Ipsilateral and Contralateral Phonetic Context Effects

Kevin Sitek

Abstract

Research in the field of speech perception has shown that the perception of certain speech sounds is dependent on their neighboring sounds. In this paper, we will show that this phonetic 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. Additionally, the results hint that the effect is greater when the formant transitions are presented to the ipsilateral ear as the context segment than when they are presented to the contralateral ear. The results therefore describe a process in which the left and right auditory streams combine before most, but likely not all, phonetic information is retrieved.

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. Like other peripheral systems, 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. In this paper we will look at phonetic context effects to test if phonetic analysis occurs in early, peripheral levels of acoustic analysis or in later, more central levels. To give us a better understanding of the present question, we will first look at key theories of speech perception and how they describe phonetic context effects, as well as an overview of the auditory pathway.

1.1 Theories of speech perception

Speech perception research is generally guided by a few central theories or frameworks, namely, Motor Theory, Direct Realist Theory, and general auditory and learning approaches (GALA) (Diehl et al. 2004).

1.1.1 Motor Theory

Motor Theory (MT) emerged as a result of Alvin Liberman's 1950s studies on the failure of blind participants in a study to correctly identify sounds in an acoustic alphabet (Liberman 1957; Galantucci, Fowler, & Turvey 2006). Unlike an orthographic alphabet, sounds in natural speech are not discrete—they meld together with properties of one speech sound appearing in a neighboring sound. This coarticulation motivated Liberman's first form of MT, which posited that babies mimic the speech sounds they are exposed to in order to learn the distinctions between phonemes in their language. Later, facing growing opposing research, Liberman shifted MT away from the mimicry basis to state that humans have adapted to being able to produce coarticulated sounds as well as to perceive them

(Liberman et al., 1967). The simplest way for this to instantiate is for both processes to be dependent on one processing source. In the 1980s, MT borrowed Fodor's language in describing a specific language "phonetic module" that works in both producing and perceiving human speech (Fodor, 1983; Liberman & Mattingly, 1985). Additionally, Liberman and Mattingly's newest form of MT said that coarticulation showed that intended phonetic gestures are distinct from realized gestures; the intended gestures are what the speaker and the listener converge upon in order to communicate and understand the same proposition.

MT found support in duplex perception experiments, which provided evidence for a phonetic module or mode. 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" sound presented to the left ear that is not quite complete; in the right ear is presented a "chirp" sound that includes the missing formants of the left ear input. The result is that the participant reports hearing a completed syllable in the left ear and a non-speech chirp in the right ear. This finding was taken to support a phonetic mode as being distinct from the general auditory mode, thus implying the existence of a specific, specialized language module in the brain (Mann & Liberman, 1983).

1.1.2 Direct Realist Theory

The most recent Motor Theory asserts three central claims about speech processing: (1) speech processing is special; (2) perceiving speech is perceiving gestures; and (3) the

motor system is recruited for perceiving speech (Galantucci, Fowler, & Turvey, 2006). The Direct Realist Theory of speech perception (DRT) was formed in large part by Carol Fowler et al. (1980) in opposition to some of the central tenets of Motor Theory.

First, DRT denies the existence of a "special" speech processing module in the brain. Such a module, if present, would only process phonetic signals and would be unused in times of no speech inputs. However, time and time again no such area of the brain has been found in functional magnetic resonance imaging tests, and it is unlikely that this type of structure will be discovered (Galantucci et al., 2006). Instead, research has found areas that are generally used for speech comprehension but that are also active during other forms of auditory perception. While duplex perception does support the claim that speech perception focuses on distal intended gestures instead of the proximal acoustic patterns, Fowler and Rosenblum (1990) found that duplex perception experiments also work for non-speech signals such as slamming doors. This research shows the extremely low possibility of having an auditory perceptual module devoted specifically to speech, since if there were one for speech there would also need to be one for slamming doors, which is not evolutionarily important enough to make a "slamming door mode" of the auditory system likely to exist.

As such, DRT denies that speech perception is based on the intended articulation of the speaker; instead, it is based on the actual vocal tract gestures the speaker produces. Similarly, DRT says that sounds are not coarticulated but coproduced (Diehl et al., 2004). This minor shift emphasizes that speech sounds are not *planned* to be coarticulated but that the combination of features of two neighboring phonemes occurs as a necessity during speech production. This difference allows the research of Mann and others continues to fit

in with the Direct Realist Theory. Support for such coarticulation/coproduction comes from studies where visual stimuli differ from auditory speech stimuli such as the McGurk Effect: an audio track of /ba/ is played on top of a video track of someone mouthing /ga/, and the effect is that participants hear something like /da/. If gestures did not play a role in speech perception, then the participants should report hearing what is produced in the audio track, /ba/.

1.1.3 General Auditory and Learning Approaches

The main opposition to the gesture-based theories of speech perception is presented by Lotto and Holt (2006) in response to Galantucci, Fowler, and Turvey (2006). Not a true "theory" but rather a framework that was developed in response to the gestural models, their general auditory and learning approaches (GALA) hypothesis is that all sounds are processed equally and that the gathering of spoken words happens without reference to the preceding vocal tract gestures. Thus, the acoustic signal plays the most important role in determining the speech of the language speaker. Lotto and Holt cite research showing that both humans and non-humans are capable of distinguishing between phonetic speech properties such as vowel onset times. Since animals without human speech are capable of observing minute differences in acoustic signals, language must be based on general auditory principles and not on an innately human understanding of the gestures required for speech sounds. Instead of assuming that production and perception rely on many of the same cognitive mechanisms, Lotto and Holt's GALA distinguishes between production following perception (the auditory enhancement hypothesis claims that we make speech sounds as distinct from each other as possible) and

perception following production (understanding what a speaker must do in order to produce an element of language) (Diehl et al., 2004).

1.2 How speech perception theories describe phonetic context effects

One feature described by all speech perception theories is the phenomenon where the perceptual system uses cues from neighboring speech sounds to better understand a target sound. This phenomenon is known as the phonetic context effect. Though speech perception theorists agree on the existence of phonetic context effects, their frameworks describe differing underlying mechanisms responsible for the effects.

The Motor Theory 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]. Mann's study shows that listeners use experiencial knowledge about their language as well as an undestanding of the speaker's intended phonetic articulations in order to comprehend the phonetic speech sounds in conversation.

Fowler's Direct Realist Theory rejects the specialized "phonetic module" proposed by Motor Theory (Liberman & Mattingly 1985) but maintains Mann's research as a crucial element of the theory (Diehl et al. 2004).

However, Holt and Lotto (2002) argue that phonetic context effects occur as a result of psychoaccoustic and physiological mechanisms that are not specific to human speech. 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. In previous experiments (Holt 1999; Lotto and Kluender, 1998), the authors found that a low-frequency precursor (F2 in /u/; F3 in /ar/) results in higher responses hearing a higher-frequency afterward (F2 in /da/; F3 in /da/). 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.

1.3 The Auditory System

The auditory system begins with the outer ear, which focuses external sound waves into the ear canal. Because of its narrow shape, the ear canal amplifies sound waves and sends them toward the eardrum, which vibrates at the rate of the sound waves' frequency. The eardrum is connected to the middle ear, a structure comprised of ossicles that amplify the vibrations to the cochlea's oval window and through the cochlear fluid. Floating in the cochlear fluid is the basilar membrane, a tonotopic structure that vibrates most at the location corresponding to the frequency of the sound wave. The bending membrane stimulates the outer hair cells, creating graded potentials in the inner hair cells that act as the auditory receptor cells. Depending on the membrane potential of the inner hair cells, the amount of neurotransmitter released by the hair cells onto auditory neurons varies, thus converting incoming sound waves into neural activity.

The encoded sound information then passes through the vestibulocochlear nerve and through the cochlear nuclei. It is in this brainstem structure that the left and right auditory pathways can be processed at the first time. After the cochlear nuclei, the auditory pathway continues through different structures of the brainstem until reaching the primary auditory cortex in the temporal lobe.

That the left- and right-ear percepts do not combine until the centrally-located cochlear nuclei is greatly important to the following experiments. If subjects are able to correctly identify /d/ and/g/ segments *only* when the formant transitions are presented ipsilaterally to the context, then the responsible phonetic context effects must be taking place before the left and right auditory stream interact in the cochlear nucleus. If, on the other hand, the task can be completed even in the dichotic conditions, then the phonetic context effects are largely occurring more centrally (i.e., not in the early peripheral levels of acoustic processing).

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.

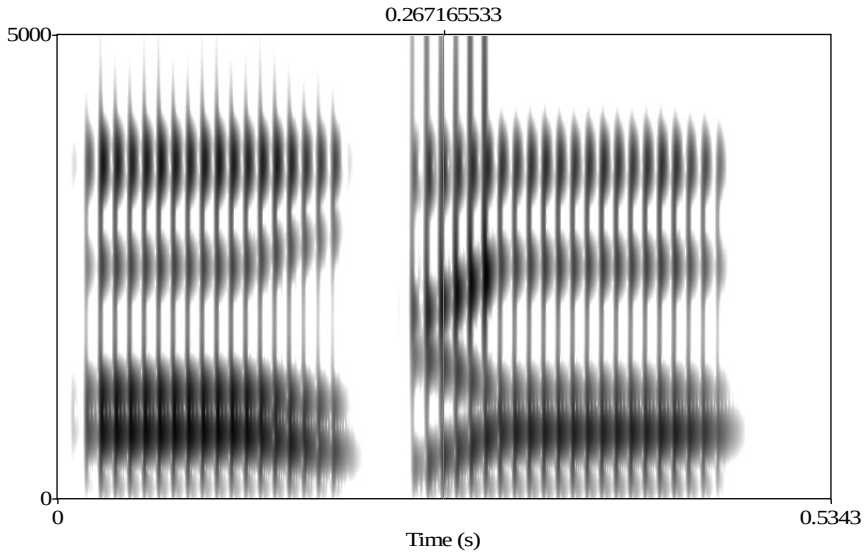

Figure 1. /al/ + /ga/ spectrogram.

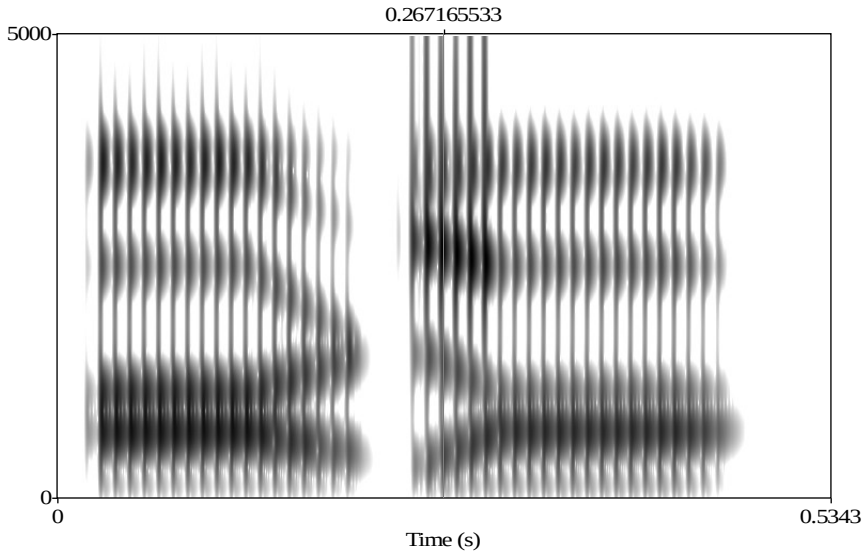

Figure 2. /ar/ + /da/ spectrogram.

3. Experiment 1

3.1 Method

3.1.1 Listeners

Fourteen 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. 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.

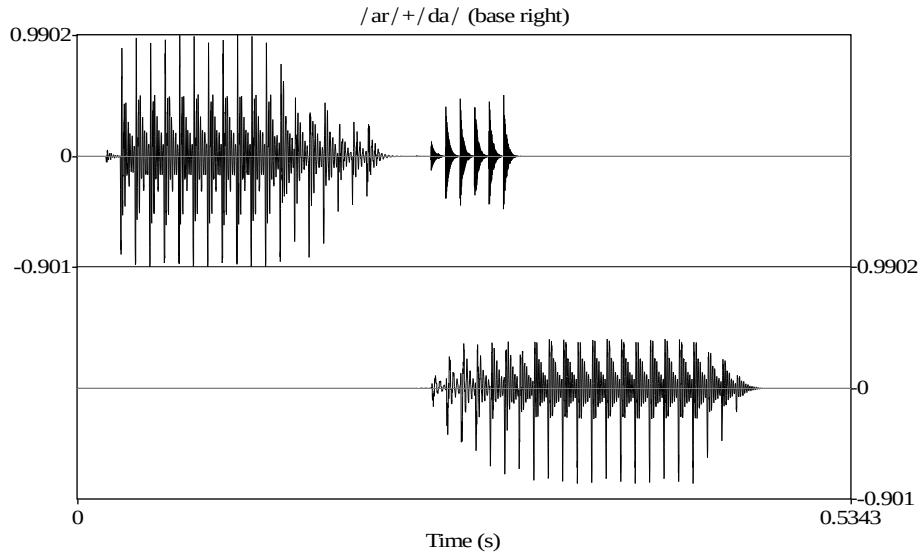

Figure 3. Experiment 1 (base right) waveform.

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 affect 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" condition responses was significant (t = -3.8263, df = 11.718, p-value = 0.003). 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 fourteen 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/.

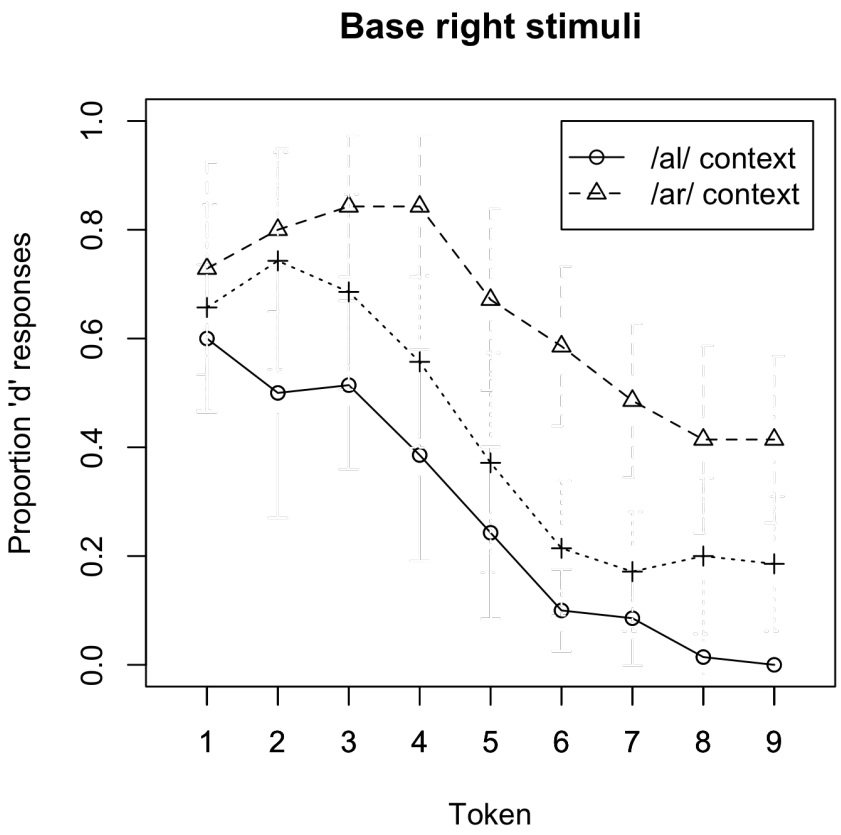

Figure 4. Each graph represents the mean percent 'd' responses. The crosses represent the null context (block 2) responses of the Experiment 1 listeners.

4. Experiment 2

4.1 Method

4.1.1 Listeners

Fifteen 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.

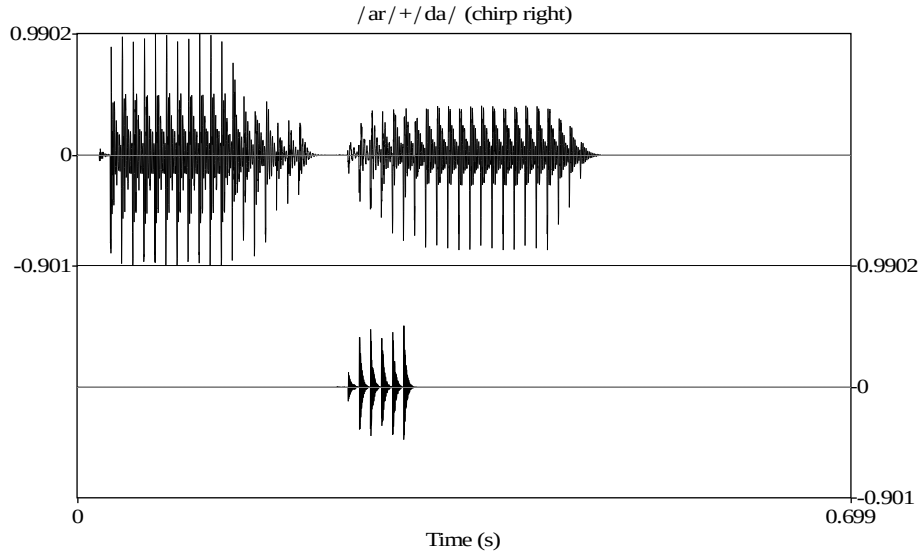

Figure 4. Experiment 2 (chirp right) waveform.

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" responses vs. "al" responses found a significant effect (t = -3.6104, df = 22.856, p-value = 0.001). Thus, phonetic context effects are preserved even when the third formant "chirp" is presented contralaterally to both the context and base segments. Similar to the results of Experiment 1, the 'd' responses were not definitively below 50% in the /al/ context until token 4. However, in the /ar/ context, the threshold is earlier than in Experiment 1—in Experiment 2, subjects perceived a 'g' in the /ar/ context beginning with token 6.

To compare the magnitude of the compensation effect in Experiment 1 to that of Experiment 2, we averaged each listeners' responses in each context (/al/ or /ar/). We then subtracted the average /ar/ response from the average /al/ response, giving us the compensation effect for each listener. While the average compensation effect was 0.371 in Experiment 1 and 0.206 in Experiment 2, this difference in compensation effect was not statistically significant (t = 1.5755, df = 26.303, p-value = 0.127).

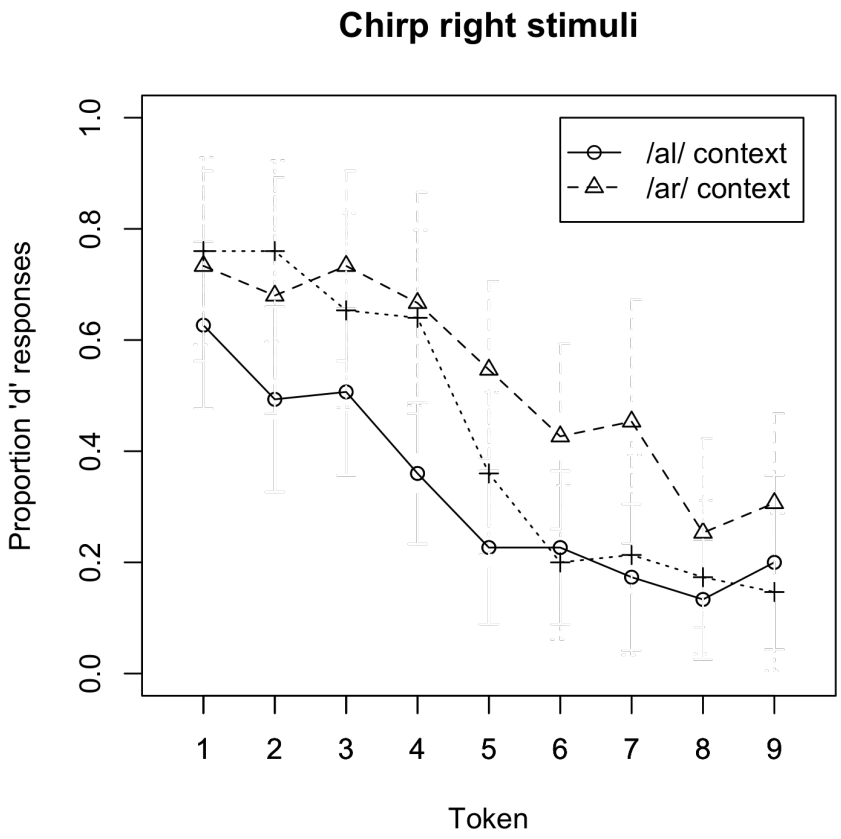

Figure 5. Each graph represents the mean percent 'd' responses. The crosses represent the null context (block 2) responses of the Experiment 2 listeners.

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 initial results indicating a difference in compensation effects between Experiment 1 and Experiment 2 (where the third formant transition is presented contralaterally to the context segment) hint that some sort of acoustic processing or binding takes place along the auditory nerve before the acoustic signal reaches the cochlear nucleus. If this turns out to be the case, then MT would have to reconsider its proposition that all of the speech percept is analyzed in a specific phonetic module in the brain. 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 compensation for coarticulation 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 research is needed, especially with an eye on the auditory neural pathway and the premotor cortex associated with speech production.

As mentioned above, preliminary results point to there being less difference between the results of the /al/ and /ar/ contexts in Experiment 2 (chirp right) than in Experiment 1 (base right). This possible difference in the compensation effect suggests that separating the *formant transition chirp* from the context may be 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 streams, may be more influential within one stream. 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 streams merge into one auditory signal. However, because these results were not statistically significant, we cannot make any certain claims until more research is done on this topic.

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 '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. Secondly, having more participants and more trials per participant would likely smooth out our data so that we would not see the strange bumps in the curves that we see presently. Additionally, having more participants would allow us to see whether the compensation effect changes between Experiment 1 and Experiment 2 or if the observed (statistically insignificant) difference is anomolous. 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 of phonetic context effects (and acoustic analysis in general) along the auditory pathway, it will be advantageous to run a high-resolution brain imaging test that incorporates phonetic analysis. 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.

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 (and 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.

Acknowledgements

Special thanks is due to Keith Johson, my thesis advisor and constant source of help, as well as to Susanne Gahl, my second reader.

References

Diehl, R.L., Lotto, A.J., Holt, L.L. (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., & Rosenblum, L. D. (1990). "Duplex perception: A comparison of monosyllables and slamming doors". *Journal of Experimental Psychology: Human Perception & Performance*, 16, 742-754.

Galantucci, B., Fowler, CA, & Turvey, M.T. (2006). "The motor theory of speech perception reviewed".

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., & Lotto, A.J. (2002). "Behavioral examinations of the level of auditory processing of speech context effects." *Hearing Research* 167, 156-169.

Klatt, D.H., 1980. "Software for a cascade/parallel formant synthesizer." *Journal of the Acoustical Society of America* 67 (3): 971-995.

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

Liberman, A.M. (1957). "Some results of research on speech perception". *Journal of the Acoustical Society of America* 29 (1): 117–123.

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

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

Lotto, A.J., & Kluender, K.R. (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., & Holt, L.L. (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., & Liberman, A. M. (1983). "Some differences between phonetic and auditory modes of perception." *Cognition*, 14, 211-235.