Published in final edited form as:
Psychon Bull Rev. 2009 February ; 16(1): 74–79. doi:10.3758/PBR.16.1.74.

A critical examination of the spectral contrast account of compensation for coarticulation

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

Abstract

Vocal tract gestures for adjacent phones overlap temporally, rendering the acoustic speech signal highly context dependent. For example, following a front place of articulation, a posterior segment is pulled frontward, and listeners' category boundaries shift appropriately. Some theories assume that listeners perceptually attune or compensate for coarticulatory context. An alternative is that shifts result from spectral contrast. Indeed, shifts occur when speech precursors are replaced by pure tones frequency matched to the formant offset at the assumed locus of contrast (Lotto & Kluender, 1998). However, tone analogues differ from natural formants in several ways, raising the possibility that conditions for contrast may not exist in natural speech. When we matched tones to natural formant intensities and trajectories, boundary shifts diminished. When we presented only the critical spectral region of natural speech tokens, no compensation was observed. These results suggest that conditions for spectral contrast do not exist in typical speech.

Keywords

Compensation for coarticulation; spectral contrast; speech perception

When individuals speak, sequences of vocal tract gestures are *coarticulated*; the gestures for one phone temporally overlap those for adjacent phones. Coarticulation makes the acoustic manifestations of phones highly context dependent. For example, if a speaker produces the sequence [lg], as in [alga], the point of constriction during [g] is pulled forward from the velar place of [g] owing to temporal overlap (coproduction) of the anterior constriction for [l]. The point where the tongue hits the palate during [g] approximates a weighted vector sum of the constriction locations of the coarticulated segments (Fowler & Smith, 1986). Thus, during the interval in which [g] has primary but not exclusive control of the vocal tract, the speech signal has an acoustic structure shifted away from canonical [g] (and toward [d]), reflecting the more front constriction location. The converse applies when a back-front sequence is produced, as in [arda].

Mann (1980) found that listeners behave as if they compensate for coarticulation: category boundaries between phones shift systematically depending on coarticulatory context. For example, the front-back boundary between [d] and [g] shifts forward after a segment with a back place of articulation, and vice-versa, suggesting listeners' sensitivity to coarticulation. Two sorts of explanations have been proposed for this phenomenon.

The first posits that perceivers attune to the event of coarticulation (via its acoustic consequences) or its acoustic correlates themselves. For example, the direct realist theory of speech perception claims that compensation for coarticulation occurs because listeners

Department of Psychology University of Connecticut 406 Babbidge Road, Unit 1020 Storrs, CT 06269-1020 Ph: (860) 486 2549 navin.viswanathan@uconn.edu.

directly perceive vocal tract gestures, and are sensitive to acoustic effects of coarticulation. (See Fowler, 2006, for an overview.) Thus, for the [alga] example, the direct realist account is that listeners are able to distinguish acoustic information for anterior [l] from that for [g], leading to appropriate context dependent responses. Mann and Repp (1981) suggest other explanations that also appeal to sensitivity to coarticulation, including a perceptual learning account on which compensation reflects perceptual experience with acoustic contingencies resulting from coarticulation.

A second sort of explanation does not assume that listeners are directly sensitive to coarticulation. Mann (1980) speculated that sensory contrast based on the relative spectral distribution of energy between the precursor segment and target might account for compensation effects. Specifically, F3 is relatively high for [l] and [d], but relatively low for [r] and [g]. Hearing the high F3 during [l] might produce a contrastive effect, causing the F3 of the following segment to be heard as lower (and hence more “g” like). Conversely, hearing a low F3 during [r] might cause the following stop’s F3 to be perceived as higher (and hence more “d” like).

In support of this account, Lotto and Kluender (1998, Experiment 3) found a similar [da]-[ga] boundary shift when natural speech precursors were replaced by pure steady tones at the F3 offsets of [l] and [r]. These results have been extended to a variety of contexts, including vowel-consonant coarticulation (Holt, Lotto, & Kluender, 2000), anticipatory coarticulation (Wade & Holt, 2005), and effects of speech precursors on perception of non-speech targets (Stephens & Holt, 2003). These findings question the interpretation that boundary shifts caused by speech precursors reflect true perceptual compensation for coarticulation. Instead, they seem to suggest that compensation for coarticulation is just another example of the general phenomenon of spectral contrast.

This conclusion is viable only if the acoustic conditions for spectral contrast are present in natural speech. However, this assumption is untested: indeed, non-speech precursors have typically differed from the speech formants they model, along three major dimensions. First, tone analogues are typically matched to the intensity of the speech syllable rather than to the formant they represent. Second, steady tones at formant offset frequencies are typically used instead of tones reflecting formant transitions. Third, the harmonic structure of the critical formant is not captured by the sinewave tone analogues (cf. Holt, 1999).

There is preliminary evidence that matching tones along these dimensions might affect the potency of contrast effects. Lotto and Kluender (1998, Experiment 2) found that sinewave tones tracking the transitions of the speech syllables and intensity matched to F3-region[1] (matched along two of three dimensions of differences listed above) produced numerically *smaller* effects than the original speech syllables. In contrast, effects with steady tones matched to entire syllable intensity generally led to effect magnitudes comparable to those observed with natural speech. However, Lotto and Kluender (1998) did not report statistical tests of the different effect magnitudes. Similarly, Mitterer (2006; Experiment 2B) investigated whether sinewave analogues of fricatives were sufficient to produce contrastive effects on judgments of a following vowel. He used steady tone analogues of fricatives, but with intensities matched to F3, and failed to find contrast effects. We will test directly whether compensation for coarticulation with natural speech can be attributed to spectral contrast, by examining whether the conditions for spectral contrast are provided by natural speech materials.

[1]Note that Lotto and Kluender (1998) used tones modeling the formant transitions alone (F3 glides), while our tones reflected both the steady and transient parts of F3.

From a spectral contrast perspective (e.g., Lotto & Kluender, 1998), compensation for coarticulation in the disyllables used by Mann (1980) is due to the spectrally contrastive nature of their third formant transitions. We test this explanation directly by providing only the F3 region of natural speech syllables as precursors in Experiment 1. If the conditions for spectral contrast for the Mann (1980) items are provided by F3 in natural speech, the critical region by itself should be as effective as the whole speech syllable in inducing boundary shifts.

Experiment 1

Two groups identified syllables from a [da]-[ga] continuum. For one group, the stops were preceded by natural tokens of [al] or [ar], to replicate compensation studies using natural speech precursors. Another group heard precursors only including the critical F3 region of these syllables, providing a direct test of whether a natural F3 provides conditions for spectral contrast.

Method

Participants—Forty-four University of Connecticut undergraduates (twenty-two in each group) participated for 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] via the source-filter method using *Praat* (Boersma, 2001). 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.

In one condition, stops were preceded by natural tokens of [al] and [ar]. The precursors were matched in duration (375 ms) and intensity. The critical F3 offsets were approximately 2600 Hz for [al] and 1820 Hz for [ar]. F2 and F4 offsets were approximately 930 Hz and 3530 Hz for [al], and 1400 and 3050 Hz for [ar]. These acoustic characteristics are similar to those of materials used in previous studies (e.g., Lotto & Kluender, 1998).

In the second condition, the precursors were filtered to isolate the critical F3 region of the natural syllables. We used a Hanning band pass filter passing frequencies between 1600 Hz and 3000 Hz with a smoothing of 100 Hz to isolate the F3 region. The frequency window was selected by examining the acoustic profile of the natural syllables and ensuring that the F3 information in each liquid was preserved even though we used the same window for both syllables.

VC precursors and CV targets were combined, separated by a silent gap of 50 ms, to form 22 unique disyllables (2 precursors X 11 target) per group. The stimuli were presented at an 11 kHz sampling rate with 16 bit resolution. Spectrograms of the precursors in each condition with a sample target syllable are shown in panels A and B of Figure 1.

Procedure—The task was two-alternative forced-choice: participants pressed keys labeled “d” or “g” to indicate their identification of the target. There were two blocks of trials. The first block consisted of practice trials presenting only the [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 of the 22 disyllables was presented eight times, resulting in 176 trials. No feedback was provided in this block.

Results

We excluded data from two participants in each condition with accuracy less than 80%[2] in the endpoint identification block, leaving 20 participants in each condition. The data from the second block were submitted to a 2 (groups) × 2 (precursor) X 11 (step) mixed ANOVA. The main effect of precursor was significant ($F(1, 38) = 20.20$, $p < .001$, $\eta^2_p = 0.35$), with more "g" responses following [al] precursors (51.3%) than [ar] precursors (47.3%; see Figure 1, panels A & B). The main effect of condition was not significant ($F < 1$, $\eta^2_p = 0.02$). Most critically there was a strong interaction between precursor and group ($F(1, 38) = 16.38$, $p < .001$, $\eta^2_p = 0.30$), suggesting a difference in compensation across precursor conditions. Planned contrasts confirmed that the natural syllables produce the expected compensation effect (7.6% more "g" responses following [al] than [ar]; $F(1, 19) = 32.70$, $p < .001$, $\eta^2_p = 0.63$), replicating earlier studies (e.g. Mann, 1980) however the ostensible critical region alone did not produce a compensation-like effect (only 0.4% more "g" responses following the F3 region of [al] than that of [ar]; $F < 1$, $\eta^2_p = 0.07$). Although the absence of a reliable shift with filtered precursors must be interpreted with caution, these items did contain the F3 region of the very syllable precursors that produced a strong boundary shift effect. Moreover, *any* reduction of the effect (let alone the absence observed here) sufficient to yield a reliable interaction is problematic for a contrast account in which compensation for coarticulation is directly attributed to the contrastive effects of precisely the critical region these precursors isolate.

We designed Experiment 2 to investigate why, despite robust boundary shift effects following single-formant tone analogues matched to F3 (e.g., Lotto & Kluender, 1998), the filtered F3 region of speech by itself does not produce category shifts. Perhaps the higher intensity and energy concentration at the offset frequency in typical tone analogues, compared to the critical F3 region of natural speech, provides the conditions needed for spectral contrast. If so, contrast effects should diminish when tones are matched to additional natural formant characteristics. We test this hypothesis in Experiment 2.

Experiment 2

Three groups participated in this experiment. One group heard typical steady tone analogue precursors matched in frequency to F3 offsets, and matched to the intensity of whole syllables (as in Lotto & Kluender, Experiment 3). The precursors for the second group were the same, except that intensity was matched to the critical formant rather than the whole syllable. The precursors for the third group were matched both in intensity and trajectory to the critical third formant. If steady, high intensity tones provide the strongest conditions for spectral contrast, we should find that effects weaken as tones are progressively matched to additional F3 characteristics.

Method

Participants—Sixty-eight University of Connecticut undergraduates (24 in group 1, 21 in group 2, and 23 in group 3), who reported normal hearing, participated for course credit. None had participated in Experiment 1.

Materials—The [da]-[ga] continuum from Experiment 1 was used. Two steady state sinewave tones were used as precursors for the first group: a high F3 tone at 2600 Hz (F3

[2]Using a cutoff of 90% (cf. Lotto & Kluender, 1998) did not qualitatively change our results.

offset of [l]), and a low F3 tone at 1820 Hz (F3 offset of [ar]). Following Lotto and Kluender (1998, Experiment 3), the intensities and durations of the precursor tones were matched to the overall intensities and durations of precursor syllables used in Experiment 1.

The second group heard steady tones at the same frequencies as group 1 with a crucial difference: tone intensities were matched to the intensity of the formant they were designed to represent (as in Mitterer, 2006, Experiment 2B). The intensity of the higher frequency [al] tone was set to 48 dB and the lower frequency tone to 52 dB, obtained by measuring the intensities of the filtered [al] and [ar] critical regions used in Experiment 1.

The third group heard tonal analogues with intensities and trajectories matched to those of the critical formant from the syllable they were designed to represent. These tones were generated by tracking the center frequency of the formants in the natural speech tokens used in Experiment 1.

Procedure—The procedure of Experiment 1 was used.

Results

We excluded eight participants with accuracy less than 80% in endpoint identification, leaving 20 participants in each condition. Data from the second block (shown in Figure 1, panels C, D, and E) were submitted to a 3 (group) × 2 (precursor) X 11 (step) mixed ANOVA. The main effect of precursor was significant ($F(1, 57) = 50.18$, $p < .001$, $\eta^2_p = 0.47$), as there were more "g" responses following [al] analogues (56.3%) than [ar] analogues (47.4%). The effect of group was not significant ($F(2, 57) = 2.113$, $p = 0.13$, $\eta^2_p = 0.06$). However, the critical group × precursor interaction was significant ($F(1, 57) = 7.83$, $p = .001$, $\eta^2_p = 0.22$), with the effect of precursor diminishing (see Figure 2) as precursors were incrementally matched to conditions in speech. Pairwise comparisons indicated that tones matched to F3 in both their frequency trajectory and intensity, produce smaller effects than syllable-intensity steady tones at critical formant offsets ($p = .045$). Other comparisons were not significant. A comparative analysis across both experiments is reported at the end of this section.

Although these findings agree in pattern with past studies using liquid-stop contexts (Lotto & Kluender, 1998; Experiments 2 and 3), they do not appear to generalize to a consonant-vowel-consonant context. For instance, Holt et al. (2000) used tone analogues matching *F2* formant trajectory (the origin of contrast in this context) that appear to produce *larger* effects than steady tones at F2 offset. To explain this difference, Holt et al. (2000) speculate that perhaps steady tones placed at formant offsets (similar to those in our study and in that of Lotto & Kluender, 1998, Experiment 3) would produce weaker effects on vowel categorization than those placed at the center frequency of the critical formant.

We agree that the specific choice of frequency for steady tone analogues is likely to be crucial in affecting the magnitude and occurrence of contrast effects across different coarticulatory contexts. One approach, as Holt et al. (2000) suggest, is to construct an empirically-based psychoacoustic model to guide this decision. However, a more direct approach is to test the critical region claimed to be causal, as we did in Experiment 1 (by filtering to isolate the critical region). On this approach, one can be sure neither to underestimate nor overestimate the conditions present in natural speech for spectral contrast.

Omnibus analysis

We conducted an omnibus analysis to compare results across conditions in the two experiments. We calculated "amount of compensation" by subtracting the mean percentage

of "g" responses to [ar] (or, in non-speech conditions, its analogue) from the mean percentage of "g" responses to [al] (or its analogue), at each step of the continuum, for each subject. The results for each condition are shown in Figure 2. The figure suggests that typically-used non-speech tones yield the numerically largest contrast effects, and, as these analogues are made more speech-like, their effects diminish.

We submitted the difference scores to a 5 (condition) X 11 (step) mixed ANOVA. The effect of condition was significant ($F(4, 95) = 9.19$, $p < .001$, $\eta^2_p = 0.28$), indicating that the amount of compensation varied across types of precursors. The effect of step ($F(10, 950) = 24.38$, $p < .001$, $\eta^2_p = 0.20$) was also significant, because compensation varied across the continuum. We conducted planned comparisons on the amount of compensation in the speech condition (7.6%) versus that with each analogue pair. There was greater compensation for the speech condition than for intensity-matched transient tones (2.8% compensation; $F(1, 38) = 9.92$, $p < .003$, $\eta^2 = 0.21$), and the critical region (0.4% compensation; $F(1, 38) = 16.22$, $p < .001$, $\eta^2 = 0.30$). Compensation in the typical steady tone condition (15%) was marginally greater than in the speech condition ($F(1, 38) = 4.07$, $p = .051$, $\eta^2 = 0.10$). The difference between formant-intensity steady tones (9.1% compensation) and the speech condition was not reliable ($F < 1$, $\eta^2 = 0.01$).

General Discussion

Whereas other accounts of the phenomenon of compensation for coarticulation appeal to attunement to acoustic information for coarticulation itself (Fowler, 2006) or to learning acoustic contingencies that coarticulation causes (Mann & Repp, 1981), the spectral contrast account proposes that "compensation" is a result of a general contrast phenomenon. As Mann (1980) pointed out, F3 differences between [r] and [l] relative to those of [d] and [g] correlate with the boundary shifts observed in [d]-[g] identification in the context of the liquids. When Lotto and Kluender (1998) first tested this hypothesis, their evidence appeared clear: pure tones matched in frequency to the F3 offsets of [ar] and [al] had similar effects on [d]-[g] identification as natural tokens of [ar] and [al]. Such results seem to imply that contrast between liquid and stop F3 frequencies was the likely cause of boundary shifts observed with natural speech precursors. However, the validity of this account requires that energy in natural liquid formants provides the conditions for spectral contrast. Heretofore, this has not been tested.

In Experiment 1, we replicated compensation for coarticulation with natural speech precursors. To test directly whether the F3 region of the natural precursors provides a sufficient basis for spectral contrast, we filtered the natural precursors such that only the F3 region remained. Listeners presented with these materials did not exhibit compensation for coarticulation. In Experiment 2, we found that, as tone precursors are made progressively more speech-like, contrast effects diminish. In fact, the tone analogues *most unlike* speech (high-intensity steady tones, of the sort typically used in contrast studies) produced responses similar in magnitude to those elicited by speech. We found progressively weaker results as tones were matched in intensity and trajectory to the formant they modeled compared to steady, syllable-intensity tones.

The most direct implication of the absence of compensation given F3 alone is that compensation effects obtained with speech syllables cannot be attributed to the spectrally contrastive effects of a natural speech precursor's F3. Moreover, compensation-like boundary shifts found in typical contrast experiments using tones may follow from characteristics of tone precursors that are unlike relevant characteristics of natural speech. This implies that compensation-like shifts following tone precursors that appear to indicate spectral contrast cannot be automatically generalized to effects in natural speech. Thus, the

current results pose a substantial challenge for the spectral contrast account of compensation for coarticulation. Our results do not provide direct evidence for accounts that explicitly posit *compensation* for *coarticulation* (e.g., direct realism, Fowler, 2006, or perceptual learning, Mann & Repp, 1981). These coarticulatory accounts do not directly address effects of tonal precursors on speech categorization. On the other hand, the spectral contrast account provides clear, testable predictions concerning boundary shifts that follow such tonal precursors.

Similar responses in tonal and speech conditions in compensation experiments need not imply identical bases of these effects (cf. Fowler, 1990). Indeed, compensation for coarticulation occurs in cases in which no spectrally contrastive segments are present. For example, precursor differences leading to compensation effects may be visually specified (Fowler, Brown, & Mann, 2000; Mitterer, 2006, Experiment 3; for a debate about visual influences on compensation, see Holt & Stephens, 2003; Fowler, 2006; Lotto & Holt, 2006). In addition, compensation occurs when coarticulating gestures are cotemporal (e.g., Silverman, 1986), in the absence of contrastive contexts.

Furthermore, in a recent study, we dissociated effects of place of articulation and F3 in a liquid precursor context (Viswanathan, Magnuson, & Fowler, under review). We utilized a trilled “r” from Tamil with an alveolar place of articulation (comparable to the English “l”) but a low F3 (comparable to the English “r”). Our listeners’ responses to this Tamil liquid patterned with the English “l” that shared place of articulation, rather than the English “r”, despite its similarly low F3 offset. Thus, when F3 and place of articulation of the precursor are dissociated, listeners appear to compensate according to place of articulation – even in a direction opposite to that predicted by sensory contrast.

In summary, compensation for coarticulation occurs both in the direction predicted by spectral contrast and in the opposite direction, and it occurs in the absence of spectrally contrastive relations. Moreover, even when the direction of compensation is in the direction predicted by spectral contrast as is the case of Experiment 1, the region assumed to provide conditions for contrast is insufficient to produce compensation. Given these dissociations, the general claim that spectral contrast underlies compensation for coarticulation appears unwarranted.

Acknowledgments

This research was supported by NSF grant 0642300 to JSM, CAF, and NV, NIH grant DC00565 to JSM, and NIH grant HD01994 to Haskins Laboratories. We thank Mark Pitt, Steve Goldinger, and an anonymous reviewer for valuable comments.

References

Boersma P. Praat, a system for doing phonetics by computer. Glot International. 2001; 5:341–345. 9/10.

Fowler CA. Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. Journal of the Acoustical Society of America. 1990; 88(3):1236–1249. [PubMed: 2229661]

Fowler CA. Compensation for coarticulation reflects gesture perception, not spectral contrast. Perception & Psychophysics. 2006; 68:161–177. [PubMed: 16773890]

Fowler, CA.; Smith, MR. Speech Perception as “Vector Analysis”: An Approach to the Problems of Invariance and Segmentation. In: Perkell, JS.; Klatts, DH., editors. Invariance and variability in speech processes. Erlbaum; Hillsdale, NJ: 1986. p. 123-139.

Fowler CA, Brown J, Mann V. Contrast effects do not underlie effects of preceding liquid consonants on stop identification in humans. Journal of Experimental Psychology: Human Perception & Performance. 2000; 26:877–888. [PubMed: 10883999]

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

Holt LL, Lotto AJ, Kluender KR. Neighboring spectral content influences vowel identification. Journal of the Acoustical Society of America. 2000; 108:710–722. [PubMed: 10955638]

Holt LL, Stephens JD, Lotto AJ. A critical evaluation of visually-moderated phonetic context effects. Perception & Psychophysics. 2005; 67:1102–1112. [PubMed: 16396017]

Lotto AJ, Holt LL. Putting phonetic context effects into context: A commentary on Fowler (2006). Perception & Psychophysics. 2006; 68:178–183. [PubMed: 16773891]

Lotto AJ, Kluender KR. General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. Perception & Psychophysics. 1998; 60:602–619. [PubMed: 9628993]

Mann VA. Influence of preceding liquid on stop-consonant perception. Perception & Psychophysics. 1980; 28:407–412. [PubMed: 7208250]

Mann VA, Repp BH. Influence of preceding fricative on stop consonant perception. Journal of the Acoustical Society of America. 1981; 69:548–558. [PubMed: 7462477]

Mitterer H. On the causes of compensation for coarticulation: Evidence for phonological mediation. Perception & Psychophysics. 2006; 68:1227–1240. [PubMed: 17355045]

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

Stephens JDW, Holt LL. Preceding phonetic context affects perception of nonspeech. Journal of the Acoustical Society of America. 2003; 114:3036–3039. [PubMed: 14714784]

Viswanathan N, Magnuson JS, Fowler CA. Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. (under review).

Wade T, Holt LL. Effects of later-occurring non-linguistic sounds on speech categorization. Journal of the Acoustical Society of America. 2005; 118:1701–1710. [PubMed: 16240828]

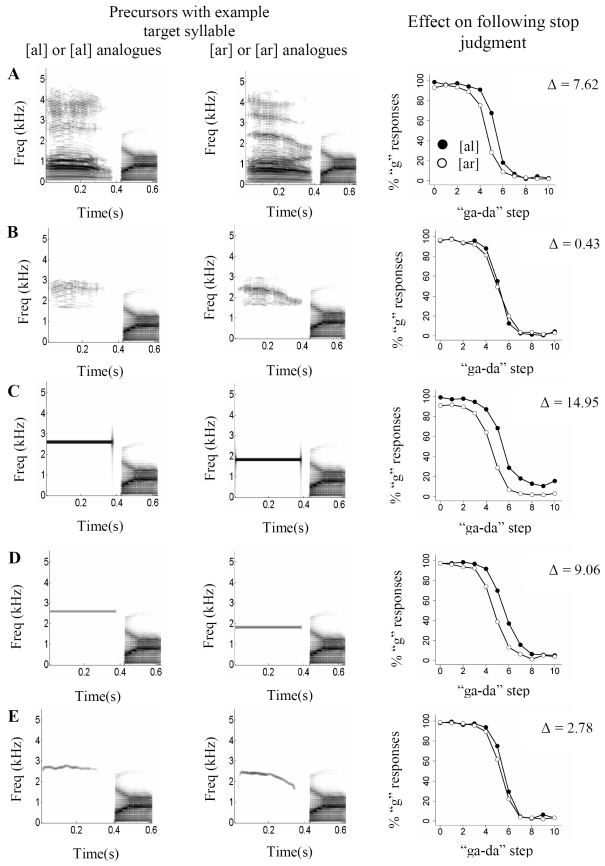

Figure 1.

Spectrograms of representative precursors before sample target syllable in each condition. Precursors and results of Experiment 1 are depicted in panels A & B and of Experiment 2 in panels C~ E. Filled circles indicate responses to [al] or [al] analogues and open circles indicate responses to [ar] or [ar] analogues. Δ indicates the average percentage shift in categorization caused by the precursors averaged over participants and step.

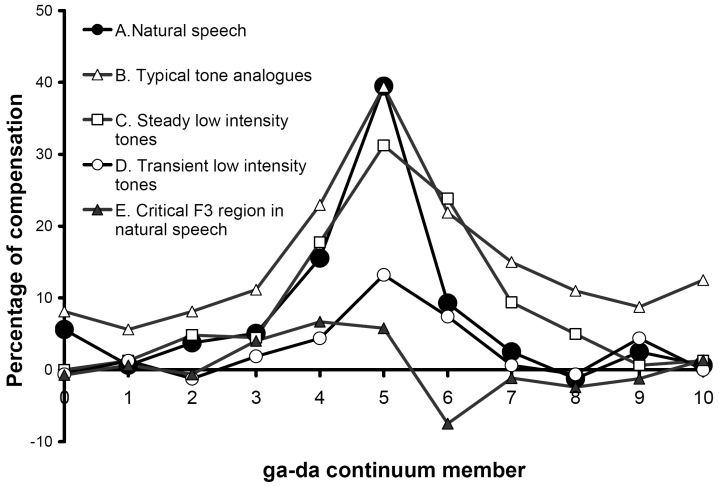

Figure 2.

Across experiment comparison of the amount of compensation as a function of relatedness of tone analogues to speech. As tones are incrementally matched to conditions in speech, contrast effects diminish.