Retroflex versus bunched [r] in compensation for coarticulation

Keith Johnson
UC Berkeley

This paper presents data from two experiments that are problematic for the auditory spectral contrast theory of compensation for coarticulation, but are perfectly compatible with the gesture recovery theory. The experiments compare compensation for coarticulation effects in the perception of a [dɑ]-[ɡɑ] continuum produced by retroflex [r] context, versus the lack of an effect for a bunched [r] context. The key acoustic feature referenced by the spectral contrast theory (the lowered F3) is present in both bunched and retroflex [r]. In experiment 2, the F4 trajectory (a key acoustic correlate of retroflexion) is seen to modulate compensation for coarticulation.

INTRODUCTION

“Compensation for coarticulation” is the name of a phenomenon that has two explanations. In one theory (Holt, 1999; Lotto & Kluender, 1998), the effect is due to auditory spectral contrast as a context syllable interacts with a following target syllable in the auditory system. The main competing theory gave the phenomenon its name (Mann, 1980; Fowler, 2006). In this theory, the listener perceptually compensates for the coarticulation that typically occurs between the context syllable and the target syllable. Thus, the interaction between context and target reflects a perceptual gesture recovery process.

The particular instance of compensation for coarticulation that is studied in this paper is the one reported by Mann (1980). The perceptual boundary on a stop place of articulation continuum from [dɑ] to [ɡɑ] is shifted by a preceding [ɑl] or [ɑr] context syllable. If the context syllable ends in [l] the boundary is shifted toward [d] - more of the tokens are perceived as [ɡɑ]. And, if the context syllable ends in [r], the boundary is shifted toward [g] - more of the tokens are perceived as [dɑ]. The effect can be quite striking and makes an effective classroom demonstration of context effects in perception.

Mann (1980) called this phenomenon “compensation for coarticulation” on the hypothesis that listeners “parse” the effects of coarticulation as they perceptually recover the intended gestures of the speaker. The idea is that because tongue position in [l] is more forward than it is in [r], a following ambiguous syllable on a [dɑ]-[ɡɑ] continuum will sound more like [dɑ] in the [ɑr] context because listeners can attribute some of the “retraction” of the ambiguous syllable to coarticulation with [r]. On the other hand, when the context syllable ends in [l] there is no such coarticulatory cause of the “retraction” of the ambiguous syllable, so it is heard as intentionally back and more like [ɡɑ].

Evidence for the gesture recovery interpretation (e.g. Fowler, Best, McRoberts, 1990; Fowler, Brown & Mann, 2000; Fowler, 2006) has been unconvincing. By this I mean that scholars who are not predisposed by training to see perceptual phenomena in terms of gesture recovery have remained unconvinced that the compensation for coarticulation phenomenon unambiguously indicates that a gesture recovery mechanism is at work. The main reason for this sustained skepticism is that there appears to be another explanation that does not require one to assume that compensation for coarticulation involves gesture recovery.

The alternative explanation (Lotto & Kluender, 1998; Holt, 1999) focusses on the specifics of the acoustic properties of the context and target stimuli. The primary acoustic cue distinguishing [dɑ] and [gɑ] in the synthetic stimuli used by Mann (1980) is the onset frequency of the third vowel formant (F3). When F3 starts from a relatively high frequency and glides down into the vowel [ɑ] the syllable sounds like [dɑ] (see figure 1). Conversely, when F3 starts from a low frequency and glides up into the vowel [ɑ] the syllable sounds more like [gɑ]. The precursor syllables [ɑl] and [ɑr] are also distinguished by the frequency of F3 – [ɑl] ends with a high F3, and [ɑr] ends with a low F3. So, the contrast between the ending F3 of the precursor syllable and the starting F3 of the stop-consonant results in a perceptuɑl shift. When the precursor ends with a high F3 [ɑl], ambiguous test syllables will by virtue of an auditory contrast effect tend to sound more like they start with a lower F3 than when the precursor syllable ends with a low F3. Thus, a simple and commonly observed psychoacoustic phenomenon (spectral contrast) explains why [ɑl] causes ambiguous test stimuli to sound like [gɑ] while [ɑr] causes the same stimuli to sound like [dɑ]. Evidence for this theory (Lotto & Kluender, 1998; Lotto, Sullivan & Holt, 2003) has come from the use of nonspeech stimuli, showing that "compensation for coarticulation" effects can be generated with stimuli that don't sound like speech. The idea is that if nonspeech stimuli produce the perceptuɑl effect then it can't be due to a speech-specific perceptuɑl mechanism like gesture recovery.

One interesting aspect of these findings – the relative sizes of the effects – has not been discussed very much in the literature. In one exception to this trend, Lotto et al. (2003, following Holt & Lotto, 2002) note that dichotic presentation of the context syllable to one ear and the target syllable to the other produces a smaller context effect than does diotic presentation. They argue, correctly I think, that because the context effect is produced even with the dichotic presentation it is reasonable to suggest that there is a central component involved in the compensation for coarticulation phenomenon. The converse is also true – that because dichotic presentation produces a weak context effect, there may be some peripheral interaction between the context and target syllables, that is, we can identify different component processes in the overall process of compensation for coarticulation by paying attention to the magnitude of the effect in different experimental manipulations. It is therefore important to note that nonspeech context effects (Lotto & Kluender, 1998), while present, are never as large as speech context effects. I would argue that the different magnitudes of context effects with speech verus nonspeech context tokens is informative.

This discussion goes beyond the specific rationale for the experiments described in this report, but speaks to the general approach that will be outlined in the conclusion of this paper. For now we turn to the rationale for the present set of experiments.

Viswanathan et al. (2010) have recently reported that not all "r" sounding precursors result in a greater number of "d" responses in a compensation for coarticulation experiment. One of the precursor syllables that they used in their experiment was a naturally produced [ɑr] in which the final consonant was a Tamil trilled [r] – which was apparently produced with a single tongue tap (see their figure 2). While this segment has a low F3 like the English approximant [r], unlike English [r] the F4 in the Tamil [r] remained as high as it was in the [ɑl] precursor syllable. Viswanathan et al. argued that retroflex tongue posture is the phonetic property that united the class of segments that causes compensation for coarticulation – not low F3. Their experiment 3 supported this interpretation. They showed that with nonspeech analogs of the Tamil precursors, listeners showed a compensation effect that seemed to be driven by the F3 frequency pattern of the precursor syllable rather than by its phonetic place of articulation. Thus, they suggested, compensation for coarticulation, with speech stimuli, has a gestural component. When the tongue is retroflex in the context token, the compensation pattern is obtained. And when the context segment does not have a retroflex tongue posture (even if it does have a low F3), the context effect is not observed. One criticism of Viswnathan et al.'s argument is that their experiment involved the use of non-native, and thus unfamiliar, speech sounds. Thus, although listeners identified the Tamil trilled [r] as an example of “r” in a forced choice identification task, one can still wonder if the non-nativeness of the sound had an impact on its perception.

Interestingly, the acoustic differences between Tamil [r] and English [r] are similar to the differences between the retroflex and bunched variants of the English approximant [r]. Mielke et al. (2007) and Westbury et al. (1998) have found that these two variants of [r] are found in a dialectally homogeneous groups of speakers in Arizona and Wisconsin (respectively) and Mielke et al. found within-speaker variation such that a person might have retroflex [r] in one phonetic environment and bunched [r] in another environment. So, speakers of American English have probably encountered both variants of [r] and find them both to sound natural. What is more, the acoustic differences between retroflex and bunched [r], in Zhou et al.'s (2008) MRI study, mirror the differences reported by Viswanathan et al. for their American English [r] and their Tamil [r]. Zhou et al. (2008) performed MRI analyses of tongue shape in the approximant [r] produced by two native speakers of American English. One used a bunched tongue shape and one used a retroflex tongue shape (the key figure was reprinted in Ladefoged & Johnson, 2010, p. 95). Zhou et al. also measured the formant frequencies of the bunched and retroflex [r] and found that both had very low F3, but F4 was 500 Hz lower in retroflex [r] (comparable to the F4 in Viswanathans's AE [r]). The F4 of bunched [r] was higher, and comparable to Viswanathan's Tamil [r]. The experiments reported here were conducted to learn if American English retroflex and bunched [r] produce different compensation for cooarticulation effects.

I. EXPERIMENT 1

The first experiment is an extension of Mann (1980) comparing the compensation for coarticulation effect for bunched and retroflex [r]. The synthesis parameters for the bunched and retroflex [r] stimuli were drawn from Zhou et al.'s (2008) MRI /acoustic study of American English [r]. They found that bunched and retroflex [r] differ substantially in F4 frequency but also have some smaller differences in F2 and F3. Interestingly, for their speakers, the F3 of bunched [r] dipped to a lower frequency than did the F3 of retroflex [r]. This leads the auditory spectral contrast theory and the compensation for coarticulation theory to make opposite predictions. If the auditory spectral contrast theory is correct then the bunched stimuli should produce a larger boundary shift (comparing [dɑ]-[gɑ] perception in the context of [ɑl] and [ɑr]) than will be seen with a retroflex [r] token. This is because F3 drops to a lower frequency in the bunched [r] context token. The compensation for coarticulation model predicts that the retroflex token will produce a larger boundary shift because retroflex [r] involves tongue tip retraction that is not found with bunched /r/.

A. Method

1. Participants

There were two groups of participants. The first group was composed of 16 subjects (11 women and 4 men). The were undergraduate students at UC Berkeley aged 19 to 24. This group completed a compensation for coarticulation experiment that contrasted [ɑl] and [ɑr] with a bunched /r/. Two of these subjects, with native competence in English, began learning English at the age of 5 having first learned Cantonese (one subject) or Mandarin (the other subject).

The second group was also composed of 16 subjects (9 women and 6 men). They were aged 19 to 25, undergraduate students at UC Berkeley. These subjects completed a compensation for coarticulation experiment like the one given to group one, except that the [ɑr] context token was synthesized with a retroflex /r/. One subject in group two, who had native competence in English, began learning English at the age of 6 having first learned Cantonese. All of the subjects had native-speaker competence in English and normal hearing. They were paid a nominal sum ($5) for their participation.

2. Stimuli

The stimuli for this experiment were synthesized using a software formant synthesizer (Klatt, 1980; Klatt & Klatt, 1993) with a digital sampling rate of 11025 Hz. The tokens were composed of a context syllable that sounded like either [ɑr] or [ɑl] and a target syllable drawn from a continuum from [dɑ] to [gɑ]. The steady-state formant (and bandwidth) values of the [ɑ] vowels in both the context and the target stimuli were: F1 = 700 (60), F2 = 1100 (90), F3 = 2500 (150), F4 = 3550 (200). F5 was fixed at 4200 (200) for all of the context and target syllables. The F0 of voicing was also fixed at a monotone 100 Hz for all of the syllables. The target syllables were 235 ms long and had initial formant transitions over the first 70 ms. Table 1 shows the starting formant frequencies for the stop-consonant transitions. Token 1 (see figure 1) had formants appropriate for [d] while token 9 had formants for [g].

Table 1. The onset formant frequencies of the [dɑ]-[gɑ] stimuli. By time 70 ms the formants attain their vowel [ɑ] frequencies – in a linear ramp.

	[dɑ]								[gɑ]
	1	2	3	4	5	6	7	8	9
F1	300				-----				300
F2	1500	1501	1522	1543	1565	1586	1607	1628	1650
F3	2750	2656	2562	2468	2375	2281	2187	2093	2000
F4	3400				------				3400

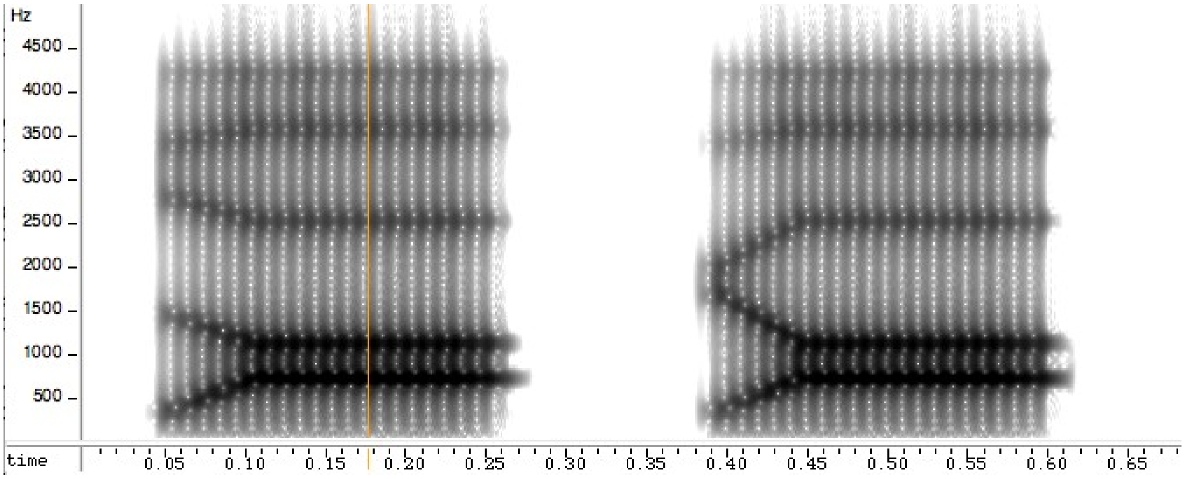

Figure 1. Spectrograms of the [dɑ] (token 1) and [gɑ] (token 9) endpoint stimuli.

The [ɑl] and [ɑr] context stimuli were 190 milliseconds long, and there was a 35 ms silent gap between the context syllable and the target syllable. The formant movements that cued [l] or [r] were 65 ms long. As indicated above, the F0 and vowel steady-state formant frequencies were the same in the context tokens as in the target continuum. The [l] and [r] formant transitions ramped linearly to the final values shown in table 2 (see figure 2 for spectrograms of these stimuli).

Table 2. Ending frequencies of the formant transition for the three context syllables.

	[ɑl]	[ɑr] bunched	[ɑr] retroflex
F1	450	425	425
F2	950	900	1250
F3	2900	1200	1550
F4	3550	3300	2800

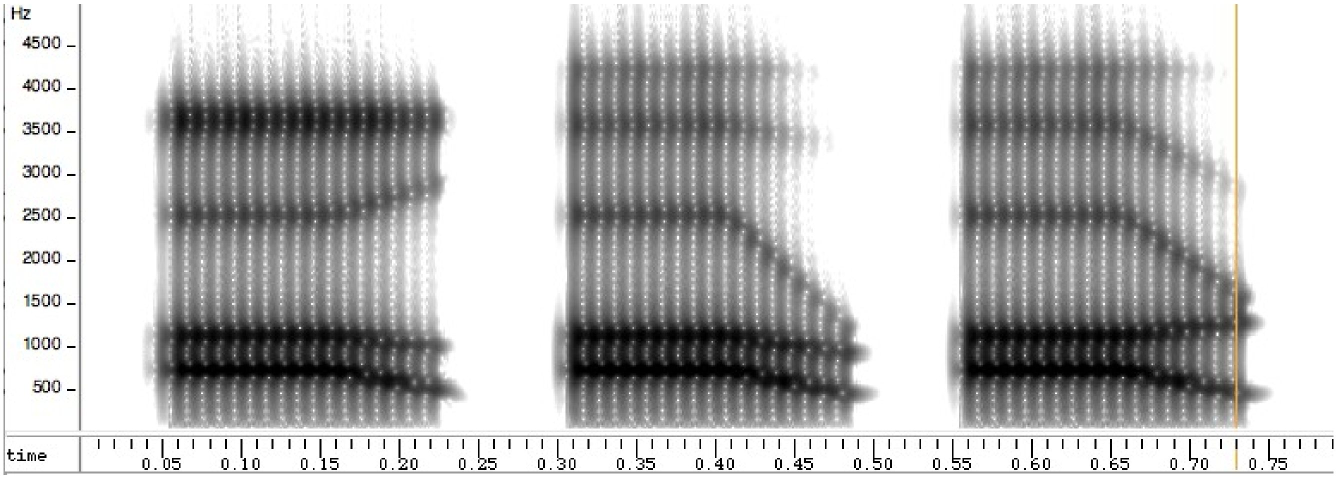

Figure 2. Spectrograms of the [ɑl], [ɑr] bunched, and [ɑr] retroflex context stimuli.

3. Procedure

Listeners in both groups heard each of the 18 stimuli (2 VC contexts by 9 CV test tokens) in randomized order 7 times (for a total of 126 trials). Listeners were instructed to identify the initial stop consonant in the target syllable using the “1” and “0” keys on a computer keyboard. For instance, a listener would hear a sequence like [ɑl dɑ] and respond by pressing the “1” key, or hear a sequence like [ɑl gɑ] and respond by pressing the “0” key. Each trial took about 2 seconds.

B. Results

The average identification contours are shown in figure 3. The 50% cross-over boundaries in these identification functions were found separately for each subject, using linear interpolation between tokens spanning the 50% cross-over. In cases where the identification function crossed 50% more than once (i.e. when the function was relatively shallow and oscillated around 50%) each 50% cross-over location in the function was found and these were then averaged to give a single estimate of the cross-over point. Multiple cross-overs occurred in 17% (11/64) of the individual identification functions. For three listeners there was no 50% cross-over in the [ɑl] context – all of the tokens in the [dɑ]-[gɑ] continuum were identified as “ga” more than 50% of the time. Two of these three subjects were in the retroflex [r] group and one was in the bunched [r] group. One subject in the bunched [r] group also did not have a 50% cross-over in the [ɑr] context – again identifying the tokens as [gɑ] most of the time.

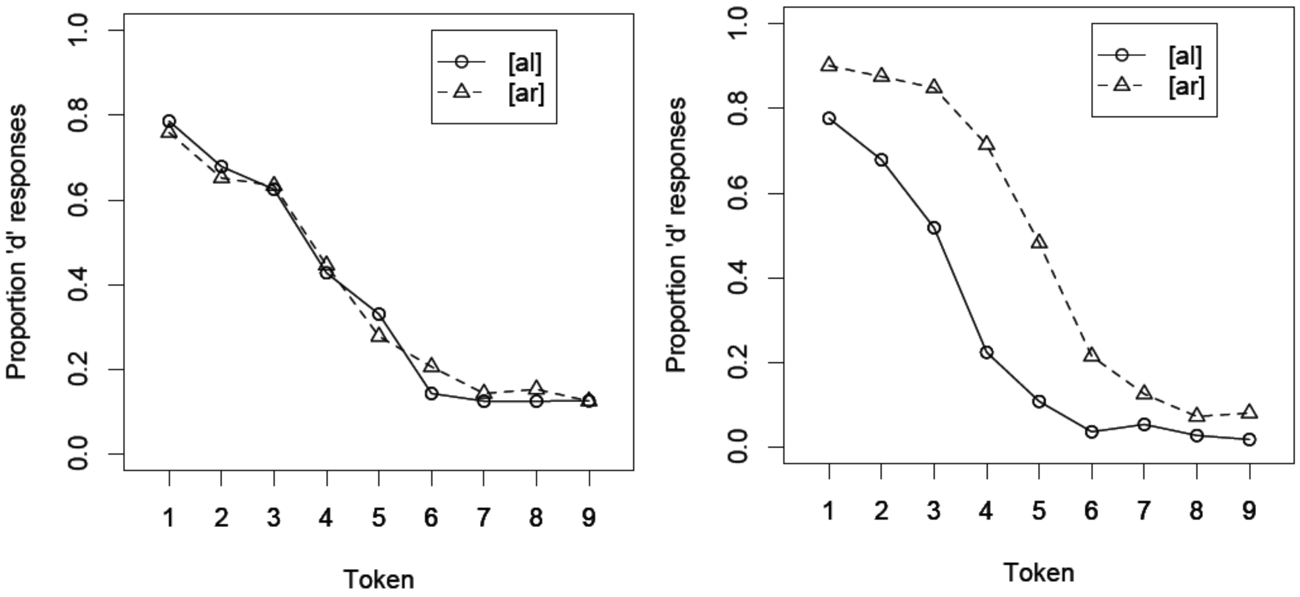

Figure 3. The average identification responses of group one (bunched [r] context) are shown in the graph on the left, and responses of group two (retroflex [r] context) are shown on the right. In both graphs responses in the [ɑl] context are plotted with open circles and a solid line, while responses in the [ɑr] context are plotted with open triangles and a dashed line.

For group one, who heard the bunched [r] context token, the [d]-[g] boundary in the [ɑl] context was at token number 3.79, and the [d]-[g] boundary in the bunched [ɑr] context was 3.84. This difference was no larger than chance [t (13)= 0.24, p > 0.8].

For group two, who heard the retroflex [r] context token, the [d]-[g] boundary in the [ɑl] context was 3.02 and the [d]-[g] boundary in the retroflex [ɑr] context was 4.79. This difference was significant in a paired comparison [t(13)= 6.83, p < 0.01].

The boundaries for the two [ɑl] contexts were marginally different from each other in an independent samples t test [t (25.7) = 1.915, p = 0.067] with a lower boundary for the retroflex [r] group (3.02) than for the bunched [r] group (3.79). The difference between the boundaries for the [ɑr] contexts was significant in an independent samples test [t (26.4) = 2.67, p = 0.013], with a higher boundary for the retroflex [r] group (4.79) than for the bunched [r] group (3.84).

The identification data were also analyzed by taking the average proportion of “d” responses across the continuum (the normalized area under the identification curve). This analysis does not rely on assumptions about how to find the identification cross-over boundary and can be sensitive to effects in overall tendency to use one or the other identification label.

The average proportion of “d” responses for group one was 0.37 when the context stimulus was [ɑl] and it was 0.38 when the context stimulus was bunched [ɑr]. This difference, like the boundary location difference, was no larger than chance [t(15)=0.08, p > 0.9]. For group two, the average proportion of “d” responses was 0.27 in the [ɑl] context and 0.48 in the retroflex [r] context – a difference that mirrored the boundary difference in being reliable [t(15) = 6.37, p < 0.01].

As with the boundary analysis, the analysis of the area under the identification curve found a tendency for the two groups of listeners to differ in both the [ɑl] context effect [t(27.1)= 2.18, p = 0.038] and the [ɑr] context effect [t(29.1)=2.48, p = 0.019].

C. Discussion

Experiment 1 found that the retroflex [ɑr] token caused a perceptual boundary shift while the bunched [ɑr] token did not. The total lack of a compensation for coarticulation effect for the bunched [ɑr] group of listeners is frankly quite puzzling because the F3 in this context token had a quite low frequency. Given prior research with nonspeech stimuli supporting a spectral contrast view of compensation for coarticulation (Lotto & Kluender, 1998; Holt & Lotto, 2002), I expected that both the bunched [r] and the retroflex [r] would cause a boundary shift, with perhaps a smaller effect in the retroflex case. F3 in the bunched [ɑr] token fell to a quite low frequency in this experiment 1200 Hz (Holt & Lotto, 2002, for example used an [ɑr] token with an F3 endpoint of 1600 Hz) so the magnitude of the F3 interaction may be dependent on the proximity of F3 in the context and the target. Further research to evaluate a spectral contrast explanation of the present findings is obviously needed, but on the whole, the findings aren't compatible with a spectral contrast explanation of compensation for coarticulation.

Although these [r] variants sound similar to each other for English listeners, and are reliably identified as ending in an [r] consonant (see the appendix), they do differ in a number of acoustic parameters and are noticeably different in an AX discrimination experiment (also in the appendix). Experiment 2 was designed to determine whether the main difference between the stimuli in experiment 1 – the ending frequency of F4 – would be enough to produce a compensation for coarticulation effect holding all other parameters constant. I was interested in whether a constellation of acoustic cues that might be associated with retroflexion would be associated with the boundary shift, or if one main acoustic characteristic in primarily involved.

II. EXPERIMENT 2

Experiment 1 used synthesis parameters that were taken from an MRI study of American English bunched and retroflex [r] (Zhou et al., 2008). The values of F1-3 were roughly comparable to each other in the bunched and retroflex variants, with a difference in F4 being the largest, most obvious difference. In particular, F3 was lower in the bunched variant than in the retroflex variant, which provided an interesting test of the auditory frequency contrast theory of compensation for coarticulation. In experiment 2 we removed all differences between the bunched and retroflex variants, except for the frequency pattern of F4. In the [ɑr] context token for one group of listeners, F4 had the higher frequency found by Zhou et al. in bunched [r]. In the [ɑr] token for a second group of listeners, F4 had the frequency found by Zhou et al in retroflex [r]. These two context tokens (higher F4 and lower F4) both sounded like a retroflex [r] (see the appendix).

The purpose of experiment 2 was to isolate the acoustic factors involved in the results found in experiment 1. We found in that experiment that an [ɑr] context token modeled on a bunched [r] pronunciation produced no boundary shift at all compared with an [ɑl] context token. One possible interpretation of the result is that the main acoustic difference between the “bunched” and “retroflex” variants of experiment 1, namely the final frequency of F4, was primarily responsible for the result. If this is so, then perhaps a modified version of the spectral contrast theory should be adopted to explain the result. Therefore, in experiment 2, all acoustic parameters were held constant except for the frequency pattern of F4 in the context [ɑr] syllables.

A. Method

1. Participants

There were two groups of participants. The first group was composed of 16 subjects (9 women and 7 men). They were students at UC Berkeley aged 18-20. This group completed a compensation for coarticulation experiment that contrasted [ɑl] and an [ɑr] token that had a high F4 offset. Three of these listeners began learning English after the age of 10 (one native speaker of Japanese who moved to Los Angeles at the age of 16, one speaker of Japanese who moved to Costa Mesa, CA at the age of 18, and one speaker of Mandarin who moved to California at the age of 10). Their data were not substantially different from the other listeners and so were retained. The second group was composed of 16 subjects (11 women and 5 men). They were students at UC Berkeley aged 19 – 24. These subjects completed a compensation for coarticulation experiment like the one given to group one except that the [ɑr] context token was synthesized with a low F4 offset. One of these listeners began learning English at the age of 9, having learned Taiwan Mandarin first. Subjects had native competence in English and normal hearing. They were paid a nominal sum ($5) for their participation.

2. Stimuli and Procedure

The only change from experiment 1, was that the two [r] context stimuli differed only in the final frequency of F4. All other parameters were taken from the “retroflex [r]” context stimulus that was used in experiment 1. In fact the “lower F4 [ɑr]” context stimulus used in this experiment was acoustically identical to the “retroflex [ɑr]” context stimulus used in experiment 1. The other [r] context token had the vowel formant trajectories of the retroflex [r] for formants 1 through 3, and the F4 trajectory of the experiment 1 bunched [r] (see table 2 for the values and Figure 4 for spectrograms of the context syllables. The [dɑ]-[ɡɑ] continuum for this experiment was the same one used in experiment 1.

The procedure was the same as in experiment 1.

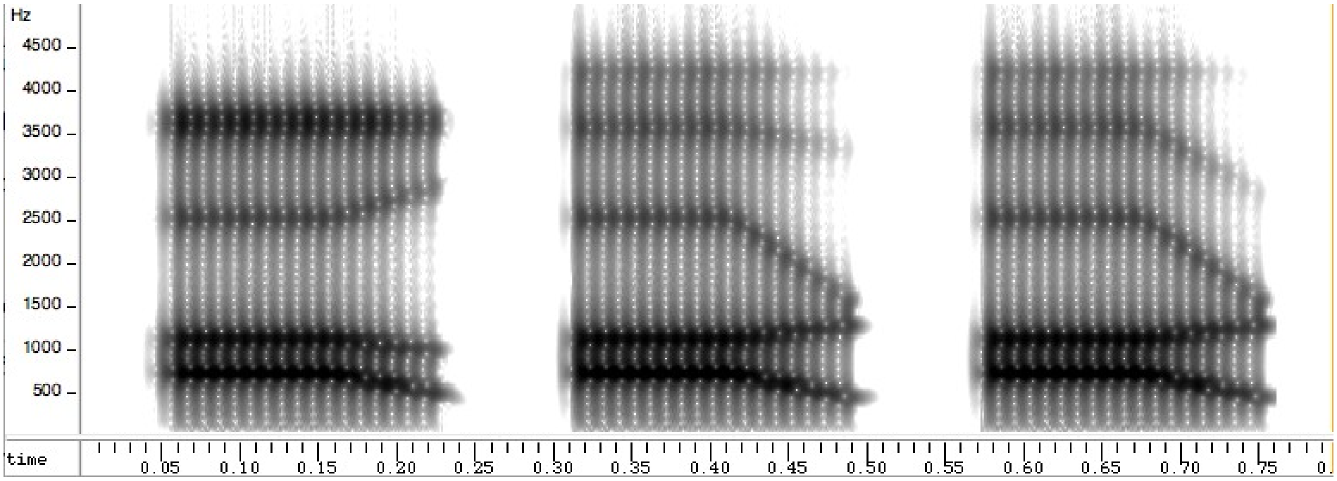

Figure 4. Spectrograms of the context syllables used in experiment 2. The left-most syllable is [ɑl], the middle syllable is the higher F4 [ɑr] token, and the right-most syllable is the same retroflex token used in experiment 1 (here called the lower F4 [ɑr] context token).

B. Results

Average identification functions for the higher F4 and lower F4 [ɑr] groups are shown in figure 5. Analysis of the data in this experiment followed the same procedure that was used in experiment 1 – with an analysis of 50% cross-over boundaries, and then an analysis of the area under the identification functions.

The 50% cross-over boundaries were found for each subject, using linear interpolation between tokens at the 50% cross-over. In cases where the identification function crossed 50% more than once (i.e. when the function was relatively shallow and oscillated around 50%) each 50% cross-over location in the function was found and these were then averaged to give a single value estimate of the cross-over point. Multiple cross-overs occurred in 14% (9/64) of the individual identification functions. For two listeners in the higher F4 [ɑr] group there was no 50% cross-over in the [ɑl] context – all of the tokens in the [dɑ]-[gɑ] continuum were identified as "ga" more than 50% of the time. One subject did not have a 50% cross-over in the [ɑr] context, and one did not have a boundary with the [ɑl] context.

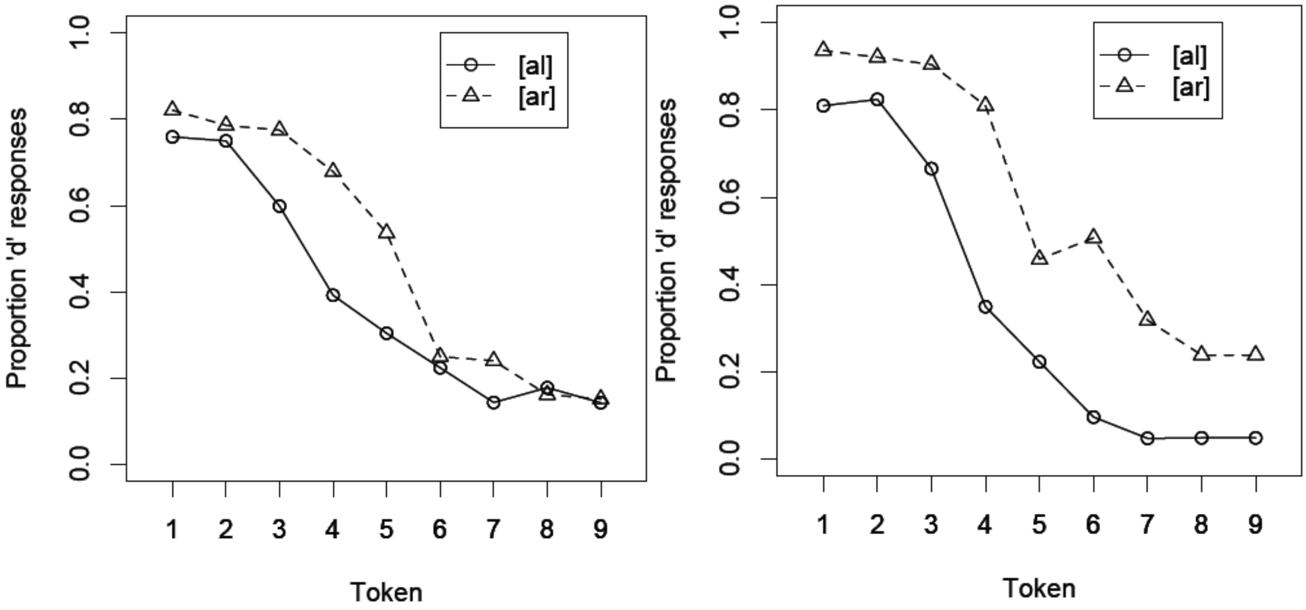

Figure 5. The average identification responses of group one (bunched [r] context) are shown in the graph on the left, and responses of group two (retroflex [r] context) are shown on the right. In both graphs responses in the [ɑl] context are plotted with open circles and a solid line, while responses in the [ɑr] context are plotted with open triangles and a dashed line.

For the listeners in the higher F4 [ɑr] group, the average boundary was located at token 3.9 in the [ɑl] context, and was at 4.86 in the [ɑr] context. This difference was marginally significant [t(13) = 2.37, p = 0.034]. The context effect was larger for listeners in the lower F4 group with the [ɑl] boundary at 3.5 and the [ɑr] boundary at 5.7. This boundary difference was significant [t(8) = 6.39, p < 0.01].

Independent samples t-tests comparing the groups showed that the [ɑl] boundaries were not reliably different [t(18.16) = 0.74] while the difference between the [ɑr] boundaries was marginally significant [t(21.2) = 1.48, p = 0.15]. Because the data are paired (with an [ɑl] boundary and an [ɑr] boundary for each listener, it was possible to compare the magnitude of the compensation for coarticulation effect more directly. The mean size of effect for the listeners in the higher F4 group was 1.19 step on the continuum, while the effect was a shift of 2.2 steps for the lower F4 group. This difference was only marginally reliable [t(20.7) = 1.67, p = 0.11].

As in experiment 1, the identification data were also analyzed by taking the sum of "d" responses across the continuum (the normalized area under the identification curve) as a measure of context effect. This analysis does not rely on assumptions about how to find the identification cross-over boundary and can be sensitive to effects in overall tendency to use one or the other identification label.

The average proportion of "d" responses for the higher F4 group was 0.39 when the context stimulus was [ɑl] and it was 0.49 when the context stimulus was bunched [ɑr]. This difference, like the boundary location difference, was larger than chance [t(15)=3.13, p < 0.01]. For the lower F4 group the average proportion of "d" responses was 0.35 in the [ɑl] context and 0.59 in the retroflex [r] context – a difference that mirrored the boundary difference in being reliable [t(8) = 5.33, p < 0.01]. Independent samples t-tests found that the groups did not significantly differ in the [ɑl] contexts [t(22.9) = 0.96, p = 0.35], but did tend to differ for the sum of "d" responses in the [ɑr] context effect [t(15.8)=2.03, p = 0.059].

C. Discussion

This experiment found that altering an acoustic cue for retroflexion (the ending frequency level of F4) modulates the compensation for coarticulation effect. When the F4 trajectory was more like that found in retroflex [r] the compensation was greater than when the F4 trajectory was more like the one found in bunched [r]. This indicates that the compensation effect in speech perception is not solely due to acoustic F3 contrast between the context and target tokens. Both [ar] contexts in this experiment sounded like they had a retroflex [r] and the difference between them was hardly noticeable to listeners. So, it is interesting that both tokens produced a compensation effect but that the token with the more typical retroflex pattern caused the stronger effect.

III GENERAL DISCUSSION

To summarize, building on research by Viswanathan, et al. (2010), we found in experiment 1 that compensation for coarticulation is strongly present when the context [ar] token is synthesized with a retroflex [r] and is completely absent when the [ar] context is synthesized with a bunched [r]. This appears to support an interpretation of compensation for coarticulation that has a role for gesture perception because the tongue-tip is retracted during retroflex [r] and thus will cause greater coarticulation on a following alveolar stop consonant. That is, the perceptual compensation happens for only those [ɑr] tokens where coarticulation is likely to happen. We speculated that detailed acoustic properties of the stimuli may leave an opening for a spectral contrast account of these data, but that further research is necessary.

In experiment 2 we found that even a very minor acoustic property in the [ɑr] context token, a property that is associated with the difference between retroflex and bunched [r], can modulate the size of the compensation for coarticulation effect. The spectral contrast theory has very little to say about this result because F4 doesn't vary at all in the [dɑ]/[ɡɑ] continuum in these experiments (table 1), and until now F4 has not been a part of the discussion on the spectral contrast causes of compensation for coarticulation. The fact that F4 trajectory is a cue for the difference between retroflex and bunched [r] suggests that any acoustic property that signals coarticulation will have an impact on compensation for coarticulation.

The natural conclusion of this study is thus that any explanation of compensation for coarticulation that fails to include a role for the use of articulatory knowledge during speech perception is missing a key component. However, a more comprehensive view of the range of research on this widely studied topic must also admit that compensation for coarticulation is a perceptual phenomenon that shows the imprint of several intersecting factors. These include (1) auditory contrast at the peripheral (Sitek & Johnson, 2011; though see Holt & Rhode, 2000), and central auditory levels (Holt & Lotto, 2002; Lotto, Sullivan & Holt, 2002; Lotto & Kluender, 1998), (2) the recruitment of articulatory knowledge during speech perception (Viswanathan et al., 2010; Fowler, 2006; Mitterer, 2006), and (3) language-specific experience with the speech sounds involved (Samuel & Pitt, 2003; Beddor, et al., 2002; Martínez et al., 2003; Johnson, in preparation). The challenge facing current research is to develop a theory of how these perceptual factors interact with each other.

ACKNOWLEDGEMENTS

Many thanks to my research assistants (Alex Beagle, Demi Chang, Greg Finley, Katherine Hall, and Kevin Sitek) who helped to collect the data reported here. Thanks also to my colleagues (including, but not limited to, Pam Beddor, Ian Maddieson, James Morgan, Meghan Sumner, Doug Whalen) for their comments and suggestions at various conference presentations and colloquium talks. This research was partly supported by NIH Grant

REFERENCES

Beddor PS; Harnsberger JD; Lindemann S (2002) Language-specific patterns of vowel-to-vowel coarticulation: acoustic structures and their perceptual correlates. *J. Phonetics 30(4)*, 591-627.

Fowler, C.A. (2006) Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics* **68**(2), 161-177.

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

Fowler, C.A., Brown, J.M. & Mann, V.A. (2000) Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *J. Exp. Psych: Human Percept. & Perform. 26*, 877-888.

Holt, L. L. (1999). Auditory constraints on speech perception: An examination of spectral contrast. Unpublished doctoral dissertation, University of Wisconsin–Madison.

Holt, L. L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science*, *16*, 305-312.

Holt, L. L. (2006). The mean matters: Effects of statistically-defined non-speech spectral distributions on speech categorization. *Journal of the Acoustical Society of America, 120*, 2801-2817.

Holt, L.L. & Lotto, A.J. (2002) Behavioral examination of the neural mechanisms of speech context effects. *Hear. Res. 167*, 156-169.

Holt, L.L., Lotto, A.J. & Kluender, K.R. (2000) Neighboring spectral content influences vowel identification. *J. Acoust. Soc. Am. 108*, 710-722.

Klatt, D.H. (1980) Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am. 67(3)*, 971-995.

Klatt, D.H. & Klatt, L.C. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am. 87(2)*, 820-857.

Ladefoged, P. & Johnson, K. (2010) *A Course in Phonetics*. Boston, Wadsworth/Cengage.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America, 102*, 1134-1140.

Lotto, A.J. & Kluender, K.R. (1998) General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. Perception & Psychophysics *60*, 602-619.

Lotto, A.J., Sullivan S.C. & Holt, L.L. (2003) Central locus for nonspeech context effects on phonetic identification (L). *J. Acoust. Soc. Am. 113(1)*, 53-56.

Mann, V.A. (1980) Influence of preceding liquid on stop-consonant perception. *Perception and Psychophysics 28(5)*, 407-412.

Martínez, Silvia, Susanna Padrosa, Irene Pascual, Andrea Pearman, Laura Riera and Susagna Tubau. (2003) The Role of Experience in the Perception of Coarticulated Speech: An Empirical Study. In *Fifty Years of English Studies in Spain* […] *Actas del XXVI Congreso de AEDEAN,* ed. Ignacio Palacios et al. Santiago de Compostela: U de Santiago de Compostela. pp. 599-605.

Mielke, J., Baker, A., & Archangeli, D. (2007) Variability and homogeneity in American English /ɹ/ allophony and /s/ retraction. Laboratory Phonology 10.

Samuel, A.G. & Pitt, M.A. (2003) Lexical activation (and other factors) can mediate compensation for coarticulation. *J. Memory & Language 39*, 347-370.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (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*, 1005-1015.

Westbury, J.R., Hashi, M. & Lindstrom, M.J. (1998) Differences among speakers in lingual articulation for American English /ɹ/. *Speech Communication 26*, 203-226.

Zhou, X., Espy-Wilson, C.Y., Tiede, M., Boyce, Holland, S.C. and Choe, A. (2008) A magnetic resonance imaging-based articulatory and acoustic study of “retroflex” and “bunched” American English /r/, *J. Acoust. Soc. Am.*, *123(6)*, 4466-4481.

APPENDIX – PERCEPTION OF THE CONTEXT TOKENS

The key manipulation in the experiments reported in this paper was of the acoustic properties of the [ɑr] context tokens. This short appendix describes a control study in which the context tokens were presented for identification and discrimination. I was mainly interested in knowing whether listeners could hear the difference between the retroflex and bunched [ɑr] context syllables of experiment 1 and between the higher and lower F4 [ɑr] context syllables in experiment 2. Additionally, it was important to know whether the syllables described as ending in [r] would actually be labelled as “r” by listeners.

So, the six context syllables that were used in experiments 1 (three context syllables) and 2 (three context syllables) were presented to listeners for identification (5 listeners) and discrimination (15 listeners) responses. The listeners in these control studies were drawn from the same population as in the main experiments, and were screened for normal hearing and native English ability, and were paid a small sum for their participation. In the identification test, the six stimuli were presented 5 times each and the 30 trials were presented in a unique random order for each participant. Listeners were asked to identify the token as “are” or “all”. In the discrimination test, the six stimuli were crossed to produce 36 pairs and each pair was presented 4 times. The resulting 144 trials were presented in a unique random order for each participant. In the discrimination paradigm, the listener saw a feedback screen after every trial that reported whether the correct answer had been given. “Same” responses were only deemed correct if the pair involved the presentation of the same sound file. That is, listeners were required to distinguish among the different kinds of [ɑr] stimuli.

Table A1. Identification results – percent "all" responses.

	al	arB	arR
Experiment 1	97	5	1
	al	high F4 ar	low F4 ar
Experiment 2	97	3	1

The results of the identification test are shown in table A1. Listeners identified the [ɑl] tokens as "all" and the [ɑr] tokens, whether retroflex or bunched, as "are".

The results of the discrimination test are shown in table A2. The false alarm rate, i.e. the percent "different" responses to acoustically identical stimuli, was about 12% on average. The retroflex and bunched [ɑr] stimuli used in experiment 1 (labeled in this table "arB" and "arR" were distinguished at a much higher rate (81%). This is not directly relevant to the results of experiment 1 because the retroflex and bunched context tokens were presented to different groups of listeners, and so were never presented in contrast with each other. The higher F4 and lower F4 [ɑr] tokens in experiment 2 (labeled in this table "arH" and "arL") were labeled "different" at a rate (13.5%) that is not substantially different from the false alarm rate (12%). However, in another pairing of the same stimuli (the higher F4 stimulus of experiment 2 (arH) with the retroflex stimulus of experiment 1 (arR) the difference may have been slightly noticeable (19% "different").This reflects the much smaller acoustic differences between the two stimuli than between the arB and arR stimuli of experiment 1. Interestingly, both arH and arL were quite reliably distinguished from the bunched [r] stimulus of experiment 1 (81% and 83% respectively versus arB).

Table A2. Discrimination results – percent of correct "different" responses.

	al	arB	arR	al	arH	arL
al	13	97	99.5	8	99	97.5
arB		13	81	95.5	81	83
arR			14	99	19.5	12.5
al				7	98.5	97.5
arH					13	13.5
arL						14

The main results of the control study are: (1) the stimuli we called "ar" in the paper were identified as such by listeners, (2) the retroflex and bunched stimuli of experiment 1 are noticeably different, while the acoustic difference between the higher F4/lower F4 stimuli of experiment 2 was not very salient, and (3) the [ɑr] context stimuli of experiment 2 sound more like the retroflex stimulus of experiment 1, than like the bunched [ɑr] stimulus.