Ipsilateral and contralateral phonetic context effects

Kevin Sitek and Keith Johnson
San Francisco VA Medical Center and UC Berkeley

Abstract
Duplex perception experiments explore how a speech unit is processed simultaneously as a speech and non-speech sound. Two adjacent speech sounds do not need to be heard in the same ear for phonetic context effects to be active. Our experiments combine these paradigms by presenting two syllables dichotically with formants of the target syllable presented contralaterally to the rest of the target. When hearing an ambiguous stimulus between /d/ and /g/, subjects are more likely to hear /g/ after /al/ and /d/ after /ar/. This context effect is significant regardless of whether the context is presented to the same ear as the base target or the isolated third formant transition, demonstrating that the auditory system must combine streams before completing all phonetic analysis. Additionally, there is a greater context effect when the formant transitions are presented to the ipsilateral ear as the context segment than when they are presented contralaterally. The difference in compensation effects between Experiments 1 and 2 (where the formant transition is presented contralaterally to the context) shows that some phonetic processing occurs before the left and right auditory streams converge. Separating the formant chirp (responsible for cuing /d/ vs. /g/) from the context is more crucial to analysis than separating the base target from the context. Phonetic context effects, while working adequately across streams, are more influential within one stream.

Key words: Speech perception; duplex perception; context effect

1. Introduction

The human auditory system is especially tuned to distinguish subtle differences between speech sounds, making communication of thoughts, ideas, and intentions possible. Yet it has also evolved to combine certain sounds that are likely to have come from the same source. Duplex perception experiments are able to test listeners' abilities to combine distinct sounds into speech and non-speech elements, which are perceived concurrently (Rand 1974; Mann & Liberman 1983).

The auditory system uses context to determine relevant signals and information. Speech perception has numerous examples of context effects, using semantic, lexical, and phonetic information to help process the speech signal. Phonetic context effects allow an ambiguous speech sound to be perceived as two different phonemes depending on the neighboring phonemes (Mann & Repp 1980; Mann 1980).

But when and where in the auditory pathway do adjacent sounds first influence each other? Duplex perception tasks cannot work unless a listener is able to combine data from the left and right auditory channels, which only occurs once the auditory signals have reached the brain. By combining duplex perception and phonetic context effect experimental paradigms, we can discover whether neighboring phonemes begin influencing each other early in the peripheral auditory pathway or later in the central processing areas. In this paper we will look at phonetic

context effects by exploiting duplex perception to test if phonetic analysis begins in early, peripheral levels of acoustic analysis or in later, more central levels.

1.1 Duplex perception

Duplex perception describes the phenomenon where a speech sound can simultaneously be processed as a speech (phonetic) sound and as a non-speech (auditory) sound. A typical duplex perception experiment has a three-formant "base" CV syllable presented to, say, the left ear that is not quite complete; in the opposite (in this case, the right) ear is presented a "chirp" sound that includes the missing formants of the left ear input (Rand 1974). For example, the syllable /ba/ is presented to the left, but without the F2 and F3 of the consonant-vowel transition, which will instead be presented to the right. Participants report simultaneously hearing a completed syllable in the left ear and a non-speech chirp in the right ear (Mann & Liberman 1983).

1.2 Phonetic context effects

An ambiguous speech sound will be identified differently by a listener in various contexts (Mann 1980). For example, a synthetic speech sound that is created to sound like halfway between a /d/ and a /g/ will be perceived as a /g/ if it follows the syllable /al/; if it comes after /ar/, it will be identified as a /d/. Though speech perception theorists agree on the existence of phonetic context effects, their frameworks describe differing underlying mechanisms responsible for the effects. Here we will look at two such mechanisms: compensation for coarticulation, which is taken as support for gestural theories of perception, and spectral contrast, which allows for a general auditory and learning approach to perception.

1.2.1 Compensation for coarticulation

Gestural theories of speech perception describes the phonetic context effect in terms of compensation for coarticulation, which seeks to explain how intended gestures are key to a listener's understanding of the speaker's language. By studying the perception of synthetic and hybrid sounds, Virginia Mann (1980) found that more 'g's were identified when the preceding consonant was /l/ than /r/ and that more 'g's were perceived when the preceding liquid had been originally produced before a /g/—this was more influential for /r/ than for /l/. Mann argued that this second finding exists because stops that follow the lateral liquid /l/ are often more forward than stops that follow the alveolar retroflex /r/. Here, the anterior-posterior boundary between /d/ and /g/ shifts forward after a segment like /l/ that has an anterior place of articulation (and vice versa) (Viswanathan, Fowler, & Magnuson, 2009). Mann's study shows that listeners use experiential knowledge about their language as well as an understanding of the speaker's intended phonetic articulations in order to comprehend the phonetic speech sounds in conversation.

Gestural theories that subscribe to compensation for coarticulation include motor theory (MT) and direct realist theory (DRT). In MT, proposed by Alvin Liberman, speech perception and production are controlled by one *phonetic module* (Liberman et al. 1967; Fodor 1983; Liberman & Mattingly 1985). Additionally, duplex perception has its roots in early MT experimentation. DRT, on the other hand, denies the existence of a "special" speech processing module in the brain (Fowler et al. 1980). Fowler and Rosenblum (1990) found that duplex perception experiments also work for non-speech signals such as slamming doors, so a speech-only brain module is unlikely to exist. However, according to DRT, because a speech listener is also a speaker of the language, a listener intuitively knows that phonemes are often coproduced,

which enables them to tease apart overlapping speech segments, even with ambiguous targets in certain environments.

1.2.2 Spectral contrast

Holt and Lotto (2002) argue that phonetic context effects occur as a result of psychoacoustic and physiological mechanisms that are not specific to human speech but instead are general auditory techniques. Instead of depending on compensation for coarticulation, which assumes that a listener understands the intended gesture by means of knowledge of the oral tract mechanisms for creating speech, Holt and Lotto describe a spectral contrast theory. Holt (1999) and Lotto and Kluender (1998) found that a low-frequency precursor (such as the low F3 in /ar/) results in subjects hearing an ambiguous target as having higher-frequency energy and identifying phonemes accordingly (so a middle F3 in an ambiguous /dga/ syllable would be heard as a high F3 of the syllable /da/). Middle frequencies are perceived as being low when they follow high frequency tones, but they sound high when they come after low frequency tones (Lotto and Holt 2006). Holt and Lotto describe phonetic context effects by means of these spectral contrasts rather than assuming that the acoustic signal carries any information about the articulatory speech gestures used to produce the utterance.

Spectral contrast explains Mann's phonetic context effect findings without necessitating speech-specific mechanisms. Instead, for an ambiguous /dga/ sound that follows the syllable /al/, the high third formant in /l/ makes the middle F3 of /dg/ sound lower *in contrast*—thus, like a /g/.

2. Experiments 1 and 2 (overview)

These experiments aim to test that phonetic analysis does not fully occur until after the left and right auditory streams have been combined into one auditory percept. Mann (1980) showed that perception of stops in synthesized speech depends on the perception of the preceding liquid. Duplex perception experiments have explored how placing the formant transition in the contralateral ear relative to the target allows listeners to correctly combine and identify the target stop consonant (Liberman et al. 1981). Holt and Lotto (2002) demonstrated that the context and target syllables do not need to be in the same ear in order for phonetic context effects to be active. In the following experiments, we combine these paradigms so that the third formant transition is always presented to the ear contralateral to the base target, which is either contralaterally (Experiment 1) or ipsilaterally (Experiment 2) placed relative to the context syllable.

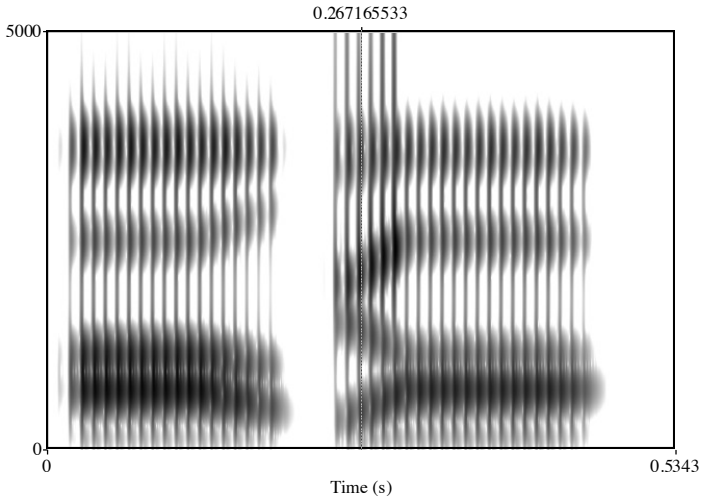

Figure 1. /al/ + /ga/ (token 9) spectrogram.

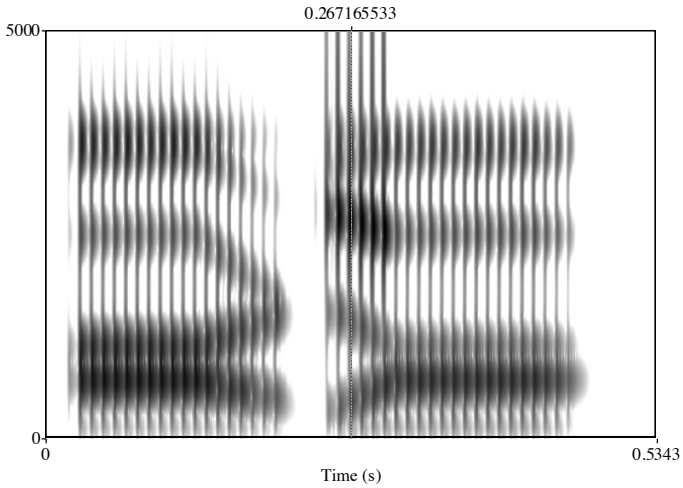

Figure 2. /ar/ + /da/ (token 1) spectrogram.

3. Experiment 1

3.1 Method

3.1.1 Listeners

Twenty-one undergraduates participated in Experiment 1. All listeners were fluent English speakers who reported no language or hearing disabilities. No participants from Experiment 1 participated in Experiment 2.

3.1.2 Stimuli

Experiment 1 and Experiment 2 use the same context, base target, and formant transition segments. These segments were created using the Klatt speech synthesizer (1980) and compiled using SoX (http://sox.sourceforge.net/). All segments have a fundamental frequency (f_0) of 100

Hz. (Unless otherwise noted, all values will be in Hz.) In both Experiment 1 and Experiment 2, the base and chirp begin 20 ms after the context ends (t = 225 ms).

There are two different context segments: /al/ and /ar/. In both segments, the /a/ was created over the first 125 milliseconds with F1 = 700, F2 = 1100, and F3 = 2500. To create /l/, F1 gradually decreased to 450 at t = 200 ms, F2 gradually decreased to 950, and F3 gradually increased to 2900. F4 stayed at 3550. For /r/, F1 fell from 700 to 450, F2 rose from 1100 to 1500, F3 fell from 2500 to 1650, and F4 fell from 3550 to 2900. In both /al/ and /ar/ segments, the amplitude rose from 0 dB at t = 0 ms to 60 dB at t = 25 ms. At t = 165 ms, amplitude began to fall until reaching 0 dB at t = 205 ms.

The base target /a/ was created with the same parameters as the initial segment of the context (F1 = 700, F2 = 1100, F3 = 2500, and F4 = 3550). It begins at t = 225 ms and ends at t = 460 ms.

Nine third formant transitions ("chirps") were used corresponding to the /da-ga/ continuum. All nine rose from silence at t = 0 ms to an amplitude of 60 dB by t = 45 ms and were cut off at t = 80 ms. Additionally, all nine chirps had the same F1, F2, and F4 movements. F1 began at 300 and increased to 700 by t = 75 ms, while F2 began at 1500 and decreased to 1100 by t = 75 ms. F4 was 3400 from t = 0 ms until t = 25 ms; at t = 30 ms it began increasing until reaching 3550 by t = 75 ms.

F3 followed the same temporal contour as F4 (steady through t = 25 ms, gradual change from t = 30 ms to t = 75 ms). The most "da"-like chirp began at 2700, with one while the most "ga"-like started at 2000. Seven additional chirps were along the continuum between these two cardinal starting frequencies at intervals of 83.3 Hz (rounded to the nearest whole number). Depending on the starting frequency, F3 began to either increase or decrease at t = 30 ms until reaching 2500 Hz at t = 75 ms.

In Experiment 1, the context segment and the formant transition ("chirp") are each presented to the left ear while the base (target) segment is presented to the right ear.

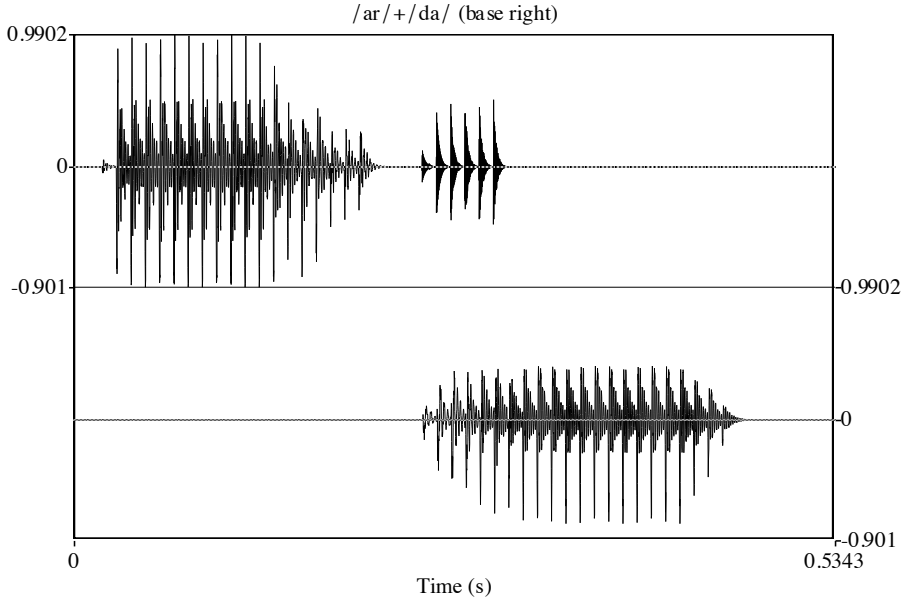

Figure 3. Experiment 1 (base right) waveform. Left channel is presented on top; right channel is on the bottom.

3.1.3 Procedure

Up to three subjects participated in the thirty-minute experiment concurrently. The experiment was divided into two blocks. The first block included the context (either /al/ or /ar/, depending on the trial) in the left ear followed by the formant transition chirp in the left ear and the base target in the right ear. Thus, the formant transition was presented contralaterally to the base target. The second block did not include a context and so only includes the target and the formant transition. Both blocks included five trials of each of the nine tokens of the /da/-/ga/ spectrum for a total of 45 trials per block and 90 trials overall. The entire experiment took approximately fifteen minutes.

3.2 Results

The results of Experiment 1 are shown below ("Base right stimuli"). Across all conditions, the results show a strong effect of trial token on listener response—the 'd' sounds were heard more often as 'd,' and the 'g' sounds were heard as 'g.' Of greatest importance is the clear separation of responses in the /al/ and /ar/ contexts. Critically, by listener, the difference between mean /ar/ and /al/ context responses was significant ($t = -5.7991$, $df = 36.621$, $p < 0.001$). This replicates the phonetic context effect findings of Mann (1980) and others who found that a listener's perception of an ambiguous synthesized phoneme can be predicted based on the synthesized phoneme's preceding context. The clearest evidence for this effect is token 9, the most /ga/-like token. In the /al/ context, all twenty-one listeners perceived token 9 as being /ga/ in each of the five trials in which it appeared, meaning it had a 100% correct response rate. In contrast, token 9 was barely below the /da-ga/ threshold when presented in the /ar/ context.

Additionally, while 'd' responses fell below 50% on token 7 in the /ar/ context, it only took until token 4 in the /al/ context for responses to definitively drop below 50%. Thus, the threshold for perceiving a 'g' in the context of /al/ is much lower than in the context of /ar/.

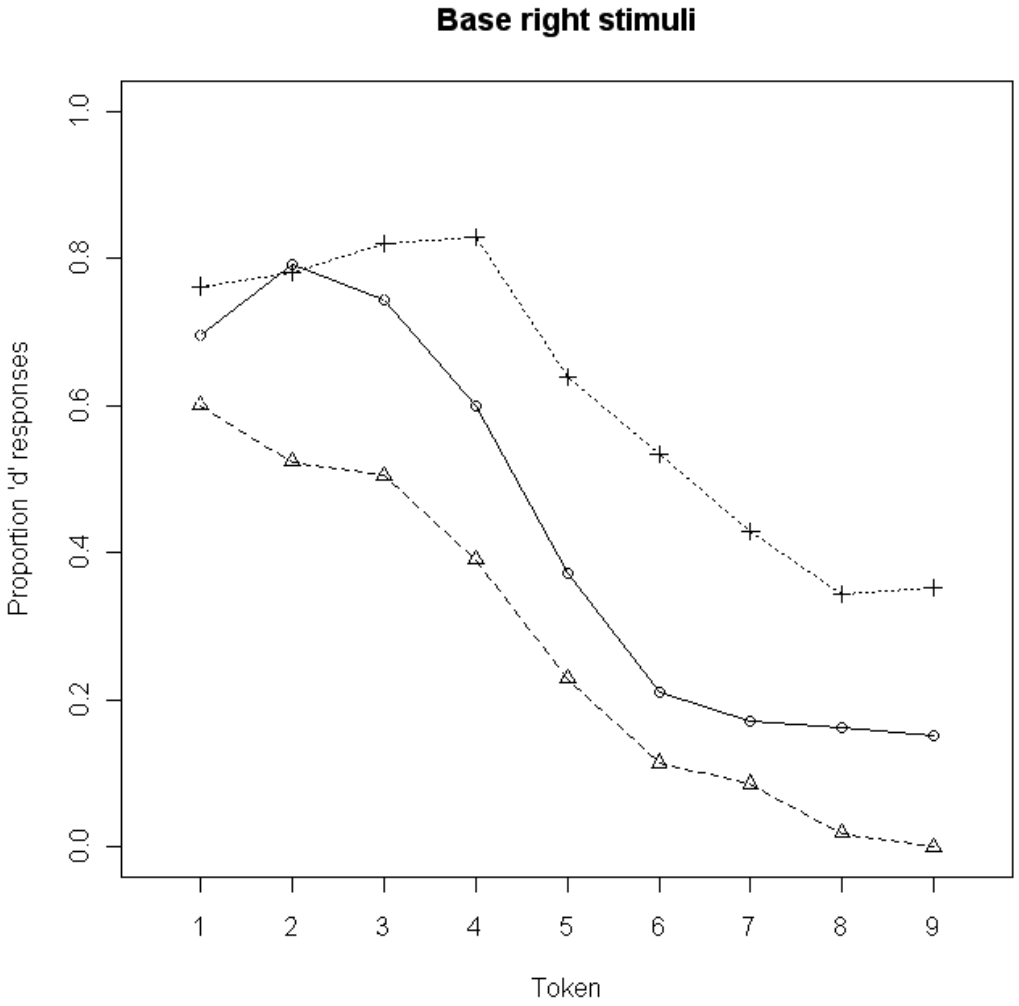

Figure 4. Each graph represents the mean percent 'd' responses. Crosses are /ar/ conditions; triangles are /al/ conditions. Circles represent the null context (block 2) responses of the Experiment 1 listeners.

4. Experiment 2

4.1 Method

4.1.1 Listeners

Twenty-one undergraduates participated in Experiment 2. All listeners were fluent English speakers who reported no language or hearing disabilities. No participants from Experiment 1 participated in Experiment 2.

4.1.2 Stimuli

The context, base (target), and formant transition (chirp) in Experiment 2 are the same as those used in Experiment 1. In Experiment 2, the context and target segments are presented to the same (left) ear while the formant transition is presented to the right ear.

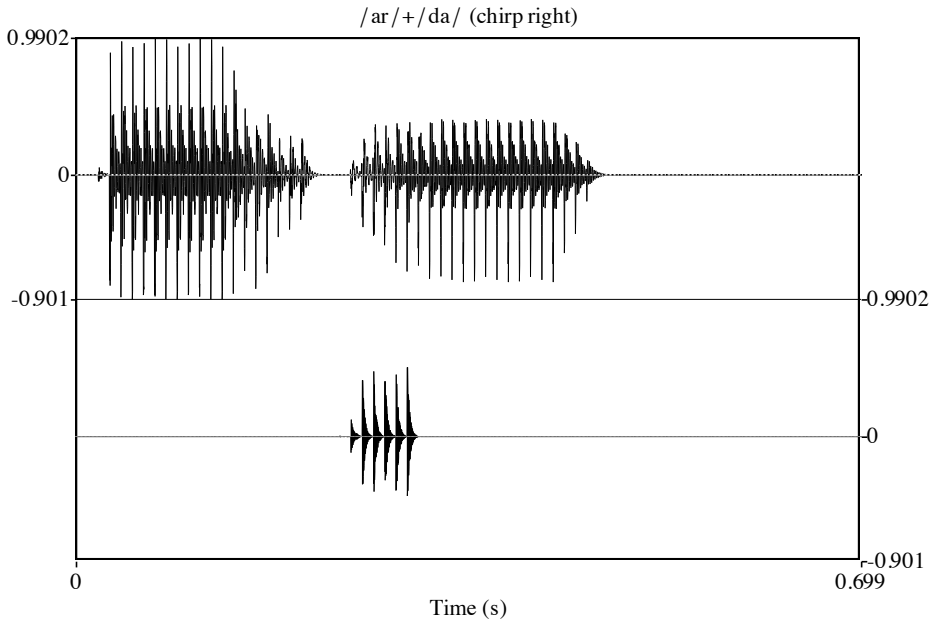

Figure 4. Experiment 2 (chirp right) waveform. Left channel is presented on top; right channel is on the bottom.

4.1.3 Procedure

Experiment 2 mirrors Experiment 1 closely. Again, up to three subjects participated in the thirty-minute experiment concurrently. The experiment was divided into two blocks. The first block included the context (either /al/ or /ar/, depending on the trial) in the left ear, but unlike in Experiment 1, in Experiment 2 the *base target* followed in the left ear while the *formant transition chirp* was presented in the right ear. Thus, the formant transition was presented contralaterally to the base target. The second block did not include a context and so only includes the base target in the left audio channel and the formant transition chirp in the right channel. Both blocks included five trials of each of the nine tokens of the /da/-/ga/ spectrum for a total of 45 trials per block and 90 trials overall. The entire experiment took approximately fifteen minutes.

4.2 Results

As in Experiment 1, all tokens exhibited a greater 'd' response in the /ar/ contexts than in the /al/ contexts. Additionally, comparisons of each listeners' /ar/ context responses vs. /al/ context responses found a significant effect ($t = -2.1718$, $df = 30.229$, $p = 0.0379$). Thus, phonetic context effects are preserved even when the third formant "chirp" is presented contralaterally to both the context and base segments. Unlike the results of Experiment 1, the 'd' responses were below 50% in the /al/ context beginning with token 3. In the /ar/ context, the threshold is also earlier than in Experiment 1—in Experiment 2, subjects perceived a 'g' in the /ar/ context beginning with token 5.

To compare the magnitude of the compensation effect in Experiment 1 to that of Experiment 2, we averaged each listener's responses in each context (/al/ or /ar/). We then subtracted the average /al/ response from the average /ar/ response, giving us the compensation effect for each listener. The average compensation effect was 0.335 in Experiment 1 and 0.148 in Experiment 2; this difference proved to be significant ($t = 2.1105$, $df = 37.593$, $p = 0.0415$).

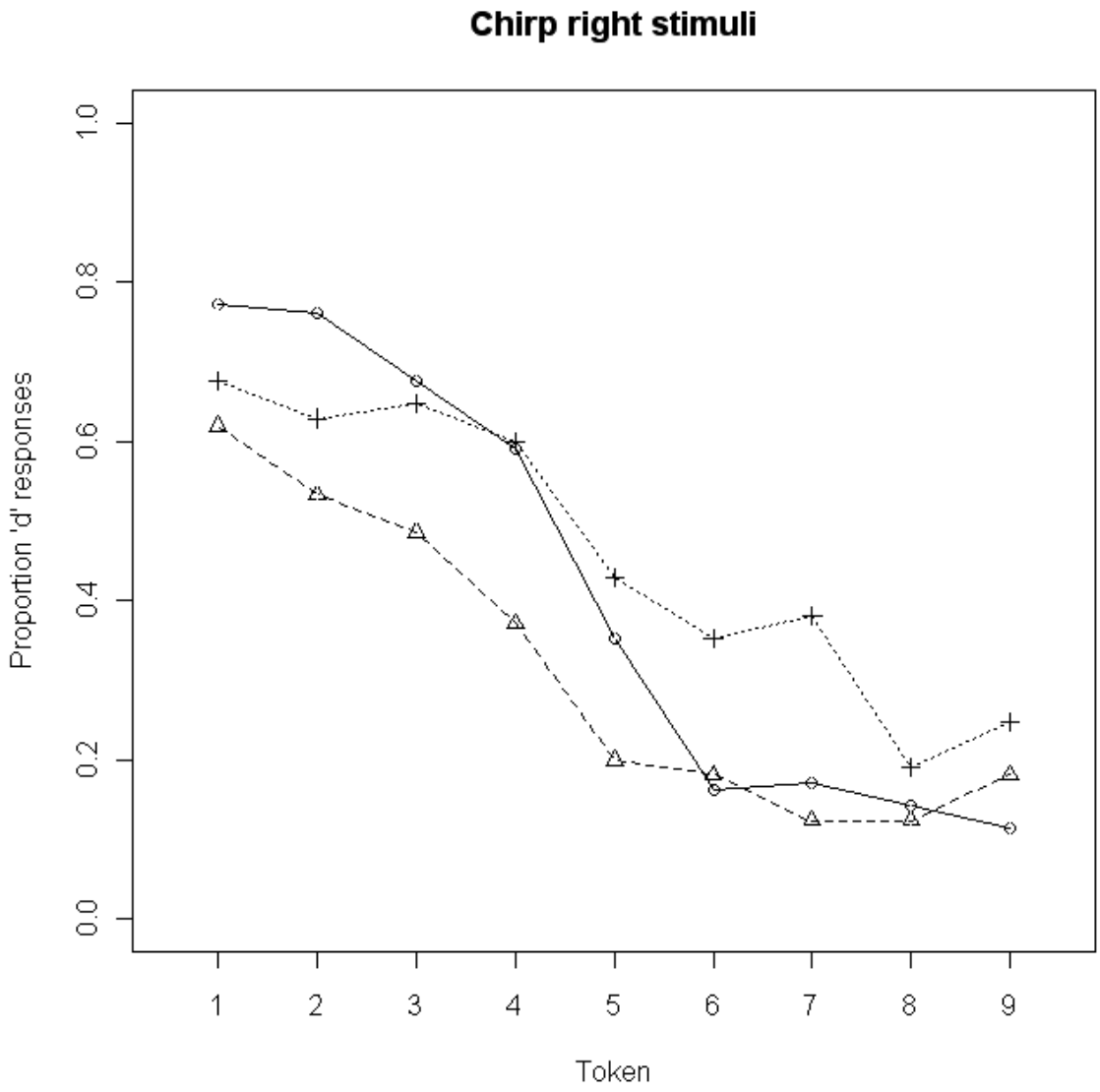

Figure 5. Each graph represents the mean percent 'd' responses. Crosses are /ar/ conditions; triangles are /al/ conditions. Circles represent the null context (block 2) responses of the Experiment 2 listeners.

5. Discussion

Because listeners are more likely to hear a /d/ in an /ar/ context than in an /al/ context, even when the segments are presented to different ears, the above experiments clearly show that phonetic context effects work across auditory streams. Thus, the perceptual system must combine auditory streams before all phonetic analysis is completed.

What is not clear is the impact these findings have on the speech perception frameworks described in the introduction. When performing similar experiments, Holt and Lotto (2002) discussed their findings in the context of the general auditory framework. The difference in compensation effects between Experiment 1 and Experiment 2 (where the third formant transition is presented contralaterally to the context segment) shows that some sort of acoustic processing or binding takes place along the auditory nerve before the acoustic signal reaches the

cochlear nucleus. However, there is no reason that these findings cannot also support the underlying gesture-based themes that Motor Theory and Direct Realist Theory put forth. Both Experiment 1 and Experiment 2 replicate Mann's (1980) findings that /da/ is always heard more often in /ar/ context than /al/ context, even when the stimuli are ambiguous in the /d-g/ spectrum. That this phonetic context effect is active across left and right ear auditory streams is a significant finding, but it does less to support a general auditory framework than it does to suggest that further phonetic processing research is needed, especially focused on the auditory neural pathway.

As mentioned above, there is less difference between the results of the /al/ and /ar/ contexts in Experiment 2 (chirp right) than in Experiment 1 (base right). This difference in the compensation effect suggests that separating the *formant transition chirp* from the context is more crucial than separating the *base target* from the context. Because the chirp contains the formant transitions and is thus responsible for cuing /d/ vs. /g/, these findings also suggest that phonetic context effects, while working across auditory channels, are more influential within one channel. Thus, although we have shown that phonetic context effects are not strictly diotic, it is likely that the effects become active before the left and right auditory channels merge into one auditory signal.

While the results we found were significant, there are minor changes that could better the robustness of the data. First, an improved /da/ stimulus could clarify some of the results of the experiments. The proportion of 'd' responses barely reached above 80% in Experiment 1, and it never broke the 80% mark in Experiment 2. If we could get 100% 'd' responses for token 1 (as we did for 'g' responses in token 9 of Experiment 1), we may be able to recognize stronger patterns in the data. We would also like to see if right ear advantage (REA) contributed to the change in compensation effect by flipping the left and right stimulus channels for some subjects (Kimura 1961). Additionally, because the experiment was short (only about 20 minutes), it would not be difficult to add another block of 45 trials to the present two blocks.

Of course, in order to discover more about phonetic context effects, new experiments will need to be designed. To gain more direct insight into the location and timing of phonetic context effects (and acoustic analysis in general) along the auditory pathway, it will be advantageous to look at subject performance in these tasks combined with fMRI or electrophysiological EEG data. To begin to tease apart the gestural from the non-gestural speech perception frameworks, experiments utilizing visual cues (along the lines of the McGurk Effect) may be beneficial, especially when coupled with brain imaging techniques (McGurk & MacDonald 1976).

Phonetic analysis may begin at the peripheral level of the auditory nerves, but our experiments found that the most of the phonetic context effects occurred at more central levels. With more research, we have the promise of mapping the phonetic (as well as linguistic and acoustic) analysis process to incredible specificity. As the convergence of linguistics, neuroscience, computational modeling, and other fields happens in front of our eyes, we are in an exciting, hopeful time for the field of speech perception.

References

Diehl, R.L., A.J. Lotto, L.L. Holt. 2004. "Speech Perception". *Annual Revue of Psychology* 55: 149–179.

Fodor, J. A. 1983. *The modularity of mind*. Cambridge, MA: MIT Press.

Fowler, C. A., & D.P. Rosenblum. 1990. “Duplex perception: A comparison of monosyllables and slamming doors”. *Journal of Experimental Psychology: Human Perception & Performance*, 16: 742-754.

Galantucci, B., C.A. Fowler, & M.T. Turvey. 2006. “The motor theory of speech perception reviewed”. *Psychon Bull Rev.* 13(3): 361-377.

Holt, L.L. 1999. “Auditory constraints on speech perception: An examination of spectral contrast.” Dissertation, Department of Psychology, University of Wisconsin, Madison, WI.

Holt, L.L., & A.J. Lotto. 2002. “Behavioral examinations of the level of auditory processing of speech context effects.” *Hearing Research* 167: 156-169.

Kimura, D. 1961. “Cerebral dominance and the perception of verbal stimuli”. *Canadian Journal of Psychology*. 15: 166-171

Klatt, D.H. 1980. “Software for a cascade/parallel formant synthesizer.” *J. Acoust. Soc. Am.* 67 (3): 971-995.

Kolb, B., & I. Whishaw. 2006. *An Introduction to Brain and Behavior*, 314-320. Worth; New York.

Liberman, A.M. 1957. "Some results of research on speech perception". *J. Acoust. Soc. Am.* 29 (1): 117–123.

Liberman, A.M., F.S. Cooper, D.P.Shankweiler, & M. Studdert-Kennedy. 1967. “Perception of the speech code”. *Psychological Review* 74 (6): 431–461.

Liberman, A.M. & I.G. Mattingly. 1985. "The motor theory of speech perception revised”. *Cognition* 21 (1): 1–36.

Lotto, A.J., & K.R. Kluender. 1998. “General auditory processes may account for the effect of preceding liquid on perception of place of articulation.” *Perception & Psychophysics* 60: 602-619.

Lotto, A.J., & L.L. Holt. 2006. “Putting phonetic context effects into context: A commentary on Fowler (2006)”. *Perception & Psychophysics*, 68 (2):178-183.

Mann, V.A. 1980. “Influence of preceding liquid on stop consonant perception”. *Perception & Psychophysics* 28: 407-412.

Mann, V.A., & A. M. Liberman. 1983. “Some differences between phonetic and auditory modes of perception.” *Cognition*, 14: 211-235.

McGurk H., & J. MacDonald. 1976. "Hearing lips and seeing voices". *Nature*. 264 (5588): 746–8.

Rand, T.C. 1974. “Dichotic release from masking for speech.” *J. Acoust. Soc. Am.* 55 (3):678-680.