Partial compensation in speech adaptation

Shira Katseff
Revised June 6, 2008

1 Background

Just as our eyes adjust to wearing colored glasses, our speech articulators are able to adjust to altered speech input. In a string of previous experiments (beginning with Houde & Jordan, 1998), subjects were asked to speak into a microphone while wearing headphones configured such that their speech feedback came primarily through the headphones. Speakers whose feedback is manipulated so that, for example, they hear the pitch of their voices shifted 200Hz higher change the way they speak in order to sound more normal to their own ears. This phenomenon is called compensation. Compensation has also been induced (to various degrees) by manipulating the first and second formants of vowels.

1.1 Motivation

The compensation response is surprising because speech is as much a motor activity as an acoustic one. While each vowel has a set of characteristic frequencies, those frequencies come from characteristic movements of our speech articulators. The compensation experiments just cited manipulate auditory feedback while leaving the articulators in place. This means that speech planning must involve an acoustic component, contrary to motor theories of speech perception (Lieberman & Mattingly, 1985). But details of the compensation response hint that speech planning is not based wholly on acoustics, and that when based on acoustics, speech is neither treated as a whole, undecomposed waveform, nor as individual formants.

This paper explores variation in the compensation response within and across subjects. The goal is to determine what part of the speech goal is acoustic and what part sensorimotor, and how much individual formants or formant ratios can account for the responses we see.
1.2 Goals in reaching experiments

Sensorimotor adaptation research has been replicated in a variety of domains. For a sense of how these experiments are structured and what they have shown about the structure of neural feedback, it will be useful to look at variations on the theme of reaching experiments (see Welch, 1986 for a review).

In a standard reaching experiment, subjects put on prism glasses that shift their visual field. Then they reach for a target object on a table. At first, subjects reach in the direction that their vision is shifted, and miss; their visual map no longer corresponds to the physical world. But over time, subjects learn a new map from their visual map to the world and are able to grasp the object. This is compensation. When the prism glasses are removed, subjects do not immediately readjust; instead, they overreach and again miss the object. After a brief period, the adaptation is lost and subjects’ visual map returns to normal.

![Illustration of sensorimotor adaptation](image)

Figure 1: Illustration of sensorimotor adaptation. When subjects’ visual fields are shifted, they slowly relearn to reach a target object. When their vision returns to normal, they initially overcompensate by reaching in the opposite direction.

In speech, a subject listens to his own resynthesized speech as he talks. There is no delay in the resynthesis, and so with no shift, the experiment roughly mimics natural speaking conditions. Over the course of the experiment, the resynthesized speech is altered. Just as the subject at first overreaches in the visual version of this experiment, the subject
in the speech version speaks normally for a short period, producing speech that sounds strange due to the resynthesis. But after that short period, the subject changes his speech production, analogous to successfully grasping the object. When resynthesized speech returns to normal, the subject again “overreaches”, continuing to produce hypernormal speech in the altered direction.

There is a clear analogy between these two experiments, but one major difference. Little attention has been paid to the fact that in speech experiments, the subject never “grasps the object”. Though subjects change the way they speak, the amount of change in production is always less than the amount that their speech was altered. Their compensation is always partial.

1.3 Goals in speech adaptation experiments

The fact that speech adaptation happens tells us that somewhere we are storing the following:

1. the sound we hear ourselves say (feedback)
2. the sound we expect ourselves to say (expectation)
3. a method of comparing feedback to expectation
4. a method of correcting mismatches between feedback and expectation

We cannot store this expectation simply in gestural terms; during these perturbation experiments, one’s acoustic feedback is altered while motor feedback stays the same. The fact that this induces compensation implies that some part of our speech expectation is acoustic. Importantly, some of the expectation is motoric; parallel experiments with jaw perturbation (e.g., Nasir & Ostry, 2006) testify to a contributing role for motor expectations.

The same is true of aerodynamic expectations. There is an interesting literature examining the degree to which speech is controlled through regulation of subglottal pressure (beginning with Warren, 1986). While there may well be a role for aerodynamics in speech control, especially in well-practiced motor routines, aerodynamic factors cannot be the primary explanation here. Compensation is seen in these experiments even though intraoral pressure remains undisturbed.

Thus expectations must be in large part acoustic. The purpose of this paper is to explore the components of this acoustic expectation and the relative importance of these components. Because speech is a complex task, there are two possibilities. First, our expectation may be acoustic and parametrized. In this case, acoustic signals would be analyzed and repackaged before being stored in memory. In this case, we should react only to changes in the stored
portions of the acoustic signal. More specifically, we would be able to gauge the degree to which any acoustic feature is stored in memory by examining the amount of compensation it induces. Second, our expectation may be acoustic and holistic. If this were the case, we should be equally sensitive to any perturbation in the acoustic signal, and the degree of compensation should equal its distance from the acoustic target.

Because a parametrized signal may seem difficult to extract from a complex time-varying signal like speech, we will deal in this paper (as in the previous literature) with compensation for vowels. Vowels are much more stable over time than other sorts of speech sounds and are thus much better candidates for a parametrization theory.

1.3.1 Parametrized acoustic targets

We can use source filter theory to come up with hypotheses about what constitutes our memory for vowels. According to this theory, a vowel is produced by air that flows through vibrating vocal folds and gets filtered by the vocal tract. The position of the tongue, lips, and teeth, in conjunction with the shape of the vocal tract, determine the vowel’s formant frequencies. In other words, the vocal folds provide the vowel’s harmonic frequencies, and the vocal tract provides the spectral envelope.

With this model of the vocal tract in mind, formant frequencies are the most obvious candidates for a parametrized target. We could, for example, store the first five formants for each of our vowels. The trouble with this sort of a theory is that no two people have exactly the same sets of formants, and so identifying a vowel based on a vector of five formant frequencies would be impossible without an additional layer of processing.

To clarify what this extra processing might entail, much research has gone into identifying vowel “invariants”, characteristics of vowels that remain approximately the same across talkers. If we could identify a set of invariants for any vowel, we could reasonably propose that that vowel was stored with exactly those invariants. Invariants would also simplify vowel perception. Because of the salience of formant frequencies in vowels, theories along these lines tend to identify functions of formants as invariants. Various incarnations of this theory have been around for over a hundred years (for a detailed discussion, see Miller, 1989).

Most of the models proposed in the last 50 years hinge on Potter and Steinberg’s (?) hypothesis, that vowels are recognized based on a pattern of activation on the basilar membrane. This hypothesis has shaped our understanding of vowel recognition. Since one aim of this project is to investigate our internal expectations for vowel sounds, and since vowel recognition is intimately tied to these internal expectations, we can test the validity of these models via speech adaptation. Following is a brief review of the more prominent models.
Peterson and Barney (Peterson & Barney, 1952) measured formants in men, women, and children. They found that vowel spaces differed greatly among the three groups. Peterson observed that formant profiles from the three groups could be minimized by simple translation.

Traummüller (1981) started his search for invariants with a perceptual task. He showed that varying the Bark distance between F1 and F0 (essentially a logarithmic distance metric) changed the way Central Bavarians categorized vowels. He used synthesized, isolated vowels. Some were single formant vowels, and some had a fixed F2, F3, and F4. Traummüller made two key observations. One was that, all other formants being equal, the perceived vowel lowered as F1 increased. The other was that the distance between F1 and F0 was used for perception when they were less than 3 Bark apart. When F1 was far from F0, listeners’ judgments corresponded to the distance between F2 and F1 instead.

Syrdal and Gopal similarly depended on Bark distances between formants in their talker normalization model. They claimed that there are two stages of vowel processing. In the first stage, Bark distances between F1 and F0 are computed along with distances between higher formants. In the second stage, differences are classified as greater than or less than 3 Bark. Vowel identity is based on these differences, with vowel height corresponding to F1-F0, and vowel place of articulation corresponding to F3-F2 (Syrdal & Gopal, 1986).

Miller (1989) was able to explain vowel category boundaries using formant ratios. In his “auditory-perceptual” theory, Miller proposed an auditory reference frame composed of functions of F0, F1 and F2. His favored auditory space used log differences between formants. He accounted for differences between speakers with what he calls a “sensory reference” (SR), which uses the geometric mean of the first three formants.

Hoemeke and Diehl (1994) synthesized vowel continua from /i/ to /ɪ/, /ɪ/ to /ɛ/, and /ɛ/ to /æ/ by varying F1-F3. F0 was varied independently within each of these tokens. By asking subjects to categorize each of these vowels with different synthesized F0, they were able to show that F1-F0 was only a better predictor of vowel category than F1 alone for the /ɪ/-/ɛ/ continuum.

Later, Fahey and Diehl (1996) synthesized vowels along an /ɪ/-/ɛ/ continuum with a fixed F4 and F5 and a varying F0, F1, F2, and F3. They filtered the fundamental frequency out of some of their stimuli. They found that the distance between F1 and F0 was correlated with vowel category when F0 was present. When absent, identification of vowels corresponded to where F0 would have been had the signal been left unfiltered, a “phantom” F0. In other words, F0 is used for categorization even when it must be inferred. Recent computational methods of vowel classification take into account exactly these features. Hillenbrand, Clark, and Nearey (2001), for example, use F0 along with duration.

Diehl (2000) tries to reconcile some of these conflicting results. He accounts for perceptual category boundaries with three linear combinations of the first three formants: F1-F0,
F3-F2, and F2 itself. However, no one measure can account for all vowels, and it is not clear that these results carry over to natural speech; the stimuli at issue in this paper, like those in most of the previous papers, were synthetic vowels.

What is the relevance of these studies to expectations of one’s own voice? In a theory of talker normalization like those summarized above, one’s own voice must be broken down and recast in the same terms as the voices of all other talkers.

On the whole, studies conducted over the last 15 years increasingly point out the differences between vowels based on both consonant environment and speaker. In spite of a large volume of related research, the critical features of isolated vowels are still in dispute.

1.4 Compensation constrained by acoustic invariants

A speech adaptation experiment is one way to shed light on this dispute. If we depend on a parametrized vowel memory, apparent incompleteness in compensation is actually an artifact of the way the vowel is stored. If we are storing \( F1 \) itself, we should expect complete compensation for alteration in \( F1 \) feedback. But if we are storing, for example, \( F1 \) modulo 150, a large range of \( F1 \) values should not cause any compensation. If we are storing a function of several formants, our expectations are more complicated. For example, if we are storing \( F2-F1 \), there is a trading relation between those two formants. We might compensate for feedback alteration where \( F1 \) is lowered in three ways: (1) raising the \( F1 \) we produce, (2) lowering the \( F2 \) we produce, or (3) both lowering \( F2 \) and raising \( F1 \). If we choose options (2) or (3) and we only monitor changes in \( F1 \), compensation will appear to be incomplete. Thus if we are aiming for a parametrized acoustic target in our speech, we can distinguish between these possibilities by investigating the relationship between several candidate variables involved in compensation.

We can do this by repeating the experiments of past researchers for several values of vowel features that we know to be relevant to compensation. For example, rather than altering \( F0 \) by 1 semitone, we can alter it by 0.25, 0.5, 0.75, 1.0, and 1.25 semitones in 5 successive experiments. We can then draw a “\( F0 \) compensation curve” showing the degree of compensation for each of our 5 values of \( F0 \). It is likely that the degree of compensation will differ depending on the magnitude of the alteration. It is also likely that the difference in percent compensation for a given change in alteration will be different for each variable.

We can get an idea of the contribution of each variable to the total compensation by comparing “compensation curves” like those pictured in Figure 1.4 for each of our 3 manipulated variables. If \( F2-F1 \) is stored, for example, we might expect \( F1 \) and \( F2 \) to show a curve of the same shape over a range of formant alterations. This measurement is valuable even if we are storing an acoustic feature closely correlated with our formants of interest.
Figure 2: Possibilities for relationship between amount of formant shift and degree of compensation. (a) One possibility is a linear relationship between the amount a variable has been altered and the amount a subject compensates. (b) It is also possible that there is a logarithmic relationship between the amount a variable has been altered and the amount a subject compensates. (c) A third possibility is no dependence between the degree of compensation and the amount the variable is altered.

rather than the formants themselves because we are looking at differences between degrees of compensation for different formant manipulations. Even if we have encoded a correlate of, say, F1, any change in degree of F1 compensation for a change in F1 shift must reflect a parallel change in the correlated “true” parameter. That is, even if a correlate of F1 is stored rather than F1 itself, the same relationship will hold between the true parameter and F2.

For example, if we find that compensation for F1 is linear, like in Figure 1.4(a), and compensation for F2 is in the shape of a 1/x curve, we can conjecture that F1 and F2 (or their correlates) are stored in a F1/F2-like relationship.

While a vowel feature based representation of vowels is possible and is likely bound up with relative formant values, it is worth considering other possible representations as well.

1.5 Holistic acoustic targets and DIVA

If we instead assume that in memory, we have access to a more faithful replica of a speech signal, we do not need to be as concerned with vowel invariants. We can instead assume that for each vowel, we store a large set of acceptable waveforms, and that we can compare these waveforms to our incoming signal to determine whether what we are hearing is an acceptable rendition of the vowel we expect to hear. With the original waveforms for reference and a reasonable method of computing acoustic similarity, we ought to be perfect judges of how the vowel we hear differs from the vowel we expect to hear. In
In this case, any incompleteness in compensation must be due to sensorimotor factors. To understand how sensorimotor factors are involved in compensation, let us consider how speech is planned.

In order to speak, we need both a speech goal and a plan for reaching that goal. The plan gets translated into muscle activations, which then produce motor movements. The usual term for a phonemic goal is a target, and the literature is split on whether those targets are articulatory or acoustic. There is an analogous debate in the motor control literature: researchers believe that when one reaches for an object with one’s arm, that goal may be phrased in terms of space (the location of the goal) or joint parameters (the joint and limb angles required to reach the goal). Compelling linguistic evidence is present to support both articulatory and acoustic goals. A theory with auditory goals is helpful for explaining, for example, the fact that American English /ʃ/ is produced with multiple articulatory configurations, even within a single speaker, while a gestural theory can be useful for explaining realtime speech phenomena like coarticulation. Because speech adaptation is a response to an acoustic perturbation without motor perturbation, it is more consistent with an acoustic target. However, any theory relying on auditory goals has the burden of explaining two pieces of evidence for gestural or motor targets. First, when planning a sequence of motor movements, people have “preferred articulator configurations; that is, there is much more consistency in motor articulation than is necessary to meet acoustic goals. Therefore there must be some bias toward producing particular sounds in particular ways (Guenther & Barreca, 1997). Second, loss of orosensory feedback renders speech unintelligible, so this sort of information is certainly transmitted and somehow used in speech production (MacNeilage, Rootes, & Chase, 1967).

One solution is to combine these approaches. For example, one might aim for an auditory goal, but constrain the motor movements used to get there. More complete models of the speech production system tend to work this way, for example, the Directions Into Velocities of Articulators (DIVA) model (Guenther, 1995). This is a neural computational model that starts with acoustic speech targets and maps them to motor movements via orosensory variables. An internal model for what sound to expect from particular articulator movements is learned in infancy via babbling. In this model, Guenther argues that in order for a gestural target theory to work, one must have good, frequently updated information about the current shape of the vocal tract. Without this information, it is impossible to plan gestural movements optimally. He claims that we do not have access to this information. While we have access to muscle length information, extrapolating from muscle lengths to constriction shape, size, and location is very hard. The calculations are specific to the individual, and muscle lengths are interdependent and have nonlinear effects on the constriction. It is possible to learn them, but there must be constant relearning over time as the muscles and vocal tract develop. Acoustic signals, on the other hand, can be yield useful information about constriction location.
This class of speech planning theories posits that when we produce a given phoneme, we have a target sound in mind. Compensation happens when we aim for but do not reach a sound within that target region. Because targets are constant, compensation should always bring production to some point in the target region. This sort of a theory makes two predictions with respect to compensation. One, onset of compensation should be sudden. There should be a range of acceptable productions (corresponding to points within the target region. Compensation should be present in all productions perceived to be outside this region, and compensation should increase as the distance from the target region increases.

To clarify, imagine that your baseline production is near the center of your target region, as in Figure 3(a). Then imagine that, as in Figure 3(b), your speech is shifted to a location 10 Hz from the baseline. This theory predicts that you would compensate, changing your production so that you hear a phoneme in the center of that target region, 10 Hz away. You are allowed to miss that target as long as your speech falls within that target region; any phoneme within its borders is equally acceptable. Compensation would be \( \frac{6-4}{10} \), or 60-100% of the perturbation. Now imagine that, as in Figure 3(c), speech is perturbed to a location 30 Hz from the baseline, or 25 Hz from the border of the target region. The theory now predicts that you will compensate \( \frac{25-30}{30} \), or 83-100% of the perturbation.

![Figure 3: Increasing degree of compensation. In (a), the baseline production, x, falls within the acoustic target region (indicated by the circle) and no compensation is required. In (b), the production, y, falls slightly outside the acoustic target region. The subject compensates by producing a sound within the acoustic target region, represented by z. In (c), the production, y, falls far outside the acoustic target region. The subject compensates by producing z, a sound falling within the acoustic target region. This sound is a larger percentage of the distance between the production and the baseline than is the z in panel (b).](image-url)
Researchers involved with acoustic speech perturbation have primarily been interested in the adaptation response. Compensation is widely acknowledged but incompletely understood. So far, experimental designs have implemented a single maximum feedback change, called a *maximum hold*. As a result, we have an idea of how much speakers will compensate for 200 Hz shifts in F1 and F2 and to 1 semitone shifts in F0. The results are intriguing: none of the experiments ever show more than 50% compensation, on the average, by subjects. That is, if an average subject’s F1 is shifted up by 200 Hz, that subject will lower his/her own F1 by about 100 Hz. As far as the DIVA model is concerned, this means that the target region is very large.

In general, these experiments work by asking subjects to repeat a target word while hearing their voice through a set of headphones. Their speech is played back in realtime by taking in small time windows of sound and quickly resynthesizing them before outputting them to the headphones. The process is fast enough that the speech stream is continuous and does not sound delayed. The method employed for this paper is McAulay-Quatieri synthesis (Quatieri & McAulay, 1986). This method can replicate a speech signal with very good accuracy from a small amount of data.

Different research groups have observed subjects from a couple of language backgrounds adapting to a variety of speech attribute changes. Jones & Munhall performed an adaptation experiment with English speakers, shifting the pitch of their voices by 1 semitone (Jones & Munhall, 2000). On average, these speakers compensated about 25%. Several years later, the same research group performed an adaptation experiment with speakers of Chinese. Their voices were pitch shifted by 1 semitone, and they compensated for about 40% of this change (Jones & Munhall, 2005).

Several experiments manipulated vowel formants. Purcell & Munhall shifted the first formant up to a maximum hold of 200 Hz. They observed 25% compensation (Purcell & Munhall, 2006b). Some members of this group also ran an experiment shifting either the first or the second formant up to 200 Hz. They observed 22.5% compensation (Pile, Dajani, Purcell, & Munhall, 2007).

Clearly, compensation during speech adaptation is incomplete and variable. The reasons for this might be motoric or perceptual. It is possible that people have entrenched motor programs for their /ɛ/ vowels and are not willing to change them much no matter what they hear. On the other hand, it is also possible that stimuli shifted by one formant sound odd or do not square with what speakers have stored for /ɛ/, so as an acoustic matter they simply don’t know how to compensate.

There is intriguing evidence to support the latter interpretation. One experiment has attempted to change both F1 and F2. In this experiment (Houde & Jordan, 2002), subjects’
feedback was altered in a straight line along their “vowel triangle” from /e/ up to /i/. Though intersubject variability ranged from 10% to 90%, mean compensation was 50%, much higher than compensation for either formant alone. An important caveat, however, was that the maximum hold for the dual formant change was different from that for the single formant changes.

Summary of previous speech adaptation experiments

<table>
<thead>
<tr>
<th>Research Group</th>
<th>Formant</th>
<th>Amount of Shift</th>
<th>Language</th>
<th>% Compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jones &amp; Munhall, 2005</td>
<td>F0</td>
<td>1 semitone</td>
<td>Mandarin</td>
<td>40%</td>
</tr>
<tr>
<td>Jones &amp; Munhall, 2000</td>
<td>F0</td>
<td>1 semitone</td>
<td>English</td>
<td>25%</td>
</tr>
<tr>
<td>Purcell &amp; Munhall, 2006b</td>
<td>F1</td>
<td>200 Hz</td>
<td>English</td>
<td>25%</td>
</tr>
<tr>
<td>Pile et al., 2007</td>
<td>F1 or F2</td>
<td>200 Hz</td>
<td>English</td>
<td>22.5%</td>
</tr>
<tr>
<td>Houde &amp; Jordan, 2002</td>
<td>F1 &amp; F2</td>
<td>200 Hz</td>
<td>English</td>
<td>50%</td>
</tr>
</tbody>
</table>

There is a growing literature on adaptation to sudden changes as well. In this sort of experiment, subjects are asked to hold out a vowel for as long as possible. As they phonate, formants in their vowel are perturbed suddenly, and changes in production are measured. Generally, the amount of total shift is less in this sort of experiment, and there is less compensation. Classically, Burnett et al. altered F0 in subjects and observed compensation behavior after F0 shifts of varying sizes, from 25 to 300 cents (Burnett, Freedland, Larson, & Hain, 1998). With larger shifts, subjects had a greater tendency to “follow” the direction of the pitch change rather than compensate for, or “oppose” it. More recently, Purcell & Munhall perturbed subjects’ first formants by a single large shift of 135Hz (Purcell & Munhall, 2006a). On average, subjects responded with 10-16% compensation.

2.1 Interim conclusions

In spite of significant intersubject variation, the average percentage compensation is remarkably consistent across experiments. People compensate for changes in F0, F1, F2, and a combination of F1 and F2. On the whole, it seems that compensation for F0 in Mandarin and for a combination of F1 and F2 in English seem to induce the most compensation. Shifts of F1 or F2 alone cause less compensation.

3 Questions

The fact that we compensate and adapt to alterations in F0, F1, and F2 means that each of them must be part of our memory for speech. Subjects who change their production must first notice that the speech they are hearing is not the speech that they expect to hear. We do not know exactly why this compensation happens.
It is possible to take a thoroughly sensorimotor view of the task and attribute all compensation to details of articulatory planning. Such a model would assume that we store a perfect replica of what we hear and have a foolproof method of matching it against a static expectation. We must further assume that our expectation contains the formant that we are manipulating. By this I mean that if we are shifting F2, our expectation must contain F2, and not a function of F2 or something roughly correlated with F2.

Conversely, we could take a thoroughly perceptual view of this task and attribute all compensation to details of perceptual memory. That is, partial compensation is due entirely to the way we store vowels. In order to ascribe to this sort of a model, we would have to assume that subjects do not store exactly what we are shifting during the experiment, and so we are not fully aware of the shift. That is, if we are not storing F2, but some quantity roughly correlated with F2, our awareness that something in speech has been shifted will be correspondingly rough. Intuitively, this accords with findings from post-session interviews with subjects. In general, subjects were totally unaware of all gradual formant shifts as they occurred, but many did notice the sudden drop to normal at the end of the experiment. It is worth mentioning that this was at best a dim awareness, as subjects tended to characterize the change as “pitch”, even when the quantity being shifted was not pitch, but F1 or F2.

To choose between these two competing models, this paper asks the following two questions:

1. How can we acoustically describe our representations of the /ɛ/ in ‘head’? We can do this by causing our subjects to compensate and looking at the degree to which their compensatory production of one formant affects their production of other formants. While the dependence between formants might vary between subjects, there should be a single representation within a single person. Consequently, we should see the same relative amount of formant change for all five amounts of formant shift.

2. To what degree is our representation affected by motor plans? We should expect entrenched motor routines to damp the compensation response. Compensation that decreases for increasing amounts of formant shift can be explained by interference from motor feedback.

4 Method

4.1 Procedure

Participants (n=8, all males) are seated in a soundproof booth and wear an AKG HSC-271 Professional headset. Their speech is routed from the headset microphone through a
Delta 44 sound card. Speech is analyzed and resynthesized in real-time with using feedback alteration software written by John Houde. Resynthesized speech output is played back through the headset’s earphones.

In order to become accustomed to hearing themselves through headphones, subjects are first asked to read a short passage. No recording of speech or speech alteration occurs during this acclimatization period.

Figure 4: Schematic of Experimental Setup. Subjects speak into the microphone portion of a headset. Their speech is analyzed, then resynthesized (and shifted, if necessary) and fed back into the headset’s earphones.

Once subjects are accustomed to hearing their resynthesized voices, they are asked to produce a series of CVC English words containing a variety of vowels. Productions of these words are recorded, but speech is not altered. These words are said 5 times apiece and used to establish baseline F0 and a baseline vowel space.

Baseline words include /hid/, /hid/, /hæd/, /hæd/, /hæd/, /hoʊd/, /hun/. Formant values are used to fine-tune the online formant shift algorithm: the algorithm is sensitive to formant location.

After baseline values are set, subjects are told that they will be prompted to say a word, and that their task is to respond to the prompt as quickly as possible. They are not informed that their speech will be altered. A full explanation of the study and its purpose
is given at the end of the experiment.

To complete the main portion of the experiment, subjects interact with a MATLAB program that leads them through a multistage experiment. It instructs them to say ‘head’ at regular intervals. Over the course of their repetitions, their feedback is altered very slowly. The procedure in this study diverges somewhat from previous studies because its purpose is to document compensation behavior for a set of changes in feedback rather than adaptation behavior for a return to normal feedback. Rather than ramp up to a single maximum hold and return to normal five different times, feedback is ramped to the five plateaus of interest in a stepwise fashion, resulting in a staircase pattern of formant changes. For clarity, the top of each “stair” will be referred to as a perturbation plateau.

![Change in feedback over the course of each experiment.](image)

Figure 5: Change in feedback over the course of each experiment. There are 360 trials in each experiment. These 360 trials are composed of 6 regions of equal formant alteration (plateaus) connected by ramps of slowly increasing feedback alteration.

Each experiment ramps from baseline (no alteration) to a large maximum amount of alteration (20 Hz for F0; 250 Hz for F1 and F2). The ramp levels off at 4 “plateaus”, evenly spaced between the baseline and the maximum alteration. Perturbed feedback remains constant for 20 trials at each plateau.

Maximum changes (in Hertz) for each variable:

<table>
<thead>
<tr>
<th>Formant</th>
<th>First Plateau</th>
<th>Second Plateau</th>
<th>Third Plateau</th>
<th>Fourth Plateau</th>
<th>Fifth Plateau</th>
</tr>
</thead>
<tbody>
<tr>
<td>F0</td>
<td>4 Hz</td>
<td>8 Hz</td>
<td>12 Hz</td>
<td>16 Hz</td>
<td>20 Hz</td>
</tr>
<tr>
<td>F1</td>
<td>50 Hz</td>
<td>100 Hz</td>
<td>150 Hz</td>
<td>200 Hz</td>
<td>250 Hz</td>
</tr>
<tr>
<td>F2</td>
<td>50 Hz</td>
<td>100 Hz</td>
<td>150 Hz</td>
<td>200 Hz</td>
<td>250 Hz</td>
</tr>
</tbody>
</table>

The multistage experiment described above is repeated once each for F0, F1, and F2 over the course of four days. Order is pseudorandomized. One acoustic variable is altered on each of the first three days. A control experiment in which feedback is not altered at all takes place on day four.
For interpretability, results will be reported in both Bark and Hertz, but the choice of measurement scale is not particularly important to the paper. Subjects do not seem to be sensitive to units of step size on the scale used in this experiment (4/15 Hz/trial for F0; 3/7 Hz/trial for F1 and F2). More extensive testing to confirm this observation is in progress.

4.2 Measurements taken

Degree of compensation (how much a subject compensates relative to the maximum altered feedback) for F0, F1, F2, and F3 is measured after each trial. Acoustic measurements were measured using PRAAT (?, ?)

4.3 Formant shifting

Speech synthesis is not generally an instantaneous procedure. The task of the analysis-synthesis system is to take in the speech that is produced and recreate it immediately, with or without formant modifications. The system used in this experiment is called a "Feedback Alteration Device", or FAD, by its author, John Houde. It is based on McAulay-Quatieri synthesis, which has analytically been shown to produce excellent reproductions of time-varying signals.

When a subject speaks, he produces a continuous sound wave. The FAD continually stores and analyzes small chunks of this wave. Each chunk is turned via FFT into a narrow band spectrum. Rather than store the entire spectrum, only certain spectral parameters are saved, which makes resynthesis faster and more efficient. Pitch is extracted by averaging the distance between spectral peaks. The first formant is estimated by the center of mass of peaks in a set F1 range. Higher formants are estimated by spectral peaks in the appropriate regions of the spectrum, and the temporal envelope is estimated from the relative peak heights. Results of this analysis method have been compared with LPC analysis and yielded similar results. The key to this method is connecting peaks from one window to the corresponding peaks in the next window. In most cases a peak will continue from one window to the next. If it does not, it either is the beginning (a peak “birth”) or the end (a peak “death”). During synthesis, peaks are connected, and the formants and spectral envelope are combined to create a spectrum. This spectrum is converted back into a waveform and fed through a sound card to the subject’s headphones.

The delay incurred by the analysis and resynthesis is approximately 10ms. This delay is small enough that it is not noticeable; the synthesis is effectively realtime. As mentioned earlier, all subjects were specifically asked about the quality of feedback in post-session interviews, and none noticed any sort of delay.
4.4 Considerations

While the method of synthesis is meant to imitate natural speech as closely as possible, both practical and theoretical considerations make perfect replication impossible. For example, lower pitch is physiologically associated with a longer vocal tract. Such a vocal tract would have slightly different resonances, and consequently, slightly different formant values. When we artificially shift pitch, we can take an atheoretical stance and maintain other formants at their current location, or we can shift the other formants according to a particular model of the vocal tract. In this study, we chose to perturb only single formants, leaving all other formants the same. This was the best way to control feedback across subjects: otherwise, subjects with different baseline spectral characteristics would have different feedback shifts. However, this choice also contributes to unnaturalness in feedback.

In addition, moving a formant necessarily requires adding amplitude to some location in the spectrum and subtracting it from another location. The quality of synthesis depends on modeling the underlying spectrum so that we know what remains when a formant is removed. In this experiment, the underlying spectrum is estimated by the spectral envelope. We must likewise have a model of the bandwidth of synthesized formants. In order to optimize processing speed, formant bandwidths are fixed at a constant value.

Finally, practical considerations of space and speed make it prudent to store a minimal amount of information about the spectrum. In this study, we stored limited information about spectral noise.

Together, these choices should be a matter of some concern. The quality of synthesis varied between speakers. Subjects varied in the degree that they believed their synthesized feedback sounded like them, and all agreed that their feedback sounded “computery”. These concerns were noted but were disregarded in the analysis. To minimize inaccuracies in formant synthesis, this study used only male talkers, whose formants were easier to shift. In addition, subjects’ input and output formants were measured and compared to ensure that formant shifting was working as expected.

5 Results

Compensation behavior varied between subjects and between formants. F1 compensation was the most consistent; every subject compensated for changes in F1. All subjects compensated for F2, though the degree of compensation varied more between subjects. For F0, two subjects (out of nine) failed to compensate at all.

Results from a typical subject are detailed below. Most subjects showed compensatory
behavior in the expected direction: that is, when their pitch was raised, they spoke with a lower pitch, and when the F1 they heard was raised, they lowered their F1 in production.

The following are the three compensation curves detailing change along a single dimension in response to a change in that dimension alone.

Figure 6: A typical subject’s response to F0 perturbation. (a) Percent compensation in F0 production for increasing F0 perturbation. Percent compensation is the difference between the formant produced during perturbed trials and the first 15 unaltered trials, divided by the amount of formant perturbation. If F0 was perturbed by 10 Hz, an F0 that was produced at 5 Hz below normal exhibits 50% compensation. (b) Absolute F0 compensation. This graph shows the F0 produced at each perturbation plateau.

Figure 7: A typical subject’s response to F1 perturbation. (a) Percent compensation in F1 production for increasing formant perturbation. Percent compensation is the difference between the F1 produced and the F1 baseline divided by the amount of F1 perturbation. If F1 was perturbed by 200 Hz, an F1 that was produced at 100 Hz below normal exhibits 50% compensation. (b) Absolute F1 compensation. This graph shows the F1 produced at each perturbation plateau.

5.1 Patterns in F0

There is some variability in the exact shape of the curve and in the maximum degree of compensation.
Figure 8: A typical subject’s response to F2 perturbation. (a) Percent compensation in F2 production for increasing formant perturbation. Percent compensation is the difference between the F2 produced and the F2 baseline divided by the amount of F2 perturbation. If F2 was perturbed by 200 Hz, an F2 that was produced at 100 Hz below normal exhibits 50% compensation. (b) Absolute F2 compensation. This graph shows the F2 produced at each perturbation plateau.

Interestingly, one subject (s07) showed the opposite pattern, at least for F0 alteration. When this subject heard his F0 raise, he spoke with a higher, not a lower F0. This phenomenon has been documented for pitch and is called *pitch following* (Burnett et al., 1998). Pitch following tends to be less common for small F0 shifts, and more common for larger F0 shifts. If subjects in this experiment followed this pattern, they would have opposed the direction of pitch shift at the beginning of the experiment, and at some critical point in the middle of the experiment, they would have started pitch following in the direction of F0 shift. This was not quite what happened in s07’s case: this subject followed the direction of F0 shift for shifts less than 5Hz, and opposed the direction of F0 shift for shifts greater than 5Hz.

5.2 Patterns in F1

All subjects compensated for changes in F1, though due to a procedural glitch, one subject’s F1 was not recorded. There were three patterns of compensation found among the remaining seven subjects:

1. Logarithmically decreasing absolute compensation. Most subjects (5 of 7) decreased their percentage compensation over the course of the experiment. These subjects compensated almost completely for small F1 shifts, but compensated less for larger F1 shifts. Their percent compensation decreased quickly for smaller amounts of formant alteration, and steadied for larger amount of formant alteration.

2. Linearly decreasing absolute compensation. Some subjects (2 of 7) decreased their percentage compensation over the course of the experiment. They also compensated more for small alterations to F1 and less for larger alterations to F1. However, their
absolute compensation decreased steadily over the entire range of formant alterations.

3. Constant absolute compensation. No subject had the same absolute compensation for every perturbation plateau.

Surprisingly, by the end of the experiment, where F1 was being altered by 250 Hz, percent compensation hit 40% for all subjects, regardless of the path they took to get there. Some started at 100% compensation and decreased to 40% by the end, and some were at 40% the entire time, but all were at about 40% for the last formant alteration plateau. While this may be a coincidence, it suggests a common compensation mechanism, to be discussed further in the next section.

5.3 Patterns in F2

All subjects also compensated for changes in F2. Compensation curves for F2 could be logarithmic, linear, or constant, just like F1. Unlike F1, however, there was not a consistent percentage that all subjects reached by the end of the experiment. Some subjects reached 60% compensation for the 250 Hz alteration to F2, while some were at 30% or even 0%. If subjects are aiming for a particular acoustic target such that it does not overlap with another acoustic target, it should be easier to compensate for F2 than F1. In English, there are more vowels along the F1 dimension than along the F2 dimension.

One subject (s03) showed following behavior for F2, something that has not been previously documented. Consistent with F0 following results, this subject had strong following behavior for small changes in F2, but weak or negligible following for larger changes in F2.
Figure 9: Change in production for increasing formant perturbation. Subject s07’s production of the /ɛ/ in ‘head’ for each of the five F0 shift plateaus. Each box represents an average of 20 trials. The thick line in the middle of each box represents the mean F0 produced during those 20 trials, and the borders of the box correspond to the 25th and 75th percentile in F0. s07 follows the direction of pitch shift for the 4Hz plateau, but opposes the direction of shift or does not compensate for larger shifts.

Figure 10: Change in production for increasing formant perturbation. Subject s03’s production of the /ɛ/ in ‘head’ for each of the five F2 shift plateaus. Each box represents an average of 20 trials. The thick line in the middle of each box represents the mean F2 produced during those 20 trials, and the borders of the box correspond to the 25th and 75th percentile in F2. Notice that s03 strongly follows the 50Hz shift in F2, but does not compensate once F2 shifts reach 150Hz.
5.4 Target regions

If we are to test the DIVA model, we have to compare subjects’ compensation to the their target regions. Compensation only “counts” for this model if formant production lies outside of what subjects would normally accept as an /ɛ/. Target regions were estimated from the control condition of the experiment. In this condition, subjects said ‘head’ 360 times while hearing their feedback resynthesized but not shifted. Target regions are surprisingly large, as shown in Figure 11:

Figure 11: Subject s02’s target region. Subject s02’s production of the /ɛ/ in ‘head’ during a control experimental session. Each point represents the vowel from one production.
When compensating for F1, subjects clearly produced vowels outside of their target region, as pictured in Figure 12.

Figure 12: Subject s02’s F1 compensation with respect to his target region. Again, this subject clearly compensates for F1. This graph shows how far the subject strays from his baseline. A convex hull has been drawn around the baseline region in black.

However, when compensating for F2, subjects largely did not produce vowels outside of their target region.

Contrary to the expectations of the DIVA model, not all points within the target region are equal. Subject s02’s F2 compensation is displayed in Figure 13. Though this subject’s /e/ productions just barely leave his target region, he is systematically decreasing the F2 he produces. He is sensitive to where his vowel is within the target region, and he is able to aim for a particular point within the target region.
Figure 13: Change in production for increasing formant perturbation. Subject s02 also clearly compensates for F2. However, it is clear that the subject does not stray nearly as far from his target region. A convex hull has been drawn around this region for clarity.
5.5 Single formant shifts induce compensation in multiple formants

Though formant shifts always involve a single formant, subjects do not seem to change only single formants in their production. At first, F1 shifts often result in exclusively F1 changes, but eventually all subjects change both F1 and F2. F0, on the other hand, does not change much in response to F1 or F2 perturbation.

Figure 14: Change in F1 production for increasing formant perturbation. Subject s02’s production of the /e/ in “head” for each of the five F1 shift plateaus. The subject’s first formant feedback is shifted by 0-250Hz. Colors correspond to the amount of formant shift: trials with no shift are in white, and trials with maximum formant shift are in black.

For F1 shifts of 50, 100, and 150 Hz (orange, yellow, and light green in Figure 14), the subject lowers F1 without substantial change in F2. The subject raises F2 as he lowers F1 for F1 shifts of 200 and 250 Hz (dark green and blue in the figure above).

When the subject’s F2 is shifted by 50, 100, or 150 Hz (light gray, medium gray, and medium dark gray in Figure 14), the subject lowers F2 without substantially changing F1. However, the subject raises F1 as he lowers F2 for F2 shifts of 200 and 250 Hz (dark gray and black in Figure 14).
Figure 15: Change in production for increasing formant perturbation. Subject s02’s production of the /e/ in “head” for each of the five F0 shift plateaus. The subject’s production of the second formant is shifted by 0-250Hz. Colors correspond to the amount of formant shift: trials with no shift are in white, and trials with maximum formant shift are in black.

6 Analysis

6.1 F1 depends on F2

Surprisingly, the compensation “curves” are, for the most part, curves. Percentage compensation differs from plateau to plateau. For an acoustic target region theory, we would expect to see no compensation for formant shifts within the target region, and increasing compensation outside the target region: the speaker’s goal is to produce a sound anywhere within that target region, and that region is increasingly far away as perturbation increases. This is not what we observe.

Instead, we see full compensation (or even overcompensation) for small formant shifts, and partial compensation for larger shifts. The degree of compensation decreases as shifts diverge from baseline. Why would this be?

Earlier, we noted that if our vowel target is a function of several formants, such as $F2-F1$, there is a trading relation between $F2$ and $F1$, and we can compensate for feedback alteration in three ways: (1) changing the $F1$ we produce, (2) changing the $F2$ we produce, or (3) changing both $F2$ and $F1$. All three options lead us to the same percept in a
parametric theory.

Does such a trading relation exist for subjects in this study? Let us consider this question with respect to F1 compensation. Recall first that all subjects showed some F1 compensation. In compensating for F1, all of these subjects changed their production of both F1 and F2. Over the course of the experiment, they lowered their production of F1 and raised their production of F2. According to a parametric theory of compensation, we would say that F1 is somehow bound up with F2 in vowel representation.

This makes a clear prediction for our subjects’ responses to F2 feedback shifts. Because we have already shown that F1 and F2 are dependent, a change in F2 should be accompanied by change in F1. Furthermore, if the function is invertible, we should expect consistency in the amount and direction of F1 compensation. That is, we should expect that decreases in F2 should be accompanied by increases in F1.

In general, this is what we find. Figure ?? shows F1 and F2 feedback shift data for two representative subjects we have already considered, s02 and s07.
Figure 16: Change in production for increasing formant perturbation. Production of the /ɛ/ in “head” by s02 and s07 for each of the five F1 and F2 shift plateaus. Colors correspond to the amount of formant shift: trials with no shift are in white, and trials with maximum formant shift are in black. In the left-hand graphs, F1 feedback is shifted from 0-250 Hz; notice the decrease in F1 production accompanied by an increase in F2 production. In the right-hand graphs, F2 feedback is shifted from 0-250 Hz; notice the decrease in F2 production accompanied by an increase in F1 production.
6.2 There are no vowel invariants

With all of this information, we are now in a position to test the status of the vowel invariants outlined at the beginning of this paper. We have seen that subjects change the way they produce F1 and F2 in response to changes in feedback in those formants. We have also seen that some formants tend to move in concert; when subjects raise F1, for example, they tend to lower F2.

The major vowel invariant theories involve formant distances measured in Bark units. Logarithmic differences correspond to arithmetic quotients, so these distances are actually ratios between F1 and F0, and between F3 and F2. Below are graphs of those vowels for subject s03, who we saw earlier has a large F1 compensations. If F1-F0 is truly a predictor of /ɛ/ quality, then this subject’s large F1 increase should be associated with a proportional increase in F0 in order to keep the F1-F0 distance the same.

This is not what we find. The graph on the left shows the formants from the vowel that the subject heard over the course of the experiment. If the subject is aiming for a particular ratio of F0, F1, F2, and F3, those quantities should be roughly constant in the signal that the subject hears. In the graph, the F3-F2 Bark distance does not change substantially, but there is a clear increase in the distance between F1 and F0 as the F1 feedback shift increases. The vowel that the subject hears does not have a constant F1/F0 ratio. The right hand graph lends further support to this view. This graph shows the vowels that the subject produced over the course of the experiment. Though once again, the F3-F2 distance remains roughly constant, the F1-F0 distance decreases over the course of the experiment.
Figure 17: Change in production for increasing formant perturbation. Subject s03’s production of the /ɛ/ in ‘head’ for each of the five F0 shift plateaus. On the left, F1 is plotted against F2. On the right, F1 is plotted against F0. The subject’s production of the second formant is shifted by 0-250Hz. Colors correspond to the amount of formant shift: trials with no shift are in red, and trials with maximum formant shift are in blue.

F1 and F0 do not move in concert, and there are no invariants in either perception or production for this subject. The correlation between them is 0.09. No other subject had an F1 – F0 correlation exceeding 0.22.

6.3 Choosing between theories

6.3.1 Trouble for DIVA

These results are not consistent with the DIVA model’s conception of vowel target regions. That model predicts that increasing feedback alteration along the same dimension should cause an increase in the percentage of compensation. Subjects in this experiment show the opposite behavior: their compensation is a larger percentage of feedback perturbation for smaller perturbations, and as perturbation increases, subjects compensate proportionally less.

The F2 compensation results are also a problem for acoustic target regions. Though subjects clearly compensate for F2 feedback shifts – their F2 productions oppose the direction of the shift – the magnitude of their compensation is small enough that they never leave their target region. Clearly, subjects are able to aim for particular points within the target region. They are not all identical.
Alternatively, these results can be explained by a model that assumes an acoustic target and depends on an internal model to map from acoustics to articulation. As Guenther (1998) observes, the relationship between articulation and acoustics is nonlinear and complex. It is impossible to predict one from the other without extensive training. This observation leads to two explanations for complete compensation for small perturbations. First, a predictable, linear mapping between acoustics and articulation is much more likely for small changes than for large ones. With regard to this experiment, if we are using an internal model to predict the necessary changes in articulation to compensate for the change in acoustics that we just heard, our prediction is more likely to be accurate at small distances from our baseline than for large distances. Second, in the course of everyday speech, we have had a good deal of practice correcting for small deviations from baseline, but we have very little practice correcting for large deviations from baseline. This means that our model is quite accurate for small deviations from baseline, but poor for large deviations from baseline. Because the relationship between acoustics and articulation is irregular, and because we lack experience in making the particular corrections required in this experiment, we would expect compensation to be very good at small perturbations, and increasingly poor compensation for larger perturbations.

Overall, this theory would predict that as the formant shift increased, degree of compensation would decrease, but it would also predict undirected compensation. If the articulatory system is really taking wild guesses, the degree of compensation should alternate unsystematically between very good and very poor. In fact we see a fairly steady decrease in compensation that would pose some trouble for this theory.

One might argue that guesses are not exactly wild: in the DIVA model, the relationship between acoustics and articulation is learned by infants as they babble. They move their articulators, produce a sound and take note of the relationship between them. Then they shift their articulator position slightly and vocalize again, comparing their prediction for the change in sound to the sound that was actually produced. It is theoretically possible to sample quite a lot of the vowel space over the course of a year and thereby build a reasonably complete model. Perhaps, then, infants have a very good sense of how to compensate fully for any missed target. If they do, why don’t adults have access to this mapping?

Furthermore, the morphology of an adult vocal tract is very different from a child vocal tract. Even the relative sizes of the pharyngeal and oral cavities differ (Menard, Schwartz, Boë, Kandel, & Vallée, 2001). Thus not only is the mapping between acoustics and articulation difficult to learn in the first place, but it also changes over time. As the infant develops into a child and then an adult, the vocal tract’s shape, size, and length develop as well. If an adult retains the acoustic-articulatory mappings that s/he learned while babbling as an infant, those mappings will be worthless when applied to an adult vocal tract.
In any case, choosing a target region theory means that there must be a compensation threshold: below some critical amount of formant shift, subjects should not compensate at all. The literature is mixed on the presence of a compensation threshold. However, I have been able to show compensation in single subjects for 30 Hz formant shifts, well below the putative threshold value (Purcell & Munhall, 2006a).

6.3.2 Trouble for bone conduction

It is also possible that incomplete compensation is due to bone conduction. While the volume in the headphones was set quite loud in order to mask as much bone conduction as possible, there was no explicit masking noise. It is therefore possible that subjects took acoustic signals derived from bone conduction into account when deciding how much to compensate. However, this bone conduction effect is systematic error: it should have been equally present at all formant shifts, and it should have had the same effect on compensation at all formant shifts. Therefore it cannot account for differing degrees of compensation across formant shifts.

6.3.3 Trouble for parametrized acoustic theories

The symmetry in F1 and F2 compensation is evidence for an invertible relationship between those two formants. This type of relationship could explain our partial compensation results. Perhaps subjects are in fact compensating fully for large F1 shifts, and their compensation is looks partial because it is split between F1 and F2. However, two pieces of evidence speak against this interpretation. First, there is significant intersubject variability in the F1-F2 relationship. It is not necessary that all speakers of the same language store the same formant dependencies, but if we adopt this theory, we bear the burden of explaining how different speakers came to different conclusions about these two formants based on similar evidence. Second, the relative amount of F1 and F2 change is different for the each of the five formant shift plateaus. This suggests that if there is any relationship at all between these two formants, it is difficult to describe.

Fortunately, even without a specific closed form function, we can test for invertibility. This is a useful exercise because if the test fails, then the function relating F1 and F2 is not invertible, eliminating all linear relationships, and, in fact, all of the normalization relationships posited by the research groups mentioned earlier in this paper. A description of this experiment is included in the Further Research section.

We have seen that compensation for F0, F1, and F2 all exhibit full compensation for small perturbations, and partial compensation for large perturbations. I have explained these results by appealing to the complex and nonlinear mapping between acoustics and
articulation. As the sound that is heard increases in distance from the acoustic target, finding the correct compensatory articulation is increasingly difficult, resulting in poor compensation for large perturbations from baseline.

6.4 A solution

We are still left with the problem of explaining partial compensation. A plausible explanation lies in the conflicting feedback received from sensorimotor and auditory sources. During the experiment, subjects received abnormal acoustic feedback, which pushes the articulatory system to compensate, but normal somatosensory feedback, which pushes to articulatory system to remain in its current configuration. If the compensation mechanism takes both types of feedback into account, one would expect incomplete feedback due to the conflicting motor commands. There is evidence that some feedback is sensorimotor; we are able to compensate for changes in motor feedback with unperturbed acoustic feedback (Nasir & Ostry, 2006).

Due to noise in the environment and in sensory receptors, small deviations from normal acoustic feedback coupled with normal somatosensory feedback are probably common. Motor commands are unlikely to be changed on account of small mismatches like these. However, very large deviations from normal acoustic feedback with normal somatosensory feedback (as with the 250Hz shift in this study) are very uncommon, and there is significant conflict between motor commands from the two sources. In this situation, the amount of compensation is determined by the amount that a talker is willing to trust acoustic feedback over parallel somatosensory feedback. This amount is apparently proportional to the amount of mismatch between the two signals.

6.5 Reasons for individual variation

The significant variation between individuals in this study is typical of compensation experiments. Any theory that wishes to explain these results must account for this variation. There are at least two possibilities.

If we are to maintain a modified acoustic target theory, one place for variation is in the size of the acoustic target region. There is evidence to suggest that talkers with larger degrees of compensation are also sensitive to smaller changes in perceived speech (Perkell, 2007). In this study, I measured perceptual acuity by asking subjects to make discrimination judgments for tokens of their own voice. Ten tokens were resynthesized with minimally different F0, ranging from 5 Hz above their baseline to 5 Hz below their baseline. There was only a loose correlation between degree of compensation and F0 discrimination ability.
### Table 1: Discrim Threshold and Compensation ability

<table>
<thead>
<tr>
<th>Subj</th>
<th>Discrim Threshold</th>
<th>Compensation ability</th>
</tr>
</thead>
<tbody>
<tr>
<td>s03</td>
<td>2 Hz</td>
<td>Fair</td>
</tr>
<tr>
<td>s04</td>
<td>3 Hz</td>
<td>Fair</td>
</tr>
<tr>
<td>s01</td>
<td>3 Hz</td>
<td>Poor</td>
</tr>
<tr>
<td>s07</td>
<td>4 Hz</td>
<td>Good</td>
</tr>
<tr>
<td>s08</td>
<td>4 Hz</td>
<td>Good</td>
</tr>
<tr>
<td>s02</td>
<td>5 Hz</td>
<td>Good</td>
</tr>
<tr>
<td>s05</td>
<td>5 Hz</td>
<td>Poor</td>
</tr>
<tr>
<td>s06</td>
<td>7 Hz</td>
<td>Fair</td>
</tr>
</tbody>
</table>

The conflicting sensorimotor/auditory feedback theory makes a different prediction. In this theory, the degree of compensation is proportional to the amount of credence that the subject gives to acoustic feedback. A subject that trusts exclusively acoustic feedback and ignores somatosensory feedback would compensate completely, and a subject to trusts exclusively somatosensory feedback would not compensate at all. The degree of credence given to either sensorimotor or auditory feedback would need to be learned, and thus experience-sensitive. According to this theory, the degree of compensation should indicate the amount of authority that auditory feedback has over somatosensory feedback.

This theory offers an appealing solution to the mystery of consistent F1 compensation for the largest (250Hz) formant shift. Because people learn to trust their acoustic feedback through experience, they differ in the amount they will trust acoustic feedback for more small formant shifts. Such shifts are more common in their experience. But they are more consistent in the amount they trust acoustic feedback for very large formant shifts. This suggests a default amount of trust conferred to acoustic feedback for unknown signals – 40%.

### 6.6 Vowel generalization

Adaptation experiments that ask subjects to reach for an object find that adaptation generalizes from reaching to walking (Morton & Bastian, 2004). So far, however, the results in the speech literature are mixed: there is one instance where adaptation in the production of one vowel, /ɛ/ generalizes to adjacent vowels (Houde & Jordan, 1998), and one where adaptation to the production of /ɛ/ does not generalize to production of other vowels (Pile et al., 2007).

This experiment sought to replicate those results. Vowel spaces from each subject were measured before and after they participated in this experiment. In some cases, we can see evidence of compensation generalization. Subjects adapt their compensation behavior to other vowels. Due to the small number of tokens, this observation is best considered a trend.
Consider subject s09’s results below. His vowel space was recorded before and after his F1 feedback was shifted upwards by 250Hz. After this manipulation, the subject lowers F1 and raises F2 for all mid and low vowels, whether they are front or back vowels: this includes /ɛ/, /æ/, /ɑ/., and /o/. High vowels are not affected.

Figure 18: Subject s09’s generalization of /ɛ/ adaptation. Production of ‘heed’, ‘hid’, ‘head’, ‘had’, ‘hod’, ‘hoed’, and ‘who’d’ by subject s09. ASCII renderings of the relevant vowels are written on the vowel chart at their formant frequencies. ‘Before’ trials are in blue. ‘After’ trials are in red. Recordings were taken just before and just after the experiment.

For this subject, the motor plan affected is more abstract than a single exemplar, lexical item, or even vowel category. Future subjects will be tested for generalization in enough detail to warrant statistical testing.

6.7 Future work

This project is intended to be the first step in a series of experiments. Several concerns that have presented themselves in the course of writing this paper can be addressed with a few simple experiments.

6.7.1 Generalization of speech adaptation

Some subjects who have participated in this experiment exhibit a tendency toward generalization of adaptation behavior. If adaptation truly generalizes to other vowels, then
some our motor plans are dependent not solely on individual exemplars, but on contrasts between other phonemes and lexical items we have learned.

Future subjects will be asked to produce 15-20 tokens of each of the seven /hVd/ words mentioned earlier in order to get a more reliable vowel space. The amount of formant variation in the control version of this experiment, in which subjects were simply asked to produce ‘head’ 360 times, is evidence that the target region for vowels is quite large.

6.7.2 Compensation threshold

Purcell & Munhall found that subjects did not compensate for formant shifts below approximately 75 Hz (Purcell & Munhall, 2006a). The large compensation at 50 Hz formant shift plateaus is not consistent with this finding, however, it is still possible that there is a compensation threshold below 50 Hz. To answer this question, I am in the process of adding an experimental condition. Rather than shifting formants by 50, 100, 150, 200, and 250 Hz, I will use a 2-plateau experiment, one below Purcell & Munhall’s threshold at 30 Hz, and one above their threshold at 90 Hz. If there is indeed a compensation threshold around 75 Hz, we should expect no compensation at the 30 Hz plateau and measurable compensation at the 90 Hz plateau.

I also plan to run a set of subjects with smaller formant shift increments. In this paper, I have assumed that differences in degree of compensation were due to the amount of formant shift and perceptual factors. However, it is possible that subjects’ sensitivity decreases over the course of the experiment: at the beginning, they are willing compensators, but by the 200th trial they are less willing to compensate no matter how large the formant shift. This possibility is unlikely, but can be dismissed by adding a smaller step size condition. In the small step size condition, subjects will experience a 50 Hz formant shift at the 200th trial, while in the large step size condition they will experience a 150 Hz formant shift at that trial.

6.7.3 Invertibility

Results indicate a relationship between F1 and F2. If this relationship is invertible, as would be the case for many of the functions we might suggest, then any vowel produced during this experiment should be produced in two ways:

Happening upon equal compensation by chance is, frankly, unlikely. But we can test for invertibility in the following way. Consider a subject whose baseline vowel V is at point $P_0$. Due to formant shift, this subject is currently producing a vowel at point $P_1$. Starting from $P_1$, let us shift F1 by 40 Hz. This is represented by the top leg of the parallelogram. The subject compensates by producing $P_2$. If the compensation function is invertible, we
should be able to undo this change with the opposite compensation. This is the bottom leg of the parallelogram. Starting from vowel \( P_2 \), we can shift F1 by -40 Hz. To compensate, we would expect the subject to produce the vowel at \( P_1 \).

![Figure 19: Illustration of invertibility experiment](image)

In some sense, this is what we are doing when we drop the formant shift back to normal at the end of the experiment. The differences are that when returning to baseline (1) formant shifts are large, and therefore probably more difficult to compensate for on short order, and (2) there are confounding motor factors that will cause subjects to tend toward their baseline production. Thus, in spite of subjects’ ability to return to normal speech, it is not obvious that subjects can return to any particular point in acoustic vowel space.

The following linguistic considerations are also important avenues for future research.

### 6.7.4 Native language

It is likely that a speaker is disposed by his/her native language to care more about some linguistic variables than others. For example, Mandarin speakers may respond more strongly to manipulation of F0, and languages with larger or smaller vowel inventories may be more or less sensitive to formant changes. Sensitivity to formant feedback is also likely to be linked to the importance of that formant as a cue, which is language specific. For example, front and back vowels in English are separated not only by frontness, correlated with F2, but also by roundedness, which is linked to F3. Thus F2 is less informative in English than it would be in a language with unrounded back vowels. Speakers of such a language are likely to compensate more for F2 shifts. Additionally, lexical neighborhoods for any given word will be a different size for English and non-English speakers. There are several interesting manipulations possible with non-English lexical items and non-native speakers which might further inform these results.
6.7.5 Size of lexical neighborhood

It is also likely that compensation is greatest when the subject understands him/herself to have said a different word. Thus compensation should be greater in dense lexical neighborhoods (heed, hid, head, had) than in sparse lexical neighborhoods (teeb, tib, teb, tob). This is an important possibility deserving a follow-up experiment.

6.7.6 Natural communicative context

However, it is difficult to argue that there is lexical access occurring with each instance of ‘head’. Consequently it is worth considering how subjects would react to compensation in a communicative situation. In the current experiment, subjects reported that ‘head’ ceased to sound like a word to them at all. Subjects might keep a tighter rein on their vowels if the word they were saying mattered. One paper has taken a step in this direction. Chen et al. asked subjects to repeat either a vowel or a sentence while their pitch was suddenly perturbed. Overall compensation for speech was greater than that for a vowel alone (Chen, Liu, Xu, & Larson, 2007). To continue along this theme, it is worth investigating how people compensate when trying to ask or answer a narrative, tell a story, or perhaps just hear their interlocutor’s speech being altered.

7 Conclusions

Speech adaptation is fundamentally different from other sorts of sensorimotor adaptation because the goal is not a concrete object in space. If we accept that compensation is always complete, as it is in reaching experiments, then the vowel that each subject hears after compensating for altered speech feedback must be that speaker’s goal. The task of the researcher is to understand what that goal is, since it is never precisely equivalent to the baseline vowel. In the reaching literature, no one ever fails to reach the target after practice.

In answer to the questions posed at the beginning of this paper, we find that:

1. Subjects compensate in the expected direction. However, compensation is partial and, in the case of F1 and F2, affects unshifted formants. A shift in F1 feedback causes an opposing change in F1 production, and a parallel change in F2 production. That is, in response to an increase in F1 feedback, subjects lower their F1 and raise their F2. Alteration of F2 causes an opposing change in F2 production and a parallel change in F1 production. That is, in response to an increase in F2 feedback, subjects lower their F2 and raise their F1. This trend is not as strong as for F1. There is not a monotonic relationship between F0 and F1 or F2.
Intriguingly, subjects have fine control over their productions within their baseline region when compensating for F2. This is a problem for the DIVA model, or any model which depends on acoustic targets.

It will always be possible to come up with a function relating speakers’ productions of F1 and F2. A number of research groups have put forward a variety of such functions to equate vowels between speakers. Even without a specific closed form function, we can test for invertibility. This is a useful exercise because if the test fails, then the function relating F1 and F2 is not invertible, eliminating all linear relationships, and many other functions one might posit to relate F1 and F2.

2. Proposed vowel invariants are poor descriptors of the vowels produced in this experiment. In particular, F1 and F0 covariance.

3. Though people are able to accept a wide range of acoustic signals as /ɪ/, as indicated by the wide range of baseline vowels. However, complete compensation for small formant shifts indicates that we have a target that is much smaller than this baseline region.

4. For small amounts of perturbation, subjects compensate fully even with normal motor feedback. For large amounts of perturbation, subjects will not compensate fully on account of normal motor feedback, which is in conflict with the abnormal acoustic feedback. Motor feedback seems to be thresholded; subjects do not compensate at all below some minimum amount of perturbation.

These results suggest that internal models for both motor and acoustic productions are at work during speech. Individual variation provides evidence that the model parameters are learned rather than innate.

References


conditions of altered auditory feedback: Does adaptation of one vowel generalize to other vowels?


