Journal of Experimental Psychology:
Human Perception and Performance
2010, Vol. 36, No. 4, 1005–1015

0096-1523/10/$12.00 DOI: 10.1037/a0018391

Compensation for Coarticulation: Disentangling Auditory and Gestural Theories of Perception of Coarticulatory Effects in Speech

Navin Viswanathan, James S. Magnuson, and Carol A. Fowler
University of Connecticut and Haskins Laboratories, New Haven, Connecticut

According to one approach to speech perception, listeners perceive speech by applying general pattern matching mechanisms to the acoustic signal (e.g., Diehl, Lotto, & Holt, 2004). An alternative is that listeners perceive the phonetic gestures that structured the acoustic signal (e.g., Fowler, 1986). The two accounts have offered different explanations for the phenomenon of *compensation for coarticulation* (CfC). An example of CfC is that if a speaker produces a gesture with a front place of articulation, it may be pulled slightly backwards if it follows a back place of articulation, and listeners' category boundaries shift (compensate) accordingly. The gestural account appeals to direct attunement to coarticulation to explain CfC, whereas the auditory account explains it by spectral contrast. In previous studies, spectral contrast and gestural consequences of coarticulation have been correlated, such that both accounts made identical predictions. We identify a liquid context in Tamil that disentangles contrast and coarticulation, such that the two accounts make different predictions. In a standard CfC task in Experiment 1, gestural coarticulation rather than spectral contrast determined the direction of CfC. Experiments 2, 3, and 4 demonstrated that tone analogues of the speech precursors failed to produce the same effects observed in Experiment 1, suggesting that simple spectral contrast cannot account for the findings of Experiment 1.

Keywords: compensation for coarticulation, speech perception, direct realism, articulatory, Tamil

The objects of speech perception have been a matter of considerable debate for the past 50 years. One prominent view of the basis of speech perception is that vocal gestures are perceived (e.g., Fowler, 1986; Liberman & Mattingly, 1985). The other is that the perceptual objects are auditory (e.g., Diehl et al., 2004). Here, we briefly lay out the claims of an instance of each kind of account and describe the empirical evidence cited in their support. We then focus on *compensation for coarticulation*, a context effect in speech perception that has been seen as support for both theories and has been the object of much debate regarding its basis.[1]

Two Views of the Basis of Speech Perception

The direct realist and auditory accounts of speech perception differ in how they address several fundamental phenomena of speech perception. In this section, we outline the basic tenets of each viewpoint and describe briefly the kinds of phenomena these viewpoints have marshaled as support.

The Direct Realist Account

The general ecological view of perception (e.g., Gibson, 1979), from which the direct realist theory of speech perception derives, is that events in the environment lawfully structure informational media such as light and air. This information, in the form of a lawfully structured informational array (optical, acoustic, haptic, etc.), is available to perceivers' sensory systems, which have evolved to detect this information (Gibson, 1966; Reed, 1996). Speech is viewed as an environmental event in which the speaker's phonetic gestures structure the acoustic array; listeners perceive the speech events (the distal phonetic gestures) from the information about them in the proximal acoustic signal (Fowler, 1986). Therefore, speech perception is not special in that it is neither restricted to humans, nor is the perception of distal environmental events unique to speech. In this view, the close correspondence between speech perception and speech production is attributed to the fact that objects of perception and production are identical *viz.* phonetic gestures of the vocal tract.

The most striking evidence for a gestural account comes from Liberman and his colleagues. Liberman, Delattre, and Cooper (1952) showed that the same burst of energy centered at 1440 Hz before [i] or [u] is heard as [p], whereas the same burst before [a]

Navin Viswanathan, James S. Magnuson, and Carol A. Fowler, Department of Psychology, University of Connecticut and Haskins Laboratories, New Haven, Connecticut.

Navin Viswanathan is now at the Department of Psychology, State University of New York, New Paltz.

This research was supported by NSF grant 0642300 to JSM, CAF, and NV; National Institutes of Health (NIH) grant DC00565 to JSM; and NIH grant HD01994 to Haskins Laboratories. We thank Douglas Honorof and Adam Jacks for help with stimulus creation and Annie Olmstead and Paula Silva for their helpful comments. We also thank Arthur Abramson for advice on formant offset measurement, and Arthur Samuel, Paul Luce, and an anonymous reviewer for helpful comments on an earlier draft of this article.

Correspondence concerning this article should be addressed to Navin Viswanathan, Department of Psychology, State University of New York, New Paltz, NY 12561. E-mail: viswanan@newpaltz.edu

[1] We note that there are additional accounts of CfC (see Samuel & Pitt, 2003), but we focus on the two accounts that have been invoked to explain the kind of findings we investigate in our article.

is heard as a [k]. That is, the same bit of the acoustic signal is perceived differently depending on coarticulatory context. A complementary finding showed that very different acoustic cues, a high rise in F2 in [di] and a low fall in [du] that, in natural speech would be caused by the same consonantal constriction gesture, are both heard as [d] in syllable context (Liberman, Delattre, Cooper, & Gerstman, 1954). This pair of findings prompted Liberman and colleagues to conclude, "there is typically a lack of correspondence between acoustic cue and perceived phoneme, and in all these cases it appears that perception mirrors articulation more closely than sound"(Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967, p. 453). Several other results seem to support a gestural account. For example, the McGurk effect (McGurk & McDonald, 1976) demonstrates that listeners' perception of speech is affected by visual information about vocal tract movements. From a direct realist account, this finding is evidence that perceivers utilize articulatory information irrespective of the modality in which it is available. (For an alternative explanation, see Diehl et al., 2004.) Fowler and Dekle (1991) tested the generality of this explanation by using a combination of haptic and auditory information and showed that listeners' perception of speech was similarly influenced when they felt the experimenter's vocal movements while they heard speech.

The General Auditory Approach

The principal claim of the general auditory approach is that listeners perceive the acoustic signal itself rather than speakers' vocal tract gestures. Speech perception, according to this view, is served by domain-general auditory processes and perceptual learning mechanisms that have evolved to perceive relevant environmental sounds. Diehl et al. (2004) write, "In contrast to MT [Motor Theory] and DRT [Direct Realist Theory], GA [the General Auditory framework] assumes that listeners' recovery of spoken messages from the acoustic signal (whether these messages are construed as distinctive features, phonemes, words, or some higher level linguistic units) is neither equivalent to nor mediated by the perception of gestures" (p. 154). Thus, listeners perceive speech by combing the acoustic signal for cues that enable them to activate the mental categories in their particular language that provide the best match for the signal. As the quotation above makes evident, the specific nature of these categories (if they are subphonemic, phonemic, lexical or other) is left unspecified. The close correspondence between speech production and perception is ascribed to the co-constraining goals of the perception and production systems. According to this view, the gestural production system maximizes the acoustic distinctiveness of phonemes to aid the task of an acoustically attuned general perceptual system. The perceptual system, apart from constraining the production system, is itself sensitive to the regularities that the speech production system imparts to the acoustic signal.

The empirical evidence cited in support of the general auditory account thus far falls into three different sets. The first consists of evidence that listeners perceive speech and non-speech similarly (e.g., Pastore, Li, & Layer, 1990; Pisoni, 1977). One interpretation of these findings is that listeners perform similarly during speech and nonspeech perception because they extract statistical patterns in the acoustic signal, irrespective of the nature of their source (e.g., Holt, 2005, 2006). This suggests that listeners perceive the acoustics of speech much like any other environmental sounds, and appeals to articulatory processes are unnecessary. The second set of findings demonstrates that animals can be trained to perform phonetic classification tasks with speech stimuli and post-training pattern of responses closely resembles that of human listeners (e.g., Kluender, Diehl, & Killeen, 1987; Kuhl & Miller, 1979) These findings, in addition to refuting any claims of specialization posited for speech perception mechanisms, have been interpreted as demonstrating that general learning mechanisms are responsible for the specific nature of phonetic categories (Lotto, Kluender, & Holt, 1997). The third set of findings show that phonemic contrasts tend to lie along natural auditory boundaries (e.g., Kuhl & Miller, 1978) and support the claim that the phonemic inventories of languages are largely determined by considerations of maximizing auditory distinctiveness (e.g., Diehl, Lindblom, & Creeger, 2003). This claim, termed the *auditory enhancement hypothesis* (e.g., Diehl, Kluender, Walsh, & Parker, 1991), suggests that the specific gestures of speech production are selected to maximize acoustic distinctiveness.

Distinguishing the Two Accounts

At first glance, the empirical evidence cited in support of both positions appears compelling. However, this evidence has been subjected to multiple interpretations and disagreement. One phenomenon that has been at the center of much debate is compensation for coarticulation (for other examples, see Diehl et al. [2003] and Fowler [1996]), as it has been seen as a possible avenue to adjudicate between accounts. Interestingly, *compensation for coarticulation* (CfC) has been cited as support by members of both theoretical camps. In the next section, we describe this phenomenon, briefly visit each side's interpretation of it, and describe past efforts to utilize it to distinguish between the two accounts.

Compensation for Coarticulation

Mann (1980) reported that listeners classifying members of a [da]-[ga] continuum report more "g" responses following the syllable [al] than [aɹ]. Mann' explanation for this finding is that these responses reflect listeners' perceptual adjustment to coarticulation. For instance, when producing [alga], the gestures for [l] and [g] overlap in time. That is, as a speaker is producing [l], the gesture for [g] is already underway, such that physical constraints of producing each segment conditions the realization of the other. This conditioning is referred to as *coarticulation*. In consequence of the overlap, the point of constriction during [g] may be forward of its canonical place of articulation. Similarly, after [ɹ], which has tongue body and tip gestures behind the constriction location for [d], the point of constriction during [d] may be farther back than the canonical place of articulation for [d]. Thus, on this gestural account the pattern of responses (more "g" responses after [al] and more "d" responses after [aɹ]) reflects the listener's tendency to attribute the ambiguous points of constriction (those not at the endpoints of the continuum) to coarticulation (see Figure 1).

CfC has been used as evidence for the direct realist theory (e.g., Fowler, 2006). From this perspective, CfC follows logically from a perceptual system that is attuned to the temporal overlap of the gestures of speech production. From a direct realist viewpoint, a

listener attuned to such a production system is able to adjust to the effects of this gestural overlap during perception. Thus, the boundary shifts observed in typical CfC studies (e.g., Mann, 1980) are attributed to listeners' tendency to perceive non-endpoint members of the continuum (that would correspond to intermediate constrictions) as instances of [ga] that have been pulled forward after [al], or as instances of [da] pulled back after [aɹ].

Several studies provide support for the gestural perspective. For instance, listeners show compensation irrespective of their linguistic experience (e.g., Japanese listeners cannot reliably identify [al] and [aɹ] [Mann, 1986], and 4-month-old infants whose perceptual categories may not be strongly learned [Fowler, McRoberts, & Best, 1990]). From a gestural perspective, this is evidence that listeners are attuned to the speakers' gestures and that this attunement is not dependant on specific linguistic experience (see Diehl et al., 2003, for an alternative interpretation—specifically, that a contrast account makes the same prediction). While these findings deal with compensation for left-to-right coarticulatory effects (carryover coarticulation), other findings show that listeners compensate for anticipatory (Mann & Repp, 1980) and simultaneous (Silverman, 1986) coarticulation as well. Furthermore, visually presented articulatory information appears to drive CfC even in the absence of disambiguating acoustic signal (e.g., Mitterer, 2006). While Mitterer interpreted these results as evidence for phonological basis of the effects, from a gestural perspective, such findings provide evidence that articulatory information is utilized by perceivers in their perception of speech irrespective of the modality by which it is perceived. In general, the fact that listeners appear to be sensitive to patterns of speakers' vocal tract gestures and exhibit responses that reflect their attunement to this articulatory information is seen as support for a direct realist viewpoint.[2]

An alternative explanation for CfC was advanced by Mann (1980) herself in the discussion of her results. In her words:

> The design of the present experiment does not eliminate the possibility that the results are due to some auditory interaction involving the VC offset and CV onset spectra. For example, the contrasting effects of [l] and [ɹ] could conceivably be the consequence of some form of *auditory contrast* between the concentration of energy in the F3 region at the end of the preceding VC and that in the F3 region in the beginning of the following CV. Perhaps the relatively higher F3 offset frequency in [l] led to the perception of a lower F3 onset frequency in the following CV syllable, leading to more [g] percepts. (pp. 410–411, emphasis added)

That is, just as perception of the warmth of lukewarm water changes after placing one's hand in cold vs. hot water, listeners' responses ("d" or "g") might depend on the frequency profiles of the preceding and following segments in CfC studies. In the particular example of a [da]-[ga] continuum following [al] or [aɹ], the crucial contrast is in the frequency of the F3 offset of the preceding syllable relative to the F3 onset of the CV. [al] and [da] both have high F3 offsets as compared to [aɹ] and [ga]. Thus, according to a contrast account, after hearing a high F3 in [al], the following stop's F3 is heard as lower and consequently more [ga]-like, whereas hearing a low F3 in [aɹ] has the opposite effect of making the following stop's F3 seem higher and more [da]-like. Lotto and Kluender (1998; Experiment 3) put this explanation to empirical scrutiny by replacing the precursor syllables [al] and [aɹ] with a steady high tone at the F3 frequency offset of [al], or a steady low tone at the F3 frequency offset of [aɹ]. Listeners exhibited a pattern of responses qualitatively like those found with contexts [al] and [aɹ], suggesting that this effect can be explained on the basis of the contrastive properties of the acoustic signal itself (specifically their interaction with the auditory system), instead of invoking an underlying gestural cause. Thus, according to the general auditory view, the spectral characteristics of the precursor and target syllables are responsible for the resulting pattern of judgments.[3]

In addition to the remarkable effects of offset frequency matched tones obtained by Lotto and Kluender (1998), several other findings appear to support a contrast viewpoint. For instance, Japanese quail trained to identify [da] and [ga] tokens demonstrate a CfC-like pattern of responses when made to identify the stops in a liquid context (Lotto, Kluender, & Holt, 1997). Furthermore, Wade and Holt (2005) show that later occurring tones affect the perception of precursor speech, demonstrating that these effects are not restricted to left-right contexts. Stephens and Holt (2003) showed that hearing a preceding speech syllable has similar effects on the discrimination of the target stimuli whether it is speech or non-speech. This appears to strengthen the auditory account further by showing the converse effect that the perception of non-speech is affected by hearing speech. Finally, Holt (2005) showed that spectral contrast effects occurred in temporally non-adjacent contexts computed over multiple events, suggesting a role for cortical mechanisms rather than solely the temporary habituation of the auditory nerve. She showed that contrast effects occur when the mean frequency of a set of tones was either manipulated to be high or low. These effects persist when the several preceding tones were separated from the target either by silence (as long as 1,300 ms), a standard tone (a tone intermediate between the high and low frequencies that occurs in both conditions) or by as many as 13 such standard tones.

Proponents of auditory contrast and the direct realist accounts continue to debate the basis of CfC in the light of these empirical studies. Even though each account has marshaled several empirical findings in its support, these have not resolved the issue. Of the findings that are deemed problematic for either account, some have been amenable to alternative explanations from the account they are said to contradict (e.g., animal studies of CfC are interpreted by gestural theories as unsurprising confirmation that animals are sensitive to auditorily specified events in their environment, and visual effects on speech perception from a general auditory perspective are attributed to associative learning). Others have been explained away as being outside the purview of the accounts they are meant to evidence against (e.g., tonal context effects are claimed to be outside the scope of a gestural viewpoint and

[2] By definition, this information is in the acoustic signal but is about the underlying articulations. This contrasts with seeking explanations at the level of surface acoustic properties (e.g., auditory contrast) or at the level of learned contingencies in the acoustical signal without considering their causes (e.g., statistical learning accounts like that of Elman & McClelland, 1986).

[3] Note that other explanations are possible within the "general auditory" account. If the contrast account were known to be untenable, one could appeal to statistical regularities: the acoustic cues to place of articulation changes systematically as a function of coarticulatory context, and listeners have learned these contingencies.

compensation for simultaneous coarticulation outside that of a contrast account). Finally, the entire class of findings employing a speech precursor and a speech target (e.g., Mann, 1980, 1986) are predicted equally well by both the direct realist and the general auditory accounts, albeit for different reasons. This poses a dilemma, in that it suggests it may be virtually impossible to distinguish the two accounts empirically.

To decide between them, we could use the original paradigm and attempt to find a context in which the articulatory and acoustic properties are not correlated. In this article, we identify such a context in which the accounts make opposite predictions for the specific coarticulatory context both accounts were proposed to explain in the first place. Our first experiment involves a speech-speech context in which the accounts differ in their predictions and our second, third, and fourth experiments are designed to investigate whether the results of our first experiment are replicated using analogous tonal precursors.

Experiment 1

Our starting point is a pair of Tamil liquids: a trilled "r" that has a frontal, alveolar place of articulation (phonetic symbol), and a retroflex liquid [ɭ] with a palatal place of constriction similar to the American English (AE) [ɹ] (see Narayanan, Byrd, & Kaun, 1999, for complete articulatory and acoustic descriptions of the Tamil liquids). In terms of place of articulation, we find that Tamil [ɾ] and English [l] both have a front place of articulation, in contrast to Tamil [ɭ] and English [ɹ] that have a back place of articulation (see Narayanan et al. 1999). However, these syllables are grouped differently by their F3 values. [ɹ], [ɾ], and [ɭ] have a low F3, whereas [l] has a high F3 (see Table 1, also Narayanan et al., 1999). The F3 offset values of both the Tamil liquids are close to that of the AE [ɹ] and are expected to produce a contrastive effect on the onset F3 of the stop consonant. Thus, these phones allow a disentangling of place of articulation and F3 as possible causes of CfC. In addition to the Tamil liquids, AE liquids [ɹ] and [l] were also presented to verify that we are able to replicate previous studies and to allow comparison of responses to the two sets of liquids.

Predictions

Auditory contrast and gestural predictions are different for the four liquids, [ɹ], [l] (English), and [ɾ], [ɭ] (Tamil). Auditory contrast predicts that responses to the [da]-[ga] continuum following the liquids should pattern according to the F3 offset of the liquids: [ɹ], [ɾ], and [ɭ] should all result in increased "da" responses, while [l] should result in increased "ga" responses. The gestural prediction is that the results should pattern according to place of articulation: [ɹ] and [ɭ], with back places of articulation, should lead to more "da" responses, while [l] and, with more frontal places of articulation, should lead to more "ga" responses (Figure 1).

Method

Participants. Thirteen University of Connecticut undergraduates, who were native speakers of American English with no prior exposure to Tamil, participated for course credit. All reported normal hearing.

Materials. The initial vowel consonant (VC) syllables were produced by a 25-year-old male (NV) trilingual speaker of Indian English, Tamil, and Hindi (coached on AE liquids by a phonetician, who also verified that the results were native quality). The Tamil liquids were judged as native quality by four native Tamil speakers naive to the purpose of our study. Four VC syllables ([al], [aɹ], [ar], and [a(ɭ)]) were used as precursors. Each syllable was 375 ms in duration, shared the same vowel onset, and matched in intensity to the CV target syllable. Equal duration was achieved by choosing natural tokens of the different precursors within 30 ms of each other and editing the steady vowel portion of the syllable to exactly same durations. Spectrograms of these precursors are presented in Figure 2.

An 11-step series of resynthesized CV syllables varying in F2- and F3-onset frequency and varying perceptually from [da] to [ga] was created using the source-filter method with the Praat software package (Boersma & Weenink, 2006). For this continuum, F3-onset frequencies varied linearly from 2200 Hz ([ga]) to 2400 Hz ([da]). The F2-onset frequencies varied linearly from 2000 Hz ([ga]) to 1400 Hz ([da]). The first and fourth formants were the same for all members of the continuum (F1 varied from 450 Hz to 600 Hz and F4 from 3200 Hz to 3500 Hz). Each of the four formants had 45 ms of transition before reaching steady state values of 600 Hz, 1200 Hz, 2400 Hz, and 3500 Hz, respectively, to signal the vowel. The overall duration of the CV syllables was 300 ms. The filter characteristics of our target continuum were based on analysis of endpoint [da] and [ga] utterances by the same speaker. The stop continuum was generated by exciting the filters by the LPC residual of a neutral utterance [a] produced by the speaker ensuring that the precursor and target were matched in source.

The syllables were combined with a closure interval of 80 ms between the VC and CV syllables. The stimuli were presented at an 11 kHz sampling rate with 16 bit resolution diotically over headphones at approximately 70 dB SPL. Although these parameters are slightly different from those employed in previous studies

Table 1
Formant Offset Frequencies, Place of Articulation, and Categorization Responses for the English and Tamil Liquids

Variable	Formant offsets in Hertz				Place of articulation	Heard by native listeners as
	F1	F2	F3	F4		
[al]	570	1060	2600	3600	Front	"l"
[aɹ]	600	1350	1800	3050	Back	"r"
[ar]	650	1440	2010	3610	Front	"r"
[a(ɭ)]	470	1600	1780	3100	Back	"r"

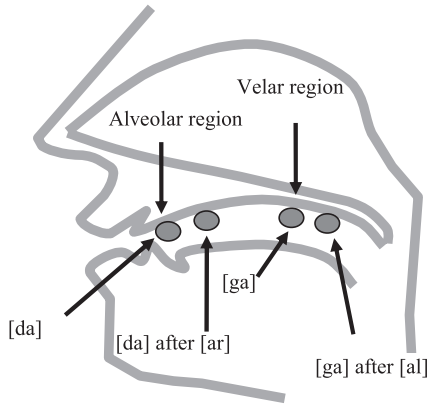

Figure 1. Effects of coarticulation on place of articulation.

(e.g., Lotto & Kluender, 1998), they were modeled after natural endpoints produced by the speaker. We also ensured that the resulting endpoints were clearly discernable and were judged as satisfactory instances of [d] and [g] by pilot participants. We used a narrow band spectrogram (512 points) to measure the formant offset values of our syllables. Offsets were measured at 340 ms, where the critical offsets were visible in the spectrogram and stable across the syllables. If not apparent (e.g. F4 of [ar]), the measurement was made at the point closest to 340 ms where the formant was visible.

Procedure. The task was a two-alternative forced-choice: participants pressed keys labeled "d" or "g" to indicate their identification of the stop. The session consisted of three blocks of trials. The first block consisted of practice trials on the [da] and [ga] endpoints with feedback. There were 9 trials with each endpoint, presented in random order. This familiarized participants with the task and syllables, and provided a basis for ensuring that they could identify the endpoints accurately.

In the second block, all items from the 11-step [da]-[ga] continuum were presented in liquid contexts without feedback. Following the procedure used by Mann and Repp (1980), the stop continuum items were presented in ratios of 1-1-2-2-3-3-3-2-2-1-1, such that midpoints in the continuum were presented more often. This provided more responses for the ambiguous steps where the strongest shift is expected, increasing power there while restricting the total number of trials to a fairly modest number. The second block consisted of trials in four conditions differing in precursor syllable ([al], [aɹ], [a(l̩)], and [ar]). Each condition had 42 trials (twice the sum of the ratios). Thus, each participant was presented with the endpoints (Steps 1, 2, 10, and 11) twice; Steps 3, 4, 8, and 9 four times each; and Steps 5, 6, and 7 six times each, in each of the four liquid context conditions. The entire set of 168 trials (42 × 4 conditions) was randomized, and participants could take a break after every 42 trials.

In the final block, we established liquid categorization by presenting the precursor syllables in isolation and changing the task to a forced choice between "l" and "r". Each precursor was presented

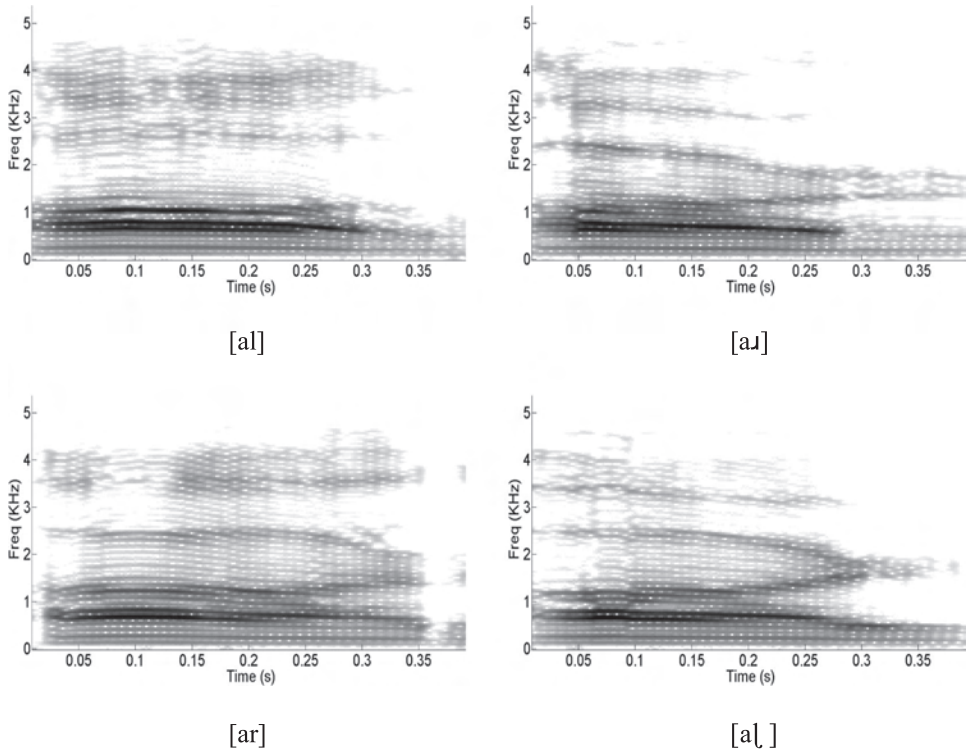

Figure 2. Spectrograms of native and nonnative precursors used in Experiment 1. The top two panels display the English precursors ([al] and [aɹ]) and the Tamil precursors ([ar] and [al̩]) in the bottom two panels.

four times, and the set of 16 was randomized. The experiment lasted about 25 min.

Results

We excluded data from one participant, whose accuracy was less than 80% in the stop endpoint task.[4] Participants exhibited 100% agreement in their classification of the precursor syllables; [aɹ], [ar] and [a(l̩)], were classified as [ɹ], and only [al] was classified as [l].

Figure 3 shows the results of the critical second block, in which the VC-CV stimuli were presented without feedback. Consistent with a gestural account, the results pattern according to place of articulation, with more "g" responses following the liquids with front place of articulation ([al] and [ar]) than the liquids with back place of articulation ([aɹ] and [a(l̩)]). The auditory contrast prediction based on spectral contrast with F3, that [ar] and [a(l̩)] should pattern with [aɹ], and all three should differ from [al], was not supported.

We used a 4 (precursor) × 11 (step) within-subjects analysis of variance (ANOVA) to evaluate percentage of "g" responses. There were significant effects of precursor, $F(3, 33) = 11.011$, $p < .001$, $\eta_2^p = 0.50$, indicating that listeners' responses depended on the liquid precursor. The effect of step was significant, $F(10, 110) = 163.374$, $p < .001$, $\eta_p^2 = 0.94$, indicating that percentage of "g" responses changed across the stop continuum. An interaction, $F(30, 330) = 1.49$, $p = .051$, $\eta_p^2 = 0.13$, was also observed, indicating that the effect of precursor was different at different points along the continuum (see Figure 3, where it is apparent that identification does not vary systematically between precursors near the endpoints).

We further investigated the results with three planned comparisons designed to test predictions of gestural and contrast theories. First, we tested the gestural prediction that the basis for compensation is place of articulation by comparing segments with front place of articulation ([al] and [ar]) and those with back place of articulation ([aɹ] and [a(l̩)]). The effect of place of articulation was reliable ($p = .001$, $\eta_p^2 = 0.65$). Next, we tested two additional comparisons in which auditory contrast and gestural predictions conflict. First, we compared [al] and [ar]. Because these have the same place of articulation, an articulatory account predicts no difference. Since they differ in F3 by as much as [ɹ] and [l], a contrast account predicts a reliable difference. The contrast was not significant ($F < 1$, $\eta_p^2 = 0.07$). Second, we compared [ar] and [aɹ]. These differ in place of articulation, so a gestural account predicts a reliable difference. They are approximately matched in F3 (see Table 1), so auditory contrast predicts a null result. The comparison was significant ($p = .008$, $\eta_p^2 = 0.49$). These planned comparisons confirm the pattern visible in Figure 3: place of articulation predicts the direction of compensation effects; spectral contrast does not.

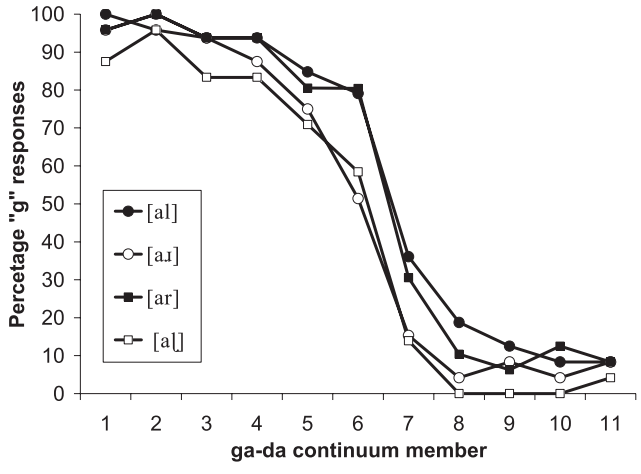

Figure 3. Experiment 1: Effect of native and non native precursors on stop judgments. Closed symbols indicate frontals ([al] and [ar]), and open symbols indicate items with back place of articulation ([aɹ] and [a(l̩)]). Circles indicate English precursors and squares indicate Tamil precursors.

We studied the spectral characteristics of the precursors to see whether there were other parts of the spectrum that could cause a contrast effect that would lead to the pattern of results observed in Experiment 1. The first possibility is whether our results reflect spectral contrast effects of the precursor's F3 on the stop judgment. Consider the English syllables: the F3 offset (at 2600 Hz) of [al] is relatively higher than the range of F3s for the stop continuum (2,200 Hz to 2400 Hz) whereas that of [aɹ] is 1,800 Hz and relatively lower than the F3 onsets of the target stops. Thus, consistent with previous studies we expect pure tone analogues of these syllables (e.g., Lotto & Kluender, 1998) to produce response patterns similar to those elicited after natural speech precursors. However, consider the Tamil syllables: both [ar] and [a(l̩)] have F3 offsets of 2,010 Hz and 1,780 Hz, respectively, that are both lower than the F3 onsets of the stop continuum. Pure tone analogues of these syllables are both expected to produce fewer "g" responses and this patterning would be unlike those obtained with speech tokens of the Tamil liquids as precursors. We test these predictions in Experiment 2.

Experiment 2

In Experiment 2, we adopted the analogous pure tone precursor method used by Lotto and Kluender (1998) to isolate F3 frequency offsets of the materials from Experiment 1 as a potential source of contrast.

Method

Participants. Thirteen University of Connecticut undergraduates, who reported normal hearing, participated for course credit. None had participated in Experiment 1.

Materials. The [da]-[ga] continuum from Experiment 1 was used. A steady state sinewave tone at the F3 offset of each liquid was used as its analogue, yielding a steady tone at 2600 Hz for [al], 1800 Hz for [aɹ], 2010 Hz for [ar] and 1780 Hz for [a(l̩)]. Following Lotto and Kluender (1998), the intensities and durations of the precursor tones were matched to the overall intensity (70 dB) and duration of precursor syllables (375 ms) used in Experi-

[4] The results of our three experiments were unaltered both qualitatively and in terms of their statistical significance when none of the subjects were excluded. Adopting a 90% cutoff as used by Lotto and Kluender (1998) excluded one more subject in Experiment 3 but did not alter the general pattern of our results.

ment 1. This choice ensured that our parameters were consistent with previous studies, and also provided conditions most favorable for observing boundary shifts with tonal analogues (Viswanathan, Fowler, & Magnuson, 2009, found that effects were stronger when "matched" to syllable intensity and duration—as in Lotto & Kluender, 1998—than when they were matched to formant-specific parameters).

As in Experiment 1, forty-two tokens (two each of continuum members with each precursor in the ratios, 1-1-2-2-3-3-3-2-2-1-1) were presented for each precursor tone for a total of 168 trials, presented in random order.

Procedure. The procedure of Experiment 1 was used, except that the liquid identification task was not included.

Results

Twelve subjects made the 80% accuracy cutoff in the stop endpoint task, and their data were included in the analysis (see note 3). Figure 4 shows the pattern of responses. We used a 4 (precursor) × 11 (step) within-subjects ANOVA to evaluate percentage of "g" responses. There were significant effects of precursor, $F(3, 33) = 10.19$, $p < .001$, $\eta_p^2 = 0.48$, indicating that listeners' responses depended on the tone analogue they heard. The effect of step was significant, $F(10, 110) = 98.48$, $p < .001$, $\eta_p^2 = 0.90$, indicating that percentage of "g" responses changed across the stop continuum. An interaction, $F(30, 330) = 1.67$, $p = .018$, $\eta_p^2 = 0.13$, was also observed indicating that the effect of precursor was different at different points along the continuum (see Figure 4).

A planned comparison indicated that we replicated the results of Lotto and Kluender (1998) with the English liquids: the tone analogue of [al] produced more "g" responses than the analogue of [ar] ($p < .001$) However, as expected from the frequency relations, neither of the Tamil liquids produced responses that were reliably distinct from English [aɹ] or from each other ($p > .3$ for all comparisons). Thus, the F3 tone analogues fail to replicate the patterning of responses observed with speech precursors observed in Experiment 1, indicating that a solely F3-based contrast explanation of our results with speech precursors is untenable.

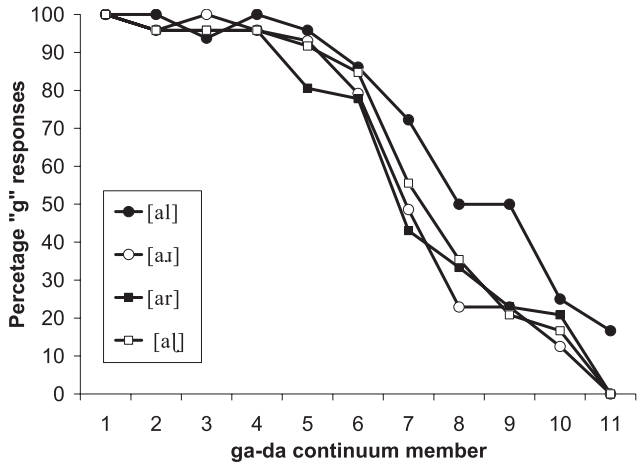

Figure 4. Experiment 2: Effect of pure tone precursors at F3 offsets of precursor liquids from Experiment 1. Closed symbols indicate frontals ([al] and [ar]), and open symbols indicate items with back place of articulation ([aɹ] and [a(|)]). Circles indicate English precursors and squares indicate Tamil precursors.

However, one crucial difference between our speech stimuli and those of Lotto and Kluender (1998) is that our stops varied in both F2 and F3 along the continuum. This leads to the possibility that relevant source of contrast may not be F3 alone, but a combination of F3 with F2 and/or F4. In fact, Lotto, Sullivan, and Holt, (2003) showed that tone analogues consisting of a combination of pure sine tones (sinewave dyads) at F2 and F3 offsets of the English liquids [al] and [aɹ] produce responses indistinguishable from the natural syllables themselves. In Experiment 3, we test a contrast account for our syllables based on a combination of F2 and F3 using sinewave dyad analogues, and in Experiment 4, we consider the possible influence of F4.

Experiment 3

In this experiment, we used sinewave dyads composed of two pure sine tones at the F2 and F3 offsets of each liquid. Specifically, we tested whether simultaneously presenting tones reflecting the appropriate F2 and F3 relations might replicate the Tamil context effects of Experiment 1.

Method

Participants. Eighteen University of Connecticut undergraduates who reported normal hearing participated for course credit. None had participated in Experiment 1 or Experiment 2.

Materials. Four precursor sinewave dyads were designed to serve as the tone analogues of the syllables used in Experiment 1 were used. For each syllable ([al], [aɹ], [ar] and [a(|)]), a combination of precursor tones at their second and third formant offset frequencies was synthesized (see Table 1 for offset values). The tones were matched for intensity and duration of the precursor syllables used in Experiment 1. Forty-two tokens (two each of continuum members with each precursor in the ratios used for Experiments 1 and 2, 1-1-2-2-3-3-3-2-2-1-1) were presented for a total of 168 trials.

Procedure. The procedure of Experiment 2 was used.

Results

All participants made the 80% accuracy cutoff in the stop endpoint task (see note 3). Figure 5 shows the results. The relative ordering of responses to [al], [aɹ], [ar] and [a(|)]. was the same as in Experiment 2 and distinct from the ordering in Experiment 1 (see Figure 5). A 4 × 11(precursor tone × step) ANOVA was used to analyze the data. Again, the effects of precursor tone, $F(3, 51) = 12.82$, $p < .001$, $\eta_p^2 = 0.43$), step ($F(10, 170) = 229.01$, $p < .001$, $\eta_p^2 = 0.93$), and the interaction ($F(30, 510) = 2.67$, $p < .001$, $\eta_p^2 = 0.14$) were significant. Planned comparisons indicated again that the [al] analogue produced more "g" responses than the [aɹ] analogue ($p < .001$), replicating (Lotto et al., 2003). The analogues to the Tamil liquids did not differ from each other or from the analogue of [aɹ] ($p > .15$, for all comparisons). It is clear that dyads matched to F2 and F3 fail to replicate the pattern of results observed in Experiment 1. This indicates that a contrast based on a combination of F2 and F3 does not underlie compensation observed in Experiment 1. Before discarding a contrast account,

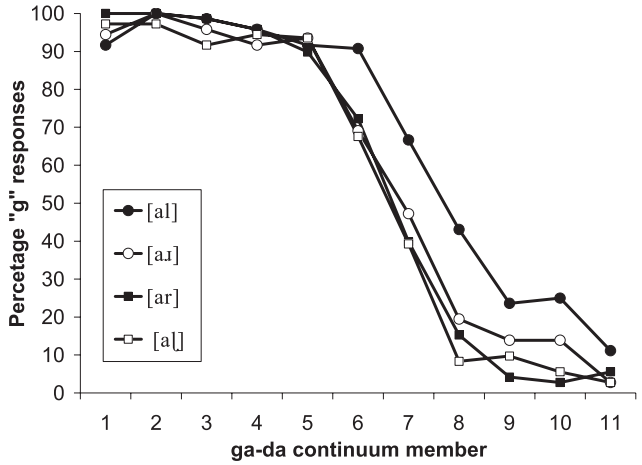

Figure 5. Experiment 3: Effect of pure tone dyads at F2 and F3 offsets of precursor liquids from Experiment 1. Closed symbols indicate frontals ([al] and [ar]), and open symbols indicate items with back place of articulation ([aɹ] and [a(ɭ)]). Circles indicate English precursors and squares indicate Tamil precursors.

we examine one final possibility: if we group the liquids by F4 (High F4: [al] and [ar]; Low F4: [aɹ] and [a(ɭ)]), we get the same groupings as in Figure 3. Therefore, an alternative interpretation of the results is that perhaps this is a reflection of an auditory contrast caused by F4 offset differences rather than F3 offset differences. However, F4 tones cannot differentially affect the stop judgments by producing a contrast in the perceived F4 of the stops because [da] and [ga] in our continuum have identical F4. Any effect of F4 would have to be on F3 or F2 of the stops. In Experiment 4, we investigate whether providing a combination of tones (chords) at F2, F3, and F4 is sufficient to replicate the results of Experiment 1.

Experiment 4

In this experiment, we used sinewave chords composed of pure sine tones at the F2, F3, and F4 offsets of each liquid to test the role of F4 in conjunction with F2 and F3 from a contrast perspective.

Method

Participants. Thirteen University of Connecticut undergraduates who reported normal hearing participated for course credit. None had participated in any of the previous experiments.

Materials. Four precursor sinewave chords were used as analogues of the liquids. For each syllable ([al], [aɹ], [ar], and [a(ɭ)]), a combination of pure sine tones at their second, third and fourth formant offset frequencies was synthesized (see Table 1 for offset values). The tones were matched for intensity and duration of the precursor syllables used in Experiment 1. Forty-two tokens were presented for each precursor tone in the same ratios as Experiments 1, 2, and 3, and were presented for a total of 168 trials.

Procedure. The procedure of Experiment 2 was used.

Results

All participants made the 80% accuracy cutoff in the stop endpoint task, and their data were submitted for analysis (see note 3). Figure 6 shows the results. Again, the relative ordering of patterning was similar to the pattern of results obtained in Experiments 2 and 3 and distinct from that of Experiment 1. A 4 × 11(precursor tone × step) ANOVA was used to analyze the data. As in the previous experiments, the effect of precursor tone, $F(3, 36) = 13.02$, $p < .001$, $\eta_p^2 = 0.52$; step, $F(10, 70) = 144.14$, $p < .001$, $\eta_p^2 = 0.92$; and the interaction, $F(30, 360) = 2.14$, $p = .001$, $\eta_p^2 = .15$, were significant. Planned comparisons indicated that the [al] analogue produced more "g" responses than [aɹ] ($p < .001$), similar to our previous tone-analogue experiments. Responses to the analogues to the Tamil liquids did not differ from each other ($p = .70$). However, unlike Experiments 2 and 3 both Tamil liquids produce fewer "g" responses than the analogue of [aɹ] ($p < .05$, for all comparisons). Although it is unclear why this occurs, critically for our purposes, it is clear that sinewave chords matched to the F2, F3, and F4 offsets of the liquids do not produce results similar to those of Experiment 1.

General Discussion

The general auditory and direct realist accounts of CfC have been difficult to distinguish because F3 (the acoustic cue that proponents of auditory contrast accounts have claimed drives "compensation" effects for materials like ours) was correlated with place of articulation in previous studies that used American English [ɹ] and [l]. The Tamil liquids [ɾ] and [ɭ] we used in Experiment 1 along with [ɹ] and [l] provide a crucial test case in which F3 and place of articulation are disentangled. Tamil has a front place of articulation (like English [l]) but its F3 is similar to that of English [ɹ] (which has a relatively back place of articulation). Tamil [ɭ] has a back place of articulation and also has an F3 similar to that of English [ɹ]. In Experiment 1 with English and Tamil precursor syllables, we found that place of articulation, rather than F3, predicts compensation effects on responses to stimuli along a following stop continuum. Experiments 2, 3, and 4 tested whether other candidate acoustic cues (pure sine tones at F3 offsets, pure

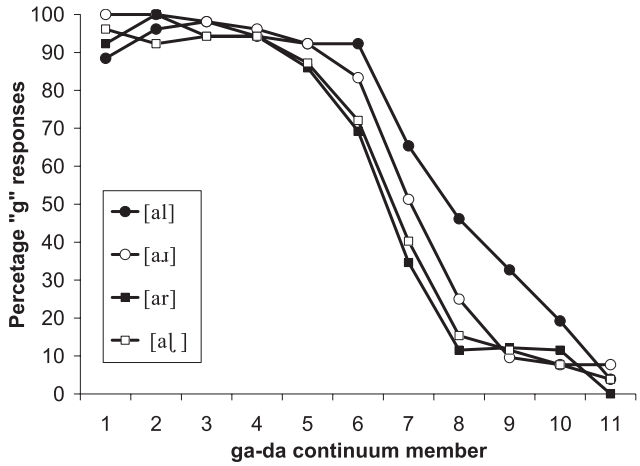

Figure 6. Experiment 4: Effect of pure tone chords at F2, F3, and F4 offsets of precursor liquids from Experiment 1. Closed symbols indicate frontals ([al] and [ar]), and open symbols indicate items with back place of articulation ([aɹ] and [a(ɭ)]). Circles indicate English precursors and squares indicate Tamil precursors.

sine dyads at F2 and F3 offsets, or pure sine chords at F2, F3, and F4 offsets) might provide a spectral contrast explanation for Experiment 1. None of the cues did. In addition, if we attempt to appeal to contrast based on other as yet unspecified portions of the acoustic stream to explain the results of Experiment 1, we would face additional challenges. For example, the contrast produced due to these acoustic characteristics, whatever they might be, would have to overcome the contrast that is predicted by the spectral contrast account due to the presence of the particular formant offsets we investigated in Experiments 2 through 4. Before completely ruling out an auditory contrast account of CfC, we further examine two candidate explanations for why pure tone analogues did not replicate effects obtained with natural speech.

The first possibility is that pure tone analogues do not faithfully represent the formants in natural speech that causes the auditory contrast. Indeed, these tones differ from the critical speech formants in intensity (their intensity is syllable-matched rather than formant matched), frequency profile (do not capture the transitions), and bandwidth (they do not reflect the complexity of speech formants). In a recent study, we investigated the effects of matching such tone analogues to appropriate conditions in speech (Viswanathan et al., 2009). We used only the English precursors [al] and [aɹ] or their pure tone analogues as precursors. Our target stops varied only in their F3 onset along the [da]-[ga] continuum. We found that the magnitude of category shifts induced by these pure tone precursors diminished as they were made more like speech-like. In fact, while the F3 region of natural syllables must be assumed to contain a sufficient basis for contrast on the auditory account, when we presented only the F3 region of natural tokens of these syllables, we failed to induce *any* boundary shifts. These findings again pose serious challenges for an auditory contrast account of CfC and disconfirm the possibility that our results are due to our tones being insufficiently analogous to formants in natural speech.

The second possibility is that listeners' sensitivity to the relative spread of F2 and F3 in these liquids causes the category shifts obtained with our speech precursors.[5] The suggestion here is that contrast operates not by changing the perceived onset frequency of the stops as has been claimed so far, but by changing the perceived spread of energy in the critical region (i.e., the distance between F2 and F3). While we have not directly tested this possibility, a couple of points must be noted. First, sine-tone dyads at F2 and F3 offsets of these liquids fail to replicate our findings with natural speech precursors. Perhaps it could be argued that pure tones do not engage the mechanism(s) responsible for this sensitivity. If we examine the distance between F2 and F3 in our natural precursors using the difference between their offsets as an approximate measure, we arrive at following values of spread: 1540 Hz for [al], 450 Hz for [aɹ], 570 Hz for [ar] and 180 Hz for [a(l̩)]. From these values, it is unclear why liquids [aɹ] and [ar] that differ by less than 100 Hz of spread would pattern differently, whereas [al] and [ar] that differ by almost 1000 Hz of spread would pattern together. While these post-hoc examinations do not directly disconfirm such a contrastive account of spread and direct tests of this explanation is needed, in our judgment, it is unlikely that our findings could be explained by such an account.

Thus, our experiments suggest that auditory contrast is insufficient to explain CfC. Although it accounts for effects of [l] and [ɹ], it does not generalize to Tamil [l̩] and [r]. The gestural assumption that CfC is literally *compensation for coarticulation* predicts that compensatory behavior should pattern with place of articulation rather than any single cue in the acoustic signal that varies independently of place of articulation. This prediction was borne out by the current studies. However, numerous past studies (e.g., Lotto & Kluender, 1998) show clearly that pure tone precursors (that presumably are devoid of articulatory information) produce robust shifts in identification. This was shown as well in our Experiments 2 through 4. The direct realist account does not naturally account for such tonal effects.

However, we speculate (see also Fowler, Brown, & Mann, 2000) that hearing pure tones may interfere with the ability of the listener to pick up information in the vicinity of the tone-frequency about oral gestures. In other words, we suggest that these pure tones might induce a form of informational masking. Some indirect support for this speculation comes from the study we outlined earlier in which we found that the potency of the boundary shifts induced by the pure tone precursors depended on how *unlike* speech they were (Viswanathan et al., 2009). Specifically, high-intensity (matched to the intensity of the whole syllable) steady tones produced much stronger effects on speech categorization than low-intensity (matched to the F3-formant-intensity) tones with time-varying properties matched to natural formant transitions, even though both kinds of tones shared their offsets. Similarly, Mitterer (2006) failed to find effects of pure tone precursors on subsequent speech categorization when the precursor tones were matched to the formant rather than syllable intensity. However, when Holt and Lotto (2002) examined one masking account of these tonal effects (also Lotto et al. 2003), they showed that shifts in target categorization occur both when the target and the precursor were presented diotically and when they were presented dichotically. These studies indicate that the context effects cannot be solely peripheral. However, they do not show that central masking does not occur or that there are no peripheral masking effects (see Lotto et al., 2003).

Another interesting finding is that even though our participants classified both Tamil liquids as instances of [ɹ], they compensated according to their place of articulation. This finding replicates, in a different language setting, Mann (1986)'s finding that a group of Japanese listeners who could not identify [l] and [ɹ] accurately, nonetheless showed compensation for their coarticulatory effects. However, there is a crucial difference. In Mann (1986) the relevant liquids were different in both their place of articulation and the F3 offsets. Our liquids differed in place of articulation but had nearly identical F3 offsets.

The explanation from a general auditory viewpoint is that language-universal compensation effects reflect spectral contrast and, because the different language users possess the same auditory system, no difference among them is expected. However, this cannot explain the language-universal compensation exhibited by our listeners. Even though our American listeners categorize both the Tamil liquids (that both have low third formant offsets) as instances of "r", they compensate differently for them according to the liquid place of articulation.

Alternatively, from a gestural perspective, this behavior reflects listeners' general ability to perceive environmental events (in this

[5] We thank Arthur Samuel for suggesting this possibility.

case, vocal tract gestures) and unsurprisingly this ability is language independent (given, for example, that listeners show compensation even in the context of non-native speech contrasts they cannot reliably categorize; Mann, 1986). This explanation, we believe, is still applicable to our results. Our listeners, when exposed to unfamiliar Tamil phonemes, still compensate for the effect of precursor place of articulation on the following stop despite having no prior experience with these particular liquids.

The finding that place of articulation predicts compensation patterns is interesting whether or not one subscribes to a direct realist account, and it raises different challenges for the two accounts under consideration here. Minimally, our results demonstrate that the acoustic patterns that underlie compensation for coarticulation are more complex than the simple contrasts proposed under the general auditory account. Furthermore, those patterns likely involve more complex properties of the speech signal than are typically extracted using current acoustic analysis techniques – such as, perhaps, the dynamic effects of gestures over extended and overlapping spans of time. As we mentioned earlier, direct realism assumes that the acoustic speech signal contains information that specifies the gestures that caused it. An outstanding challenge, then, is to identify the information in the acoustic speech signal that specifies place of articulation, which in turn drives CfC.

We started out creating an experiment with the potential to falsify either account. Had judgments in the CfC paradigm patterned with spectral contrast predictions when they were uncorrelated from place of articulation, the gestural account would have been falsified. Instead, judgments patterned with place of articulation (that is, with the a priori predictions of the direct realist account), falsifying a spectral contrast account of CfC in human speech. Even though our results naturally follow from the theoretical perspective of direct realism, we can state the empirical challenge raised by our results in a theory-neutral fashion. No account of speech perception has yet provided a satisfactory explanation of how listeners achieve stable phonetic percepts despite rampant variability in the acoustic signal. It is undeniable that gestures are what cause the acoustic structure available to listeners; the theories differ on the perceptual status of acoustics (information for gestures in direct realism, perceptual object in the general auditory account). When judgments pattern with place of articulation, both sides must assume the judgments are guided by information in the acoustic signal specifying (on a direct realist account) or at least correlated with (on the general auditory view) place of articulation. Thus far invariant acoustic information for articulation has eluded us using conventional acoustic analyses of speech. Perhaps gestures may provide a theory-neutral organizing principle. That is, it may be that both the general auditory and direct realist programs can be advanced by searching for ways to characterize the dynamics of the acoustic consequences of gestural sequences. Such descriptions might be the key to understanding phonetic constancy.

Investigating whether there are systematic patterns linking acoustics and the dynamics of underlying articulations also has the potential to clarify theoretical debates regarding the objects of speech perception. Again, direct realism posits that there is an invariant mapping from acoustics to gestures; this must be so for the proximal stimulus to specify the causal event that is perceived (however, see Fowler, 1994). This implies that if mental categories hypothesized by indirect accounts exist, the mapping from the acoustics to these categories must also be invariant. The general auditory account, and cognitive theories more generally, begin with the lack of invariance assumption–the premise that the mapping from acoustics to phonetic categories is nondeterministic– and therefore, phonetic constancy must depend on inferential pattern recognition mechanisms (Nusbaum & Magnuson, 1997). But we would echo Stevens and Blumstein (1981) and note that the failure thus far to find invariance may indicate that we have not been looking in the right places. Establishing a deterministic gestural-acoustic-phonetic mapping would be a large step forward for direct realism, but would be of clear theory-neutral benefit as well. On the other hand, if it turns out that we are unable to pinpoint an invariant mapping, our investigations would still serve to constrain the inferential mechanisms that are posited by current cognitive accounts. This is an important consideration, given the absence of an inferential account that specifies mechanisms that lead to phonetic constancy given our current acoustic characterization of speech.

References

Boersma, P., & Weenink, D. (2006). *Praat: Doing phonetics by computer* (Version 4.4.16) [Computer program]. Retrieved from http://www.praat.org/

Diehl, R. L., Kluender, K. R., Walsh, M. A., & Parker, E. M. (1991). Auditory enhancement in speech perception and phonology. In R. Hoffman & D. Palermo (Eds.), *Cognition and the symbolic processes: Applied and ecological perspectives* (pp. 59–76), Hillsdale, NJ: Erlbaum.

Diehl, R. L., Lindblom, B., & Creeger, C. P. (2003). Increasing realism of auditory representations yields further insights into vowel phonetics. In *Proceedings of the 15th International Congress of Phonetic Science* (Vol. 2, pp. 1381–1384). Adelaide: Causal Publications.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55,* 149–179.

Elman, J. L., & McClelland, J. L. (1986). Exploiting the lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 360–380). Hillsdale, NJ: Erlbaum.

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

Fowler, C. A. (1994). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception & Psychophysics, 55,* 597–610.

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

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

Fowler, C. A., Brown, J., & Mann, V. (2000). Contrast effects do not underlie effects of preceding liquid consonants on stop identification in humans. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 877–888.

Fowler, C. A., & Dekle, D. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 816–828.

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

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

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

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

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the neural mechanisms of speech context effects. *Hearing Research, 167,* 156–169.

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese Quail can form phonetic categories. *Science, 237,* 1195–1197.

Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63,* 905–917.

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

Liberman, A. M., Delattre, P. C., & Cooper, F. S. (1952). The role of selected stimulus variables in the perception of the unvoiced-stop consonants. *American Journal of Psychology, 65,* 497–516.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs,* 68, (8, Whole No. 379).

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

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects of 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.

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for non-speech effects on phonetic identification. *Journal of the Acoustical Society of America, 113,* 53–56.

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

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r," *Cognition, 24,* 169–196.

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

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

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

Narayanan, S., Byrd, D., & Kaun, A. (1999). Geometry, kinematics, and acoustics of Tamil liquid consonants. *Journal of the Acoustical Society of America, 106,* 1993–2007.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego: Academic Press.

Pastore, R. E., Li, X. F., & Layer, J. K. (1990). Categorical perception of nonspeech chirps and bleats. *Perception and Psychophysics, 48,* 151–156.

Pisoni, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America, 61,* 1352–1361.

Reed, E. S. (1996). *Encountering the world.* New York: Oxford University Press.

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

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

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.

Stevens, K. N., & Blumstein, S. E. (1981). The search of for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin and Review, 16,* 74–79.

Wade, T., & Holt, L. L. (2005). Effects of later-occurring non-linguistic sounds on speech categorization. *Journal of the Acoustical Society of America, 118,* 1701–1710.

Received October 30, 2007
Revision received May 11, 2009
Accepted June 19, 2009 ■