Journal of Experimental Psychology:
Human Perception and Performance
2014, Vol. 40, No. 3, 1228–1236

0096-1523/14/$12.00 DOI: 10.1037/a0036214

Information for Coarticulation: Static Signal Properties or Formant Dynamics?

Navin Viswanathan
State University of New York, New Paltz
and Haskins Laboratories, New Haven, Connecticut

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

Perception of a speech segment changes depending on properties of surrounding segments in a phenomenon called *compensation for coarticulation* (Mann, 1980). The nature of information that drives these perceptual changes is a matter of debate. One account attributes perceptual shifts to low-level auditory system contrast effects based on static portions of the signal (e.g., third formant [F3] center or average frequency; Lotto & Kluender, 1998). An alternative account is that listeners' perceptual shifts result from listeners attuning to the acoustic effects of gestural overlap and that this information for coarticulation is necessarily dynamic (Fowler, 2006). In a pair of experiments, we used sinewave speech precursors to investigate the nature of information for compensation for coarticulation. In Experiment 1, as expected by both accounts, we found that sinewave speech precursors produce shifts in following segments. In Experiment 2, we investigated whether effects in Experiment 1 were driven by static F3 offsets of sinewave speech precursors, or by dynamic relationships among their formants. We temporally reversed F1 and F2 in sinewave precursors, preserving static F3 offset and average F1, F2 and F3 frequencies, but disrupting dynamic formant relationships. Despite having identical F3s, selectively reversed precursors produced effects that were significantly smaller and restricted to only a small portion of the continuum. We conclude that dynamic formant relations rather than static properties of the precursor provide information for compensation for coarticulation.

Keywords: compensation for coarticulation, sinewave speech, speech perception

A critical challenge for any account of speech perception is to explain perceptual stability despite a highly variable speech signal. For instance, the acoustic manifestation of a given phonetic segment can be different depending on the rate of produced speech (e.g., Miller & Baer, 1983), physical characteristics of the talker (e.g., Peterson & Barney, 1952), the dialect of the talker (e.g., Clopper & Pisoni, 2004), and the coarticulatory influences of surrounding segments (e.g., Mann, 1980). The research that we report here was designed to identify the nature of information that listeners use to achieve context-appropriate perception in different coarticulatory contexts.

Mann (1980) showed that when listeners categorize members of a [da]-[ga] (anterior-to-posterior place of articulation) continuum following liquid syllables [al] (anterior) and [aɹ] (posterior), they make more "g" responses after the syllable [al] than after [aɹ]. Mann suggested that these context-dependent responses reflect listeners' compensation for the acoustic effects of coarticulatory overlap between the syllable-final liquid and syllable-initial stop consonants. That is, in [alga], because of the forward pull of the tongue tip gesture of the preceding [l], the point of constriction during [g] is more forward than in a neutral context. In contrast, in [aɹda], the pharyngeal constriction of [ɹ] pulls the point of constriction during [da] farther back than in a neutral context. In both cases, coarticulation with the preceding liquid affects the point of constriction during production of the target segments and consequently their acoustic realization. Therefore, by this account, listeners "compensate for coarticulation," and their consonant category boundaries shift in accord with acoustic consequences of coarticulation. Although the gestural interpretation was originally proposed from a motor theory perspective (Liberman & Mattingly, 1985), it is also consistent with a direct realist theory of speech perception[1] (e.g., Fowler, 1986, 2006; Best, 1995; Viswanathan, Magnuson, & Fowler, 2010), which posits that the coarticulatory overlap is directly perceived.

An alternative explanation, also suggested by Mann (1980, p. 410–411) and put to test by Lotto and Kluender (1998), is that the

This article was published Online First April 14, 2014.

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

This research was supported by NSF Grant 0642300 to JSM, CAF and NV, NIH Grant R15DC00565 to NV, and P01HD-001994 to Haskins Laboratories.

Correspondence concerning this article should be addressed to Navin Viswanathan, Department of Psychology, State University of New York, 600 Hawk Drive, New Paltz, NY 12561-2440. E-mail: viswanan@newpaltz.edu

[1] Technically, the direct realist account is that the perceptual changes reflect listeners' attunement to coarticulation. Specifically, the acoustic change because of coarticulation provides information for the listener to detect rather than variability that the listener should "compensate for" as the motor theory assumes.

response changes in target consonant perception are because of spectral contrast between the precursor and the target segments. Specifically, [al] and [da] have high F3 offset and onset frequencies, respectively, compared with [aɹ] and [ga], both of which have lower F3s at offset and onset, respectively. This allows for the possibility that spectral contrast from a high-offset F3 in [al] causes listeners to hear the onset F3 in the target segment as lower in frequency, leading to more "ga" (low F3) responses. For the same reason, after hearing a precursor syllable with a low F3 offset, such as [aɹ], listeners report hearing more "da"s (high F3 responses). That is, just as a bucket of warm water feels hot after immersion in cold water, but cold after immersion in hot water, an ambiguous F3 is effectively high after a low F3, and effectively low after a high F3. On this account, the contingency between F3 and place of articulation is immaterial, as it is F3 that drives the result prelinguistically. This spectral contrast explanation is consistent with auditory accounts of speech perception and has been cited as strong support for the general auditory approach (Diehl, Lotto, & Holt, 2004). The general auditory approach is a framework that is seen as an alternative to gestural accounts of speech perception. Diehl et al. (2004) note that this approach is yet to be fully specified:

> G[eneral] A[uditory approach] is labeled an approach rather than a theory because, as summarized in preceding paragraphs, it is quite abstract, defining itself mainly by its opposition to key claims of M[otor] T[heory] and D[irect] R[ealist] T[heory]. At this level of abstraction, GA has too little content to be falsifiable. However, it does provide a general framework within which particular theoretical claims may be formulated and tested (p. 155).

However, despite these limitations, the GA spectral contrast explanation for compensation for coarticulation is well-specified and directly testable.

Although several studies have focused on the aforementioned liquid-stop context, findings of compensation for coarticulation have been obtained for other contexts such as fricative-stops (e.g., Mann & Repp, 1980), vowel-stops (Holt, 1999), stop-vowel-stops (Holt, Lotto, & Kluender, 2000), and fricative-vowel contexts (Mitterer, 2006). Although these extensions have demonstrated the ubiquity of these effects, none is capable of dissociating the two accounts, because they make identical predictions in all cases. This is because the critical acoustic signal properties that the general auditory theory predicts to be the locus of each contrast effect correlate with constriction locations that direct realism claims drive compensation.

However, Viswanathan et al. (2010) examined a liquid-stop context that did dissociate the two accounts. We used Tamil liquids [ar] and [aɹ] in addition to the English liquids [al] and [aɹ]. In the English segments, F3 correlates with place of articulation, whereas in Tamil it does not. Crucially, [ar] has a low F3 offset relative to the F3 onsets of the following continuum members, leading the spectral contrast account to predict *fewer* target "g" responses, but it has a frontal alveolar constriction, leading the gestural account to predict *more* target "g" responses. In support of the gestural account and against the predictions of spectral contrast, the Tamil [ar] patterned with the English [al] (with which it shares constriction location) producing more "g" responses than the English [aɹ] (with which it shares a low F3). That is, perception followed articulation rather than F3, despite the unfamiliarity of the Tamil segments for English speakers. Furthermore, we conducted follow-up experiments designed to extend Lotto and Kluender's (1998) findings that pure tone analogues matched to F3 offsets were sufficient to produce speech precursor-like effects, but we found that no combination of pure tones (single tones at F3 offsets, ditones at F2 and F3 offsets, or tritones at F2, F3 and F4) replicated the response pattern obtained with natural non-native speech precursors. This suggested that the spectral contrast account cannot be salvaged by appealing to contrast produced by other components of the precursor (Viswanathan et al., 2010). Compatibly with the findings of Viswanathan et al. (2010), Johnson (2011) dissociated the contrast and gestural accounts by looking at listeners' compensation to the bunched, relatively anterior variant of American English [ɹ] and the relatively posterior retroflexed variant of American English [ɹ] that share a low F3 offset. Similar to findings of Viswanathan et al. (2010), he found that the relatively anterior segment produced more "g" responses than the posterior segment. This pair of findings presents strong challenges for a spectral contrast account of compensation for coarticulation. For other challenges to the contrast account of compensation, please see Viswanathan, Fowler, and Magnuson (2009) and Viswanathan, Magnuson, and Fowler (2013).

Debates regarding the competing explanations of coarticulatory compensation have focused on whether the objects of perception are the acoustic signal itself (e.g., Diehl et al., 2004) or are the vocal tract gestures that produce the acoustic signal (e.g., Fowler, 1986, 2006). In this paper, we focus on another implication of each competing account of compensation for coarticulation regarding the nature of information that drives it. Although the general auditory and direct realist accounts agree that the information that listeners use to compensate for coarticulation is present in the acoustic signal, this information[2] is of a fundamentally different nature in the two accounts. The spectral contrast account is that the acoustic properties driving compensation are static properties (e.g., F3 offset, average F3 frequency; e.g., Lotto & Kluender, 1998; Holt, 2006) that are not restricted to speech, and indeed, changes observed in speech perception result from prelinguistic effects. For instance, the finding that nonspeech tones matched to the frequency offsets of the critical precursor speech syllables produce similar shifts to those produced by precursor syllables (Lotto & Kluender, 1998; but see Viswanathan et al., 2009) is used as support for this account. The explanation for nonspeech tone effects from the direct realist account is as follows. Although nonspeech tones can sometimes produce qualitatively similar effects to speech precursors, these effects have different origins. Specifically, nonspeech tones produce their effects by masking information in the target precursor. This assertion is supported by the findings of Viswanathan et al. (2013) that demonstrate that effects of tonal precursors occur because of masking of specific frequencies in the F3 region of target syllables by precursor tones. Furthermore, Viswanathan et al. (2009) showed that as tones are made more like the formants to which they are supposed to be analogs (by matching them to formant amplitude, trajectory, and bandwidth), "compensation" effects actually weaken, lending further support to the hypothesis of distinct origins of speech and

[2] Strictly speaking, from the general auditory perspective, these are properties of the acoustic signal that change the sensitivities of the auditory system.

nonspeech effects. For the purposes of the current discussion, the critical distinction from the spectral contrast account is that according to the direct realist account, information for compensation for coarticulation is the consequence of coarticulating gestures. Because this coarticulatory gestural overlap occurs over time, the information for compensation for coarticulation cannot come from static, isolated segments of the acoustic signal as suggested by the spectral contrast account. Instead, this information is necessarily dynamic and likely higher-order (e.g., the changing relationship between F2 and F3 formants over time).

In the present pair of experiments, we investigate the nature of information that underlies listeners' apparent compensation for coarticulation. Specifically, we ask whether static acoustic properties (such as formant offsets, average frequency of formants), as assumed by the spectral contrast account, are sufficient to produce compensation effects, or whether dynamic (unfolding over time) information about gestures, as assumed from the direct realist account, is required. This information, from this perspective, is about gestural overlap and is dynamic and higher-order rather than static and lower-order (e.g., F3 at offset or average F3). To do so, in Experiment 1, we use sinewave speech (Remez, Rubin, Pisoni, & Carrell, 1981) as precursors in the original liquid-stop contexts from Mann (1980). Sinewave speech is synthesized by replacing the formants in natural speech by pure sinewave tones that track the center frequencies of the formants and mimic their trajectories and intensities. Therefore, sinewave speech (rather than typically used static tones) retains the higher-order, dynamic information about gestural overlap critical for compensation for coarticulation. In Experiment 1, we ask whether the information preserved by the sinewave speech precursors is sufficient to produce shifts in the categorization of following speech. In addition, because sinewave speech may be heard by listeners as either speech or nonspeech (Remez et al., 1981), we investigate whether listeners' perception of the precursor as speech or nonspeech influences the categorization of the target speech continuum. In Experiment 2, we investigate whether static signal properties (such as F3 offsets of sinewave precursors used in Experiment 1) are sufficient, or whether dynamic spectrotemporal information is required to produce compensation for coarticulation.

Experiment 1

In this experiment, we investigate whether sinewave speech versions of natural precursors [al] and [aɹ] produce shifts in the perception of the target speech continuum. From the contrast perspective, given that the sinewave precursors have the same F3 mean and offset frequencies as in natural speech, spectral contrast between the sinewave precursors' offset frequencies and the target's onset F3s should produce shifts in target perception similar to those observed after natural speech precursors (although we can predict from Viswanathan et al. [2009] that the effects will be weaker than with typical static tone precursors that have had unrealistically high amplitudes).

From the direct realist perspective, listeners can recognize sinewave speech because time-varying sinusoids at formant-centers preserve sufficient dynamic information in the acoustic signal about the gestures of speech production (but do not preserve much other information, e.g., voice quality). On the assumption that the preserved information is available to the listener, we expect that it will influence the perception of the following target continuum members despite the difference from them in voice quality. In part, this assumption is supported by Lotto and Kluender's (1998; Experiment 1) finding, which is difficult to explain from the gestural perspective (but see the account offered by Lotto & Kluender, 1998). Specifically, they showed that listeners' perception of a target [da]-[ga] continuum members (presented in a male voice) is influenced by a preceding liquid (presented in a female voice) despite an obvious change in vocal source and therefore speaker identity. We defer further discussion of the limitations of the direct realist account until the discussion section. For now, we note that both the direct realist and the spectral contrast accounts expect that shifts in target perception, similar to those after natural precursors, will occur, but for different reasons.

Method

Participants. Twelve male and 10 female undergraduate students at the University of Connecticut participated for partial course credit. All reported being monolingual, native speakers of American English.

Materials. We used the same continuum as Viswanathan et al. (2009). The 11-step continuum of resynthesized CV syllables varied perceptually from [da] to [ga] and was created by varying the F3-onset frequency of the syllables. F3-onset frequencies varied in 100 Hz steps from 1800 Hz ([ga]) to 2800 Hz ([da]), changing linearly to a steady state value of 2500 Hz over an 80 ms transition. The first, second and fourth formants were the same for all members of the continuum. Over the 80 ms transition, F1 shifted from 500 Hz to 800 Hz, F2 shifted from 1600 Hz to 1200 Hz, and F4 was held steady at 3500 Hz. The overall duration of each CV syllable was 215 ms. The sinewave precursors were synthesized based on the first three formants of natural syllables. Their center frequencies were traced at 10 ms intervals, refining automatically generated LPC values, to prepare a synthesis table (see Remez et al. [2011] for a comparison of automated LPC prediction and hand tracing). Each sinewave precursor consisted of three sinewave tones designed to be analogues of the first three formants. The amplitudes of the F1, F2 and F3 analogues at each time interval were set respectively at 0.7, 0.4 and 0.2 times the overall amplitude of the original syllable for that interval, yielding relative amplitudes of formants and mean syllable amplitude matched to properties of the original speech tokens on which these were based. The critical F3 of the precursors started at a steady state value of 2400 Hz and transitioned up to 2800 Hz for [al] and transitioned down to 1800 Hz for [aɹ]. The overall intensities of the sinewave precursors were matched to those of the precursor syllables (which were themselves matched to those of the target syllables). The sinewave analogue of the [aɹ] precursor is shown in the left panel of Figure 1.The precursor and target were separated by 50 ms of silence and were presented diotically over headphones (Sennheiser HD-595) at approximately 70 dB SPL.

Procedure. Participants completed a pretest before the main test of the experiment. They heard six sinewave sentences, each repeated three times. Participants were not told that they would be hearing sinewave sentences. They were asked to listen to the first three stimuli to get used to the sinewave stimuli that would be presented in the main experiment. For the next three, they described what they heard. This pretest was used to determine how

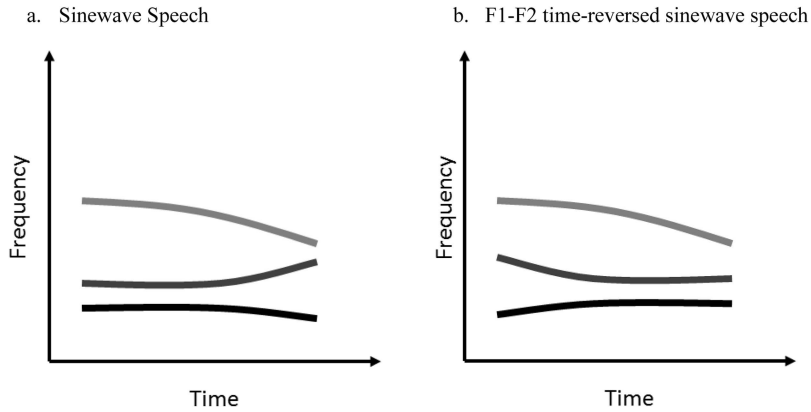

Figure 1. Schematic depictions of three formant sinewave speech precursor [aɹ] used in Experiment 1 on the left and its selectively reversed analogue on the right. Note that the first two formants in panels A and B are temporally reversed with respect to each other. The third formant is identical in both panels.

each participant categorized sinewave speech at the beginning of the experiment. At the end of the experiment, participants were interviewed by the experimenter to determine whether or not they had heard the precursors in the experimental block as speech. Participants who reported hearing the precursor syllables [al] and [aɹ] in the experimental block were classified as being in a *speech* mode. Participants with other responses (e.g., "space sounds," "birds chirping," "no idea") were classified as being in a *nonspeech* mode. This information was used to analyze whether listeners' perception of the target continuum depended on their categorization of the sinewave speech precursors. Exclusion or reclassification of these subjects based on their pretest reports (instead of their posttest reports) did not alter the pattern of results.

The identification test started with a practice block in which only the clear [da]-[ga] endpoint tokens were presented in isolation 10 times each in random order. Subjects received feedback on their categorizations. The purpose of this block was to familiarize subjects with the task and to ensure that they were able to classify the unambiguous endpoint tokens. In the experimental block, participants classified the target syllable in each sequence by button press to indicate whether they heard [da] or [ga] and were not given feedback. No explicit instructions were given to attend to or ignore the precursors. Instead, participants were told that during each trial they would hear two sounds and that they should classify the second sound as "ga" or "da." Each of the 11 continuum members followed the two precursors for nine cycles resulting in 198 trials per participant, presented in random order. Each session lasted approximately 25 minutes.

Results and Discussion

Data from all subjects were included in the analyses. The left panel of Figure 2 shows that sinewave precursors produced a sizable shift, similar to those produced by natural speech precursors and in the same direction. The average percentage "g" responses following the [al] analogue was 54.42 and following [aɹ] analogue was 46.99, indicating an average shift of 7.43%. Of the 22 listeners, 9 reported hearing the precursor as speech whereas 13 reported hearing various nonspeech sounds not associated with the human vocal tract in the postexperiment interview.

Proportions of "g" responses, shown in Figure 2, panel A, were first transformed to logit values before being submitted to 2 (precursor) × 2 (mode; quasi-experimental factor, between-subjects) × 11 (step; within) mixed ANOVA. The proportions of 0 and 1 were replaced by 0.1 and 0.99 resulting in bounded outcomes of (0, 1) to avoid singularities in the transformed data. The effect of precursor was significant ($F(1, 20) = 18.49$, $p < .0001$, $\eta_p^2 = 0.48$) indicating an increased likelihood of responding "g" following the sinewave [al] than [aɹ. The expected effect of step was significant ($F(10, 200) = 98.39$, $p < .0001$, $\eta_p^2 = 0.83$) whereas the effect of mode ($F < 1$) did not approach significance. No interaction was observed between precursor and mode ($F < 1$) indicating that whether the listeners heard the precursor as speech or as nonspeech did not alter the effect of the precursor on target categorization. There was an interaction between precursor and step ($F(10, 200) = 3.44$, $p < .0001$, $\eta_p^2 = 0.15$) because of the stronger effects of the precursor in the middle (ambiguous region) of the continuum than the end points (see Figure 2).

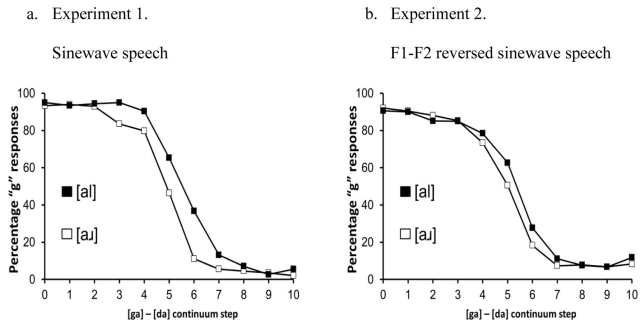

Figure 2. A comparison of the effects of sinewave precursors (Experiment 1 in panel A) and selectively reversed sinewave precursors (Experiment 2 in panel B). Whereas sinewave precursors produce strong (average mean difference across conditions = 7.43%, $p < .0001$), the precursor conditions are not reliably different in Experiment 2 (average mean difference across conditions = 2.57%, $p = .21$). The weak trends noticeable between continuum steps 4 through 7 may be explained because of the masking produced by the concentration of energy in F3 (see Figure 3).

From the direct realist gestural perspective, this finding is interpreted to indicate that the dynamic information about gestural overlap is present in the time-varying sinewave analogues of natural speech. The absence of interaction between mode and the boundary shifts implies that listeners still attune to gestural information even if they are unable to report it consciously. A current limitation of this account is that it does not explain why mode does not matter. For instance, it does not specify what structure in sinewave speech listeners must attune to in order to perceive it as speech and whether this differs from the information for coarticulation.

From the spectral contrast perspective, the result of Experiment 1 is expected because boundary shifts are determined by the static signal properties retained in the sinewave precursor; specifically, by the contrast between F3 offset of the precursors and the F3 onset in the target continua. The absence of interaction between mode and the boundary shifts suggests, from this perspective, that this is a purely auditory effect (also see Lotto, Kluender, & Holt, 1997). However, as outlined earlier, several findings (e.g., Viswanathan et al., 2010) have disconfirmed the spectral contrast explanation of boundary shifts along a [da]-[ga] continuum because of static precursor tones. Given these findings, and the crucial question of the nature of information in the acoustic signal for coarticulatory overlap between precursors and target syllables, in Experiment 2, we investigate whether static F3 offsets or dynamic, overtime relations among formants are responsible for the boundary shifts obtained in Experiment 1.

Experiment 2

In this experiment, we investigate whether the critical signal information responsible for compensation for coarticulation behavior (like that found in Experiment 1) is provided by the static acoustic signal properties claimed by the spectral contrast account (F3-offsets and/or mean F3; e.g., Lotto & Kluender, 1998), or if the critical signal information is instead provided by higher-order dynamic acoustic relationships (that provide information for gestural overlap). We investigate this question by modifying sinewave precursors to distort higher-order formant relationships (overtime dynamic relations among formants) while leaving F3 (and therefore F3-offsets and average formant frequencies) intact. Specifically, we temporally reverse the purportedly noncritical formants (F1 and F2 from the contrast perspective) of the precursor, while leaving its F3 intact (also see Viswanathan, Dorsi, & George, in press). The rationale behind this manipulation is that, if only contrast with the static F3-offsets matters, then reversing F1 and F2 analogues should not change the results compared with Experiment 1. That is, even with F1 and F2 reversed, the sinewave speech precursors should still produce boundary shifts identical to shifts because of sinewave precursors in Experiment 1. However, if the dynamic formant interrelationships (the information for constriction location in sinewave speech) are important, then the disruption of these relations should destroy information for coarticulation. Therefore, the gestural account predicts that compensation for coarticulation should not be observed when F1s and F2s are temporally reversed in the sinewave precursors despite their intact F3s.

Method

Participants. Thirteen male and 10 female undergraduate students at the University of Connecticut participated in the experiment for partial course credit. All reported being monolingual, native speakers of American English. None had participated in Experiment 1.

Materials. The target continuum from Experiment 1was used. The sinewave precursors from Experiment 1 were modified in the following manner. The analogues of the first and second formants from each of the two sinewave precursors in Experiment 1 were temporally reversed such that onset frequencies of the first two formants of the sinewave precursors of Experiment 1 became offset frequencies in the F1-F2-reversed sinewave precursors of the present experiment. The third formant analogue was left unmodified. The [aɹ] analogue precursor is shown in the right panel of Figure 1.

Procedure. The procedure was identical to that of Experiment 1.

Results and Discussion

Data from three subjects with accuracy less than 80% in the endpoint stop judgment task were excluded from the analysis.[3] The right panel of Figure 2 shows the results of Experiment 2. For comparison, the left panel shows the results of Experiment 1, with data collapsed over speech mode. None of the participants of Experiment 2, including those who reported hearing sinewave sentences in the preexperimental block, reported hearing the selectively reversed sinewave precursors as speech. The average percentage "g" responses following the [al] analogue was 50.60 and following the [aɹ] analogue was 48.03, indicating an average shift of 2.57%.

The data were submitted to a 2 (precursor) × 11 (step) within subject ANOVA. Importantly, there was no effect of precursor ($F(1, 19) = 2.61$, $p = .13$, $\eta_p^2 = 0.12$). The expected effect of step ($F(10, 190) = 57.37$, $p < .0001$, $\eta_p^2 = 0.75$) was significant, indicating that listeners' responses changed along the continuum. A marginal interaction between precursor and step was detected ($F(10, 190) = 1.89$, $p = .079$, $\eta_p^2 = 0.08$) indicating that the effect of the precursor was different at different steps of the continuum. A closer examination of the effects in Figure 2, panel B, reveals a separation of the curves in steps 4, 5 and 6. Restricting our analyses to these steps shows that the effect of the precursor in this region is indeed reliable ($F(1, 19) = 7.80$, $p = .012$, $\eta_p^2 = 0.29$).

Figure 3 presents a comparison of resulting shifts in responses produced by the precursors used in Experiments 1 and 2 (and, for comparison, in Viswanathan et al., 2009). On the ordinate, we plot percentage compensation as calculated by percentage "g" responses after [al] minus percentage "g" responses after [aɹ]. Despite having identical F3 sinewave analogues, intact sinewave precursors produce significantly stronger compensation compared with selectively reversed sinewave precursors ($F(1, 40) = 4.70$, $p = .036$, $\eta_p^2 = 0.11$). Furthermore, restricting this cross-

[3] This criterion is consistent with past studies (Viswanathan et al., 2009, 2010). In comparison, Lotto and Kluender (1998) used a stricter criterion of 90% accuracy in the endpoint categorization to ex. The rationale for using the endpoint classification task is twofold. First, it ensures that the listener is able to perform the two alternative force choice tasks accurately with clear unambiguous endpoints. Second, it affords a principled method to determine which subjects must be included in the final analyses.

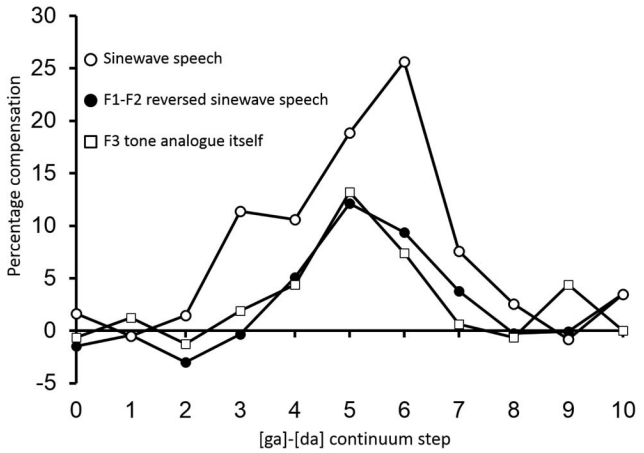

Figure 3. A comparison of the average compensation (expressed as the % "g" responses to the [al] analogue - % g responses to [aɹ] analogue, at each step, for each subject) produced by sinewave speech (open circles), selectively reversed sinewave speech (filled circles) in Experiment 2 and F3 sinewave in isolation (open squares) in Viswanathan et al. (2009, Experiment 2). Although three conditions all contain identical F3 analogues, sinewave speech produces robust effects that are stronger than the other two conditions, which themselves produce comparable effects. These data suggest that the strong effects of the sinewave speech precursors are because of their spectro-temporal acoustic structure rather than their F3 offsets.

experimental comparison only to those steps of the continuum that revealed a reliable precursor effect for Experiment 2 (steps 4, 5 and 6) did not alter the results ($F(1, 40) = 4.62$, $p = .038$, $\eta_p^2 = 0.103$). This finding presents a strong challenge for a spectral contrast account. Even though the energy relationships in the purportedly critical F3 region were held constant across Experiments 1 and 2, shifts comparable to those elicited by natural precursors were observed only when the dynamic information about F1 and F2 was left intact. Minimally, this result shows that the boundary shifts are not solely caused by the contrastive F3 region as assumed by the contrast account (also see Viswanathan et al., 2009, 2013).

Could the contrast account appeal to other contrastive relations between the precursors and targets? Viswanathan et al. (2010) ruled out, among many options, the contrast produced by a combination of F2 and F3 as a possible explanation for boundary shifts. Moreover, it is relevant that the overall mean frequencies of F1 and F2 are unaffected by the manipulation of temporal reversal. Holt (2006, 2005) showed that spectral contrast effects are sensitive to the overall mean frequency of the precursor and that this average is not temporally weighted. Because temporal reversal does not alter the overall spectral average, following this account, there should be no difference between the precursors of Experiments 1 and 2 even when the non-F3 regions are considered. Thus, the spectral contrast account cannot be salvaged by appealing to other portions of the acoustic signal. From the gestural perspective, these findings suggest that the information for vocal tract gestures is carried by the time varying change in the signal. When this dynamic information is distorted so that formant analogues do not have a trajectory reflecting coarticulated speech gestures, listeners should fail to show boundary shifts.

What then explains the weak precursor effects observed in Experiment 2? This finding may be explained by placing it in the context of findings of Viswanathan et al. (2009) and Viswanathan et al. (2013). Specifically, the F3 of the sinewave precursors of Experiments 1 and 2 consist of sinewave tones in which the entire energy in the analogous natural speech formant is concentrated in the formant center frequency. Viswanathan et al. (2009) studied the effects of tones that were identical to the F3s of the sinewave precursors in both of the present experiments. As shown in Figure 3, these tones are matched in intensity and trajectory but not bandwidth to F3 of natural speech precursors. These F3 tones produced weak effects, similar to those obtained in Experiment 2, presumably because of the tight concentration of energy around the formants' center frequency. Viswanathan et al. (2013) showed that this energy concentration of the precursor in the F3 region produces energetic masking effects (rather than spectral contrast) on the categorization of the target continuum. In other words, hearing these tones with an energy concentration that is unlike speech makes listeners temporarily insensitive to acoustic information in specific frequencies of the subsequent speech target.

In order to evaluate this explanation, we compared the shifts elicited by the F1-F2-reversed sinewave precursor of the present experiment with those elicited by the isolated F3s from Viswanathan et al. (2009). This is justified because the precursor F3s in that experiment and the present Experiment 2 were identical, and the same target continuum was used in both experiments. Figure 3 clearly shows that effects with the sinewave F3 alone and with the modified sinewave precursors of the present experiment are highly similar. These conditions are statistically indistinguishable ($F < 1$). In other words, this shows that when the dynamic information is destroyed, the unnatural concentration of energy in F3 center frequency may still produce weak boundary shifts because of masking (Viswanathan et al., 2013). Accordingly, by this argument, the effects obtained in Experiment 1 of the present study are also due in part to masking. However, when dynamic information from multiple formants is left intact in sinewave speech precursors, they produce stronger, more robust perceptual shifts that are comparable to those elicited by natural speech precursors as indicated by Figure 3.

General Discussion

Accounts of compensation for coarticulation differ on whether static properties of the precursor (such as formant offset or average frequency from the spectral contrast account) or higher-order signal properties such as interformant relationships (from the direct realist gestural account) underlie compensation for coarticulation.[4]

[4] We should note that these two perspectives do not exhaust possible explanations. For example, one author on our team (JSM) favors what might be called a "generic cognitive" view of compensation for coarticulation, in which the listener has learned through experience the acoustic contingencies that result from coarticulation (e.g., posterior shifts in segments with anterior place of articulation following segments with posterior constriction). Such an account has no trouble accommodating the finding that dynamic formant relations, rather than static formant details, drive compensation for coarticulation. However, unlike the direct realist account, it does not generate predictions about the nature of information in the signal that listeners' compensation behavior depends upon. From this generic cognitive perspective, the current results highlight the utility for any account of examining how the gestures of speech production shape dynamic aspects of the speech signal.

We attempted to identify some of the properties in the acoustic signal that produce compensation for coarticulation using a pair of experiments.

In Experiment 1, we used sinewave speech equivalents of typically used speech precursors [al] and [aɹ] and examined whether they produced shifts in the perception of a following resynthesized speech [da]-[ga] target continuum. We found that sinewave precursors produced robust shifts in the perception of members of the target speech continuum that did not depend on whether listeners heard the sinewave speech precursors as speech. From the spectral contrast account, the boundary shifts are attributed to the F3 offsets that are preserved in sinewave speech precursors. From the direct realist account, the target perception shifts indicate that the formant analogues of sinewave speech preserve sufficient dynamic information about gestural overlap to permit attunement to coarticulatory effects on target [da]-[ga] syllables.

In Experiment 2, we dissociated spectral contrast and gestural accounts of Experiment 1. The former holds that compensation behavior can be driven by static properties of the sinewave precursor such as F3 offset or mean formant frequency, whereas the latter holds that compensation results from higher-order dynamic acoustic patterning such as interformant relationships that specify coarticulated gestures. In order to dissociate the two explanations, we created sinewave precursors in which the F1 and F2 analogues of the sinewave precursors from Experiment 1 were temporally reversed whereas F3 analogues were untouched. This manipulation preserved the F3 offsets of the precursors as well as the overall average frequency of F1, F2 and F3 analogues. The results of Experiment 2 indicated that the F1-F2-reversed sinewave precursors, unlike the sinewave precursors of Experiment 1, did not produce similar shifts in the perception of following target segments. This shows that when the dynamic relationships among the components of the sinewave speech precursor are disrupted, target perception shifts are greatly diminished (and may result from masking), despite the fact that static properties of the signal the spectral account claims cause compensation behavior are preserved.

A Critical Assessment of the Two Competing Accounts

In this section, we examine the adequacy of the spectral contrast and direct realist accounts.

Spectral Contrast

The spectral contrast explanation of compensation for coarticulation is that energy relations between particular frequency regions in the precursor and the target are responsible for the resulting boundary shifts. This explanation is buoyed by the observation that pure tone precursors, seemingly bereft of any articulatory information, produce similar effects to natural speech as long as they are matched in frequency to the assumed critical regions (F3 offsets in the current liquid-stop context). Therefore, by this account, an appeal to articulatory information is unwarranted.

A series of recent findings call this explanation into question. First, the assumed critical region in natural speech does not, by itself (i.e., with all other regions of a natural speech stimulus removed), produce boundary shifts (Viswanathan et al., 2009). Second, in contexts in which place of articulation and F3 offsets are dissociated (Viswanathan et al., 2010; Johnson, 2011), listeners' boundary shifts occur in the direction predicted by place of articulation and *opposite* to the direction predicted by spectral contrast. Third, boundary shifts occur in other contexts despite the absence of spectral contrast between the coarticulating segments (visual coarticulatory contexts, e.g., Mitterer, 2006; simultaneous coarticulatory contexts, e.g., Silverman, 1986). Finally, there is reason to question whether the tone-speech effects are in fact because of contrast. Viswanathan et al. (2009) found that as properties of nonspeech F3-analog tone precursors are progressively matched to the speech regions that they are designed to be analogues of (in terms of amplitude, bandwidth, and frequency transition over time), their effects weaken rather than mimic the effects of natural speech precursors. In a follow-up investigation, Viswanathan et al. (2013) examined whether nonspeech tone effects result from energetic masking rather than spectral contrast. We found that nonspeech tones farther from the critical F3 region produced weaker effects despite a greater contrast in frequencies between the precursor tones and the target continuum. Furthermore, we filtered the target continuum in either the high- or low-frequency regions to mimic the assumed effects of energetic masking produced by high- or low-frequency precursor tones. The perception of these filtered continua presented without precursors patterned similarly to those after nonspeech tones; listeners reported more "g" responses to the continuum with the high-frequency region filtered than the one with the low-frequency region filtered. These findings suggest that energetic masking is a plausible alternative explanation of nonspeech tonal effects to spectral contrast (e.g., Fowler, Brown, & Mann, 2000).

In sum, because perceptual boundary shifts occur in the direction of spectral contrast (e.g., Lotto & Kluender, 1998), against the direction of spectral contrast (e.g., Viswanathan et al., 2010), in the absence of spectral contrast (e.g., Mitterer, 2006) and sometimes do not occur despite the presence of spectral contrast (e.g., current Experiment 2, Viswanathan et al., 2009), we consider the spectral contrast explanation for compensation for coarticulation sufficiently falsified. Other acoustic explanations of compensation for coarticulation may be possible (e.g., Mitterer, 2006). For these accounts, based on the current experiments and the other results we have reviewed, we suggest that the acoustic information driving compensation for coarticulation must involve time-varying combinations of covarying formants.

The Direct Realist Account

The direct realist account of compensation for coarticulation is that listeners attune to gestural overlap in speakers' coarticulated productions through informative structure in informational media (acoustic, optical, haptic). By this account, the information in the acoustic signal is about the causal source of this structure, that is, the vocal tract gestures. Thus far, we have highlighted the range of findings that support the gestural account. In this section we focus on a key limitation of this account as currently specified. In particular we note that the direct realist account must specify the acoustic information for gestural overlap that listeners use to compensate for coarticulation. As outlined earlier, by this account,

compensation for coarticulation occurs because listeners use information in the acoustic signal to perceive coarticulated gestures. However, exactly what aspects of acoustic signal carry this information needs specification.[5] In the experiments reported in this paper, we took a small step in identifying the nature of this information (that it is not static or lower-order). In the theory, it is the acoustic consequences over time of overlapping speech gestures. For a direct realist account to be complete, identification of this information and demonstrating its specificity to speech gestures is mandatory.

Finally, in addition to the findings discussed thus far, there exists a set of compensation for coarticulation findings that is not adequately addressed by either account. Specifically, these studies investigate the question of whether listeners' prior phonological knowledge influences compensation for coarticulation and have yielded two apparently incompatible sets of results. One set that has been addressed by both accounts appears to indicate that compensation for coarticulation can be independent of phonological learning (e.g., Fowler et al., 2000; Lotto & Kluender, 1998). However, there exists a second set of findings that appear to indicate a strong role for learning in compensation for coarticulation (e.g., lexical compensation for coarticulation, Elman & McClelland, 1988). In addition, in an investigation of a related phenomenon of compensation for phonological assimilation, Darcy, Peperkamp, and Dupoux (2007) demonstrate a strong role for language-specific phonological attunement in addition to language-independent compensation. Such findings have not been adequately addressed by either account. From the spectral contrast perspective, these findings are problematic because these effects occur despite the lack of contrastive properties in the signal (i.e., the same signal is treated differently by listeners, e.g., Darcy et al., 2007). From a direct realist viewpoint, one could interpret these findings as indicating a role for language-specific attunement. However, the challenge for this account is to specify under what conditions compensation involves language-independent perception of vocal gestures and distinguish these situations from those in which language-specific attunement plays a strong role.

Conclusion

Broadly, our results suggest that listeners use information in time-varying acoustic signals, including interformant relationships, to attune to coarticulatory variability. Although this does not automatically imply that listeners perceive gestural overlap, it is clear that the spectral contrast account (which appeals to low-level auditory effects of simple, static signal properties such as formant offset or mean frequencies) is untenable in light of the current result, as well as several others we have already reviewed (Johnson, 2011; Viswanathan et al., 2009, 2010, 2013). In general, any account must acknowledge that higher-order, time-varying relations in the acoustic signal are crucial for compensation for coarticulation.

This higher-order patterning in the acoustic signal provides information about vocal tract gestures, and, from a direct realist account, this is how listeners make use of it. Specifically, the direct realist gestural account is that during coarticulation, dynamic gestures of speech production overlap in time, causally structuring the resulting acoustic signal. In this study we demonstrate that, compatibly, perceptual information for coarticulatory overlap is dynamic and higher-order.

[5] It is clear that gestural overlap during the production of these disyllables produces these specific patterns of formant changes (Mann, 1980). The critical question is the one of inversion: Does the acoustic signal specify the gestural overlap and, therefore, the point of constriction? (Many researchers argue that it is not [e.g., Atal & Hanauer, 1971; Diehl et al., 2004]). However, for some positive evidence, see Iskarous, 2010; Iskarous, Fowler, & Whalen, 2010).

References

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America, 50,* 637–655. doi:10.1121/1.1912679

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience*. Baltimore, MD: York Press.

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics, 32,* 111–140.

Darcy, I., Peperkamp, S., & Dupoux, E. (2007). Bilinguals play by the rules: Perceptual compensation for assimilation in late L2-learners. In J. Cole & J. Hualde (Eds.), *Laboratory phonology 9* (pp. 411–442). Berlin: Mouton de Gruyter.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology, 55,* 149–179. doi:10.1146/annurev.psych.55.090902.142028

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143–165. doi:10.1016/0749-596X(88)90071-X

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. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics, 68,* 161–177. doi:10.3758/BF03193666

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. doi:10.1037/0096-1523.26.3.877

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

Holt, L. L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science, 16,* 305–312. doi:10.1111/j.0956-7976.2005.01532.x

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America, 120,* 2801–2817. doi:10.1121/1.2354071

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. doi:10.1121/1.429604

Iskarous, K. (2010). Vowel constrictions are recoverable from formants. *Journal of Phonetics, 38,* 375–387. doi:10.1016/j.wocn.2010.03.002

Iskarous, K., Fowler, C. A., & Whalen, D. H. (2010). Locus equations are an acoustic expression of articulator synergy. *Journal of Acoustical Society of America, 128,* 2021–2032. doi:10.1121/1.3479538

Johnson, K. (2011). Retroflex versus bunched in compensation for coarticulation. *UC Berkeley Phonology Lab Annual Report, 2011,* 114–127.

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

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception & Psychophysics, 68,* 178–183. doi:10.3758/BF03193667

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. doi:10.3758/BF03206049

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. doi:10.1121/1.419865

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28,* 407–412. doi:10.3758/BF03204884

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on the perception of [ʃ]-s] distinction: I. Temporal factors. *Perception & Psychophysics, 28,* 213–228. doi:10.3758/BF03204377

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America, 73,* 1751–1755. doi:10.1121/1.389399

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics, 68,* 1227–1240. doi:10.3758/BF03193723

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175–184. doi:10.1121/1.1906875

Remez, R. E., Dubowski, K. R., Davids, M. L., Thomas, E. F., Paddu, N. U., Grossman, Y. S., & Moskalenko, M. (2011). Estimating speech spectra by algorithm and by hand for synthesis from natural models. *Journal of the Acoustical Society of America, 130,* 2173–2178. doi:10.1121/1.3631667

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947–950. doi:10.1126/science.7233191

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

Viswanathan, N., Dorsi, J., & George, S. (in press). The role of speech-specific properties of the background in the irrelevant sound effect. *Quarterly Journal of Experimental Psychology*.

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. doi:10.3758/PBR.16.1.74

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception and Performance, 36,* 1005–1015. doi:10.1037/a0018391

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2013). Similar response patterns do not imply identical origins: an energetic masking account of nonspeech effects in compensation for coarticulation. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 1181–1192. doi:10.1037/a0030735

Received April 15, 2013
Revision received January 28, 2014
Accepted February 4, 2014 ■