Perception & Psychophysics
2006, 68 (2), 161-177

Compensation for coarticulation reflects gesture perception, not spectral contrast

CAROL A. FOWLER
Haskins Laboratories, New Haven, Connecticut
and University of Connecticut, Storrs, Connecticut

This article reports three experiments designed to explore the basis for speech perceivers' apparent compensations for coarticulation. In the first experiment, the stimuli were members of three /da/-to-/ga/ continua hybridized from natural speech. The monosyllables had originally been produced in disyllables /ada/ and /aga/ to make Continuum 1, /alda/ and /alga/ (Continuum 2), and /arda/ and /arga/ (Continuum 3). Members of the second and third continua were influenced by carryover coarticulation from the preceding /l/ or /r/ context. Listeners showed compensation for this carryover coarticulation in the absence of the precursor /al/ or /ar/ syllables. This rules out an account in which compensation for coarticulation reflects a spectral contrast effect exerted by a precursor syllable, as previously has been proposed by Lotto, Holt, and colleagues (e.g., Lotto, Kluender, & Holt, 1997; Lotto & Kluender, 1998). The second experiment showed an enhancing effect of the endpoint monosyllables in Experiment 1 on identifications of preceding natural hybrids along an /al/-to-/ar/ continuum. That is, coarticulatory /l/ and /r/ information in /da/ and /ga/ syllables led to increased judgments of /l/ and /r/, respectively, in the precursor /al/-to-/ar/ continuum members. This was opposite to the effect, in Experiment 3, of /da/ and /ga/ syllables on preceding tones synthesized to range in frequency from approximately the ending *F*3 of /ar/ to the ending *F*3 of /al/. The enhancing, not contrastive, effect in Experiment 2, juxtaposed to the contrastive effect in Experiment 3, further disconfirms the spectral contrast account of compensation for coarticulation. A review of the literature buttresses that conclusion and provides strong support for an account that invokes listeners' attention to information in speech for the occurrence of gestural overlap.

Talkers temporally overlap the phonetic gestures of speech, a behavior known as *coarticulation* or *coproduction*. Gestures characteristically begin in the domain of other, earlier gestures (*anticipatory* coarticulation) and characteristically end in the domain of later gestures (*carryover*, or *perseveratory*, coarticulation). This renders the acoustic speech signal highly context sensitive. However, remarkably, listeners do not, in general, hear it that way. Rather, they *compensate for coarticulation*. To take a much-studied example, in natural speech, constriction gestures of /l/ and /r/ of a precursor syllable, /al/ or /ar/, can carry over into those of a following /d/ or /g/ gesture. One outcome can be that the point of constriction achieved during closure for /d/ and /g/ reflects a blending of their own constriction gestures with those of /r/ or /l/. The pharyngeal constriction for /r/ may pull the point of articulation during the stop consonants back. The tongue tip constriction for /l/ may pull the point of constriction during /g/ forward. In addition, /r/ is a rounded consonant, and the rounding constriction gesture carries over into the following syllable, further adding to the context sensitivity of the acoustic signals during /da/ and /ga/. However, Mann (1980) showed that listeners compensate for the context sensitivity. She found that identification of ambiguous members of a synthetic /da/-to-/ga/ continuum as /da/ was increased when the precursor syllable was /ar/, as opposed to when it was /al/. The backing and rounding gestures of /r/ will lower the high *F*3 for /da/; fronting of /g/ by coarticulation with /l/ will raise *F*3 during /ga/. By reporting more /da/ responses for ambiguous members of the continuum following /ar/, listeners behaved as if they had ascribed the relatively low *F*3 for /da/ to the lowering effect of the /r/ context. By identifying some of those same syllables as /ga/ in the context of /al/, they behaved as if they had compensated for the fronting effects of /al/ on /ga/.

There are other findings of compensation as well. For example, listeners identify more of the ambiguous members along a /ta/–/ka/ continuum as /t/ when they follow /ʃ/ than when they follow /s/ (Mann & Repp, 1981). This may reflect compensation for the backing effects that /ʃ/ should have on /t/. As a final example, Mann and Soli (1991) found compensation for rounding effects of the vowel /u/ in identification of final fricatives as /s/ or /ʃ/.

Preparation of the article was supported by Grants HD-01994 and DC-02717 to Haskins Laboratories. I thank Brook Swainson for making the stimuli in Experiment 1 and Doug Honorof both for serving as the second talker in Experiment 1 and for providing expertise for the acoustic analyses reported in Table 1. I thank Justin Bates, who asked to be identified as the model speaker employed in the research in Fowler, Brown, and Mann (2000) and in Holt, Stephens, and Lotto (2005). His image appears in Figures 1 and 2 of the present article. Correspondence concerning this article should be addressed to C. A. Fowler, Haskins Laboratories, 300 George Street, New Haven, CT 06511 (e-mail: carol.fowler@haskins.yale.edu).

At issue is how perceptual compensation comes about. There are five kinds of account that have been put forward, just three of which are of central relevance here. Compensation for coarticulation can be lexically mediated (Elman & McClelland, 1988; Samuel & Pitt, 2003) or mediated by knowledge of more and less frequent phonotactic patterns in the language (Pitt & McQueen, 1998). However, these effects are unlikely to be important in the present study. The experiments under review and the ones that I present here for the first time do not use words for stimuli. Fowler, Brown, and Mann (2000) claimed to have ruled out a phonotactic account of their findings, which we will pursue here.

Of the remaining three accounts of perceptual compensation, one is consistent with the motor theory of speech perception (cf. Mann, 1980). Motor theorists have proposed that speech perception is achieved by a specialization of the brain, a phonetic module (e.g., Liberman & Whalen, 2000). The module incorporates motor competence regarding speech, and specifically, it incorporates knowledge of coarticulation and its acoustic consequences (e.g., Liberman & Mattingly, 1985). By this account, compensation for coarticulation occurs when listeners make use of motor competence in the perceiving of speech. Listeners perceive the intended phonetic gestures of the speaker, using the phonetic module's knowledge of coarticulation and its acoustic consequences. To perceive intended gestures, listeners must track gestures separately during intervals in which coarticulation causes acoustic consequences that blend information for the overlapping gestures. They parse the acoustic signal along gestural lines. In the case of Mann's /arda/–/arga/ disyllables, that means ascribing the $F3$-lowering consequences of /r/'s gestures during /d/ and /g/ to the gestures of /r/. This gives compensation for coarticulation. In the motor theory, tracking gestures requires a process of analysis by synthesis, using an "innate vocal tract synthesizer" (Liberman & Mattingly, 1985).

A second account is provided by direct realist theory (e.g., Fowler, 1986, 1996). On that account, listeners use acoustic structure in speech utterances as information for the causal sources of that structure—namely, the phonetic gestures that produced the signal. They compensate for coarticulation precisely because they track gestures and, so, parse the signal along gestural lines. For example, any acoustic consequences of the backing effect of /r/ in the domain of the constriction gesture for /d/ is ascribed to /r/, not to /d/. In respect to perceivers' use of structure in proximal stimulation to perceive properties of the world of distal events, speech perception reflects a presumed universal function of perceiving in the world—that is, of perceiving components of the ecological niche, not of proximal stimulation at the sense organs. In contrast to the motor theory, direct realist theory supposes that acoustic structure specifies its gestural sources, obviating the need for analysis by synthesis and motor involvement in speech perception.[1]

Whereas the motor theory of speech perception invokes a brain mechanism that is special to speech and direct realism invokes the supposed universal function that perception serves, the third account, spectral contrast, invokes a very general perceptual process that applies to any acoustic signal with the right properties. In addition, its proponents (e.g., Coady, Kluender, & Rhode, 2003; Lotto & Kluender, 1998; Lotto, Kluender, & Holt, 1997) suggest that spectral contrast is one of very many perceptual processes that yield contrast across the perceptual modalities. It offers a general solution to a general perceptual problem, although the nature of that general problem remains an object of speculation. Spectral contrast refers to a process whereby input to the auditory system temporarily renders the system less sensitive than it otherwise is to the spectral properties of that input. For example, the very low ending $F3$ of /ar/ renders the auditory system temporarily relatively insensitive to frequencies near that $F3$. Therefore, an $F3$ just above that for /ar/ is perceived to be composed only of the higher of its frequencies. A syllable ambiguous between /da/ and /ga/ that follows /ar/ thereby sounds more like /da/ than it does presented in isolation or after /al/, with its high $F3$. More generally, as Lotto and Kluender put it,

> Contrast may be a rather general solution to the effects of phonemic context on identification. Coarticulation tends to be assimilative, and contrastive processes could compensate for much of the lack of invariance in speech acoustics due to articulatory dynamics. (p. 615)

Spectral contrast will reduce or eliminate the assimilatory acoustic consequences of coarticulation, thereby compensating for it.

It will become clear in the following literature review that none of the three accounts of compensation for coarticulation explains all of the research findings deemed relevant by at least a subset of the theorists who have contributed to this literature. However, the spectral contrast account stands out in this regard. I will suggest that the spectral contrast account handles very little of the relevant literature and, indeed, that it was ruled out on empirical grounds well before it was first proposed. A more general auditory account of compensation for coarticulation that invokes spectral contrast to explain just a subset of the findings requires an eclectic set of explanations for a coherent set of findings. Both the motor theory and direct realism provide a single account that handles the set of findings.

A series of three experiments underscores the inadequacy of the contrast account to handle critical research findings.

Findings Consistent With the Contrast Account That Are Not Predicted by the Other Two Theories

Especially consistent with the contrast account are largely findings that, in some cases, coarticulatory contexts can be replaced by nonspeech sounds without eliminating the compensation-like effects, even though the nonspeech sounds have no phonetic properties. For example, in one experiment, Lotto and Kluender (1998) replaced the /al/ and /ar/ of disyllables similar to those in

Mann (1980) with sine wave tones at the ending F3 of /al/ or /ar/. They got a significant context effect qualitatively like that obtained with /al/ and /ar/ as precursor syllables. That is, more /da/ responses were given to ambiguous syllables along a /da/-to-/ga/ continuum after the low tone than after the high tone. Fowler et al. (2000) replicated this effect and showed that it was contingent on the energy relations between the tones and the critical F3s of the /da/–/ga/ syllables.

Holt, Lotto, and Kluender (2000) performed a conceptual replication of earlier findings of Lindblom and Studdert-Kennedy (1967) and of Williams (1986). In that earlier research, vowels were more likely to be identified as /ɪ/ (rather than /ʊ/) in a /w/–/w/ context than in a /j/–/j/ context. Lindblom and Studdert-Kennedy suggested that this was compensation for coarticulation-based vowel undershoot ("vowel recognition thus compensates for vowel production"; p. 842), invoking a process of analysis by synthesis. A direct realist account is that listeners ascribe acoustic evidence of rounding in the vowel to rounded /w/ in the /w/–/w/ context. Holt et al. (2000) speculated, rather, that the effect was due to spectral contrast. The rising F2 transition of the initial labial consonant induced a contrast effect whereby the F2 of the vowel sounded higher and more /ɪ/-like than it was. To test this idea, they synthesized vowels ranging from /ɛ/ to /ʌ/ in either a /b/–/b/ context (with initial rising F2 and final falling F2) or in a /d/–/d/ context (falling F2 initially, rising finally). More /ɛ/ responses occurred in the labial context, as if the rising transitions of initial /b/ caused the F2 of the vowel to sound higher and, so, more like /ɛ/. Next, they replaced the flanking consonants with single sine waves tracking the center frequencies of the consonants' F2 transitions. This also led to more judgments that the vowel was /ɛ/ in the context of the rising sine waves than in the context of the falling ones.

Recently, Stephens and Holt (2003) have shown a complementary result to the finding that nonspeech contexts can replicate the effects of phonetic ones. In that instance, the phonetic precursor syllables were /al/ and /ar/; they were followed either by members of a /da/–/ga/ continuum or by transitions, heard as nonspeech, that tracked the frequencies of the F2 and F3 of members of the /da/–/ga/ continuum. Stimuli were presented in pairs for a same–different judgment. On each AX trial, if the precursor syllable of A was /al/, that of X was /ar/, and vice versa. Same–different judgments were to be based on the following /da/–/ga/ syllables or the transitions.

Precursor syllables were paired with members of the two continua in each of two ways. In one condition, a contrast effect of the precursor syllable should enhance the distinctiveness of the following syllables or transitions. That is, /al/ was paired with the more /ga/-like syllable or transition and /ar/ with the more /da/-like one. In the other condition, the pairing was the opposite, which should reduce the discriminability of the members of an AX pair of /da/–/ga/ syllables or of F2 and F3 transitions. The findings were that discrimination performance was better in the first condition than in the second, both when syllables were discriminated and when transitions were discriminated.[2]

Findings such as these are not predicted by the motor theory, because the speech module, if any, is not expected to process most sine waves (but see, e.g., Remez, Rubin, Pisoni, & Carrell, 1981). They are not predicted by direct realist theory, because the sine waves do not provide information about coarticulatory overlap between phonetic gestures. I will suggest possible interpretations of these findings from the perspective of the motor theory and direct realism in the General Discussion section.

Findings Inconsistent With the Spectral Contrast Account

Findings inconsistent with the spectral contrast account fall into four categories. I will consider each in turn.

Compensation for coarticulation is not restricted to left-to-right effects; contrast is. Kluender, Lotto, and Holt have suggested, in a number of articles (e.g., Holt et al., 2000; Lotto & Kluender, 1998), that their contrast account accrues plausibility because contrast effects occur very broadly across and within perceptual modalities. It provides a general solution to a general perceptual problem, albeit one whose precise nature remains to be worked out. Contrast effects are early-to-late context effects,[3] and accounts of them, such as Warren's (1985) well-known criterion shift account, assume that they are left-to-right effects:

> Perceptual criteria are displaced in the direction of recently encountered stimuli with the more recent exemplars having a greater effect. (p. 582)

Lotto (1996; see also Lotto et al., 1997) agrees:

> Due to the variables of inertia and mass, physical systems tend to be assimilative across time. The configuration of a system at time t is significantly constrained by its configuration at time $t - 1$. The set of possible transformations from time $t - 1$ to t is very limited. Rapid change is the exception for physical systems. . . . A perceptual system that respects the assimilative nature of physical laws may emphasize these changes from stability. One way to emphasize these differences is through contrast effects. (p. 128)

Contrast effects eliminate or reduce the assimilative effects of the configuration of the system at time $t - 1$ that are manifest in the configuration of the system at time t. For example, spectral contrast might reduce or eliminate perceptual sensitivity to assimilative effects of /l/ and /r/ during /d/ and /g/ of Mann's (1980) disyllables.

Holt (1999) suggests, further, that neural adaptation may provide a plausible mechanism underlying spectral contrast. In neural adaptation, neurons having preferred firing frequencies become adapted, after responding, to input that includes their preferred frequencies. They are then less responsive to additional input having those frequencies. This is a left-to-right effect.

However, compensation for coarticulation is not restricted to compensation for carryover coarticulation. It occurs when coarticulation is anticipatory in direction and when gestures are produced concurrently.

Compensation for anticipatory coarticulation. As for anticipatory coarticulation, for example, Mann and Repp (1980) found that identification of fricatives along the /s/–/ʃ/ continuum varied depending on the following vowel. Specifically, more /s/ responses occurred in the context of their following /u/ than when they followed /a/. This is compensation for the spectrum-lowering effects that coarticulatory anticipation of lip rounding has on a preceding fricative. More recently, Mann and Soli (1991) found compensation for coarticulatory effects of vowels on fricatives both in VC syllables, where coarticulation occurred in the carryover direction, and in CVs, where the effects were anticipatory. Fowler (1984) found compensation for anticipatory coarticulation in /ɡ/V syllables.

Of course, confronted with these kinds of findings, the contrast account can be revised to permit context effects to work in both directions. However, to make this revision, theorists have to acknowledge that what they call *spectral contrast* is not one among many other contrast effects. It is exceptional in working in both directions. Nor is it a general process that emphasizes change over time; right-to-left effects work backward in time. Also, the account of contrast in terms of neural adaptation is ruled out in cases of compensation for anticipatory coarticulation.

In relation to this, any new account of compensation for coarticulation devised to accommodate compensation for anticipatory coarticulation has to explain how the perceiving mechanism knows in which direction to compensate. Why does compensation work left to right in vowel–fricatives but right to left in fricative–vowels?

There *is* a generalization that holds across all of the findings of compensation for coarticulation, but it is not a generalization that can be captured in a spectral contrast account. The generalization is that the effects of coarticulating gestures (i.e., of aggressor or encroaching gestures) are compensated for whether the domain on which they encroach is that of a serially earlier or later gesture. Compensation is a parsing of effects of the coarticulating gesture from the domain of the encroached-upon gesture.

Compensation for coarticulation that is neither anticipatory nor carryover in direction. Anticipatory and carryover coarticulations do not exhaust the kinds of coarticulation that occur. Sometimes, different gestures have converging effects on common acoustic dimensions at the same time. Silverman (1987) studied one such instance. Other things being equal, high vowels, such as /i/, are associated with higher fundamental frequencies (*f*0s) than are low vowels, such as /a/. The difference in *f*0 is called *intrinsic f*0. Why this association occurs is not known, although the literature offers a number of speculative ideas (e.g., Sapir, 1989). Silverman (1987) presented listeners with sentence pairs that included two intonational pitch accents, one on an /i/ vowel and one on an /a/ vowel—for example, "They only feast before fasting" and "They only fast before feasting." (Speakers of the sentences spoke dialects of British English, so that the vowel was /a/, not American English /æ/.) In one experiment, the first pitch accent was held fixed, and the second varied in seven steps, centered at the level of the first accent. Listeners had to judge which intonational peak was the more prominent, as indicated by intonational peak height. Silverman (1987) found that, in their judgments of peak height, listeners compensated for intrinsic *f*0. That is, to be judged equal in pitch to a pitch accent on /a/, an accent on /i/ had to be higher in *f*0. This evidence of compensation is not amenable to an account in terms of spectral contrast. One reason is that there is no left (or right) context to induce spectral contrast. Another reason will be offered next.

Not all coarticulatory effects have assimilative acoustic consequences. Spectral contrast can account only for those that do. Listeners compensate for coarticulation whether or not coarticulation effects are assimilative. The overlap of gestures that realize vowel height and gestures that realize an intonation contour does not give rise to assimilative acoustic consequences in the way that /r/, with its low *F*3, lowers the *F*3 of a following /da/. That is, coproduction of high and low vocalic gestures with intonational gestures does not generate an acoustic outcome somewhere else that is similar to the outcome of the intonational gesture or an outcome somewhere else that is similar to acoustic consequences of the vowel gestures. It has an effect that is contemporaneous with the acoustic consequences of intonational gestures and, therefore, cannot assimilate to them. A spectral contrast account can handle only instances in which there are assimilative acoustic consequences and in which those consequences occur at a different point in time from their coarticulatory cause. In the following, another example is given in which coarticulation does not cause assimilative acoustic consequences but does have consequences for which listeners compensate.

Production of an unvoiced obstruent can cause a high falling tone to occur on a following vowel. The reason for this (see, e.g., Löfqvist, Baer, McGarr, & Story, 1989) is probably that the vocal folds are tensed during closure for the consonant, in an effort to keep the vocal folds from being pulled shut by the airflow through the glottis. When that tensing gesture carries over into the following vowel, with vocal folds now adducted, *f*0 is transiently raised. This is a second example in which the acoustic consequences of coarticulation are not assimilatory. A high falling tone on a vowel does not make the vowel acoustically more like an unvoiced consonant than it is without the tone. After all, the unvoiced consonant has no *f*0; it is unvoiced. However, listeners compensate for that coarticulatory effect. They do not hear the tone as a tone on the vowel (e.g., Pardo & Fowler, 1997; Silverman, 1986). To hear vowels after voiced and voiceless obstruents as equal in pitch, the vowel following the voiceless obstruent has to be higher in *f*0 than is that following the voiced obstruent. Rather than hearing the obstruent-induced tone as pitch, listeners use the tone as information for the voicelessness of the obstruent (Pardo & Fowler, 1997).

Compensation for coarticulation occurs in other instances in which spectral contrast is ruled out. Fowler et al. (2000) reported an audiovisual version of compensation for coarticulation. They synthesized a syllable that they judged ambiguous between /al/ and /ar/.

This ambiguous syllable was a precursor to members of a /da/-to-/ga/ continuum. The VCCVs were dubbed onto both of two videos of a speaker hyperarticulating either /alda/ or /arda/. Speech was hyperarticulated to make the /l/–/r/ distinction as visible as possible. In this experiment, the only information distinguishing /al/ from /ar/ was optical. The only information distinguishing /da/ from /ga/ was acoustic. Any compensation for coarticulation in identifying /da/ and /ga/ syllables then had to be due to a *McGurk effect* (e.g., McGurk & MacDonald, 1976); it could not be due to spectral contrast. Significant compensation for coarticulation occurred.

Holt, Stephens, and Lotto (2005) have recently proposed that the critical experiment of Fowler et al. (2000) was flawed in that video frames showing /da/ articulation at /da/ or /ga/ acoustic onset were not matched. That is, the stimulus materials confounded the critical variable, whether the precursor syllable was /al/ or /ar/, with whatever visible differences in the articulation of /d/ might be present in the two video syllables. In particular, in a way that the authors do not specify, the /d/ in the /arda/ video looked subtly more like a /d/ and less like a /g/ than did that in the /alda/ video. That, rather than audiovisual compensation for coarticulatory effects of the precursor syllable, may have led listener/viewers to report more /da/s, given the video of /arda/, than when given the video of /alda/. To test their account, they presented participants with only the second syllable of each disyllable stimulus item. Their outcome was very much like that of their experiment in which both audiovisual syllables were presented. They concluded that it was not the precursor context but the subtle /d/-likeness in the /arda/ video that gave rise to the outcome. It was an ordinary McGurk effect.

The idea that the effect depended on subtle differences between the videos, so that the CV after /ar/ looked more like a /d/ than did that after /al/, is implausible on its face, because /d/ and /g/ are considered to be members of the same viseme class (see, e.g., Walden, Prosek, Montgomery, Scherr, & Jones, 1977). That is, given a video of a /d/ or /g/ articulation, viewers generally are unable to judge which articulation is being made. Accordingly, anything that would make one video look subtly more like /d/ would also make it look subtly more like /g/. The implausibility of the idea is enhanced by the investigators' apparent inability to specify what the critical subtle difference was.

More telling than the viseme consideration, however, is Figure 5 in Holt et al. (2005), showing the alleged confounding, and the present Figures 1 and 2. The videos in Figure 5 in Holt et al. (2005) show single frames at the release of the stop consonant in the CV syllables of /alda/ and /arda/ in the stimuli in Fowler et al. (2000). I reproduce those frames as the top displays in Figures 1 and 2. The figures reveal one subtle and one salient difference in the frames. The subtle difference is that the model's teeth are parted in the frame from /alda/ but closed in the frame from /arda/. Through the parted teeth, the alveolar constriction for /d/ is dimly visible (more so on the computer screen than in the figure). The bottom displays in each of Figures 1 and 2 show frames just after those in the top displays. Again, the alveolar constriction is visible in the video of /alda/, but even though the teeth are parted in the video of /arda/, no constriction is visible. The tongue lies on the floor of the oral cavity. For observers who notice the alveolar constriction in the /alda/ video, the consonant has to be /d/; it cannot be /g/. There is no way to tell what the constriction was in the video from /arda/.

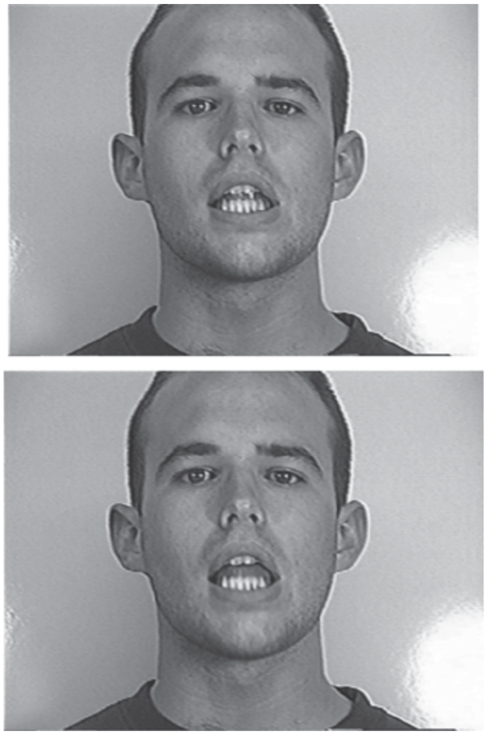

Figure 1. Model speaker's visible vocal tract configuration at stop consonant release in /alda/ (top) and shortly thereafter (bottom). The tongue tip constriction is visible, particularly in the bottom display.

Most likely, the alveolar constriction was not typically noticed. The constriction is not easy to see, and the response pattern—more /ga/ responses to acoustic stimuli dubbed onto the video /da/ from /alda/ than to stimuli dubbed onto /da/ from /arda/—suggests, at most, a weak effect of the visible /d/ articulation. But that is the only subtle bias favoring /d/ in the stimuli that I can see, and it favors /d/ responses to the wrong audiovisual condition for the argument in Holt et al. (2005).

What, then, underlies the findings in Holt et al. (2005)? Consider the salient difference between the videos in the top displays of Figures 1 and 2. It is that the lips are protruded in the video of /da/ from /arda/, but not in the video of /da/ from /alda/. This is not a confounding; it is, rather, the coarticulatory effect whose perceptual effects we are studying. It is a carryover of lip rounding from the /r/ in /arda/. This visible rounding must underlie the response patterns in Experiment 4 in Holt et al. (2005). Listener/viewers were compensating for the effects of the visible rounding gesture that overlaps with the /da/–/ga/ syllables.

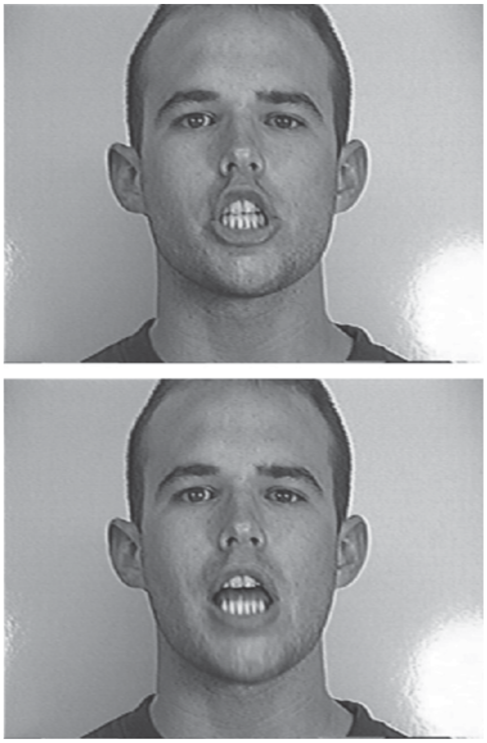

Figure 2. Model speaker's visible vocal tract configuration at stop consonant release in /arda/ (top) and shortly thereafter (bottom). No stop constriction is visible.

In rejecting the idea that Experiment 4 might reflect compensation for coarticulation, Holt et al. (2005) commented that the silence before the syllable specifies production of an isolated syllable; it tells listener/viewers that there was no left context that the listeners needed to compensate for. This may or may not be the case, but it does not matter. Imagine this sampling of possibilities: (1) Contrary to the inference of Holt et al. (2005) from the silence preceding the syllable, participants realize that the syllable has been excised from a rounding context; (2) they guess that, before saying /da/ or /ga/, the model speaker was whistling a merry tune or kissing a loved one; (3) they guess that the model was about to whistle a merry tune or about to kiss a loved one; or (4) the participants were mystified by the lip protrusion. It does not matter which scenario is accurate, if any, because it does not matter *why* the lips were rounded; it only matters *that* they were rounded and, therefore, would lower the $F3$ of the syllable that the gesture overlapped with temporally. In a very interesting experiment, Holt et al. (2005) have verified the finding of audiovisual compensation for coarticulation.

Compensation for coarticulation does not reflect a loss of sensitivity to the coarticulatory information: A companion finding to compensation for coarticulation. Research on compensation for coarticulation has focused, in large part, on listeners' failures to hear phonetic properties signaled by a highly context-sensitive acoustic structure as context sensitive. According to a spectral contrast account, that occurs because of a transient loss of sensitivity to the spectral information that is similar to the spectral properties of a preceding stretch of acoustic signal. However, that account fails to explain a companion finding to compensation for coarticulation in this research domain. Listeners are highly sensitive to coarticulatory information. They use it as information for the coarticulatory context, not for the perceived context sensitivity of targets of coarticulatory overlap.

For example, listeners do not hear coarticulatory information that is in the domain of a schwa vowel as context sensitivity of the schwa vowel (e.g., Fowler, 1981; Fowler & Smith, 1986). This was shown by extracting a /bə/ syllable that had been produced in the context of flanking /i/ vowels (i.e., from /ibəbi/) and splicing it both into another /i/–/bi/ context (a *spliced* trisyllable) and into an /a/–/ba/ context (a *cross-spliced* trisyllable). Corresponding spliced and cross-spliced trisyllables using /bə/ from an /a/–/ba/ context were constructed. Listeners took part in a 4IAX discrimination task. They heard two pairs of the trisyllables and had to judge in which pair the trisyllables sounded more alike. On critical trials, the flanking vowels were different within a pair, so the task was difficult, and the discrimination had to be based on the schwa vowels. An example trial is /ibəbi/(spliced)--/abəba/(spliced)---------/ibəbi/(spliced)--/abəba/(cross-spliced). In this example, the schwa vowels in the first pair are acoustically different, due to coarticulation from the flanking vowels, but the acoustics of each schwa are appropriate for their coarticulatory context. In the second pair, the schwa vowels are acoustically identical, because that in /abəba/ was originally produced in an /i/ context. Listeners judged the members of the first pair to be more similar than the members of the second pair. The schwa vowel in /abəba/ in the second pair sounded high (somewhat like /ɪ/).

This outcome can be explained as being due to spectral contrast. That is, the formants during schwa will be affected by a vowel-to-vowel coarticulation that is both carryover (from initial /i/ or /a/) and anticipatory (from final /i/ or /a/) in direction, so that, for example, $F2$ during schwa is higher if flanking vowels are /i/ than if they are /a/. If the flanking vowels induce spectral contrast effects on schwa, an /i/ context will effectively lower $F2$ during schwa; /a/ will not. This will lead to appropriate compensation for coarticulation during schwa for schwas embedded in contexts similar to the ones in which they were originally produced. But in cross-spliced contexts, acoustically identical schwas will sound different, and that was the outcome.

However, Fowler and Smith (1986) showed that, despite findings of compensation for coarticulation during schwa, listeners nonetheless use coarticulatory information during schwa as information for its coarticulatory source (see also Fowler, 2005). That is, they identify the final vowels of the trisyllables above (presented with fillers /abəbi/ and /ibəba/) more quickly in spliced contexts, where schwas provide accurate anticipatory information, than in cross-spliced contexts, where they provide misleading information.

A second pair of outcomes is similar. Listeners compensate for the coarticulatory effects of /u/ and /a/ in a preceding /s/ or /ʃ/ (Mann & Repp, 1980). That is, they report more /s/s in the context of a following /u/ than in the context of an /a/, as if they are compensating for the spectrum-lowering effects of /u/'s anticipatory rounding gesture. However, despite compensation, they use the coarticulatory information for the forthcoming vowel that occurs in /s/ or /ʃ/ to identify the vowel (Whalen, 1984, Experiment 2). Whalen showed this by using the splicing/cross-splicing procedure. He found that, when the fricative provided misleading coarticulatory information for the vowel, listeners identified the syllables more slowly and less accurately than when the coarticulatory information was appropriate.

This pair of findings—compensation for coarticulation, leading to near or full context independence of phonetic perception, but, nonetheless, use of the eliminated context sensitivity as enhancing information for its coarticulatory source—is exactly the prediction of both the motor theory and direct realism. Compensation for coarticulation occurs because parsing occurs along gestural lines. The companion finding occurs for the same reason.

The pair of findings is obtained also in the domains described above in which the spectral contrast account does not predict contrast but compensation occurs. Listeners do not hear an *f*0 contributed by a high vowel as part of a pitch accent (Silverman, 1987), but they do use it as information for vowel height (Reinholt-Peterson, 1986). Listeners do not hear an *f*0 perturbation on a vowel, due to a preceding unvoiced obstruent, as a pitch contour, but they do use the perturbation as information that the consonant is unvoiced (e.g., Pardo & Fowler, 1997). In these cases and others, not hearing a coarticulated phonetic property as context sensitivity does not mean that listeners have lost sensitivity to it. Rather, in every instance tested, they have used the acoustic sources of context sensitivity as information for their causal source in a coarticulating gesture.

Resumé

In short, the spectral contrast account is too narrow in scope to explain the broad range of findings of compensation for coarticulation. Spectral contrast, as it has been described by its proponents, can account only for effects of carryover coarticulation, and only when that coarticulation has assimilative acoustic consequences. However, compensation for coarticulation comes in many other varieties.

Moreover, spectral contrast does not predict the companion finding that listeners do not lose sensitivity to sources of context sensitivity in speech. If it invokes the idea that contrast only reduces, not eliminates, sensitivity to the assimilative acoustic consequences of coarticulation, it must predict contrastive, not enhancing, effects of that residual information.

In order for the general auditory theory that stands behind the contrast account to explain the variety of compensation for coarticulation and companion findings, a multiplicity of different accounts, in addition to contrast, would have to be invoked. Instead, gesture theories provide a single cohesive account of all of the findings of true compensation for coarticulation (i.e., excluding those involving nonspeech contexts) and the companion findings.

The following experiments provide final examples in which the spectral contrast account fails to predict listener performance. They are like those in the research in which vowel height and intonational gestures have had converging effects concurrently on *f*0, and listeners have compensated for the vowels' intrinsic *f*0 in judging peak height (Silverman, 1987), and they are like Experiment 4 in Holt et al. (2005). In those cases and in the present Experiment 1, compensation for coarticulation occurs with no left or right context to induce a contrast effect. In Experiment 1, however, I sought this outcome with the VCCV disyllables of Mann (1980), because these stimuli have been used often by Holt, Lotto, and colleagues to support their contrast account.

Having found compensation for coarticulation in Experiment 1, I sought and found the companion finding in Experiment 2. Listeners used information about /l/ or /r/ in /da/–/ga/ syllables as information for /l/ or /r/. Accordingly, in contrast to claims of spectral contrast, listeners did not lose sensitivity to coarticulatory effects of /r/ and /l/ on /da/ and /ga/. By the logic of speech–nonspeech comparisons in the literature, Experiment 3 ruled out an account of the companion finding that would be general to speech and sine waves. It also showed a right-to-left contrastive effect of a speech syllable on nonspeech tones.

EXPERIMENTS 1A AND 1B

In judging the synthetic /da/–/ga/ syllables first used by Mann (1980), listeners needed the precursor syllables /al/ and /ar/ in order to compensate for coarticulation, because the synthetic /da/–/ga/ syllables provided no specific information that coarticulation with /l/ and /r/ had occurred. They provided only the general information that place of articulation was not canonical for either /da/ or /ga/ for most continuum members. The precursor syllables were needed to account for the source of the place shifts. However, in natural /da/–/ga/ syllables originally produced in the contexts of precursor /al/ and /ar/, according either to the motor theory or to direct realism, compensation may not require presence of the precursors. Specifically, the precursors may not be required if information for /r/ or /l/ is sufficiently salient in the /da/ and /ga/ syllables. For the spectral contrast view, however, the precursor context is necessary to induce left-to-right contrast and cause the assimilative acoustic consequences of coarticulation in the /da/ and /ga/ syllables to be reduced or eliminated by contrast.

Experiment 1 was designed to distinguish these predictions by using /da/–/ga/ continua generated from natural, rather than synthetic, speech. In the experiment, the listeners heard members of hybrid continua made from natural speech. Each continuum consisted of /da/–/ga/ hybrids that had originally been produced in the context of a precursor syllable, /a/, /al/, or /ar/. The monosyllabic hybrids contained carryover coarticulatory information for their original context. That is, hybrids from the /ar/ context, for example, will contain carryover information about /r/.

From the perspective of a spectral contrast account, the response pattern should look opposite to that typical in previous investigations of these disyllables. That is, acoustic consequences of CVs originally produced in an /ar/ context will have a lowered spectrum due to the /r/ coloring. Those produced in the /l/ context will have *F*3 effectively raised. Because there is no context to induce spectral contrast, there is no way to compensate for those effects. Accordingly, the listeners should report more "ga" responses when the coarticulatory context had been /ar/ and more "da" responses when it had been /al/, opposite to findings with synthetic speech.

Predictions from the gestural overlap account are, unfortunately, much weaker. If the listeners were ideal parsers, the predictions would be clear.[4] For each continuum member, information for overlapping /r/ or /l/ gestures would be parsed out, leaving "pure" acoustic information for each /da/–/ga/ hybrid. The result of parsing /l/ and /r/ gestures should leave identical residuals, so that response curves for "da"–"ga" identifications should sit on top of one another.

It is well established, however, that listeners are not ideal parsers (see Fowler, 2005, for a review and discussion). Sometimes, they pull out too much; sometimes, they pull out too little (see Fowler & Brown, 1997, for both outcomes). The conditions under which each outcome is observed have not been determined.

In Experiment 1, if the listeners underparsed, their response pattern would not be distinguishable from predictions of spectral contrast. If they were ideal parsers, their responses in all three conditions (/a/, /al/, /ar/ contexts) would be indistinguishable. If they overparsed, their response patterns would resemble those in Mann (1980), with more /g/ responses when the overlapping gestures were those of /l/, rather than /r/.

Because the predictions of the gestural theory are weak, encompassing any of three possible outcomes, Experiment 1 is best seen as a test of the claim of the spectral contrast account that a context is needed for compensation for coarticulation, not as a test of gestural parsing. The prediction was that, as in Experiment 4 in Holt et al. (2005), compensation for coarticulation would occur when a left context was absent. However, in the present experiment, in contrast to that in Holt et al. (2005), coarticulatory information was acoustic, rather than visual.

Method

Participants. The speech stimuli presented to the listeners were produced by two native speakers of American English with phonetics training. The listeners were 38 undergraduates at the University of Connecticut, who participated for course credit. They were native English speakers who reported normal hearing. Eighteen listened to the speech of one talker, and 20 listened to the speech of the other talker. Data from 1 of the 20 participants were excluded from the analyses because his responses were random. Data from another were excluded in order to achieve even counterbalancing. Of the possible participants who could be excluded, one was selected who heard no /d/s in the /alda/–/alga/ condition—that is, whose data would bias the outcome toward the predictions of direct realism and the motor theory.

Stimulus materials. Each of two speakers (one the author, one a phonetician who was unaware of the purpose of the research) produced at least 10 tokens each of six disyllables: /ada/, /aga/, /alda/, /alga/, /arda/, and /arga/. For the latter four, the speakers attempted to produce the disyllables with substantial carryover coarticulation between the consonants. That is, they allowed the /l/ to pull the tongue forward during the stop constriction and allowed the /r/ to pull it back. They made sure that lip rounding for /r/ persisted during the stop consonants. Disyllables were produced in isolation and had an appropriate intonation contour for isolated utterances. That is, they were declarative contours.

From the speech of each talker, six /da/-to-/ga/ continua were made from the three pairs of disyllables. For each, the initial V (from /ada/ and /aga/) or VC was spliced from the CV syllable. Tokens of /da/ and /ga/ were selected for each continuum that had as similar an *f*0 contour and duration as possible. Formant values at four time points during the endpoint stimuli are provided in the Appendix. The time points are the middle of the first /a/, the end of the /al/ or /ar/ syllable, the onset of voicing of the second /a/, and the midpoint of the same vowel.

Initially, an assistant, naive as to the purposes of the research, made a hybrid continuum from speaker C.A.F.'s speech by taking as one end of the continuum 100% of the amplitude of a /ga/ syllable from an /al/ context. The next continuum member had 90% of /ga/'s amplitude and 10% of /da/'s from an /al/ context. Eleven members were created in steps of 10% change. These syllables were unsatisfactory, because the *f*0s, although similar, were not similar enough and, often, two voices were heard. Accordingly, the assistant used a different method to make the continua from the speech of both speakers. She took the stop burst and the first approximately 65 msec (cutting the signal at a zero crossing) from /da/ and /ga/ syllables, one syllable of each type produced in the context of /a/, of /al/, and of /ar/. She hybridized those syllable fragments as described above to make an 11-step continuum in which gradually more of /da/'s amplitude and less of /ga/'s contributed to the hybrid. Then she pasted the remainder of a vowel either from the /da/ syllable or from the /ga/ syllable to complete the syllable. This provided six hybrid continua per speaker. Two of them combined syllables originally produced in an /a/ context, one in which the final pitch pulses of a /da/ syllable completed the vowel and one in which the final pitch pulses of the /ga/ syllable completed the vowel. Corresponding pairs of hybrid continua were constructed from a /da/ and a /ga/ originally produced in the context of /al/ and from a /da/ and a /ga/ from an /ar/ context.

The syllables, particularly those of speaker C.A.F., did not sound good. Many included artifactual clicks that the best efforts of the individual who constructed them could not eliminate. However, other than possibly introducing noise into the response patterns, this should have no effect on the outcome.

Four test orders were constructed out of the syllables of each speaker. One consisted of five tokens each (110 trials total) of members of the two monosyllabic continua constructed from the /ada/ and /aga/ contexts. Two others consisted of corresponding monosyllables from the /ar/ and /al/ contexts. The fourth test order consisted of one token of each member of all six continua. The stimuli were randomized in each test order.

Procedure. The stimuli were presented to the listeners using a MATLAB program (MathWorks, Natick, MA). The participants listened over headphones as they took four tests. The order of three of the tests was counterbalanced across participants. In these tests, the listeners heard monosyllables from the /ada/–/aga/, /alda/–/alga/, or /arda/–/arga/ series. Their task on each trial was to identify the consonant of the syllable as /d/ or /g/. All the participants took a final test in which each token from the six continua was presented once. Their task was to decide whether the missing first syllable was /a/, /al/, or /ar/. The purpose of this test was to confirm our impression that the syllables that were originally produced after /al/ or /ar/ did not sound as if they began with /l/ or /r/.

Results and Discussion

The results of the first three tests are shown in Figures 3A and 3B, which plot the proportion of "g" responses across the continua separately for the /ada/–/aga/, /alda/–/alga/, and /arda/–/arga/ series. In the speech of C.A.F. (Figure 3A), there is a large effect of coarticulatory context, with substantially more "g" responses when the precursor syllable had been /al/ than when it had been /ar/. The data were submitted to an ANOVA with factors of precursor syllable (/a/, /al/, or /ar/), continuum member (1–11), and vowel. The last factor reflects whether the final pitch pulses of the vowel had come from a /da/ or a /ga/ syllable. In the ANOVA, the effects of precursor syllable [$F(2,34) = 49.85$, $p < .001$] and continuum member [$F(10,170) = 266.05$, $p < .001$] and their interaction [$F(20,340) = 2.42$, $p < .001$] were significant, as was the three-way interaction [$F(20,340) = 2.47$, $p < .001$]. Planned contrasts on the precursor syllable factor showed that all pairwise differences were significant [/al/–/ar/, $F(1,34) = 91.75$, $p < .0001$; /al/–/a/, $F(1,34) = 52.31$, $p < .0001$; /ar/–/a/, $F(1,34) = 5.50$, $p < .05$]. The two-way interaction reflects the finding that differences across precursor syllable conditions were absent or reduced at the edges of the continuum. The three-way interaction is significant because, whereas the curves for the /ada/–/aga/ and /alda/–/arda/ stimuli were nearly identical across the vowel factor, those for /arda/–/arga/ paradoxically showed more "ga" responses for two continuum members when the end of the vowel had come from /da/, rather than from /ga/.

In Figure 3A, the /ada/–/aga/ curve reveals an anomalous outcome on Continuum Member 10, where there are substantially more "ga" responses than to Continuum Members 7, 8, and 9. Most likely, the hybrid syllable fragment for this continuum member was made incorrectly. The error is most likely on the fragment, because the unexpected outcome occurs both when the vowel remainder is from /da/ and when it is from /ga/.

The results for the second speaker (Figure 3B) were the same in the most critical respect. More "g" responses occurred to members of the continuum constructed from /alda/ and /alga/ utterances than to members of the continuum constructed from /arda/ and /arga/. In addition, as was expected, more "g" responses occurred to /alda/–/alga/ members than to /ada/–/aga/ members. However, unexpectedly, numerically more "g" responses occurred to /arda/–/arga/ members than to /ada/–/aga/ members. In an ANOVA, effects of precursor syllable [$F(2,34) = 9.19$, $p < .001$], continuum member [$F(10,170) = 139.2$, $p < .001$], and their interaction [$F(20,340) = 9.64$, $p < .001$] were significant. In pairwise comparisons across the continua, the difference in "g" responses to the /ada/–/aga/ and /alda/–/alga/ continuum members and to the /alda/–/alga/ and /arda/–/arga/ continua were significant [$F(1,34) = 16.73$, $p < .001$, and $F(1,34) = 9.99$, $p < .01$, respectively]; that between /ada/–/aga/ and /arda/–/arga/ did not approach significance ($F < 1$). The interaction reflected the fact that differences across continua were confined to intermediate continuum members. There was also a significant interaction between precursor syllable and vowel remainder [$F(2,34) = 7.56$, $p < .01$], reflecting that "g" responses increased when the vowel remainder was from /ga/ rather than from /da/, but only in monosyllables constructed from /ada/ and /aga/.

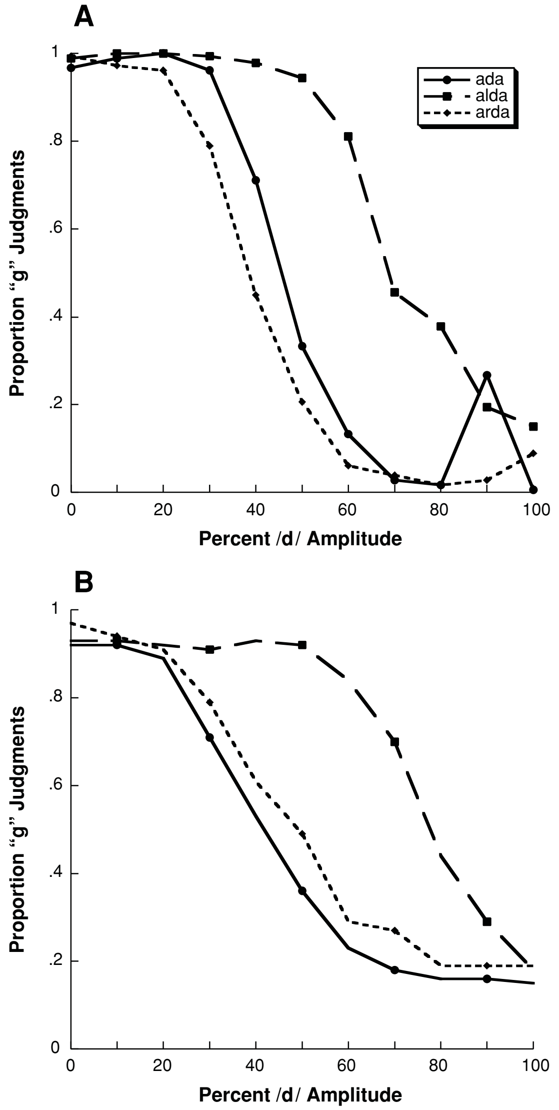

Figure 3. Results of Experiment 1. The listeners reported hearing more /g/ consonants in continuum members having coarticulated with precursor /al/ than with /a/ and in members coarticulated with precursor /a/ than with /ar/. (A) Speech of C.A.F. (B) Speech of D.H., a male phonetician who was naive as to the purposes of the research.

The next question was whether compensation for coarticulation occurred because listeners heard syllables such as /rda/ or /lga/. They did not. For the speech of C.A.F., performance identifying the precursor syllable did not exceed chance in any condition (the proportion correct, as compared with chance [.33], for /ada/–/aga/ was .37, for /alda/–/alga/ was .40, and for /arda/–/arga/ was .33; all *p*s $>$ .10). Likewise, performance on the speech of the second talker revealed no significant departures from chance (/ada/–/aga/, .40; /alda/–/alga/, .31; /arda/–/arga/, .33; all *p*s $>$ .10). It is not paradoxical that listeners both used the information about /l/ and /r/ in their "d"–"g" identifications and failed to be able to determine consciously whether the missing syllable was /a/, /al/, or /ar/. Much of perception occurs outside of awareness.[5]

The experiment revealed a response pattern opposite to that predicted by a spectral contrast account. As in research by Mann (1980) and in follow-up studies, the listeners reported more "g" responses when the coarticulatory information was for /l/ than when it was for /r/ or for no overlapping consonant. The spectral contrast account has to predict more "g" responses in the /ar/ context, because coarticulation effectively lowers $F3$ and there is no left context to induce compensation.

The only alternative account to date is true compensation. The listeners extracted the coarticulatory effects of /r/ and /l/ from the domain of /da/–/ga/ continuum members. However, the parsing that the experiment reveals was not ideal. Had ideal parsing occurred, response functions would have overlaid one another. Yet only in responses to /da/–/ga/ continuum members originally produced in the context of /r/ did the listeners respond as if /r/ information had been accurately parsed, and only then in the speech of D.H. The other results suggest overparsing—that is, pulling out more than coarticulation had put into the /da/–/ga/ continuum members.

Despite the inaccurate parsing, Experiment 1, like Experiment 4 in Holt et al. (2000), shows that compensation for coarticulation need not reflect "action at a temporal distance." To compensate for coproduced phonetic segments, listeners need to know what the temporally overlapping gestures are that they need to compensate for. However, they can discover that in a variety of ways. When speech is synthetic, as in the research by Mann (1980), listeners require the left context to tell them what the overlapping gestures are. This is because hearing members of the /da/–/ga/ continuum, they can detect that place of articulation is not canonical for either /d/ or /g/, but they cannot know whether the /d/ was pulled back or the /g/ pulled forward without knowing whether the overlapping consonant gestures are those for /l/ or /r/. However, in the research by Holt et al. (2000), no left context was required even though the speech was synthetic, because viewers could *see* the overlapping rounding gesture from /r/. In Experiment 1, they could hear it.

Of course, this experiment does not rule out the possibility that spectral contrast occurs in speech perception when there *is* a left context present, so that the effects of Experiment 1 would have been larger had the context been present.[6] However, the present experiment does show that very large (even too large) compensation effects occur when contrast is ruled out. Given that the same effects also occur in situations (such as the experiments on compensation for $f0$) to which spectral contrast cannot apply, the conclusion is warranted that spectral contrast, if it contributes at all to compensation for coarticulation, plays an insignificant role.

Experiment 2 was designed to test for the companion finding—that is, to ask how, if at all, listeners use coarticulatory information about /l/ or /r/ that is parsed from the /da/ and /ga/ syllables. If information about the consonant at the end of the precursor syllable is ambiguous between /l/ and /r/, listeners may use coarticulatory information about the consonant in the domain of a following /d/ or /g/ to reduce the ambiguity. In two ways, this experiment predicts the reverse of a contrast effect. First, it predicts an effect of a gesture that starts later in time on one that starts earlier. Second, it predicts that information for /l/ in the domain of /d/ or /g/ will *increase* identifications of a preceding consonant as "l."

EXPERIMENT 2

Method

Participants. The participants were 16 undergraduates at the University of Connecticut, who received course credit for their participation. They were native speakers of English, who reported having normal hearing.

Stimulus materials. As in Experiment 1, the stimulus materials were created from natural speech. The speech was that of C.A.F. used to construct the materials for Experiment 1. Three final /da/ syllables and three final /ga/s were selected, one token each from the /ada/, /aga/, /alda/, /alga/, /arda/, and /arga/ tokens. They were the tokens used to construct the hybrid syllables in Experiment 1. Corresponding first syllables of the VCCV disyllables were used to make hybrid continua ranging from /al/ to /ar/.

The final approximately 102 msec of each precursor syllable (with cuts made at zero crossings) were used to construct the hybrid consonants. An /al/ from a /da/ context was mixed with an /ar/ from a /da/ context in proportions of 100%–0%, 80%–20%, 60%–40%, 40%–60%, 20%–80%, and 0%–100%. Corresponding hybrids were made using /al/ and /ar/ from /ga/ contexts. These were appended to the initial portions of vowels from original /al/ and /ar/ syllables from both /da/ and /ga/ contexts. Finally, each of these hybrid precursor syllables was spliced before /da/ and /ga/ syllables from /a/, /al/, and /ar/ contexts. This made 144 unique disyllables. Two tokens of each disyllable were randomized to make a test order.

Even more so than the hybrid syllables in Experiment 1, these VCs were poor in quality. In this case, I was not successful in finding a section of the VC to hybridize that included all of the /l/ or /r/ information. Accordingly, some of the VCs sounded like /alr/ or /arl/. However, because the same precursor syllables occurred before /da/ and /ga/ from all disyllable types, this could have no biasing effects on the outcome. It could only reduce the magnitude of any effect that might occur.

Procedure. The participants listened over headphones to the stimuli, which were presented by a MATLAB program. They were instructed to listen to each disyllable and to identify the consonant at the end of the first syllable as "l" or "r," guessing if necessary. They were warned that sometimes they might hear both consonants. In that case, they were to choose the consonant that was more clearly audible or, if neither was, to pick a response at random.

Results and Discussion

The responses were collapsed over counterbalancing variables, leaving two variables of interest: original coarticulatory context (/a/, /al/, /ar/) and continuum member. Figure 4 shows the results. The figure plots the proportion of "l" judgments across members of the continua. Long dashed lines represent /da/ or /ga/ syllables that had coarticulatory information for /l/ in them; short dashed lines are /da/s or /ga/s from an /ar/ context; the solid line represents /da/s and /ga/s from /a/ contexts. The important finding is that /da/s and /ga/s with information for /l/ in them promoted /l/ judgments. Those with information for /r/ in them promoted /r/ judgments.

In an ANOVA with factors of original coarticulatory context of the CV (/a/, /al/, or /ar/) and continuum members, both the main effects and the interaction were significant [coarticulatory context, $F(2,30) = 7.36$, $p < .005$; continuum members, $F(5,75) = 101.39$, $p < .0001$; interaction, $F(10,150) = 2.90$, $p < .005$]. Bonferroni tests revealed a significant difference between the /al/ and the /ar/ contexts ($p < .001$) and a marginal difference between the /a/ and the /al/ contexts ($p = .022$). The significant interaction reflects the changing relation between responses to /da/s and /ga/s in the /a/ context and those to /da/s and /ga/s in the other two contexts. In addition, CVs originally produced after /al/ were associated with more "l" responses than were those originally produced after /ar/ everywhere except at the last continuum member, where responses were as close to the floor as responses got in any condition.

The experiment revealed evidence of the companion finding. Information that Experiment 1 had shown was parsed from the /da/ and /ga/ syllables was used in Experiment 2 as information for its coarticulatory source, /r/ or /l/. Notably, these are effects of a gesture's later acoustic consequences on identification of the gesture's beginnings in a preceding syllable. They are opposite in outcome to a contrast effect. That is, /l/ information in /da/ or /ga/ promotes, rather than reduces, /l/ identifications.

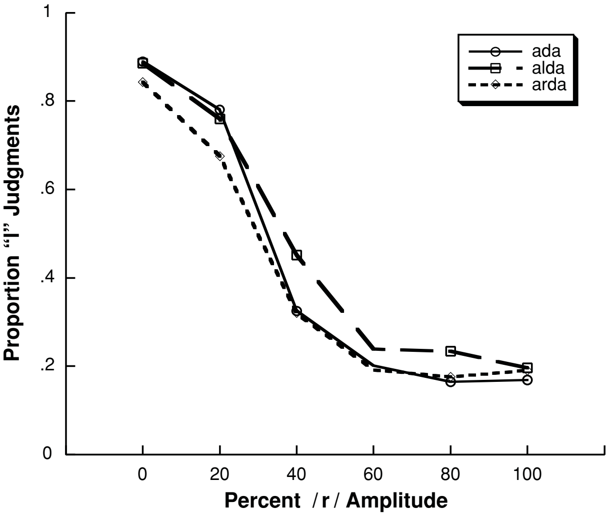

Figure 4. Results of Experiment 2. The listeners reported hearing more /l/s in VC syllables followed by /da/ or /ga/ that include coarticulatory information about /l/ than in other contexts. This is an enhancing effect of the second syllable's coarticulatory information on perception of the first syllable.

The third experiment in the series was designed to look for an analogue of the companion finding when tones are substituted for the precursor syllables. Although spectral contrast cannot underlie the outcome of Experiment 2, it does not follow that no auditory process common to speech and tone perception underlies the findings.

Interpretation of Experiment 3 depends on a logic that I reject, as I will explain in the General Discussion section. However, it is a logic widely used in the field in these kinds of tests. If the behavioral patterning of responses to nonspeech signals mirrors that to speech signals, an inference is drawn that processing applied to the nonspeech and speech signals is the same. In Experiment 3, I expected to show that, by this logic, the processing applied to speech in Experiment 2 and to nonspeech in Experiment 3 is different.

EXPERIMENT 3

Method

Participants. The listeners were 18 undergraduates at the University of Connecticut, who participated for course credit. They were native English speakers, who reported normal hearing.

Stimulus materials. The ending $F3$ of /al/ in the speech of C.A.F. averaged 3059 Hz. That of /ar/ averaged 2152 Hz. Six tones were synthesized to span that range approximately. Tones were 300 msec long and had frequencies from 2000 to 3000 Hz in steps of 200 Hz. The amplitude of the tones were ramped up over the first 15 cycles and down over the last 15 cycles of each tone. The RMS amplitudes of the tones were matched to those of the initial syllables in Experiment 2.

The tokens of /da/ and /ga/ used in Experiment 2 were appended to the tones, with no temporal gap between tone offset and closure onset of the syllables. This mimicked the time course of the /al/ and /ar/ precursors in Experiment 2.

There were 36 unique tone–syllable sequences. In the test order, four tokens of each stimulus type were randomized, with one token of each type appearing in each quarter of the test order.

Procedure. The participants heard the tone continuum endpoints five times each and were told to call the 3000-Hz tone the high tone and the 2000-Hz tone the low tone by hitting the appropriately labeled key on the computer keyboard. They were then told that they would hear a variety of tone–syllable stimuli. On each trial, they were to classify the tone as more like the high tone or more like the low tone by hitting the appropriately labeled key, guessing when necessary. They heard the stimuli over headphones. A program written in MATLAB presented the stimuli and collected the responses.

Results and Discussion

The results are shown in Figure 5, which presents proportions of *high* judgments across members of the tonal continuum. Note that the tones are presented with the high tone leftmost on the x-axis, so that Figures 4 and 5 may be easily compared. The figure does reveal an effect of the CV syllable on tone judgments, but the effect, unlike that in Experiment 2, is contrastive in direction.

An ANOVA with factors of coarticulatory context of the CV syllable (/a/, /al/, or /ar/) and continuum member showed all the factors to be significant [coarticulatory

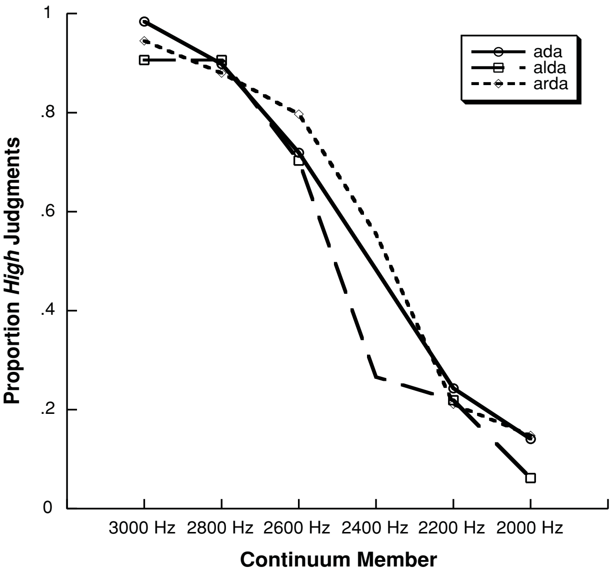

Figure 5. Results of Experiment 3. The listeners reported hearing fewer high tones when continuum members were followed by /da/ or /ga/ syllables that included coarticulatory information for /l/ than when they were followed by CVs that coarticulated with /a/ or /ar/. This is a contrastive effect.

context, $F(2,30) = 7.76$, $p < .002$; continuum member, $F(5,75) = 113.63$, $p < .0001$; interaction, $F(10,150) = 4.38$, $p < .0001$]. Bonferroni tests revealed significant differences between the /al/ and the /ar/ coarticulatory contexts ($p = .002$) and between the /a/ and the /al/ contexts ($p = .007$).

This experiment, like that in Stephens and Holt (2003), shows an effect of a speech context on judgments of nonspeech signals. However, in contrast to that work, the effects are opposite to those of speech contexts on judgments of speech signals.

GENERAL DISCUSSION

Experiments 1–3 further refute the spectral contrast hypothesis, showing that a contrast account does not suffice to explain compensation for stimuli such as those used by Holt, Lotto, and colleagues in many of their experiments. Experiment 1 showed compensation for coarticulation with no left or right context to induce contrast. Experiment 2 showed the companion outcome, whereby an effect of information later in time on the perception of earlier phonetic information occurred. Moreover, the qualitative effect was opposite to that of contrast. Information about /l/ or /r/ in the second syllable of a VCCV disyllable fostered identification of ambiguous consonants in the precursor syllable as /l/ and /r/, respectively. Experiment 3 showed no corresponding effects on identification of precursor tones. Rather, contrastive effects occurred. As was outlined in the introduction, ample evidence dating from 1980 onward already has disconfirmed the spectral contrast account. The present experiments show that the spectral contrast account does not explain effects on the disyllables to which Lotto, Holt, and colleagues have frequently applied it, those first used by Mann (1980).

There remain the findings reviewed in the introduction that are predicted by a contrast account, but by neither gesture theory. These are findings that nonspeech contexts can affect phonetic judgments in the same way as speech contexts and that speech contexts can affect nonspeech judgments in the way that they affect speech judgments. In the light of the evidence that disconfirms the contrast account of compensation for coarticulation, how should these findings be viewed?

The short answer is that I do not know, but I doubt very much that the findings have any bearing on speech perception.

Kuhl (1987) has suggested that, as tests specifically for special-to-speech processing, experiments comparing speech with nonspeech perception are not convincing, because nonspeech signals selected to resemble speech sufficiently to warrant comparison may be processed by the speech system. That is, that research may be about the tuning of a phonetic module, if there is one.

I do not find that account likely to be applicable to the stimuli used by Lotto, Holt, and colleagues, because, were the signals handled by a phonetic module, one would expect listeners to extract some phonetic information from them. Although this is not usually tested for in this literature, when it has been tested for (e.g., Fowler, 1992; Stephens & Holt, 2003), listeners have generally failed to detect any phonetic information in the nonspeech signals.

A different criticism that can be raised about the speech–nonspeech comparisons is that the logic required for their interpretation is very weak. Theorists infer from qualitatively similar response patterns to speech and nonspeech signals that qualitatively similar perceptual processes are being applied to them. However, that reasoning can go very far wrong, as I have shown (Fowler, 1990). The idea in the research by Fowler (1990) was to obtain responses to nonspeech events that were qualitatively either similar to or different from the findings in Miller and Liberman (1979) on rate normalization in /b/–/w/ perception. The nonspeech events were ball bearings running along two kinds of tracks. Listeners judged an initial sloping part of each track as shallow or steep in slope. With one kind of track, the listeners' slope judgments patterned like those of listeners making /b/–/w/ classifications over rate variation. With the other, the response pattern was opposite. It is not likely, however, that the processing of ball bearings running down one kind of track is like the identifying of consonants as /b/ or /w/, whereas the processing of ball bearings running down another kind of track is different.

The logic of research comparing speech and nonspeech perception aside, there is a difference between the speech and the nonspeech sounds used in the comparisons made by Holt, Lotto, and colleagues that may be relevant to their findings. When listeners are presented with sequences such as /alga/ and so forth, they hear a single event of talking. When they hear a sine wave tone followed by /ga/

and so forth, they hear two events very closely sequenced. Possibly, the latter event pairing leads to interference between the perceived events, perhaps due to spectral contrast, whatever that may be.

Whatever the account for the nonspeech context effects may turn out to be, I suggest that the similar outcomes that have occurred when speech and nonspeech context effects have been compared are unfortunate (because misleading) consequences of qualitatively similar response patterns arising for different reasons. And, of course, not all outcomes are the same (e.g., Fowler, 1990; the present Experiments 2 and 3). The previous literature and the present findings show that compensation for coarticulation is not and cannot be due, to any significant degree, to spectral contrast.

Nor is a general auditory account viable that invokes spectral contrast to explain just a subset of the relevant findings. First, the central findings used to support spectral contrast are findings that elaborate on the research of Mann (1980), and the present research shows that contrast cannot account even for those findings. Second, a general auditory account has to invoke an eclectic set of accounts to handle a coherent pair of findings that occurs in multiple domains: Listeners compensate for coarticulation, and they use information extracted in the course of compensation as information for its phonetic source. Gestural theories of speech perception have just one account of the pair of findings: Listeners parse acoustic speech signals along gestural lines. This leads both to compensation for coarticulation and to the companion finding.

Rebuttal of "Putting Phonetic Context Effects Into Context" by Lotto and Holt (2006)

Errors. The commentary by Lotto and Holt (2006) includes a number of errors. A sampling follows.

1. Lotto and Holt (2006) comment that I criticized their account on grounds of parsimony, but I did not. I criticized it on grounds of generality. An account is unparsimonious if it invokes explanatory constructs that are unnecessary. Lotto and Holt have not done that to my knowledge, and I do not anticipate that they would do so were they to develop a more comprehensive account of compensation for coarticulation. However, were they to develop a more comprehensive account, they would have to add explanatory constructs beyond that of spectral contrast. My criticism was that the set of findings of compensation for coarticulation and their companion is coherent. Listeners are doing the same thing in every experiment showing true compensation for coarticulation or the companion finding. I offer one account for the whole set of findings—namely, that listeners track temporally overlapping phonetic gestures. In contrast (so to speak), a comprehensive account of the findings by Lotto and Holt would necessarily offer an eclectic set of constructs.

2. Contra Lotto and Holt (2006), I did not criticize their contrast account because it does not account for "all speech perception phenomena." That would, as the authors wrote, be unreasonable. I criticized their account because it does not account for much of the domain of compensation for coarticulation. There is far more to speech perception than compensation for coarticulation.

3. Lotto and Holt (2006) do not understand that, because I recognize that coarticulation is coproduction, I do not see compensation for coarticulation as a context effect in the way that they do. Therefore, I am less concerned by the failure of Holt et al. (2005) to get compensation when they restricted the video to the precursor syllable (but see a different result below) than they believe that I should be. In their view, compensation as contrast occurs when something earlier in time (or, these days, later in time as well) has an impact on the perception of something later (or, these days, earlier). However, in my view, compensation is parsing from one domain—say, that of /da/ or /ga/—information for the *temporally overlapping* production of, say, /al/ or /ar/. In the experiments in which the video was dubbed only onto the acoustic precursor syllable, compensation should occur only if the audiovisual information specifies to the listeners/viewers that the gestures of that syllable will overlap temporally with those of the following CV.

4. Of the finding by Holt et al. (2005) that compensation occurred with the stimuli of Fowler et al. (2000) when the stimuli were audiovisual members of the /da/-to-/ga/ continuum without the precursor syllable, I did not suggest that "when there was no context listeners parsed the context from the subtle video clues present during the target." It was Holt et al. (2005) who invoked subtle clues. Moreover, Lotto and Holt (2006) have confused the terms *subtle* and *salient*. And they are unaware that direct realists do not traffic in perceptual "clues."

I was dismayed that the reviewers did not require Holt et al. (2005) to specify what the subtle differences were in the videos of /da/ from the /al/ and /ar/ contexts that they contended underlay their findings and those of Fowler et al.'s (2000) Experiment 3. So I took a look myself. I did find a subtle difference (a visible tongue tip gesture for /d/), but it was in the video that, according to Holt et al. (2005), should have looked subtly less like /d/. However, I remarked that that subtle difference was unlikely to have had much impact on performance. It occurred in the condition in which the listeners/viewers reported more /g/s. And it was subtle. I referred to the difference that I am confident underlies their outcome as *salient*, not subtle. The model speaker's lips were very clearly rounded during the /da/ from /arda/. The rounding was gestural overlap from /r/. The listeners/viewers compensated for that. I very much like this experimental outcome. Holt et al. (2005) do not understand it.

5. Lotto and Holt (2006) allege that they provide an account, involving spectral contrast, of a companion finding by Whalen (1984). However, I do not see how their account generates the findings. Why is a syllable in which there is less spectral contrast identified more slowly than one in which there is more?

6. Lotto and Holt (2006) comment that experiments such as Experiment 1 in the target article and Experiment 4 in Holt et al. (2005) do not test a contrast account, because there is no context to induce contrast. But this

reasoning is erroneous. The account that Lotto, Holt, and colleagues offer for compensation for coarticulation is that it is a contrast effect. If that is correct, no compensation should occur when the context is removed. But it does in Experiment 1 in the target article and in Experiment 4 in Holt et al. (2005). Lotto and Holt are left with the (unparsimonious) explanation that compensation for coarticulation is the result of two redundant processes: contrast and something else. When the context is present, both processes lead to compensation. When it is absent, something else does the job.

Criticism 1: Contrast Effects Are Restricted to Left-to-Right Effects

There were two reasons why I suggested that the contrast effects invoked by Holt, Lotto, and colleagues are left to right only. Most important, the authors had written as if they are. For example, Lotto (1996) did, in the quotation that I provided in the target article that refers to the constraint on the state of a system at time t imposed by its state at time $t - 1$. Nearly the same words appear in Lotto et al. (1997). Second, to underscore the generality of contrast effects, Lotto, Holt, and colleagues (e.g., Lotto & Kluender, 1998; Lotto et al., 1997) cited the literature on contrast effects in perceptual judgments generally (e.g., Warren's [1985] review), and, in that literature, also accounts of contrast presume that they are left to right. Warren's criterion shift account is that contrast effects reflect recalibration of perceptual systems on the basis of *recently encountered* stimuli. It has become clear to me that the word *contrast* is polysemous in the field of perception. In their commentary, Lotto and Holt (2006) have retained the word but shifted to a different meaning having nothing to do with the states of systems at times $t - 1$ and t, to accommodate their contrast account to their belated discovery that compensation for anticipatory coarticulation occurs.

Criticism 2: Some Speech Effects Are Simultaneous or Are Not Contrastive

As for "the nongestural approach provides one explanation for the context effects summarized in Table 1, whereas gestural theories require multiple explanations to cover these observations," au contraire. This gesture theorist offers just one. The findings in the top quarter of the table reflect gesture perception. (Those in the remainder of the table are uninterpretable.) In any case, Lotto and Holt (2006) have once again missed the main point of the target article. It is true that they offer a single account of the findings in their Table 1, but the findings are not coherent; they are eclectic. So they must offer an eclectic account of a coherent set of findings (compensation for coarticulation), and they do offer a coherent account of an eclectic set of findings.

Criticism 3: Spectral Contrast Results in a Loss of Sensitivity to Coarticulatory Information

I wrote that spectral contrast results in a loss of sensitivity to frequencies near the source of the contrast, on the basis of information I received from Andrew Lotto. Fowler et al. (2000) refer to that personal communication:

> Lotto (personal communication, May 8, 1998) has augmented the account of spectral contrast offered [by Lotto & Kluender, 1998]. Spectral contrast occurs when presentation of a tone reduces the effective amplitude of that tone's frequency, and perhaps nearby frequencies, in a subsequently presented acoustic stimulus. (p. 881)

Criticism 4: Contrast Effects Can Occur Without Changes in the Makeup of the Context

I have dealt with most of Lotto and Holt's (2006) comments on this topic under the first heading. The main thing I wish to add is that, in my laboratory, we have twice obtained the result that Holt et al. (2005) have twice failed to get. In one of these experiments, as in those in Holt et al. (2005), an audiovisual /al/ or /ar/ preceded audio-only members of a /da/ to /ga/ continuum. The acoustic signal for /al/ and /ar/ was the same; only the visible gestures determined the syllable-final consonant. I replicated the procedure in Holt et al. (2005), and the stimuli were identical to those in one of their two experiments. My outcome was different, as it had been in an earlier pilot experiment with different stimuli generated for another project. With 16 participants, there was a significant difference in the percentage of /ga/ responses, depending on the precursor video [$F(1,15) = 7.86$, $p < .05$], with 61.9% /ga/ responses when /al/ was the precursor syllable and 54.6% when /ar/ was the precursor. When scoring was made contingent on correct identification of the precursor syllable, response percentages were 63.0% and 53.8%, also a significant difference [$F(1,15) = 9.83$, $p < .01$]. I have no idea why our outcomes were different.

As for Lotto and Holt's (2006) comments on Experiment 1 in the target article, they are mostly right on. That experiment puzzles me. But the effects are large, they hold for two speakers, and they cannot be contrast effects.

The Adequacy of Gestural Theories

Lotto and Holt (2006) offer three pieces of evidence opposing theories that listeners perceive speech gestures. The first one (Lotto & Kluender, 1998), that compensation for coarticulation occurs when the precursor and target syllables are produced by a female and (synthetic) male speaker, respectively, was an odd choice. Lotto and Kluender themselves offered an account that a direct realist might provide of those findings.

It is true that I would not predict the findings in Holt et al. (2000).

As for the findings of Aravamudhan and Lotto (2004, 2005), they are unavailable in the archival literature, and so I can only guess at them. I do not see why they cannot be handled by gesture theories. People with cochlear implants appear to get enough gestural information to make an accurate labeling response, but not enough to show so-called context effects. To show the latter effects, they would have to be sensitive to the gestures in the region of gestural overlap—that is, in the domain of a preceding or following segment.

The Viability of Contrast in a General Perceptual and Cognitive Account of Speech Perception

Contra Lotto and Holt (2006), I have not written that spectral contrast is not real. I have written that it does not have sufficient generality, in the domain of speech, to account for most findings of compensation for coarticulation or the companion findings. I stand by my story.

REFERENCES

Aravamudhan, R., & Lotto, A. J. (2004). Perceptual overshoot in listeners with cochlear implants. *Journal of the Acoustical Society of America*, **116**, 2523.

Aravamudhan, R., & Lotto, A. J. (2005). *Phonetic context effects in adult cochlear implant users*. Paper presented at the 10th Symposium on Cochlear Implants in Children, Dallas.

Coady, J. A., Kluender, K. R., & Rhode, W. S. (2003). Effects of contrast between onsets of speech and other complex spectra. *Journal of the Acoustical Society of America*, **114**, 2225-2235.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**, 2154-2164.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory & Language*, **27**, 143-165.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, **15**, 399-402.

Fodor, J. A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech & Hearing Research*, **46**, 127-139.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, **36**, 359-368.

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

Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America*, **88**, 1236-1249.

Fowler, C. A. (1992). Vowel duration and closure duration in voiced and unvoiced stops: There is no contrast effect here. *Journal of Phonetics*, **20**, 143-165.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, **99**, 1730-1741.

Fowler, C. A. (2005). Parsing coarticulated speech in perception: Effects of coarticulation resistance. *Journal of Phonetics*, **33**, 199-213.

Fowler, C. A., & Brown, J. M. (1997). Intrinsic *f*0 differences in spoken and sung vowels and their perception by listeners. *Perception & Psychophysics*, **59**, 729-738.

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. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 877-888.

Fowler, C. A., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123-139). Hillsdale, NJ: Erlbaum.

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

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, **108**, 710-722.

Holt, L. L., Stephens, J. D. W., & Lotto, A. J. (2005). A critical evaluation of visually moderated phonetic context effects. *Perception & Psychophysics*, **67**, 1102-1112.

Kuhl, P. K. (1987). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 355-386). Cambridge: Cambridge University Press.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Liberman, A. M., & Whalen, D. H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, **4**, 187-196.

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, **42**, 830-843.

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *Journal of the Acoustical Society of America*, **85**, 1314-1321.

Lotto, A. J. (1996). *General auditory constraints in speech perception: The case of perceptual contrast*. Unpublished doctoral dissertation, University of Wisconsin.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics*, **68**, 178-183.

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., 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.

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

Mann, V. A., & Liberman, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, **14**, 211-235.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]–[s] distinction. *Perception & Psychophysics*, **28**, 213-228.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America*, **69**, 548-558.

Mann, V. [A.], & Soli, S. D. (1991). Perceptual order and the effect of vocalic context on fricative perception. *Perception & Psychophysics*, **49**, 399-411.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, **25**, 457-465.

Pardo, J. S., & Fowler, C. A. (1997). Perceiving the causes of coarticulatory acoustic variation: Consonant voicing and vowel pitch. *Perception & Psychophysics*, **59**, 1141-1152.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory & Language*, **39**, 347-370.

Reinholt-Peterson, N. (1986). Perceptual compensation for segmentally conditioned fundamental frequency perturbations. *Phonetica*, **43**, 31-42.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory & Language*, **48**, 416-434.

Sapir, S. (1989). The intrinsic pitch of vowels: Theoretical, physiological and clinical observations. *Journal of Voice*, **3**, 44-51.

Silverman, K. (1986). F_0 cues depend on intonation: The case of the rise after voiced stops. *Phonetica*, **43**, 76-92.

Silverman, K. (1987). *The structure and processing of fundamental frequency contours.* Unpublished doctoral dissertation, Cambridge University.

Stephens, J. D. W., & Holt, L. L. (2003). Preceding phonetic context affects perception of nonspeech. *Journal of the Acoustical Society of America*, **114**, 3036-3039.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K.,

& Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech & Hearing Research*, **20**, 130-145.

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, **92**, 574-584.

Whalen, D. H. (1984). Subcategorical mismatches slow phonetic judgments. *Perception & Psychophysics*, **35**, 49-64.

Williams, D. R. (1986). *Role of dynamic information in the perception of coarticulated vowels.* Unpublished doctoral dissertation, University of Connecticut.

NOTES

1. There may be no need for listeners to access their speech motor system when they hear speech. However, there begins to be evidence that they do (e.g., Fadiga, Craighero, Buccino, & Rizzolatti, 2002).

2. This finding may appear contradictory to an earlier one by Mann and Liberman (1983), showing no effect of /al/ and /ar/ on *F*3 transitions. There are many differences between the studies that may be relevant to the different outcomes. However, I suspect that the most relevant is that, within a trial, Mann and Liberman did not vary the precursor syllable, whereas Stephens and Holt (2003) did. That may mean that on most trials, the contrastive effect of the precursor syllable in Mann and Liberman's test would move the target transitions in the same direction, having little impact on their discriminability. That does not explain why /al/ and /ar/ did have an impact on /da/–/ga/ discriminability. However, here the effect was to shift the discrimination peak. There was no discrimination peak in the function for the *F*3 transitions, presumably because the listeners did not form two categories along the continuum. In short, the design of Stephens and Holt appears to be superior to that of Mann and Liberman. Most likely, the context effect is a real one.

3. A reviewer commented that there are contrast effects that work in a right-to-left direction. It is true that there are findings in which the effects are qualitatively contrastive in character and that are right-to-left context effects (e.g., Diehl & Walsh, 1989; Pisoni, Carrell, & Gans, 1983). However, in my view, these effects in the speech literature should never have been given the name *contrast* and never should have been linked to the classic contrast effects underlying theoretical treatments of them, such as that of Warren (1985), precisely because they do not have the properties of those effects. Spectral contrast will turn out to be another case in point (see Experiment 3 below, in which right-to-left effects occur).

4. The predictions were not clear to me until Andrew Lotto pointed them out. I am grateful to him for setting me straight.

5. Consider, for example, expectations of motor theorists. (Direct realists, such as myself, have no account of processing to offer.) In that theory, phonetic perception is served by a module. Modules have "limited central access" (Fodor, 1983), meaning that "central" cognitive systems cannot monitor the internal workings of the module. Accordingly, in Experiment 1, the participants' modules should have, as it were, spit out "da" or "ga," without offering any insight into the processes that led to those identifications.

6. Of course, had results been larger with a left context, spectral contrast would not be the inevitable cause. A left context provides even better information than does the coarticulatory information within /da/ or /ga/ as to the nature of the coarticulatory influence.

APPENDIX

Table A1
Formant Values (Hz) of Stimuli Used in Experiment 1 Taken at Four Time Points in the Speech of the Female Speaker: Mid-Vowel and End of the First Syllable and Onset and Mid-Vowel of the Second Syllable

	Mid-Vowel	Syllable 1 End	Syllable 2 Onset	Mid-Vowel
/ada/				
*F*1	875	703	562	812
*F*2	1,375	1,125	1,906	1,531
*F*3	2,906	2,750	3,046	2,609
/aga/				
*F*1	922	625	562	906
*F*2	1,500	1,562	1,642	1,500
*F*3	2,734	2,906	2,859	2,906
/alda/				
*F*1	968	531	500	906
*F*2	1,265	1,125	1,642	1,375
*F*3	3,031	3,187	3,250	3,187
/arda/				
*F*1	843	515	562	859
*F*2	1,228	1,375	1,921	1,484
*F*3	2,343	1,718	2,921	2,687
/alga/				
*F*1	912	500	671	968
*F*2	1,237	1,406	1,515	1,406
*F*3	3,265	3,015	3,109	3,015
/arga/				
*F*1	859	406	546	875
*F*2	1,265	1,640	1,531	1,422
*F*3	2,450	2,031	2,562	2,718

APPENDIX (Continued)

Table A2
Formant Values (Hz) of Stimuli Used in Experiment 1 Taken at Four Time Points in the Speech of the Male Speaker: Mid-Vowel and End of the First Syllable and Onset and Mid-Vowel of the Second Syllable

	Mid-Vowel	Syllable 1 End	Syllable 2 Onset	Mid-Vowel
		/ada/		
*F*1	750	575	516	700
*F*2	1,275	1,425	1,734	1,300
*F*3	2,675	2,800	2,984	2,575
		/aga/		
*F*1	725	625	640	703
*F*2	1,200	1,125	1,453	1,406
*F*3	2,900	2,875	2,456	2,593
		/alda/		
*F*1	658	475	480	750
*F*2	984	1,125	1,375	1,328
*F*3	3,046	2,968	3,046	3,062
		/arda/		
*F*1	609	406	546	734
*F*2	1,125	1,468	1,546	2,187
*F*3	1,812	1,687	2,187	2,437
		/alga/		
*F*1	609	546	531	797
*F*2	984	1,562	1,593	1,422
*F*3	2,937	2,546	2,546	2,453
		/arga/		
*F*1	703	453	562	750
*F*2	1,140	1,422	1,437	1,344
*F*3	1,891	1,593	2,515	2,718

(Manuscript received April 16, 2004;
revision accepted for publication April 11, 2005.)