Disentangling gestural and auditory contrast accounts of compensation for coarticulation

Navin Viswanathan, James S. Magnuson, Carol A. Fowler

University of Connecticut, Storrs, CT, 06269-1020, U.S.A
Haskins Laboratories, New Haven, CT, 06511, U.S.A
navin.viswanathan@uconn.edu, james.magnuson@uconn.edu, carol.fowler@haskins.yale.edu

ABSTRACT

Compensation for coarticulation (CfC), a context effect in which the articulatory characteristics of one segment influence the perception of a neighboring segment [1], has been a matter of considerable debate between proponents of gestural [2] and auditory theories of speech perception [3]. We set out to distinguish the two accounts by using non-native liquids (Tamil with American English listeners) that have distinct articulatory and acoustic characteristics from the native phoneme categories to which they are assimilated. We report three experiments that show that the auditory contrast account of CfC cannot explain compensatory effects with our non-native stimuli. We argue that these context effects reflect perceptual compensation for coarticulation, as predicted on a gestural account, but discuss problems for both theories.

1. INTRODUCTION

Mann [1] reported that classification of members of a [da-ga] continuum shifts toward more [ga] responses following the syllable [al] and more [da] responses following [ar]. The typical explanation for this finding is that when speakers must transition from relatively front ([l]) to back ([g]) places of articulation, they will be unlikely to reach the canonical place of articulation for [g]. Thus, after [l], [g] is likely to be produced farther forward than usual due to coarticulation. After [r], with a back place of articulation, [d] is likely to be produced farther back than usual.

CfC has been used to argue for gestural theories, and Fowler's *direct perception* theory in particular [2, 4]. On this view, hypothesized lawful relations between events that generate speech (vocal gestures) and perceptual categories provide an invariant basis for speech perception. Perceiving the gestural causes that underlie speech is crucial, as their effects are mixed nonlinearly in the acoustic features of speech signals *as analyzed by speech researchers* (though the lawful relations must be present in the signal as experienced by the listener), making the acoustic parameters analyzed by speech researchers an insufficient (under-determined) basis for phonetic perception (but see [5] for an alternative view that embraces non-determinism). From the view of direct perception, CfC follows logically from a system attuned to the gestures of speech production. An intermediary place of articulation is perceived from tokens at the middle of the [da-ga] continuum, and the system attributes this place of articulation to coarticulation in a context-dependent fashion.

Lotto and Kluender [3] have proposed a much simpler explanation for CfC, which they call *auditory contrast*. On this view, CfC phenomena follow from low-level sensory effects. The idea is that just as one would call a pail of lukewarm water cold or hot depending on the temperature of water sampled before, listeners' responses depend on contrasts between the preceding and following segments in CfC studies. In the particular example of a [da-ga] continuum following [al] or [ar], the crucial contrast is in the frequency of the third formant of the preceding syllable. Lotto and Kluender showed that when the precursor syllables [al] and [ar] were replaced by a steady high tone at the F3 frequency offset of [al], or a steady low tone at the F3 frequency offset of [ar], listeners exhibited a pattern of responses identical to CfC, suggesting all that is required to account for CfC is indeed that isolated frequency contrast.

Proponents of auditory contrast and direct perception continue to debate the basis of CfC. Three results in particular have been claimed to be problematic for coarticulatory accounts in general and gestural accounts in particular. First, Japanese quail exhibit the same CfC effect as human listeners [6]. Second, the converse effect is also found, with speech influencing the perception of following nonspeech [7]. Third, CfC is found when the context and target syllables are produced by speakers of different genders [3], which was argued to be incompatible with an articulatory basis for CfC, on the basis that coarticulatory influences should not be able to carry over between talkers.

From the gestural perspective, none of these is strong evidence. Converging results with quail and humans do not prove identical bases for perception (and one could interpret the result as demonstrating that quail and humans are similarly sensitive to lawful relations between physical events and their acoustic consequences). The arguments that stem from the other two results apparently assume the gestural theory cannot account for perception of impossible stimuli such as the tone-syllable sequences presented in the laboratory. Whereas the full acoustic basis for the gestural account has not yet been discovered, it does not predict a rigid perceptual system incapable of experiencing illusions when presented with non-articulatory information. Rather, the system is expected to be greedy, and to strive constantly to provide the best "explanation" for information that appears to have a physical cause.

There are also several problematic results for the auditory contrast account. For instance, CfC has been demonstrated in the visual modality [2] (however, see [8]). Moreover, due to the acoustic consequences of preceding syllables on stop consonants, under some conditions the obtained response pattern is the opposite of that predicted by auditory contrast [4]. Furthermore, contrast effects are not always found when the precursors and targets both consist of isolated tones [2].

Our goal is to find a basis for disentangling the gestural and contrast accounts. Our starting point is a pair of Tamil liquids: a trilled "r" that has a frontal, alveolar place of articulation (phonetic symbol [r] henceforth denoted by [R] to prevent confusion with American English (AE) "r"; we will follow the American convention of transcribing AE "r" as [r], rather than with the correct IPA symbol, [ɹ]) and a retroflex liquid with some similarity to AE [l] (phonetic symbol [ɭ] henceforth denoted by [L] to prevent confusion with AE [l]). The Tamil liquids, unlike their English counterparts, have the necessary acoustic and articulatory properties that elicit different predictions from the two theories. Moreover it is interesting to see if the predictions of the two accounts generalize to previously untested contexts.

AE listeners hear both Tamil phones as "r" (see Experiment 1, below), although Tamil listeners group [R] with American English [r] and [L] with American English [l]. Most interestingly, the place of articulation of [R] is similar to that of [l], while that of [L] is similar to [r]. However, F3 in both [R] and [L] is close to that of [r] (see Table 1). Thus, these phones allow a partial disentangling of place of articulation, F3, and eventual percept as bases for CfC. We now turn to three experiments that test the predictions of the various accounts.

2. EXPERIMENT 1

This experiment was designed to replicate previous effects of [al] and [ar] contexts on a following [da-ga] continuum, and to examine two new cases with interesting articulatory and acoustic properties, the Tamil liquids [R] (same place as [l], same F3 as [r], perceived as "r" by American listeners) and [L] (same place as [r], same F3 as [r], heard as an odd "r" by American listeners). If compensation is actually a sensory effect (specifically, the contrast of the liquid and following stop F3s), as proposed under auditory contrast accounts [3], the Tamil liquids should both pattern with [r] (i.e., with more "d" judgments compared to the [l] context, due to their relatively low third formants). The gestural account predicts that place of articulation should determine compensation. Thus, Tamil [R] should pattern with [l], and Tamil [L] should pattern with [r].

2.1 Method

Participants. 13 University of Connecticut undergraduates, who reported normal hearing, participated for course credit.

Materials. An 11-step series of resynthesized CV syllables varying in F2 and F3-onset frequency and varying perceptually from [ga] to [da] was created using the source-filter method with the Praat software package [9].

For this continuum, F3-onset frequencies varied linearly from 2200 Hz ([ga]) to 2390 Hz ([da]). The F2-onset frequencies varied from 2000 Hz ([ga]) to 1400 Hz ([da]) in steps of 150 Hz. The first and the fourth formants were the same for all members of the continuum. The initial VC syllables were produced by a 25 year old male, trilingual speaker of Indian English, Tamil, and Hindi (coached on AE liquids by a trained phonetician, who also verified that the results were native quality). Four VC syllables ([al], [aL], [ar] and [aR] were used. The utterances were combined with a gap of 80 ms between the VC and the CV syllables. The stimuli were presented at 11 kHz sampling rate and 16 bit resolution.

Table 1: Formant offset frequencies and place of articulation for the English and Tamil liquids.

	Formant				Place of
	F1	F2	F3	F4	articulation
[al]	536	1050	2637	3598	Front
[ar]	492	1465	1818	3016	Back
[aR]	521	1448	1946	3591	Front
[aL]	411	1686	1935	3146	Back

Procedure. The task was a two-alternative forced-choice: participants pressed "d" or "g" to indicate their perception of the stop. The session consisted of three blocks. In the first, the [da-ga] endpoints were presented 9 times each without a preceding liquid, in random order, with feedback. This familiarized participants with the task and syllables, and provided a basis for ensuring they could perceive the endpoints.

In the second block, all items from the 11-step [da-ga] continuum were presented in liquid contexts without feedback. Following the procedure used in [10], the stop continuum items were presented in ratios of 1-1-2-2-3-3-3-2-2-1-1, such that mid-points in the continuum were presented more often. This provided more responses for the ambiguous steps where the strongest shift is expected. To understand the design, think of the stop series as a set of 21 items (the sum of the ratios). Each of those was presented 8 times with each of the 4 liquid syllables ([al], [ar], [aL], and [aR]), such that there were 168 trials. The order of the entire set of 168 trials was randomized, and participants could take breaks after every 42 trials.

In the final block, participants heard precursor syllables in isolation and classified them as "l" or "r". Each precursor was presented 4 times, and the set of 16 was randomized. The experiment lasted about 25 minutes.

2.2 Results

Following [3], one participant, with accuracy less than 80% in the stop endpoint task, was excluded Figure 1 shows the results of the second block. Consistent with a gestural account, the results pattern according to place of articulation, with more "g" responses following front place of articulation contexts ([al] and [aR]) than back place of articulation contexts ([ar] and [aL]). The auditory contrast predictions (that [aR] and [aL] should pattern with [ar] and all three should differ from [al]) are not observed.

We used a 4 x 11 (precursor syllable X step) within-subjects ANOVA to evaluate percentage of "g" responses. There were significant effects of precursor (F (3, 30) = 12.87, p <.001) and step (F (10, 100) = 138.29, p <.001), and a reliable interaction (F (30, 300) = 1.54, p =.038). We investigated the interaction with three planned contrasts designed to test

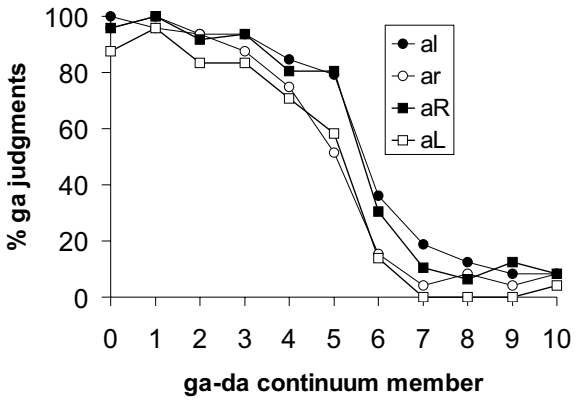

Figure 1: Results of Experiment 1. Closed symbols indicate frontals ([al] and [aR]), and open symbols indicate items with back place of articulation [ar] and [aL]). Circles indicate English phonemes and squares indicate Tamil phonemes.

predictions of gestural and contrast theories. First, we tested the gestural prediction that the basis for compensation is place of articulation by comparing segments with front place of articulation ([al] and [aR]) and those with back place of articulation ([ar] and [aL]). The effect of place of articulation was reliable (F (1, 10) = 27.48, p <.001).

Next, we tested two comparisons in which auditory contrast and gestural predictions conflict. First, we compared [al] and [aR]. Because these have the same place of articulation, a gestural account predicts no difference. Since they differ in F3 by as much as [r] and [l], a contrast account predicts a reliable difference. The contrast is not significant (F (1, 10) = 1.274, p = .285). Second, we compared [aR] and [aL]. These differ in place of articulation, so a gestural account predicts a reliable difference. They are approximately matched in F3, so auditory contrast predicts a null result. The comparison was significant (F (1, 10) = 11.48 p = .007). In the isolated liquid identification task, there was unanimous agreement: both [aL] and [aR] were classified as "r."

The pattern of results is precisely that predicted by a gestural account: curves are grouped by place of articulation. The results are problematic for auditory contrast and acoustic cue accounts: the acoustic characteristic that is expected to drive compensation effects, F3, would predict that the curve for [aR] would pattern with [aL] and [ar] rather than with [al]. However, consider F4 in Table 1. If we group the liquids by F4, we get the same groupings seen in Figure 1. Perhaps F4 provides a basis for auditory contrast. Experiments 2 and 3 explore whether F4 or some other simple acoustic characteristics might rescue the auditory contrast explanation.

3. EXPERIMENT 2

Experiment 2 uses the pure tone precursor method used in [3] to isolate F3 and F4 as potential sources of contrast. While F3 cannot explain the results of Experiment 1, we include it here to ensure that we can replicate the basic findings using this method, and to compare its effects with those of F4.

3.1 Methods

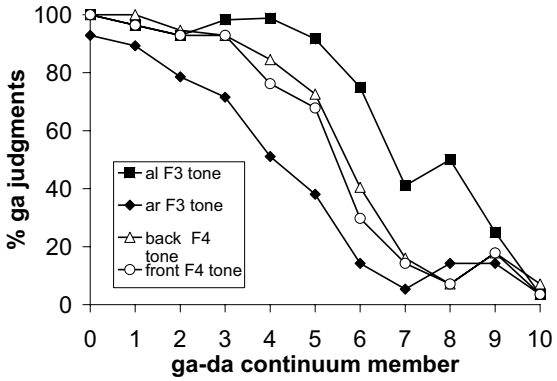

Figure 2: Experiment 2: effects of pure tone precursors.

Participants. 15 University of Connecticut undergraduates, who reported normal hearing, participated for course credit. None had participated in Experiment 1.

Materials. The [da-ga] continuum from Experiment 1 was used. Precursors were steady state sine tones synthesized at the third formant offsets of [al] and [ar], and the average fourth formant offsets of the front and back liquids used in Experiment 1. Following [3], the intensities and durations of the precursor tones were matched to the overall intensities and durations of precursor syllables used in Experiment 1.

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

3.2 Results

11 subjects out of 15 made the 80% accuracy cutoff in the stop endpoint task and were included in the analysis. Figure 2 shows the pattern of responses. The F3 tones have a strong effect, and replicate the results of [3]. We conducted separate 2 (precursor tone) x 11 (step) ANOVAs. In the case of the F3 tones, the effects of precursor tone (high vs. low) (F (1, 10) = 64.32, p < .001) and step (F (10, 100) = 81.24, p < .001) were highly significant. The effect of the precursor tone replicates [3]. In the case of F4 tones, only the effect of step was significant (F (10, 100) = 113.62, p < .001). The effect of precursor tone was nearly significant (F (1, 10) = 3.79, p < .08), but the trend was in the wrong direction for a contrast account: the F4 associated with back place of articulation led to *more,* rather than fewer, "g" responses.

Thus, neither the offset frequency of F3 nor that of F4 can provide an auditory contrast explanation of the results with the Tamil phones used in Experiment 1. Experiment 3 examines whether a more complex cue might suffice.

4. EXPERIMENT 3

Offset frequencies of F4 for the four liquids (Table 1) correlate with the results of Experiment 1 (Figure 1). In Experiment 2, we used the pure tone precursor method developed by proponents of auditory contrast [3]. We successfully replicated effects of F3 tone analogs for American English liquids. However, precursor tones matched to F4 had no reliable effect, and sensory contrast of F4 cannot explain the compensatory effects of Tamil liquids in Experiment 1. Perhaps a more complex form of contrast is required. This experiment tested

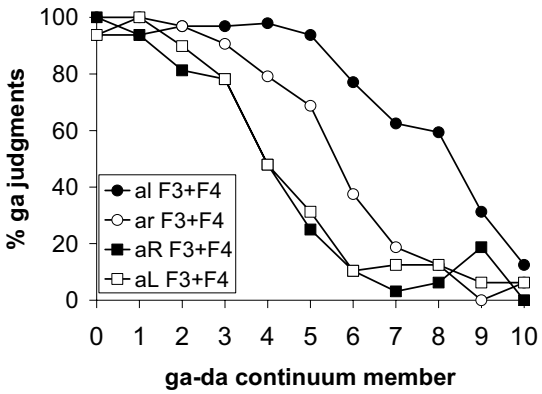

Figure 3: Effect of complex tones that mirror the third and fourth formant offsets of precursor syllables from Experiment 1.

whether presenting the F3 and F4 tones simultaneously might replicate the Tamil context effects of Experiment 1.

4.1 Methods

Participants. 11 University of Connecticut undergraduates who reported normal hearing participated for course credit. None had participated in Experiment 1 or Experiment 2.

Materials. For each syllable, a combination of precursor tones at the third and fourth formant offset frequencies was synthesized. The tones were matched for intensity and duration of the precursor syllables used in Experiment 1.

Procedure. The procedure of Experiment 2 was used.

4.2 Results

8 of 11 participants made the 80% accuracy cutoff in the stop endpoint task. Figure 3 shows the results. The relative ordering of [al] and [ar] was the same as in Experiment 1: there were more "g" judgments following [al] tone analogs than following [ar] tone analogs. This result again replicates previous findings by Lotto and Kluender [3]. However, the pattern for the Tamil tone analogs did not resemble the results found with natural liquids in Experiment 1. The expected relative Tamil ordering was not observed (more "g" responses after [aR] than after [aL] analogs), and [aR] did not pattern with [al] as it did in Experiment 1. A 4 x 11(precursor tone x step) ANOVA was used to analyze the data. Again the effect of precursor tone (F (3, 21) = 44.34, $p < .001$), step (F (10, 70) = 85.64, $p < .001$), and the interaction (F (30, 210) = 4.957, $p < .001$) were found to be highly significant. Further comparisons were not made as it is clear from Figure 3 that the pattern of results for the Tamil liquids is substantially different from that of Experiment 1.

5. DISCUSSION

Gestural and auditory contrast accounts of CfC have been difficult to distinguish because F3 (the acoustic cue that auditory contrast accounts [3] claim drives "compensation" effects) was correlated with place of articulation in previous studies that used American English [r] and [l]. The Tamil liquids, [R] and [L], provide crucial test cases in which F3 and place of articulation are disentangled. [R] has a frontal place of articulation (like [l]) but its F3 is similar to that of [r] (which has a back place of articulation). [L] has a back place of articulation and also has an F3 similar to that of [r]. In Experiment 1, we found that liquid place of articulation, rather than F3, predicts compensation effects on the following stop continuum. Experiments 2 and 3 tested whether other acoustic cues (F4 or F3 and F4 simultaneously) might provide a contrast explanation for Experiment 1. Neither cue did.

These experiments suggest that auditory contrast is insufficient to explain compensation. While it accounts for effects of [l] and [r], it does not generalize to Tamil [L] and [R]. Instead, place of articulation, consistent with a gestural account, predicts results with those phones. However, several follow up experiments must be conducted to test the adequacy of this viewpoint. One important shortcoming of this perspective currently is that even though the assertion is made in gestural accounts that information about articulation is recovered from the acoustic input and used in speech perception, the exact nature of this information has not yet been identified. We would argue that this gap does not falsify or make a gestural account implausible; however, uncovering the nature of the mapping between articulation and the consequent acoustic pattern would enhance the credibility of gestural accounts of speech perception. Follow-up studies are currently underway to identify the nature of the acoustic information that drives context effects with our Tamil liquids.

6. ACKNOWLEDGMENTS

We thank Adam Jacks and Douglas Honorof for their enthusiastic support with stimulus creation. This work was supported by NIDCD grant R01DC005765 to JSM and HD-01994 to Haskins Laboratories.

7. REFERENCES

[1] Mann, V. A., ''Influence of preceding liquid on stop-consonant perception," *Perception & Psychophysics, 28*, 407–412, 1980.

[2] Fowler, C. A., 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 and Performance, 26*, 877-888, 2000.

[3] Lotto, A. J., & Kluender, K. R., "General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification," *Perception & Psychophysics, 60*, 602-619, 1998.

[4] Fowler, C. A., "Compensation for coarticulation reflects gesture perception, not spectral contrast," *Perception & Psychophysics*, (in press).

[5] Nusbaum, H. C., & Magnuson, J. S., "Talker normalization: Phonetic constancy as a cognitive process," In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 109-132), San Diego: Academic Press, 1997.

[6] Lotto, A. J., Kluender, K. R., & Holt, L. L., "Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*)," *Journal of the Acoustical Society of America, 102*, 1134-1140, 1997.

[7] Stephens, J., & Holt, L., "Preceding phonetic context affects perception of nonspeech," *Journal of the Acoustical Society of America, 114*, 3036-3039, 2003.

[8] Holt, L. L., Stephens, J. & Lotto, A. J.,"A critical evaluation of visually moderated phonetic context effects," *Perception & Psychophysics* (in press).

[9] Boersma, P., & Weenink, D., "Praat: doing phonetics by computer" *(Version 4.4.16) [Computer program]. Retrieved April 1, 2006, from http://www.praat.org/,* 2006.

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