

**RESEARCH ON SPOKEN LANGUAGE PROCESSING**

Progress Report No. 28 (2007)

*Indiana University*

**Integrating Auditory and Visual Information in Speech Perception:  
Audiovisual Phonological Fusion<sup>1</sup>**

**Joshua L. Radicke, Susannah V. Levi, Jeremy L. Loebach and David B. Pisoni**

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

---

<sup>1</sup> This work supported by NIH-NIDCD Training Grant T32-DC00012 and NIH-NIDCD Research Grant R01-DC00111. We would like to Luis Hernandez for providing technical assistance and advice in the design and implementation of the experimental procedures.

## **Integrating Auditory and Visual Information in Speech Perception: Audiovisual Phonological Fusion**

**Abstract.** Phonological Fusion is a phenomenon in which different phonemes are presented to each ear, prompting the listener to perceive a blend of the two (e.g., /ba+/la/=/bla/). The present study assessed whether Phonological Fusion has an audiovisual analogue. That is, when listeners are presented with video clips containing visual stop consonants paired with auditory liquids, do they integrate information across the two modalities as they do in unimodal Phonological Fusion (e.g., visual “back” + auditory “lack” = “black”)? The three experiments presented here demonstrate that Audiovisual Phonological Fusion does occur, but primarily for visual bilabial stop consonants (e.g., /b/ and /p/) paired with the auditory liquid /l/. Moreover, the overall rate of fusion is determined by the lexicality of the target word. Taken together, the results of the present study suggest that similar processes underlie both unimodal and multimodal fusion, suggesting a general mechanism for conflict resolution in speech perception.

### **Introduction**

Integration of sensory information, whether within or across modalities, is necessary for successful interaction with the environment. The fusion of two separate pieces of information to form a single object or event is common to perceptual tasks. Fusions frequently occur in speech perception, and can be unimodal or multimodal. Dichotic listening fusions, such as Phonological Fusion, are a form of auditory unimodal fusion, and demonstrate how the integration of information between the two ears can modify the percept of an auditory stimulus. When each ear is presented with a different speech sound (e.g., /ba/ and /la/), the resulting percept is a combination of the two streams of phonetic information (e.g., /bla/)(Cutting, 1976). Similarly, multimodal fusions in speech perception occur when different sources of visual and auditory phonetic information are integrated (e.g., visual /ba/ auditory /ga/), resulting in the perception of an average of the two streams (e.g., /da/)(McGurk & MacDonald, 1976). Although most forms of multimodal fusions in speech perception have a unimodal analogue, suggesting that there is a general set of rules that governs both, not all fusions have been tested in both domains. The present study, therefore, assessed whether auditory Phonological Fusion has a multimodal analogue, and investigated some of the conditions under which it may arise.

Much of the work on unimodal fusions in speech perception comes from the work of Cutting (1976). Utilizing the dichotic listening paradigm in which different sources of auditory information are presented to each ear of a listener, Cutting described six prominent unimodal fusions in speech perception: Sound Localization, Psychoacoustic Fusion, Spectral Fusion, Spectral/temporal Fusion, Phonetic Feature Fusion, and Phonological Fusion. Although many of these unimodal fusions also have multimodal analogues (Sound Localization, and Phonetic Feature Fusion), not all have been investigated. The present study sought to assess whether Phonological Fusion has a multimodal analogue.

Sound Localization Fusions occur when two speech sounds, whose onsets are temporally asynchronous, are presented dichotically, resulting in the percept of a single sound originating from a particular location in space (Cutting, 1976). In this case, the perceived azimuthal location of the speech sound is determined by the temporal synchrony of the two stimuli: if the sounds arrive at each ear at a different time, the resulting percept is of two separate sounds originating in two separate locations (Cutting, 1976). The multimodal analogue of the Sound Localization Fusion is the Ventriloquist Effect

(Bertelson, Vroomen, Gelder & Driver, 2000). The Ventriloquist Effect is a multimodal fusion in which the perceived location of an auditory stimulus is altered by the presence of a salient visual stimulus (Bertelson *et al.*, 2000). In this case, an auditory stimulus is produced by one sound source (e.g., the human ventriloquist) but attributed to originating from a different location due to the salient visual information (e.g., the movements of the inanimate doll's mouth). In both cases, two separate pieces of information are combined to determine the percept of the location of the sound source.

Psychoacoustic Fusion (Cutting, 1976) is a dichotic listening fusion in which two different phonemes are presented to each ear, resulting in the percept of a single phoneme that is the “average” of the other two. For example, /ba/ is presented to one ear, and /ga/ to the other, resulting in the fused percept of /da/ (Cutting, 1976). An analogous multimodal fusion is the well-known McGurk Effect (McGurk & MacDonald, 1976), in which auditory and visual speech information mismatch, eliciting the percept of something that is an average of the two. When subjects were presented simultaneously with a video clip of a talker producing a velar stop (e.g., “gaga”) and an audio track of a talker producing a bilabial stop (e.g., “baba”), the most common percept reported is an alveolar stop that is an average of the two (e.g., “dada”). The perceptual fusions under Phonetic Feature Fusion and the McGurk Effect do not simply combine information across the two sources of input: rather subsegmental information is fused to form a single segment that is not completely specified by either source alone.

Another dichotic listening fusion reported by Cutting (1976) is Phonological Fusion, which occurs when two different sounds are presented to each ear resulting in a percept that is a combination of the two. In this case, one consonant is presented to one ear (e.g., /ba/), and a different consonant to the other (e.g., /la/) resulting in the percept that is the combination of the two (e.g., /bla/). Cutting defines Phonological Fusion as “when two inputs, each of  $n$  phonemes, yield a response of  $n + 1$  phonemes.” (Cutting, 1976, p 121). In the case of the example, each of the inputs has 2 phonemes, but the response has 3 ( $n+1$ ) phonemes. In other words, two consonants presented in two different auditory streams can be fused to form the percept of a consonant cluster.

Although Phonological Fusion has not been experimentally assessed in the multimodal domain, some evidence for its existence comes from McGurk and MacDonald (1976). In the original configuration, visual velar consonants paired with auditory bilabials result in an averaged percept (e.g., an alveolar consonant). However, when the configuration was reversed, and subjects were presented with a visual bilabial (e.g., “baba”) and an auditory velar (e.g., “gaga”), a Combination Response occurred (such as “gabga”, “bagba”, “gaba”, or “baga”). Although it is a perceptual fusion, the Combination Response appears to be governed by different rules than the true McGurk Effect. Combination Responses were given much less frequently than true McGurk Effect responses (90% for the McGurk response, versus 49% for the Combination Response), and could evoke four different possible percepts as opposed to just one. The frequency of occurrence and number of possible unique percepts suggests that the level at which the information is integrated is different for the McGurk Effect and the Combination Response. In the McGurk Effect, neither the auditory nor visual information is present in the final response; instead a fusion occurs at the featural level. The Combination Response, on the other hand, contains a sequence of phonetic segments that are specified by both visual and auditory streams, and appears to be more similar to auditory Phonological Fusion rather than Psychoacoustic Fusion.

Although the Combination Response was observed by McGurk and MacDonald (1976), a true multimodal analogue to unimodal Phonological Fusion has not been documented. If other unimodal fusions (e.g., Sound Localization and Psychoacoustic Fusion) have multimodal analogues (e.g., Ventriloquist Effect and McGurk Effect), it appears likely that a multimodal fusion corresponding to unimodal Phonological Fusion also exists. The purpose of the present study was to investigate whether

Phonological Fusion does indeed have a true multimodal analogue (Audio Visual Phonological Fusion or AVPF).

Understanding the conditions under which different fusions occur may provide additional insight into the integration of conflicting information for speech perception. In the case of unimodal Phonological Fusion, conflict between the two ears leads to a blending of the auditory streams, whereas in the McGurk effect, conflict between the auditory and visual modalities leads to an averaging of the auditory and visual streams. Understanding how the brain resolves conflicting information may lead to a better understanding of the operation of the perceptual processes underlying both unimodal and multimodal speech perception, and may suggest that a common domain general substrate.

Here, we report three experiments that pair visual stop consonants (e.g., “back”) with auditory liquids (e.g., “lack”) to determine whether true phonological fusions are reported (e.g., “black”). The first experiment used an open set recognition task to assess the occurrence of AVPF by presenting listeners with initial stop consonants at three places of articulation (bilabial, alveolar and velar), paired with one of two liquids (/l/ and /r/). In the second experiment, we assessed the effects of lexicality on open set recognition by comparing the rate of AVPF in words and nonwords. In the third experiment, we replicated the findings of the first two using a closed-set of response alternatives to further constrain the possibility of observing AVPF in other contexts.

## Experiment 1

### Methods

#### Stimulus Materials

The words and nonwords used as stimuli in all three experiments were modeled after those used by Cutting and Day (1975) and Cutting (1975; 1976) in dichotic listening experiments and by McGurk and MacDonald (1976). These specific stimuli were selected because they had been shown to successfully elicit either unimodal Phonological Fusion or the McGurk Effect. The specific stimulus set used is shown in Table 1. These stimuli were selected to compare fusion rates across three different places of articulation (bilabial, alveolar, velar) and across the two liquids /l/ and /r/.

Place of Articulation	Visual Stop	Auditory liquid		Fused response	
		/l/	/r/	stop + /l/	stop + /r/
Bilabial	back	lack	rack	black	brack
	pay	lay	ray	play	pray
Alveolar	dead	led	red	dled	dread
	tie	lie	rye	tly	try
Velar	go	low	row	glow	grow
	camp	lamp	ramp	clamp	cramp

**Table 1:** Stimulus set used in Experiment 1

All stimuli were recorded using a Canon GL1 video camera and lapel microphone (Shure mx-100). The talker was a male, native English speaker, who reported no history of speech or hearing disorders at the time of testing. A computer screen located below the camera displayed the words for the talker to read. One researcher operated the camera, while another monitored for pronunciation errors.

Any errors in pronunciation were noted, and the talker asked to repeat the mispronounced words at the end of the session. Two repetitions of the stimulus materials were originally recorded.

Stimuli were edited using Final Cut Pro 5.0.1 on a Macintosh Powerbook G4. The beginning and ending of each stimulus were identified using both visual inspection of the waveform and auditory discretion. To create the experimental files for presentation, an additional 15 frames (approximately 500 ms) was added before and after the target stimulus. When this method resulted in an unusual beginning or ending of the visual display (e.g., blinking) an additional frame was appended or an extra frame was removed.

Congruent stimuli were used as control items, and contained the same auditory and visual target (e.g., auditory “lack” and visual “lack”). Incongruent stimuli were created by pairing the auditory stimulus from one recording (e.g., “lack”) with the visual signals from a different recording (e.g., “back”). All permutations of auditory and visual signals were created for each syllable (5 congruent and 20 incongruent stimuli). All incongruent stimuli attempted to splice utterances with the closest duration. When durations did not match exactly, the beginnings of the stimuli were aligned, and overall differences in duration for the two constituent portions of a stimulus never exceeded two frames (approximately 67 ms). Additionally, audiovisual and dichotic listening fusions have been shown to be robust across a relatively large window of asynchrony, from 30 milliseconds auditory lead to 175 milliseconds visual lead (van Wassenhove, Grant & Poeppel, 2006); thus slight temporal asynchronies in the onset and offset of the stimuli should not affect of the experimental results.

### **Participants**

Twenty-five undergraduate students at Indiana University participated in the study. Each received partial course credit for their participation. All were native speakers of English, reported no history of speech or hearing impairment, and had normal or corrected-to-normal vision.

### **Procedure**

The experiment took place in a quiet room with multiple testing stations. The experiment was run using a custom script written for PsyScript on four Macintosh G3 computers. Participants viewed stimuli on fifteen-inch CRT monitors and listened through Beyer Dynamic DT-100 headphones. Stimuli were presented at a comfortable listening level (approximately 65 dBA) for all participants.

Stimulus presentation was blocked in order to eliminate the potential for repetition priming. Stimuli were randomized within each block, but the blocks were always presented in the same order. The first block consisted of 36 Incongruent stimuli in which neither constituent contained a cluster (e.g., visual “back”, auditory “lack”). The second block consisted of 84 Incongruent stimuli in which one or both of the constituents contained a cluster (e.g., visual “black”, auditory “lack”). The third block served as a control, and consisted of 30 Congruent stimuli (e.g., visual “lack”, auditory “lack”).

Participants were instructed to both watch and listen during the experiment and great care was taken not to bias their attention to either modality. A fixation cross at the center of the screen preceded each stimulus and was followed by a 500 ms delay. A dialog box appeared directly after each stimulus, prompting subjects to type what they thought the talker said. They were told that their responses could be real words or nonwords, and were encouraged to check that they had typed their intended response to minimize typographic errors. No time limit was imposed for a participant’s response, and each trial was separated by a 1000 ms intertrial interval.

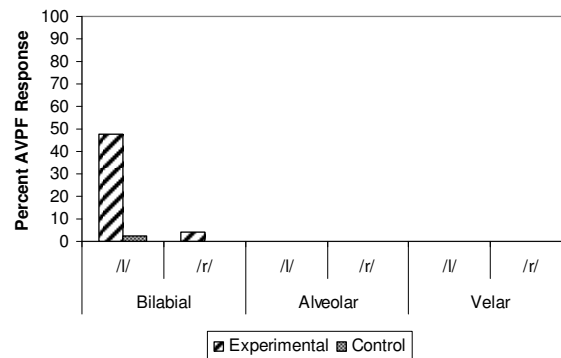
## Data analysis

For the purposes of data analysis, only incongruent stimuli containing a visual stop (e.g., “back”, “tie”) and an auditory liquid (e.g., “lack”, “rye”) were included in the experimental stimuli (n=12). Congruent stimuli containing a liquid (e.g., “lack”, “rye”, etc.) served as the control stimuli (n=12). These specific stimuli were selected in order to assess the frequency of occurrence of AVPF and to explore the effects of place of articulation and liquid category on fusion rate. Subjects’ responses were coded as either containing a consonant cluster or not. Because the perception of stop voicing is virtually imperceptible from visual information alone, both voiced and voiceless clusters were considered acceptable fusions, regardless of whether the actual stimulus was a voiced or voiceless bilabial (e.g., “black” or “plaque” would both be coded as a fused response for visual “back” and auditory “lack”). Other responses that did not contain a consonant cluster such as “back”, “lack”, or “mack” were coded as containing no fusion.

## Results

A repeated measures ANOVA with stop Place of Articulation (bilabial, alveolar, velar), Liquid Type (/l/, /r/), and Experimental Condition (Congruent, Incongruent) as within-subjects factors was conducted on the data. The ANOVA revealed significant main effects of Place of Articulation ( $F(2,46) = 36.33, p \leq 0.001$ ), Liquid Type ( $F(1,23) = 26.01, p \leq 0.001$ ), and Experimental Condition ( $F(1,23) = 30.67, p \leq 0.001$ ). More fusions were reported for bilabial stops than for alveolar or velar places of articulation (Figure 1). Similarly, across all places of articulation, more fusions occurred for /l/ than for /r/. Finally, more fusions were reported for the experimental condition (e.g., Incongruent stimuli) than the control condition (e.g., Congruent stimuli), indicating that listeners do reliably perceive AVPF in incongruent stimuli with visual stops and auditory liquids.

In addition to these main effects, all two-way and three-way interactions reached significance (all  $p$ -values  $\leq 0.001$ ). All of these results, however, were driven by the presence of fusions in stimuli with a visual bilabial and an auditory /l/, as illustrated in Figure 1. Post-hoc paired-sample t-tests revealed that fusions in the bilabial + /l/ experimental condition were significantly greater than the fusions in the control /l/ + /l/ condition (48% vs. 2% respectively) ( $p \leq 0.001$ ). For all other places of articulation and liquids, the difference between experimental and control conditions did not reach significance (all  $p$ -values  $\geq 0.162$ ). Although some fusions were recorded in the bilabial + /r/ experimental stimuli, the fusion rate did not differ significantly from the controls (4% vs. 0%).



**Figure 1.** Percent of AVPF responses for visually presented stops at different places of articulation and auditorily presented liquids compared to controls in Experiment 1.

## Discussion

Experiment 1 demonstrated the existence of AVPF, but its occurrence was shown to be limited to certain specific conditions. In particular, when a visual bilabial stop (i.e., /b/ or /p/) is presented simultaneously with an auditory /l/, a consonant cluster, /bl/ or /pl/, was perceived. AVPF was not observed when the visual signal contained stop consonants at the other two places of articulation (i.e., /d/, /t/, /g/, or /k/). Additionally, AVPF did not occur with /r/ even when paired with a visual bilabial stop. It is possible that the lack of fused responses with alveolar and velar stops is due to the absence of salient visual cues to these places of articulation. More importantly, stops at these places of articulation are not visually distinctive from the liquid /l/. All of the results found in Experiment 1, and those of other audiovisual fusions, can be explained by the degree and type of conflict between the auditory visual information.

Results from McGurk and MacDonald (1976) and this experiment suggest that AVPF occurs when the information from the two modalities conflict. The resolution of these conflicts yields a robust percept that includes phonetic attributes of both the visual and auditory information. In AVPF, a visual /b/ conflicts with an auditory /l/. The visual information of /b/ strongly specifies the presence of a bilabial, but no bilabial cues are present in the auditory stream of information. This conflict results in the percept of a cluster of phonemes, a serial combination of the phonemes from the visual and auditory streams of information.

Audiovisual conflict, or lack thereof, also explains the absence of AVPF with velars. The visual cues to /k/, /g/, and /l/ are not as perceptually distinct from each other as bilabials and /l/ are. Because the auditory and visual cues do not conflict, AVPF is not observed with visual velar stops paired with auditory /l/. Similarly, a lack of conflict can explain the absence of AVPF with alveolar stops, however, the lack of AVPF with alveolars and /l/ is more likely due to the phonotactics of English which prohibit clusters of this type in word-initial position (i.e., \*/tl/ and \*/dl/). This explanation suggests that listener's responses are highly constrained by their knowledge of phonology.

In addition to examining the effect of place of articulation on AVPF, Experiment 1 also assessed differences between the two liquids /r/ and /l/. Whereas AVPF was present for /l/ paired with bilabials, it was not observed at a significant rate for /r/ paired with any stops. The lack of AVPF with the liquid /r/ may have been due to its visual properties. In English, /r/ is produced with lip rounding. The bilabial stops are also produced with a labial gesture, although in this case, the gesture is complete lip closure. Because both bilabial stops and /r/ are produced with a labial articulation, less conflict is present between these segments than there is between the bilabial stop and /l/. We hypothesize that the occurrence of AVPF is dependent on the degree of difference between expected and actual visual information. Both bilabial stops and /r/ are produced with labial gestures, so there is less conflict, and thus no fusion is observed.

Although some degree of conflict exists between auditory /r/, which includes a labial gesture, and velar and alveolar stops, which lack a labial gesture, the type of conflict is quite different than for visual bilabial stops and auditory /l/. In the latter combination, where AVPF occurs reliably, the cues to labiality come from the visual information. In the former, similar to the McGurk Effect, the cues to labiality are specified in the auditory stream. In these cases, an auditory "labial" conflicts with a visual "non-labial". The visual cues provide information for a non-visually salient articulation, which is highly similar to the visual information of other segments (e.g., /d/). In the McGurk Effect, the conflict between auditory and

visual information results in the percept of a phoneme that is not labial, matching the visual information, but is acoustically similar (i.e., /da/).

The results from Experiment 1 and previous audiovisual studies support an interpretation of perceptual fusion that results from perceptual conflict. When conflict is high, observers perceive an utterance that fuses the information either at a subsegmental, featural level or at a segmental, syllable level. In Experiment 2, we sought to further explore the effect of perceptual conflict in AVPF and replicate the results found in Experiment 1. Since some fusions with /r/ were reported, we increased the number of trials in Experiment 2 to examine the frequency of AVPF with /r/. In addition, some of the fused responses in Experiment 1 produced valid English words. In Experiment 2 we examined the effect of lexicality on AVPF by including both word and nonword stimuli. Since no fusions were reported for the alveolar and velar stop consonants, Experiment 2 examined only performance on bilabial stops.

## Experiment 2

### Methods

#### Stimulus Materials

Stimuli for Experiment 2 (Table 2) were drawn from the same set of materials used in Experiment 1, but included both words and nonwords in order to test the effects of lexicality. These stimuli were selected to compare the rate of AVPF in the two liquid conditions (/l/ and /r/) and in both words and nonwords.

Lexicality	Visual Stop	Auditory liquid		Fused response	
		/l/	/r/	stop + /l/	stop + /r/
Word	back	lack	rack	black	brack
	pay	lay	ray	play	pray
Nonword	baba	lala	rara	blabla	brabra
	papa	lala	rara	plapla	prapra

**Table 2:** Stimulus set used in Experiment 2.

#### Participants

Twenty-six participants took part in this experiment. All participants were undergraduate students at Indiana University and either received partial course credit or were paid \$10.00 per hour for their participation. All participants were native speakers of English, reported no history of speech or hearing impairment, and had normal or corrected-to-normal vision.

#### Procedure

The experiment took place in a quiet room with multiple testing stations. The program for the experiment was written in PsyScript and was run on Macintosh G3 computers. Participants viewed stimuli on CRT monitors and listened through Beyer Dynamic DT-100 headphones. Stimuli were presented at a comfortable listening level (approximately 65 dBA) for all participants.

Experiment 2 was divided into three blocks. Blocks were very similar to those in Experiment 1, except that each stimulus was presented twice within each block. The first block consisted of 48 incongruent stimuli in which neither constituent contained a cluster (e.g., visual “back”, auditory “lack”). The second block consisted of 112 incongruent stimuli in which one or both of the constituents contained a cluster (e.g., visual “black”, auditory “lack”). The third block consisted of 40 congruent stimuli and served as the control block (e.g., visual “lack”, auditory “lack”). Stimuli were randomized within these blocks, but the blocks were always presented in the same order to eliminate the potential for repetition priming by cluster-initial words earlier in the experiment.

Participants were instructed to both watch and listen during the experiment and great care was taken not to bias their attention to either modality. A fixation cross at the center of the screen preceded each stimulus and was followed by a 500 ms delay. A dialog box appeared directly after each stimulus, prompting subjects to type what they thought the talker said. They were told that their responses could be real words or nonwords, and were encouraged to check that they had typed their intended response to minimize typographic errors. No time limit was imposed for a participant’s response, and there was a one second intertrial interval.

### Data Analysis

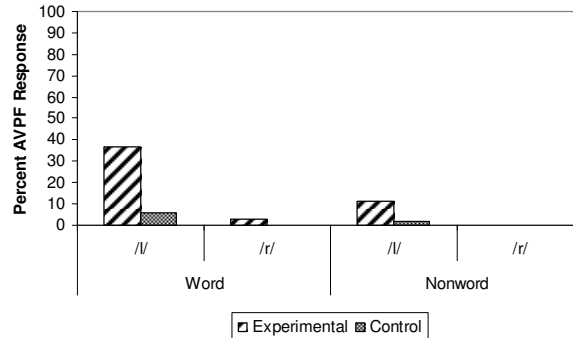
For the purposes of data analysis, only incongruent stimuli with a visual stop (e.g., “back”) and an auditory liquid (e.g., “lack” or “rack”) were included in the experimental stimuli ( $n=8$ ). Congruent stimuli with a liquid (e.g., “lack”, “ray”, etc.) served as the control stimuli ( $n=8$ ). These two sets were selected to determine the degree of AVPF in the two liquid conditions and the two lexical conditions since all fused responses are valid English words. For words, participant responses were coded as either resulting in a consonant cluster (ignoring voicing alternations) or resulting in no consonant cluster. Other responses with no consonant cluster such as “pay” or “lay” were coded as non-fusions. For nonword stimuli, participant responses were coded using both a stringent and a lenient measure. The stringent measure scored a fusion as successful only if fusions were reported at both word-initial and intervocalic positions (e.g., “blabla”), whereas the lenient standard scored a fusion as successful if a fusion occurred at either position (e.g., “blaba”, “babla”, and “blabla”). Both measures are reported below, although we focus on the lenient measure.

### Results

A 2x2x2 repeated measures ANOVA with Lexicality (word, nonword), Liquid Type (/l/, /r/), and Experimental Condition (experimental, control) as within-subjects factors was conducted on the data. The ANOVA revealed significant main effects of Lexicality ( $F(1, 25) = 10.155, p = 0.004$ ), Liquid Type ( $F(1, 25) = 29.436, p < 0.001$ ), and Experimental Condition ( $F(1, 25) = 35.708, p < 0.001$ ). More fusions were reported in response to words than to nonwords. Similarly, more fusions occurred for /l/ than for /r/. Finally, more fusions were reported in the experimental condition than in the control condition, indicating that listeners perceived AVPF when presented with incongruent stimuli consisting of visual bilabial stops and auditory liquids. Results are summarized in Figure 2.

In addition to the main effects, all two-way interactions reached significance (all  $p$ -values  $\leq 0.016$ ). The three-way interaction also reached significance ( $p = 0.027$ ). As in Experiment 1, all effects were driven by the large number of reported fusions in the word bilabial plus /l/ experimental condition. Post-hoc paired samples t-tests revealed significant differences between the experimental and control conditions only for the word ( $p \leq 0.001$ ) and nonword ( $p = 0.015$ ) stimuli consisting of visual bilabials and auditory /l/ (36.54% vs. 5.77% for words and 11.54% vs. 0.00% for nonwords). However, using the

stringent coding method described above, the rate of fusion in response to nonwords containing visual bilabials and auditory /l/ was not significantly greater than in the control condition ( $p = 0.212$ , 3.85% vs. 0.00%). No other conditions produced fusions that were significantly different from zero (all  $p$ -values  $\geq 0.185$ ). Although some fusions were observed in the word bilabial + /r/ experimental condition, they did not differ significantly from the control condition (2.88% vs. 0%). This result supports the finding from Experiment 1 that auditory /r/ does not induce the perception of AVPF. Additionally, post-hoc comparisons between the four experimental conditions revealed that the rate of AVPF in words with bilabials + /l/ condition was greater than in all other experimental conditions (all  $p$ -values  $\leq 0.004$ ).



**Figure 2.** Percent of fused (cluster) responses for visual stops and auditory liquids in monosyllabic words and disyllabic nonwords in Experiment 2.

## Discussion

The findings from Experiment 2 replicate and extend the findings from Experiment 1. AVPF was again observed with visual bilabial stops and auditory /l/. Lexicality affected the rate of fusion as AVPF was perceived more often in response to words than to nonwords. AVPF did not occur at a significant rate in response to words or nonwords with the liquid /r/. However, for the trials containing words with the liquid /r/ some fusions were reported, but not enough to reach statistical significance.

Although AVPF did occur in response nonwords, the fusion was much less prevalent than it was in response to words. High-frequency words are perceived more easily than low-frequency words (Broadbent, 1967), so by extension, nonwords, which essentially have a lexical frequency of zero, should produce fewer fusions than words. Interestingly, some of the monosyllabic words resulted in fused percepts that were nonwords (e.g., “blay”) even though their constituents were words (e.g., “pay”, “lay”). This discrepancy suggests that the lower rate of fusion is not caused by the target fusions being nonwords, but rather because the constituents are nonwords.

Since both Experiments 1 and 2 used an open-set response format, it is possible that additional fusions could occur, but that they were not being reported frequently enough to reach significance with the current methodology. Subjects may be less likely to type a non-word response than a word even if their fused percept was closer to a nonword. The third experiment was designed to replicate the findings from the second experiment using a closed-set response paradigm. In this procedure, participants could only select one of six response options (selected from the most common responses indicated in the open set, as well as other possible target fusions) rather than freely typing their response. We hypothesized that

the closed-set response methodology would result in a larger number of fused responses and yield a more stable estimate of AVPF.

### Experiment 3

#### Methods

##### Stimulus Selection

Stimuli for Experiment 3 were the same as those used in Experiment 2, (word and nonword sets of bilabials stops, Table 3) and were selected to compare fusions of the liquids /l/ and /r/ and words and nonwords.

Lexicality	Visual Stop	Auditory liquid		Fused response	
		/l/	/r/	stop + /l/	stop + /r/
Word	back	lack	rack	black	brack
	pay	lay	ray	play	pray
Nonword	baba	lala	rara	blabla	brabra

**Table 3:** Stimulus set used in Experiment 3

#### Participants

Thirty participants took part in this experiment. All participants were undergraduate students at Indiana University and they either received partial course credit or were paid \$10.00 for their participation. All participants were native speakers of English and reported no history of speech or hearing impairment and had normal or corrected-to-normal vision.

#### Procedure

The procedure for Experiment 3 was the same as that for Experiment 2 except for the response method. Whereas Experiments 1 and 2 obtained open-set responses from participants, Experiment 3 required participants to choose from among six response options selected based on the responses that were most common for those conditions in Experiments 1 and 2. Response options included each constituent (i.e., the auditory and visual components) (e.g., “lack” and “back”), the predicted AVPF (e.g., “black”), and the theoretical feature-level fusion that might occur if the auditory and visual components were presented in the other modality (e.g., “dack”). Two additional response options were included. For nonwords, response options included only those with fusions at both locations (e.g., “blabla”). The response options for each stimulus are provided in Table 4.

A fixation cross at the center of the screen preceded each stimulus and was followed by a 500 ms delay. Immediately after the end of the stimulus, six boxes of equal size containing a possible response alternative appeared on the screen. Participants used a mouse to select the response option that was closest to what they thought the talker said. Participants were instructed to both watch and listen during the experiment and that their responses could be either real words or nonwords. Great care was taken not to bias the subjects’ attention to either modality. No time limit was imposed and a 1 second intertrial interval separated each trial.

Condition	Lexicality	Video/Audio	Fused Responses	Unfused Responses
Experimental	Word	back / lack	black	lack, back, dack, rack, brack
		back / rack	brack	rack, back, dack, lack, black
		pay / lay	blay, play	pay, lay, tay, day
		pay / ray	bray, pray	ray, pay, tay, bay
	Nonword	baba / lala	blabla	lala, baba, dada, rara, brabra
		baba / rara	brabra	rara, baba, dada, lala, blabla
		papa / lala	blabla, plapla	lala, papa, tata, dada
		papa / rara	brabra, prapra	rara, papa, tata, dada
Control	Word	lack / lack	black	lack, back, dack, rack, brack
		rack / rack	brack	rack, back, dack, lack, black
		lay / lay	blay, play	lay, pay, tay, day
		ray / ray	bray, pray	ray, pay, tay, lay
	Nonword	lala <sub>n</sub> / lala <sub>n</sub>	blabla	lala, baba, dada, rara, wawa
		rara <sub>n</sub> / rara <sub>n</sub>	brabra	rara, baba, dada, lala, wawa
		lala <sub>n</sub> / lala <sub>n</sub>	blabla,	lala, papa, tata, rara, wawa
		rara <sub>n</sub> / rara <sub>n</sub>	brabra	rara, papa, tata, lala, wawa

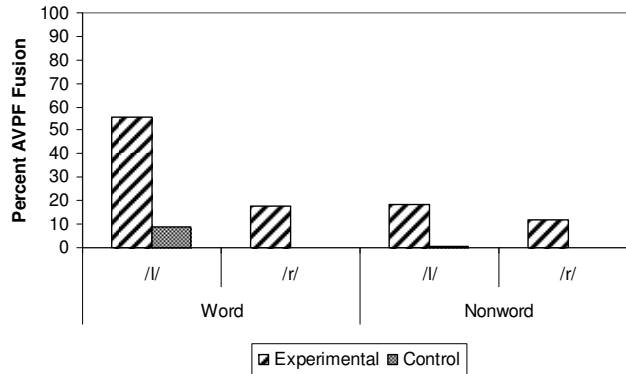
**Table 4:** Response options for stimuli of interest in Experiment 3.

### Data Analysis

As with Experiment 1 and 2, only incongruent stimuli with a bilabial stop (e.g., “back”) and a liquid (e.g., “lack”) were included in the experimental stimuli (N=8). Congruent stimuli with a liquid in both the visual and auditory domains (e.g., “lack”) served as the control stimuli (N=8). These two sets were selected to determine the degree of AVPF. Participant responses were coded as either resulting in a consonant cluster (ignoring voicing alternations). Other responses with no consonant cluster, such as “pay”, “lay”, or “ray”, were coded as no fusion.

### Results

A 2x2x2 repeated measures ANOVA using fusion rate as the dependent variable and Lexicality (word, nonword), Liquid Type (/l/, /r/), and Experimental Condition (experimental, control) as within-subjects factors was conducted on the data. The ANOVA revealed main effects of Lexicality ( $F(1, 29) = 24.794, p \leq 0.001$ ), Liquid Type ( $F(1, 29) = 33.740, p \leq 0.001$ ), and Experimental Condition ( $F(1, 29) = 35.263, p \leq 0.001$ ). More fusions were reported for words than for nonwords and for /l/ than for /r/. Additionally, more fusions were reported in the experimental condition than in the control, indicating that listeners perceived AVPF in incongruent stimuli comprised of visual bilabial stops and auditory liquids. Results are presented in Figure 3.



**Figure 3:** Percent of fused (cluster) responses for visual stops in monosyllabic words and disyllabic nonwords and different auditory liquids in Experiment 3.

In addition to the main effects above, all two-way interactions were significant (all  $p$ -values  $\leq 0.001$ ), as well as the three-way interaction ( $p = 0.009$ ). Post-hoc paired samples t-tests revealed that fusions in the word bilabial + /l/ experimental condition were significantly greater than in its corresponding control condition (55.83% vs. 9.17% respectively) ( $p \leq 0.001$ ). In addition to these expected fusions, AVPF also occurred more in the experimental condition than the control condition for words + /r/ (17.50% vs. 0.00%), for nonwords + /l/ (18.33% vs. 0.83%) and for nonwords + /r/ (11.67% vs. 0.00%) (all  $p \leq 0.011$ ). These findings show that the rate of reported fusions in each individual experimental condition was significantly greater than the corresponding control condition. Thus, AVPF can occur in more conditions than suggested from Experiments 1 and 2. Additional paired-samples t-tests were conducted to examine the rate of AVPF in the four experimental conditions. Results revealed that fusions in the word + /l/ condition were reported significantly more often than in the other conditions ( $p < 0.001$ ) but that fusions in response to the other experimental conditions (e.g., word + /r/, nonword + /l/ or nonword + /r/) did not significantly differ from each other (all  $p$ -values  $\geq 0.147$ ).

## Discussion

The results of Experiment 3 replicated and extended the findings obtained in the previous two experiments. The data reported in Experiment 3 revealed that AVPF occurred most frequently in response to real words that contained a visual bilabial and an auditory liquid /l/ as was found in Experiments 1 and 2. However, changing the response format from open-set to closed-set resulted in listeners reporting fusions in other contexts as well; AVPF was also observed in words with the auditory liquid /r/ and in disyllabic nonwords with /l/ and /r/. AVPF was found significantly more frequently for words compared to nonwords and for the liquid /l/ compared to the liquid /r/. These differences are driven by the strength of the effect in the monosyllabic word + /l/ condition.

Stimuli for this experiment were created to emulate the stimuli that were used successfully in previous McGurk effect studies (McGurk & MacDonald 1976) and as such, the nonwords that we used were disyllabic (e.g., “baba”, “lala”). As a result, we cannot determine from our data whether the differences in rate of fusion between the word and nonword conditions were due to the length of the nonwords (disyllabic vs. monosyllabic) or whether they were due to the lexical frequency of the targets or constituents.

Although the rate of AVPF reported for stimuli with an auditory /r/ reached significance with a closed-set response format, it was still significantly lower than that of stimuli with an auditory /l/. As discussed in Experiment 1, the rate of perceived fusions may be related to the conflict between the visual and auditory stimulus. The most likely context to elicit perceptual fusions is when the conflict between the visual and auditory input is greatest, as with visual bilabials and the auditory liquid /l/. While the conflict between visual bilabials and auditory liquid /r/ is decreased, due to the presence of labial gestures for both segments, it is not eliminated; thus AVPF is found in these conditions as well but to at a lower rate.

## General Discussion

In three perceptual experiments, we demonstrated the existence of a novel audiovisual fusion. AVPF is a multimodal fusion in speech perception in which the simultaneous presentation of a viseme and different auditory phoneme results in the perception of a permissible cluster of two phonemes. In addition to revealing the existence of AVPF, the results of our experiments revealed three major findings about the phenomenon. First, the rate at which AVPF is perceived depends critically on the place of articulation of the visual stop consonant, and occurred only with visual bilabial stops that have highly distinctive visual attributes. Second, AVPF occurs more often with the liquid /l/ than with the liquid /r/ when paired with visual bilabial stops. Third, AVPF is affected by the lexical status of the composite stimuli, occurring more frequently with real words than with nonwords.

We believe that an account of this novel multimodal perceptual fusion (AVPF), as well as the McGurk effect and the unimodal, dichotic listening fusions, must appeal to the degree of conflict between the two inputs, whether they are within a single modality or across separate modalities. In Experiment 1, we found that AVPF was limited to visual bilabial stops paired with an auditory /l/. In this case, the auditory input indicated that a non-labial sound had occurred, but the visual display indicated that labial sound had been produced, thus specifying “labial” to the perceiver. Because visual information for labials is the most salient type of visual speech information (Walden, Prosek, Montgomery, Scherr, & Jones, 1977), the degree of conflict between the auditory input (“not labial”) and the visual display is high. Perceivers resolve the conflict between two salient perceptual cues by combining both in their responses (e.g., “black”). By similar explanation, AVPF does not occur with visual alveolars and velars paired with /l/. Because both the auditory and visual signals suggest “non-labial” targets, the perceptual system has nothing to resolve and the final percept does not contain a consonant cluster. Generally in this case, subjects’ responses are of the auditory signal.

Not only does the notion of conflict account for the differences observed for place of articulation, it can also account for the differences observed between the two liquids. AVPF was found to be less robust for visual labial stops paired with auditory /r/ than auditory /l/. Because /r/ in English is produced with lip rounding, the associated visual gesture is a concomitant labial articulation. Although the labial gestures for /r/ and bilabial stops are not identical, the conflict between the two inputs is reduced, thus decreasing the rate of AVPF. Indeed, in Experiment 3 where AVPF occurred in more contexts than in the previous studies, it still occurred less often with /r/ than with /l/.

Visual alveolar and velar stops paired with /r/ also did not result in AVPF, although the potential consonant clusters (e.g., /dr/, /gr/) are legal sequences in English. In these particular pairings, the auditory input is associated with a labial gesture from /r/, but the visual input is clearly non-labial. In these cases where the visual input is strongly negative for labiality, perceivers frequently report only the auditory signal, but crucially do not report a sequence of two segments.

This final example where the auditory information implies “labial” and the visual information implies “non-labial” resembles the stimulus configuration that results in the McGurk illusion. Similar to AVPF, the McGurk Effect and Combination Responses (e.g., “bagba”, see Introduction) reported by McGurk and MacDonald can be explained by the degree of conflict between the auditory and visual information. The McGurk Effect is observed when the visual cue is “not labial”. On these types of trials, listeners report a single segment which conforms to the visual input (“not labial”) and actually alters the auditory input of /b/ to yield the most anterior non-labial stop, namely /d/. In contrast, AVPF, and the combination responses, occur in the opposite configuration of inputs where the visual cue is “labial”. In these cases, the salient visual labial articulation strongly specifies the presence of a bilabial stop even though one is not present in the auditory stream of information. The auditory information is not altered, but the visual information is added to the auditory stream because nothing in the visual domain contradicts the stronger auditory percept as is the case in the McGurk effect. Thus, for both AVPF and the combination response, the only suitable resolution to the mismatch between the auditory and visual inputs is an output response that contains both consonants. In the McGurk effect the visual information “excludes” certain phonemes, whereas in AVPF the visual information strongly indicates the presence of certain phonemes.

The explanation of fusions as resulting from conflict between two sources of phonetic information also accounts for some results of unimodal fusions. Because dichotic listening experiments are conducted within a single perceptual domain (i.e., audition), there is no inherent inequality between the two inputs. Both AVPF and the McGurk effect result from combining inherently unequal types of input. Although visual information can enhance auditory speech perception (Sumbly & Pollack, 1954), the auditory stream is the dominant perceptual channel in normal-hearing listeners. It is no surprise that both of these effects show that audiovisual conflict depends on the one cue in which the visual domain may be superior (Summerfield, 1987), namely labial vs. nonlabial.

In contrast, unimodal fusions observed in dichotic listening studies have no inherent dominance of information. Thus, the unimodal analogue to AVPF extends from bilabial stops paired with /l/ to alveolar and velar stops and to /t/, yielding a larger set of clusters (e.g., /gl/, /gr/, /dr/, /br/). In these cases of auditory Phonological Fusion, a conflict exists between the manner of articulation between the two inputs, creating a perception that includes both segments. In the unimodal analogue to the McGurk effect (i.e., Psychoacoustic Fusion), the manner of articulation is the same; all that conflicts is the place of articulation. In these cases, the percept is of the same manner (i.e., a stop) and the perceptual system resolves the conflict by perceiving a stop at an intermediate place of articulation. In addition to stop-stop and stop-liquid conflicts, Cutting (1975) conducted dichotic listening studies pairing /s/ with stops. He found that these combinations also resulted in Phonological Fusions (Cutting, 1975). If perceptual conflict is at work in resolving the previous AVPFs, we would expect that auditory /s/ paired with a visual bilabial stop would also result in the perception of /sp/ clusters.

The other main finding of the current studies was the presence of an interaction between rate of AVPF and lexicality. Experiments 2 and 3 revealed that monosyllabic words fused more readily than disyllabic nonwords. In Experiment 2, the more stringent measure of fusions requiring AVPF in both positions of the two syllable nonwords did not reveal a statistically significant increase in fusions over the control condition. In contrast, the more lenient measure, which counted all responses containing at least one AVPF, did result in a significant number of fusions. Even with this latter measure, however, the rate of AVPF in nonwords was significantly lower than in words. The different findings observed for words vs. nonwords may result from their lexical status, although other differences between these two types of stimuli exist.

The most salient difference between the words and nonwords used in this study was length; the real words were monosyllabic and the nonwords were disyllabic. Because longer words or nonwords have more segments that must be aligned, it is more likely that the timing or synchrony between some segments may not coincide precisely. Although disyllabic nonwords are more likely to be asynchronous, this may not be problematic for perceiving Phonological Fusions. Previous studies have reported a large window of asynchrony over which audiovisual speech stimuli are judged as synchronous (Conrey & Pisoni, 2006; van Wassenhove *et al.*, 2006). If the auditory and visual portions of the stimulus come from the same utterance but are presented with one modality temporally ahead of the other, participants still perceive the stimuli as synchronous: at least 131 ms of asynchrony in monosyllabic words (Conrey & Pisoni, 2006) and at least 74 ms of asynchrony for /da/ (van Wassenhove *et al.*, 2006). Furthermore, the McGurk Effect can be perceived over a similar window of audiovisual asynchrony (30 ms auditory lead to 175 ms visual lead) over which identical stimuli are judged as synchronous (van Wassenhove *et al.*, 2006). Finally, dichotic listening Phonological Fusion is also observed over a large window of asynchrony, up to 150 ms delay between the two auditory presentations (Cutting, 1976). Thus, we conclude that the difference in the rate of AVPF for words and nonwords is not due to differences in overall synchrony of the stimuli, but is due to lexical status.

The difference in length between words and nonwords in this study also yields a difference in the location of the target fusions. Dichotic listening Phonological Fusion occurs not only for consonants in word-initial position, but has also been observed for intervocalic and word-final consonants. In the current study, we demonstrated that AVPF can occur word initially and we have provided some evidence that it can occur intervocalically. Further experiments are needed to fully explore the occurrence of AVPF intervocalically and to test the existence of AVPF word finally. The findings obtained in Experiment 2 suggest that initial position is the most favorable location to perceive AVPF, where it was perceived with /l/ in 10.58% of responses; intervocalic AVPF was only perceived with /l/ in 4.81% of responses.

In addition to lexicality, word frequency could also affect the rate of perceived fusions. In the monosyllabic word conditions, all the constituents were words, while the target fusions were mostly words (e.g., “black”), but included a few nonwords (e.g., “blay”). In the disyllabic nonword condition, the stimuli were nonwords, but were most likely utterances familiar to the participants (e.g., “baba”, “papa”, “lala”, “rara”). In spoken word recognition, high frequency words are perceived more easily and recognition errors tend to be of higher frequency than the frequency of the stimulus (Broadbent, 1967). We would expect the word frequency of the target fusion relative to the frequency of the constituents to have an effect on fusion rate. To our knowledge, no experiments have reported that explore the effect of word frequency on the perception of either multimodal or unimodal fusions. This would be a productive topic of investigation and could potentially provide additional information on whether fusions depend upon the prior linguistic experience of the observer.

Prior linguistic experience has also been demonstrated to play a role in the susceptibility to the McGurk effect. In Finnish, for example, the McGurk effect occurs at a rate similar to English (Sams, Manninen, Surakka, Helin, & Katto, 1998). However, in Japanese, perceptual fusions occur much less frequently (Sekiyama & Tohkura, 1991). A much lower signal-to-noise ratio is required in the auditory stream for native Japanese speakers to exhibit the McGurk effect. In other words, the relative dominance of auditory information must be severely degraded for Japanese listeners to incorporate visual information and perceive a multimodal fusion. Additionally, native speakers of Japanese show no evidence of a combination response when presented with a visual bilabial /b/ simultaneously with an auditory liquid /r/ (Sekiyama & Tohkura, 1991). Since stop-liquid clusters are prohibited in Japanese, the

combination response of /br/ is not a valid percept. Similarly, we would expect that AVPF would be nonexistent in Japanese due to the phonotactics of the language.

The data presented here and in other AV experiments illustrate several parallels between multimodal and unimodal speech perception. Studies of sound localization show that delayed auditory input (unimodal) or a displaced visual input (multimodal, ventriloquist effect) both result in a change in the perception of the location of a sound source. Similarly, fusing the two inputs to yield a single intermediate segment occurs in both unimodal (Psychoacoustic Fusion) and multimodal (McGurk effect) perception. The experiments presented here document another similarity between unimodal and multimodal fusions; the sequential perception of stop-liquid clusters occurs both in unimodal, dichotic listening and in this newly documented AVPF.

Despite these similarities, there is a critical difference that is related to the relative importance or weighting of the two inputs. In dichotic listening, both inputs are auditory and thus equal in terms of the type and amount of information that is conveyed. In AV perception, there is an inherent asymmetry in the relative dominance of the two cues; auditory information carries more robust phonetic information than visual (viz. near ceiling performance of auditory-only speech perception and comparatively low performance – 15.69% of key words in CUNY sentences – for visual-only speech perception; Conrey & Pisoni, 2006).

More specifically, the kind of information that can be contrasted in the visual domain differs, possibly being limited to labial vs. nonlabial articulations. Whether information from the two inputs is equal in importance (unimodal vs. multimodal), this asymmetry between unimodal and multimodal fusions corresponds to differences in the responses. For example, the perception of fused clusters containing alveolars and velars paired with liquids occurs only in unimodal fusions – and not multimodal fusions – where cues to stop place of articulation are retained. Similarly, the combination response versus the McGurk effect depends on which information is presented in each modality. Unimodal, dichotic presentation of two stops does not result in a combination response because the information presented to both ears carries the same perceptual salience (Cutting, 1976).

In summary, we reported three experiments that documented the existence of a novel audiovisual fusion, audiovisual phonological fusion (AVPF). When visual bilabial stops (i.e. /b/ or /p/) and auditory liquids (i.e. /l/ or /r/) are presented simultaneously, observers often perceive a consonant cluster composed of both the stop and liquid (e.g., /bl/ or /br/). Furthermore, we have shown that AVPF depends on both the place of articulation of the visual stop and the identity of the liquid. We argued that the presence of AVPF is directly related to the degree of conflict between the auditory and visual sources of information. Our account of perceptual conflict also explained the earlier results found in other multimodal (e.g., McGurk Effect) and unimodal fusions. Because of the relative importance of the two input streams, what counts as ‘conflict’ is different for unimodal and multimodal fusions. Here we presented evidence suggesting that fusions occur for a basic reason – conflict – which is resolved by mechanisms that are similar for both unimodal and multimodal perception.

## References

- Bertelson, P., Vroomen J., Gelder B.D., & Driver J. (2000). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception and Psychophysics*, 62(2), 321-332.
- Broadbent, D.E. (1967). Word-Frequency Effect and Response Bias. *Psychological Review*, 74(1), 1-15.

- Conrey, B. & Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America*, *119*(6), 4065-4073.
- Cutting, J. E. (1975). Aspects of Phonological Fusion. *Journal of Experimental Psychology*, *104*(2), 105-120.
- Cutting, J. E. (1976). Auditory and Linguistic Processes in Speech Perception: Inferences from Six Fusions in Dichotic Listening. *Psychological Review*, *83*(2), 114-140.
- Cutting, J. E. & Day, R.S. (1975). The perception of stop-liquid clusters in phonological fusion. *Journal of Phonetics*, *3*, 99-113.
- McGurk, H. & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.
- Sams, M., Manninen, P., Surakka, V., Helin, P., & Katto, R. (1998). McGurk effect in Finnish syllables, isolated words, and words in sentences: Effects of word meaning and sentence context. *Speech Communication*, *26*, 75-87.
- Sekiyama, K. & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, *90*, 1797-1805.
- Sumby, W. H. & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212-215.
- Summerfield, Q. (1987). Some Preliminaries to a Comprehensive Account of Audio-visual Speech Perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. (pp. 241-289). Hillsdale, NJ: Lawrence Earlbaum & Associates.
- van Wassenhove, V., Grant K.W., & Poeppel, D. (2006). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607.
- Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., & Jones, C.J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, *20*, 130-145.