Journal of Experimental Psychology:
Human Perception and Performance
2013, Vol. 39, No. 4, 1181–1192

0096-1523/13/$12.00 DOI: 10.1037/a0030735

Similar Response Patterns Do Not Imply Identical Origins: An Energetic Masking Account of Nonspeech Effects in Compensation for Coarticulation

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

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

Nonspeech materials are widely used to identify basic mechanisms underlying speech perception. For instance, they have been used to examine the origin of *compensation for coarticulation*, the observation that listeners' categorization of phonetic segments depends on neighboring segments (Mann, 1980). Specifically, nonspeech precursors matched to critical formant frequencies of speech precursors have been shown to produce similar categorization shifts as speech contexts. This observation has been interpreted to mean that spectrally contrastive frequency relations between neighboring segments underlie the categorization shifts observed after speech, as well as nonspeech precursors (Lotto & Kluender, 1998). From the gestural perspective, however, categorization shifts in speech contexts occur because of listeners' sensitivity to acoustic information for coarticulatory gestural overlap in production; in nonspeech contexts, this occurs because of energetic masking of acoustic information for gestures. In 2 experiments, we distinguish the energetic masking and spectral contrast accounts. In Experiment 1, we investigated the effects of varying precursor tone frequency on speech categorization. Consistent only with the masking account, tonal effects were greater for frequencies close enough to those in the target syllables for masking to occur. In Experiment 2, we filtered the target stimuli to simulate effects of masking and obtained behavioral outcomes that closely resemble those with nonspeech tones. We conclude that masking provides the more plausible account of nonspeech context effects. More generally, we suggest that similar results from the use of speech and nonspeech materials do not automatically imply identical origins and that the use of nonspeech in speech studies entails careful examination of the nature of information in the nonspeech materials.

Keywords: spectral contrast, energetic masking, compensation for coarticulation, nonspeech context effects, speech perception

A common and productive strategy in the study of human perception is to attempt to pinpoint the causal mechanisms of perceptual phenomena by varying different aspects of the perceiver and environment. For example, demonstrations that nonhuman species such as quail or chinchillas exhibit categorical perception of consonants (Kluender, Diehl, & Killeen, 1987; Kuhl & Miller, 1975) falsify claims that categorical perception is unique to humans, and demonstrations that musicians exhibit categorical perception of plucked versus bowed strings falsify claims that categorical perception is unique to speech (Cutting & Rosner, 1974). Such demonstrations tempt one to surmise that similar mechanisms underlie each case. Indeed, it could be argued that the assumption of similar causes of perceptual response patterns is most parsimonious. However, such demonstrations can serve only as starting points for further investigation; they do not establish the nature of the mechanisms nor their underlying similarity. They establish only a surface similarity in performance. Establishing similar (let alone identical) causal mechanisms requires much deeper investigation.

The allure of similarity is especially strong in cases where an idealized stimulus, which retains only hypothesized critical properties, replaces natural speech and produces qualitatively similar responses. In this study, we address such a situation in investigations of the phenomenon of *compensation for coarticulation*. Compensation for coarticulation is the finding that listeners' perception of a given segment can change depending on the properties of the preceding segment. For example, when listeners classify members of a [da]-[ga] continuum, they report more "g" responses after [al] than after [aɹ] (Mann, 1980). Mann noted that the response difference might reflect compensation for effects of coarticulatory overlap between the syllable-final and syllable-initial consonants. Dur-

This article was published Online First November 12, 2012.

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

This research was supported by NSF Grant 0642300 to JSM, CAF, and NV, NIH Grant DC00565 to NV, and NICHD Grant HD-001994 to Haskins Laboratories.

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

ing the production of [ga] following [al], speakers may not reach the canonical constriction location for [ga] because of coarticulatory overlap of [g]'s velar constriction gesture of the tongue body with the more front tongue tip gesture of [l]. This leads to a point of constriction during [g] that is shifted in the direction of that of [l]. The opposite may happen when the alveolar constriction gesture for [da] overlaps with the pharyngeal constriction gesture of [ɹ]. In both cases, because of coarticulation with the preceding syllable, the point of constriction during the consonant and the resulting acoustic signal are affected. In this context, listeners' categorization shift, described earlier, appears to take into account coarticulatory overlap in production. Hence, Mann dubbed this phenomenon "compensation for coarticulation." From a theoretical account in which listeners perceive speech gestures (Best, 1995; Fowler, 1986; Liberman & Mattingly, 1985), Mann's findings provide evidence that listeners are sensitive to the effects of gestural overlap on the target's acoustic realization, leading to context-appropriate target identification.[1] In other words, compensation for coarticulation is a perceptual effect demonstrated by the listener to enable detection of talkers' phonetic intents despite coarticulatory overlap.

In the original report of compensation for coarticulation, Mann also offered a radically different explanation. She observed that the specific acoustic correlate of the preceding phoneme that covaried with the context effect was the third formant (F3) transition offset. This suggests the possibility of a *spectral contrast* account. Specifically, [al] has a high F3 offset relative to the F3 onset of the target [da]-[ga] continuum steps, while [aɹ] has a relatively low F3 offset. This frequency difference between the precursor and continuum members produces spectral contrast such that, when listeners hear [al], they are likely to hear the following segment's onset F3 as lower (and more [ga]-like). The converse occurs after [aɹ]. Contrast effects are pervasive in perception (e.g., Warren, 1985). A salient example is the subjective experience that lukewarm water feels hot after dipping one's hand in cold water, but cold after dipping one's hand in hot water. She speculated that behavior in a compensation for coarticulation experiment could have a low-level auditory cause because of sensory contrast. In the case of compensation for coarticulation, the analogous contrast would be that, after a high-frequency F3, an ambiguous F3 frequency (middle range) would sound low, while after a low-frequency F3, it would sound high.

Lotto and Kluender (1998) put this hypothesis to test. They reasoned that if sensory contrast were the cause of compensation behavior, one should be able to replicate compensation for coarticulation effects by replacing the natural context syllables (e.g., [al] and [ar]) with sinusoidal tones at the natural syllables' F3 center frequencies. This is precisely what they found. They reasoned that, because a pure tone at the crucial F3 frequency has qualitatively similar effects as a speech stimulus with F3 at that frequency, it is the energy at the F3 frequency that is driving perceptual performance. Furthermore, proponents of the spectral contrast account have demonstrated that such qualitative similarity in the effects produced by pure tones versus natural speech generalizes across other coarticulatory contexts. For a comprehensive list of context effects that can be described as contrastive, see (Lotto & Holt, 2006, Table 1).[2] In addition, because gestural accounts of speech perception (e.g., Fowler, 1986; Liberman & Mattingly, 1985) do not predict any effect of nonspeech tones on speech perception, these findings have been taken as strong support for the spectral contrast account of compensation for coarticulation in particular, and as argument against gestural accounts of speech perception in general (e.g., Diehl, Lotto, & Holt, 2004).

In this article, we evaluate the spectral contrast account by asking whether similar responses implicate identical origins. Let us examine the rationale of this account more closely.

(a) The hypothesis is that sensory contrast between precursor F3 and target F3 causes compensation for coarticulation.

(b) Therefore, even a nonspeech, pure tone precursor with energy at the critical frequency should result in compensation for coarticulation.

(c) Because (b) is observed empirically, the most parsimonious explanation is that sensory contrast is the cause of compensation for coarticulation both for nonspeech and natural precursors.

In the context of the earlier discussion, this assumption of identical cause stems from the observation of similar responses in speech and nonspeech contexts. Therefore, by the logic presented previously, while this assumption is justified by the principle of parsimony, it still must be carefully investigated before being accepted as an explanation for compensation for coarticulation.

Recently, Viswanathan, Fowler, and Magnuson (2009) questioned whether compensation in speech contexts could be attributed to spectral contrast of F3s. They directly tested the spectral contrast explanation of compensation for coarticulation by presenting listeners either with the standard [al]-[aɹ] speech precursors or only their third formant region (by using a bandpass filter that only preserved that region). The rationale behind the manipulation was that, if spectral contrast effects caused by F3 are indeed responsible for compensation for coarticulation, then the two sets of precursors (complete syllables or the filtered F3 regions) should produce similar effects on the perception of the target continuum. However, although the intact precursors produced the typically observed compensation for coarticulation effects, the F3 regions by themselves did not. This suggests that compensation for coarticulation effects cannot be because of spectral contrast produced by the F3s of precursor disyllables as suggested by Lotto and Kluender (1998). Viswanathan et al. also found that, because nonspeech tones are progressively more closely matched to the characteristics of F3s in real speech (i.e., by matching the tones to F3's amplitude rather than to the whole-syllable mean amplitude, by having them track F3 changes over time rather than holding steady at the F3 offset frequencies, and by giving them bandwidth), their effects on identification of a following [da]-[ga] continuum weaken rather than strengthen. Furthermore, Viswanathan, Magnuson and Fowler (2010) demonstrated that when F3 and place of articulation of precursors are dissociated, compensation follows place. They did so by using a non-native liquid (Tamil alveolar trill) [ar] that had a low F3 despite a front place of articulation. Consistent with the gestural hypothesis, they found that Tamil [ar] produced *more* "g" responses than consonants with more back places of articulation (Tamil [al] and English [aɹ]).

[1] For a review of the crucial differences between the direct realist and the motor theories of speech perception, see Fowler (1996).

[2] Note that while the spectral contrast account would obviate the need for any form of compensation, for simplicity, we will continue to label the identification shifts observed in these paradigms as "compensation for coarticulation effects."

Critically, for the questions at hand, this result demonstrates that speech context effects can occur in a direction that is opposite to the direction expected based on spectral contrast.

In other words, to summarize, similarity of response pattern *is consistent with* but is not sufficient to demonstrate similarity of cause. These findings together show that the parsimonious cause of responding is not the true cause or causes because (a) the information present in typical nonspeech tones is not matched to the information actually present in the natural speech stimulus it is meant to idealize, (b) it cannot be demonstrated that similar compensation behavior results when the assumed critical F3 energy is presented in isolation, and (b) speech context effects can occur in a direction *opposite* to the direction predicted by spectral contrast.

However, the findings reviewed above leave unexplained how nonspeech pure tones matched to F3 offsets of [al] and [aɹ] (e.g., Lotto & Kluender, 1998, Experiment 3), produce shifts in the perception of following speech. In this study, we evaluate whether changes in target speech identification after nonspeech tone precursors have a different origin than changes after speech precursors despite the similarity of responses. In our discussion, we highlight a more general implication for the study of perception. It is that the temptation to interpret similar response patterns to the natural and the manipulated signal as indicating common causality must be resisted until it has been established that information critical to the perceiver (rather than the experimenter) is preserved.

Fowler, Brown, and Mann (2000) suggested an alternate explanation of nonspeech context effects that nonspeech effects are a result of *forward energetic masking* produced by the tone precursors on the speech precursors. In general, energetic masking is an increase in the threshold for the detection of a target stimulus because of the presence of competing acoustic energy in the same frequency regions (Moore, 1995). Specifically, the masking explanation of nonspeech effects on speech is that tones affect perception of following speech by interfering with detection of gestural information in the target speech syllables in the frequency region near the tone frequencies. Fowler et al. (2000) offered support for this explanation by demonstrating that when the F3s of the target syllables were made unnaturally high in intensity (so as to resist energetic masking), preceding nonspeech tones produced no response shifts, supporting a masking explanation of these nonspeech effects.

However, other researchers have disputed this claim. For instance, on the (mistaken) assumption that masking effects are restricted to the auditory periphery (see Moore, 1995, for evidence against such a restriction), Holt and Lotto (2002) investigated whether solely peripheral mechanisms are involved in compensation for coarticulation. They found that typical speech contexts produced boundary shifts even when the contexts and the target syllables were presented to opposite ears, implying that nonspeech effects do not have solely peripheral origins (e.g., at the level of the cochlea or the auditory nerve). Lotto, Sullivan, and Holt (2003) later extended this finding to nonspeech contexts and concluded that, because nonspeech effects, like speech context effects, persist when the context and the target are presented dichotically, their origins must also involve central auditory mechanisms. Furthermore, Holt (2005, 2006; also see Holt & Lotto, 2002) showed that nonspeech contexts produce effects that persist over several hundred milliseconds. This is inconsistent with typical masking effects, which largely diminish within about 50 ms (Elliott, 1971). Taken together, these findings have been used by proponents of the contrast account to dismiss the masking account suggested by Fowler et al. (2000).

However, for at least two reasons, the dismissal is premature. First, even though peripheral processes are often implicated in masking, central masking effects are also well documented (e.g., see Zwislocki, Buining, & Glanz, 1968). Second, and more importantly, there is suggestive empirical support for a masking account of nonspeech contexts even in the original findings of Lotto and Kluender (1998). In their Experiment 2, Lotto and Kluender synthesized sinewave glides that were matched to the third formant intensities of syllable precursors and tracked the transition frequencies of the precursor syllables' F3s. They obtained boundary shifts qualitatively similar to those because of syllable precursors but numerically smaller than the speech effects. However, the investigators did not test statistically whether the differences in boundary shift between the tone and speech conditions were reliable. In an effort to ensure that these tones did not capture any speech-specific property, in their Experiment 3, Lotto and Kluender generated steady state tones that were matched to the higher intensity of the total syllable rather than to that of F3.These tones produced robust boundary shifts that were larger than in their Experiment 2. Again, the authors did not report quantitative tests of the effects of increasing the intensity and changing the frequency contour of the precursor tones. If the untested conditions differ reliably, this would suggest that, in addition to frequency offset, the frequency trajectory and intensity of the precursor with respect to the target's critical frequency region may determine the size of the categorization shifts.

Findings of Viswanathan et al. (2009) described earlier provide support for this possibility. We found that, as nonspeech tones are progressively more closely matched to the characteristics of F3s in real speech, including in intensity, their effects on identification of a following [da]-[ga] continuum weaken rather than strengthen. This finding quantitatively confirms the trends toward weaker effects of transient, formant-intensity (rather than syllable-intensity) matched tones observed in Lotto and Kluender (1998) and agree with the finding of Fowler et al. (2000) that nonspeech tones have no effect on speech targets that have unnaturally high-intensity F3s.

The empirical facts presented so far suggest that compensation for coarticulation found in experiments using speech and nonspeech precursors derive from different causes. From the perspective of the gestural theory, a gestural account of compensation applies in speech contexts, whereas masking may underlie the effects of nonspeech contexts. Although it is less parsimonious to invoke two different explanations for the qualitatively similar speech and nonspeech context effects instead of the unitary explanation offered by spectral contrast, the data appear to require distinct accounts. In the present study, we directly test competing predictions from spectral contrast and masking accounts of nonspeech effects on speech. In Experiment 1, we ask whether nonspeech effects are always contrastive in direction. If not, then the utility of spectral contrast to explain nonspeech context effects, irrespective of its (as-yet unspecified) underlying mechanisms, is unclear. In Experiment 2, we focus on typical nonspeech effects that have been ascribed to spectral contrast and investigate whether these effects may, in fact, be attributable instead to masking.

Experiment 1

In Experiment 1, we test whether effects of nonspeech contexts are always contrastive. In other words, when effects on "da"-"ga" identifications produced by two tones are compared, does the higher frequency tone always produce more "g" (low-F3) responses? To conduct this test, we manipulate the frequency relations between the nonspeech tones and the target speech continuum members and examine the effects on resulting shifts in target perception.

Holt (1999, Experiment 7) made a similar comparison in a precursor-stop-vowel context. She synthesized a series of single formant precursor stimuli that had a center frequency ranging from 500 to 3,200 Hz in 300 Hz steps. Although these stimuli retained the harmonic structure of speech, they were not heard as speech by listeners. Her participants classified each member of a following [ba]-[da] continuum with F2 onset frequencies ranging from 1,000 Hz to 1,500 Hz presented after one of the precursors. From a contrast account, higher frequency precursors should produce more "b" responses, because listeners will perceive the F2 onset as lower and more like [ba]. This is exactly what she found. Importantly, the size of the effect (by her account, a contrast effect) steadily increased for precursor center frequencies ranging from 800 Hz to 2,300 Hz after which it asymptoted, demonstrating a broad range of frequencies across which the influence of the precursor could be detected. She concluded that nonspeech effects could not be because of low level processes, such as cochlear masking or auditory nerve adaptation, because, in that case, the frequency range across which contrastive effects would have been observed would have been much smaller.

However, findings by Viswanathan et al. (2010) cast doubt on the relevance of Holt's vowel-stop-vowel findings to the liquid-stop contexts that have been used in many studies of compensation for coarticulation. Viswanathan et al. found that listeners' categorization of members of a [da]-[ga] continuum differing in F3 was unaffected by precursor tone energy in the F4 region of speech precursors in nonspeech conditions in which F3 and F4 were presented simultaneously. Instead, listeners were strongly influenced only by precursor tone energy in the critical F3 region of speech precursors. If the findings of Holt (1999) are relevant to a [da]-[ga] context, F4 tone analogues and, by extension, F4 in natural speech, should have had a strong influence on the speech categorization task and should have uniformly produced more "g" responses owing to their relatively high frequency compared with the stops' F3s.

Our masking account of nonspeech effects is that precursor tones interfere with the pick-up of phonetic information in the target speech signal through energetic masking. If this is correct, consistent with typical masking effects, a nonspeech tone will exert a stronger effect when it is closer in its frequency composition to the region of the speech target continuum that carries information critical for categorization (see Moore, 1995, for a review of the literature on masking). For our experiment, we use a [da]-[ga] continuum varying only in its F3, making the F3 region critical for categorization of continuum members. Therefore, it follows from a masking account, that tones farther away from this F3 region should produce weaker shifts than those that are closer. The typical nonspeech context effects on categorization of the following [da]-[ga] targets is that tones at frequencies corresponding to high F3 offset frequencies produce fewer "g" responses (implicitly more "low F3" responses by the spectral contrast account) than those at lower F3 offset frequencies.

In Experiment 1, we use a variety of nonspeech tone precursors, including a standard low F3 ([aɹ]-analogue 1,800 Hz), and a standard high F3 ([al]-analogue 2,600 Hz). The standard tones will provide a basis for replicating previous nonspeech effects, thus they have a flat frequency trajectory (one frequency value) and are matched to syllable intensity rather than formant intensity (as in the materials used in Viswanathan et al., 2009). By the spectral contrast account, the high F3 tone precursors should produce more "g" (low target F3) responses than the low F3 precursor because of the energy contrast differences between the tone precursor and the target (Lotto & Kluender, 1998). The same pattern is also expected from the masking account but for a different reason. The low F3 precursor interferes with the detection of low F3 energy in the target and reduces the number of "g" responses. Similarly, the high F3 precursor interferes with the detection of high F3 energy in the target and therefore reduces the number of "d" responses (or increases "g" responses). Thus, the high F3 tone precursors produce more "g" responses than the low F3 tone.

To distinguish these accounts, we also include higher frequency tones at typical low (3,000 Hz) and high F4 (3,400 Hz) offset frequencies. If masking underlies nonspeech context effects, then F4 tones that are more distant from the F3 onset of the [d] endpoint should produce weaker effects on its detection than the standard high F3 tone. Accordingly, when the three higher frequency tones are considered (high F3, low F4, and high F4) from the masking account, we expect the pattern of "g" responses, relative to the lowest frequency tone (low F3), to align inversely to their respective distances from the F3 onset of [d] (high F3[2,600 Hz] $>$ low F4[3,000 Hz] $>$ high F4 [3,400 Hz] $>$ low F3[1,800 Hz]). The low F3 tone produces fewest "g" responses because of the direct interference of this tone in the pickup of information for [g]'s F3 onset.

The contrast account makes very different predictions. Because the low and high F4 tones are both higher than the high F3 tone, the contrast account predicts that they should induce larger shifts toward more "g" responses than the high F3 tone, and that the shift should increase or plateau with distance from the target F3. For instance, in an investigation of durational contrast, Diehl, Elman, and McCusker (1978), show that "stronger" contexts (in their case, voiceless contexts with Voice Onset Time [VOT] of $+100$ ms or voiced contexts of -100 ms VOT) produce greater contrastive effects on judgments of stimuli with intermediate VOTs compared with "weaker" contexts (voiceless VOT of 40 ms or voiced VOT of $+10$). That is, the farther away the context VOT, the greater was its effect on the test stimuli (so in terms of "g" responses: low F3 [1,800 Hz] $<$ high F3 [2,600 Hz] $<$ low F4 [3,000 Hz] $<$ high F4 [3,400 Hz]). Another possibility is that, similar to findings of Holt (1999) the size of the contrast effect will increase up to a critical point and increase no further. This would suggest the possibility that, relative to the low F3 tone, both higher frequency F4 tones would produce increases in "g" responses comparable to the high F3, without an increase in "g" responses from the lower F4 tone to the higher F4 tone (e.g., high F3 [2,600 Hz] $\cong$ low F4 [3,000 Hz] $\cong$ high F4 [3,400 Hz] $>$ low F3 [1,800 Hz]). Critically, the spectral contrast account does not predict that a higher frequency tone would produce *fewer* "g" responses than a lower frequency

tone as suggested by the masking account. The predictions of both accounts are schematically depicted in Figure 1.

Method

Participants. Eleven male and 17 female University of Connecticut undergraduates, 18–23 years old, participated for partial course credit. All reported normal hearing.

Materials. We created an 11-step continuum of resynthesized CV syllables varying in F3-onset frequency and varying perceptually from [da] to [ga] using the source-filter method of the *Praat* software package (Boersma & Weenik, 2006). F3-onset frequencies varied in 100 Hz steps from 1,800 Hz ([ga]) to 2,800 Hz ([da]), changing linearly to a steady state value of 2,500 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 1,600 Hz to 1,200 Hz, and F4 was held steady at 3,500 Hz. The overall duration of each CV syllable was 215 ms. This continuum was used by Viswanathan et al. (2009), who replicated typical compensation findings with speech and tone-analogue precursors.

Four steady state sinewave tones at 1.800 Hz, 2,600 Hz, 3,000 Hz, and 3,400 Hz, matched to overall syllable intensity, were used as precursors. The first two tones were designed to mimic the typical F3 offsets of [aɹ] and [al], respectively. The third and fourth tones were synthesized at frequencies progressively farther from the critical F3 region and at the typical F4 offsets of these liquids (3,000 Hz for [aɹ] and 3,400 Hz for [al]). Each of the four precursor tones was combined with each member of the 11-step

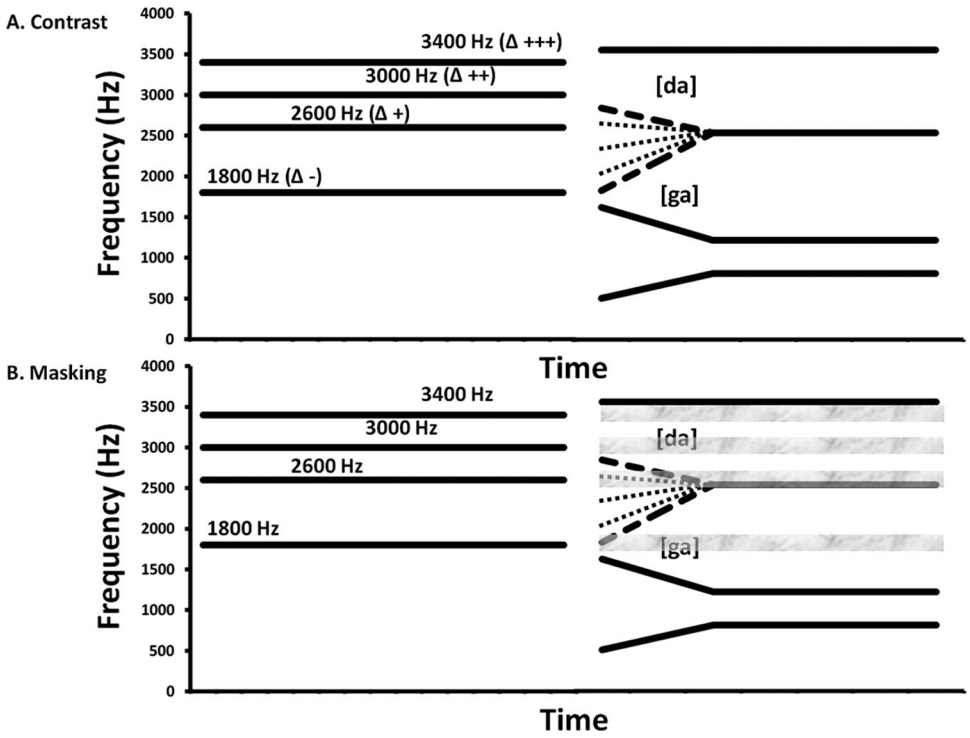

Figure 1. The predictions of the competing accounts are depicted. Several precursor tones are displayed, but the predictions apply to the presentation of one precursor tone in isolation. The contrast account (top panel) predicts effects on the manipulated F3 transition that results in the [da-ga] continuum. The contrast account predicts that precursor tones at a frequency greater than that of F3 should cause F3 to be perceived as lower than its true frequency, resulting in more "g" responses. Thus, the 2,600-, 3,000-, and 3,400-Hz tones should all increase "g" responses. (The direction of the frequency difference between precursor and target is indicated by the sign after Δ and the magnitude by the number of signs.) The 3,000- and 3,400-Hz tones should produce at least as many "g" responses as the 2,600-Hz tone (or possibly more if the change in "g" responses is proportional to the frequency difference). The 1,800-Hz tone is predicted to produce the fewest "g" responses because of its lower frequency relative to the targets' F3 onsets. The masking account (bottom panel) provides an alternative explanation, and predicts that the precursor tones' effects are to mask information in a frequency band centered at the tone frequency. The gray bars schematize the hypothesized regions of maximum influence of each tone. The masking account predicts that the two highest frequency tones' interference in detecting [da]s diminishes progressively with increased distance from the F3 onset of [da]. (The bands extend throughout the entire syllable, because their actual duration is unknown; what really matters is their effect near syllable onset.) The higher tones should thus produce *progressively weaker shifts* relative to the 2,600-Hz tone (which should interfere strongly with detecting [da]). The 1,800-Hz tone interferes strongly with the detection of [ga] and thus produces the fewest "g" responses.

continuum with an interval of 50 ms between them. This resulted in 44 distinct tone-syllable combinations. The stimuli were presented at an 11-kHz sampling rate, with 16-bit resolution diotically over headphones (Sennheiser HD-595) at approximately 70 dB SPL.

Procedure. The task was two-alternative forced-choice: participants pressed keys labeled "d" or "g" to indicate their identification of the target consonant. There were two blocks of trials. The first block consisted of practice trials presenting the [da] and [ga] endpoints with feedback. There were 12 trials with each endpoint, presented in random order. In the second block, each of the 44 tone-syllable combinations was presented five times, resulting in 220 trials. Participants were asked to classify the initial consonant in each CV as "d" or "g." No feedback was provided in this block.

Results and Discussion

Data were excluded from four participants with accuracy less than 80% in the endpoint identification block. Figure 2 shows results from the experimental block. The mean percentage of "g" responses averaged across steps of the continuum was lowest for the low-F3–1,800 Hz tone (43.5%) and highest for the high-F3–2,600 Hz tone (59.1%). The corresponding value for the low-F4–3,000 Hz tone was 52.7% and, important, higher than that for the high-F4–3,400 Hz tone (47.9%). The order of tones that produced the most "g" responses is as follows: high-F3–2,600 Hz tone > low-F4–3,000 Hz tone > high-F4–3,400 Hz tone > low-F3–1,800 Hz tone, the pattern predicted by the masking account. As we noted earlier, this pattern is inconsistent with spectral contrast. That account predicts a different ordering of conditions: high-F4–3,400 ≥ low-F4–3,000 > high-F3–2,600 > low-F3–1,800. To further examine the observed pattern, the data were submitted to a 4 (precursor) × 11 (step) within-subject analysis of variance (ANOVA). The main effect of precursor was significant ($F(3, 69) = 29.50$, $p < .0001$, $\eta_p^2 = 0.56$) indicating that percent "g" responses shifted according to the frequency of the precursor (as we unpack statistically below). The expected main effect of continuum step was also significant, $F(10, 230) = 268.41$, $p < .0001$, $\eta_p^2 = 0.92$, indicating that listeners' categorization responses changed across the continuum. The interaction, $F(10, 23) = 5.80$, $p < .0001$, $\eta_p^2 = 0.2$, was significant because the effect of precursor was stronger in the ambiguous portion of the continuum. This expected interaction was not explored further because visual inspection of Figure 2 reveals that, consistent with past studies of compensation for coarticulation, the largest effects of precursor are observed in the middle steps of the continuum (e.g., Viswanathan et al., 2009).

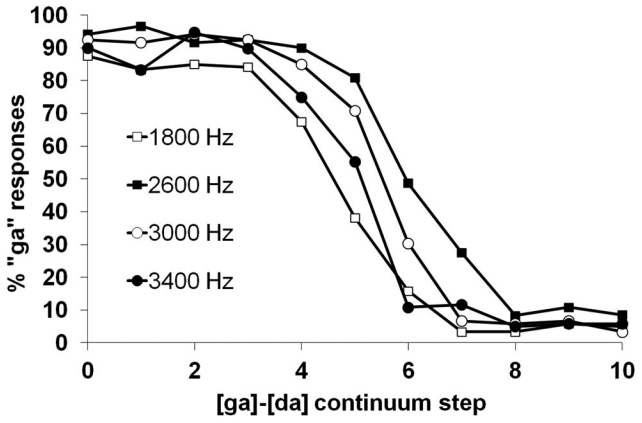

Figure 2. "Ga" responses as a function of precursor tone and continuum step. Square symbols denote tones in typical F3 region that produce strong boundary shifts. In this case, the higher F3 analogue (2,600 Hz, filled square) produces the most "g" responses. Circular symbols denote tones that are placed away from the F3 region (F4 analogues), and they produce comparatively weaker effects. In this case, the higher F4 tone (3,400 Hz, filled circle) produces *fewer* "g" responses than the lower F4 tone (3,000 Hz, open circle).

We investigated the main effect of precursor using a pair of planned comparisons for tones in each formant region. The first planned comparison of interest is between the two tones in the critical F3 regions. The high-frequency-F3 tone at 2,600 Hz produced more "g" responses than the low-frequency-F3 tone at 1,800 Hz, $F(1, 23) = 42.40$, $p < .0001$, $\eta_p^2 = 0.65$s, replicating numerous previous findings (e.g., Lotto & Kluender, 1998; Viswanathan et al., 2009). The second planned comparison of interest is between tones in the F4 regions; it confirmed the result, surprising from a contrast perspective, that the lower-frequency-F4 tone at 3,000 Hz produced *more* "g" responses than the higher-frequency-F4 tone at 3,400, $F(1, 23) = 17.75$, $p < .0001$, $\eta_p^2 = 0.44$. In fact, all pairs of post hoc comparisons were found to be significant (Figure 3) by a repeated measures extension of Tukey's HSD test, confirming statistically the pattern of results apparent in Figure 2. As predicted by a masking account, the tone closer to the critical F3 region produced more "g" responses than the tone farther and higher in frequency. The tone at 2,600 Hz (at the typical endpoint F3 of [al]) produced the greatest increase in "g" responses relative to the tone at 1,800 Hz. The overall pattern of results suggests that the farther a tone is from the critical F3 region, the smaller its influence on categorization.

Our finding is contrary to that reported by Holt (1999) that response shifts increase (and eventually asymptote) with increases in the frequency separation between a precursor and a target. It is unclear what causes the difference in results although there are some differences between the experiments. First, our precursors were pure tones rather than single formant analogues, and they did not preserve the harmonic information present in the formant analogues. It is possible that single formants provided vowel information influencing the resulting [ba]-[da] categorization; however, Holt reported that her precursors were not heard as speech.[3] Moreover, it is unclear why the effects should increase and then asymptote. Second, within the range of frequencies tested by Holt (1999), masking effects could have occurred such that lower frequency precursors masked F2s, and higher frequency precursors masked the F3s of the following target [ba]-[da] continuum. The results of Experiment 1 suggest that, even as a description (underlying mechanism aside), contrast is consistent only with effects of tones within the critical F3 region. When tones

[3] Other studies have shown that information for categorization can be different from the information for compensation (e.g., Mann, 1986; Viswanathan et al., 2010). In a related study, we found that listeners can show compensation even if they do not hear the precursor as speech (Viswanathan, Magnuson, & Fowler, 2012).

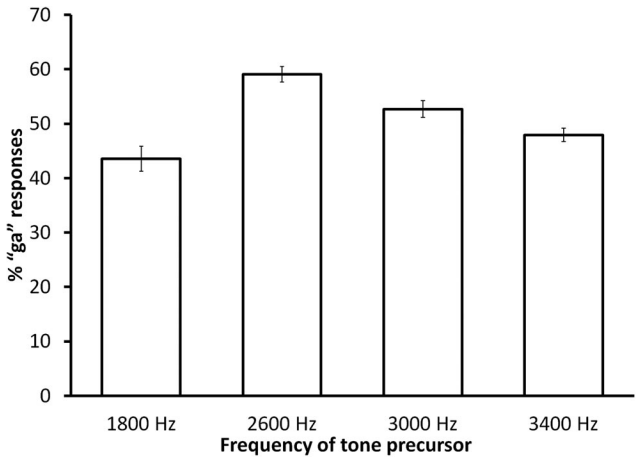

Figure 3. Mean "g" responses collapsed across continuum steps produced by each tone precursor. All differences except (between 1,800-Hz and 3,400-Hz tones; $p = .10$) were significant at $p < .01$ level.

farther away from the critical frequency region are considered, the effects they exert on speech categorization are not contrastive.

However, the pattern of results of Experiment 1 is as expected from a masking perspective. Tones closer in frequency to the critical region of the following syllable produce the strongest effects by interfering with detection of the critical frequencies (in the F3 region in our case) that carry information for identification of the following speech. When tones have frequencies increasingly far from this critical region, the masking effects get weaker because the region of interference shifts progressively away from the information required to identify target speech.

Can the spectral contrast and the masking accounts be reconciled? Could it be that spectral contrast and masking both exert influences on speech perception? Directly, this would mean that, in Experiment 1, in the F4 region, spectral contrast either plays no part or plays a weaker part compared with masking and that is why we failed to observe contrast-like effects, whereas in the F3 region, both masking and contrast worked together. However, coexistence of contrast and masking would mean that spectral contrast loses power as an explanatory principle. In fact, one of the appeals of the spectral contrast explanation is that it provides a single account for speech and nonspeech effects, and is thus more parsimonious (see Lotto & Holt, 2006) than the alternative of having a gestural explanation for speech effects and a masking explanation for nonspeech effects. Accepting coexistence of masking and spectral contrast would undermine this parsimony.

This leaves us with another possibility: Perhaps the spectral contrast account should be wholly rejected as an explanation for findings involving speech and nonspeech in compensation for coarticulation. We know that, given a preceding context (speech or nonspeech), some findings occur in the direction of spectral contrast (nonspeech contexts: Lotto & Kluender, 1998; critical (F3) region tones in our Experiment 1; speech contexts: Mann, 1980), some findings go against the direction of spectral contrast (nonspeech contexts: Mitterer, 2006, Experiment 2A; noncritical (F4) region tones in our Experiment 1; speech contexts: Tamil liquid [ar] in Viswanathan et al., 2010) and sometimes no effects are observed (nonspeech contexts: Fowler et al., 2000; speech contexts: excised F3 region from natural speech in Viswanathan et al., 2009). In short, the spectral contrast description of context effects neither accounts for all apparently relevant nonspeech effects nor for all apparently relevant speech effects.

Before we can reject the spectral contrast account, we must investigate whether contrast-like effects produced by F3 tones can also be explained by masking. In other words, we need to establish that masking and contrast accounts are not both required to explain nonspeech effects in the critical (F3) and the noncritical (F4) regions. We investigate this issue in Experiment 2.

Experiment 2

The masking account is that effects of nonspeech analogues of liquid contexts on target speech perception occur when the contexts interfere with listeners' ability to detect information in particular frequency regions (but see Holt, 2005, 2006). In Experiment 1, we found that, while tones in the critical F3-region can be described as contrastive (a higher frequency tone produces *more* "g" [low F3] responses), tones in the noncritical F4 regions show opposite effects (a higher frequency tone produces *fewer* "g" [low F3] responses). These effects are naturally explained from a masking perspective when the relative distance of each tone from the critical region is considered; as distance between a precursor tone and the critical region of a target stimulus increases, the precursor's ability to mask the target region diminishes. On a masking explanation, hearing a low F3 tone masks lower onset frequencies of [ga] near its own frequencies, leading to a perceived stimulus with effectively higher frequency in the F3 region and, therefore, more "d" judgments in the two-alternative-forced-choice task. Similarly, hearing a high F3 tone masks the higher onset frequencies of [da] leading to more "g" judgments. Thus, even though the resulting pattern of shifts appears contrastive, it may be the result of masking.

Furthermore, from a masking account, the smaller shifts following tones matched in F3-intensity relative to those matched to syllable-intensity (Viswanathan et al., 2009, Experiment 2) may be interpreted as the nonspeech precursor masking a smaller range of frequencies owing to the weaker concentration of energy in the critical frequency regions. This is consistent with the observation that the range of frequencies that are masked by a tone varies directly with its intensity (e.g., Plack & Oxenham, 1998).

These observations led us to make two testable predictions that follow from the masking account. First, if effects attributed previously to contrast are in fact masking effects, we should be able to simulate precursor effects without using a precursor by *removing* (via filtering) information in the target syllables that we assume to be masked by nonspeech precursors. Second, recall that our masking explanation of stronger compensation effects following higher-intensity tone precursors is that higher intensity precursors produce greater energetic masking; if so, simulating a wide masking field (consistent with a high-intensity precursor) should produce larger compensation effects than simulating a narrow masking field (consistent with a lower-intensity precursor).To test these predictions, we filtered out frequencies in the vicinity of each tone analogue in the target [da]-[ga] continuum. We created four sets of filtered [da]-[ga] tokens. The filter bands were centered at either a relatively high (2,600 Hz) or relatively low (1,800 Hz) frequency, and were either relatively wide (400 Hz) or narrow (100

Hz). The two sets of syllables with high-frequency filters were designed to resemble the frequencies in target syllables hypothesized to be available to a listener on the masking account following a typical high-frequency [al] tone analogue.[4] The two sets with lower frequency filters were designed to resemble the frequencies hypothesized to be available to a listener on the masking account following a typical low-frequency [aɹ] tone analogue. The wide band filters mimic the greater masking effects of higher amplitude precursor tones, while the narrow band filters simulate the lesser masking effects of lower amplitude precursor tones.

Figure 4 shows the unfiltered ambiguous member of the [da]-[ga] continuum in the middle panel flanked by the same token subjected to a high 400-Hz filter in the left panel and low 400-Hz filter in the right panel. If filtering captures the hypothesized effects of the tone analogues (high and low intensity) qualitatively on the perception of the speech targets by the masking account, we should observe response shifts in target perception that vary systematically depending on the region filtered. In particular, there should be more "g" responses with targets filtered in the high region than with those filtered in the low region. This would provide evidence that, even in the critical F3 region where effects are consistent with a contrast account, masking is a candidate explanation for the nonspeech effects. From Figure 4 it is apparent that the low filter (right panel) has the effect of attenuating the intensities of frequencies at the upper and lower edges of the second and third formant onsets respectively. Therefore, the F2 and F3 onset frequencies are separated more than in unfiltered syllables, reducing evidence for the "velar pinch" that serves as information for a velar place of consonant articulation (see, e.g., Ladefoged, 1993). This should result in *fewer* "g" responses in this condition compared with the high filter condition in which this separation for formant onsets does not occur (Figure 4, left panel). The resulting pattern therefore, should be consistent with those obtained after pure tones (more "g" responses after the high tone than low tone) even though the source of the effect is masking rather than spectral contrast.

Method

Participants. Eighteen male and 24 female University of Connecticut undergraduates, in the age range of 18–23, participated for partial course credit. All reported normal hearing.

Materials. We created four sets of continua starting with the continuum from Experiment 1 and manipulating two variables: Filter Location (high or low) and Filter Width (broad or narrow). One continuum was created by filtering out frequencies from the entire CV syllable between 2,400 Hz and 2,800 Hz (400 Hz Hanning bandpass window with 100 Hz smoothing, centered at 2,600 Hz) to mimic hypothesized effects of high-frequency [al] analogues. A second continuum, designed to mimic the hypothesized masking effects of typical low-frequency tone analogues of [aɹ], was created by filtering out frequencies between 1,600 Hz and 2,000 Hz (400 Hz Hanning bandpass window with 100 Hz smoothing, centered at 1,800 Hz). Two additional continua were generated to mimic the hypothesized (weaker) effects of masking produced by transient tones matched to the intensity of F3 (rather than the whole syllable, see Viswanathan et al., 2009, Experiment 2). For these continua, a narrower Hanning band stop filter of 100 Hz width with 25 Hz smoothing was used one centered at the F3 offset of [al] (filtering out frequencies between 2,550 Hz and 2,650 Hz), and the other centered at the F3 offset of [aɹ] (filtering out frequencies between 1,750 Hz and 1,850 Hz) for the fourth. These filter windows are within the typical frequency range of masking produced by nonspeech tones (Moore, 1995), and we expected them to qualitatively approximate the hypothesized masking characteristics of the tones used in nonspeech studies of compensation for coarticulation. The stimuli were presented at an 11 kHz sampling rate with 16-bit resolution diotically over headphones (Sennheiser HD-595) at approximately 70 dB SPL.

Procedure. The task was two-alternative forced-choice: participants pressed keys labeled "d" or "g" to indicate their identification of the initial consonant of each syllable. The location of the filtered region (high vs. low) was manipulated within subjects and the width (broad vs. narrow) between subjects. Each group participated in two blocks of trials. The first block consisted of practice trials presenting unfiltered [da] and [ga] endpoints with feedback. There were 12 trials with each endpoint, presented in random order. This block familiarized participants with the task and target syllables, and provided a basis for ensuring that they could identify the endpoints accurately. In the second block, each member of the 11-step continuum from two sets of filtered stops (high vs. low) was presented to each group (broad vs. narrow) 10 times, resulting in 220 trials presented in a random order. No feedback was provided in this block.

Results and Discussion

We excluded data from four participants with accuracy less than 80% in the endpoint identification block. This left 17 participants in the wide filter condition and 21 in the narrow filter condition. Figure 5 shows the data from the experimental block. Panel A shows the performance of the broad-window group and panel B the narrow-window group. For the sake of clarity, we present each group's data separately (we discuss group effects in a subsequent analysis). Figure 6 shows the percentage of "g" responses in the different filter conditions averaged across the continuum members.

The data were submitted to 2 (Filter Region) × 11 (Continuum Step) within-subject ANOVAs, one per subject group, to confirm response shifts. For both groups, the main effect of Filter Location was significant, low region $M = 42.0\%$; broad: high region $M = 58.4\%$; $F(1, 16) = 114.24$, $p < .0001$, $\eta_p^2 = 0.87$; narrow: low region = 50.6%; high region $M = 53\%$; $F(1, 20) = 14.32$, $p = .001$, $\eta_p^2 = 0.42$, with more "g" responses when the higher region was filtered (mimicking the effects of [al] tone precursors) than when the lower region was filtered (mimicking [aɹ] tone precursors). The expected main effect of Continuum Step was also significant for both groups, broad: $F(10, 160) = 268.63$, $p < .0001$, $\eta_p^2 = 0.94$; narrow: $F(10, 160) = 244.79$, $p < .0001$, $\eta_p^2 = 0.92$, indicating listeners' categorization changed across the continuum. The interaction, $F(10, 160) = 8.86$, $p < .0001$, $\eta_p^2 = 0.35$, was only significant for the broad group ($F < 1$ for the narrow group), indicating that the effect of the filtered region changed

[4] Strictly, the masking account is that the threshold for detecting the energy in the masked region is raised with the magnitude of increase determined by the strength of the masker. We mimic this effect by filtering out this energy.

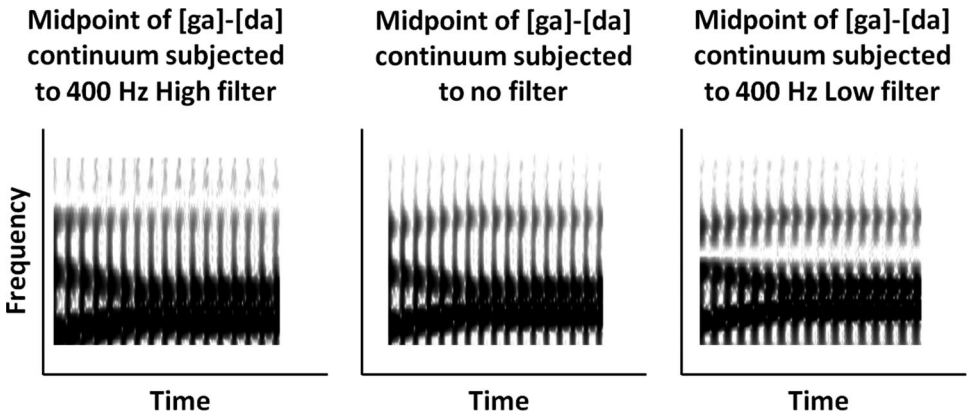

Figure 4. Spectrograms depicting effects of filters with width of 400 Hz on the midpoint of [da]-[ga] continuum. Note that the complete effects are not apparent because of smoothing. The spectrogram in the center panel shows the unfiltered syllable. The high 400-Hz filter (centered at 2,600 Hz) and low 400-Hz filter (centered at 1800 Hz) effects are shown on the left and right, respectively. (The 100-Hz filters are not depicted because their weaker effects are not clearly discernible on the spectrogram).

across the continuum, with effects larger in the more ambiguous middle region and in the [g]-end of the continuum.

Panel A shows an unpredicted result. Stimuli at the [ga] end of the continuum that had been filtered in the low [aɹ] analogue region are only labeled "g" about 60% of the time. This is most likely because the 400-Hz filter (range = 1,600 Hz to 2,000 Hz) has a disproportionate effect on the onset of the [ga] endpoints owing to its low F3 (1,800 Hz for the first member and 1,900 Hz for the second). However, restricting our analyses to continuum steps 3 to 11 did not qualitatively alter the results, $F(1, 16) = 91.56$, $p < .0001$, $\eta_p^2 = 0.85$, and this is confirmed by the clear separation of the curves observed in Panel A of Figure 5. The atypical categorization is restricted to the [ga] end of the continuum. However, this anomaly shows that our choice of parameters may have led to more drastic effects than the hypothesized masking effects of the tonal precursors. Despite this limitation, the original question of whether hindering information pickup results in similar boundary shifts to those obtained with typical nonspeech precursors is answered in the affirmative. Irrespective of filter size, target members filtered in the high-frequency region produce more "g" responses than those filtered in the low-frequency region.

Next, a separate analysis on difference scores was performed to analyze the interaction between filter width and step. For each group, the extent of the difference in percent "g" responses depending on the filter was determined by calculating the signed difference between percentage "g" judgments to the high filtered region and the low filtered region for each participant at each step. These were submitted to an 11 (step) × 2 (Filter Width, narrow or broad) mixed ANOVA. There was a strong effect of Filter Width (Broad $M = 16.4\%$ vs. Narrow $M = 2.3\%$), $F(1, 36) = 84.22$, $p < .0001$, $\eta_p^2 = 0.71$, indicating a significant decrease in the effect of the filters on difference scores in the Narrow filter group. In other words, a reduction in response shift was observed for narrow versus broad filters that is qualitatively similar to that observed

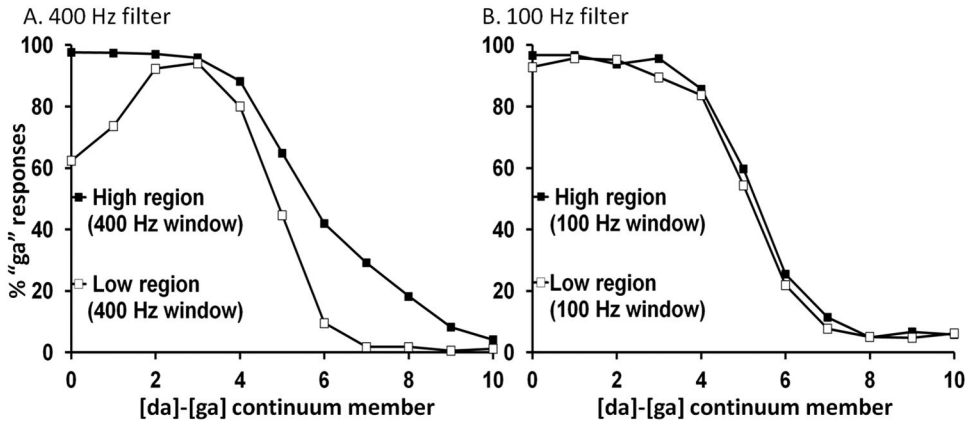

Figure 5. Effects of removing hypothesized masked frequencies from the target continuum on target categorization. Panel A shows stronger effects of using a 400-Hz filter window to represent masking effects that follow typical high intensity tone analogues. Panel B shows weaker effects of using a 100-Hz filter window to represent masking effects of lower F3 intensity tone analogues.

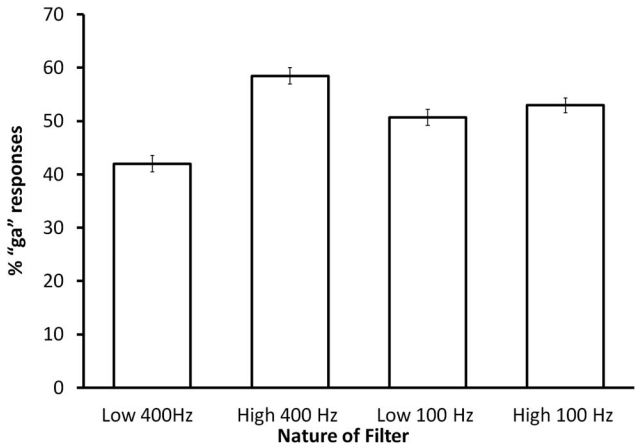

Figure 6. Mean "g" responses collapsed across continuum steps produced in each filter condition. Continua filtered in high-frequency regions produced more "g" responses than those filtered in low-frequency regions and size of this effect depended on the width of the filter used.

when precursor tones are matched to F3 intensity and trajectory rather than being matched in intensity to the whole precursor syllable. This suggests that reduction in masking provides a plausible explanation for the reduction in potency of the tones matched to the intensity of the formant rather than the syllable (Lotto & Kluender, 1998, Experiment 2; Mitterer, 2006, Experiment 2B; Viswanathan et al., 2009, Experiment 2). The significant effect of Continuum Step, $F(10, 360) = 7.72$, $p < .001$, $\eta_p^2 = 0.176$, indicates that the size of the effect of a high versus low-frequency filter on percent "g" responses was different for different steps of the continuum. The Continuum Step × Filter Width interaction was also significant, $F(10, 360) = 6.01$, $p < .001$, $\eta_p^2 = 0.143$, indicating that the effect of changing the filter width (broad to narrow) was different at different steps of the continuum (clearly driven by the nonuniform differences for high and low filters for the Broad filter group seen in Panel A of Figure 5).

These findings confirm the predictions of a masking account that contrast-like shifts in the target categorization can be obtained in the absence of contrast-producing precursors. Therefore, they show that a masking explanation of effects of nonspeech tone analogues is viable and sufficient to account for context effects in the vicinity of F3.

General Discussion

The gestural account of compensation for coarticulation is that listeners' shifts in perceptual category boundaries reflect their attunement to coarticulatory gestural overlap in speech production. A strong argument put forth against this account has been the observation that nonspeech tones, bereft of gestural information, produce shifts in responses that are qualitatively similar to the effects of speech precursors.

In the present experiments, we considered the possibility that speech and nonspeech effects may have different origins. The gestural explanation for speech context effects received support in an earlier study in which its predictions were directly dissociated from those of a contrast account put forward to explain both speech and nonspeech effects (Viswanathan et al., 2010). In the present project, we examined an energetic masking account of nonspeech context effects. In short, the explanation is that nonspeech tones differ in crucial ways from the speech regions to which they are supposed to be analogous (Viswanathan et al., 2009). Typically, they are more intense, steady, and have their energy concentrated in a single frequency region unlike the corresponding speech formants. As argued by Fowler et al. (2000) and supported by the current experiments, these qualities of the tones produce *energetic forward masking* that prevents listeners from detecting gestural information in the frequency regions in the following speech that surround the frequency of the masking tone. In previous research (Viswanathan et al., 2009), we found that the F3 regions of natural speech precursors presumed to be critical for contrastive effects had no effect on phoneme identification when presented by themselves (with noncritical regions removed). They were not heard as speech (but see footnote 3), and they did not have the right acoustic properties to serve as maskers (because they were not more intense than acoustic energy in the speech stimuli that followed them), although they did have the right frequency properties to produce contrast effects.

Additional support for the masking hypothesis comes from findings demonstrating that tone precursor effects diminish when tone analogues are more closely matched to actual formant properties such as amplitude and trajectory (Viswanathan et al., 2009) or when the F3s of following speech are made unnaturally intense (Fowler et al., 2000). To test whether masking is a better explanation of nonspeech effects than spectral contrast, we adopted a two-part strategy. In Experiment 1, we investigated whether nonspeech effects are always contrastive. We generated tones at increasing distances (2,600 Hz, 3,000 Hz, and 3,400 Hz) from the typical low-frequency [aɹ] analogue (at 1,800 Hz). The 2,600-Hz tone was at the F3 offset frequency for [al] used in previous studies, and we included two more tones matched to the typical F4 offset frequencies of [aɹ] and [al] (3,000 Hz and 3,400 Hz, respectively), and above the critical F3 region of either [aɹ] or [al]. The results of Experiment 1 were conclusive. The farther the tone frequencies were from the critical region, the weaker were the effects on categorization, as predicted by a masking account, but not by a contrast account. Furthermore, a direct examination of the effects of higher tones (2,600 Hz vs. 3,000 Hz vs. 3,400 Hz) showed that the higher the tone, the *lower* the proportion of "g" responses it elicited, revealing an effect that is not amenable to a contrast description (in which higher frequency tones induce fewer "g" [low F3] responses). While these results are problematic for a spectral contrast account, they are consistent with the prediction from a masking account that tones placed farther from the critical region should have weaker effects on target categorization. Critically, it demonstrates that nonspeech effects are not always contrastive, calling into question the generality of a contrast explanation.

When Holt (1999) investigated the same issue in a vowel-stop context she found a different outcome. As tones increased in frequency, the size of the shift increased until it reached a constant size. This seemingly contradictory finding deserves attention. One possibility is that the masking account applies strongly in the liquid-stop context in which the F3 is relatively weak in intensity and therefore more susceptible to masking than Holt's more intense F2 stimuli. If this is the right explanation, it would raise questions about the generality of a masking account. An alternative

explanation discussed earlier is that in the frequency region that Holt explored (600 Hz to 2,700 Hz), the tones masked either F2 or F3 of the following stop, producing shifts throughout the range. Of course, this hypothesis requires empirical confirmation.

In Experiment 2, we asked directly whether a masking account can explain the contrast-like shifts elicited by the F3 tonal precursors in previous studies. Specifically, we asked whether simulating effects of masking by filtering target continuum members would give results qualitatively like those attributable to precursor tones. The filter location was manipulated (high region for [al] tone analogues vs. low region for [aɹ] tone analogues) to mimic the hypothesized masking effect of preceding high- and low-frequency tones. An additional manipulation of filter size (broad vs. narrow) was included to simulate the effect of varying tone intensity on the nonspeech effects (where greater intensity would result in a broader masking region). The results were clear. Targets filtered in the high-frequency region consistently produced more "g" reports than those filtered in the low-frequency region. This suggests that contrast-like effects can be obtained because of masking in the critical F3 region. Furthermore, the size of this difference was less when a narrower region was removed from the target, qualitatively mimicking the effect of lowering the intensity of the nonspeech tone precursor—a result that does not follow from a spectral contrast account.

From these experiments, we draw two primary conclusions directly relevant to the question of what underlies compensation for coarticulation. First, spectral contrast does not explain all nonspeech context findings. In particular, it fails to explain why a higher F4 tone in Experiment 1 produces fewer "g" responses than a lower F4 tone or why the high F3 tone, which is lower than either of the F4 tones, produces the most "g" responses. Furthermore, the fact that tone effects are not always contrastive rules out the possibility that masking is a mechanism that underlies spectral contrast (also see Holt, 1999). Second, we conclude that typical contrast-like effects of nonspeech precursors is most likely because of masking. Masking, but not contrast, explains why lower intensity F3 analog tones (typical of speech) produce weaker effects than those with higher intensities (i.e., greater intensity than the formant the tone is meant to simulate). Generally, our findings augment earlier ones challenging a spectral contrast explanation of compensation for coarticulation (see Viswanathan et al., 2009; 2010), suggesting that spectral contrast explains neither speech nor analogous nonspeech context effects satisfactorily. On the contrary, we suggest that a masking account of the nonspeech context effects on speech perception that we have been addressing is viable, given our results in addition to others described in the introduction (Fowler et al., 2000; Lotto & Kluender, 1998; Viswanathan et al., 2009). It provides a plausible explanation of nonspeech effects; that is, it explains why nonspeech precursors sometimes mimic speech effects that are best explained by gestural accounts.

In addition to their relevance to the debate regarding the origin of compensation for coarticulation, we note that these studies have implications for studying speech perception in general. Critically, we suggest that the use of nonspeech materials in studies of speech perception can be tricky, and that the results of such studies must be interpreted cautiously. In particular, when nonspeech materials (or nonhuman listeners; Lotto, Kluender, & Holt, 1997) are used, qualitatively similar response patterns may arise from fundamentally different sensory and perceptual processes. In such cases, we suggest that these results can serve only as starting points for further investigation rather than as automatic confirmation of identical underlying causation. The results of our current experiments add to previous ones (Fowler, 1990; Viswanathan et al., 2009) that demonstrate that such investigation often reveals different sources of qualitatively similar behavior.

For the study of speech perception more generally, the described research on compensation for coarticulation illustrates the caution with which inferences based on similarity must be approached. Nonspeech materials are often used to implicate domain-general mechanisms in speech perception by designing nonspeech stimuli that retain properties of natural speech deemed critical by the experimenter. We suggest that when such an approach is adopted, the ecological validity of the nonspeech materials must be kept in mind. Failure to consider information critical to the perceiver may mislead the experimenter regarding the true cause of the resulting effects.

Conclusions

Although we have shown that some studies using nonspeech precursors for speech syllables have provided misleading results, we do not intend to suggest that, in general, studies using nonspeech materials cannot contribute to an understanding of speech perception. Studies of masking in both auditory and visual processing, for example, are useful in providing information about the operating characteristics of these perceptual systems. However, such effects as masking should not mislead investigators as to the factors most relevant to phonetic perceivers or to the nature of perception itself. Acoustic signals, reflected light, and other media serve as information for perceivers about events in the world that they need to know about. Their task, as perceivers, is to extract information about those events from informational media. Compensation for coarticulation appears to reflect their doing just that in the domain of speech perception. Nonspeech studies, in this context, clearly demonstrate that perceptual systems are not infinitely sensitive to information. In hearing, listeners are not sensitive to all frequencies of vibrating air, and extraction of information in frequency ranges to which the ear *is* sensitive can be impaired by such effects as masking. A complete picture of speech perception, then, cannot end with the observation of similar responses under different informational conditions; instead, it requires an understanding of the acoustic information in speech signals, as well as its transduction by auditory systems, under different contextual conditions.

References

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.

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

Cutting, J. E., & Rosner, B. S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics, 16,* 564–570. doi:10.3758/BF03198588

Diehl, R. L., Elman, J. L., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Hu-*

man Perception and Performance, 4, 599–609. doi:10.1037/0096-1523.4.4.599

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

Elliott, L. L. (1971). Backward and forward masking. *Audiology, 10,* 65–76. doi:10.3109/00206097109072544

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

Fowler, C. A. (1990). Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. *Journal of the Acoustical Society of America, 88,* 1236–1249. doi:10.1121/1.399701

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America, 99,* 1730–1741. doi:10.1121/1.415237

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*(5 Pt 1), 2801–2817. doi:10.1121/1.2354071

Holt, L. L., & Lotto, A. J. (2002). Behavioral examinations of the neural mechanisms of speech context effects. *Hearing Research, 167,* 156–169. doi:10.1016/S0378-5955(02)00383-0

Kluender, K. R., Diehl, R. L., & Killeen, P. R. (1987). Japanese quail can learn phonetic categories. *Science, 237,* 1195–1197. doi:10.1126/science.3629235

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosives. *Science, 190,* 69–72. doi:10.1126/science.1166301

Ladefoged, P. (1993). *A course in phonetics* (3rd ed.). New York, NY: Harcourt, Brace, Jovanovich.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21,* 1–36. doi:10.1016/0010-0277(85)90021-6

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

Lotto, A. J., Sullivan, S. C., & Holt, L. L. (2003). Central locus for nonspeech effects on phonetic identification. *Journal of the Acoustical Society of America, 113,* 53–56. doi:10.1121/1.1527959

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

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. doi:10.1016/S0010-0277(86)80001-4

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

Moore, B. C. J. (1995). *Hearing*. San Diego, CA: Academic Press.

Plack, C. J., & Oxenham, A. J. (1998). Basilar membrane nonlinearity and the growth of forward masking. *Journal of the Acoustical Society of America, 103,* 1598–1608. doi:10.1121/1.421294

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. (2012). Compensation for coarticulation in sinewave-speech contexts. Manuscript in preparation.

Warren, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review, 92,* 574–584. doi:10.1037/0033-295X.92.4.574

Zwislocki, J. J., Buining, E., & Glantz, J. (1968). Frequency distribution of central masking. *Journal of the Acoustical Society of America, 43,* 1267–1271. doi:10.1121/1.1910978

Received February 16, 2012
Revision received September 20, 2012
Accepted September 25, 2012 ■