# Partial effects of perceptual compensation need not be auditorily driven

Gregory Finley
April 7, 2012

## Introduction

A major question in the study of speech perception is the mechanism by which listeners process a stream of acoustic information and convert this varied signal into phonemes, words, and even larger linguistic units. Several theoretical approaches have been made to address the relationship between signal and understanding, and many studies, including this one, have been designed to test the predictions of these frameworks in novel contexts. The most broad division is between general auditory approaches (Diehl, Holt, & Lotto 2004), according to which speech perception is explicable entirely through auditory perception, and approaches for which speech perception requires the decoding of articulatory gestures responsible for speech sounds; these latter approaches most notably include Motor Theory and Direct Realist Theory (Liberman *et al.* 1967; Fowler 1986).

Many studies have aimed to evaluate these approaches by testing the predictions that each would make for speech perception in different cases. Beginning with Mann's original 1980 study, one test condition that has received considerable attention in evaluating these hypotheses is compensation for coarticulation (or perceptual compensation). Several cases have been demonstrated in which gesture recovery (henceforth GR) appears to underlie this phenomenon (e.g., Fowler *et al.* 2000, Johnson 2011, Viswanathan *et al.* 2010). Evidence also exists in the general speech perception literature that the perception of speech is at some level linked to the parsing of articulatory gestures (see Möttönen & Watkins 2009 for neurological data). However, a number of studies have also been conducted showing compensation for coarticulation in cases where GR would not be possible: by speakers without the context segments in their phonetic inventory (Mann 1986), by prelinguistic infants (Fowler et al. 1990), and even by nonhuman listeners (Japanese quail, as studied in Lotto et al. 1997). Credit for compensation in these cases is often attributed to the phenomenon of spectral contrast, in which listeners are desensitized to frequencies similar to those of an immediately preceding stimulus (Lotto & Kluender 1998, Holt & Lotto 2002). Other acoustic context phenomena have also been shown to apply, including cases in which nonspeech context following speech has been shown to affect that speech's categorization (Wade & Holt 2005).

Given the existence of evidence for both GR and general auditory perception, a natural question is whether they contribute additively to speech perception. If so, then we would expect effects of both sources to be evident in

processes such as perceptual compensation. Several researchers have proposed that the effects of compensation owing to each mechanism can be measured, and that differences in magnitude of compensation are meaningful in identifying the source; furthermore, different sources could contribute additively to the overall effect, and it may be possible to isolate them in certain conditions (Johnson 2011, Holt & Lotto 2002, Mitterer 2006).

With the goal in mind of teasing apart factors underlying perceptual compensation, I designed an experiment to measure compensation for nonspeech sounds. Though not identifiable as speech, these sounds contained the acoustic information necessary to determine their articulatory features that would cause coarticulation. I hypothesized that in this nonspeech condition, compensation would be present but not as strong as in a speech condition, owing presumably to a failure of speech-specific processes such as GR to apply. This intuition follows from considering both motor and auditory perceptual processes as necessary components in the processing of acoustic input into speech.

**Experiment 1**

The coarticulation condition tested was anticipatory lip rounding on English /s/ before round vowels. Lip rounding on /s/ lowers its spectral mean (or centroid frequency), making it perceptually closer to /ʃ/. Though English /s/ and /ʃ/ differ in a number of acoustic dimensions, the most salient of these is in spectral mean (Li et al. 2009). When before rounded vowels, then, listeners expect to hear lower energy for /s/ than before unrounded vowels; hearing an ambiguous sound somewhere between the two, they would be more likely to perceive it as /s/ before a rounded vowel than unrounded, attributing the lower-than-usual centroid to rounding coarticulation.

Because all front vowels in English are unrounded and all back non-low vowels rounded, the roundedness of any given non-low English vowel can effectively be determined by its backness, which is measurable in its F2. Furthermore, because rounding lowers F2, a very high F2 (such as that of [i]) is practically unattainable if the lips are rounded, and a very low F2 ([u] or [o]) unattainable without rounding. Therefore, extreme values of F2 should alone carry enough information regarding the status of rounding for an English speaker.

With this in mind, I designed several sets of experimental stimuli that contained F2 information but no other formants. Consequently, these stimuli sound incomplete and are not identifiable as human speech. Four distinct conditions were tested: one in which the vowels were entirely speechlike; two in which the vowel contained only the second speech formant but otherwise featured a harmonic profile similar to vocal fold vibration; and one in which vowels were replaced only by a sine-wave tone at F2. The prediction was that compensation for coarticulation would be weaker in cases where GR is not possible—all but the speechlike case. However, given that the relevant acoustic cues are present and audible, we would expect to see some effect if GR and general auditory processes are indeed both responsible for compensation.

## Method

*Participants*

There were 20 college-age participants in this experiment. All participants reported English as a native language and did not report any history of hearing or language disorders. All participants performed the same experimental task.

*Stimuli*

I tested subjects on four sets of stimuli. Each set consisted of 18 CV monosyllables, which were constructed by concatenating nine different onsets and two different vowels in all possible combinations. The onsets were [s], [ʃ], and seven other fricatives along a continuum between them. All fricatives along the continuum were synthesized in the Klatt Speech Synthesizer (Klatt 1980), with all synthesis parameters interpolated linearly and with equal differences between each step. Refer to the appendix for detailed synthesis parameters for these fricatives as well as for the vowels discussed below.

The vocalic sounds were based on [i] and [o] and varied from set to set. All contained an acoustic cue to the F2 of the vowel, but they differed in the other information present:

| Stimuli set | Description |
|---|---|
| Speech | All formants present; amplitude and pitch were dynamic, matched as closely as possible to human speech. |
| F0 + F2 | Impoverished vowels, consisting only of F0 (which was held constant at 100 Hz) and F2 (equal to the F2 in the Speech condition). Sounds from this set resembled buzzes with a simultaneous chirp. |
| Contour F0 + F2 | Same as above, but with an F0 contour identical to the Speech set. These sounded slightly more natural but still unlike human speech. |
| Sine at F2 | A single dynamic sine wave matched to the frequency of F2 taken from natural speech. These nuclei bore no resemblance to speech. |

The first three sets above were synthesized using Klatt. As such, the nonspeech conditions still had a speechlike harmonic makeup, although only one formant was present. (Said another way, the sound source resembled vocal fold vibration, but the filter function was not identifiable as coming from a vocal tract.) The Sine at F2 set was created in Praat based off of natural speech formants. All tokens were sampled at 22,050 Hz and adjusted to match the RMS amplitude of natural speech for each vowel. Fricatives were also synthesized using Klatt at 22,050 Hz. The endpoint fricative tokens were synthesized

independently to match natural speech as closely as possible; the formant amplitudes and frequencies were then interpolated linearly to seven intermediate steps to generate the continuum.

*Setup*

Subjects completed the study seated at a computer running E-Prime, receiving auditory stimuli over headphones and seeing text instructions on the computer monitor. During the experimental trials, subjects saw a static screen reminding them that the button on their left was for 's' and the one on their right for 'sh'. Responses were given using a button box.

*Task*

The task consisted of five separate blocks, each of which included stimuli drawn entirely from one of the above sets. The conditions were presented in the following order (note that the F0 + F2 set was presented twice):

| **Block 1** | **Block 2** | **Block 3** | **Block 4** | **Block 5** |
|---|---|---|---|---|
| F0 + F2 | Speech | F0 + F2 | Sine at F2 | Contour F0 + F2 |

For each block, subjects heard one stimulus at a time and were asked to judge whether the fricative was /s/ or /ʃ/ by pressing one of two buttons. Within each block, stimuli were presented in random order, with 7 tokens of each, for a total of 126 trials per block. The entire experiment took approximately 30 minutes.

After Block 1, subjects were asked briefly to associate four of the F0 + F2 stimuli with English words: after hearing the endpoint-fricative tokens based off of /si/ and /so/ (the two were presented separately), they were asked to match them to one of the words *see*, *say*, *saw*, *so*, or *sue* (/i/, /e/, /ɑ~ɔ/, /o/, /u/); after those based off of /ʃi/ and /ʃo/, they were asked to choose between *she*, *shay*, *shah*, *show*, and *shoe*. Between Blocks 2 and 3, they were presented with a written message informing them that the F0 + F2 stimuli were 'derived from' the Speech stimuli. They were tested on this set of stimuli twice, on either side of the speech block, to test if learning the association between the F0 + F2 nonspeech sounds and speech would enhance the degree of compensation.

*Calculating the boundary*

The crossover point, or boundary, between /s/ and /ʃ/ identification was calculated for each of the two vocalic nuclei in each condition; the degree of compensation was measured as the difference between the /o/ and /i/ boundaries. These boundaries were calculated by interpolating the point at which majority identification as one fricative would cross over into majority identification of the other. (Put another way, the point was found at which the subject would identify each fricative 50% of the time.) This crossover point can be thought of as a hypothetical token somewhere along the nine-step continuum, somewhere between two that were actually synthesized and tested.

This crossover was calculated by fitting a logistic function to the responses, with continuum token (1 through 9) along the *x*-axis and fricative identification (/s/ = 1, /ʃ/ = 0) along the *y*-axis. The formula fitted was:

$$y = 1 - \frac{1}{1 + \left(\frac{x}{a}\right)^b}$$

where *a* is the identification boundary (*a* being equal to the *x*-value at which *y* = 0.5), and *b* is a measure of the steepness of boundary crossing. The calculated value of *a* was used for the boundary itself; the values of *b* have not yet been analyzed. As there were 20 subjects tested on five blocks with two different vowels, a total of 200 boundaries were calculated.

There were two boundaries for which a fit using the above formula could not be found due to extreme fricative identification on either side of the boundary (that is, there were either one on no tokens with identification of anything but pure /s/ or /ʃ/). In these cases, the boundary was calculated by interpolating linearly between the two tokens on either side of the boundary.

**Results**

Between Blocks 1 and 2, subjects reported what vowels they thought the F0 + F2 stimuli resembled. In response to two stimuli with the low-F2 ([o]-based) nucleus, 100% of responses identified the vowel as a back rounded vowel; 90% identified the high-F2 nucleus as a front vowel. Despite these strong preferences, anecdotal reports suggest that the sounds in Set A were certainly not identifiable as speech.

In all sets, fricatives on either end of the continuum were perceived almost entirely as one fricative, with significant inconsistency observed in only a few of the steps between them. Steps 1–4 on the continuum were overwhelmingly identified as /s/ (step 4 had 90% identification as /s/ across all trials), and 7–9 as /ʃ/ (93% of step 7). Steps 5 and 6 were sites of the categorical boundary and showed the most variation.

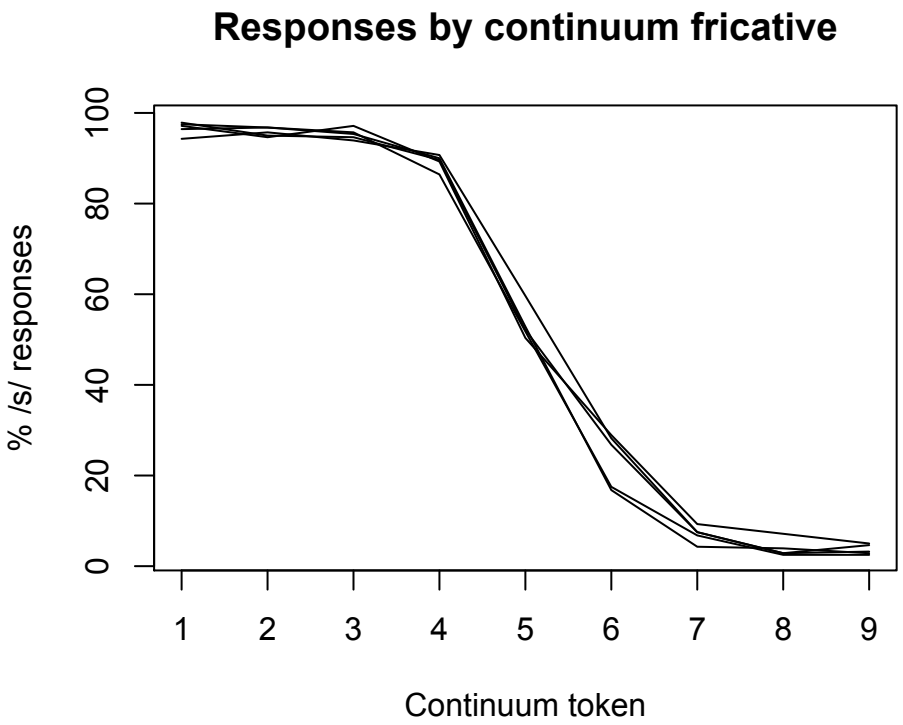


FIGURE 1: Percentage of /s/ responses for each step on the fricative continuum. Each condition is an overlaid line and includes both vowels and all subjects.

If the boundary location for the rounded vowel stimulus is measured significantly higher than that for the unrounded vowel, it can be taken as evidence for compensation: in this event, more of the ambiguous tokens are judged to be /s/ due to compensation for rounding coarticulation on the fricative. For every condition, the /i/ boundaries were subtracted from the /o/ boundaries, and I conducted a zero-mean *t*-test on these differences. (Note that this method captures the dependent relationship between the two samples, as would a paired *t*-test.) All *p*-values in the table below are adjusted by Holm–Bonferroni correction; all very significant results remained very significant, and all non-significant results remained non-significant. I have also included the mean difference, which can be interpreted as the number of continuum steps between the /i/ and /o/ boundaries, as well as the standard deviation. These data are visualized in Figure 2 below. The boundaries for each vowel, before taking the difference between them, are shown in Figure 3.

**Difference between /i/ and /o/ boundaries**

| | *t* | *p* | mean | σ |
|---|---|---|---|---|
| F0 + F2 | 1.90 | 0.14 | 0.42 | 0.98 |
| Speech | 5.33 | < 0.01 | 1.36 | 1.14 |
| F0 + F2 (2) | 4.42 | < 0.01 | 0.80 | 0.81 |
| Sine at F2 | 1.47 | 0.16 | 0.43 | 1.30 |
| Contour | 3.56 | < 0.01 | 0.54 | 0.68 |

TABLE 1: Significance tests of difference between /i/ and /o/ boundaries. As there are 20 subjects, all tests are carried out on 19 degrees of freedom. Significant results are shown in shaded rows.

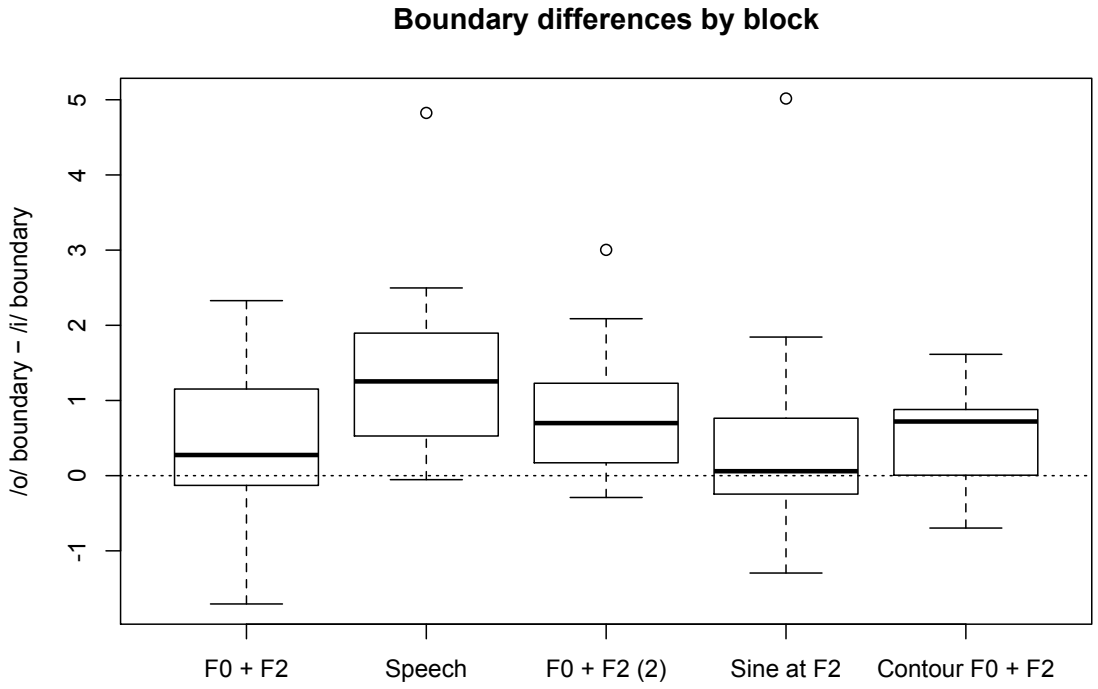


FIGURE 2

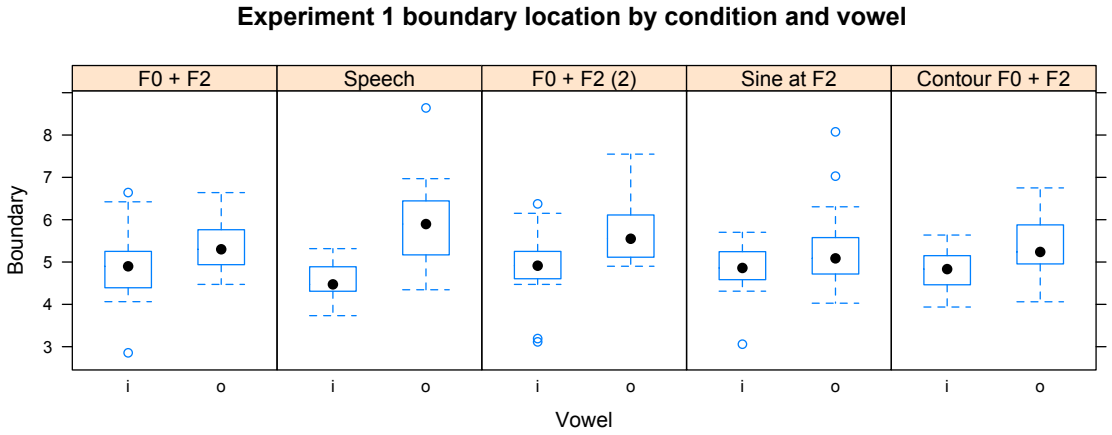


FIGURE 3

These results show that a reliable separation is happening in most conditions, but not in the Sine at F2 block or in the initial F0 + F2 block (although it is when subjects face this condition for a second time). Note also that all blocks shown to be significant have a higher mean boundary difference than all the non-significant blocks, and that the non-significant blocks also show a standard deviation of more than double the mean (recall that a mean of zero indicates no compensation). This high SD suggests that although subjects *on average* did in fact give more /s/ responses in front of the rounded vowel, they did so unreliably. Indeed, out of 20 subjects, 10 showed a higher boundary for /i/ in the Sine at F2 block, and 7 in the initial F0 + F2 block; compare this to just one in the Speech block and two in the second F0 + F2 block (and five in the Contour F0 + F2 block).

Given the fact that boundary separation is shown to be significant only in certain blocks, and given the dramatic difference in stimuli between conditions, we would expect condition to be a significant predictor of boundary difference. To confirm that the variation in boundary between conditions was statistically significant, a repeated measures ANOVA was conducted for each group. The most relevant test was on the interaction between vowel and condition—deviation by vowel is expected in most or all conditions. This interaction was indeed found to be highly statistically significant, $F(4,76) = 4.13$, $p < 0.01$. (Additionally, vowel alone was a statistically significant predictor: $F(1,19) = 24.37$, $p < 0.01$.)

To determine how all five blocks differed from each other, a series of post-hoc ANOVAs was carried out for every possible pairing of conditions. The table below shows all unique pairings along with the $F$ score and $p$-values, both raw and adjusted, for a test of the interaction between vowel and condition. All significant results are shaded; other non-significant results which did show statistical significance or near-significance before $p$-correction are boxed in a dotted line.

**Post-hoc ANOVAs between block pairs: interaction of vowel and block**

| Blocks compared | | $F(1,19)$ | unadjusted $p$ | adjusted $p$ |
|---|---|---|---|---|
| F0 + F2 | Speech | 12.55 | < 0.01 | < 0.05 |
| F0 + F2 | F0 + F2 (2) | 2.02 | 0.17 | 0.91 |
| F0 + F2 | Sine at F2 | 0.002 | 0.97 | 1 |
| F0 + F2 | Contour F0 + F2 | 0.24 | 0.63 | 1 |
| Speech | F0 + F2 (2) | 4.03 | 0.059 | 0.41 |
| Speech | Sine at F2 | 6.53 | < 0.05 | 0.15 |
| Speech | Contour F0 + F2 | 17.94 | < 0.01 | < 0.01 |
| F0 + F2 (2) | Sine at F2 | 2.23 | 0.15 | 0.91 |
| F0 + F2 (2) | Contour F0 + F2 | 1.89 | 0.19 | 0.91 |
| Sine at F2 | Contour F0 + F2 | 0.11 | 0.75 | 1 |

TABLE 2: Post-hoc ANOVA differentiating blocks

These results suggest that although condition has a significant impact on fricative identification when vowel is controlled, there are very few conditions that can be said to differ reliably from one another. The Speech block differs reliably from the F0 + F2 and Contour F0 + F2 nonspeech blocks, but not from any other nonspeech blocks. Additionally, no nonspeech conditions differ from each other. The implications of these results, as well as of considering the near-significant results, are discussed below.

**Discussion**

The results above indicate that compensation for coarticulation is occurring in most blocks. Only certain conditions met a threshold of statistical significance that would lead us to conclude this; namely, any condition with both F0 and F2 that can be linked to speech somehow. I specify 'linked to speech' because when introduced to the F0 + F2 stimuli for the first time, listeners do not perform compensation; only following the Speech block is there a significant boundary separation. (Recall also that between the first and second F0 + F2 conditions, listeners were asked to associate these stimuli with English words.) Despite the difference observed in the boundary difference *t*-tests, our post-hoc tests did not point to a significant difference between these two blocks. However, if we conduct a dependent samples sign test on the two groups—that is, the boundary differences of F0 + F2 pre- and post-speech—we see that 15 of the 20 ($p < 0.05$) post-speech F0 + F2 boundary differences are higher than for pre-speech. Note that the sign test does not test the magnitude of difference between two groups, as a *t*-test or ANOVA do, and is therefore more sensitive to weaker effects. We have reason to claim that there is some effect, if weak, of hearing a speech condition between two F0 + F2 blocks.

The significance of the difference between F0 + F2 conditions is that we can see the introduction of top-down effects if listeners are consciously aware of

the source of the sound they are hearing. That is, listeners will perform perceptual compensation, a process common in speech perception, upon stimuli that are clearly not speech if they are primed to think of them as being speech-related. It also appears that these effects can be introduced onto some types of nonspeech, but not to others: although the pure sine stimuli contained the requisite information to judge vowel roundedness, they are a far cry from speech, and apparently either the presence of F0 or the speechlike harmonic pattern (or both) is key to making this effect possible.

What then of the fact that even the non-statistically significant blocks do show a higher /o/ boundary than /i/? Notice that the mean boundary difference for both of these conditions is roughly 0.4 (compare to 1.4 for speech); is it possible that there is some purely auditory trigger for perceptual compensation that is applying in these two cases? We might then adopt the position that GR-driven and general auditory compensation effects can be additive: general auditory effects are responsible for the boundary separation observed in Block 1, and we see the addition of perception driven by GR in Blocks 2, 3, and 5. However, given the lack of demonstrable significance in these cases, no conclusions can be drawn at this time.

And unfortunately, there are also a few findings that cast doubt on the usefulness of some of these data. The failure to find significant differences between most of the conditions is not totally unexpected, given the high amount of variation in certain blocks, especially Speech and Sine at F2 (see Table 1). Note that the high standard deviation of boundary separation in both of these conditions may be to blame for the lack of effect in demonstrating their difference in post-hoc tests, given that Speech and F0 + F2 (which had evidently stronger compensation than Sine at F2) *were* shown to be significantly different. However, the testing of my hypothesis depends on showing these differences. If we could consider the dashed-line cases in Table 2, which were near-significant before *p*-correction, then it would be possible to claim that Speech differed from all nonspeech conditions, none of which differ from each other. And given the relative weakness of compensation in the sine condition, why was the difference between it and speech deemed to not quite be statistically significant? It is still unclear whether perceptual compensation is happening in the sine and initial F0 + F2 conditions, as there is indeed some boundary separation. Furthermore, this question will be very difficult to answer indeed because there was no control condition for the Sine at F2 block, and we cannot be sure if the minimal effect we saw in this block was due to prior exposure to the other stimuli or to an actual auditorily driven compensation effect.

Additional procedural shortcomings were discovered during the course of conducting the experiment. The method of constructing the interpolated continuum may be a problem, as a linear scaling of frequencies between the fricative stimuli does not translate to a constant human-perceptual distance between each step on the continuum. There are also a few highly suspect outliers in the data (see Figure 2). The length of the experiment and number of conditions may also be a concern, as we see compensation in the final (Contour at F2) block waning, even to the point that it is significantly different from Speech compensation, despite being objectively more like speech than the plain F0 + F2 condition.

With all of these concerns in mind, a second experiment was designed to re-assess and buttress these findings. I outline the method and results of Experiment 2 below before returning to a discussion of the theoretical implications. As the task in Experiment 2 hews closely to that of Experiment 1, the following section will discuss only the differences between the two.

**Experiment 2: Method**

*Participants*

Participants for Experiment 2 were divided into two groups. Each contained 16 participants aged 17–22 who spoke English as a native language and did not report any history of hearing or language disorders. Each group was tested on a different protocol, both of which are discussed below.

*Stimuli*

Stimuli similar to those from Experiment 1 were used, with a few modifications. All "vocalic" nuclei were the same as before, but fricatives were resynthesized. The nine-step continuum was recalculated from different endpoints—steps 2 and 8 from the old continuum, which allowed for a slightly higher resolution in locating the categorical boundary.

Additionally, the interpolated continuum stimuli were calculated using the Bark scale, which should correspond more accurately to human perception. Formant values for the fricatives were converted to Bark before scaling, then converted back into Hz. Amplitude was still interpolated linearly. See the appendix for further details on these stimuli's acoustic parameters.

*Task*

Experiment 2 involved two different protocols, with no subjects performing both. Subjects in Group 2 performed a task similar to the task in Experiment 1, while Group 1 performed a shorter and slightly different task. The Contour F0 + F2 condition was removed entirely for this experiment, as it did not seem to differ interestingly from the plain F0 + F2 condition.

| | **Block 1** | **Block 2** | **Block 3** | **Block 4** |
|---|---|---|---|---|
| Group 1: | F0 + F2 | Speech | F0 + F2 | Sine at F2 |
| Group 2: | Sine at F2 | Speech | Sine at F2 | |

All subjects saw a message between the second and third blocks informing them that the stimuli they were about to hear were derived from the speech sounds they just heard, although they were not asked to match the F0 + F2 stimuli with English words. The Sine at F2 condition was given the same treatment as the F0 + F2 condition—that is, testing before and after a speech block—to determine if association with speech would affect the results for the former in the way we saw it affecting the latter.

As in Experiment 1, stimulus presentation was random, and 7 instances of each of the 18 distinct stimuli were presented for 126 trials per block. Each block

took approximately 5 minutes. A slight modification was also made to the procedure of the experiment: in addition to the labeled screen they saw in Experiment 1, subjects were given a modest visual feedback in the form of a blank screen (for 200 ms) after responding to a stimulus.

**Results**

As in Experiment 1, tokens at the ends of the continuum were strongly identified as one fricative or the other, suggesting a categorical response. Steps 1–3 and 7–9 were largely unambiguous, while the boundary tended to fall near steps 4–6. (Recall that step 4 was less ambiguous in Experiment 1, owing presumably to the frequency range being narrower in Experiment 2.) The figure below shows the overlay of all conditions for both groups.

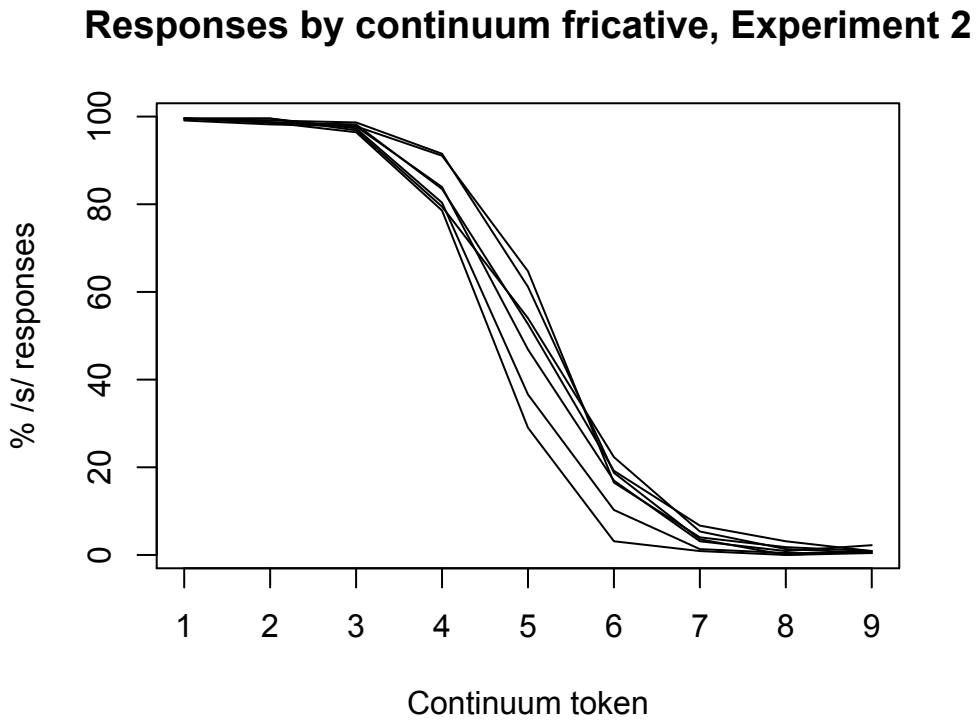


FIGURE 4: Percentage of /s/ responses for each step on the fricative continuum. Each condition is an overlaid line and includes both vowels and all subjects.

The Experiment 2 data were subjected to the same tests used for Experiment 1. The table below tests the significance of the separation between /i/ and /o/ boundaries. All conditions except for pure sine showed a highly significant ($p < 0.01$) effect of vowel. Values of $p$ are adjusted as they were in Experiment 1. (Correction is applied in acknowledgment of the fact that there are seven tests, not two groups of three and four tests corrected for separately; significant values are all very robust.)

**Boundary difference: Group 1**

| | *t* | *p* | mean |
|---|---|---|---|
| Sine F2 | –1.10 | 0.82 | –0.15 |
| Speech | 7.83 | < 0.01 | 1.36 |
| Sine F2 (2) | –0.38 | 0.82 | –0.06 |

**Boundary difference: Group 2**

| | *t* | *p* | mean |
|---|---|---|---|
| F0 + F2 | 3.70 | < 0.01 | 0.40 |
| Speech | 4.56 | < 0.01 | 1.22 |
| F0 + F2 (2) | 4.30 | <0.01 | 0.55 |
| Sine F2 | 1.14 | 0.82 | 0.12 |

TABLE 3: Significance tests of difference between /i/ and /o/ boundaries. As there are 16 subjects, all tests are carried out on 15 degrees of freedom. Significant results are shown in shaded rows.

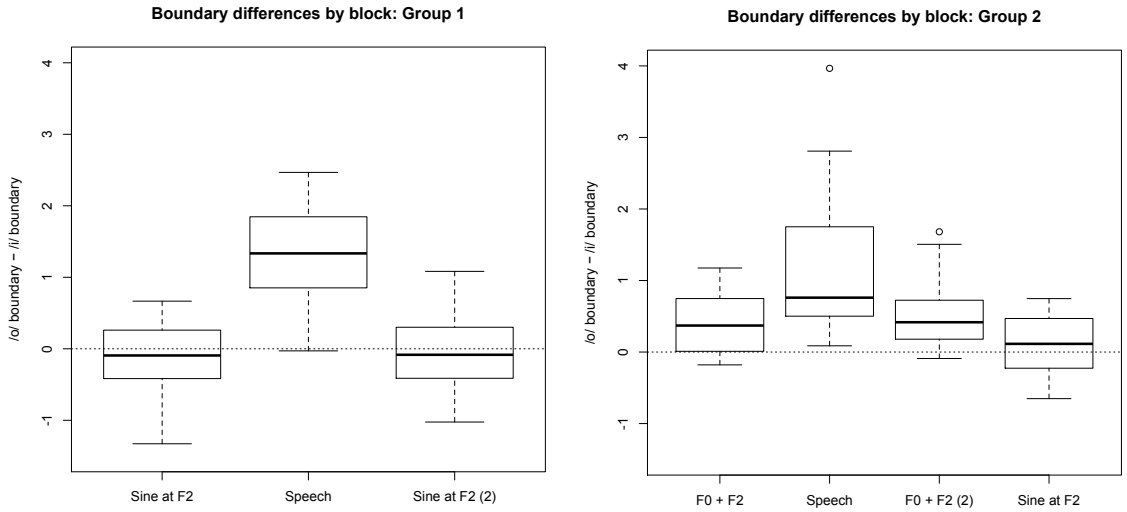


FIGURE 5

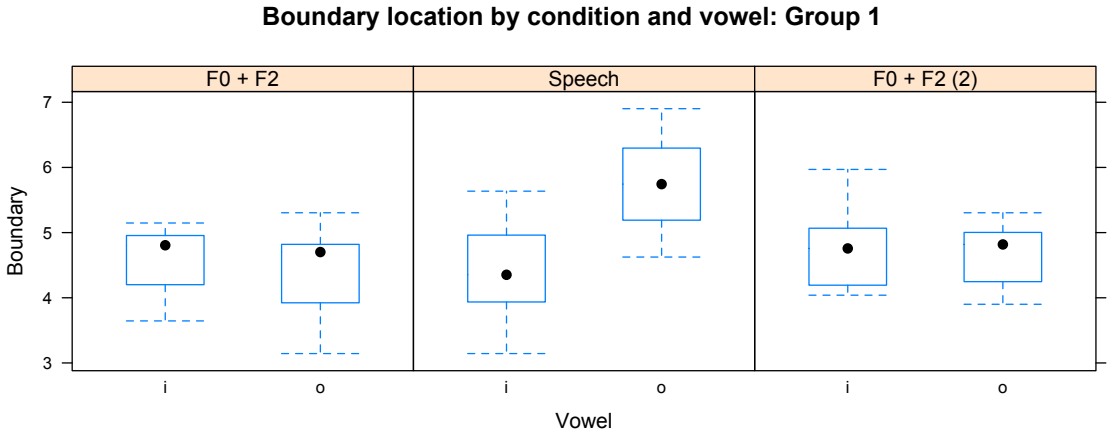


FIGURE 6a

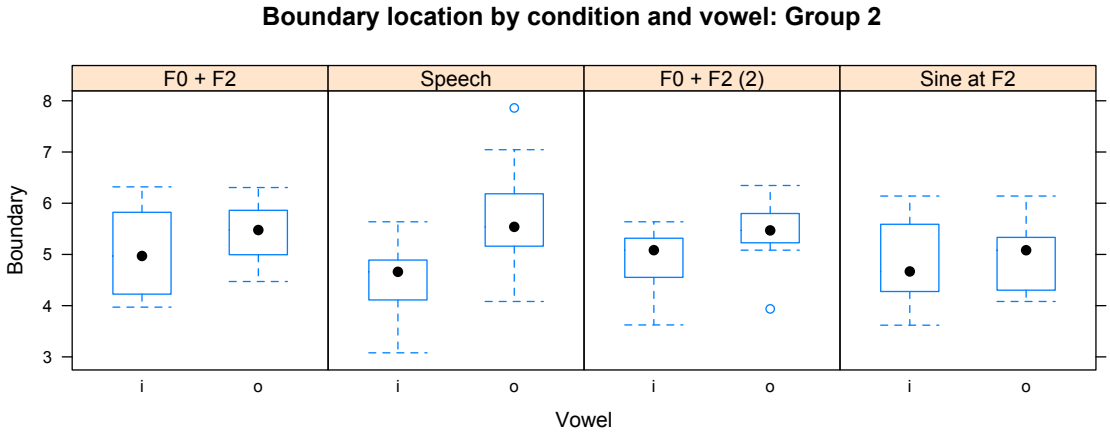


FIGURE 6b

Table 3 and Figures 5–6 above show a statistically significant effect of vowel on boundary separation in all cases except for the Sine at F2 condition.

An ANOVA determined the interaction between vowel and condition as predictors to be statistically significant: $F(2,30) = 44.14$, $p < 0.01$ for Group 1, $F(3,45) = 10.94$, $p < 0.01$ for Group 2. (Vowel alone was also a statistically significant predictor for both groups: $F(1,15) = 11.72$, $p < 0.01$ for Group 1, $F(1,15) = 26.20$, $p < 0.01$ for Group 2.) Post-hoc repeated measure ANOVAs were again conducted, comparing all blocks within each group; results are given in Table 4.

**Post-hoc ANOVAs between block pairs: interaction of vowel and block**

**Group 1**

| Blocks compared | | $F(1,15)$ | unadjusted $p$ | adjusted $p$ |
|---|---|---|---|---|
| Sine at F2 | Speech | 57.96 | < 0.01 | < 0.01 |
| Sine at F2 | Sine at F2 (2) | 0.44 | 0.52 | 0.52 |
| Speech | Sine at F2 (2) | 51.72 | < 0.01 | < 0.01 |

**Group 2**

| Blocks compared | | $F(1,15)$ | unadjusted $p$ | adjusted $p$ |
|---|---|---|---|---|
| F0 + F2 | Speech | 8.18 | < 0.05 | < 0.05 |
| F0 + F2 | F0 + F2 (2) | 1.66 | 0.22 | 0.22 |
| F0 + F2 | Sine at F2 | 4.89 | < 0.05 | 0.09 |
| Speech | F0 + F2 (2) | 7.43 | < 0.05 | < 0.05 |
| Speech | Sine at F2 | 23.38 | < 0.01 | < 0.01 |
| F0 + F2 (2) | Sine at F2 | 11.68 | < 0.01 | < 0.05 |

TABLE 4

Post-hoc tests show all Speech blocks to differ from all nonspeech conditions (F0 + F2 and Sine at F2); note that this is a different grouping from what we saw in the *t*-tests above, which showed significant boundary difference in the Speech and F0 + F2 conditions, but not in Sine at F2.

**Discussion**

The results of Experiment 2 corroborate many of the results observed in Experiment 1. I begin by addressing what has been confirmed or strengthened and subsequently discuss how the results diverge. I then synthesize the findings and discuss their implications for speech perception theory.

As in Experiment 1, there is a strong and significant degree of perceptual compensation to speechlike sounds. And as before, there is also observable and statistically significant compensation to certain nonspeech stimuli. Once again it appears that sounds similar enough to speech are inducing top-down effects, causing hearers to interpret the sounds, to some degree, as if they were speech—perhaps through mechanisms normally associated with speech perception. These results are important even given what we already knew from Experiment 1—not only do we know that the compensation effect for nonspeech is repeatable, but it appears that with the right methods we can demonstrate a statistically significant difference from speech. The Experiment 1 results led us to suspect this, and these results make a stronger case for speech and nonspeech being processed differently by the listener. This is especially interesting given that our tests do not find the F0 + F2 results to be as discernible from the pure sine condition as they are from speech. Yet, we know that something is different about them: in Experiment 2, there was never a case that would lead us to believe that compensation is occurring for the Sine at F2 blocks. In Group 1, both sine conditions yielded what looked like anti-compensation. This allows us to rule out two scenarios: that partial effects in compensation may be due to purely spectral cues, and that a bare F2 would trigger top-down processing the same way the F0 + F2 condition does.

Looking further into the top-down effects, and specificially into the question of their learnability, this is also an area in which the Experiment 2 results deviate in a remarkable way from what we discovered in Experiment 1. In Experiment 1 we only saw statistically significant boundary divergence post-speech, and a sign test hinted that although the difference between the two F0 + F2 blocks was small, the proportion of subjects who showed a stronger compensation effect in the second block suggests that they were treating the stimuli more as speech. In Experiment 2, we see compensation happening with the very first set of stimuli that Group 2 encounters, and while the boundaries do seem to diverge slightly more with the second block, this difference is not significant, whether we appeal to a *t*-test (Table 3) or a sign test—which shows that only 10 of the 16 subjects compensated more strongly in the second block ($p = 0.45$). Why the apparent difference between the two experiments, then? I imagine that the more obvious forced association of the F0 + F2 stimuli to speech in Experiment 1 is responsible—recall that subjects were asked to match the sounds to English words in that experiment. However, given that participants in Experiment 1 and Experiment 2 treated the F0 + F2 condition differently even upon their first exposure, it may simply be a chance difference in the groups. This is still largely speculative, and should be confirmed with an experimental condition similar to Experiment 2 but with this association task.

The results of Experiment 2 diverge in other ways from Experiment 1. Recall that in the discussion of Experiment 1, there was a hint that some compensation may be occurring in the Sine at F2 conditions, although slight and

not demonstrably significant. In my previous discussion I speculated that some general auditory process may be responsible for what may have been a low degree of compensation for the F0 + F2 and Sine at F2 conditions, but it appears that this is not the case due to anti-compensation by Group 1. Was the effect in Experiment 1 due entirely, then, to random error, or perhaps to deficiencies in the experimental method? Given the non-significance, error is a plausible explanation. Whatever the cause, it appears that this effect was not repeatable in the second experiment, and we should therefore not consider it to be an essential component of the phenomena we are testing. It is true that Group 2 appeared to show higher compensation than Group 1, and that Group 1 is the only group of subjects not to have been exposed to the F0 + F2 stimuli, perhaps raising the question of whether learned association with speech can happen only when the listener hears a somehow 'intermediate' stimulus; however, this effect also is very small and should probably not be the cause of too much speculation at this stage. An experimental condition could be designed to motivate listeners to induce top-down effects to pure F2 stimuli given enough guided transitional steps from a speech condition.

I do believe it worth noting that a recurrent theme throughout this study is the effect that is observable and intuitive but not deemed significant through common statistical methods. Part of the problem is that the hypotheses I am testing predict partial effects, which can be very difficult to measure reliably, especially when a true partial effect is easily mistakable for insignificant error. One simple solution to this difficulty may be to maximize the number of data points. I conducted this study with 16–20 participants because such a number is usually sufficient for behavioral linguistic experiments, but a partial effects study may very well have stricter requirements. There may also exist other solutions, such as a slightly different testing condition; however, other methods that I piloted to determine boundary location generated at least as much error as the /s~ʃ/ identification task I used. It is possible that a higher *n* would indicate more statistically reliable differences in places where I suspect that true perception-related differences exist.

However, going from the combined results of both experiments, and especially Experiment 2, we can see plainly that it is possible for a nonspeech stimulus to invoke a compensatory response. Compensation for nonspeech has been shown many times before, but this case is especially interesting because I have shown a diminished effect that cannot be explained as belonging to a processing mechanism different from speech or as being one component of an aggregate effect. That is, it does not appear that a speech-unrelated psychoacoustic effect drives compensation for F0 + F2 'vowels'. It is unlikely that some acoustic property of the vowel is affecting perception of the consonant frequencies, as the frequencies differentiating [s] and [ʃ] are all well above the F2 of vowels. And the experimental evidence agrees: given Group 1's complete failure to compensate for sine-wave stimuli, it appears that the presence of F2 does not license compensation; rather, the reminiscence of F0 + F2 to speech invokes speech-processing mechanisms, but to a lesser degree than sounds clearly identifiable as speech. Subjects in Experiment 1 even largely agreed on which vowels to associate these stimuli with, yet these subjects too showed a reduced effect, even with the second F0 + F2 block. Somewhere in the application

of top-down effects is what we might characterize as an unwillingness by the speech perception program to view these sounds as speech, yet it seems that the mere fact that these sounds are reminiscent of speech leads human listeners to accept the possibility that *vocal articulation* may be the source of what they are hearing. It is possible that the harmonic makeup of sounds made in the Klatt synthesizer are reminiscent enough of vocal fold vibration, even without the proper formants, to cause this effect. The results achieved in Experiment 2 do not support the hypothesis stated before that partial effects are achieved by the isolation of perceptual mechanisms; that is, a weaker effect may simply be a weakened effect, not an effect with one of its components removed.

Another implication of this study, although an admittedly weaker one, is the possibility for top-down perception effects to be learnable, speaking about the difference in the F0 + F2 conditions pre- and post-speech. A number of previous studies have shown that aspects of linguistic competence can mediate categorical speech perception, including phoneme frequency (Yoneyama, Johnson, & Kataoka 2011) and lexical probability (Elman & McClelland 1988). In my study, subjects used recently acquired conscious knowledge to shift the location of a perceptual boundary; this situation may be different from those cited above, as the effect is not a property of the listener's linguistic ability but rather a recent, short-term association. This is certainly an aspect of my findings that bears further and deliberate investigation.

**Conclusion**

The discussion of two experiments presented in this paper has focused on explaining compensation for coarticulation performed on both speech and nonspeech. When presented with certain nonspeech sounds, subjects did perform perceptual compensation, but to a lesser degree than when exposed to speech. Although earlier indications suggested value in the hypothesis that this difference in degree was related to component processes in speech production, later results appeared to challenge that view, pointing rather to a general weakening of speech perceptual processes when they are applied in an artificial, top-down fashion, without reliance on a general auditory explanation. It appears that the observance of minor effects of processes such as compensation for coarticulation does not necessarily indicate that some component processes have been tidily isolated.

## References


Diehl, R.L., A.J. Lotto and L.L. Holt. 2004. Speech Perception. *Annual Review of Psychology* 55, 149–79.

Elman, Jeffrey and McClelland, James. 1988. Cognitive Penetration of the Mechanisms of Perception: Compensation for Coarticulation of Lexically Restored Phonemes. *Journal of Memory and Language* 27, 143–165.

Fowler, C.A. 1986. An event approach to a theory of speech perception from a direct-realist perspective. *Journal of Phonetics* 14, 3–28.

Fowler, C.A., C.T. Best and G.W. McRoberts. 1990. Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception and Psychophysics* 48 (6), 559–570.

Fowler, C.A., J.M. Brown and V.A. Mann. 2000. Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance* 26, 877–88.

Johnson, Keith. 2011. Retroflex versus bunched [r] in compensation for coarticulation. *UC Berkeley Phonology Lab Annual Report*.

Holt, L.L. and A.J. Lotto. 2002. Behavioral examination of the neural mechanisms of speech context effects. *Hear. Res* 167, 156–169.

Klatt, Dennis H. Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America* 67 (3), 971–995.

Liberman, A.M., F.S. Cooper, D.P. Shankweiler, and M. Studdert-Kennedy. 1967. Perception of the speech code. *Psychological review* 74, 431–461.

Li, Fangfang, Jan Edwards and Mary E. Beckman. 2009. Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics* 37 (1), 111–124.

Lotto, A.J., K.R. Kluender and L.L. Holt. 1997. Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America* 102, 1134–40.

Lotto, A.J. and K.R. Kluender. 1998. General contrast effects in speech perception: Effectof preceding liquid on stop consonant identification. *Perception and Psychophysics* 60, 602–619.

Mann, Virginia. 1980. Influence of preceding liquid on stop-consonant perception. *Perception & Psycholinguistics* 28 (5), 407–412.

Mitterer, Holger. 2006. On the causes of compensation for coarticulation: Evidence for phonological mediation. *Attention, Perception, & Psychophysics* 68 (7), 1227–1240.

Möttönen, Riika and Watkins, Kate E. 2009. Motor Representations of Articulators Contribute to the Categorical Perception of Speech Sounds. *The Journal of Neuroscience* 29 (31), 9819–9825.

Pitt, Mark and McQueen, James. 1998. Is Compensation for Coarticulation Mediated by the Lexicon? *Journal of Memory and Language* 39, 347–370.

Viswanathan, Navin, James S. Magnuson, and Carol A. Fowler. 2010. Compensation for Coarticulation: Disentangling Auditory and Gestural Theories of Perception of Coarticulatory Effects in Speech. *Journal of Experimental Psychology: Human Perception and Performance* 36 (4), 1005–1015.

Wade, Travis and Lori Holt. Effects of later-occurring nonlinguistic sounds on speech categorization. ). *Journal of the Acoustical Society of America* 118 (3), 1701–1710.

Yoneyama, Kiyoko, Keith Johnson and Reiko Kataoka. 2011. An effect of phoneme frequency on stop place perception by English-speaking and Japanese-speaking listeners. 161st Meeting of the Acoustical Society of America, Seattle, Washington, May 23 (poster session).

## Appendix

*Vowel synthesis*

Below are tables for parameters fed to the Klatt synthesizer to generate the vocalic nuclei used. Table A1 shows information for all five formants utilized for the speechlike vowels as well as the peak amplitude. All of these were synthesized with an F0 (Klatt parameter 'f0') starting at 190 and falling to 100. The Klatt master gain parameter 'g0' was constant at 60. All vowels were 300 ms in length.

| | F1 | F2 | F3 | F4 | F5 | p1 | p2 | p3 | p4 | p5 | amp (dB) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| i | 300 | 2219-2445-2208 | 3139-3362-2801 | 4289 | 3700 | 80 | 200 | 350 | 500 | 600 | 71.8 |
| o | 480 | 1620-860 | 2773-2568 | 3354 | 4000 | 60 | 90 | 150 | 500 | 600 | 73.4 |

TABLE A1: Formant frequencies and amplitude parameters used in the Klatt speech synthesizer for speechlike vowels.

F0 + F2 vowels were synthesized with only one formant, which was set equal to the F2 of vowels in the Speech condition. For these stimuli, F0 was held constant at 100 Hz.

| | F1 | p1 | amp (dB) |
|---|---|---|---|
| i | 300 | 2219-2445-2208 | 75.0 |
| o | 480 | 1620-860 | 74.2 |

TABLE A2: Formant frequency and amplitude for F0 + F2 vowels.

Stimuli from the F0 + F2 Contour condition differed from F0 + F2 only in that the F0 value followed the 190–100 Hz contour used in the Speech block.

Sounds from the Sine at F2 set were not synthesized using Klatt, but rather in Praat by extracting F2 values from natural speech. The natural speech tokens used were from the same speaker as the tokens upon which Klatt synthesis was modeled. The maximum F2 value for /o/ especially was lower than for the Klatt-synthesized conditions, although the mean is similar.

| | min pitch (Hz) | max pitch (Hz) | mean pitch (Hz) | peak amp (dB) |
|---|---|---|---|---|
| i | 1800 | 2515 | 2397 | 71.0 |
| o | 688 | 1012 | 918 | 75.0 |

TABLE A3: Pitch and amplitude values for the Sine at F2 stimuli

*Fricative synthesis*

Klatt parameters for fricative synthesis are given below. Token 1 is endpoint /s/ and 9 is endpoint /ʃ/. All fricatives are 240 ms.

| | F2 | F3 | F4 | F5 | F6 | a3 | a4 | a5 | a6 | g0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3250 | 4661 | 5875 | 4812 | 9625 | 35 | 44 | 58 | 53 | 66 |
| 2 | 3011 | 4341 | 5775 | 7661 | 9343 | 38 | 47 | 60 | 55 | 64 |
| 3 | 2790 | 4042 | 5677 | 7514 | 9062 | 42 | 51 | 62 | 57 | 62 |
| 4 | 2584 | 3764 | 5581 | 7369 | 8781 | 46 | 54 | 64 | 59 | 61 |
| 5 | 2392 | 3504 | 5487 | 7227 | 8500 | 50 | 58 | 67 | 62 | 59 |
| 6 | 2214 | 3262 | 5394 | 7088 | 8212 | 53 | 61 | 69 | 64 | 57 |
| 7 | 2048 | 3036 | 5303 | 6952 | 7937 | 57 | 65 | 71 | 66 | 56 |
| 8 | 1894 | 2825 | 5213 | 6818 | 7656 | 61 | 68 | 73 | 68 | 54 |
| 9 | 1750 | 2628 | 5125 | 6687 | 9395 | 65 | 72 | 76 | 71 | 53 |

TABLE A4: Formant frequencies and amplitudes for fricatives.