Perceptual Compensation for Coarticulation as a General Auditory Process

Andrew J. Lotto
Boys Town National Research Hospital

1. Introduction

One of the most significant challenges for models of speech perception is explaining the effect of context on phonemic identification. The mapping from the acoustics of the speech signal to identified phonemes is neither simple nor transparent. It has been demonstrated that acoustic, phonological, semantic and syntactic context can all shift the perceived identity of a speech sound (e.g., Ganong, 1980; Repp, 1982; Connine, 1987; Borsky, Tuller, & Shapiro, 1998). As a result, two identical acoustic segments can be labeled as different phonemes and two different acoustic segments can be labeled with identical phonemic labels.

For example, phonetic context has been shown to result in substantial shifts in perceived phonemic identity. Mann (1980) presented listeners with a series of syllables that varied from /ga/ to /da/ (the target stimuli). Each syllable was preceded by natural productions of the syllable /al/ or the syllable /ar/ (the context stimuli). This change in context resulted in a shift in the identification of the ambiguous /ga/-/da/ syllables. Syllables following /al/ were more often identified as /ga/ than were the same syllables preceded by /ar/. Thus, identical acoustic segments were labeled differently depending on phonetic context. This perceptual shift is in the opposite direction of the influences of /l/ and /r/ on the acoustics of the production of /d/ and /g/ due to coarticulation. Due in part to physical constraints of mass and inertia, a /g/ produced following the production of /l/ is drawn toward the anterior of the oral cavity. This context-dependent changes in production is reflected in the resulting acoustics and a /g/ following /l/ in natural speech becomes acoustically more similar to /d/ (anterior place of articulation). Thus, the perceptual shift appears to compensate for coarticulation (consonant after /l/ identified more often as /g/) allowing identification of intended phonemes despite context-specific acoustics. The evidence for influence of phonetic context on phonemic categorization is extremely rich. Such effects have been documented across many contexts (e.g., Lindblom & Studdert-Kennedy, 1967 [vowels]; Mann & Repp, 1981 [fricatives]; Mann, 1980 [stop-consonants]; see Repp, 1982 for a review), as well as developmentally (Fowler, Best, & McRoberts, 1990) and cross-linguistically (Mann, 1986). What aspect of the context is culpable for these robust shifts in phonemic identification?

The answer that seems most plausible is that it is the phonemic content of the context that determines the shift in identification of the target sound. This answer appears explicitly in the TRACE model of word recognition offered by McClelland and Elman (1986; Elman & McClelland, 1986). The TRACE model accounts for phonetic context effects by allowing the activation of phoneme representation "nodes" to modulate the mapping from feature values to phoneme nodes for subsequent sounds in the speech wave. In the case of the context effect described by Mann (1980), the presence of the context syllable /al/ would lead to the activation of the /l/ phoneme node. This activation would change the strength of the connection between the distinguishing feature node (presumably the onset frequency of the third formant, in this case) and the phoneme nodes for /d/ and /g/. The identification of the target consonant would, as a result, be biased toward /g/. The activation of the /r/ node by the /ar/ context would presumably alter the connection strengths in a complementary fashion leading to

 Proceedings of the 2003 Texas Linguistics Society Conference, ed. Augustine Agwuele, Willis Warren, and Sang-Hoon Park, 42-53. Somerville, MA: Cascadilla Proceedings Project.

more /d/ identifications. In either case, it is the activation of the *phoneme* node that would drive the context effect. For the TRACE model, the perceived phonemic content of the context determines the shift.

In addition to phonemic identity, there is another aspect of contextual content that may play a role in identification shifts of target sounds. In classic demonstrations of phonetic context effects, the spectral content (i.e., the acoustics of the speech sound) is manipulated in order to achieve a shift in the phonemic content of the context. For example, in the Mann (1980) study the offset frequencies for the second (F2) and third (F3) formants differed between /al/ and /ar/. It is possible that these changes in spectral energy, independent of phonemic status, can cause shifts in the perceived identity of subsequent speech sounds. In fact, there have been several recent demonstrations that nonspeech context sounds can cause changes in perceived identity of temporally proximate speech sounds. For example, Lotto and Kluender (1998) were able to obtain shifts in identification of a /ga/-/da/ series by preceding the syllables with simple tones that mimicked the frequency trajectory of the F3 offset for /al/ and /ar/. This context had no recognizable phonemic content but resulted in a context effect in the same direction as obtained for speech contexts (Mann, 1980). Holt, Lotto, and Kluender (1998) demonstrated that the perceived identity of ambiguous /ba/-/da/ syllables could be shifted by preceding the syllables with the vowels /i/ or /u/ (more /ba/ responses after /i/). Again, target identification shifts in the same direction could be elicited by context tones that followed the frequency of F2 for the vowels. Holt, Lotto, and Kluender (2000) offered analogous results for nonspeech sounds shifting the identification of vowels. These nonspeech effects must have been driven by the acoustic structure of the context, as they were not perceived as speech or any other recognizable environmental sound. Presumably the only perceptual representation for these sounds would consist of their spectral composition (as opposed to a lexical label or other discrete representation). In fact, Holt (1999) established that the size of the nonspeech effect on /ba/-/da/ identifications was a strict function of tone frequency with no evidence of any discrete boundary effects that would be a symptom of discrete label representations.

These results have been interpreted by Lotto and Kluender (1998) as indicative of general auditory interactions that affect the processing of the spectral content of the target syllable. That is, the interactions are at a sub-phonemic level. If phonetic context effects are due in part to the spectral content of the context, this would have significant implications for models of speech perception. In particular, comprehensive models would be required to incorporate more realistic representations of spectral content.

To date, the demonstrations of nonspeech context effects have not completely clarified the respective roles of phonemic and spectral content in phonetic context effects. The nonspeech context shifts obtained by Lotto and Kluender (1998; Experiment 2) were not as large as those obtained for speech contexts. These experiments were designed in the psychoacoustics tradition of speech perception research in which individual stimulus attributes are manipulated to determine their effect on identification or discrimination functions. As a result, the difference in speech and nonspeech effects may have been due to either the lack of phonemic content in the nonspeech stimuli or because the spectral content wasn't similar enough. In particular, /al/ and /ar/ differ substantially in the offset frequencies of F2 and F3, whereas Lotto and Kluender only modeled the F3 frequency trajectory.

The two experiments presented here were designed to examine the respective roles of phonemic and spectral content in phonetic context effects by first manipulating phonemic status while holding important spectral characteristics similar and then by holding labeled identity constant while changing spectral composition. In both experiments, attempts were made to match the speech and nonspeech signals on important spectral qualities.

2. Experiment 1

The most salient distinguishing features of the liquids in /al/ and /ar/ are the offset frequencies of F2 and F3. The consonant /l/ is characterized by a high-frequency-F3 offset and a low-frequency-F2 offset. The consonant /r/, in contrast, has a lower-frequency-F3 offset and a higher-frequency-F2 offset. To create reasonable nonspeech analogs, steady-state sine wave tones were matched in frequency to the offset of both F2 and F3 for each consonant. Thus, some spectral information

concerning both of these distinguishing features was preserved in the nonspeech context. Nevertheless, the context sounds are not recognizable as speech (they sound like warbles or simple chords) and, thus, contain no perceivable phonemic content. The effects of this nonspeech context on target syllable identification will be compared to synthesized speech contexts. Thus, there will be a comparison between contexts that differ in spectral and phonemic content (speech /al/ versus /ar/) and contexts that differ in spectral content alone (nonspeech tone dyads modeling /al/ versus /ar/).

2.1 Method

2.1.1 Participants

Seventeen undergraduate psychology students volunteered to participate for course credit. All participants were native English speakers who reported no hearing deficits.

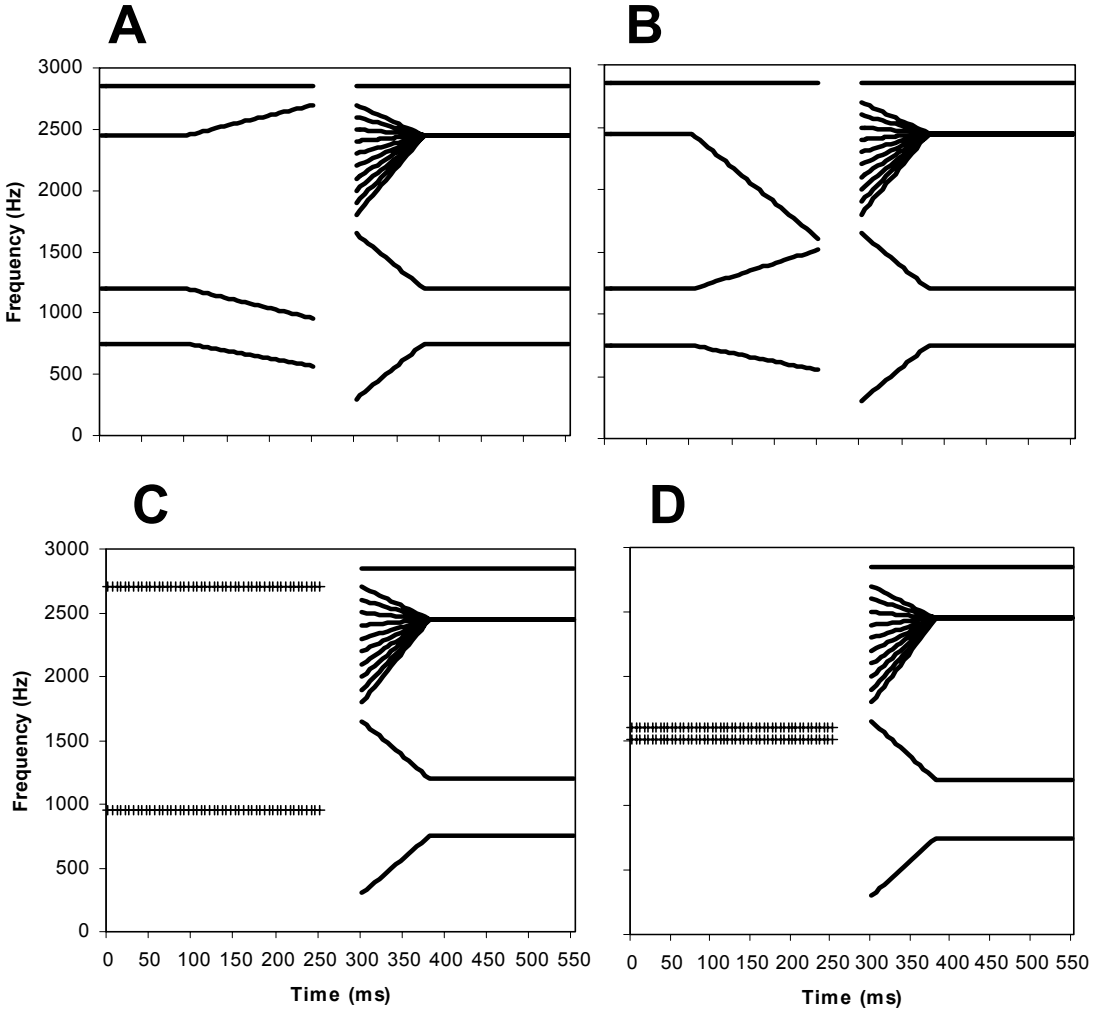

Figure 1. Pseudospectrograms of Experiment 1 stimuli. In each panel, /ga/-/da/ series members are preceded by context stimuli. Panel A, /al/; Panel B, /ar/; Panel C, tones modeling /al/; Panel D, tones modeling /ar/. Solid lines indicate full-formant speech stimuli whereas stippled lines indicate sine-wave tones. F3 transitions for each of the 10 /ga/-/da/ series members are displayed on the same graph for convenience.

2.1.2 Stimuli

A 10-member series of synthetic speech varying acoustically in F3 onset frequency and varying perceptually from /ga/ to /da/ was created using the cascade branch of the Klatt (1980) synthesizer. All synthesized speech tokens were based on acoustic measurements of natural productions by the same

adult male. These stimuli were similar to those used in Lotto and Kluender (1998; Experiment 2). The onset frequency of F3 varied from 1800 to 2700 Hz in 100-Hz steps. From onset, F3 frequency changed linearly to a steady-state value of 2450 Hz across 80 ms. All other synthesis parameters were constant across series members. The first formant frequency (F1) increased linearly from 300 to 750 Hz and F2 frequency declined from 1650 to 1200 Hz across 80 ms. The fourth formant had a steady-state value of 2850 Hz. Fundamental frequency (*f*0) was 110 Hz over the first 200 ms and decreased to 95 Hz over the last 50 ms. Total stimulus duration was 250 ms.

Four additional stimuli were created to serve as precursor context to the /ga/-/da/ series members. Two of these stimuli were synthesized using the Klatt (1980) speech synthesizer and were similar to the /al/ and /ar/ precursor stimuli used by Lotto and Kluender (1998). The first 100 ms of each syllable was identical, with steady-state formant frequencies of 750, 1200, 2450, and 2850 for the first four formants, respectively. After this 100-ms vowel, each stimulus had a 150-ms linear formant transition. Offset frequencies for the first four formants of the /al/ syllable were 564, 956, 2700, and 2850 Hz, respectively. For the /ar/ syllable, formant frequencies changed linearly to arrive at 549, 1517, 1600, and 2850 Hz at stimulus offset. Each syllable was 250-ms long and had a constant *f*0 of 110 Hz. Pseudospectrograms of /al/ and /ar/ preceding /ga/-/da/ series members are shown in Figure 1a and 1b.

Two additional precursor stimuli were modeled after the salient spectral characteristics differentiating the speech (/al/ and /ar/) context stimuli. These nonspeech stimuli consisted of steady-state sinusoids at the offset frequency of F2 and F3 for /al/ and /ar/. For each of these nonspeech stimuli, the tone modeling F2 and the tone modeling F3 were created separately, equated in RMS energy, and added in phase. A 5-ms linear amplitude ramp was applied to the beginning and end of each 250-ms tone dyad. Finally, total energy of the tone combinations were matched to the RMS energy of the /al/ and /ar/ syllables. Pseudospectrograms of the stimuli are shown in Figure 1c and 1d. For all contexts, a 50-ms silent gap was placed between the offset of each context stimulus and the onset of each target syllable. Thus, the presented stimulus was 550 ms in duration.

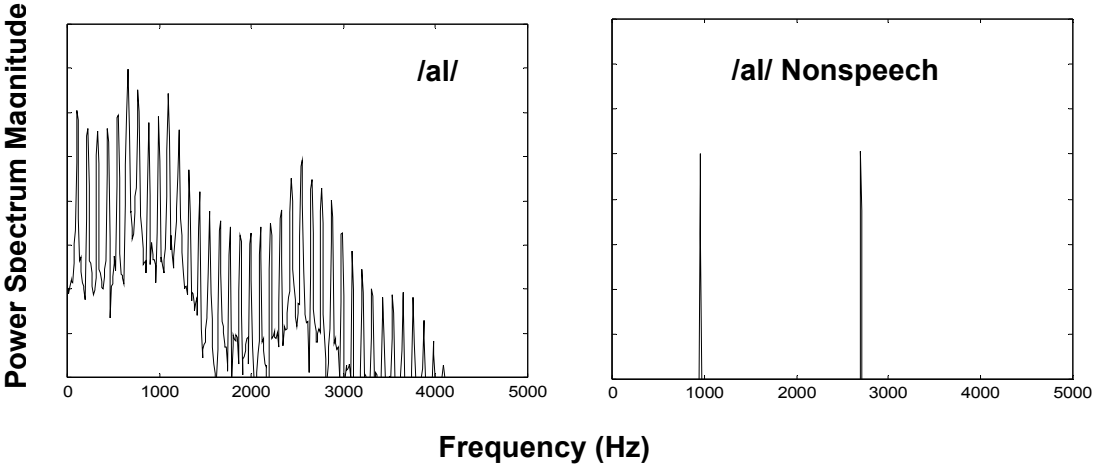

Figure 2. The top panel depicts a fast-Fourier transform (FFT) power spectrum of a single slice in time at the end of the consonant /l/. In the bottom panel, the corresponding tone dyad modeling its F2 and F3 frequency is shown. Notice especially the fine harmonic spectrum of the speech (with energy at each of the multiples of the fundamental frequency, 110 Hz) versus the concentration of energy solely in the regions of F2 and F3 for the nonspeech analog.

It is important to note that the nonspeech precursors, although modeling distinguishing spectral characteristics of the /al/ and /ar/ syllables, are acoustically quite different from speech and are not perceived as speech-like. Figure 1 blurs this distinction by presenting only frequency trajectories of the most intense regions of stimulus spectra. However, speech syllables possess a rich harmonic structure, with energy at each multiple of the fundamental frequency (*f*0=110 Hz). In contrast, the tone pairs

possess energy only at their nominal frequencies (e.g., 956 and 2700 Hz for the /al/ tones), with no fine harmonic structure. This distinction, concealed by Figure 1, is clear in Figure 2, which presents a slice of the spectrum at a single moment in time for /al/ and its nonspeech counterpart. Here the spectral distinctions between speech and nonspeech are quite apparent. Despite these acoustic differences, however, the very simple tone combinations capture some of the putatively important spectral energy in the region of /al/ and /ar/ F2 and F3.

All stimuli were synthesized with 16-bit resolution at a 20-kHz sampling rate and stored on computer disk following synthesis. Stimulus presentation was under the control of a microcomputer and Tucker Davis Technologies (TDT) hardware. Following D/A conversion (TDT, DD1), stimuli were low-pass filtered at a 9.8-kHz cutoff frequency (TDT, FTG2), amplified (TDT, PA4), and presented over headphones (Sennheiser HD285) at 75 dB.

2.1.3 Procedure

Listeners heard the /ga/-/da/ series members preceded by speech (/al/ and /ar/) and nonspeech (tone dyads modeling /al/ and /ar/) in separate blocks. Presentation order of blocks was counterbalanced across participants. Within a block, each of the 20 stimuli (2 precursor stimuli x 10 /ga/-/da/ series members) was presented 10 times in a randomized order. Listeners responded to stimuli in a two-alternative forced-choice (2AFC) identification task, identifying the target syllable as /ga/ or /da/ by pressing labeled buttons on an electronic response box. Each block took approximately 11 to 13 minutes to complete.

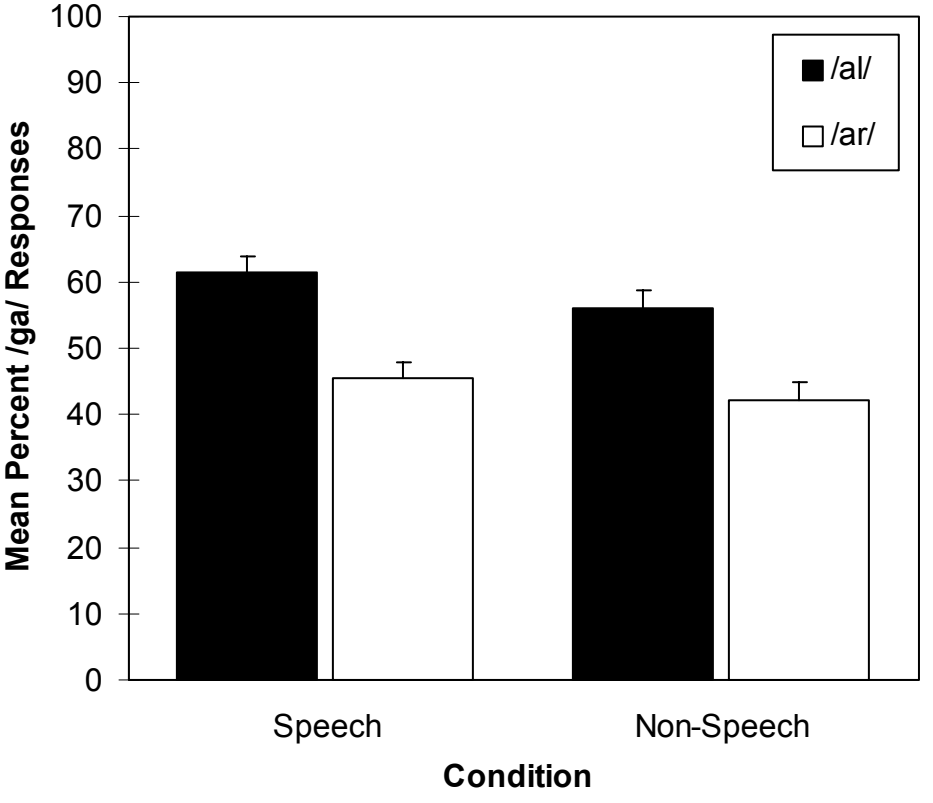

Figure 3. Average percent /ga/ responses as a function of preceding context. Dark bars illustrate responses to /al/ and its nonspeech analog. Light bars correspond to /ar/ and the tones that mimic it. Error bars represent the standard error.

2.2 Results and Discussion

Data from three listeners were withheld from data analyses because they failed to correctly identify 80% of the series endpoints. Average percent "ga" identifications as a function of context for

the remaining 14 listeners are presented in Figure 3. Data were submitted to a 2 x 2 x 10 (Speech vs. Tone Condition, Preceding Context [e.g., /al/ vs. /ar/], Consonant-Vowel Series Member) analysis of variance (ANOVA). Overall, there was a robust effect of preceding context on listeners' consonant labeling ($F(1,13) = 47.94$, $p < .0001$). That is, across speech and nonspeech conditions, precursors influenced /ga/-/da/ identification. As has been reported in previous experiments (Mann, 1980; Lotto & Kluender, 1998), there was a strong effect of a preceding /al/ or /ar/ on identification of /ga/-/da/ series members. Listeners were more likely to label stimuli as /ga/ when preceded by /al/. The same stimuli were more often identified as /da/ when preceded by /ar/.

Like the speech syllables they modeled, the tone dyad stimuli produced an effect of context on consonant identification. Moreover, this effect was in the same direction as for the speech syllables they modeled. When syllables were preceded by tone pairs mimicking the offset frequencies of /al/ F2 and F3, they were more often labeled as /ga/. Preceded by tones that caricatured /ar/, the same syllables were more often identified as /da/. The influence of speech and the influence of nonspeech on listeners' consonant identification were statistically indistinguishable (16% and 14% identification shift for speech and nonspeech, respectively), with no Condition x Context interaction ($F(1,13) <1$). That is, the inclusion of phonemic content in the context did not result in a larger shift in identification function for the target sound.

These results suggest that it is the spectral content of the context that drives this particular phonetic context effect with little indication of a role for phonemic content. When asked to describe the nonspeech context sounds, subjects refer to them as "computer sounds", "warbles" or "buzzes". No subject has ever referred to them as speech sounds. In spite of this, these context sounds effect a shift in the identification of speech sounds.

The use of the nonspeech analogs allows one to match spectral content (to a particular degree) while manipulating phonemic content. It can also be used to examine the effects of maintaining phonemic content while changing the spectral content of the context. In the field of speech perception, researchers have often examined the role of phonemic information by convincing listeners to perceive nonspeech sounds as speech (e.g., Best, Morrongiello, & Robson, 1981; Remez, Rubin, Pisoni, & Carrell, 1981). These manipulations are performed with the notion that listeners may only make use of special speech processes when they explicitly perceive the signal as speech. Experiment 2 applies an empirical manipulation in this spirit to further examine the roles of spectral and phonemic content in phonetic context effects.

3. Experiment 2

Experiment 2 exploits the context effect of Experiment 1 to address the relative roles of spectral and phonemic content of context and their possible interaction. This experiment closely models Experiment 1, with identical stimuli and similar methods. However, prior to participating in an identification task to assess the phonetic context effect, listeners were trained to label the nonspeech precursors as /al/ or /ar/. That is, the nonspeech stimuli were imbued with phonemic content through the use of a training procedure. One group of listeners was trained in a manner consistent with the spectral characteristics of the nonspeech stimuli, labeling tones that model /al/ as "al" and tones that model /ar/ as "ar". Another group of listeners labeled stimuli in an inconsistent manner, with labels assigned in the opposite way. Thus, the spectral content of the signal that was assigned the /al/ label was different for the two groups. In this way, phonemic content is maintained while spectral content shifts.

During an identification test very similar to that of Experiment 1, listeners explicitly labeled the nonspeech context as "al" or "ar" in a manner that reflected their training and then identified the following target syllable as "ga" or "da". Across consistent and inconsistent labeling conditions, the spectral energy of nonspeech precursors was held constant, only the phonemic labels associated with precursors varied. Thus, if phonemic labeling plays a role in the perceptual interactions responsible for the phonetic context effect of Experiment 1, listeners in the two training groups should differ in their demonstration of the effect. However, if spectral content is the fundamental determinant of the context

effect, then the results observed in Experiment 1 should persist in the present test, irrespective of pre-test training to phonemically label the nonspeech stimuli.

3.1 Method

3.1.1 Participants

Fifteen undergraduate psychology students participated in return for course credit. All students were native English speakers and reported normal hearing.

3.1.2 Stimuli

The stimuli for this experiment were identical to those used in the nonspeech condition in Experiment 1 (Figure 1c and 1d). Two two-tone nonspeech stimuli served as precursors to a synthetic speech series that varied perceptually from /ga/ to /da/.

3.1.3 Procedure

3.1.3.1 Training

The subjects were told that they were part of a study to examine the perception of degraded speech sounds. They were informed that they would be learning to label actual productions of the syllables "all" and "are" that had been degraded by computer processing. The training session consisted of 60 trials. During training, listeners heard 30 repetitions of each of the two-tone precursor stimuli. On each presentation, listeners labeled the nonspeech stimulus as /al/ or /ar/ by pressing labeled buttons on an electronic response box. After responding, listeners received feedback via a light above the "correct" response button. Listeners in the *Consistent Label* condition were given feedback that was congruent with the speech token on which the nonspeech stimulus was modeled. For example, the light above the "al" button lit following presentation of the nonspeech stimulus with tones situated at the F2 and F3 offset frequencies of /al/. Listeners in the *Inconsistent Label* condition received the opposite feedback; presentation of the nonspeech stimulus modeling /al/ led to feedback that "ar" was the correct response. Assignment of listeners to the conditions was random. During training, each stimulus could occur at most twice in a row. Training took approximately 4-6 minutes. Assignment of phonemic labels to the nonspeech stimuli was rather easy for listeners. Most participants learned the mapping (whether consistent or inconsistent) in several trials and performed perfectly across the remaining training trials.

3.1.3.2 Identification Task

Immediately following training, listeners participated in two blocks of identification trials that resembled the nonspeech condition of Experiment 1. In each block, listeners heard the two nonspeech precursors precede each of the 10 /ga/-/da/ series members 6 times for a total of 120 trials per block (12 total responses to each stimulus). On each trial, listeners heard the entire stimulus complex consisting of nonspeech precursor followed by CV stimulus and then labeled the nonspeech precursor as "al" or "ar" using labeled buttons on an electronic response box. Listeners then identified the target syllable as "ga" or "da" using two other labeled buttons on the same response box. The task took approximately 7-9 minutes per block.

3.2 Results

Two listeners were eliminated from data analyses because they failed to identify endpoint /ga/-/da/ series members correctly 80% of the time. Data from a third listener were kept out of the analyses because the listener consistently identified the context stimuli incorrectly. Of the remaining listeners, 6 participated in the *Consistent Label* condition and 6 participated in the *Inconsistent Label* condition. Listeners readily learned to label two-tone stimuli as /al/ and /ar/ and were very accurate in their identifications. Average mean correct identification for all training trials was 89.69% for the *Consistent Label* condition and 87.02% for the *Inconsistent Label* condition. There were no significant differences between the conditions in labeling accuracy ($t(10)=0.25$, $p=0.81$).

The question most germane to the present experiment is whether training listeners to phonemically label the tone pairs consistently versus inconsistently (with respect to the spectral content) influenced the context effect. If phonemic content of the context is important for determining the context effect, then one would expect that more "ga" responses should be obtained for contexts labeled "al" in both groups. This would show up as an interaction between the labeling condition (*Consistent* versus *Inconsistent*) and spectral content (/al/-modeled tones versus /ar/-modeled tones). On the other hand, if spectral content was the only determining factor for the context effects then the nonspeech stimulus

derived from acoustic characteristics of /al/ should lead to more "ga" responses regardless of the labeling condition (Spectral Content main effect, no interaction).

Average percent "ga" identifications as a function of context for the 12 listeners are presented in Figure 4. A Spectral Content (/al/-modeled tones vs. /ar/-modeled tones) x Labeling Condition (Consistent vs. Inconsistent) x Target Syllable (10 members of the CV series) mixed ANOVA revealed a significant effect of Spectral Content on the labeling of the target syllables ($F(1,11) = 6.57$, $p<.05$). Following tones modeled after spectral properties of /al/, more /ga/ responses were obtained. The interpretation of this main effect is fairly straightforward because the interaction between Spectral Content and Labeling Condition was not significant ($F(1,11)<1$). The context effect was determined by the spectral qualities of the preceding context regardless of labeling condition. These results further support the conclusion from Experiment 1 that phonetic context effects are driven by the spectral content of the context with little role for the phonemic content.

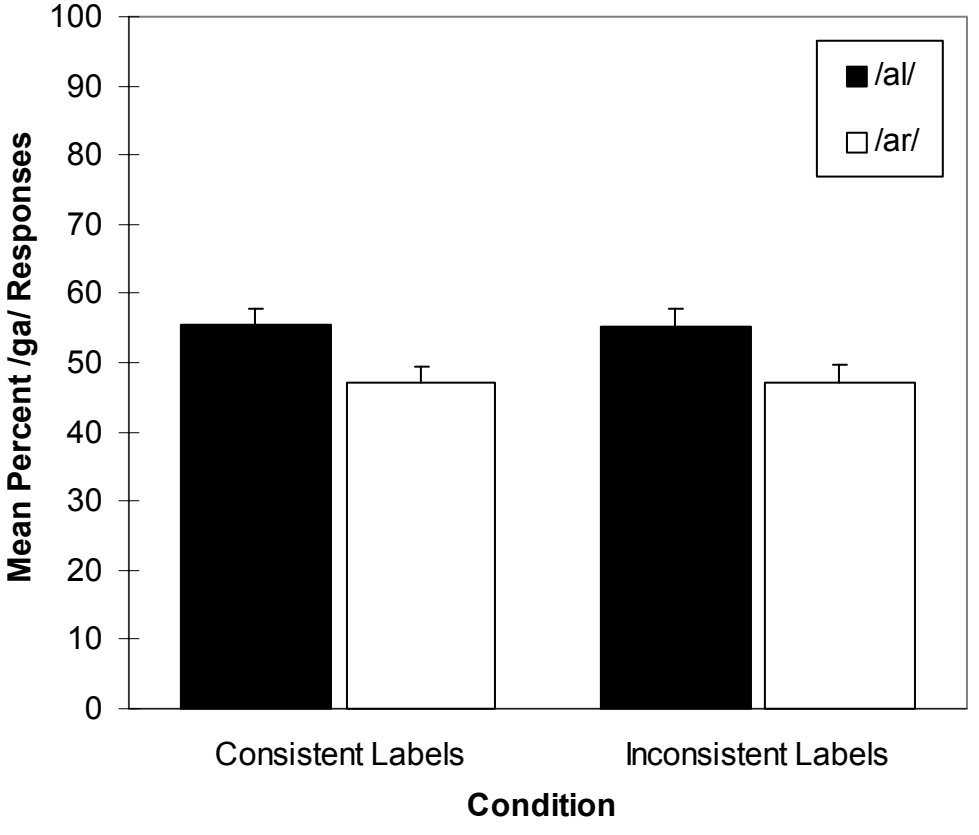

Figure 4. Average percent /ga/ responses as a function of preceding context and labeling condition. Dark bars illustrate responses when tones matching the spectral properties of /al/ preceded the syllables (irrespective of label). Light bars show responses to tones modeling /ar/. Listeners in *Consistent Labels* condition received training to label the sounds consistent with the speech after which they were modeled. Listeners in *Inconsistent Labels* condition received training to label the sounds in the complementary inconsistent manner. Error bars represent the standard error.

4. General Discussion

The two experiments reported here examined the respective roles of spectral and phonemic content of contexts in a classic phonetic context effect. In Experiment 1, the phonemic content was manipulated by creating speech and nonspeech contexts that are similar in some spectral characteristics but vary to the extent that they are perceived as speech. These two types of context led

to equivalent shifts in the identification of subsequent /ga/-/da/ stimuli despite the fact that the nonspeech context led to no explicit perception of phonemic content. These results agree with previous demonstrations of speech identification shifts caused by nonspeech contexts (Diehl, 1976; Kat & Samuel, 1984; Lotto & Kluender, 1998; Holt *et al*., 1998, 2000). In the present case, there was no difference in the size of the identification shift between the two kinds of context. This leaves little room for a role for phonemic content in these context effect demonstrations.

In Experiment 2, listeners were trained to identify the nonspeech signals with speech labels. This technique has been used previously to direct listeners to use specialized speech processes or phonemic information. The results indicated that the labels used by the listeners for the contexts had no effect on the resulting shift in target syllable identification. Again, there was no evidence for a strong role for phonemic content in the context effect.

Taken together, the results of these two experiments indicate a substantial role for spectral content in phonetic context effects. These effects do not appear to require phonemic labels for the context sounds. Thus, even listeners who fail to apply phonemic labels to context speech sounds should exhibit effects of context on perception of subsequent sounds. In agreement with this prediction, Fowler, Best and McRoberts (1990) found shifts in the discrimination responses of 4-month-olds to /ga/-/da/ stimuli that were preceded with /al/ or /ar/. Lotto, Kluender and Holt (1997) trained Japanese quail (*Coturnix japonica*) to "label" /ga/-/da/ syllables by pecking a button. The birds' labeling functions shifted in the predicted manner when the syllables were preceded by /al/ and /ar/. Mann (1986) examined the same phonetic context effect for Japanese speakers who could not discriminate /l/ and /r/. For native Japanese speakers, the liquids are perceived as members of the same phoneme. Nevertheless, the /ga/-/da/ identification functions for these listeners exhibited the same context-dependent shift obtained from native English speakers. The results of these three studies are predictable if the context effects are elicited by spectral content as opposed to phonemic content. In each case, the context sounds would presumably not be perceptually labeled as different phonemes. Nevertheless, the different sounds led to predictable shifts in responses to the target stimuli. Given the nature of the subjects involved, it may also be concluded from the studies that these effects do not require substantial experience with contextual dependencies or statistical regularities of a particular language.

4.1 Implications for Cognitive Models

As discussed in the introduction, the TRACE model (Elman & McClelland, 1986 McClelland & Elman, 1986) explains phonetic context effects by appealing to the phonemic content of the context. The results of the two experiments presented here call into question the plausibility of this approach. Context stimuli that contained no perceivable phonemic content led to equivalent shifts in identification of target stimuli.

By appealing to discrete linguistic representations such as phonemes as the causal agents in phonetic context effects, these models are unable to account for the interactions due to the continuous spectral content. One way to deal with this problem is by allowing interactions between spectral representations in the model. For example, in models like TRACE one could add links between nodes at the feature level across time slices. That is, the activation of a feature node in one time slice could modulate the activity feature nodes in surrounding time slices. Of course, to account for the data from the experiments presented here, the feature nodes would have to represent general acoustic features. Linguistically-based distinctive feature nodes would fail to account for the data for the nonspeech condition.

Lotto and Kluender (1998) describe the results of speech and nonspeech context effects as examples of *frequency contrast* (or *spectral contrast*). One can describe the results of the speech and nonspeech contrast effects in the following way: Subsequent to *high*-frequency energy (high-frequency F3 of /al/ or high-frequency tone), target syllables are labeled with more *low*-frequency responses (/ga/ with its lower F3 frequency). In order to mimic these effects, *high*-frequency nodes (associated with energy around F3) that are activated during the presentation of /al/ or the nonspeech /al/-analog would send activation to *low*-frequency nodes in subsequent time slices (or, alternatively, inhibit high-frequency nodes in subsequent time slices). This modulation at the feature level would lead to shifts in the identity of the most-activated node at the phoneme level. These results could also be accomplished with recurrent networks (Norris, 1993) at the level of spectral representation.

Regardless of the mechanism, the evidence supports a role for spectral content in phonetic context effects. However, there is evidence from several studies that shifts in target sound identification can occur with no change in the acoustics of neighboring context. Elman and McClelland (1988; see also, Pitt & McQueen, 1998) ran a series of studies in which the perceived phonemic label for the context was changed despite constant spectral content. They presented listeners with a series of syllables varying from a good version of "dates" to a good version of "gates" preceded by words ending in /s/ (e.g., "ridiculous") or /S/ (e.g., "Spanish"). These context sounds have been demonstrated to be effective at eliciting identification boundary shifts in /d/-/g/ series (Mann & Repp, 1981). After replicating this result, Elman and McClelland created an ambiguous fricative midway (acoustically) between /s/ and /S/. This sound was appended to edited versions of the context words (e.g., "ridiculou" and "Spani"). The word frames resulted in changes in the perceived phonemic label (e.g. "ridiculous" and "Spanish") despite no changes in the acoustics of the fricative. As a result, phonemic content changes while spectral content is maintained. When these edited words were presented as context before the "dates"-"gates" series, there was a significant shift in the identification function that was similar (though smaller) to that obtained for the original unambiguous fricatives

How can one accommodate the results of these studies demonstrating phonetic context effects resulting from little or no change in the acoustics of the context with the results of the current study demonstrating a significant role for spectral content in these same effects? One possibility is that there are two independent levels of context influence on target identity. It may be that both spectral content and phonemic content play a role in the context effects. In natural speech, these levels would be in general agreement. However, in the laboratory, we can manipulate the signal such that spectral changes in context are not accompanied by phonemic changes in context (Experiment 1 and 2 of the present report; Mann, 1986) or we can change phonemic content and maintain spectral content (Elman & McClelland, 1988; Fowler *et al*., 2000). Changes present at one of these levels could be sufficient to determine a change in target sound identification. It is possible that these levels of effect map onto the information processing levels II (integrative acoustic) and III (phonetic) proposed by Samuel and Kat (1996).

Another provocative possibility is that changes in the perceived phonemic content result in changes in the effective spectral content. Elman and McClelland (1988) account for their phonetic context effects in the TRACE model by presuming that word level representations interact with (or "feed down" to) phonemic representations, which in turn modulate the weights between feature and phoneme representations in subsequent time slices. For example, the stem "ridiculou" activates the "ridiculous" word node, which activates the /s/ node, which leads to a biasing of the identification of following /d/-/g/ sounds. Given this interactive mechanism, it is a logical step to suggest that the activation of a phoneme could feedback and modulate the spectral representation of the sound in the same time slice. For example, the stem "ridiculou" activates the "ridiculous" node, which activates the /s/ node, which changes the spectral representation of the fricative by increasing the center frequency of the fricative noise (one of the acoustic distinctions between /s/ and /S/). This spectral representation would then effect a change in the processing of subsequent spectral information in a contrastive manner. In this proposal, the perceptual interactions would occur at the general spectral level. Thus, the mechanism for the context shifts would be the same for speech contexts in humans and birds and for nonspeech contexts.

There is considerable debate whether interactive models such as TRACE are the best way to account for speech perception phenomena (Massaro, 1989; Samuel, 1997, 2001; Pitt & McQueen, 1998; Norris, McQueen, & Cutler, 2000). Some context effects can be effectively explained by purely bottom-up or "autonomous" models that include recurrent feedback within each level (Norris, 1994; Norris *et al*., 2000). The current data probably can be handled with the same types of mechanisms at the spectral level. As a result, the current experiments do not decide these major modeling issues. However, the results *do* broaden the discussion to include roles for more "low-level" representations that are closer in structure to the acoustic input. Cognitive models of speech perception often ignore realistic spectral representations in favor of questionable discrete features or basic phoneme representations. Any complete model of speech perception will have to respect the general characteristics of the auditory system and the representation of spectral patterns in order to account for all effects of context on speech sound identification.

5. Author Note

The author would like to thank Pam Mueller and Amina Habib for their assistance in collecting the data and to Lori Holt for helpful comments and assistance in preparing this manuscript. This research was supported in part by the Parmly Hearing Institute. Address correspondence and reprint requests to A. J. Lotto, Boys Town National Research Hospital, 555 North 30th Street, Omaha, NE 68131. E-mail: lottoa@boystown.org.

References

Best, C. T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception & Psychophysics, 29, 191-211.

Borsky, S., Tuller, B., & Shapiro, L. P. (1998). "How to milk a coat:" The effects of semantic and acoustic information on phoneme categorization. Journal of the Acoustical Society of America, 103, 2670-2676.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. Journal of Memory and Language, 26, 527-538.

Diehl, R. L. (1976). Feature analyzers for the phonetic dimension stop versus continuant. Perception & Psychophysics, 19, 267-272.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and variability in speech processes (pp. 360-385). Hillsdale, NJ: Erlbaum.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. Journal of Memory and Language, 27, 143-165.

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.

Ganong III, W. F. (1980). Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance, 6, 110-125.

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

Holt, L. L., & Lotto, A. J. (2001). Auditory context effects for nonequivalent sources. Journal of the Acoustical Society of America, 109, 2311.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (1998). Spectral contrast in perception of VCV syllables. Journal of the Acoustical Society of America, 104, 1759.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. Journal of the Acoustical Society of America, 108, 710-722.

Kat, D., & Samuel, A. G. (1984). More adaptation of speech by nonspeech. Journal of Experimental Psychology: Human Perception and Performance, 10, 512-525.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 971-995.

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. Journal of the Acoustical Society of America, 42, 830-843.

Lotto. A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. Perception & Psychophysics, 60, 602-619.

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

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., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. Journal of the Acoustical Society of America, 69, 548-558.

Massaro, D. W. (1989). Testing between the TRACE model and the Fuzzy Logical Model of Speech Perception. Cognitive Psychology, 21, 398-421.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. Cognitive Psychology, 18, 1-86.

Norris, D. G. (1993). Bottom-up connectionist models of "interaction". In G. Altmann & R. Shillcock (Eds.), Cognitive models of speech processing: The second Sperlonga Meeting (pp. 211-234). Hillsdale, NJ: Erlbaum.

Norris, D. G. (1994). Shortlist: A connectionist model of continuous speech recognition. Cognition, 52, 189-234.

Norris, D. G., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. Behavioral and Brain Sciences, 23, 299-370.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? Journal of Memory and Language, 39, 347-370.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. Science, 212, 947-950.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

Samuel, A. G. (1997). Lexical activation produces potent phonemic precepts. Cognitive Psychology, 32, 97-127.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. Psychological Science, 12, 348-351.

Samuel, A. G., & Kat, D. (1996). Early levels of analysis of speech. Journal of Experimental Psychology: Human Perception and Performance, 22, 676-694.

Proceedings of the 2003 Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception

edited by Augustine Agwuele, Willis Warren, and Sang-Hoon Park

Cascadilla Proceedings Project Somerville, MA 2004

Copyright information

ISBN 1-57473-402-4 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Lotto, Andrew J. 2004. Perceptual Compensation for Coarticulation as a General Auditory Process. In *Proceedings of the 2003 Texas Linguistics Society Conference*, ed. Augustine Agwuele et al., 42-53. Somerville, MA: Cascadilla Proceedings Project.

or:

Lotto, Andrew J. 2004. Perceptual Compensation for Coarticulation as a General Auditory Process. In *Proceedings of the 2003 Texas Linguistics Society Conference*, ed. Augustine Agwuele et al., 42-53. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #1066.