury.ac.nz

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

receptors provide information about muscle lengths, forces, and the spatial positions of articulators (Wyke, 1983; Perrier, Lœvenbruck, & Payan, 1996; Guenther & Barreca, 1997; Sanguineti, Laboissière, & Ostry, 1998; Shiba et al., 1999). Auditory feedback reaches the brain via the ear and auditory nerve. These two types of feedback are integrated at an early stage of processing (Schroeder et al., 2001). Models of speech motor control, for example DIVA (Guenther, 1995, 2003; Perkell et al., 2000), explicitly include components by which feedback is used to adjust articulators during speech. Such models are based on results of experiments that measure articulatory changes in response to modified somatosensory and auditory feedback.

1.1 Responses to abnormal feedback

Experimental evidence suggests that talkers are sensitive to perturbations in both somatosensory and auditory feedback. When somatosensory feedback is altered with no acoustic consequences, talkers can show sensitivity to the location of their articulators. In a representative experiment, Tremblay, Shiller, and Ostry (2003) pulled the jaw forward as talkers produced speech and non-speech. They found that, while compensation is eventually complete for speech sounds, non-speech sounds do not show complete adjustment.

Talkers compensate for altered auditory feedback in response to both sudden *perturbations* and longer-term *adaptation* designs. Our focus here is on adaptation experiments, which use real-word stimuli and demonstrate long-term learning effects. In these experiments, talkers wear a headset whose microphone is connected to its earphones via a computer. As talkers produce target words or sounds, they hear their own voices played back to them in real time. During the experiment, the computer alters the talker's incoming voice before sending it back out to the headphones. Talkers tend to compensate by opposing these feedback shifts (Houde & Jordan, 2002; Purcell & Munhall, 2006). For example, a talker with a baseline /æ/ F1 of 800 Hz who hears her /æ/ F1 shifted up to 900 Hz typically responds by producing /æ/ with a lower F1. Most such experiments have been performed in English, though this general result has also been replicated in Mandarin (Jones & Munhall, 2005; Cai et al., 2010). Although compensation can also be induced by sudden *perturbations* to pitch or formant feedback (Burnett et al., 1998), compensation for short-term changes may proceed by a different mechanism than the one used for long-term adaptation.

In general, speech adaptation experiments proceed by slowly altering incoming auditory feedback from no change up to some maximum amount of change. These experiments describe compensation for single shifts across category boundaries in F1 and F2 (typically 200–300 Hz), and over a range of changes, from 25 cents up to 1 semitone, in F0. Notably, subjects never compensate for formant shifts, on average, more than 40%. That is, an average subject whose F1 feedback is raised by 200 Hz will produce vowels with an F1 no more than 80 Hz lower than usual. Compensation during speech adaptation is incomplete and variable (Table 1).

The fact that, on average, subjects consistently compensate for altered auditory feedback is compelling evidence that feedback is monitored. But if talkers are both sensitive to their auditory feedback and able to change their vowel formants to oppose the feedback shift, why should they stop at 16% or 20% compensation? Several plausible explanations can be dismissed on the basis of existing work.

One such explanation is that the altered signal is mixed with accurate feedback from bone conduction, attenuating subjects' responses to auditory feedback shifts. This explanation implies that speech with minimal bone conduction ought to exhibit more complete compensation than voiced speech. A comparison of feedback shift studies testing sibilants or whispered speech (Shiller et al., 2009; Houde & Jordan, 2002) to studies testing voiced speech shows that, contrary to this expectation, compensation is similarly incomplete with or without bone conduction.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

Table 1. Summary of recent speech perturbation and adaptation experiments.

Research group	Formant	Amount of shift	Language	Compensation
Larson et al., 2008	F0	1 semitone	English	20%
Jones & Munhall, 2005	F0	1 semitone	Mandarin	36%
Jones & Munhall, 2000	F0	1 semitone	English	32%
Purcell & Munhall, 2006	F1	to /ɛ/ or /æ/	English	16% / 11%
Pile et al., 2007	F1 & F2	to /æ/	English	20%
Houde & Jordan, 2002	F1 & F2	to /i/ or /æ/ (whispered)	English	28%

Another possibility is that subjects produce different vowel formants not because they are perceiving a discrepancy between observed and expected feedback, but because performing the task changes their perceptual boundary between vowels. This sort of perceptual adaptation is common in psycholinguistic experiments (Diehl, 1981). It is possible that hearing shifted /ɛ/ changes the criteria used to distinguish it from /ɪ/ and /æ/, damping compensatory changes in production. Some support for this possibility comes from fricative adaptation experiments, in which talkers experience a boundary shift of approximately 10% on the /s/-/ʃ/ continuum as a result of hearing their own shifted sibilants (Shiller et al., 2009). Using this study as a benchmark, we might predict that an adaptation experiment shifts perceptual boundaries by about 10%. This possibility will be considered in the discussion of our own results.

Two remaining explanations for partial compensation are the focus of the experiment described here. The first is that vowel targets are large regions in acoustic space, and thus, a large set of formant values are equally good representatives of a vowel. Major models of speech motor control, e.g., the DIVA model (Perkell et al., 2000), subscribe to this view. Large target regions imply two predictions that are tested here. First, if all vowels within a region are functionally equivalent, there should be no compensation for shifts that do not push the token outside of the vowel target region. Second, larger formant shifts ought to exhibit more complete compensation than smaller formant shifts. If a vowel at the center of the baseline region is shifted to a position just outside of it, the shift should elicit a production change just large enough to bring it back within the target region. If, on the other hand, the vowel is moved to a position far from the border of the baseline region, the production change required to hear a vowel within the baseline region is almost as large as the feedback shift.

We test these predictions by measuring the completeness of compensation relative to the vowel's baseline region in acoustic vowel space. Standard measures of the speaker's starting location, such as the average formants produced when auditory feedback is unaltered, are inadequate for this purpose because vowels produced in succession are autocorrelated. Over 15 unaltered trials, a typical number for this sort of experiment, the subject's vowels will cluster in a small area of the subject's baseline vowel region. We propose a more flexible estimation method.

Measures of compensation are also dependent on the units of measurement. While F1 and F2 changes to auditory feedback are typically performed in Hertz (F0 is typically altered in cents), perceptual distances between vowels are a better fit to an auditory scale. If discrepancies between observed and expected auditory feedback are mediated by the perceptual system, it is important to measure compensation in Hertz as well as other auditory perceptual measures.

The second explanation for partial compensation, also tested in this experiment, is based on the fact that vowel targets are defined in terms of both acoustic/auditory and somatosensory dimensions. The assumption here is that the speaker's response optimizes deviation from the auditory and somatosensory target values so that the overall production is as close as possible in a

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

multidimensional space (defined by both auditory and somatosensory dimensions) to the intended vowel. Altered auditory feedback causes the vowel to deviate from the intended target in only the acoustic/auditory dimensions. The speaker's response, however, is constrained by somatosensory feedback to be only partial in the acoustic domain. There is already strong evidence for this hypothesis. In Larson et al. (2008), talkers produced the vowel /u/ under F0 perturbations before and after their vocal folds were numbed by local anesthesia. There was a significant difference in degree of compensation before and after the anesthesia was administered; beforehand, subjects produced an average of 20 cents compensation for a 100 cent change in auditory feedback, whereas after the anesthesia was administered, subjects compensated for a 100 cent feedback perturbation by about 25 cents. Larson et al. cite two possible interpretations for these results. First, somatosensory feedback might oppose auditory feedback. Second, the lack of somatosensory feedback might trigger the monitoring system to attend more to auditory feedback.

If these two sources of feedback always have equal weight, then the completeness of compensation ought to be independent of the size of the formant shift. But if the speech motor control system makes online changes to the weighting of the two types of feedback so as to downweight unreliable feedback sources, then compensation should decrease as the amount of formant shift increases. This experiment sought to distinguish between these two accounts of partial compensation for altered auditory feedback by comparing compensation for shifts of various sizes.

It is important to note that these two explanations for partial compensation, large target regions and conflicting sources of feedback, are not mutually exclusive. In addition to the predictions outlined above, it is possible for both explanations to hold. In this case, small shifts might elicit no compensation because shifted auditory feedback would fall within the large target region, and large shifts might elicit minimal compensation because shifted auditory feedback would fall outside of the large target region, but would also be downweighted for deviating considerably from typical auditory feedback.

2 Methods

Participants were seated in a soundproof booth and wore an AKG HSC-271 Professional headset. Their speech was routed from the headset microphone through a Delta 44 sound card and into a computer, where it was analyzed and re-synthesized in real time. Re-synthesized speech was played through the headset's earphones in place of normal auditory feedback. Both analysis and re-synthesis were performed with a real time feedback alteration device (FAD) designed by the second author. Post-session interviews indicated that subjects did not notice either formant shifts or delays in re-synthesized feedback.

2.1 Apparatus

Analysis and re-synthesis were performed by a feedback alteration device (FAD). Our current FAD is based on the method of sinusoidal synthesis developed by Quatieri and colleagues (McAulay & Quatieri, 1986, 1991; Quatieri & McAulay, 1986, 1992; Quatieri, 2002). In this FAD, an analysis-synthesis process repeatedly digitizes, from the microphone, 3 ms frames of the subject's speech (32 time samples at an 11.025 kHz sampling rate = a 3 ms frame rate). These frames are analyzed, modified, and re-synthesized into new frames making up the modified speech output. Each input frame is shifted into a 400 sample (36 ms) buffer, and this buffer is analyzed by computing a narrow-band magnitude frequency spectrum. For voiced speech, a narrow-band spectrum has a comb-like appearance, with narrow, regularly spaced peaks whose heights slowly vary in amplitude over the length of the spectrum. These peaks are harmonics, located at integer multiples of the fundamental frequency, and thus the spacing between successive harmonics is equal to the fundamental

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

frequency. From the envelope of the peak heights, the so-called spectral envelope, we are able to estimate formants and total frame energy (i.e., the current value of the temporal envelope of the speech). Thus, from the narrow-band magnitude spectrum alone, we are able to separate the pitch, formant, and temporal envelope characteristics of the speech. These features can then be independently modified before being recombined to make a new narrow-band magnitude spectrum. This new spectrum is used to synthesize the next frame of output speech using sinusoidal synthesis. This synthesis method does not require the phase spectrum of the original input speech; instead, each harmonic peak in the new narrow-band magnitude spectrum specifies the frequency and amplitude of a sinusoid, and these sinusoids are then simply added together to create the next frame of output speech. The output speech is converted back to an analog signal which is fed to the subject's earphones. The phase of each sinusoid used to create the current frame of output speech is tracked in order to avoid discontinuities, which are possible when sinusoids are continued on to the next output frame. The double buffering scheme used to implement this process in real time, with the previous frame being outputted while the current frame is processed, would ideally only incur a 6 ms, or two frame, feedback delay. However, because of additional pipeline delays in the sound card, the entire process, from input to output, incurs a feedback delay of 12 ms.

2.2 Procedures

Before the experiment began, subjects read a short passage in order to become acclimatized to hearing themselves through headphones. No recording of speech or speech alteration occurred during this period. In the first baseline stage, 7 *hVd* words appeared on a computer screen for 20 seconds: /hid/, /hɪd/, /hɛd/, /hæd/, /had/, /hoʷd/, /hud/. Subjects were recorded while reading the list of words until they disappeared from the screen. Feedback was not altered during this stage. During the alteration stage, a MATLAB program displayed the prompt 'Say HEAD now' on a computer screen for 1000 ms. Subjects were instructed to say 'head' when the prompt appeared. During each trial, the subject's formants were shifted in real time using the FAD. The size of the formant shift was specific to the trial, as described below. Both the word that the subjects produced and the shifted word that subjects heard were recorded for 750 ms. There were a total of 360 trials, split into 24 15-trial blocks. Subjects were permitted to take a break after each block. The entire procedure took approximately 15 minutes.

To disguise the purpose of the task, talkers were told that their reaction time would be recorded as they followed the instructions given by the prompt. They were not informed that their speech would be altered, though a full explanation of the study was given at the end of the experiment.

2.2.1 Replication. To verify that our setup yielded similar results to previous experiments, we first replicated the procedures of Purcell and Munhall (2006). Subjects in the replication experiment had their F1 raised from baseline to a maximum of 200 Hz. Two college-aged male subjects participated in this initial experiment. The time course of this effect for one of the initial subjects is illustrated in Figure 1. The other subject compensated similarly.

As talkers heard their F1 increase from trial to trial, they began to produce /ɛ/ with a lower F1 (such that their vowels sounded more like /ɪ/). That is, they *compensated* for the change in auditory feedback, closely mirroring the formant patterns observed in previous formant shift experiments.

For the main experiment, formant alteration proceeded in three phases over a total of 360 trials, as shown in Figure 2.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

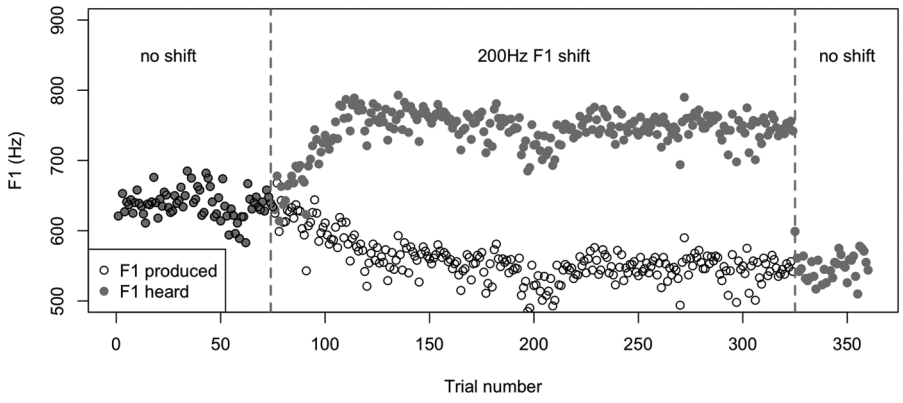

Figure 1. F1 produced by a typical subject (black open circles) and F1 heard by this subject (gray filled circles) over the course of the experiment. Each gray circle/black circle pair represents one trial.

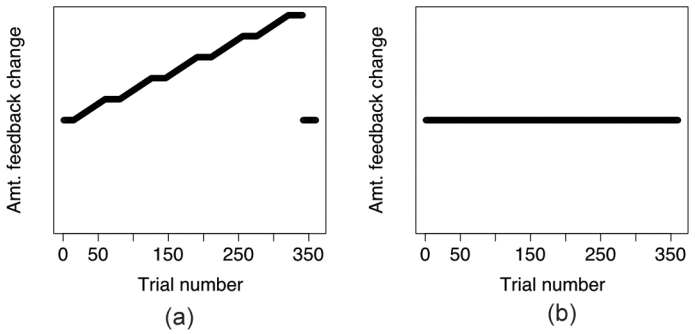

Figure 2. Change in F1 feedback over the course of each experiment. During a formant shift session (a), the 360 trials were composed of 5 regions of equal formant shift steps connected by ramps of slowly increasing or decreasing feedback shift. During the control session (b), there was no change in feedback over the 360 trials.

Baseline: No formant shift (15 trials).
Shift: Formant feedback was slowly shifted up to 5 different feedback shift steps: 50Hz, 100Hz, 150Hz, 200Hz, and 250Hz (20 trials on each step).
Adaptation: Formant feedback returned to normal (25 trials).

This procedure was repeated for F0, F1, and F2 on three different days. On a fourth day, subjects participated in a control condition in which their feedback was not altered. The order of these four sessions was randomized.

2.3 Participants

Seven subjects participated in this experiment. All were native speakers of English with normal hearing and ranged in age from 18 to 49. Because the quality of formant re-synthesis was better for modal voices with low pitch, all subjects were males. The experimental protocol was approved by the UC Berkeley Committee for the Protection of Human Subjects.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

2.4 Analysis

Only responses to upward shifts in F1 feedback (from /ɛ/ toward /æ/) are considered here. Responses to shifts in F0 and F2 feedback were less consistent, with many subjects either following the feedback shift or failing to compensate, and require a different sort of analysis. The vowel formants of each token were calculated from the average formant values during the 50 ms surrounding the amplitude peak of the vowel, as measured by Entropic's/ESPS Xwaves software installed in the Phonetics Laboratory at the University of California at Berkeley. To verify the accuracy of this method, 5 tokens per speaker were randomly measured by hand using PRAAT (Boersma & Weenink, 2009).

3 Results

3.1 F1 shift data

All subjects in this experiment compensated for the shift in F1 feedback. A typical subject's F1 in /ɛ/ over the course of the experiment is illustrated in Figure 3.

The F1 in this talker's 'head' vowel clearly decreased for increasing formant shifts. Relating these raw formant values to the amount of F1 compensation requires taking into account the baseline F1, which we recovered from the control condition of the experiment.

3.2 Control condition

As noted in the introduction, estimating a subject's baseline vowel region using the first 15 trials tends to underestimate the size of the baseline region because successive trials are autocorrelated.

The baseline condition for this study, during which subjects produced 360 /ɛ/ vowels in 'head' with no formant shift, catalogued the acoustic variety in /ɛ/ formants typically produced by that subject. The convex hull surrounding the F1 and F2 produced in each of these vowels is shown in Figure 4. The standard deviation of these baseline vowel regions is approximately 30 Hz, which is in line with other, similar studies (e.g., Purcell & Munhall, 2006, *inter alia*).

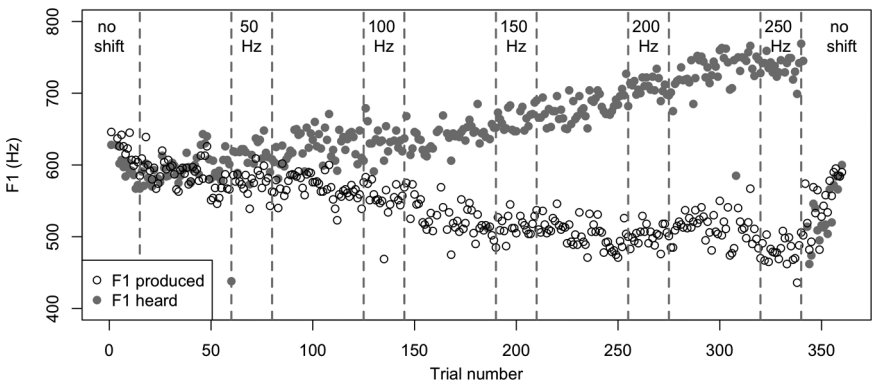

Figure 3. F1 from the /ɛ/ in 'head' over the course of the experiment (one typical subject shown here). Black circles mark F1 from the vowels that the subject produced at each trial, and gray filled circles show the altered F1 heard by the subject at each trial. Each gray circle/black circle pair represents one trial.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

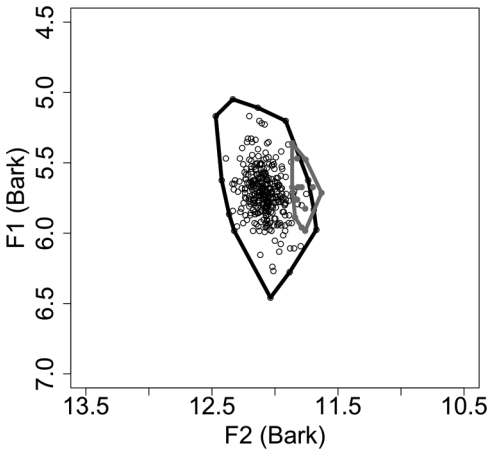

Figure 4. A typical subject's baseline region for the /ɛ/ in 'head'. Open circles mark the vowel formants extracted from the 360 vowels produced during the control condition, and the solid black line is the convex hull surrounding them. Gray circles mark the vowel formants produced during unaltered trials of the F1 shift experiment, and the smaller, gray convex hull outlines them.

This and all subsequent analyses are performed in Hz and also in Bark, a psychoacoustic scale based on the frequency response of the cochlea (Zwicker, 1961).

Figure 4 shows that the first 15 unaltered /ɛ/ formants from the F1 shift condition (outlined in gray) occupy a small portion of the /ɛ/ vowel space recorded during the baseline condition for the same subject (outlined in black). The means of the gray region and the black region differ because high variance and autocorrelation conspire to make those first 15 trials poor representatives of the subject's true baseline. The first /ɛ/ recorded in the F1 shift condition might lie anywhere within the large vowel region, and the formants recorded during the next 15 trials are influenced by that first vowel's location. Because calculation of compensation is crucially dependent on an accurate calculation of the subject's baseline, we augmented those first 15 trials with the additional 360 baseline trials to build a more comprehensive target region for each subject.

Using this augmented baseline, we estimated compensation within and across subjects using a mixed effects linear model. Although the model estimates only a single baseline for each subject, that point is estimated using formant measurements that cover the large, control vowel space rather than the small 15-trial baseline from the beginning of a particular session.

This method allowed us to use all of a subject's baseline trials when deciding on his baseline during a given session. The model allows for straightforward estimates of compensation along with confidence intervals for those estimates. Two major results arise from this analysis: (1) compensation is almost never complete, and (2) compensation decreases for increased formant shift.

$$\mathrm{F1} = \mathrm{baseline}_i + \mathrm{compensation}_j + \mathrm{error}_{ij}$$

$$\text{where } 1 \leq i \leq \#\text{subjects, and } 1 \leq j \leq \#\text{shifts.}$$

We modeled the F1 produced as a function of a baseline F1, which was permitted to vary by subject, and the formant shift. Each subject is assumed to have an idiosyncratic baseline F1, estimated from their 360 baseline vowels. Compensation at each formant shift step was the estimated distance between the baseline and the formants produced at that shift step. Ninety-five percent

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

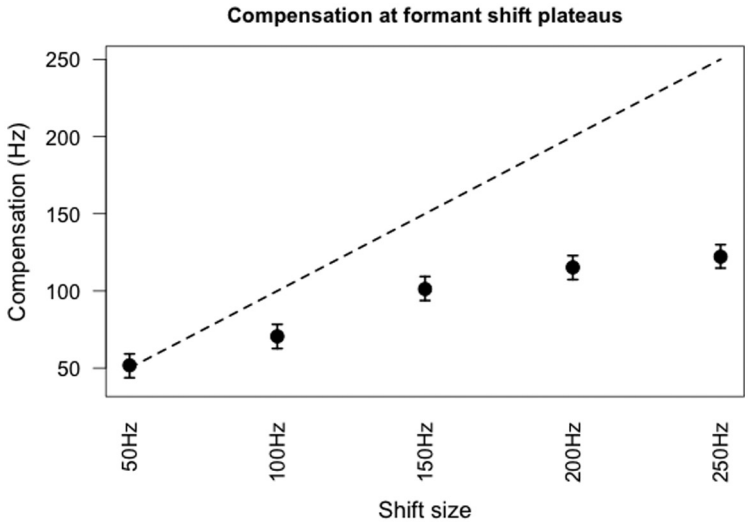

Figure 5. Raw compensation in Hertz, averaged across subjects, at each of the five formant shift steps. Error bars mark 95% confidence intervals for each plateau. For reference, the dotted line indicates what would be 100% compensation at each formant shift step.

confidence intervals for these estimates were obtained with 20,000 iterations of Markov Chain Monte Carlo sampling (see Baayen, Davidson, & Bates, 2008 for a clear account of how these confidence intervals are calculated). Using this model, we estimated each subject's baseline and compensation at each of the five formant shift steps.

The mean estimates of compensation at each of the five formant shift steps, along with their confidence intervals, are shown in Figure 5. For reference, complete compensation is marked with a dotted line. This figure demonstrates that compensation is approximately complete at a formant shift of 50 Hz, but is partial for all shifts greater than 50 Hz. Although the raw amount of compensation increases nonlinearly with increasing formant shift, the increase in compensation for altered auditory feedback does not nearly keep pace with the increasing shift in auditory feedback. Indeed, the amount of additional compensation is smaller at each successive formant shift after 150 Hz. Given this trend toward less compensation at larger formant shifts, it is possible that compensation would approach an asymptote for a sufficiently large formant shift (see MacDonald, Goldberg, & Munhall, 2010).

The trend toward less complete compensation at greater formant shifts is characteristic of individual subjects as well, as shown by Figure 6. This figure shows a representative subject's baseline and experimental /ɛ/ vowels. Each of the graphs in Figure 6 shows the formants during tokens produced at each of the five F1 feedback shift steps: 50 Hz, 100 Hz, 150 Hz, 200 Hz, and 250 Hz. The dashed shape in each graph outlines the vowels produced during control trials. The dark solid line in each graph is the convex hull of vowels produced when F1 feedback was shifted by the amount shown in the graph title. For example, the dark, solid shape in the leftmost graph outlines the vowels produced during trials with F1 feedback shifted by 50 Hz. The gray shape in each graph outlines these vowels after they have been shifted by 50 Hz; these are the vowels that subjects heard. As an example, a vowel produced with an F1 of 550 Hz in the leftmost graph would fall within the solid black shape, but after its 50 Hz shift, that vowel would be heard with an F1 of 600 Hz, which is within the solid gray shape. Notice that in the leftmost graph, the gray shape falls almost completely within the dotted shape, indicating that the shifted vowels that the subject heard

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

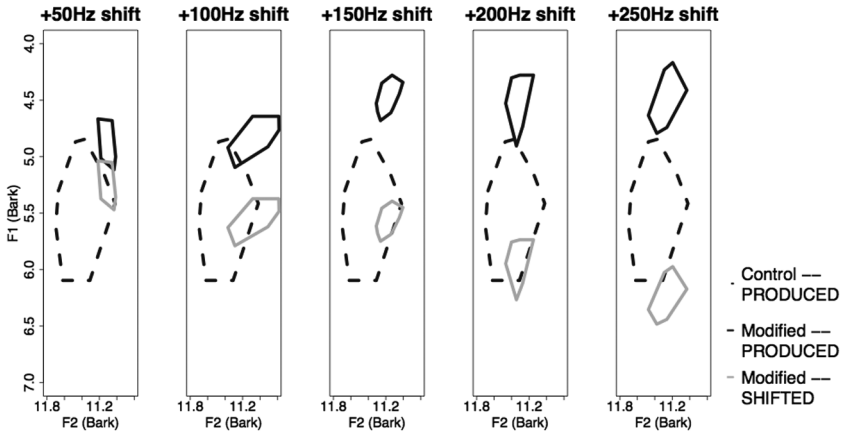

Figure 6. Productions of /ɛ/ during Experiment 1 plotted in F1-F2 Bark space against productions of /ɛ/ during control trials. Results for a typical subject are shown. For small feedback shifts, the light gray shape (formants heard as a result of the feedback shift) falls entirely within the dashed shape (the subject's baseline range), indicating that the vowels that the subject heard were all within his baseline region and that compensation was complete. As the amount of feedback shift increases (the dark solid shape), compensation is less and less complete.

were almost all within his baseline region, and that compensation was nearly complete. Compensation is likewise nearly complete for F1 feedback shifts of 100 Hz. As the amount of feedback shift increases to 150 Hz and beyond, the vowels that the subject hears are no longer within his baseline region, and compensation is less and less complete.

4 Discussion

This experiment used a stepwise feedback alteration design and a novel method of quantifying baseline vowel regions to measure compensation for feedback alterations of five sizes. Results demonstrated that percent compensation decreases monotonically as the formant shift increases. In particular, compensation was approximately complete for small shifts in auditory feedback and partial for large shifts in auditory feedback. This behavior is consistent with a speech motor control system that monitors both auditory and somatosensory feedback.[1]

Returning to the candidate explanations considered in the introduction, we find that this pattern of compensation cannot be explained by large vowel target regions. Large target regions predict no compensation for feedback shifts small enough that the altered feedback falls within the baseline region, and compensation that is increasingly complete as the amount of formant shift increases. Even if targets were larger than the baseline measured in this experiment, we would expect a wide range of unsystematic responses to changes in auditory feedback rather than the consistent decrease that was observed. The fact that we found compensation at even the 50 Hz shift, which sometimes falls within subjects' target regions, is additionally surprising because it is barely perceptible: untrained listeners generally cannot detect F1 changes smaller than 35–40 Hz (Kewley-Port, 2001).

We suggest that the decreasing completeness of compensation is a consequence of the integration of unusual auditory feedback with normal somatosensory feedback. For small shifts in F1, auditory feedback is slightly deviant and somatosensory feedback is normal. Because there is only a small discrepancy between the two types of feedback, subjects take both into account and compensate for the altered acoustic feedback. For large shifts in F1, auditory feedback is highly unusual while somatosensory

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

feedback remains normal, resulting in a large discrepancy between the two feedback sources. When the auditory-somatosensory discrepancy is large, the unusual auditory feedback might be downweighted, allowing the normal somatosensory feedback to attenuate the compensation response.

This explanation is consistent with results both from this experiment and from a recent study by MacDonald, Goldberg, and Munhall (2010), which measured compensation in response to a similar set of stepwise formant shifts. They found partial compensation for shifts in F1 and F2 auditory feedback, and suggested that speakers may reach a compensation asymptote, beyond which they are unwilling to stray any further from their baseline regions. There are several differences between the results of that study and this one, most importantly that MacDonald et al. find a linear relationship between compensation and feedback shift size. This difference may be due to their calculation of formant baseline, or perhaps to their formant manipulation, which involved a combination of F1 and F2 rather than a single formant.

Another possibility is that the autocorrelation between successive trials limits the amount of production change that can be achieved by the end of the 360 trials in this experiment. We quantified the degree to which the F1 of a given token can predict the F1 of the next token by measuring the autocorrelation at lag 1 for every subject, and found that the average autocorrelation across subjects is 0.23. If the amount of per-trial compensation is fixed by this autocorrelation, then subjects whose feedback is altered less ought to compensate more completely at their maximum shift than subjects in this experiment. Pilot data suggest that compensation is no more complete for 150 Hz feedback shifts than it is for 250 Hz feedback shifts, though investigation is ongoing. Certainly the lag between feedback shift and compensation for that shift is interesting and worthy of further study.

Finally, there are two ways in which perception might drive the decrease in compensation for increasing changes to auditory feedback. First, it is possible that listening to so many altered tokens of /æ/-like /ɛ/ vowels stretches the /ɛ/ baseline region toward /æ/. Such a boundary change would make formant-shifted feedback sound closer to the baseline region, and perhaps reduce the speaker's tendency to compensate. As Shiller et al. found in their /s/- /ʃ/ formant manipulation, it is unlikely that this effect dampens the compensation response appreciably, both because we observed compensation for auditory feedback even for tokens very close to the border of the baseline region, and because we observed compensation of 40–100%, which is much greater than the 10% change in phoneme boundary that was previously observed. To determine whether the effect of one vowel on surrounding vowels is amplified in a more ecologically valid setting, we are currently investigating whether compensation for altered auditory feedback differs when subjects receive altered feedback during a two-person object naming game.

Second, it is possible that percent compensation appears to be decreasing when measured in Hertz, but not when measured in the proper set of perceptually realistic coordinates. The main analysis considered Bark coordinates. In Figure 7, we consider three additional measurement systems that have been proposed based on properties of speech perception: (1) log F1; (2) the Bark-transformed difference between F2 and F1 (as suggested by Syrdal & Gopal, 1984); and (3) the difference between log (F1) and log (F0) (as suggested by Miller, 1989). In each case, we estimate compensation by re-fitting the model with the function of formants indicated rather than with F1 alone.

Compensation for increasing formant shifts in each of the subfigures of Figure 7 is nonlinear and decreasing. This is the same pattern as in Figure 5, which showed compensation in absolute F1 alone. That is, none of these perceptually realistic formant coordinate systems show complete compensation for increasing formant shifts, either. It remains possible that some other function of F1 or of multiple formants does fit the data linearly, but such a function would not be theoretically motivated. Whether we are storing F1 or some function correlated with F1, we argue that a purely auditory target is unlikely given that changes in somatosensory feedback alone generate compensatory responses, and that compensation for shifts in pitch feedback changes when somatosensory feedback is removed.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

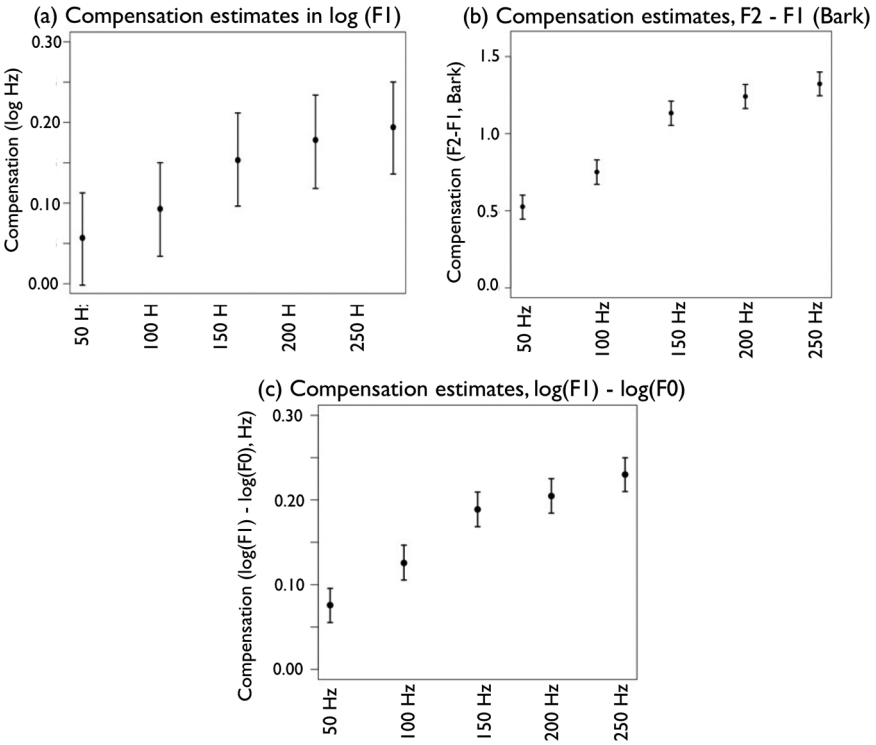

Figure 7. Absolute compensation model fit with data transformed according to several perceptual theories. To accommodate increased hearing sensitivity in lower frequencies, (a) plots absolute compensation on a log F1 scale. To accommodate the purported perceptual dependence of F1 on F2, (b) plots absolute compensation modeled with (F2 (in Bark) - F1 (in Bark)), as proposed by Syrdal and Gopal (1984). To accommodate effects of pitch on vowels, (c) plots absolute compensation modeled with (log (F1) - log (F0)), as proposed by Miller (1989).

This finding brings speech motor control models closer to some models of arm motor control, which also tend to incorporate multiple sources of feedback (Sober & Sabes, 2005). These arm motor control models already consider contributions from proprioceptive and visual feedback sources. In our case, a similar interplay of somatosensory and auditory feedback could act to stabilize the speech production system. Because there are large individual differences in degree of compensation, the relative weighting of somatosensory and auditory feedback may be specific to individual talkers. In the arm motor control literature, the learning of optimal contributions from multiple sensory sources has recently been modeled using Bayesian inference (Wolpert & Kawato, 1998; Körding & Wolpert, 2004). Future work could test whether retuning of the auditory system operates in this way.

There were two patterns of compensation that cannot be explained with the model we have just suggested. First, as shown in Figure 7 above, subjects changed their production of both F1 and F2 in response to a feedback shift in only F1, if only by a small amount. This may be a consequence of perception, in which vowels that are heard are mapped to near neighbors, or to the relative salience of somatosensory feedback in different regions of vowel space. The reason may also be articulatory; expected feedback might be compared to actual feedback in a way that is sensitive to adjacent vowels' articulation, or it may be difficult to produce a vowel that exactly opposes the shift in feedback. Second, subjects compensated more fully for shifts to higher F1 than for shifts to

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

lower F1. This, too, may result from perceptual, processing, or articulatory factors. Current work is investigating the source of these observations.

While multiple sources of feedback may have some effect on online speech monitoring and planning, this study makes clear that auditory and somatosensory feedback are two major but variably-weighted factors in determining whether an immediate articulatory correction is needed. For small discrepancies between auditory and somatosensory feedback, auditory feedback takes precedence, and for large discrepancies between auditory and somatosensory feedback, somatosensory feedback takes precedence.

Acknowledgements

This work was supported by an NSF GRFP grant to the first author, as well as NSF grant BCS-0926196 and NIH grant R01-DC010145. Thanks to Richard Hahn for valuable discussions regarding data analysis and to Ronald Sprouse for technical support.

Note

1 The experiment reported here has close ties to perturbation experiments involving the sidetone amplification effect, in which talkers compensate for changes in the perceived loudness of their voices (Chang-Yit, Pick, & Siegel, 1975). The profile of these responses closely parallels responses to our formant changes. When loudness is perturbed by a small amount, less than 1 dB, compensation is nearly complete. Large perturbations result in much smaller proportional changes (Heinks-Maldonado & Houde, 2005; Bauer, Mittal, Larson, & Hain, 2006). These results were broadly similar in spite of the fact that these were somewhat less linguistic tasks; subjects in these studies held out the vowel /u/ rather than saying real words.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Bauer, J. J., Mittal, J., Larson, C. R., & Hain, T. C. (2006). Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude. *Journal of the Acoustical Society of America*, *119*, 2363–2371.

Boersma, P., & Weenink, D. (2009). *Praat: Doing phonetics by computer* (Version 5.1.12) [Computer program].

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice F0 responses to manipulations in pitch feedback. *Journal of the Acoustical Society of America*, *103*, 3153–3161.

Cai, S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2010). Coordination of the first and second formants of the Mandarin triphthong /iau/ revealed by adaptation to auditory perturbations. *Journal of the Acoustical Society of America*, *127*(3), 2018–2018.

Chang-Yit, R., Pick, H. L., Jr., & Siegel, G. M. (1975). Reliability of sidetone amplification effect in vocal intensity. *Journal of Communication Disorders*, *8*, 317–324.

Diehl, R. L. (1981). Feature detectors for speech: A critical reappraisal. *Psychological Bulletin*, *68*, 1–18.

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*, 594–621.

Guenther, F. H. (2003). Neural control of speech movements. In A. Meyer and N. Schiller (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities* (pp. 209–240). Berlin: Mouton de Gruyter.

Guenther, F. H., & Barreca, D. M. (1997). Neural models for flexible control of redundant systems. In P. Morasso and V. Sanguineti (Eds.), *Self organization, computational maps and motor control* (pp. 383–421). Amsterdam: North Holland, Elsevier Science B.V.

Heinks-Maldonado, T. H., & Houde, J. F. (2005). Compensatory responses to brief perturbations of speech amplitude. *Acoustics Research Letters Online*, *6*, 131–137.

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, *45*, 295–310.

Jones, J. A., & Munhall, K. G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015

Jones, J. A., & Munhall, K. G. (2005). Remapping auditory-motor representations in voice production. *Current Biology*, *15*, 1768–1772.

Kewley-Port, D. (2001). Vowel formant discrimination II: Effects of stimulus uncertainty, consonantal context, and training. *Journal of the Acoustical Society of America*, *110*(4), 2141–2155.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247.

Larson, C. R., Altman, K. W., Liu, H., & Hain, T. C. (2008). Interactions between auditory and somatosensory feedback for voice F0 control. *Experimental Brain Research*, *187*, 613–621.

MacDonald, E. N., Goldberg, R., & Munhall, K. G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *Journal of the Acoustical Society of America, 127*(2), 1059–1068.

McAulay, R. J., & Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP, *34*(4), 744–754.

McAulay, R. J., & Quatieri, T. F. (1991). Low-rate speech coding based on the sinusoidal model. In S. Furui & M. M. Sondhi (Eds.), *Advances in speech signal processing* (Vol. 76, pp. 165–208). New York, NY: Marcel Dekker.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, *85*, 2114–2134.

Perkell, J. S., Guenther, F. H., Lane, H., Matthies, M. L., Perrier, P., Vick, J., Wilhelms-Tricarico, R., & Zandipour, M. (2000). A theory of speech motor control and supporting data from speakers with normal hearing and with profound hearing loss. *Journal of Phonetics*, *28*, 233–272.

Perrier, P., Lœvenbruck, H., & Payan, Y. (1996). Control of tongue movements in speech: The Equilibrium Point Hypothesis perspective. *Journal of Phonetics*, *24*, 53–75.

Pile, E. J. S, Dajani, H. R., Purcell, D. W., & Munhall, K. G. (2007). Talking under conditions of altered auditory feedback: Does adaptation of one vowel generalize to other vowels? In *Proceedings of the International Conference of Phonetic Sciences* (Vol. XVI, pp. 645–648).

Purcell, D. W., & Munhall, K. G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *Journal of the Acoustical Society of America*, *120*, 966–977.

Quatieri, T. F. (2002). *Discrete-time speech processing: Principles and practice*. Upper Saddle River: Prentice Hall PTR.

Quatieri, T. F., & McAulay, R. J. (1986). Speech transformations based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP, *34*, 1449–1464.

Quatieri, T. F., & McAulay, R. J. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, *40*, 497–510.

Sanguineti, V., Laboissière, R., & Ostry, D. J. (1998). A dynamic biomechanical model for neural control of speech production. *Journal of the Acoustical Society of America*, *103*, 1615–1627.

Schroeder, C. E., Lindsley, R. W., Specht, C., Marcovici, A., Smiley, J. F., & Javitt, D. C. (2001). Somatosensory input to auditory association cortex in the macaque monkey. *Journal of Neurophysiology*, *85*, 1322–1327.

Shiba, K., Miura, T., Yuza, J., Sakamoto, T., & Nakajima, Y. (1999). Laryngeal afferent inputs during vocalization in the cat. *NeuroReport*, *10*, 987–991.

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *Journal of the Acoustical Society of America*, *125*(2), 1103–1113.

Sober, S. J., & Sabes, P. N. (2005). Flexible strategies for sensory integration during motor planning. *Nature Neuroscience*, *8*, 490–497.

Syrdal, A., & Gopal, H. (1984). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, *79*, 1086–1100.

Tremblay, S., Shiller, D. M., & Ostry, D. J. (2003). Somatosensory basis of speech production. *Nature*, *423*, 866–869.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks*, *11*, 1317–1329.

Wyke, B. (1983). Neuromuscular control systems in voice production. In D. M. Bless and J. H. Abbs (Eds.), *Vocal fold physiology: Contemporary research and clinical issues* (pp. 71–76). San Diego, CA: College Hill Press.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands. *Journal of the Acoustical Society of America*, *33*, 248.

Downloaded from las.sagepub.com at UNIV CALIFORNIA BERKELEY LIB on January 12, 2015