# CHANGES IN CONSONANT PERCEPTION DRIVEN BY ADAPTATION OF VOWEL PRODUCTION TO ALTERED AUDITORY FEEDBACK

Will Schuerman[1], Srikantan Nagarajan[2], & John Houde[2]

[1]Max Planck Institute for Psycholinguistics; [2]University of California, San Francisco
Will.Schuerman@mpi.nl

## ABSTRACT

Adaptation to altered auditory feedback has been shown to induce subsequent shifts in perception. However, it is uncertain whether these perceptual changes may generalize to other speech sounds. In this experiment, we tested whether exposing the production of a vowel to altered auditory feedback affects perceptual categorization of a consonant distinction. In two sessions, participants produced CVC words containing the vowel /i/, while intermittently categorizing stimuli drawn from a continuum between "see" and "she." In the first session feedback was unaltered, while in the second session the formants of the vowel were shifted 20% towards /u/. Adaptation to the altered vowel was found to reduce the proportion of perceived /S/ stimuli. We suggest that this reflects an alteration to the sensorimotor mapping that is shared between vowels and consonants.

**Keywords:** Altered auditory feedback, compensation for coarticulation, speech perception, phonetics, sensorimotor mappings

## 1. INTRODUCTION

There is no clear one-to-one mapping between the acoustic signature of a speech sound and how it will be perceived in a given context. For example, [5] demonstrated that lexicality can influence the perception of speech sounds; an ambiguous sound between /t/ and /d/ is more likely to be perceived as a /t/ in the context of "?-ask", while the same sound is more likely to be perceived as /d/ in the context of "?-esk." Another factor which has been shown to influence categorical perception is vowel context. The same ambiguous fricative is more likely to be classified as /s/ when followed by an unrounded front vowel, but as /S/ when followed by a rounded back vowel [4]. These sorts of "compensation for coarticulation" effects have sparked numerous debates about the nature of the production-perception interface.

One technique which has broken ground in studying the production-perception interface is altered auditory feedback. This technique modulates the acoustic properties of a speaker's voice (e.g. pitch or formant frequencies) to create a mismatch between the articulatory gesture used to generate an intended acoustic output and what the speaker actually hears [6]. This leads to adaptation, in which the speaker modulates their articulations in order to accurately produce an intended sound. Adaptation to an altered fricative has been found to cause subsequent changes in fricative categorization during perception tasks [8]. Interestingly, these changes in perception occurred when the participants had to produce the target words but not during passive listening to shifted recordings. This suggests that changes in perception may be driven by a shift in the sensorimotor-mappings between a given articulation and its acoustic output.

In this study, we investigated whether adaptation to an altered vowel may affect the perception of contextually dependent speech sounds. Specifically, we examined whether altering the sensorimotor mapping for the vowel /i/ may affect the categorization of a preceding ambiguous fricative. We tested participants' identification of ambiguous fricative stimuli after production of fricativeless words containing the vowel /i/ under unaltered and altered feedback, utilizing a within-subjects design. In the altered feedback condition, the formants of the speaker's vowels were shifted in the vowel space towards the speaker's average productions of /u/. The speaker would therefore have to "hyper-articulate," make an articulation that is more fronted or more unrounded than normal, in order to counter these effects and produce a clear /i/.

Based on previous research [8], we hypothesized that under the condition of unaltered feedback, repeated production of the vowel /i/ may lead to fatigue or satiation, in effect reducing or degrading the perceptual space for the vowel /i/. We expected that this could lead to a subsequent decrease in the number of "she" responses when categorizing stimuli in Identification blocks. With regard to altered feedback, we tested two alternative hypotheses: if

a speaker adapts to the altered feedback by hyper-articulating, this suggests that the original acoustic signal is now mapped to a more fronted articulation. If this shift does not generalize to consonants, then we would expect an increase in the proportion of stimuli perceived as "she," as vowel now suggests a more fronted place of articulation while the fricative remains constant. However, if exposure to the altered feedback generalizes to the consonant, then a place of articulation which once had produced /S/ would now map to a more /s/ like sound, reducing the proportion of stimuli perceived as "she." Crucially, under altered feedback any changes in perception should correspond to the shift in auditory feedback, whereas with unaltered feedback any changes in perception should gradually accumulate over the course of the session.

## 2. METHODS AND MATERIALS

### 2.1. Participants

Twenty-seven participants (11M, 16F, mean age: 29.04, range: 21-36) native speakers of English took part in the study. None of the participants reported any hearing difficulties or speech impairments. Participants provided written consent and were paid for their participation.

### 2.2. Materials

A list of thirteen English words was created for the Production task ('peep', 'beep', 'deep', 'keep', 'peat', 'beet', 'bead', 'deed', 'keyed', 'peak', 'beak', 'teak', 'geek'). Each word consisted of a stop consonant, followed by the vowel /i/, followed by another stop consonant. This forced the participants to actively pay attention and read each word, while minimizing segmental variation.

Stimuli for the Identification task were drawn from a 100-step continuum between "see" and "she." To construct these stimuli, a female native speaker of English was recorded producing the word "see" in the sentential context "say the word see." The most natural sounding elicitation was selected from three recordings. Using Praat [1], we extracted the prosodic contour of this recording into a format readable by Mbrola, a text-based diphone synthesis program[3]. This enabled us to create synthesized versions of the words "see" and "she" with identical pitch and duration. Stimuli were normalized by root-mean-squared amplitude, then additive synthesis was used to generate the 100-step continuum.

### 2.3. Design

#### *2.3.1. Procedure*

The experiment consisted of a pretest, an Unaltered Feedback (UF) session, and an Altered Feedback (AF) session. The UF and AF sessions were separated by a minimum of two weeks. UF always preceded AF in order to avoid any possible after-effects of the altered auditory feedback.

Both UF and AF sessions followed the same format. There were seven Identification blocks interleaved by six Production blocks. In Production blocks, each word was presented twice. Order was pseudo-randomized to ensure that no words were presented twice in a row and that all words had been presented once before repetition. In the AF session, feedback was unaltered in block one. Beginning in block two, the values of the first and second formants of the participant's voice were shifted gradually over 26 trials to the maximum of 20% perturbation toward the participant's average values for /u/. Maximum perturbation continued in blocks three and four, then was ramped down back to baseline in block five. Block six consisted of unaltered feedback.

In Identification blocks, stimuli were pseudo-randomized into sets of four trials, consisting of two ambiguous ($\pm$ 1, 3, 5, 7, 9 steps from boundary) and two clear ($\pm$ 11, 13, 15, 17, and endpoints) stimuli above and below the participant's identification boundary. Each stimulus step was presented twice per block for a total of 40 trials, with all steps presented once before repetition began.

#### *2.3.2. Pretest*

A modified adaptive staircase procedure was used to determine each participant's identification boundary. The procedure began with four clear practice stimuli drawn from the endpoints of the continuum. Step size between trials was initially set at 100, and began with step 0 (clear "see"). Step size diminished by 50% with each reversal. In pilot tests, repeated exposure to ambiguous stimuli tended to disorient participants, therefore each real trial was interleaved with a randomly selected stimulus chosen from one of the two endpoints. Procedure terminated after 24 trials.

#### *2.3.3. Production Task*

In Production blocks, participants were instructed to read visually presented words aloud in a clear, normal voice. Words were presented in 30 point white

font on a black background. Each word was presented for a duration of 1.742 sec, with a minimum inter-stimulus interval of 500ms.

*2.3.4. Identification Task*

In Identification blocks, participants listened to the synthesized stimuli and reported via keyboard whether they heard "see" (button 1) or "she" (button 2). Valid responses were recorded only after presentation of the sound file was completely finished. This ensured that participants listened to both the fricative noise as well as the following vowel. Early responses elicited a pause in the experiment, during which the researcher verbally instructed the participants to respond more slowly.

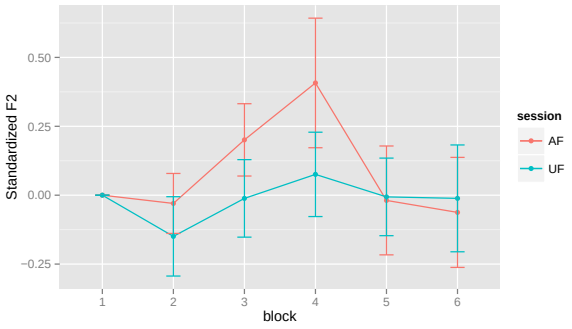
## 3. RESULTS

### 3.1. Production

Due to software failure, two participants (22 and 16) failed to complete Production block six and Identification blocks five and six in the UF session, though their data was included as these blocks followed the end of perturbation in the AF session. For the Production task, vowel frequencies were measured at the closest time-point to the midpoint of the vowel with good formant tracking. Trials in which formant tracking failed were discarded. Trials in which vowel frequency was greater or less than three standard deviations from the participant's average were also discarded, leaving 7051 trials for analysis.The average value of the second formant (F2) was calculated. This measure correlates with the degree to which a vowel is fronted in the vowel space. Due to large differences in frequency values between participants, average F2 values were first centered with respect to the average of the first production block, which constituted the Production baseline. These centered scores were then standardized by participant within each session with respect to the baseline by dividing by the standard deviation of the baseline block. This enabled us to compare changes in production over the course of each session while accounting for differences in vocal tract dimensions between participants.

We analyzed F2 in each block using two-tailed, one-sample t-tests corrected for multiple comparisons. No test in either session revealed a significant deviation from zero. However, visual inspection of the data suggested that participants did in fact adapt to the altered feedback (Fig. 1). In the UF session, F2 appeared to drift slightly below baseline, and then gradually return over the remaining blocks. In

**Figure 1:** Averaged F2 (standardized by participant and session), over the course of the experiment. Bars indicated standard error.

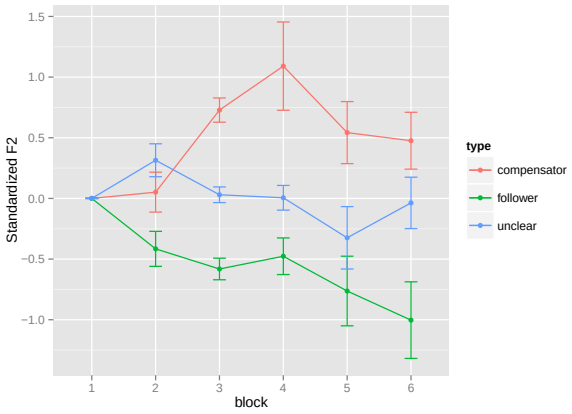



the AF session, F2 increases sharply in the third and fourth blocks before dropping below baseline in the fifth and sixth blocks. This increase in F2 almost exactly matches the contour of the AF protocol. In the individual results, in numerous cases it was found that participants had a tendency to decrease F2 in the second block, then increase F2 in subsequent blocks. To examine this further, we examined production in the AF session by separating participants into three groups (Fig. 2) based upon whether average F2 in blocks three and four was greater than baseline ("compensators", n = 12), less than baseline ("followers", n= 7), or greater in one of the two blocks and lesser in another ("unclear", n = 5). Average F2 for compensators followed the contour of the AF protocol. For followers, we see that F2 decreases but levels out in blocks 3 and 4, suggesting that while the participants did not increase F2 above baseline they appeared to compensate to the auditory feedback by returning towards baseline in these blocks. This may account for why as a group the production results do not differ significantly from zero.

### 3.2. Identification

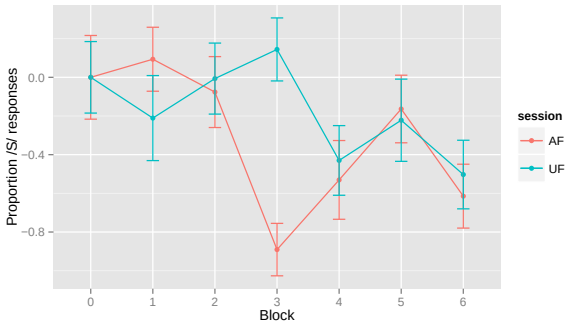
In each Identification block, the proportion of "she" responses was recorded. This proportion was first standardized within participants, within sessions, across all seven blocks. The standardized scores were then centered with respect to the first identification block, which served as a baseline for each participant (Fig. 2). The centered, standardized scores were then analyzed with two-tailed, one-sample t-tests. In order to account for multiple sampling across the seven blocks, we utilized Bonferroni correction to set a statistically significant p-value of 0.0071. Only in block three of the AF session was the proportion of "she" responses found to differ sig-

**Figure 2:** Response groups in the AF session, defined by whether average F2 in blocks 3 and 4 were greater ("compensators", n=12), lesser ("followers", n=7), or mixed ("unclear", n=5) with respect to baseline. Bars indicate standard error.




nificantly from baseline (t(23)=-3.22, $p < 0.004$). An average z-score of -0.89 indicates that participants reported less "she" responses in this block compared to baseline.

**Figure 3:** Averaged proportion /S/ responses (standardized by participant and session), over the course of the experiment. Each Identification block corresponds to a preceding Production block, i.e. Identification block 4 immediately followed Production block 4. Bars indicated standard error.




## 4. DISCUSSION

In line with our initial expectations, we found a general decrease in the proportion of stimuli identified as /S/ in the UF session. However, this decrease did not differ significantly from baseline in any of the blocks. Consistent with fronting generalizing to consonants, in the AF session we also found a *decrease* in the proportion of stimuli identified as /S/. This decrease significantly differed from baseline in the third Identification block, which was the first Identification block after 26 trials of maximum perturbation. Previous and subsequent blocks did not differ from baseline, suggesting that this perceptual shift was induced by exposure to the altered feedback.

In the AF session, participants heard themselves producing altered versions of the vowel /i/, followed by presentation of the synthesized stimuli in which the vowel quality remained constant. From the group results, we observed that stimuli which were once ambiguous were more likely to be categorized as clear /s/ after exposure to this altered feedback. Adapting to the altered feedback maps the same value for the second formant of the vowel to a greater degree of fronting than normal. During AF, the participant learns to remap acoustic outcomes to articulations in order to produce an intended output, and the results suggest that this remapping carries over during perception. Thus, when presented with the synthesized stimuli during the Identification task, the participant is now faced with the question of what amount of fronting is indicated by the acoustics of this signal. If only the sensorimotor for mapping for /i/ has been altered, such that the same vowel acoustics are indicative of a more fronted tongue position, then this should lead us to expect an increase in /S/ responses, which we did not find. However, in the articulation of both /i/ and /s/, the tongue tip and blade occupy a fronted position in the oral cavity. It is possible that due to this articulatory similarity, the quality of “frontedness” applies to both the consonant and the vowel. Therefore, when presented with the synthesized stimuli, the same acoustics for both /i/ and /s/ now correspond to a more fronted position in articulatory-acoustic space, which would lead to a decrease in the number of stimuli perceived as /S/.

Traditionally, consonants have been assigned place of articulation features such as labiodental or alveolar, while vowels are differentiated as high/mid/low, or front/back [2]. However, this result suggests that vowels and consonants that are articulated in a similar manner may share mapping features, such as relative frontedness in the oral cavity. Recent experiments with vowels have found that perceptual changes driven by adaptation to altered auditory feedback only occur when the compensatory movements occupy the same articulatory space utilized to produce the perceived stimuli [7]. Our results align with this finding, and extend it to suggest that the mappings between articulation and sound space may generalize across manner of articulation.

## 5. REFERENCES


[1] Boersma, P. 2011. Praat: Doing Phonetics by Computer.

[2] Chomsky, N., Halle, M. 1968. *Sound pattern of English*. Harper & Row.

[3] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., van der Vrecken, O. 1996. The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96* 3.

[4] Fujisaki, H., Kunisaki, O. 1978. Analysis, recognition, and perception of voiceless fricative consonants in Japanese. *IEEE Transactions (ASSP)* 26, 21–27.

[5] Ganong, W. F. 1980. Phonetic categorization in auditory word perception. *Journal of experimental psychology. Human perception and performance* 6, 110–125.

[6] Houde, J. F., Jordan, M. I. 1998. Sensorimotor adaptation in speech production. *Science* 279(5354), 1213–1216.

[7] Lametti, D. R., Rochet-Capellan, a., Neufeld, E., Shiller, D. M., Ostry, D. J. 2014. Plasticity in the Human Speech Motor System Drives Changes in Speech Perception. *Journal of Neuroscience* 34, 10339–10346.

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