

General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification

ANDREW J. LOTTO and KEITH R. KLUENDER
University of Wisconsin, Madison, Wisconsin

When members of a series of synthesized stop consonants varying acoustically in F_3 characteristics and varying perceptually from /da/ to /ga/ are preceded by /al/, subjects report hearing more /ga/ syllables relative to when each member is preceded by /ar/ (Mann, 1980). It has been suggested that this result demonstrates the existence of a mechanism that compensates for coarticulation via tacit knowledge of articulatory dynamics and constraints, or through perceptual recovery of vocal-tract dynamics. The present study was designed to assess the degree to which these perceptual effects are specific to qualities of human articulatory sources. In three experiments, series of consonant-vowel (CV) stimuli varying in F_3 -onset frequency (/da/-/ga/) were preceded by speech versions or nonspeech analogues of /al/ and /ar/. The effect of liquid identity on stop consonant labeling remained when the preceding VC was produced by a female speaker and the CV syllable was modeled after a male speaker's productions. Labeling boundaries also shifted when the CV was preceded by a sine wave glide modeled after F_3 characteristics of /al/ and /ar/. Identifications shifted even when the preceding sine wave was of constant frequency equal to the offset frequency of F_3 from a natural production. These results suggest an explanation in terms of general auditory processes as opposed to recovery of or knowledge of specific articulatory dynamics.

Despite 40 years of sustained effort to develop machine speech-recognition devices, no engineering approach to speech perception has achieved the success of an average 2-year-old human. One of the more daunting aspects of speech for these efforts is the acoustic effects of coarticulation. Traditionally, *coarticulation* refers to the spatial and temporal overlap of adjacent articulatory activities. This is reflected in the acoustic signal by severe context dependence; acoustic information specifying one phoneme varies substantially, depending on surrounding phonemes. As a result, there is a lack of invariance between linguistic units (e.g., phonemes, morphemes) and attributes of the acoustic signal. This poses quite a problem for speech-recognition devices which are designed to output strings of phonemes.¹

An example of coarticulatory influence is the effect of a preceding liquid on the acoustic realization of a subsequent stop consonant. Mann (1980) reports that articulation of the syllables /da/ and /ga/ may be influenced by the production of a preceding /al/ or /ar/. Articulatorily described, the physical realization of the phonemes /d/

and /g/ primarily differ in the place at which the tongue occludes the vocal tract. For a velar stop [g], the tongue body is raised against the soft palate at the rear of the mouth, whereas for an alveolar stop [d], the tongue tip comes in contact with the alveolar ridge toward the front of the oral cavity behind the teeth. The liquids /l/ and /r/ differ in a similar manner; an [r] is produced with the tongue raised toward the rear of the cavity, and an [l] is produced with the tongue tip nearer the front of the mouth. Mann (1980) suggests that productions following [l] will be articulated at a more anterior position than productions following articulation of [r] because of the assimilatory nature of coarticulation. Therefore, a [ga] production after [al] would be articulated at a more anterior position in the direction of the alveolar ridge and, thus, be produced at a more [da]-like place of articulation. The same reasoning holds, *mutatis mutandis*, for a [da] production following [ar], which will be produced with a more posterior ([ga]-like) place of articulation. Thus, coarticulatory effects of preceding liquids can lead to [da] and [ga] syllables that are similar articulatorily.

Of course, this assimilation of articulation affects the spectral characteristics of the subsequent consonant-vowel (CV) syllables. Figure 1 shows schematized spectrograms of four VC CV productions based loosely on the mean formant frequency values measured from one male speaker described in Mann (1980). As the [al da] and [ar ga] spectrograms depict (Figures 1A and 1D, respectively), one of the primary acoustic differences between the syllables [da] and [ga] is the onset frequency of the third formant (F_3). The anterior place of articulation for the alveolar stop [d]

This work was supported by NIDCD Grant DC-00719 and NSF Young Investigator Award DBS-9258482 to the second author and by Sigma Xi Dissertation Research Award to the first author. Some of the data were presented at the spring meeting of the Acoustical Society of America, May 1994, in Cambridge, MA, and at the International Congress on Phonetic Sciences, August 1995, in Stockholm. The authors would like to thank Lori Holt for assisting in the preparation of this manuscript. Requests for reprints or correspondences should be addressed to A. J. Lotto, Department of Psychology, 1202 W. Johnson St., Madison, WI 53706 (e-mail: ajlotto@facstaff.wisc.edu).

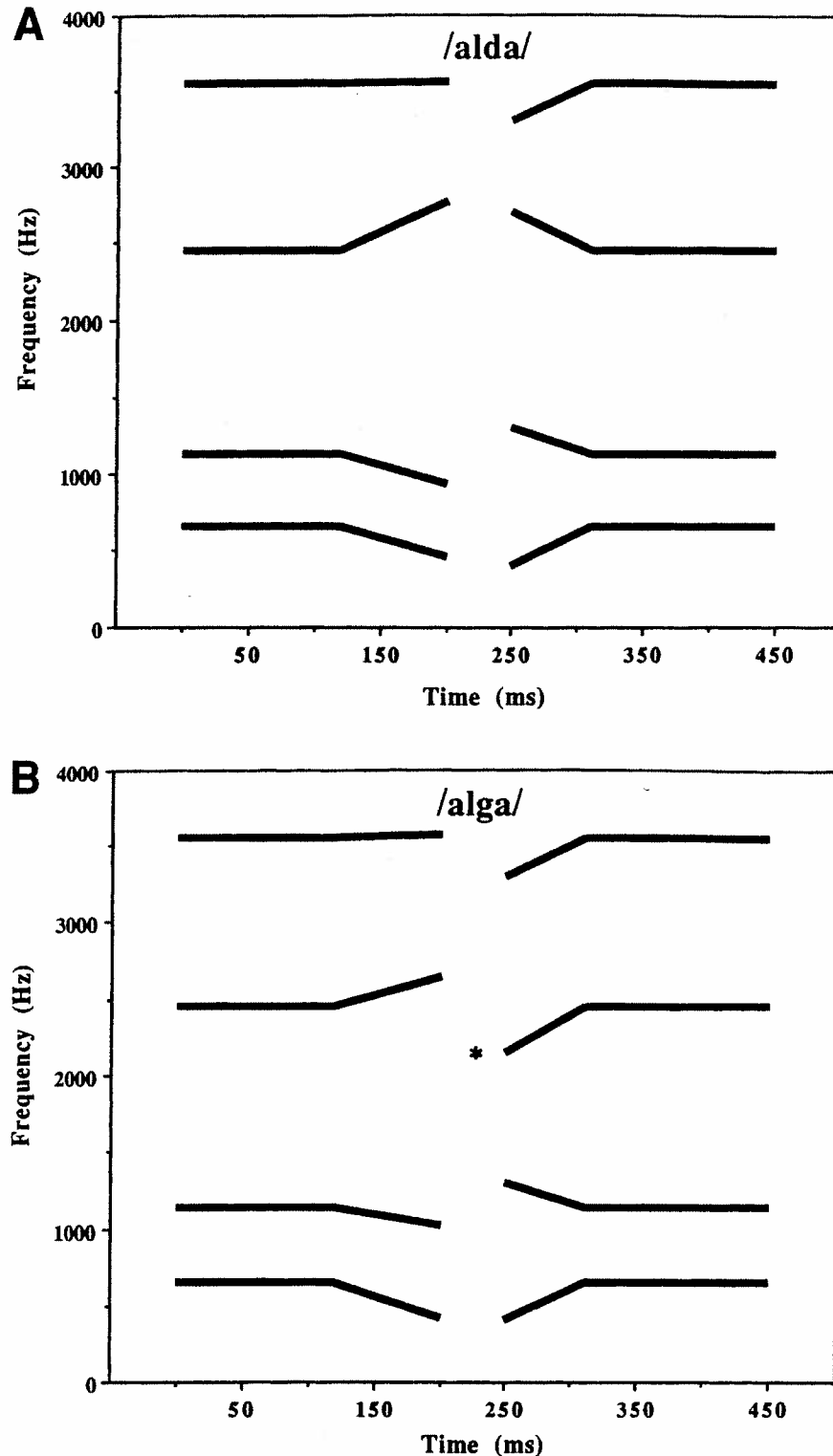


Figure 1. Schematic spectrograms of trajectories of the first four formants for VC CV productions. Based loosely on data presented in Mann (1980; Figure 4 and Table 2). The asterisks in panels B and C (next page) mark the third-formant trajectory for the CVs in the two contexts that result in nearly identical CV acoustics (/al ga/ and /ar da/). (A) /al da/. (B) /al ga/.

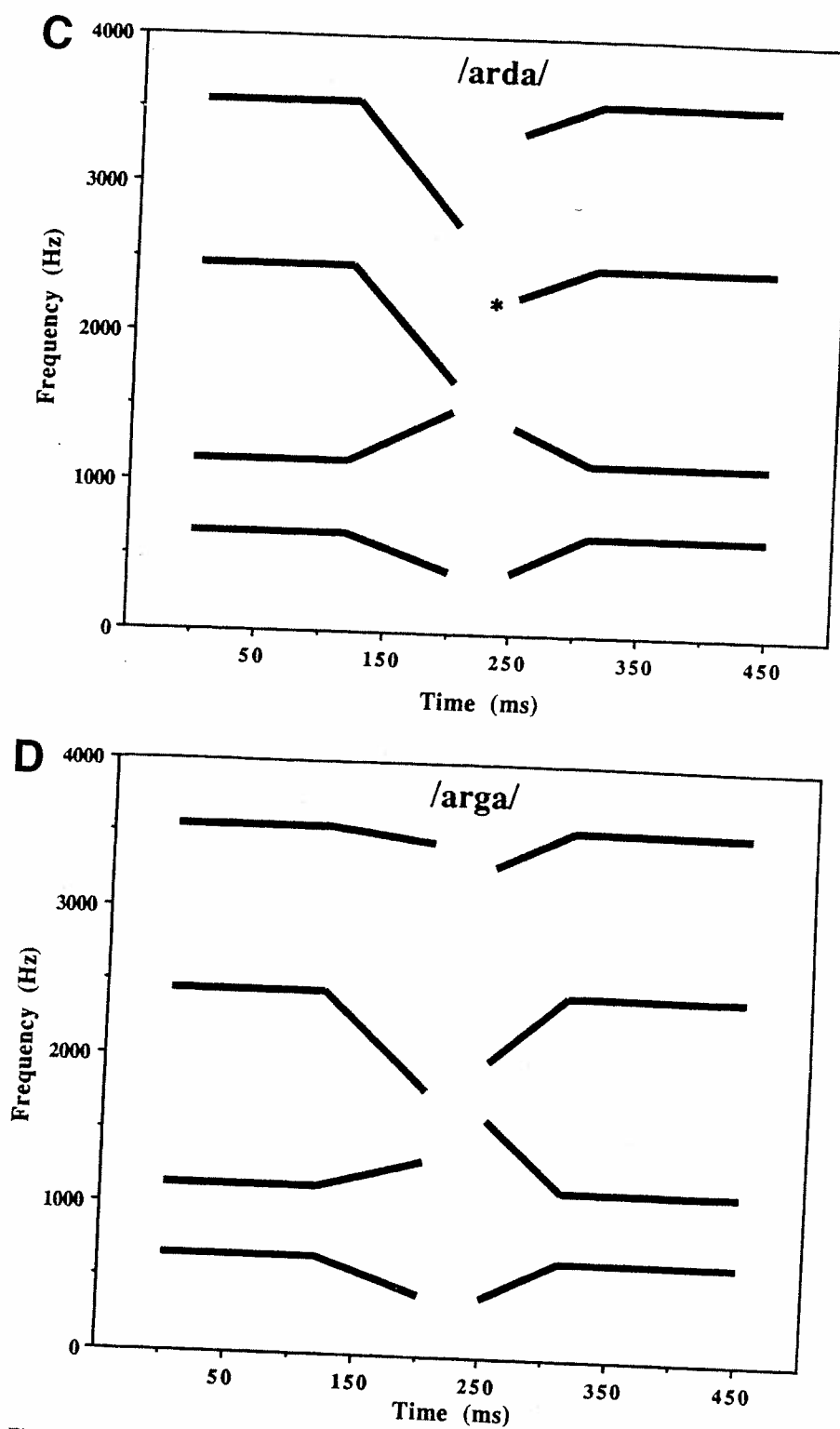


Figure 1 (cont'd). Schematic spectrograms of trajectories of the first four formants for VC CV productions. Based loosely on data presented in Mann (1980; Figure 4 and Table 2). The asterisks in panel B (previous page) and panel C mark the third-formant trajectory for the CVs in the two contexts that result in nearly identical CV acoustics (/al ga/ and /ar da/). (C) /ar da/. (D) /ar ga/.

results in a higher $F3$ onset than does the posterior occlusion of the velar stop [g]. However, this spectral distinction diminishes when one compares the onset frequency of $F3$ for the second syllable in the cases [al ga] and [ar da] (Figures 1B and 1C, respectively). In fact, these two syllables are almost spectrally identical (see also Dainora, Hemphill, Hirata, & Olson, 1996). This is due to the assimilation of place of articulation, which is generated by the liquid context. Speech recognition models that employ pattern recognition on a phonemic or syllabic base would not be able to distinguish these productions that result from instructions to produce /da/ or /ga/. For example, the linear logistic models of Nearey (1990, 1992, 1995), which label syllables on the basis of weightings of acoustic attributes without interactions at levels above the single segment, would have difficulties distinguishing these utterances. Nearey's models provide admirable simplicity and generality by relying solely on acoustic attributes for identification. However, these models cannot accommodate effects of coarticulation that span one or several syllables without losing much of this simplicity.² Yet this failing is of little consequence if human listeners also find these syllables to be indistinguishable in context, since, presumably, the goal of machine speech recognition and theoretical models is to obtain human performance.

Mann (1980) tested the effects of liquid context on the identification of stop consonants for human adult listen-

ers. A seven-step series of synthesized CV syllables varying in onset characteristics of $F3$ and varying perceptually from /da/ to /ga/ was presented to subjects for identification. Each CV syllable was preceded by a naturally produced token of the syllable /al/ or /ar/. The CV series was modeled after productions by the same male who had produced the /al/ and /ar/ tokens. The data from one of the conditions from Mann (1980) (the stressed-VC condition) are replotted in Figure 2. Interestingly, subjects were more likely to label a syllable as /da/ when it was preceded by /ar/ and as /ga/ when it was preceded by /al/. That is, perception appeared to compensate for the assimilatory effects of coarticulation. Syllables produced after [al] are acoustically more [da]-like, but they were identified more often as /ga/. Human adults can apparently "correctly" label the second syllables in Figures 1B and 1C. Mann (1980) concluded that

The fact that listeners are able to make correct use of such influences as cues to stop perception attests to the view that speech perception must somehow operate with tacit reference to the dynamics of speech production and its acoustic consequences. (p. 411)

This perceptual compensation for coarticulation was also demonstrated in listeners whose native language did not contain the /l/-/r/ contrast. Japanese speakers who could not distinguish the syllables [al] and [ar] in a discrimination task identified a CV more often as /ga/ following /al/

Data from Mann (1980)

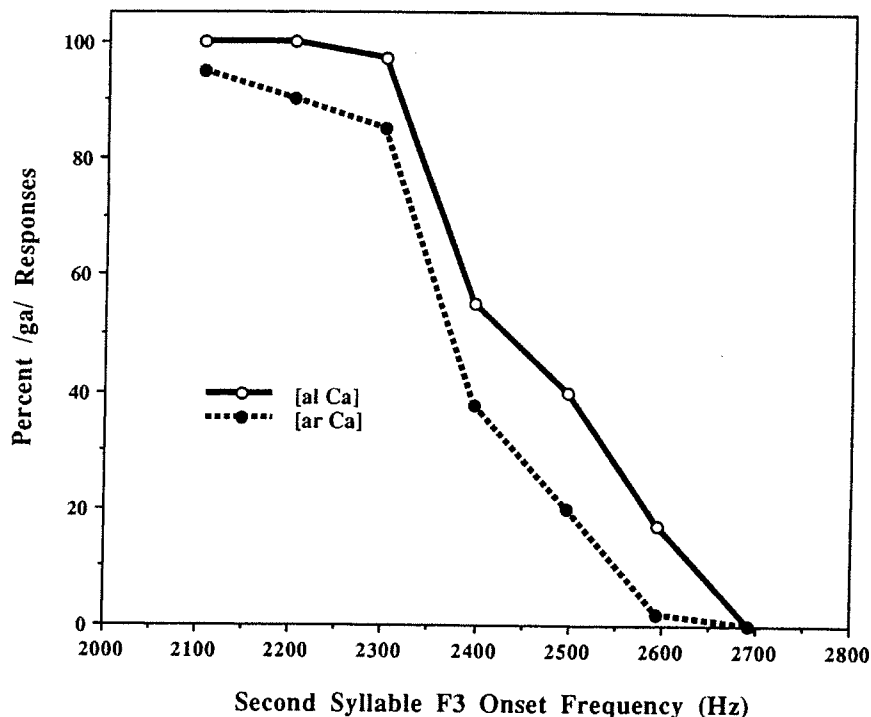


Figure 2. Data from Mann's (1980) VC-stressed conditions are replotted as percent of /ga/ responses.

and as /da/ more often following /ar/ (Mann, 1986). These results suggest that the perceived identity of the preceding liquid is not important to the context effect. Mann (1986) concluded that

Preceding a language-specific level of perception where speech sounds are represented in accordance with the constraints of a given phonological system, there may exist a universally-shared level where the representation of speech sounds more closely corresponds to the articulatory gestures that give rise to the speech signal. (pp. 169-170)

Subsequently, Fowler, Best, and McRoberts (1990) demonstrated the effect of a preceding liquids on the discrimination of CV syllables by 4-month-old infants. Again, these results were adduced as evidence that human speech perception is constrained by particulars of human vocal tracts.

The results suggest that pre-linguistic infants disentangle consonant-consonant coarticulatory influences in speech in an adult-like fashion. . . . we conclude that perception of coarticulated speech by infants indexes their recovery of talkers' linguistically significant vocal tract actions. (Fowler et al., 1990, pp. 559, 568)

These earlier research efforts would seem to direct those who are involved in automatic speech recognition to include constraints particular to vocal tracts in their algorithms. This could be achieved by building in a representation of vocal-tract dynamics and constraints as discussed by Mann (1980, 1986), or one could attempt to recover vocal-tract dynamics from acoustic information in an approach similar to the *direct realism* of Fowler (1986).

At least partially inspired by direct realism and by early analysis-by-synthesis approaches (Stevens, 1960), much effort has been made of late to recover articulatory gestures from the physical acoustic wave form. In general, the history of these efforts can be summarized in the following manner (for reviews, see McGowan, 1994; Schroeter & Sondhi, 1992). Early attempts were made to use limited acoustic information such as the first three formant frequencies to derive the area function of the vocal tract. In many cases, these efforts were not successful, because multiple area functions could be specified by the same wave form. Solutions are compromised to the extent that the vocal tract producing the speech sound is not the same length as that assumed by a model such as linear predictive coding. Errors are also introduced to the extent that solutions require relatively precise approximation of glottal wave forms. Each of these problems becomes highly evident when one attempts to specify articulator activity for more than a single specific talker, particularly across talkers of different sex who differ in terms of the proportional size of the pharynx and the overall vocal tract (Fant, 1966, 1975) and differ in characteristics of the glottal wave form (Henton & Bladon, 1985; Holmberg, Hillman, & Perkell, 1988; Klatt & Klatt, 1990; Monsen & Engebretson, 1977).³

More recent efforts have been successful to the extent that they have incorporated specific constraints on the nature of the vocal tract, together with dynamic and kinematic information. For example, McGowan (1994) used a task-dynamic model (Saltzman, 1986; Saltzman & Kelso, 1987) driving six vocal-tract variables (tongue-body location and degree, tongue-tip location and degree, lip protrusion and aperture) with transformations between tract variables and articulators derived from an articulatory model (Mermelstein, 1973). McGowan and Rubin (1994) exploited a genetic algorithm to discover relations between task-dynamic parameters and speech acoustics for six utterances by a single talker. Their results were somewhat mixed, in that whereas the model got many things right, some errors persisted, and McGowan has noted that future applications will likely require customization of the model to individual talkers. Related efforts continue to be productive (see, e.g., Schroeter & Sondhi, 1992), but one point is becoming increasingly clear. The extent to which these attempts to solve the inverse problem (recovering vocal-tract shape from acoustic information) are successful seems to depend critically on models' engendering highly realistic details of sound production specific to human vocal tracts, and often to a single human vocal tract.

Given the complexity of these solutions and the effort necessary for further success, it is important to determine what information in the speech wave form may be used by human listeners in compensating for coarticulatory effects like those presented above.⁴ How close is the agreement between dynamics of the vocal tract and labeling performance of listeners? If the explanation of human-listener abilities lies in either specific knowledge of vocal-tract dynamics or in the recovery of vocal-tract gestures and their constraints, one may predict that perceptual compensation for coarticulation relies on rather specific attributes of the wave form that are directly due to constraints particular to vocal tracts. The experiments described in this paper were designed to determine the specificity of the information responsible for perceptual compensation for acoustic effects of coarticulation, utilizing labeling tasks similar to those in Mann (1980).

EXPERIMENT 1

Presumably, the kinematics of coarticulated speech are due in part to the fact that the vocal tract is a physical system constrained by mass, inertia, and degrees of freedom. These constraints act on a single vocal tract, but different vocal tracts are independent. The vocal-tract shape of one talker at time t is not physically constrained by the vocal-tract shape of another talker at time $t-1$. If articulatory constraints are recovered by or represented in the human perceptual system, presumably these constraints will only affect the perception of speech arising from a single vocal tract. So, what happens if the talker changes midway through an utterance? Coarticulatory

influences cannot carry over between talkers, and one would not expect perceptual compensation for coarticulation for speech from two talkers if compensation is accomplished through information specific to a single vocal tract. This is especially true for a change in sex of talker, given the production differences listed above. Experiment 1 was designed to test this presumption with an identification task of VC CVs.

Method

Subjects. Fifteen college-age adults, all of whom had learned English as their first language, served as listeners. All reported normal hearing. These subjects received introductory psychology course credit for their participation.

Stimuli. A 10-step series of CV syllables varying in *F3*-onset frequency and varying perceptually from /da/ to /ga/ was synthesized on the cascade synthesizer described in Klatt (1980). Endpoint stimuli were based on isolated natural productions of a male talker. For this series of CVs, the onset frequency of *F3* varied from 1800 to 2700 Hz in 100-Hz steps, changing linearly to a steady-state value of 2450 Hz over an 80-msec transition. All other synthesizer parameters remained constant across members of the series. Frequency of the first formant (*F1*) rose from 300 to 750 Hz, and the second-formant (*F2*) frequency decreased from 1650 to 1200 Hz, over 80 msec.⁵ Fundamental frequency (*f0*) was 110 Hz from onset until decreasing to 95 Hz over the last 50 msec. The total stimulus duration of the synthesized syllables was 250 msec. These stimuli were preceded by two sets of natural speech tokens of the syllables /al/ and /ar/ with a 50-msec silent interval between syllables. One set of VCs was produced by a large 6-ft 2-in. 37-year-old male, after whom the /da/-/ga/ series was modeled, with an average *f0* of 110 Hz. The other set was produced by a small 5-ft 2-in. 19-year-old female with an average *f0* of 210 Hz. Formant frequency values averaged about 12% higher for the female VCs. The frequency values for the first three formants for each speaker are given in the Appendix, and schematized spectrograms of the four preceding contexts are displayed in Figure 3. The RMS energy of each syllable was matched for presentation.

The stimuli were synthesized (/da/-/ga/) or recorded (/al/ and /ar/) with 12-bit resolution at a 10-kHz sampling rate and stored on computer disk. Stimulus presentation was under control of a microcomputer, and concatenation of syllables (with 50-msec silent gaps) occurred on-line during the experiment. Following D/A conversion (Ariel DSP-16), the stimuli were low-pass filtered (Frequency Devices 677; cutoff frequency, 4.8 kHz) prior to being amplified (Stewart HDA4) and were played over headphones (Beyer DT-100) at 75 dB SPL.

Procedure. The subjects participated in a two-response forced-choice identification task. In each experimental session, 1-3 subjects were tested concurrently in single-subject sound-attenuated booths (Suttle Equipment). Before listening to the VC CVs, the subjects were presented each CV in isolation 10 times in order to familiarize them with the synthesized syllables and to test their ability to identify the syllable-initial stop consonants. Then, the subjects heard the CV series preceded by the male natural VC productions in a separate block from the CVs preceded by the female natural VC productions. Presentation order was counterbalanced across subjects. Within each block, the 10 CV syllables were presented preceded by each of the two VC syllables (/al/ and /ar/) 10 times each, for a total of 200 disyllables. Stimulus presentation was randomized within each block. The subjects were instructed to identify the second syllable by pressing either of two buttons on a response box which were labeled "da" and "ga." Disyllables were presented approxi-

mately every 3 sec, and the experimental session lasted approximately 45 min.

Results

Data from 3 subjects were held from further analysis because those individuals failed to identify the endpoint syllables of the CV series correctly 90% of the time when they were presented in isolation. This criterion was established because some subjects have difficulty labeling synthesized speech and sometimes respond randomly. Identification functions, averaged across the remaining 12 subjects for both of the speaker gender and preceding-liquid conditions, are presented in Figure 4.⁶

A 2×10 (identity of preceding liquid \times *F3*-onset frequency of CV) within-subjects analysis of variance (ANOVA) was performed separately for the male VC and female VC conditions, with percentage of /ga/ responses serving as the dependent variable. For the CVs preceded by VCs produced by the same talker (male), there was a significant shift in reported stop-consonant identity with change in preceding liquid [$F(1,11) = 47.20, p < .0001$]. Replicating the results of Mann (1980), more /ga/ responses were made following /al/ than following /ar/. As in Mann (1980), it appeared that the subjects perceptually compensated for coarticulation despite the fact that the first syllable was a natural production and the second syllable was synthesized. In itself, this suggests that any perceived constraints of the distal sound source cannot be perfectly valid, especially since there was no actual coarticulation between the VC and CV.

Even more troublesome for models of speech perception that rely specifically on articulatory dynamics were the data obtained for CV identification in the context of a syllable produced by another talker. For the preceding female VCs, there was also a significant effect of the identity of the preceding liquid on stop identification [$F(1,11) = 70.44, p < .0001$]. The shift in identification curves was smaller than that for the male-produced VC context. A matched-pairs *t* test indicated that this difference was statistically significant (see Table 1). Nevertheless, a change in preceding phonemic context resulted in a shift in CV identification in a manner that appears to have compensated for the assimilatory effects of coarticulation. This is troublesome for any account of speech perception that relies on strict adherence to vocal-tract dynamics or constraints. There are no physical constraints that should act across the articulations of two speakers. Although there are general correspondences between vocal tracts of different speakers, presumably with some set of constraints that operate across all vocal tracts, the shape of one vocal tract is not constrained by the shape of another vocal tract preceding it in time. Yet listeners labeled these disyllables in a manner consistent with their labeling of disyllables produced by the same talker. This nonveridical perceptual integration over talker change has some precedence in speech research. In recent reports concern-

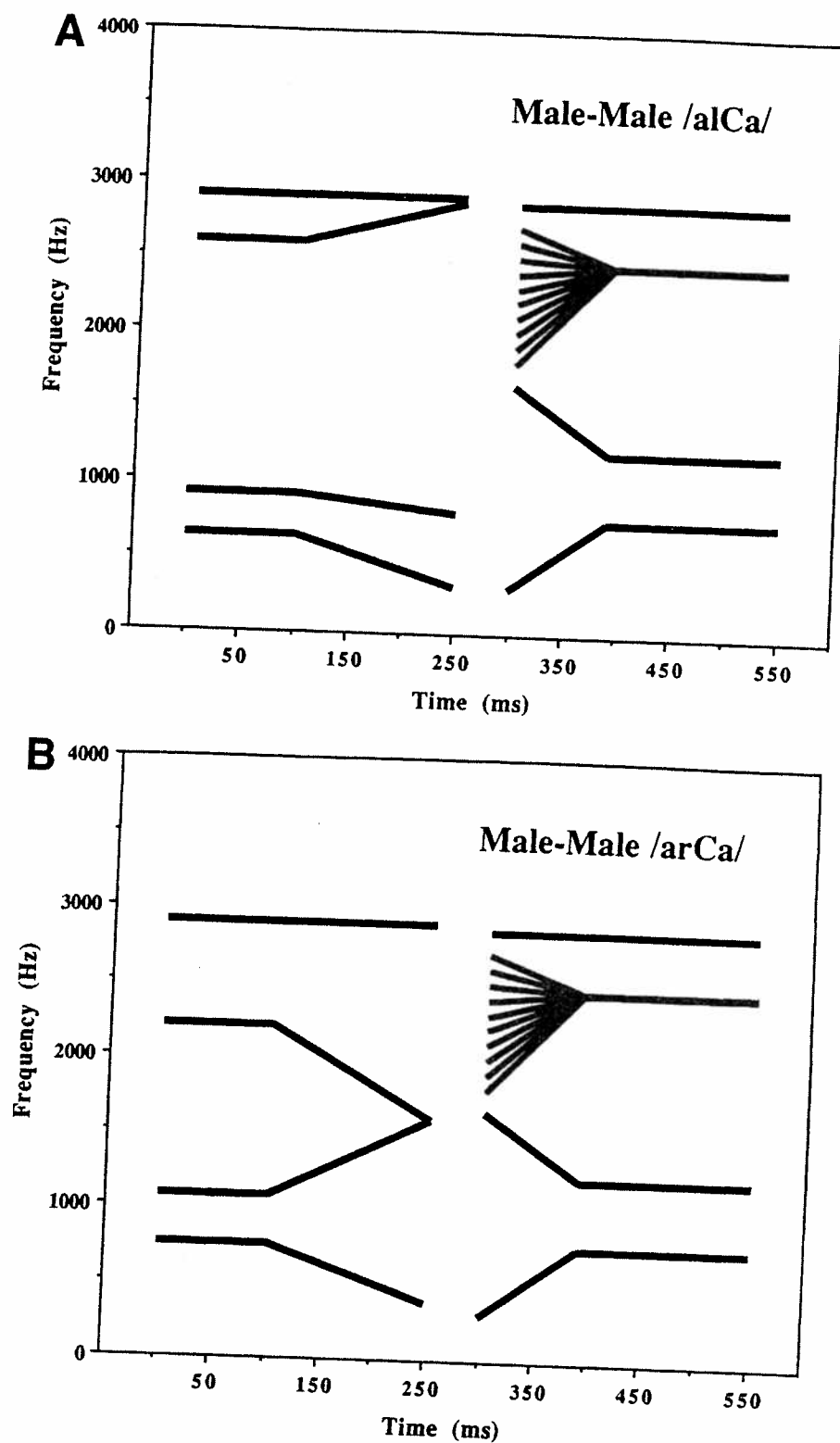


Figure 3. Schematic spectrograms of trajectories of the first four formants for all conditions presented in Experiment 1. The preceding VCs were natural productions from either a male or a female talker, and the CVs were synthesized on the basis of endpoints produced by the male talker. All 10 F_3 trajectories for the CVs are displayed in each figure. (A) Male [al] preceding /da/-/ga/ series. (B) Male [ar] preceding /da/-/ga/ series.

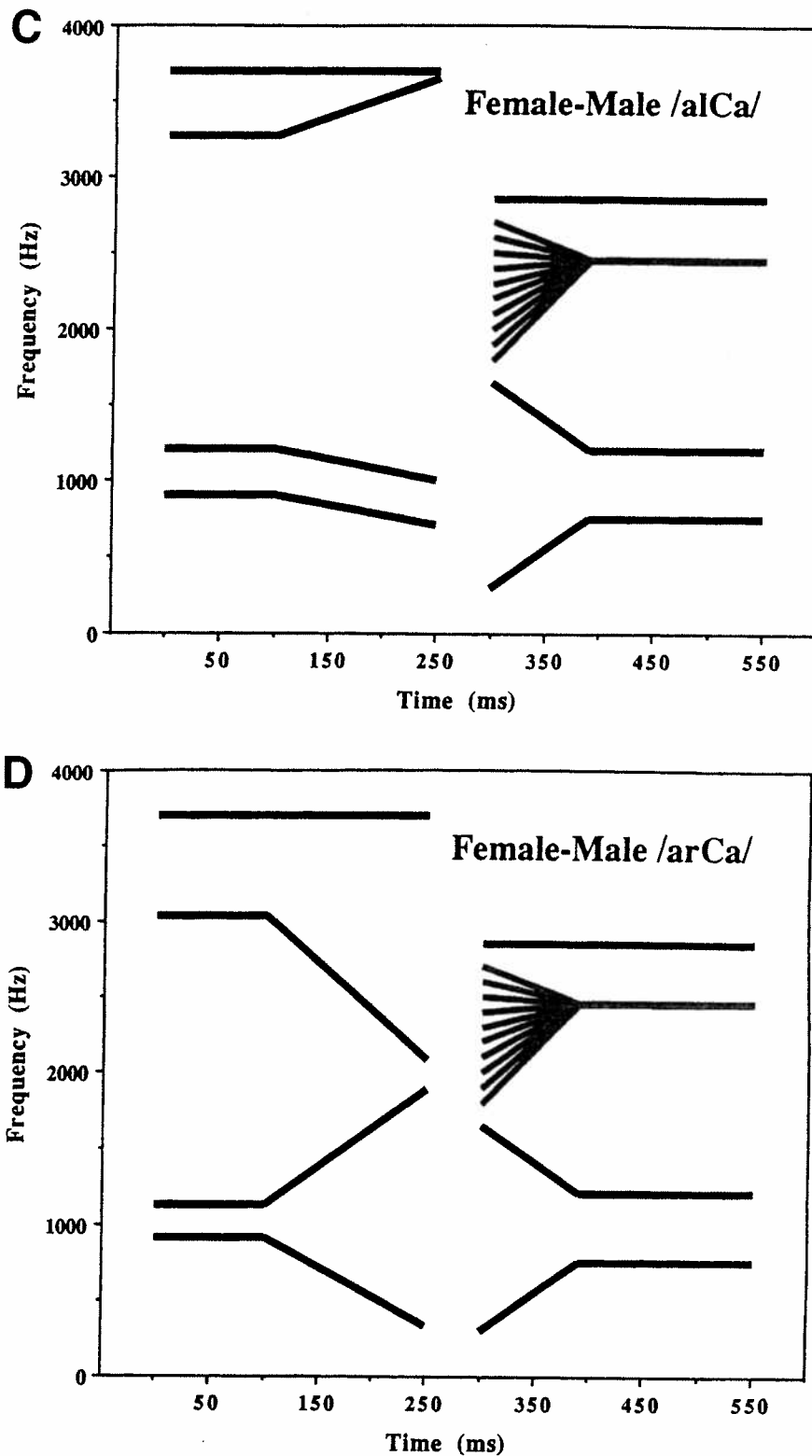


Figure 3 (cont'd). Schematic spectrograms of trajectories of the first four formants for all conditions presented in Experiment 1. The preceding VCs were natural productions from either a male or a female talker, and the CVs were synthesized on the basis of endpoints produced by the male talker. All 10 F_3 trajectories for the CVs are displayed in each figure. (C) Female [a] preceding /da/-/ga/ series. (D) Female [ar] preceding /da/-/ga/ series.

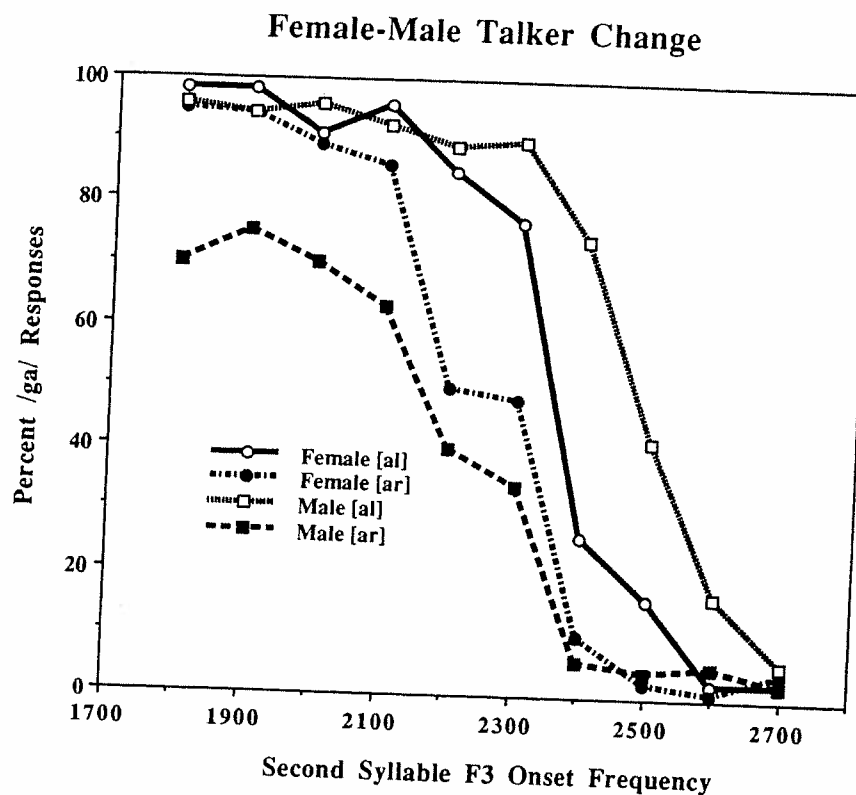


Figure 4. Mean identification functions from Experiment 1. Percent of /ga/ responses is plotted for each of the 10 CV stimuli, with preceding context serving as the parameter.

ing the integration of spectral energy for determining syllable length, as measured by the syllable length effect on stop perception, subjects treated CV syllables as wholes despite a midvowel perceived change in talker (Green, Stevens, & Kuhl, 1994; Lotto, Kluender, & Green, 1996).

It appears that the context effect of preceding /al/ and /ar/ is not critically sensitive to the entire stimulus complex being produced by a single talker or even by a modestly similar vocal tract. It is possible that putative pertinent articulatory representations are sufficiently abstract to allow the compensatory mechanism to operate across substantial acoustic change. One may wonder just how loosely tied this "compensatory" mechanism is to the speech stream.

Also, one might propose that listeners are recovering transformational invariants that are independent of the vocal tract (or tracts) producing the speech sounds. Because the speech sounds were contiguous and were presented in an otherwise quiet environment, listeners could have perceived the speech as a unified event and picked up the appropriate invariants that specified the linguistic gestures even across the change in speaker.⁷ These speech hybrids may not be sufficient to test the validity of a direct realist account of the context effect.⁸ Experiment 2 included a nonspeech analogue of an acoustic/auditory aspect of the liquids used in Experiment 1, so that we could examine further the relationship between the perceptual mechanism of compensation and acoustic characteristics

of the input and clarify the importance of the apparent source identities for the context effect.

EXPERIMENT 2

Because results from Experiment 1 indicated that precise matching of articulatory/acoustic characteristics for the initial VC and following CV was not essential for the effect on stop consonant identification, Experiment 2 was designed to test whether partial schematized nonspeech information akin to that for liquid identity could affect labeling of following CVs. It is clear from Figures 3A and 3B that among the most salient differences between the male productions of the syllables /al/ and /ar/

Table 1
Mean Difference in Percent /ga/ Responses Between /al/ and /ar/ Contexts for Each of the Conditions in the Three Experiments

	Mean /al/-/ar/ Difference % /ga/	SE
Experiment 1*		
Male	32.42	4.72
Female	11.17	1.33
Experiment 2†		
Speech	26.54	2.80
Nonspeech	6.62	1.77
Experiment 3		
Nonspeech	15.80	3.62

* $t(11) = 4.46, p = .001$. † $t(12) = 8.43, p = .000$.

are frequency characteristics of energy concentration around the $F3$ offset. It is possible that this change in relative energy across frequency could be important information for determining the identification of the following CV. To test this possibility, two frequency-modulated sine waves, created to model the $F3$ -offset characteristics of the liquids, were used as the preceding context for the /da/-/ga/ series from Experiment 1. If a shift in stop consonant identification is affected by this change, the necessity of a strict representation of or recovery of vocal-tract characteristics for perceptual coarticulation compensation would, indeed, be called into question. As a consequence of this experiment, the essential information for these context effects would be more clearly delimited.

Method

Subjects. Seventeen native-English speakers served as listeners. All reported normal hearing and received credit in an introductory psychology course for their participation.

Stimuli. The /da/-/ga/ syllables used in Experiment 1 served as target syllables for identification in Experiment 2. However, the preceding context consisted of four new sounds. For the speech context, /a/ and /ar/ syllables were synthesized on the basis of measurements of two new productions of /a/ and /ar/ by the male who produced the natural utterances from Experiment 1. Synthesized syllables were utilized in Experiment 2 to allow more precise control of formant values in order to create faithful nonspeech analogs. For these syllables, f_0 remained constant at 110 Hz throughout the 250-msec duration. Steady-state /a/ values were equivalent for both syllables for the first 100 msec, followed by 150-msec formant transitions. Nominal formant frequency values for these stimuli are presented in Table 2.

Based on these synthesized syllables, two frequency-modulated sine wave glides were created to model the $F3$ -offset characteristics of the liquids. Sine wave glides were matched in RMS energy to that present in a critical band placed at the midpoint between $F3$ and $F4$ for [a] and between $F2$ and $F3$ for [ar]. Midpoints between these formants were chosen because their offsets were very close in frequency and probably would be summated auditorially (Johnson, 1989, e.g., found perceptual integration for spectrally close $F2$ and $F3$ for vowels). The frequency of the [a]-glide increased linearly from 2523 to 2775 Hz; parallel to the $F3$ transition for [a], but with an offset midway between $F3$ and $F4$. Likewise, the [ar]-glide decreased from 2389 to 1541 Hz. Both FM sine wave glides were 150 msec long. The nonspeech conditions are represented schematically in Figure 5. All details of stimulus presentation were identical to those for Experiment 1.

Procedure. The procedure was essentially identical to that used in Experiment 1. Subjects first identified members of the /da/-/ga/

series in isolation. Then they identified the CV syllables in two separate counterbalanced blocks. In one block, the CVs were preceded by synthesized [a] and [ar] syllables. In the other block, the CVs were preceded by sine wave glides modeled after the $F3$ transitions. The subjects were told that they would be hearing computer tones in the nonspeech condition. The intervening silent interval between the CVs and the preceding context was 50 msec for all stimulus complexes.

Results

Four subjects failed to meet the 90% correct criterion. Average identification functions for the remaining 13 subjects are presented in Figure 6.

A 2×10 (preceding context $\times F3$ -onset frequency of CV) within-subjects ANOVA was performed separately for the speech and nonspeech condition. Once again, the findings of Mann (1980) were replicated. For the speech condition, there was a significant shift in percent of /ga/ identifications in accordance with a change in the identity of the preceding liquid [$F(1,12) = 89.67, p < .0001$]. Subjects identified CV syllables more often as /ga/ when they were preceded by /a/.

Although the context effect of sine wave glides was not as great as the effect for full-spectrum synthesized [a] and [ar] (see Table 1), there was a significant effect of glide type as reflected in responses to the synthesized /da/-/ga/ series [$F(1,12) = 14.02, p < .005$]. Even a preceding stimulus that consists of only a sine wave caricature of a portion ($F3$ transition) of rich full-spectrum speech is adequate to give rise to the context effect on perception of the following CV as /da/ or /ga/. The effect of FM glides may be surprising, since they sounded nothing like speech and were much less salient than the synthesized [a]-[ar] syllables. FM glides were matched to the energy within a single critical band centered on $F3$ and as a result were 12–15 dB less intense than the full /a/-/ar/ syllables and were 60% as long as the full VC syllables.

Given the caricature nature of these FM glides in comparison with the full speech stimuli, one may wonder about the specificity to speech of this "coarticulatory compensation" process. Could this be a general auditory process that happens to compensate for the dynamics of articulatory systems? If so, what might this general auditory process be? One possibility springs to mind when one redescribes the results of the previous experiments in terms of $F3$ frequency. In Mann (1980) and in the experiments reported here, the identification boundary for initial stop consonant shifted to a higher $F3$ frequency (more /ga/ or "low $F3$ " responses) when the CV was preceded by a high-frequency $F3$ or sine wave analogue. Identification boundaries shifted to a lower $F3$ frequency (more /da/ or "high $F3$ " responses) when the CV was preceded by a low-frequency $F3$ or sine wave analogue. This pattern of behavior can be recast as *auditory frequency contrast*. Effective $F3$ -onset frequency of a stop consonant is lowered following high-frequency $F3$ offset of [a]. The effective $F3$ -onset frequency of a stop consonant is raised following the low-frequency $F3$ offset of [ar].

Table 2
Synthesizer Parameter Values for [a] and [ar] Syllables
From Experiment 2

Frequency	Vowel	Consonant	
	[a]	[l]	[r]
f_0	110	110	110
$F1$	750	564	549
$F2$	1200	956	1517
$F3$	2450	2700	1600
$F4$	2850	2850	2850

Note—Values for the vowel [a] were constant over the initial 100 msec of each syllable. Entries for [l] and [r] represent the offset values after 150-msec linear transition from steady-state values.

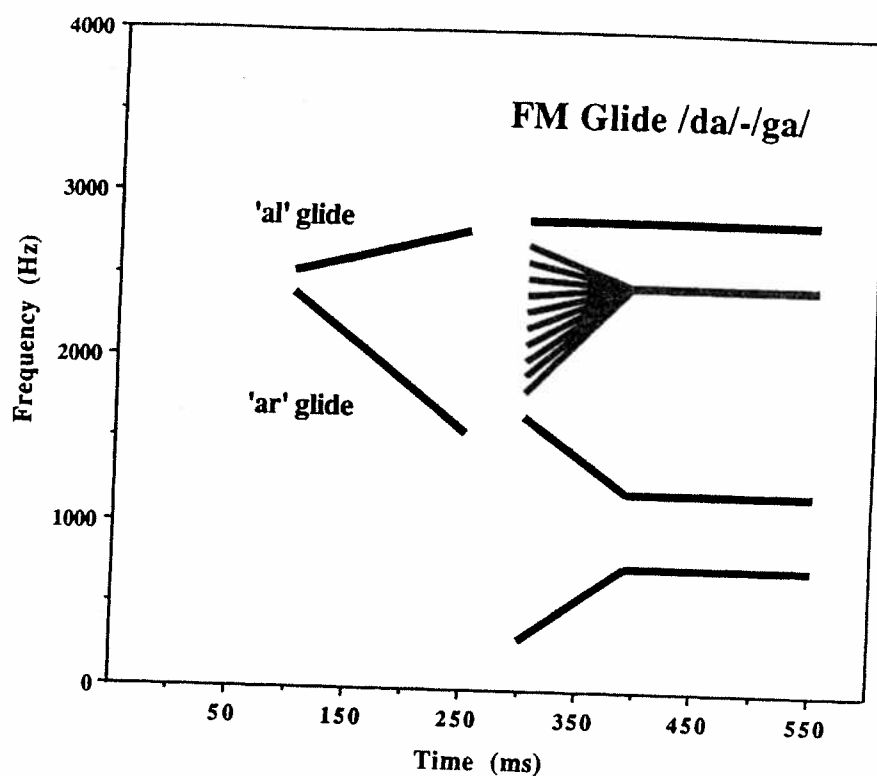


Figure 5. Schematic spectrograms of trajectories of preceding FM sine wave glides and the first four formants of following CV syllables. All 10 F_3 trajectories for the CVs are displayed.

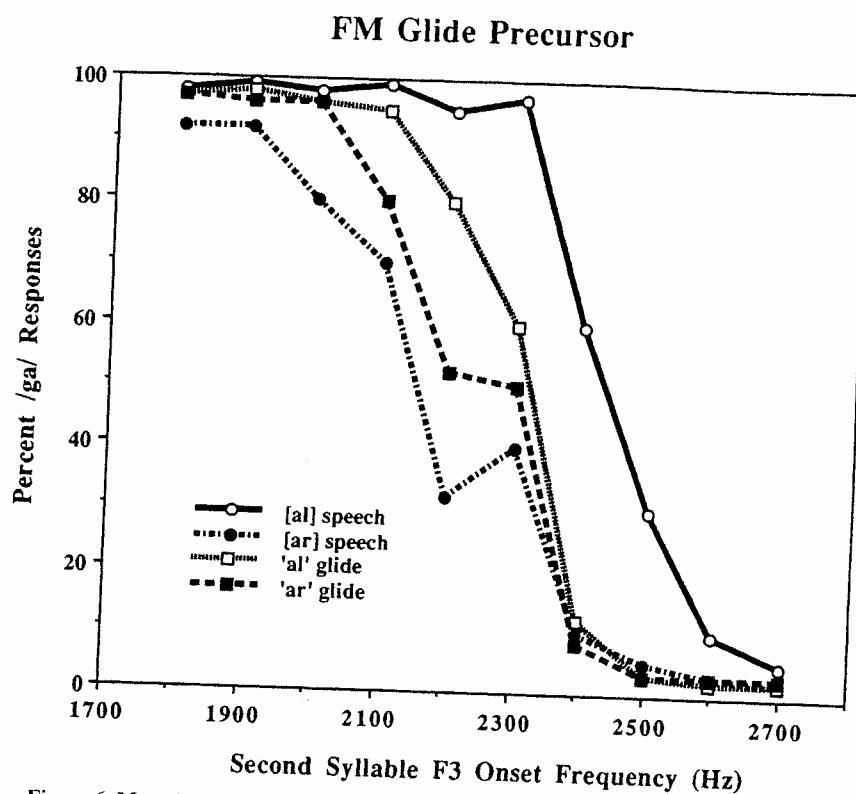


Figure 6. Mean identification functions from Experiment 2. Percent of /ga/ responses is plotted for each of the 10 CV stimuli, with preceding context serving as the parameter.

General auditory frequency contrast *has* been empirically established (Cathcart & Dawson, 1928–1929; Christman, 1954). Here, “frequency contrast” is used to denote a description of the results of a general auditory process without precise specification of a mechanism. The perceived frequency of a sound or component of a sound is shifted in a contrastive manner by the frequency of nearby (temporally and spectrally) sounds or components of sounds. Whether this result is due to peripheral auditory mechanisms or to central processes (such as Warren’s, 1985, “criterion shift”) is an empirical question that Experiments 1 and 2 do not resolve.

If, in fact, speech effects reported in Mann (1980) and in Experiments 1 and 2 are due to some general auditory frequency contrast that is not specific to speech, then contextual sounds that are not created by a human vocal tract could conceivably affect the labeling of a CV syllable. This seems to have been the case in Experiment 2 when sine wave glides influenced the labeling of subsequent stop consonants. However, it could be argued that these glide stimuli were sufficiently similar to speech sounds, at least in terms of spectral change near F_3 , for the glides to be incorporated into a mechanism designed especially to deal with the kinematics of speech sounds. This sort of argument has been expressed with regard to other non-speech results (e.g., Diehl & Walsh, 1989, suggest this potential explanation for the results of Pisoni, Carrell, & Gans, 1983). In this view, the shift in CV identification was due to processes designed specifically to compensate for the acoustic effects of coarticulation, albeit processes loosely constrained by the spectral composition of the sounds. This manner of speech-specific process which relies on more abstract kinematic consequences of articulatory constraints has been proffered to explain the finding that some subjects are able to hear complexes of sine waves modeling aspects of formants as speech (Remez, Rubin, Berns, Pardo, & Lang, 1994; Remez, Rubin, Pisoni, & Carrell, 1981).

Experiment 3 was designed to test further the stimulus specificity of contextual effects on stop consonant identification. Sine waves with constant frequencies equal to the F_3 offsets of natural /al/ and /ar/ utterances were used as the context for a series of CV syllables. If general psychophysical processes were culpable for the context effects described in Mann (1980), the frequency of these preceding sine waves might alter the effective frequency of F_3 onset for the initial stop consonant, resulting in a shift in identification functions. These sine wave tones did not offer kinematic information in the way that the glides from Experiment 2 did and, presumably, any shifts in identification boundaries engendered by these stimuli would be due solely to their dissimilar frequencies.

EXPERIMENT 3

Method

Subjects. Sixteen college-aged adults participated as listeners. All reported being native-English speakers with normal hearing.

Stimuli. To compare the present results with previous findings, it was determined that the stimuli used in Experiment 3 should resemble the stimuli utilized by Mann (1980, 1986) and Fowler et al. (1990). These earlier studies had revealed the context effect of liquids in several different populations (i.e., infants, English-speaking adults, Japanese-speaking adults), and it seemed important to establish the context effects of a nonspeech analogue with a stimulus set similar to that used in these earlier studies. In order to achieve this comparison, a seven-step /da/–/ga/ series was synthesized on the basis of specifications reported by Mann (1980). The syllables in the series differed solely in F_3 -onset frequency, which varied from 2104 to 2692 Hz in 98-Hz steps. The steady-state frequency values for the first three formants were 649, 1131, and 2448 Hz, respectively. The onset frequencies for F_1 and F_2 were set at 310 and 1588 Hz and changed linearly toward the steady-state values over 50-msec transitions. This transition duration was much shorter than the 100-msec transitions used for the stimuli in Mann (1980), because the authors were concerned that the initial consonants of the original stimuli could be mistaken for glides. Fundamental frequency was 100 Hz for the first 200 msec and declined to 88 Hz over the last 30 msec of the syllable.

The preceding context for these syllables was either of two constant-frequency sine waves with frequencies equal to the offset of F_3 for the natural [al] and [ar] utterances from the stressed-VC conditions of Mann (1980). As reported in Fowler et al. (1990), these frequencies were 2720 Hz for [al] (pronounced preceding [da] in Mann’s studies) and 1824 Hz for [ar] (preceding [da]). The sine waves were 250 msec in duration and included 5-msec amplitude ramps at the beginning and end of each stimulus. The sine waves and CV syllables were RMS matched and digitally appended on line with a 50-msec intervening silence. All other details of stimulus presentation were equivalent to those described for Experiment 1. A schematic representation of the stimulus setup is displayed in Figure 7.

Procedure. The subjects participated in a single session during which they heard each of the seven CV syllables appended to each of the two sine wave contexts 10 times, for a total of 140 trials. Again the subjects identified target syllables by pressing labeled buttons on a response box.

Results

Mean identification functions are presented in Figure 8. The difference in frequency of preceding tone resulted in a significant shift in identification functions [$F(1,15) = 19.09, p < .001$] despite the fact that the only characteristic the tones shared with [al] and [ar] was that they contained substantial energy in the region of F_3 offset. The shift is more substantial than for the sine wave glides of Experiment 2. This may have been due to the different CV stimulus construction or to the increased salience of the tones used in this experiment (the glides in Experiment 2 were attenuated 12–15 dB and had only two thirds of the duration). The size of the present effect may be somewhat surprising, considering the simplicity of the contextual stimuli. In fact, the identification boundary shift (as determined by probit analysis) was 103 Hz, whereas the boundary shift from the analogous speech experiment from Mann (1980; probit boundaries estimated from identification functions from Figure 2) was approximately 70 Hz. Thus, similar boundary shifts occur for nonspeech and speech contexts with very similar /da/–/ga/ series (presumably identical, except for the shorter transition durations).

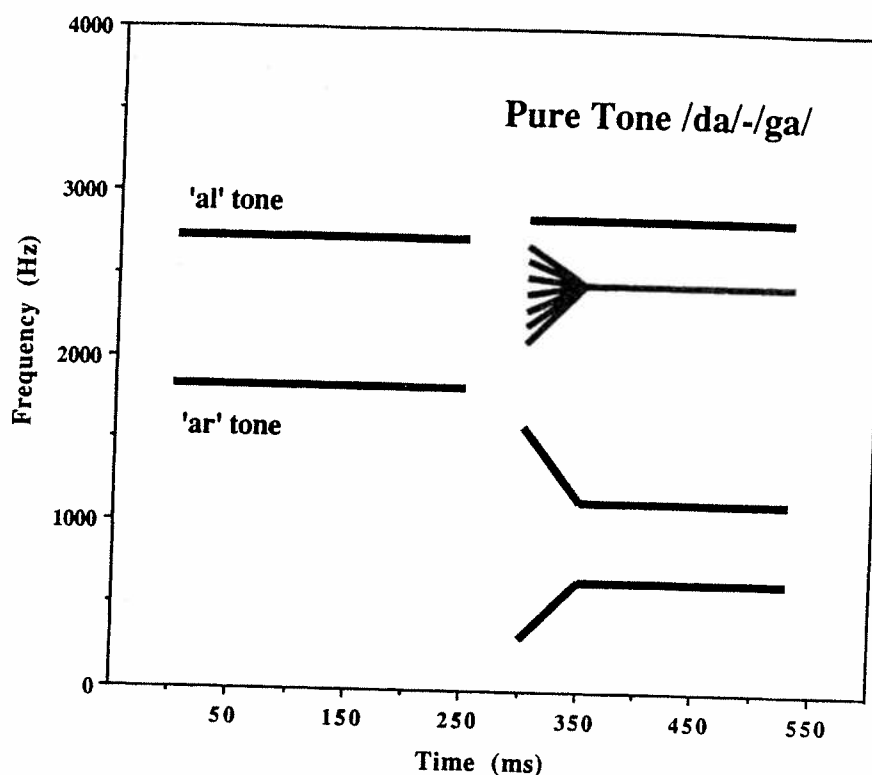


Figure 7. Schematic spectrograms of trajectories of preceding constant-frequency sine wave tones and the first four formants of following CV syllables. All seven F_3 trajectories for the CVs are displayed.

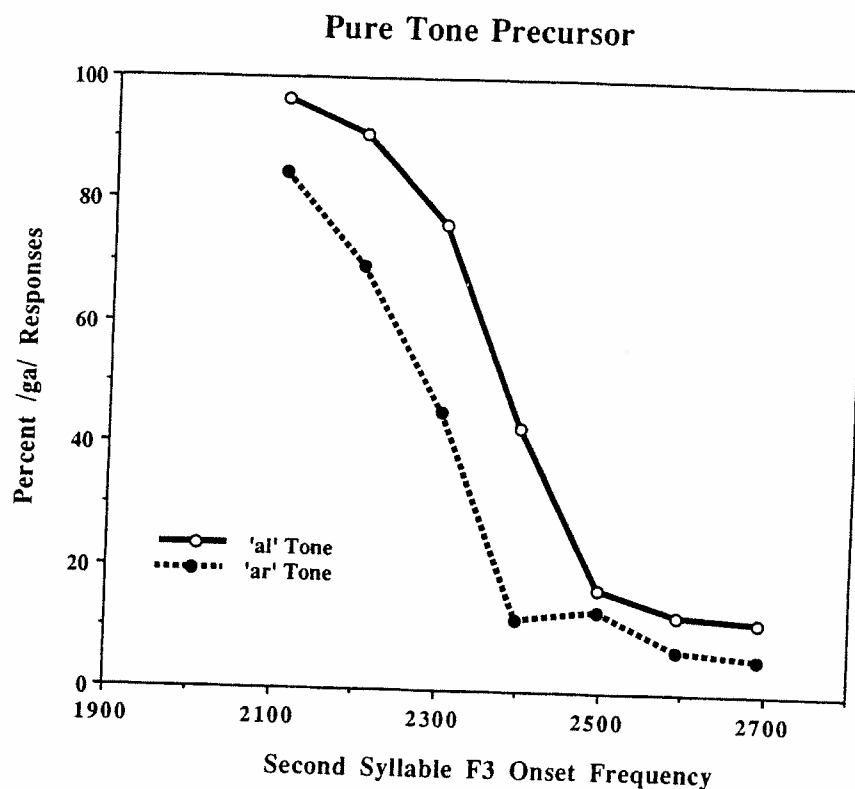


Figure 8. Mean identification functions from Experiment 3. Percent of /ga/ responses is plotted for each of the seven CV stimuli, with preceding context serving as the parameter.

Given the results of Experiment 3 along with those from the previous two experiments, one would be hard pressed to conceive of a model of "coarticulation compensation" that relied on specific constraints of vocal tracts and could account for these effects. Of course, one could maintain that separate mechanisms were responsible for the speech and nonspeech results. However, if one wields Occam's razor, explanatory generality recommends that one search for a general auditory process that can account for both effects. It is too early to specify this general auditory process exactly. However, one can describe the resulting behavior expressed in these experiments. The effective frequency of a formant (any sound?) can be changed by preceding frequencies in a contrastive manner. For example, following a concentration of high-frequency energy, the effective frequency of a formant (in terms of identification) is lowered.

If, in fact, human listeners "compensate" for coarticulation through some general process such as frequency contrast, then a rather general algorithm may be successful for practical applications. Estimations of formant frequency values derived from natural productions could be modulated by an appropriate function specifying the auditory contrastive effects of contextual sound. The results of experiments presented in this report are insufficient for the reconstruction of such a function. Programmatic research on the temporal and spectral extent of frequency contrast is necessary. Nonspeech sounds, such as the sine waves in Experiments 2 and 3, can be used effectively to expose input-output characteristics and help to define the range of frequency contrast. Such research could result in an appropriate model of speech perception that could deal with a conundrum like that posed in Figure 1.

For those interested in the ephemera of theories of actual human speech perception (including the authors of this report), the mechanism behind the pattern of behavior described as "frequency contrast" is of central concern. Possibilities include peripheral processes (such as masking) or central processes of shifted judgments. The former is consistent with processes of lateral inhibition underlying visual brightness contrast (Hartline, Ratliff, & Miller, 1961). Central processes are favored by results presented in Mann and Liberman (1983). Subjects exhibited shifted discrimination boundaries for /da/-/ga/ series preceded by [al] or [ar], even though the distinguishing F_3 transition was presented to the opposite ear from that for the [al] and [ar] stimuli. Mann and Liberman argue that this rules out peripheral auditory explanations, inasmuch as input to the contralateral ear does not affect peripheral processing.⁹ One possible concern with Mann and Liberman's results is that contrast effects involving F_2 frequency could have played a role. The F_2 offsets for the [ar] stimuli used in that study were of higher frequency than the F_2 -offset frequency of the [al] stimuli. Identification of [da]-[ga] can be affected by the onset frequency of F_2 for the CV, with a higher F_2 frequency favoring [ga] responses (Delattre, Liberman,

& Cooper, 1955). Frequency contrast of F_2 would predict the same pattern as that witnessed in Mann (1980): more /ga/ (high F_2) responses following /al/ (low F_2). Because energy in the F_2 region was delivered to the same ear for both syllables in Mann and Liberman (1983), peripheral effects cannot be ruled out. Clearly, further experimentation will be necessary for researchers to establish the mechanism responsible for the frequency contrast behavior in the cases presented above. Given the seemingly general nature of the responsible process, the unveiling of this mechanism may have important implications for our understanding of all perceptual systems.

DISCUSSION

The experiments in this report were designed to specify the information that is important for perceptual compensation for the acoustic effects of coarticulation. Given the complexity of efforts to recover vocal-tract shape from the resultant acoustic wave form, it seemed beneficial to determine whether solving the inverse problem is necessary in order to manage coarticulatory influences. Results from three experiments demonstrate that effects of context are not critically dependent on acoustic consequences specific to a single vocal tract or, indeed, to any vocal tract. As a result, *frequency contrast* was proffered as a description of the behavior of a general perceptual process that is responsible for the effect of preceding liquid identity on stop consonant identification.

A general perceptual solution to the phenomena of speech context effects has much to offer in terms of practical and theoretical simplicity. Accounts of infants (Fowler et al., 1990) and Japanese listeners (Mann, 1986) displaying contextual effects of liquids on stop consonant perception become less surprising if general perceptual processes are responsible. Because the present results with nonspeech suggest that this "compensatory" process is extralinguistic, specific linguistic experience would not be prerequisite for these context effects.

Contrast may be a rather general solution to the effects of phonemic context on identification. Coarticulation tends to be assimilative, and contrastive perceptual processes could compensate for much of the lack of invariance in speech acoustics due to articulatory dynamics. No cases of contrastive effects of coarticulation are readily apparent to the authors. These would certainly serve as important test cases, were they to exist. However, a review of Öhman's (1966) spectrographic measurements of VCV utterances, for example, does not reveal any appropriate contrastive cases. Fowler et al. (1990) state plainly that "acoustic effects of coarticulation are generally assimilative, and contrastive effects of the coarticulating segment's acoustic consequences will always work to neutralize the perceptual effects of the assimilations" (p. 567).

In terms of theoretical generality, effects of liquid identity on stop identification join a list of speech-perception phenomena that appear to be symptomatic of general au-

ditary (perceptual) processes. For example, the effect of $F1$ -onset frequency on voice/voiceless identification may be due to rather general processes, according to data from nonspeech identifications (Parker, 1988) and non-human-animal experiments (Kluender, 1991; Kluender & Lotto, 1994).

Of practical concern is the presumed computational simplicity of a frequency contrast approach to the problem of lack of invariance of speech due to coarticulation.¹⁰ If appropriate research on the limits of frequency contrast can determine the function for frequency contrast with complex signals such as speech, a model of human speech perception that varied formant values in accordance with frequency contrast prior to calling any recognition procedures could potentially disambiguate situations such as those presented in Figure 1. This algorithmic approach would appear, at this early stage, to be less complex than quixotic efforts to recover vocal-tract shape from the acoustic wave form. In addition to being computationally complex, it is possible that this inverse problem is impossible to solve, in general. Related efforts in acoustics, such as attempts by mathematicians to solve the inverse problem for simpler sound-producing devices such as drum heads (conceptualized as Riemannian manifolds; Kac, 1966), show that for many cases a unique inverse cannot be derived (Gordon, Webb, Wolpert, 1992). The results of Experiments 1, 2, and 3 challenge the necessity of recovery of vocal-tract shape or dynamics for human-like speech perception.

Similarly, the results recommend a revision of simple pattern recognition models (e.g., Nearey, 1990, 1992, 1995). As noted above, these models depend on discrete variables encoding acoustic characteristics in order to classify syllables. If the formant frequency values that are fed into these linear equations are determined by an appropriate frequency contrast function, the models may be able to achieve human-like performance on identification of syllables distinguished solely by context. The alternative of adding a variable to encode $F3$ -offset frequency of the preceding syllable seems less appealing, because the effects of these added variables on identification are, in Nearey's formulation, learned through experienced covariation. Although one ought not underestimate the contribution of general learning processes for speech perception, the results of Experiments 2 and 3 demonstrate that the effect of liquid context is not speech specific, and it is doubtful that the frequency of sine waves and CV $F3$ -onset frequencies covary in the natural environment to any great extent. In general, speech phenomena that arise from general auditory processes are not well represented in these models or in Nearey's (1991, 1995) *double-weak theory*, which de-emphasizes auditory and articulatory constraints on speech perception (Kluender & Lotto, 1997). Through the incorporation of auditory functions, such as frequency contrast, these simple pattern-recognition models might offer a more cogent theory of actual human performance by ameliorating failures to account for context effects.

It should be noted that the authors of the present report are not theoretically committed to formant extraction as part of a model of human speech perception. Although the use of the term "frequency contrast" might seem to imply some manner of formant frequency extraction, this term is meant to be a useful, and perhaps efficient, descriptive term for the results without carrying specific theoretical content (similar to the caveat concerning "phoneme" in note 1). The fact of the matter is that manipulations in formant frequency values result in shifts in the relative amplitude of harmonics and *not* in a shift in the frequency of the harmonics (assuming a stable f_0). A masking explanation would relate these context effects in terms of relative loudness levels, not pitches, of components. The resulting identification functions can be described as a contrastive change in effective $F3$ frequency, but the mechanism behind this behavior may be described better as "spectral contrast." Hopefully, future research will disambiguate the mechanisms behind this contrastive behavior. Until then, "frequency contrast" will remain a descriptive term for the mapping between stimulus manipulations and identification patterns.

There are, of course, several reservations that one may hold concerning this general perceptual framework for context effects in speech. One concern arising from the data from Experiments 1 and 2 regards the size of the effect in nontypical conditions (female preceding or glide preceding) in comparison with effect sizes for the typical condition of one talker producing the disyllable (see Table 1). It is possible that general auditory frequency contrast accounts for the two nontypical effects, but that a speech-specific process is active for the processing of the single-speaker disyllables. According to this argument, the considerable effect size for the typical condition would be due to the requirements of coarticulatory compensation, which may necessitate a greater shift in identification boundaries than is afforded by the general processes of contrast. This is certainly a reasonable statement, even in light of the findings of Experiment 3, in which the nonspeech effect was similar to that reported for a natural utterance (Mann, 1980). In defense of the general contrast account of context effects, it is possible that perceptual organization played some role in effect size. The substantial f_0 difference between the female VC and male CV productions, as well as the qualitative difference between the CVs and the preceding sine glides of Experiment 2, would favor perceptual disassociation by grouping rules such as those encapsulated in "auditory scene analysis," described by Bregman (e.g., 1981, 1990). Empirical results from the visual modality demonstrate that context effects are malleable in relation to perceptual organizations. For example, Gilchrist (1977) has reported that brightness contrast occurs only for luminances that are perceived as coplanar. Gogel (1978) has described a series of experiments that resulted in decreasing context (or "induction") effects as a function of increasing perceptual distance between objects. Auditory frequency contrast could follow an analogous function: As evidence

of disparate sources increases, the contextual effects of sounds decrease.¹¹ Thus, the change in sex of speaker in Experiment 1 would result in acoustic attributes that signal a change in source, and the resulting contrast effect would be lessened. In addition, there may be other attributes of preceding [al] or [ar] syllables that additively affect subsequent stop identification that were not represented in the sine wave glides modeled after *F3*. For example, there may be frequency contrast effects of *F2*.

A second concern about an account based on frequency contrast was raised in Fowler et al. (1990). Although the authors of that paper listed two reasons for discounting contrast effects as responsible for context effects in speech perception, the reasons can be summarized as concerning the lack of parsimony of the hypothesis. According to Fowler et al., the contrast explanation leads to a proliferation of levels of contrast, because contrast could occur at a peripheral level (but does not, according to the results of Mann & Liberman, 1983), at a high cognitive level (but does not, according to the results with Japanese listeners of Mann, 1986), or at some intermediate level. In addition, Fowler et al. claim that auditory contrast fails to account for findings in which coarticulatory influences serve as information for the surrounding context, as is shown by slower reaction times to hybrid syllables that mismatch members of coarticulated syllables (Fowler, 1984; Fowler & Smith, 1986). The authors of the present report agree with Fowler that a "direct realist" approach is quite parsimonious and capable of dealing with these previous findings. Unfortunately, the data reported here contradict the predictions of such a theory. Sounds that obviously arise from separate sources, such as sine wave glides, affect the labeling of CV syllables. If the perceptual system was picking up valid information about source identity and commensurate constraints, then data such as those reported in the present paper are inexplicable unless one can develop some nonparsimonious adjunct assumptions. Given the constellation of findings at present, the present general perceptual account *does* appear to be the most unadorned.

The generality of the purported "general" processes, however, may be called into question. Despite the results of Experiment 3, one may continue to propose that these effects are due to the special character of vocal-tract constraints. As Kuhl (1978, 1986a, 1986b) points out, data from nonspeech stimuli alone can be accommodated by advancing a "speech-special" mechanism with quite broad application. Although this brings into question the usefulness of the term "speech-specific,"¹² it does resonate with theories of sine wave speech perception (Remez et al., 1994). Another possibility, not often submitted, is that the overlearned nature of speech perception affects the perception of nonspeech events. That is, constraints inherent in speech sounds are incorporated into the auditory processing of all quasi-periodic sounds, regardless of origin. For example, Deutsch (1996) has re-

cently reported that the perception of a nonspeech auditory illusion varies with the language experience of the listener. Perhaps even by 4 months of age, the age of the infants in Fowler et al. (1990), humans have received enough speech input to affect nonspeech perception.

To test the generality of these context effects, one needs to use nonhuman animal subjects. Previous animal studies of speech perception have been used to assess general auditory processes without confounds of effects of experience (e.g., Dooling, Best, & Brown, 1995; Kluender, 1991; Kluender & Lotto, 1994; Kuhl & Miller, 1975, 1978). These animals would not be expected to possess innate speech-specific mechanisms. Thus, if nonhuman animals demonstrate effects of phonemic context on stop consonant "labeling" like those demonstrated in Mann (1980), a general auditory account becomes bolstered. Experiments of this type, with Japanese quail (*Coturnix japonica*), are being conducted by the authors of this report.

With results from nonhuman animal subjects and nonspeech analogues such as those used in Experiments 2 and 3, one can begin to see the lineaments of a general theory of "perceptual compensation for coarticulation." Because of the assimilative nature of coarticulation, general contrastive perceptual processes may be able to accommodate many of the acoustic consequences of articulatory dynamics. Practically, machine speech recognition attempts could be enhanced by incorporating frequency contrast functions. This solution, provided by nature, may prove less complex than attempts to recover vocal-tract configurations from the acoustic wave form.

REFERENCES

- BREGMAN, A. S. (1981). Asking the "what for" question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 99-118). Hillsdale, NJ: Erlbaum.
- BREGMAN, A. S. (1990). *Auditory scene analysis*. Cambridge, MA: MIT Press.
- CATHCART, E. P., & DAWSON, S. (1928-1929). Persistence (2). *British Journal of Psychology*, **19**, 343-356.
- CHRISTMAN, R. J. (1954). Shifts in pitch as a function of prolonged stimulation with pure tones. *American Journal of Psychology*, **67**, 484-491.
- DAINORA, A., HEMPHILL, R., HIRATA, Y., & OLSON, K. (1996). Effects of context and speaking rate on liquid-stop sequences: A reassessment of traditional acoustic cues. *Journal of the Acoustical Society of America*, **100**, 2601.
- DELATTRE, P. C., LIBERMAN, A. M., & COOPER, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America*, **27**, 769-773.
- DEUTSCH, D. (1996). Paradoxical music. *Echoes*, **6**, pp. 1, 4-5.
- DIEHL, R. L., & WALSH, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, **85**, 2154-2164.
- DOOLING, R. J., BEST, C. T., & BROWN, S. D. (1995). Discrimination of synthetic full-formant and sinewave /ra-la/ continua by budgerigars (*Melopsittacus undulatus*) and zebra finches (*Taeniopygia guttata*). *Journal of the Acoustical Society of America*, **97**, 1839-1846.
- FANT, G. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report* (No. 4, pp. 22-30). Stockholm: Royal Institute of Technology.

- FANT, G. (1975). Non-uniform vowel normalization. *Speech Transmission Laboratory Quarterly Progress and Status Report* (Nos. 2-3, pp. 1-19). Stockholm: Royal Institute of Technology.
- FOWLER, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics*, **36**, 359-368.
- 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., BEST, C. T., & McROBERTS, G. W. (1990). Young infants' perception of liquid coarticulatory influences on following stop consonants. *Perception & Psychophysics*, **48**, 559-570.
- FOWLER, C. A., & SMITH, M. R. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (Eds.), *Invariance and variability of speech processes* (pp. 123-139). Hillsdale, NJ: Erlbaum.
- GILCHRIST, A. L. (1977). Perceived lightness depends on perceived spatial arrangement. *Science*, **195**, 185-187.
- GOGEL, W. C. (1978). The adjacency principle in visual perception. *Scientific American*, **238**, 126-139.
- GORDON, C., WEBB, D. L., & WOLPERT, S. (1992). One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society*, **27**, 134-138.
- GREEN, K. P., STEVENS, E. B., & KUHL, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics*, **55**, 249-260.
- HARTLINE, H. F., RATLIFF, F., & MILLER, W. H. (1961). Inhibitory interaction in the retina and its significance in vision. In E. Florey (Ed.), *Nervous inhibition* (pp. 241-284). New York: Pergamon.
- HENTON, C. G., & BLADON, R. A. W. (1985). Breathiness in normal female speech: Inefficiency versus desirability. *Language & Communication*, **5**, 221-227.
- HOLMBERG, E. B., HILLMAN, R. E., & PERKELL, J. S. (1988). Glottal air flow and pressure measurements for soft, normal and loud voice by male and female speakers. *Journal of the Acoustical Society of America*, **84**, 511-529.
- JAMIESON, D. J., RAMJI, K. V., KHEIRALLAH, I., & NEAREY, T. M. (1992). CSRE: A speech research environment. In J. Ohala, T. Nearey, B. Derwing, M. Hodge, & G. Wiebe (Eds.), *Proceedings ICSLP 92* (pp. 1127-1130). Edmonton, AB: University of Alberta.
- JOHNSON, K. (1989). Higher formant normalization results from auditory integration of F2 and F3. *Perception & Psychophysics*, **46**, 174-180.
- KAC, M. (1966). Can one hear the shape of a drum? *American Mathematics Monthly*, **73**, 1-23.
- KENT, R. D., & BURKHARD, R. (1981). Changes in the acoustic correlates of speech production. In D. S. Beasley & G. A. Davis (Eds.), *Aging: Communication processes and disorders* (pp. 47-62). New York: Grune & Stratton.
- KENT, R. D., & MINIFIE, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics*, **5**, 115-133.
- KLATT, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.
- KLATT, D. H., & KLATT, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, **87**, 820-857.
- KLUENDER, K. R. (1991). Effects of first formant onset properties on voicing judgments result from processes not specific to humans. *Journal of the Acoustical Society of America*, **90**, 83-96.
- KLUENDER, K. R., & LOTTO, A. J. (1994). Effects of first formant onset frequency on [-voice] judgments result from general auditory processes not specific to humans. *Journal of the Acoustical Society of America*, **95**, 1044-1052.
- KLUENDER, K. R., & LOTTO, A. J. (1997). *Virtues and perils of an empiricist approach to speech perception*. Manuscript submitted for publication.
- KUHL, P. K. (1978). Predispositions for the perception of speech-sound categories: A species-specific phenomena. In F. D. Minifie & L. L. Lloyd (Eds.), *Communicative and cognitive abilities—early behavioral assessment* (pp. 229-255). Baltimore: University Park Press.
- KUHL, P. K. (1986a). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception* (pp. 355-386). Cambridge, MA: Cambridge University Press.
- KUHL, P. K. (1986b). Theoretical contributions of tests on animals to the special-mechanisms debate in speech. *Experimental Biology*, **45**, 233-265.
- KUHL, P. K., & MILLER, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, **190**, 69-72.
- KUHL, P. K., & MILLER, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, **63**, 905-917.
- LIBERMAN, A. M., & MATTINGLY, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.
- LOTTO, A. J., GREEN, K., & KLUENDER, K. R. (1993). Vowel continuity and perception of /ba-wa/. *Journal of the Acoustical Society of America*, **93**, 2391-2392.
- LOTTO, A. J., KLUENDER, K. R., & GREEN, K. P. (1996). Spectral discontinuities and the vowel length effect. *Perception & Psychophysics*, **58**, 1005-1014.
- MANN, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, **28**, 407-412.
- MANN, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English "l" and "r." *Cognition*, **24**, 169-196.
- MANN, V. A., & LIBERMAN, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, **14**, 211-235.
- MCGOWAN, R. S. (1994). Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, **14**, 19-48.
- MCGOWAN, R. S., & RUBIN, P. E. (1994). Perceptual evaluation of articulatory movement recovered from acoustic data [Abstract]. *Journal of the Acoustical Society of America*, **96**, 3328.
- MERMELSTEIN, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America*, **53**, 1070-1082.
- MONSEN, R. B., & ENGBRETSON, A. M. (1977). Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, **62**, 981-993.
- NEAREY, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, **18**, 347-373.
- NEAREY, T. M. (1991). Perception: Automatic and cognitive processes. In *Proceedings of the XIIIth International Congress of Phonetic Sciences* (Vol. 1, pp. 40-49). Aix-en-Provence: Publications de l'Université de Provence.
- NEAREY, T. M. (1992). Context effects in a double-weak theory of speech perception. *Language & Speech*, **35**, 153-172.
- NEAREY, T. M. (1995). Speech perception as pattern recognition [Abstract]. *Journal of the Acoustical Society of America*, **97**, 3334.
- ÖHMAN, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, **39**, 151-168.
- PARKER, E. M. (1988). Auditory constraints on the perception of stop voicing: The influence of lower-tone frequency on judgments of tone-onset simultaneity. *Journal of the Acoustical Society of America*, **83**, 1597-1607.
- PISONI, D. B., CARRELL, T. D., & GANS, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.
- REMEZ, R. E., RUBIN, P. E., BERNIS, S. M., PARDO, J. S., & LANG, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**, 129-156.
- REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL, T. D. (1981). Speech perception without traditional speech cues. *Science*, **212**, 947-950.
- SALTZMAN, E. (1986). Task-dynamic coordination of the speech articulators: A preliminary model. *Experimental Brain Research*, **15**, 129-144.
- SALTZMAN, E., & KELSO, J. A. (1987). Skilled actions: A task-dynamic approach. *Psychological Review*, **94**, 84-106.
- SCHROETER, J., & SONDH, M. M. (1992). Speech coding based on physiological models of speech production. In S. Furui & M. M. Sondhi (Eds.), *Advances in speech signal processing* (pp. 588-591). New York: Marcel Dekker.
- STATHOPOULOS, E. T., & SAPIENZA, C. (1993). Respiratory and laryngeal measures of children during vocal intensity variation. *Journal of the Acoustical Society of America*, **94**, 2531-2543.

- STEVENS, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America*, 32, 47-55.
- WARREN, R. M. (1985). Criterion shift rule and perceptual homeostasis. *Psychological Review*, 92, 574-584.

NOTES

1. This conceptualization of coarticulation assumes that discrete and invariant units, such as phonemes, are fundamental elements serving as input for articulatory motor control systems and as the output for speech perception (e.g., Kent & Minifie, 1977). While the phoneme does provide a rather efficient and powerful linguistic descriptor of speech behavior, its psychological validity has not been established unequivocally (though see Nearey, 1990). That is, the phoneme is a conceptual abstraction. It does, however, provide a useful heuristic device, given the lack of invariance between acoustics and listeners' speech-labeling behavior. Thus, despite uncertainty about the phoneme's ontological status, the helpful descriptive system of phonemic segments will be used throughout this paper.

2. These models could account for the labeling data presented in Mann (1980) if one added a weighted attribute based on the frequency of the F3 offset of the preceding syllable. This solution may be unsatisfactory when one considers all of the attributes that would have to be entered in order to account for all possible coarticulatory influences across syllables. A slight modification to Nearey's model is suggested in the General Discussion.

3. These differences between vocal tracts are compounded by introducing children as speakers. Differences in vocal-tract shape shift formant values throughout the life of a human (Kent & Burkhard, 1981), and children seem to be qualitatively different in their use of laryngeal and respiratory mechanisms when producing speech (Stathopoulos & Sapienza, 1993). Considering these individual differences, one may wonder how specific the representations of vocal-tract constraints and articulatory dynamics could be in accounts such as Mann (1980) or other "motor theory" models (Lieberman & Mattingly, 1985).

4. Of course, listeners may use lexical or probabilistic information to compensate for coarticulation in some cases. Presumably, this is not the case for the nonsense disyllables used in this study.

5. Formant transitions of 80-msec duration may seem to be rather long, but these were measured from natural productions and are comparable to the 100-msec transitions used by Mann (1980).

6. The identification function for the isolated CVs are not presented in the graph to maintain clarity. This function was nearly identical to the female-/aI/ function.

7. It should be noted once again that there was no coarticulation present in the creation of these stimuli, given that the preceding contexts were produced in isolation. Thus, there is presumably no information in both syllables, which specify the same gesture. However, it is still possible that the manner of presentation encouraged subjects to discover relational information that was not valid.

8. The authors wish to thank Catherine Best for pointing out this theoretical possibility.

9. This may be a dubious assumption, considering that efferent neural connections from the superior olivary complex (which receives contralateral input) may affect peripheral processing. Owing to the fact that—only two synapses away from the hair cell—substantial contralateral connections converge at the inferior colliculus (and superior olive), one must be very cautious in rendering conclusions from dichotic studies about the level of the auditory system at which some process occurs.

10. The function for frequency contrast could, of course, turn out to be rather complex. For example, if the pitch of a component was a function of the perceived frequency of other components, the resulting non-linear function could be computationally extravagant.

11. This organization would probably be more commensurate with what Bregman (1990) calls "primary auditory stream segregation" than

with "schema-based segregation." Lotto et al. (1996) have shown that perceived changes in speaker or phonemic identity play little role in the effects of syllable duration on stop identification. Also, Lotto, Green, and Kluender (1993) report that even perceived continuity is not essential to determining what spectral information will be included in speech perception effects.

12. This argument dilutes the falsifiability of the "speech-specific" hypothesis because of its rather circular form. The sounds that elicited "speech-like" behavior were processed by a "speech-specific" mechanism; those that did not elicit "speech-like" behavior were not processed by a "speech-specific" mechanism.

APPENDIX

The natural productions of [aI] and [aʀ] that served as the preceding context in Experiment 1 were produced in isolation and under instructions to produce clear exemplars of the syllables. Several productions of each syllable were recorded digitally and analyzed. The syllables chosen were closest to a desired duration of 250 msec. Discrete fast Fourier transforms were calculated on these syllables, and measurements of the frequencies of the first three formants were obtained with CSRE software (Jamieson, Ramji, Kheirallah, & Nearey, 1992) run on a microcomputer. Measurements for the vowels were taken from near the midpoint of the steady-state portion, where the formant frequencies were judged to be least variable. Measurements for the liquids were taken from as near the offset of the syllable as possible. Average fundamental frequency was determined at the measurement location for the vowel.

Table A1
Measurements of Formant Frequencies (in Hertz) for
Both Male and Female Productions in Experiment 1

	Female Productions	
	Vowel ([a])	Consonant ([l])
F1	896	700
F2	1209	998
F3	3260	3644
f0	217	
	Male Productions	
	Vowel ([a])	Consonant ([r])
F1	908	327
F2	1131	1885
F3	3030	2073
f0	213	
	Male Productions	
	Vowel ([a])	Consonant ([l])
F1	633	306
F2	910	795
F3	2595	2863
f0		106
	Male Productions	
	Vowel ([a])	Consonant ([r])
F1	748	370
F2	1064	1593
F3	2212	1593
f0		100

(Manuscript received April 3, 1996;
revision accepted for publication April 4, 1997.)