

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 28 (2007)
Indiana University

Cross-modal Repetition Priming in Spoken Word Recognition¹

Adam Buchwald², Stephen J. Winters³, and David B. Pisoni

*Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by grant number DC00012. The authors would also like to thank Melissa Troyer for her assistance running participants for this study. We also acknowledge Tessa Bent and Susannah Levi for helpful comments on earlier versions of this paper.

² Currently at New York University, Department of Speech-Language Pathology and Audiology.

³ Currently at University of Calgary, Department of Linguistics.

Cross-modal Repetition Priming in Spoken Word Recognition

Abstract. Multimodal speech perception has become a topic of considerable interest to speech researchers. Previous research has demonstrated that perceivers use information from the visual modality to inform the process of spoken word recognition. In this paper, we used a cross-modal repetition priming paradigm to explore questions about multimodal speech perception. First, we report that participants identified spoken words mixed with noise more accurately when the words were preceded by a dynamic video clip of the word being produced than when the words were preceded by a static image. Second, analyses of the responses indicate that both correct and incorrect responses are constrained by dynamic visual information. These complementary results indicate that perceivers integrate speech information from two different sensory modalities even when the signals are presented asynchronously. Third, we addressed the nature of multimodal integration, and found that the cross-modal repetition priming was maintained even when visual and auditory signals come from different sources. We discuss implications of these results for theories of multimodal speech perception.

Introduction

Multimodal speech perception and the cognitive processes by which individuals integrate auditory and visual speech information with linguistic knowledge have become major areas of research in the field of speech perception (Bernstein, 2005; Calvert, Spence, & Stein, 2004; Kim, Davis, & Krins, 2004; Massaro & Cohen, 1995; Massaro & Stork, 1998; Massaro, 1998; Rosenblum, 2005). As Sumbly and Pollack (1954) reported more than 50 years ago, normal-hearing listeners reliably make use of information from the visual speech signal to increase intelligibility of auditory speech over a wide range of signal-to-noise ratios. In addition, McGurk and MacDonald (1976) reported a perceptual illusion in which incongruent information from visual (e.g., [ba]) and auditory (e.g., [ga]) speech signals led to misperceptions (e.g., perceived: [da]) of the speech sounds. More recently, it has been found that presenting visual speech prior to auditory speech facilitates processing of the latter signal in a lexical decision task (Kim et al., 2004). These phenomena clearly suggest that perceptual integration – or binding – of the information from these two modalities is an integral part of speech perception, and that characterizing the nature of multimodal representations of speech is critical to a full understanding of speech perception (Summerfield, 1979).

This paper contributes to our understanding of multimodal speech perception by exploring the conditions under which information from auditory and visual signals is integrated. First, we provide critical evidence indicating that neither temporal synchrony of the auditory and visual signals nor identity in the source of the two signals is a necessary condition for this type of audiovisual integration to be observed, and that binding of these sources of information in the processes involved in spoken word recognition can occur in the absence of both of these conditions. In particular, we employed a repetition priming task in which we found that visual-only speech signals facilitated open-set recognition of subsequent noise-degraded audio-only speech signals, and that this effect persists even when the visual and auditory signals are clearly produced by different speakers (i.e., when there is a talker gender mismatch). Second, detailed analyses of the participants' responses reveal critical differences in responses to the auditory signals when participants first see a dynamic video clip compared to when they first see a static visual image, even when the signals cannot be reliably identified (i.e., when the open-set identification is incorrect). We show that these differences in responses are under stimulus control; that

is, the additional visual speech information constrains the responses in a manner consistent with the phonetic information present in the signal.

Audiovisual Integration in Speech Perception

The topic of multimodal speech perception and audiovisual integration have received attention from researchers addressing a wide variety of problems including second language acquisition (Davis & Kim, 2001; Kim & Davis, 2003; Davis & Kim, 2004), neurological processes and impairment (Skipper, Nusbaum, & Small, 2005; Hamilton, Shenton, & Coslett, 2006), speech production (Yehia, Rubin, & Vatikiotis-Bateson, 1998) and voice identity (Lachs, 2002; Kamichi, Hill, Lander, & Vatikiotis-Bateson, 2003; Lachs & Pisoni, 2004a, 2004b) as well as issues directly related to spoken word recognition (Dodd, Oerlemens, & Robinson, 1989; Kim et al., 2004; Mattys, Bernstein, & Auer Jr., 2002). These studies – combined with the pioneering work of Sumby and Pollack (1954) – all reveal that audiovisual integration is a fundamental part of speech perception which is seen in a variety of tasks and under a variety of conditions. In this paper, we focus on two possible conditions that are expected to promote audiovisual integration: temporal synchrony and source identity of the auditory and visual signals.

Previous research has shown effects of asynchronously presented visual speech on tasks involving auditory speech perception. Dodd, Oerlemans, and Robinson (1989) observed lexical repetition priming effects with visual-only primes and auditory-only targets. Using a semantic categorization task, Dodd et al. reported facilitation when participants were presented with visual-only primes of a speaker saying a word followed by an auditory target compared to a condition with no prime. This finding suggests that the visual prime and the auditory target activate common semantic representations in memory. It is worth noting that Dodd et al. used visual speech stimuli which were readily identified on their own (by at least 80% of participants in a screening task); thus, it remains possible that the facilitation in the semantic categorization task arises from separate, accurate identification of the stimuli from each modality rather than from audiovisual integration.

More recently, Kim et al. (2004) had participants perform a lexical decision task on spoken words that were preceded by visual-only speech signals. Participants' reaction times in lexical decision on trials with a consistent visual speech prime were compared to trials with inconsistent visual speech primes; they reported facilitation (i.e., faster reaction times) in the responses for trials with consistent visual speech primes, and concluded that speech perception is amodal because the priming effect suggests that visual and auditory signals activate common representations. However, it remains possible that the difference between the baseline and experimental conditions in Kim et al.'s study was due to response inhibition in the presence of inconsistent stimulus information as opposed to response facilitation in the presence of consistent information. Nevertheless, both of these explanations suggest that there is some type of common representation activated by both the visual and auditory signals which affects processes used to perform the lexical decision task. Moreover and critical to the present investigation, Kim et al. presented the two stimuli asynchronously.

Thus, these two lines of evidence suggest that speech information from auditory and visual modalities need not be presented synchronously to observe effects of perceptual binding in multimodal speech perception (see van Wassenhove, Grant, & Poeppel, 2006 for a recent discussion of temporal synchrony in audiovisual integration).

There also exists evidence that source identity is not required for audiovisual integration. Green, Kuhl, Meltzoff and Stevens (1991) reported a study in which participants readily perceived the McGurk illusion discussed above even when there was a gender mismatch between the face producing the visual

speech signal and the voice producing the auditory speech signal. Green et al. suggest that this finding supports the hypothesis that audiovisual integration occurs over abstract representations of speech and not over the detailed signals present in the environment. This unintuitive result contradicts any view of speech perception that does not allow for cognitive operations over abstract representations of speech, a strong view which has sometimes been attributed to event-based perception in general (Gibson, 1966), and Direct Realism in particular (Fowler, 1986).

This section has highlighted two findings regarding audiovisual integration in multimodal speech perception: 1) audiovisual integration has been observed in the absence of temporal synchrony (as in Kim et al., 2004, with both auditory and visual signals coming from the same speech event) and 2) audiovisual integration has been observed in the absence of source identity (but with temporal synchrony, as in Green et al., 1991). One goal of the present investigation is to determine whether audiovisual integration is observed in the absence of both temporal synchrony and source identity; that is, do observers integrate auditory and visual speech signals that are both separated in time and clearly come from different speech events in the world?

The other main component of the present investigation is to build on these previous works by investigating the nature of the audiovisual integration. To achieve this, we examined differences in correct and incorrect responses due to asynchronously presented visual speech in open-set spoken word identification.

Experiment 1

Experiment 1 sought to replicate the cross-modal priming in speech perception findings of Kim et al. (2004) using a task that allows us to explore the nature of the priming effect more directly. We employed a spoken word recognition task with auditory targets preceded by either static primes or dynamic video clips of the same speaker producing the same word. One critical goal of this experiment – beyond that reported by Kim et al. (2004) – was to try to gain a deeper understanding of the nature of audiovisual integration in a cross-modal task. We investigated not only whether the dynamic video clip prime would increase overall spoken word recognition accuracy, but also whether the responses on trials with dynamic video clips primes are more constrained than on those with static primes, and whether the nature of these constraints is predictable by (and can shed light on) the nature of visual and multimodal speech perception.

Participants

Forty Indiana University undergraduate students, ages 18-23, participated in Experiment 1. All participants were native speakers of English with no speech or hearing disorders. Participants received either course credit or monetary compensation for their participation in this study.

Materials

All stimulus materials were drawn from the Hoosier multi-talker audio-visual (AV) database (Sheffert, Lachs, & Hernandez, 1997). Monosyllabic, CVC words produced by one female speaker and one male speaker in the database were selected for this study. The stimulus set for each participant contained 96 different word tokens. In each condition, half of the stimuli were “Easy” words – high frequency words from lexically sparse phonological neighborhoods (e.g., “fool”), while the other half were “Hard” words – low frequency lexical items from lexically dense phonological neighborhoods (e.g., “hag”; see Luce and Pisoni, 1998).

Auditory Stimuli. In each condition, we used envelope-shaped noise (Horii, House, & Hughes, 1971) to reduce performance on the spoken word recognition task. The experimental stimuli were created by processing the audio files through a MATLAB program that randomly changed the sign bit of the amplitude level of 30% of the spectral samples in the acoustic waveform. Reducing auditory-only word recognition performance to below-ceiling levels is a necessary prerequisite to detect the effects of cross-modal repetition priming in the spoken word recognition task. Pilot data indicated that this level of noise degradation reduced auditory-only open-set recognition to about 50% correct.

Visual Stimuli. Two kinds of visual primes were used: Static and Dynamic. Dynamic primes consisted of the original, unedited video clips associated with each target word. Previous research has shown that the overall identification accuracy on these stimuli presented in visual-only condition was 14%, with less than 1% of the individual tokens accurately identified more than 90% of the time (Lachs & Hernandez, 1998). Thus, the specific words used in the study were not consistently identifiable in a visual-only condition. The video track of the Static primes consisted of a still shot of the speaker whose duration was identical to that of its counterpart in the Dynamic prime condition. The same still image was used in the Static condition for each target word. This image was taken from a resting state of each speaker.

Procedure

Participants were tested in groups of four or fewer in a quiet room with individual testing booths. During testing, each participant listened to the auditory signals over Beyer Dynamic DT-100 headphones at a comfortable listening level while sitting in front of a Power Mac G4. A customized SuperCard (v4.1.1) stack presented the stimuli to each participant. Participants were instructed to watch the computer monitor and then type the English word that they heard over the headphones using the computer keyboard.

On each trial (see Figure 1), participants first saw either a Static or a Dynamic visual prime. 500 milliseconds after the presentation of the visual prime, participants heard the degraded auditory target word over the headphones. A prompt then appeared on the screen asking the participant to type the word they heard. Presentation of the next stimulus was participant-controlled.

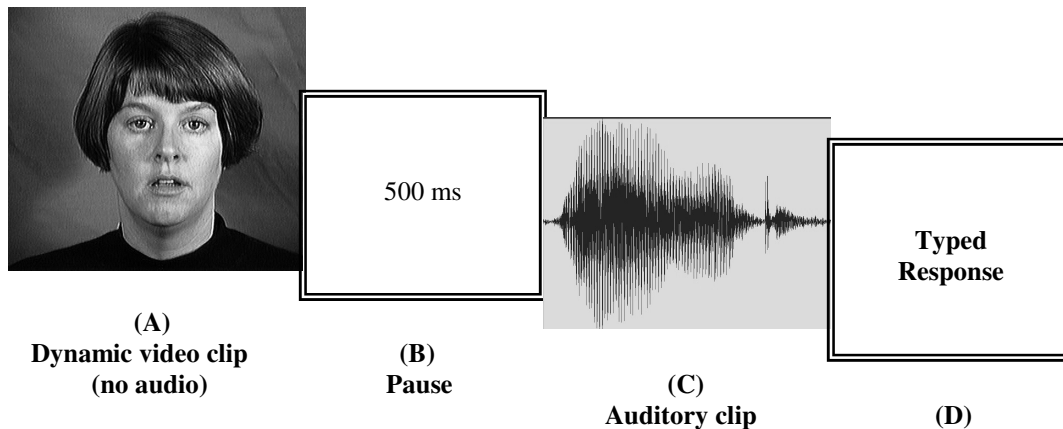


Figure 1. Schematic of trial in cross-modal repetition priming experiment. In Experiment 1, the video clip (A) and auditory clip (C) come from same token of a single speaker. In Experiment 2, (A) and (C) come from the same speaker or from different speakers producing the identical lexical item.

Participants were either presented with all female talker stimuli (both targets and primes) or all male talker stimuli. Words were presented to participants in random order, with Dynamic and Static primes randomly interleaved over the course of the experiment. Each participant responded to 48 words in each priming condition, half of which were lexically “Easy” targets and half of which were lexically “Hard” targets.

Results: Experiment 1

Word Identification Accuracy. For analyses reported in this section, the dependent variable was word recognition accuracy. The results revealed that the participants benefit from the presentation of a dynamic video identity prime when compared to a static prime. Overall, participants in Experiments 1 exhibited a 14% accuracy gain on trials in which a dynamic video prime preceded the degraded audio signal (67%) compared to trials in which a static face prime preceded the auditory target word (53%). The word recognition accuracy data for the female and male talkers were analyzed with separate 2x2 Prime Type (Dynamic/Static) vs. Target Type (Easy/Hard) repeated measures Analyses of Variance (ANOVAs).

The ANOVAs revealed a significant main effect of Prime Type for both the female speaker (Dynamic = 66.8%, Static = 49.1%; $F(1,19) = 166.8, p < 0.001$) and the male speaker (Dynamic = 67.2%, Static = 56.7%; $F(1,19) = 166.7, p < 0.001$), as well as significant main effects of Target type for both speakers (Female speaker: Easy = 67.1%, Hard = 49.9%; $F(1, 19) = 121.2, p < .001$; Male speaker: Easy = 69.9%, Hard = 52.5%, $F(1, 19) = 196.7, p < .001$). Thus, better performance was obtained on trials with Dynamic primes compared to trials with Static primes, and on trials with Easy targets compared with trials with Hard targets. The interaction between Prime Type and Target type was not significant for either speaker.

Response Analysis. To enrich our understanding of the information observers perceive and encode in the Dynamic prime condition when compared to the Static prime condition, we performed several analyses comparing the responses participants made on Dynamic trials to those made on Static trials. For the purposes of increasing power over the analyses, the data from participants who observed the Female speaker and those who observed the Male speaker were combined for all analyses reported in this section.

Collapsing over all the data, there were 1920 responses for each trial type. A total of 465 unique responses were given for Dynamic trials, whereas 610 unique responses were given for Static trials. A chi-square analysis revealed that there were significantly more unique responses to Static trials than to Dynamic trials [$\chi^2(1) = 46.40, p < 0.01$]. This finding strengthens the whole-word results reported above and indicates that the information present in the Dynamic prime acts as a constraint on the participants’ responses to the auditory word presented in noise. Additionally, there were significantly fewer unique responses on trials with Easy targets (476) compared to trials with Hard targets [599; $\chi^2(1) = 19.23, p < .01$]. The difference in the number of unique responses for Easy words with the two prime types (Dynamic: 190; Static: 286), and Hard words with the two prime types (Dynamic: 275; Static: 324) approached but did not reach significance [$\chi^2(1) = 3.68, p < .06$].

Additional analyses were geared towards exploring the nature of the constraints on the response selection process. Initially, each response was coded for the number of correct segments of the CVC word (i.e., 0-3 segments correct), and the average number of correct segments for each participant was then computed for each condition. A repeated measures ANOVA with Prime Type (Dynamic vs. Static) and Target Type (Easy vs. Hard) as independent variables and overall segmental accuracy as the

dependent variable revealed a main effect for Prime Type [$F(1,39) = 75.81, p < .001$], with higher segmental accuracy for targets with Dynamic primes (mean = 2.49, SD = 0.20) than for targets with Static primes (mean = 2.21, SD = 0.19). The ANOVA also revealed a main effect of Target Type ($F(1,39) = 69.46, p < .001$), with significantly higher segmental accuracy for Easy targets (mean = 2.44, SD = .28) than for Hard targets (mean = 2.25, SD = .24). There was also a significant interaction between Target Type and Prime Type [$F(1,39) = 12.57, p < .001$]. The locus of the interaction indicated that the effect of Target Type was larger for the Dynamic primes (Easy: 2.62; Hard: 2.34) than for the Static primes (Easy: 2.26; Hard: 2.16). Overall, these results support the claim that the responses on trials with Dynamic primes were more constrained (i.e., closer to the target) than the trials with Static primes.

To further address whether incorrect responses are also more constrained when preceded by Dynamic primes, we limited the analysis described above to responses in which the participant gave the wrong whole word response (thus giving a possible range of 0-2 segments correct). Using the number of correct segments in incorrect responses as the dependent variable, we performed 2x2 Prime Type (Dynamic/Static) vs. Target Type (Easy/Hard) repeated measures ANOVA. This analysis again revealed a significant main effect of Prime type [$F(1,39) = 15.35, p < .001$]; the number of correct segments on trials with Dynamic primes (mean = 1.47, SD = 0.18) was significantly greater than the number of correct segments on trials with Static primes (mean = 1.34, SD = 0.14). This result indicates that the information present in the dynamic video signal constrains all of the participants' responses, leading to greater accuracy even for incorrect responses.

Additionally, a main effect of Target type was obtained [$F(1,39) = 23.56, p < .001$]; however, when the analysis was limited to incorrect responses, the responses to trials with Hard targets had significantly higher overall segmental accuracy (mean = 1.49, SD = .21) than responses on trials with Easy targets (mean = 1.29, SD = .30). This result may at first appear surprising; however, it reflects a significantly higher proportion of incorrect responses with 2 segments correct on trials with Hard targets (558/927, 60.2%) than Easy targets (257/621, 41.4%; $\chi^2(1) = 52.02, p < .001$). Given the working definition of lexical neighbors as words sharing N-1 segments of an N-segment word (which was used to generate the Easy/Hard targets for this experiment; see Luce and Pisoni, 1998), participants' incorrect responses that contain two correct segments are, by definition, lexical neighbors of the target word. Thus, incorrect responses to Hard targets (words from lexically dense neighborhoods) were more likely to be neighbors of the target than incorrect responses on trials with Easy targets (words from lexically sparse neighborhoods). The interaction between Prime type and Target type was significant [$F(1,39) = 6.78, p < .05$], with the effect of Target type attenuated for Dynamic prime trials (Easy: 1.43; Hard: 1.53) compared to Static prime trials (Easy: 1.16; Hard: 1.46). Thus, the effect of Target type was stronger in the condition when there was no dynamic visual information about the target, further suggesting that this additional optical information provides a constraint on participants' open-set word identification responses.

The above results reveal that incorrect responses on trials with Dynamic primes are closer to the target than incorrect responses on trials with Static primes. To gain a more detailed understanding of how the dynamic video information constrains responses on the word recognition task, we analyzed the likelihood of correct responses for each syllable position of the CVC words as a function of Prime type.⁴ These data, presented in Table 1, show that the accuracy is greater for words in the Dynamic condition

⁴ For the remaining analyses, we report data collapsed over Easy/Hard targets, as the data within each of these Target types matches the overall pattern of the data. We return to a discussion of the effects of Target type in Experiment 2 and in the General Discussion.

compared to the Static condition for each of the syllable positions, revealing that the dynamic information helped constrain responses for all three syllabic positions of the CVC words.

	<i>Dynamic</i> % SD	<i>Static</i> % SD	<i>Analysis</i>
Onset	80.3 (8.3)	67.6 (7.4)	$t(39) = 9.18, p < .001$
Nucleus	87.1 (6.1)	78.4 (7.3)	$t(39) = 6.38, p < .001$
Coda	81.3 (7.9)	74.8 (8.3)	$t(39) = 4.39, p < .001$

Table 1. Response accuracy for each of the three syllable positions in the CVC stimuli as a function of prime type (Experiment 1).

To determine whether there was a difference in the accuracy benefit for any of the three positions, we computed a difference score (Dynamic – Static) for each syllable position. Planned comparisons indicated that the cross-modal priming effect was significantly greater for onset position than it was for either nucleus position [$t(39) = 2.51, p < .05$] or for coda position [$t(39) = 3.58, p < .01$], but there was no difference between accuracy on the nucleus position and coda position [$t(39) = 1.49, ns$].

The data analyzed in this section thus far suggest that there was a global benefit from the dynamic visual information which constrained all components of the participants' responses, and that this effect was particularly robust for onset position. However, it remains possible that the information in the dynamic video clip constrained responses by limiting specific components of the set of competing hypotheses about the target word. To address this possibility, we examined the participants' identification of particular phonological properties of the target stimulus. Specifically, we examined the likelihood that participants would correctly identify the place features, manner features, and voicing features of the onset and coda consonants in the target word. These analyses were performed by collapsing the data obtained from all forty subjects, and comparing the accuracy on these individual dimensions for target words with Dynamic primes and target words with Static primes. The results of these analyses are presented in Table 2.

		<i>Dynamic</i> % correct	<i>Static</i> % correct	<i>Analysis</i>
Place	Onset	86	76	$\chi^2(1) = 52.66, p < 0.001$
	Coda	90	85	$\chi^2(1) = 19.45, p < 0.001$
Manner	Onset	88	79	$\chi^2(1) = 58.89, p < 0.001$
	Coda	89	86	$\chi^2(1) = 7.28, p < 0.01$
Voice	Onset	98	97	$\chi^2(1) = 2.28, ns$
	Coda	97	95	$\chi^2(1) = 2.25, ns$

Table 2. Response accuracy in reporting features for onset and coda consonants as a function of prime type (Experiment 1).

The data in Table 2 indicate that the dynamic video clip primes created a robust increase in accuracy with respect to place and manner of articulation for both onset and coda consonants. This result suggests that the participants were able to use the optical information available in the dynamic video clip to limit the set of possible responses, and that this information was useful in specifying both place and

manner of articulation. With respect to voicing, we limited our analysis to those trials in which the target and response were obstruents and thus the voice feature would have to be specified as part of the response. Not surprisingly, there was no significant effect of prime type on accuracy of the voice feature, a finding that is consistent with the hypothesis that voicing is poorly specified in visual-only speech signals (Summerfield, 1979).

Discussion

The data presented above help to sharpen our understanding of the information specified in different sensory modalities used in speech perception. In particular, we presented evidence that a visual-only speech signal facilitates identification of asynchronously presented auditory speech when the latter is presented in noise. This result complements and builds on previous results in the literature indicating that speech perception is not limited to the auditory modality (e.g., Sumbly & Pollack, 1954; Massaro, 1998; Bernstein, 2005). We will return to a discussion of these broad issues in the General Discussion.

More specifically, Experiment 1 provided critical evidence suggesting that observers are able to integrate information presented in two modalities, even when the signals are separated in time; thus, temporal synchrony is not a necessary condition for audiovisual integration to be observed. This finding converges with the results reported by Kim et al. (2004), who found similar asynchronous cross-modal priming in a lexical decision task (also see Dodd et al., 1989). In an attempt to build on their earlier results, we explored an additional factor which may be a necessary feature for audiovisual integration in repetition priming: source identity of the two input modalities. In our view, the strong version of the Direct Realism approach maintains no role for abstract internal representations or cognitive operations in perception (Gibson, 1966), and should hold that identity of the source of the two streams of information is a necessary condition for audiovisual integration in speech perception. However, if we find that repetition priming due to audiovisual integration persists even when the visual and auditory signals come from different speech events that are temporally asynchronous, we must conclude that there is some additional component to audiovisual integration which relies on activation of common abstract (i.e., removed from the signal; non-episodic) representations in addition to the perception of the specific event itself.

Experiment 2

The second experiment sought to extend the findings of Experiment 1 by presenting participants with trials in which the visual signal and the auditory signal were produced by distinct talkers. To achieve this goal, we used two speech sources that would clearly be perceived as different speakers: a male face/voice and a female face/voice. This experiment represents an extension and elaboration of Green et al.'s (1991) intriguing finding that observers will perceive the typical McGurk illusion even with a gender mismatch between the face and voice. Thus, there exists some prior experimental evidence that source identity is not a necessary condition for audiovisual integration. The present study seeks to extend this finding by investigating whether audiovisual integration – indexed by repetition priming – may be achieved with neither source identity nor temporal synchrony.

Participants

Twenty-six Indiana University undergraduate students, ages 18-23, participated in Experiment 2. All participants were native speakers of English with no speech or hearing disorders. Participants received either course credit or monetary compensation for their participation in this study. None of the participants from Experiment 2 had participated in Experiment 1.

Materials

As with Experiment 1, all stimulus materials were drawn from the Hoosier multi-talker audio-visual (AV) database (Sheffert et al., 1997). Monosyllabic, CVC words produced by the same female speaker and male speaker as in Experiment 1 were selected for this study. In Experiment 2, 240 different word tokens were used. As with Experiment 1, half of the stimuli were “Easy” words – high frequency words in sparse phonological neighborhoods (e.g., “fool”), while the other half were “Hard” words – low frequency lexical items in high density neighborhoods (e.g., “hag”; Luce and Pisoni, 1998).

Procedure

The testing situation was identical to that used in Experiment 1. Each participant was presented with eight different trial types, with all permutations of prime type (Dynamic vs. Static), prime gender (Female vs. Male), and target gender (Female vs. Male). The experimental trials were analyzed as two groups: Matched (Female prime and Female target; Male prime and Male target) and Mismatched (Female prime and Male target; Male prime and Female target).

Results

Word Identification Accuracy. Data from Experiment 2 were analyzed with a 2x2x2 Prime type (Dynamic/Static) vs. Target type (Easy/Hard) vs. AV-matching (Matched/Mismatched) ANOVA. Consistent with the results reported from Experiment 1, the results indicated a significant main effect of Prime type; words from Dynamic prime trials were identified more accurately (mean = 65.6%) than words from Static prime trials (mean = 54.4%; $F(1, 25)=108.3, p < .001$). A significant main effect of Target type was also observed, with Easy targets recognized more accurately (mean = 65.1%) than Hard Targets (mean = 54.9%, $F(1, 25) = 85.6, p < .001$). No significant main effect was found for AV Matching ($F(1, 25) = 0.9, ns$), reflecting the lack of a difference in overall accuracy on AV-Matched and AV-Mismatched trials, regardless of Prime or Target type. Critical planned comparisons examined effects of Prime type separately for AV-Matched and AV-Mismatched trials. These comparisons revealed a significant effect of Prime type for both Matched (Dynamic: mean = 66.6%; Static: mean = 55.1 %; $t(25) = 3.27, p < .01$) and Mismatched (Dynamic: mean = 64.4%; Static: mean = 54.4%; $t(25) = 4.01, p < .001$), revealing that the spoken word recognition priming effect observed in the single-speaker condition does not crucially rely on the signals in the two stimulus presentation modalities coming from an identical source. No significant interactions were obtained in the two-speaker conditions (all F s < 1.6).

Response Analysis. We performed the same analyses on the set of responses in Experiment 2 as we did in Experiment 1. Collapsing over all the data, there were 3120 responses to targets with Dynamic primes and 3120 responses to targets with Static primes. A total of 1010 unique responses were given for Dynamic trials, whereas 1180 unique responses were given for Static trials. A chi-square analysis revealed that there were significantly more unique responses to Static trials than to Dynamic trials [$\chi^2(1) = 19.70, p < 0.01$]. This finding strengthens the results reported above, indicating that the information present in the Dynamic prime acts as a constraint on the participants’ responses to the auditory word presented in noise. When we examined the number of unique responses on the 1620 Matched Dynamic trials (659 unique responses) and the 1620 Mismatched Dynamic trials (700 unique responses), there was no significant difference between these two groups [$\chi^2(1) = 1.98, ns$], indicating that there was no difference in the constraint on responses for these conditions reflected by the number of unique responses for Matched trials and for Mismatched trials. Overall, there were more unique responses to Hard targets (893 unique responses) than to Easy targets [769 unique responses; $\chi^2(1) = 18.59, p < .01$]. No significant

differences were found in the proportion of unique responses to Easy and Hard words for any of the priming conditions (Dynamic Matched: Easy – 309, Hard – 350; Dynamic Mismatched: Easy – 324, Hard – 376; Static: Easy – 524, Hard – 638).

Following the analyses used in Experiment 1, each response was coded for the number of correct segments of the CVC word (i.e., 0-3 segments correct), and the average number of correct segments for each participant was computed for each condition. A repeated measures ANOVA revealed a significant main effect of Prime type [$F(1,25) = 69.26, p < .001$] on the number of correct segments, with responses on trials with Dynamic primes (mean = 2.46, SD = 0.15) having significantly more correct segments than responses on trials with Static primes (mean = 2.27, SD = 0.14). The difference between the Matched (mean = 2.48, SD = .16) and the Mismatched (mean = 2.43, SD = .17) groups was not significant, though there was a trend towards better performance on Matched trials ($t(25) = 2.03, p < .06$). When compared to the static trials, performance on Dynamic trials was significantly better for both Matched ($t(25) = 8.41, p < .001$) and Mismatched ($t(25) = 6.72, p < .001$) trials. These data provide further support for the claim that the responses are constrained by the presence of the optical information available in the dynamic video clip. This ANOVA also revealed a significant main effect of Target type [$F(1,25) = 34.71, p < .001$], with responses on trials with Easy targets having more segments correct (mean = 2.41, SD = .17) than responses on trials with Hard targets (mean = 2.30, SD = .19). The interaction between Prime type and Target type was not significant.

When the analysis was limited to responses in which the participant gave the wrong whole word response, a repeated measures ANOVA revealed a significant main effect of Prime type [$F(1,25) = 14.10, p < .05$], with the number of correct segments on trials with Dynamic primes (mean = 1.43, SD = .28) significantly greater than the number of correct segments on trials with Static primes (mean = 1.36, SD = .18). There was no significant difference between performance on Matched (mean = 1.44, SD = .21) and Mismatched (mean = 1.42, SD = .16) trials ($t(25) = 0.69, ns$). When compared to the number of segments correct in incorrect responses for the Static condition, there were significantly more segments correct for Dynamic trials in both the Matched ($t(25) = 2.34, p < .05$) and Mismatched ($t(25) = 2.10, p < .05$) AV conditions. This latter result confirms again that the information present in the dynamic video signal constrains all of the participants' responses leading to greater accuracy even for incorrect responses, and that this effect is not attenuated by having a gender mismatch between the source of the dynamic video prime and the auditory target.

The ANOVA also revealed a significant main effect of Target type [$F(1,25) = 26.39, p < .05$], with the number of segments correct in incorrect responses higher for Hard targets (mean = 1.48, SD = .14) than for Easy targets (mean = 1.31, SD = .19). As in Experiment 1, this reflects a greater number of neighbors given as responses for Hard targets (807/1406, 57.4%) than for Easy targets (521/1089, 47.8%; $\chi^2 = 22.12, p < .05$). The interaction between Prime type and Target type was not significant for Experiment 2.

Following the analyses in Experiment 1, we analyzed the likelihood of correct responses for each syllable position of the CVC words as a function of Prime type (collapsing over Target types). These data are presented in Table 3, with Matched and Mismatched conditions listed separately as well as combined. These data indicated that the overall accuracy is increased for words in the Dynamic condition compared to the Static condition for each of the syllable positions, revealing that the dynamic information helped constrain responses for all three segments of the CVC words.

		<i>Dynamic</i>	<i>Static</i>	<i>Analysis</i>
		% SD	% SD	
Onset	Total	79.9 (6.7)	70.4 (5.0)	$t(25) = 8.35, p < .001$
	<i>Matched</i>	81.3 (6.3)		$t(25) = 9.76, p < .001$
	<i>Mismatched</i>	78.3 (8.4)		$t(25) = 5.40, p < .001$
Nucleus	Total	84.6 (6.1)	78.6 (5.5)	$t(25) = 5.82, p < .001$
	<i>Matched</i>	85.4 (6.4)		$t(25) = 6.01, p < .001$
	<i>Mismatched</i>	83.7 (6.0)		$t(25) = 4.25, p < .001$
Coda	Total	81.7 (4.9)	76.8 (4.6)	$t(25) = 4.86, p < .001$
	<i>Matched</i>	81.8 (6.1)		$t(25) = 4.19, p < .001$
	<i>Mismatched</i>	81.5 (5.3)		$t(25) = 4.02, p < .001$

Table 3. Response accuracy for each of the three syllable positions in the CVC stimuli as a function of prime type (Experiment 2). Matched and Mismatched Dynamic trials are compared to overall data from Static trials.

To determine whether there was a difference in the accuracy benefit for any of the three positions, we computed a difference score for each syllable position. Overall planned comparisons indicated that the cross-modal priming effect was significantly greater for onset position than it was for either nucleus position [$t(25) = 3.07, p < .01$] or for coda position [$t(25) = 3.46, p < .01$], but there was no difference between accuracy on the nucleus position and coda position [$t(25) = 0.88, ns$]. Comparisons limited to Matched and Mismatched dynamic trials exhibit the same pattern, with onset position having significantly greater priming benefit than nucleus or coda, and with no significant difference observed between nucleus and coda.

The analyses of the data from Experiment 2 presented thus far suggest that there was a global benefit from the dynamic information which constrained all components of the participants' responses. Further, these effects were observed even when there was neither temporal synchrony nor source identity of the auditory and dynamic video speech signals. As discussed above, it is critical to investigate whether the priming benefit reflects a general benefit from the information present in the video clip, or whether the responses are constrained by the stimulus by limiting specific components of the set of competing hypotheses about the target word.

Following the analyses in Experiment 1, we examined the likelihood that participants would correctly identify particular phonological properties of the target stimulus. In particular, we examined the likelihood that participants would correctly identify the place features, manner features, and voicing features of the onset and coda consonants in the target word. The results are presented in Table 4.

The data in Table 4 indicate that the dynamic video clip primes promote a robust increase in accuracy with respect to place and manner of articulation for both onset and coda consonants. Crucially, these effects hold for both Matched and Mismatched primes; that is, the responses were significantly more accurate for both place and manner features even when the prime and target were presented asynchronously and when they came from a different source. The performance on Matched and Mismatched trials did not differ significantly for any comparisons in Table 4 other than Onset place, where the identification of place for Matched trials was significantly better than identification of place for Mismatched trials ($\chi^2 = 8.68, p < .05$). With respect to voicing, following the analyses in Experiment 1, we limited our analysis to those trials in which the target and response were obstruents and thus the voice feature would have to be specified as part of the response. As with Experiment 1, there was no

significant effect of prime type on accuracy of the voice feature. This result is again consistent with the claim that voicing is not well-specified as part of the visual-only speech signal and that the other attributes of responses are under stimulus control.

Feature	Position	Prime Type	Dynamic %	Static %	Analysis
Place	Onset	Total	86	78	$\chi^2(1) = 52.3, p < 0.001$
		Matched	87		$\chi^2(1) = 54.6, p < 0.001$
		Mismatched	84		$\chi^2(1) = 17.2, p < 0.001$
	Coda	Total	89	85	$\chi^2(1) = 26.0, p < 0.001$
		Matched	89		$\chi^2(1) = 13.8, p < 0.001$
		Mismatched	89		$\chi^2(1) = 19.4, p < 0.001$
Manner	Onset	Total	88	83	$\chi^2(1) = 29.7, p < 0.01$
		Matched	89		$\chi^2(1) = 26.7, p < 0.01$
		Mismatched	87		$\chi^2(1) = 12.4, p < 0.01$
	Coda	Total	90	88	$\chi^2(1) = 8.18, p < 0.05$
		Matched	90		$\chi^2(1) = 6.48, p < 0.05$
		Mismatched	90		$\chi^2(1) = 4.10, p < 0.05$
Voice	Onset	Total	98	97	$\chi^2(1) = 3.08, ns$
		Matched	98		$\chi^2(1) = 0.72, ns$
		Mismatched	98		$\chi^2(1) = 2.47, ns$
	Coda	Total	95	94	$\chi^2(1) = 1.80, ns$
		Matched	95		$\chi^2(1) = 3.55, ns$
		Mismatched	95		$\chi^2(1) = 1.96, ns$

Table 4. Accuracy in identifying the place, manner, and voice for onset and coda consonants. Statistical analyses compare the performance on static trials to performance on total Dynamic trials, as well as to Matched and Mismatched trials separately.

General Discussion

The experimental work reported in this paper reflects a novel application of the conventional repetition priming paradigm. Here, we used this paradigm to investigate central issues pertaining to the nature of multimodal speech perception. Participants were required to identify spoken words presented in envelope-shaped noise that were preceded by dynamic or static visual-only primes. In Experiment 1, the results indicated that participants were more accurate at identifying spoken words when the auditory stimulus was preceded by a dynamic visual stimulus of the same word compared to a static image of the speaker's face. Furthermore, detailed analyses of the participants' responses indicated that the dynamic video information constrained the responses to the auditory target even on trials where spoken word recognition was not successful. In Experiment 2, the same priming benefit was observed even when it was readily apparent that the auditory and visual signals came from different speakers.

These results raise several important issues regarding the nature of multimodal speech perception. First, we have demonstrated that cross-modal repetition priming in speech perception requires neither temporal synchrony nor source identity; the repetition priming effect was observed even when the commonality that exists between the dynamic video clip prime and auditory target was only at

the level of the lexical identity of the token being produced, and not identity of the token or specific “episode” that is being perceived. This result is consistent with a view of multimodal speech perception in which integration of auditory and visual information is part of the cognitive process(es) involved in speech perception (Bernstein, 2005; Hamilton et al., 2006; Kim et al., 2004; Massaro & Stork, 1998).⁵ According to this account, language users store and maintain in memory abstract, internal representations of the external auditory world, such as a representation of the speech sound /p/. The results of the cross-modal repetition priming experiments reported here suggest that these representations may be activated directly by an acoustic waveform containing particular sounds, and they may also be activated (either directly or indirectly) by dynamic visual displays of a speaker creating the articulatory gesture that produces the same speech sound (e.g., a labial closure).

The results reported here also reveal that the nature of the benefit observers received from the dynamic video prime was under tight stimulus control. In particular, the participants’ responses were constrained in several important ways. First, more correct responses to auditory targets were observed on trials with dynamic video clip primes. Second, across responses from all participants, there was a smaller range of responses provided on trials with dynamic primes compared to static primes. Third, the presentation of the dynamic primes increased identification of segments in all three of the syllable positions of the CVC targets, with onsets benefiting more than the nucleus and coda. Fourth, the responses on trials with dynamic primes were more likely to exhibit accurate identification for two kinds of sub-segmental information: place of articulation and manner of articulation of both onset and coda consonants. In contrast, dynamic primes did not significantly increase the likelihood of accurately reporting the correct voicing status of the target obstruents, revealing that the components of the speech signal that are not available in the visual speech stream did not receive a benefit from the dynamic visual display.

Audiovisual Integration and Cross-Modal Identity Matching

Another line of research in the multimodal speech perception literature has revealed that perceivers are able to match a video of a speaker’s face to the appropriate corresponding voice when visual and auditory stimuli are presented separately (Lachs, 2002; Lachs & Pisoni, 2004a, 2004b). The cross-modal matching task can be performed successfully even when the linguistic content of the two signals differs (Kamichi et al., 2003), suggesting that the perceptual cues used for cross-modal identity matching are independent of the idiosyncrasies of a particular utterance.

Lachs and Pisoni (2004a, 2004b) suggested that their participants’ success in cross-modal identity matching – in which the correctly matched stimuli came from the same utterance – may be rooted in event-based perception (Gibson, 1966). Lachs and Pisoni’s auditory and visual stimuli provided information about the same physical event in the world, and they argued that “integration” of the two modalities of information came from the real-world event itself, which shaped and constrained the pattern of sensory stimulation impinging on the eyes and ears. Within the direct realist event-based theoretical framework (Fowler, 1986), the locus of audiovisual integration is in the real world, and acoustic and optical speech signals are integrated seamlessly because they are two sources specifying information about the same distal event (also see Fowler, 2004).

⁵ This type of theoretical approach posits that sensory information from the world is encoded in modality-specific representations, and that these modality-specific representations are either: a) linked directly to one another (Massaro & Stork, 1998); or b) linked to a separate “multimodal” representation that integrates information from the different sources (Skipper et al., 2005; Hamilton et al., 2006). However, the difference between these proposals cannot be addressed by the research reported here.

In our view, the results of Experiment 2 – which provided clear and consistent evidence indicating that the effects of priming on both overall accuracy and in a detailed error analyses are maintained even in a condition where there was a mismatch between the speakers – suggest that an event-based perception account would need to additionally permit a level of abstraction in the multimodal speech perception process. Experiment 2 presented listeners with visual-only primes and auditory-only targets which were lexically identical (e.g., both stimuli are “cat”), but clearly produced by two different speakers (one male, one female). Thus, the prime and target stimuli came from two different perceptual events in the world. If a strong version of the event-based perspective on audiovisual integration outlined above were correct, the repetition priming effect should be absent in this condition. When the speakers differ – as in our experimental manipulation – the two sensory input modalities no longer provide the perceiver with sensory information about the same event in the world. However, if perceptual identity is defined with respect to the articulatory gestures that create the visual and auditory percept (e.g., [p] defined as voiceless labial stop, regardless of who produces it), then there is no reason to predict that the cross-modal priming effect would be absent when there is a lack of identity in the source of the two stimuli. However, it is worth noting that accounting for the data presented above requires that the identity between the two signals is processed at some abstract level of representation (e.g., identity of underlying segments without identity in the actual events producing the segments).

Open-set Identification and Lexical Access

One additional finding which emerged from this study provides further insight into the nature of lexical competition in the process of lexical access regardless of input modality. For both experiments reported here, when we looked at the incorrect whole word responses (i.e., failures of lexical access), we observed more correct segments on trials with “hard” target words (low frequency words from dense lexical neighborhoods) than on trials with “easy” target words (high frequency words from sparse lexical neighborhoods). This finding was largely attributable to a larger number of incorrect responses with two segments correct on trials with hard targets than on trials with easy targets. The definition of lexical neighbor used in this paper, based on Luce and Pisoni (1998), was a word that shares all but one segment with the target word. Thus, it was more likely that incorrect responses for “hard” targets were neighbors of the target word (i.e., sharing two of the three segments) than it was that incorrect responses for “easy” targets were neighbors of the target word. While this result follows from the Neighborhood Activation Model (NAM) of Luce and Pisoni (1998) in a straightforward manner, it is a novel empirical demonstration of a critical component of NAM.

NAM holds that the strength and number of competitors directly influences the ease with which lexical items are accessed (Luce & Pisoni, 1998). Previous attempts to understand the role of neighborhood density in lexical access have typically focused on the increase in accuracy and processing time (Luce & Pisoni, 1998; Vitevitch & Luce, 1998, 1999; Vitevitch, Luce, Pisoni, & Auer Jr., 1999) for words with strong competitors (i.e., “hard” words) compared to words with weaker competitors (i.e., “easy” words). However, previous accounts have not included detailed response analyses of the type presented in this paper. The results reported here provide further support for the fundamental claim underlying NAM by demonstrating that when lexical access fails, the response is more likely to be a lexically similar neighbor/competitor for “hard” words than it is for “easy” words.

Conclusion

We reported results from a cross-modal priming study in which identification of spoken words mixed with noise was facilitated by the earlier presentation of a dynamic video clip of the utterance compared to a static image of a speaker. The present set of findings indicate that neither temporal

synchrony in the presentation of the two signals nor identity in the source of the two signals is a necessary precondition for audiovisual integration in multimodal speech perception, suggesting that the set of neural and cognitive processes involved in multimodal speech perception includes activation of abstract representations of speech. The cross-modal repetition priming paradigm can be used in the future to provide critical new information pertaining to the nature of multimodal representations of speech by exploring the nature of the stimuli that produce this effect. We expect that these lines of research will converge to address additional issues related to multimodal perception of linguistic information, such as the time-course of audiovisual integration in speech perception processes or the neural mechanisms underlying repetition priming (see Grill-Spector, Henson, & Martin, 2006 for a recent review) and multimodal perception (e.g., see Ghazanfar & Schroeder, 2006). In addition, these lines of research are relevant to understanding the relation of the two input modalities in clinical populations such as hearing-impaired listeners who have experienced a period of auditory deprivation that may encourage reorganization and remodeling of the typical developmental processes (Bergeson & Pisoni, 2004).

References

- Bergeson, T. R., & Pisoni, D. B. (2004). Audiovisual speech perception in deaf adults and children following cochlear implantation. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.
- Bernstein, L. E. (2005). Phonetic processing by the speech perceiving brain. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 79-98). Malden, MA: Blackwell.
- Calvert, G. A., Spence, C., & Stein, B. E. (Eds.). (2004). *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.
- Davis, C., & Kim, J. (2001). Repeating and Remembering Foreign Language Words: Implications for Language Teaching Systems. *Artificial Intelligence Review*, 16, 37-47.
- Davis, C., & Kim, J. (2004). Audio-visual interactions with intact clearly audible speech. *Quarterly Journal of Experimental Psychology*, 57A(6), 1103-1121.
- Dodd, B., Oerlemans, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Language*, 22, 59-77.
- Fowler, C. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, 14, 3-28.
- Fowler, C. (2004). Speech as a supramodal or amodal phenomenon. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The Handbook of Multisensory Processes*. Cambridge, MA: MIT Press.
- Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Science*, 10, 278-285.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, 38, 269-276.
- Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Science*, 10(1), 14-23.
- Hamilton, R. H., Shenton, J. T., & Coslett, H. B. (2006). An acquired deficit of audiovisual speech processing. *Brain and Language*, 98, 66-73.
- Horii, Y., House, A. S., & Hughes, G. W. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, 49, 1849-1856.
- Kamichi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). 'Putting the face to the voice': Matching identity across modality. *Current Biology*, 13, 1709-1714.

- Kim, J., & Davis, C. (2003). Task effects in masked cross-script translation and phonological priming. *Journal of Memory and Language*, *49*, 484-499.
- Kim, J., Davis, C., & Krins, P. (2004). Amodal processing of visual speech as revealed by priming. *Cognition*, *93*(1), B39-B47.
- Lachs, L. (Ed.). (2002). *Vocal tract kinematics and crossmodal speech information*. Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., & Hernandez, L. R. (1998). Update: The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., & Pisoni, D. B. (2004a). Cross-modal source information and spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 378-396.
- Lachs, L., & Pisoni, D. B. (2004b). Crossmodal source identification in speech perception. *Ecological Psychology*, *16*(3), 159-187.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*, 1-36.
- Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.
- Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. *Current Directions in Psychological Science*, *4*, 104-109.
- Massaro, D. W., & Stork, D. G. (1998). Speech recognition and sensory integration: a 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech. *American Scientist*, *86*, 236-244.
- Mattys, S. L., Bernstein, L. E., & Auer Jr., E. T. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, *64*(4), 667-679.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.
- Rosenblum, L. D. (2005). Primacy of Multimodal Speech Perception. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 51-78). Malden, MA: Blackwell.
- Sheffert, S., Lachs, L., & Hernandez, L. R. (1997). The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 578-583). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Skipper, J. I., Nusbaum, H. C., & Small, S. L. (2005). Listening to talking faces: motor cortical activation during speech perception. *Neuroimage*, *25*, 76-89.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212-215.
- Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica*, *36*, 314-331.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2006). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*.
- Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in spoken word perception. *Psychological Science*, *9*, 325-329.
- Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic Phonotactics and Neighborhood Activation in Spoken Word Recognition. *Journal of Memory and Language*, *40*, 374-408.
- Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer Jr., E. T. (1999). Phonotactics, neighborhood activation and lexical access for spoken words. *Brain and Language*, *68*, 306-311.
- Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, *26*, 23-43.

