THE EFFECTS OF SEPARATING AUDITORY AND VISUAL SOURCES ON AUDIOVISUAL INTEGRATION OF SPEECH

Jeffery A. Jones and Kevin G. Munhall

Queen's University

Kingston, Ontario

Canada

Please send all Correspondence to:

Jeffery A. Jones,
Department of Psychology,
Queen's University,
Kingston, Ontario,
K7L 3N6
CANADA

Telephone: (613) 545-6012
Fax: (613) 545-2499

E-mail: jonesj@pavlov.psyc.queensu.ca

CITATION:

Jones, J. A., and Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. Canadian Acoustics. 25 (4), 13-19.

THE EFFECTS OF SEPARATING AUDITORY AND VISUAL SOURCES ON AUDIOVISUAL INTEGRATION OF SPEECH

Jeffery A. Jones and Kevin G. Munhall
Queen's University
Kingston, Ontario,
K7L 3N6

ABSTRACT

When the image of a speaker saying the bisyllable /aga/ is presented in synchrony with the sound of a speaker saying /aba/, subjects tend to report hearing the sound /ada/. The present experiment explores the effects of spatial separation on this class of perceptual illusion known as the McGurk effect. Synchronous auditory and visual speech signals were presented from different locations. The auditory signal was presented from positions 0°, 30°, 60° and 90° in azimuth away from the visual signal source. The results show that spatial incongruencies do not substantially influence the multimodal integration of speech signals.

SOMMAIRE

Lorsqu'on présente simultanément l'image d'une personne prononçant la bisyllabe /aga/ et le son /aba/, les participants ont tendance à dire qu'ils ont entendu /ada/. Cette illusion est connue sous le nom d'effet McGurk. La présente étude explore les conséquences perceptives de la séparation spatiale entre les sources visuelle et sonore sur l'effet McGurk. Un signal auditif était presenté a 0, 30, 60, et 90 degrés en azimuth par rapport au signal visuel. Les resultats demontrent que les paramètres spatiaux n'ont que peu d'influence sur l'intégration visuo-auditive des signaux.

1. INTRODUCTION

One of the most elegant demonstrations of multisensory integration in humans is observed in speech perception. It is well known that watching a speaker's mouth movements while listening to speech in noisy environments enhances intelligibility (Miller, Heise & Lichten, 1951; Sumby & Pollack, 1954; Walden, Prosek, Montgomery, Scherr & Jones, 1977). McGurk and MacDonald (1976) demonstrated that visual information also affects the perception of speech in situations with perfectly audible acoustic signals. When speech sounds such as /ba/ were presented in synchrony with an image of a speaker saying /ga/, subjects reported hearing a different syllable, /da/. Other combinations of auditory and visual stimuli similarly yield "blended" percepts (e.g. auditory /ba/ and visual /da/ produce the perception of a /va/ syllable). In addition, certain manipulations cause participants to perceive both the auditory and visually presented syllables. For example, showing observers a visual /ba/ while they hear a /ga/ causes them to report hearing /bga/. This class of perceptual illusions has been labeled the "McGurk effect" and is a well established phenomenon (e.g., Green & Kuhl, 1989, 1991; MacDonald & McGurk, 1978; Manuel, Repp, Studdert-Kennedy, & Liberman, 1983; Massaro, 1987; Massaro & Cohen, 1983; Munhall, Gribble, Sacco, & Ward, 1996; Summerfield & McGrath, 1984). However, the particular conditions that affect the audiovisual integration of speech, as well as how the integration occurs, remain unidentified. In this study, we address the boundary conditions governing integration by studying the influence of spatial location on the McGurk effect.

Recently, Radeau (1994) and others have suggested that the audiovisual processing of speech represents an example of modular perceptual processing. In Radeau's view, speech is not subject to the same constraints as other types of audiovisual perception. Cross-modal information regarding nonspeech events seems to be integrated based on similar rules proposed by Gestalt psychologists for visual grouping; namely common fate and proximity (e.g., Bermant & Welch, 1976; Bertelson, 1993; Bertelson & Radeau, 1981; Jack & Thurlow, 1973; Radeau & Bertelson, 1977,1978; Welch & Warren, 1980). However, audiovisual speech integration persists when the rules are violated in the temporal domain (Massaro, Cohen & Smeele, 1996; Munhall et al., 1996. Very little work has been done on the effects of spatial separations between auditory and visual sources on the McGurk effect. Fisher and Pylyshyn (1994)

report that spatial separations do not reduce the effectiveness of audiovisual stimuli in producing the McGurk effect. Bertelson, Vroomen, Wiegeraad and de Gelder (1994) confirmed this finding, however, both studies used relatively small spatial separations not exceeding 24° (B. D. Fisher, personal communication, November 22, 1995) and 37.5° (Bertelson et al. , 1994) to the right and left of the visual stimulus. Sharma (1989) did use larger spatial separations of 60° to the left and right of the visual stimulus and his experiment showed a small effect of spatial separation on the McGurk effect. However, the results were difficult to interpret because the effect was not consistent for the left and right side of the visual stimulus.

Our experiment was designed to determine whether the strength of the McGurk effect would be influenced by extreme spatial conflicts between the source of the auditory and visual stimuli. It may be that the failure of studies to find a consistent reduction in the audiovisual integration of speech signals has occurred because too small spatial discrepancies have been used. The auditory signal was presented from positions 0°, 30°, 60° and 90° in azimuth away from the visual signal source. It was predicted that if the processing that results in the McGurk effect relies on spatial congruency, then the size of the McGurk effect should decrease as the angular separation between the auditory and visual stimuli sources increases.

2. METHOD

2.1 Subjects

Thirty-six undergraduates at Queen's University, Canada, participated either for credit in an introductory psychology course or were paid volunteers. All subjects were native speakers of Canadian English who reported having either normal or corrected to normal vision and no previous history of hearing or speech disorders. The age of the subjects ranged from 18 to 63 years (M=21.9 years) .

2.2 Apparatus

Stimulus materials and equipment.

The stimuli were selected from a videodisc database created at Queen's University containing Canadian English talkers producing various vowel-consonant-vowel (VCV) bisyllables. Five talkers from the database, 3 females and 2 males between 20 and 30 years of age were used for the experiment. The visual stimuli were the bisyllables /igi/, /IgI/ and /ægæ/ and the auditory stimuli consisted of the bisyllables /ibi/, /IbI/ and /æbæ/ produced by the same talkers. The individual VCV stimuli were not counterbalanced across the five talkers because only stimuli that elicited strong McGurk effects were chosen for the experiment. As a result, 12 stimuli were used in the experiment; five /æbæ/-/ægæ/, four /ibi/-/igi/ and three /IbI/-/IgI/ audiovisual combinations. The audio stimuli were digitized from the sound track of the videodisc at a 22 kHz sampling rate using a 12-bit a/d board (DataTranslation, dt2820).

Stimulus display: Equipment and setup

Seven loudspeakers (Realistic Minumus 7's) were positioned along an arc at 30° intervals starting 2 m to the left of the subject (at 0° in azimuth) and ending 2 m to their right (at 180°). The seven loudspeakers were hidden from the subject by a white curtain hanging in a semicircle in front of the loudspeaker array. Figure 1 shows an overview of the experimental setup. The auditory stimuli were filtered with a 10 kHz cutoff using Frequency Devices (Model 901) analog filters and then amplified before playing at an average of 70 dB (SPL) through the selected loudspeaker. The visual tokens stored on the videodisc were presented by a Pioneer (Model LD-V8000) videodisc player connected to a 20 inch video monitor (Sony Model-PVM 1910).

Consonant identification responses were entered into a keyboard. The same software controlled the videodisc player, and synchronously played the auditory tokens

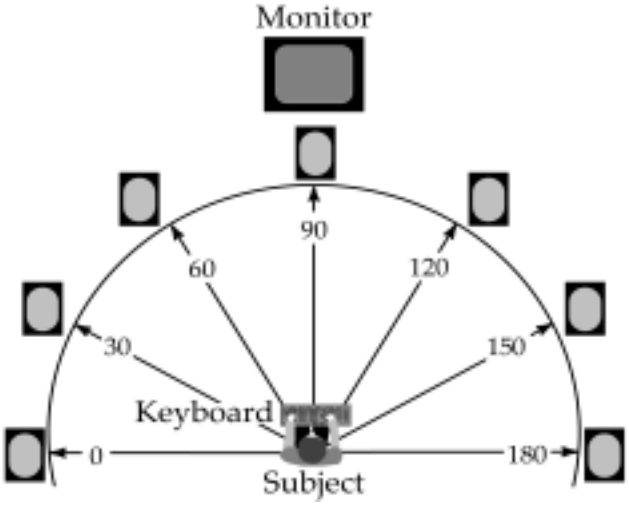

Figure 1: An overview of the experimental setup. Subjects sat facing the video monitor located directly in front of them. Consonant identification responses were made by pressing a key on a keyboard located in front of them. The seven loudspeakers were hidden behind a curtain and located at 0°, 30°, 60°, 90°, 120°, 150° and 180° in azimuth.

through the appropriate loudspeaker. The auditory and visual tokens were synchronized such that the timing of the acoustic burst onset of the /g/ on the videodisc soundtrack for the visual token was aligned with the burst onset of the /b/ of the digitized auditory token. The synchronization allowed the consonant burst alignments to be reliably reproduced (±1 ms).

2.3 Procedure

Subjects were seated 2 m from the video monitor in a 7 by 6.1m room. To minimize trial to trial differences in head position, a subject's head was held firmly in a concave head rest with a forehead strap.

Subjects were asked to report what consonant they heard within the nonsense bisyllables by pressing one of the labeled keys on the keyboard in front of them. They were given the forced-choice option of responding that they heard /b/, /g/, /d/, or some "other" consonant by pressing the B, G, D, or O labeled keys. The key order was counterbalanced across subjects. Subjects were told that they might or might not hear a particular nonsense syllable more than once during the session.

The experiment consisted of five practice and 252 experimental trials. Each auditory stimulus was presented three times from each of the seven loudspeaker locations (12 stimuli x 3 presentations x 7 loudspeakers). The auditory stimulus and the position from which it was presented was randomly selected by the computer on each trial. Following each response, the screen of the video monitor went black for two seconds before the next trial was initiated.

Design

There were three between-subject conditions in the experiment. Twelve subjects were presented with the audiovisual stimuli and required to identify the consonants[1] . Another 12 subjects were required to identify the consonants in the auditory tokens without seeing the visual stimuli. For these subjects, the video monitor was not turned on. The remaining 12 subjects were asked to identify the consonant using the visual information alone. The sound system was not activated for these subjects. To summarize, three independent conditions existed; an *Audiovisual* , *Auditory Only* , and a *Visual Only* condition.

3. RESULTS AND DISCUSSION

The percentage of /b/s that a subject reported hearing was the primary dependent measure for analysis[2] . A clear overall McGurk effect was found in the experiment. The Auditory Only group reported hearing 95.5% /b/s. In comparison, the Audiovisual Group reported hearing only 5.2% /b/s. Subjects in the Visual Only group reported seeing very few /b/s produced on the video monitor (3.4%). Although there was not a difference in percentage of /b/ responses between the Audiovisual and Visual Only conditions [$F(1,33)=0.227$, $p>0.05$], the Audiovisual group reported an entirely different response pattern across all of the possible responses than the Visual group. The overall means and standard errors of the /b/, /g/, /d/ and "other" responses for the three groups are presented in Figure 2. As can be seen in the figure, the Audiovisual group reported hearing many more /d/s than the Visual group [$F(1,33)=121.1$, $p<0.0001$] while the Visual group reported more /g/s than the Audiovisual group [$F(1,33)=44.99$, $p<0.0001$]. Thus, the observed McGurk effect was not due to a substitution of visual information for auditory information but presumably reflected influences from both auditory and visual modalities.

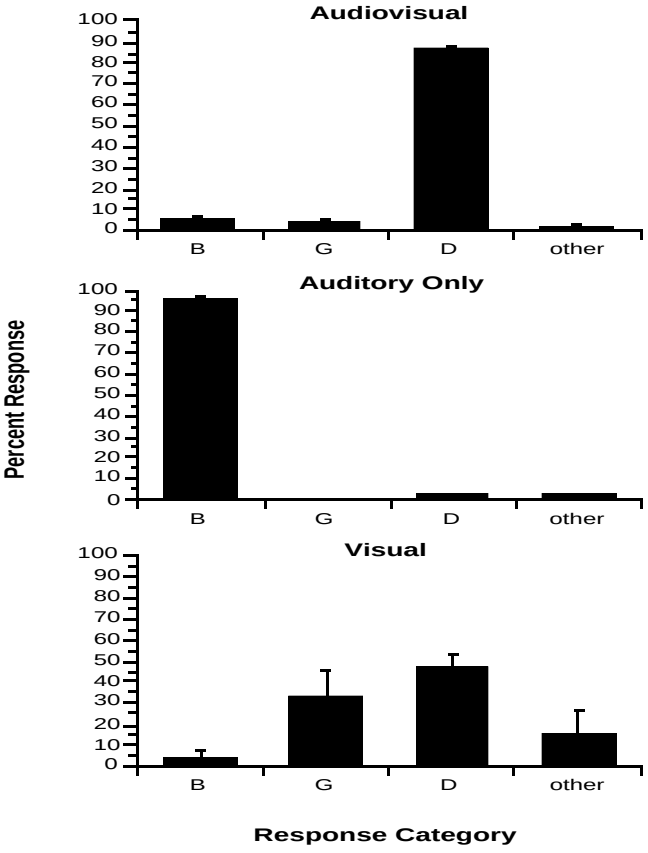

Figure 2: Means and standard errors of the consonant identification responses for the Audiovisual and Auditory Only conditions.

3.1 Analysis by Loudspeaker Location

An analysis of loudspeaker location was performed using only the Audiovisual and Auditory Only groups since the Visual Only group did not receive auditory stimulus presentations. The mean number of /b/ responses that occurred for each loudspeaker location is presented in Figure 3. As noted before, the Auditory Only group reported more /b/s overall than did the Audiovisual group. In addition, a significant location effect was found [$F(6,132)=3.686$, $p<0.01$]. There was no interaction between group and loudspeaker location [$F(6,132)=0.59$, $p>0.5$]. As can be seen in Figure 3, it appears that slightly more /b/ responses were given when the auditory tokens emanated from the right side of the subject versus the left in the Audiovisual group. However, when the mean of the responses for the three speakers on the left was compared with the mean of the three speakers on the right, this right versus left difference was not significant in either the Audiovisual group [$F(1,11)=4.71$, $p>0.05$] or the Auditory Only group [$F(1,11)=3.94$, $p>0.05$].

While no interaction between loudspeaker location and condition was observed, an examination of the means in Figure 3 reveals a small tendency in the Audiovisual condition for the more central speakers to produce a smaller number of /b/ responses. However this difference is extremely small with the difference between the smallest (the central location) and largest /b/ responses being only 1.17%. When the center location is contrasted with the means reported for the other loudspeaker locations, no difference is found [$F(1,11)=2.36$; $p>0.1$].

It appears that the McGurk effect is not greatly influenced by the magnitude of the spatial discrepancy between auditory and visual events. The results show a large McGurk effect even when the angular separation between the auditory and visual sources increases to as much as 90°. As such, these results replicate the findings both of Bertelson et al. (1994) and Fisher and Pylyshyn (1994) but with much larger spatial discrepancies.

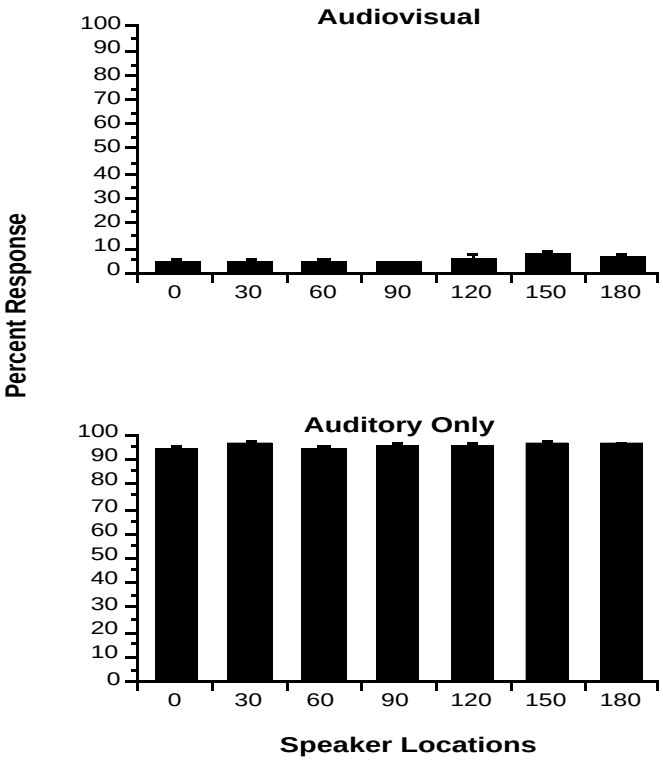

Figure 3: Mean and standard errors of /b/ responses for the Audiovisual and Auditory Only groups that occurred for each loudspeaker location.

4. CONCLUSION

The results of our experiment show that increasing the spatial separation between the auditory and visual stimulus sources has little effect on the McGurk effect. The visual influences on speech perception occur regardless of whether or not the bimodal signals are physically or merely perceptually coincident in space. This finding replicates and extends that of Fisher and Pylyshyn (1994) and Bertelson et al. (1994) by demonstrating McGurk effects for much larger spatial discrepancies. There were small influences of the spatial incongruity but the size of these influences suggests that spatial aspects of the stimuli are not the primary basis of audiovisual integration in speech.

The question arises, on what basis does audiovisual integration take place? Originally it had been our working assumption that information from the two modalities is "glued" together perceptually on the basis of shared amodal properties: The cross-modal equivalent of Gestalt grouping principles (common fate and proximity) might account for audiovisual integration in speech (Radeau, 1994). In this and a previous set of experiments (Munhall et al., 1996; cf., Massaro et al., 1996) we have manipulated coincidence in space and coincidence in time with the expectation that our measure of audiovisual speech integration, the McGurk effect, would be influenced. To our surprise, both sets of studies revealed a remarkable tolerance for incongruity. We are left with two major classes of explanations for our findings:

1. The overall redundancy of audiovisual leaves many bases on which the information from the two modalities could be linked. As Mendelson (1979) noted there is a hierarchy of amodal properties that are available to perceivers for any single object or event. Events are patterned in space and time along a number of dimensions and these patterns can provide many optical and acoustical cues (Gibson, 1966). Speech utterances are complex events that involve

multidimensional visual and auditory patterns. Syllables have onset and offset times and locations in space but they also have durations, rhythms, rates of change, et cetera. In the experiments in Munhall et al. (1996) and the present experiment we have manipulated only the most basic shared amodal properties, one at a time. In the absence of any perceptual competition, presenting the subject with a conflict situation for one property may not seriously stress audiovisual integration. This suggests that multiple property conflict experiments (e.g., manipulating temporal and spatial incongruity simultaneously) may yield more dramatic changes in the McGurk effect than observed in our experiment .

2. The second explanation is that fusion in speech occurs only following independent information processing within a modality (e.g., Kuhl, 1991; Massaro, 1987; Miller, Connine, Schermer & Kluender, 1983; Samuel, 1982; Summerfield, 1987). In this view integration would not be constrained by Gestalt grouping principles applied to the general sensory characteristics of signals. Rather, domain specific information would determine the degree of integration of signals from different modalities. For example, it has been suggested that the time-varying characteristics of speech are used for perceptual grouping and phonetic perception (Remez & Rubin, 1994; Summerfield, 1987). In this view, listeners would extract information about the rate of change of vocal tract shape from both the auditory and visual stimuli and may not be reliant on other information usually thought to be necessary for perceptual grouping.

This suggestion would account for a number of findings about the McGurk effect that indicate that a sense of perceptual unity is not necessary for vision to influence auditory speech perception. Green and Kuhl (1991), for example, have shown that the McGurk effect is present even when subjects know the voice and face don't match in gender. The knowledge that the auditory and visual signals cannot be derived from the same source does not affect the integration of speech. Similarly, Rosenblum and Saldaña (1996) have shown that point light displays of facial movement can influence auditory speech perception in subjects who do not recognize the light motions as facial movements. In both of these experiments the auditory and visual signals share a common time signature but are lacking other significant correspondences.

In closing, the finding that spatial and temporal coincidence has limited influence on the McGurk effect adds to what we feel is a growing list of uncertainties about the McGurk effect. These include individual differences in subjects' perceptions of the effect and individual differences in stimulus effectiveness in evoking the effect (Munhall et al., 1996), influences of familiarity of the faces used as stimuli (Walker, Bruce & O'Malley, 1995), cross linguistic differences (Massaro, 1987; Massaro, Cohen, Gesi, Heredia & Tsuzaki, 1993; Sekiyama & Tohkura, 1991, 1993) and attentional differences (Kuhl, Green & Meltzoff, 1988; Massaro, 1987). This diverse list suggests that we still know little about the subject variables, stimulus parameters, processing limitations and perceptual strategies that govern the McGurk effect.

ACKNOWLEDGMENTS

This research was funded by NIH grant #DC-00594 from the National Institute of Deafness and other Communicative Disorders and NSERC. We thank two anonymous reviewers for their helpful comments and suggestions.

REFERENCES

Bermant, R. I., and Welch, R. B. (1976). The effect of degree of visual-auditory stimulus separation and eye position upon spatial interaction of vision and audition. Perceptual and Motor Skills, 43, 487-493.

Bertelson, P. (1993). The time-course of adaptation to auditory-visual spatial discrepancy. In C. Bundesen and A. Larsen (Eds.), Proceedings of the 6th Conference of the European Society for Cognitive Psychology, pp. 3-4.

Bertelson, P. , and Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Perception and Psychophysics, 29, 578-584.

Bertelson, P. , Vroomen, J. , Wiegeraad, G. , and de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. Proceedings of 1994 International Conference on Spoken Language Processing, 2, 559-562.

Fisher, B. D. , and Pylyshyn, Z. W. (1994). The cognitive architecture of bimodal event perception: a commentary and addendum to Radeau (1994). Current Psychology of Cognition, 13, 92-96.

Gibson, J. J. (1966). The senses considered as perceptual systems. Boston, MA: Houghton Mifflin.

Green, K. P. , and Kuhl, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. Perception and Psychophysics, 45, 34-42.

Green, K. P. , and Kuhl, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. Journal of Experimental Psychology: Human Perception and Performance, 17, 278-288.

Jack, C. E. and Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Perceptual and Motor Skills, 37, 967-979.

Kuhl, P. K., Green, K. P., and Meltzoff, A. (1988). Factors governing the integration of auditory and visual influence in speech: the level effect. Journal of the Acoustical Society of America, 83, Suppl. 1, S86.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Perception and Psychophysics, 50, 93-107.

MacDonald, J. , and McGurk, H. (1978). Visual influences on speech perception processes. Perception and Psychophysics, 24, 253-257.

Manuel, S. Y., Repp, B., Studdert-Kennedy, M., and Liberman, A. (1983). Exploring the "McGurk effect". Journal of the Acoustical Society of America, 74, S66.

Massaro, D. W. (1987). Speech Perception by Ear and Eye. Erlbaum:Hillsdale.

Massaro, D. W. , and Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. Journal of Experimental Psychology: Human Perception and Performance, 9, 753-771.

Massaro, D. W., Cohen, M. M., Gesi, A., Heredi, R. and Tsuzaki, M. (1993). Bimodal speech perception: an examination across languages. Journal of Phonetics, 21, 445-478.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. Journal of the Acoustical Society of America, 100, 1777-1786.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Mendelson, M. J. (1979). Acoustic-optical correspondences and auditory-visual coordination in infancy. Canadian Journal of Psychology, 33, 334-346.

Miller, G. A. , Heise, G. A. and Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials, Journal of Experimental Psychology, 41, 329-335.

Miller, J. L., Connine, C. M., Schermer, T. M., and Kluender, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. Journal of the Acoustical Society of America, 73, 2124-2133.

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996). Temporal constraints on the McGurk effect. Perception and Psychophysics, 58, 351-362.

Radeau, M. (1994). Auditory-visual spatial interaction and modularity. Current Psychology of Cognition, 13, 3-51.

Radeau, M., and Bertelson, P. (1977). Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Perception and Psychophysics, 22, 137-146.

Radeau, M., and Bertelson, P. (1978). Cognitive factors and adaptation to auditory-visual discordance. Perception and Psychophysics, 23, 341-343.

Remez, R. E. and Rubin, P. E. (1994). On the perceptual organization of speech. Psychological Review, 101, 129-156.

Rosenblum, L. D., and Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. Journal of Experimental Psychology: Human Perception and Performance, 22, 318-331.

Samuel, A. G. (1982). Phonetic prototypes. Perception and Psychophysics, 31, 307-314.

Sekiyama, K., and 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.

Sekiyama, K., and Tohkura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. Journal of Phonetics, 21, 427-444.

Sharma, D. (1989). Audio-visual speech integration and perceived location. Ph. D. thesis, University of Reading.

Sumby, W. H. , and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America, **26**, 212-215.

Summerfield, Q. (1979). Use of visual information for phonetic perception. Phonetica, 36, 314-331.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (Eds.), Hearing by eye: The psychology of lip-reading (pp. 3-51). London: Erlbaum.

Summerfield, Q., and McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. Quarterly Journal of Experimental Psychology, 36, 51-74.

Walden, B. E. , Prosek, R. A. Montgomery, A. A., Scherr, C. K., and Jones, C. J. (1977). Effects of training on the visual recognition of consonants. Journal of Speech and Hearing Research, 20, 130-145.

Walker, S., Bruce, V., and O'Malley, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. Perception and Psychophysics, 57, 1124-1133.

Welch, R. B. and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. Psychological Bulletin, 88, 638-667.

ENDNOTES

[1] A between-subject design was used because pilot studies in our lab have shown that the magnitude of the McGurk effect is greatly influenced by subjects' experience with the auditory stimuli in *Auditory Only* cconditions.

[2] The rationale was that /b/ responses would indicate that the visual stimulus did not influence subject's perceptions and non-/b/ responses would indicate that visual influences existed. It is possible that non-/b/ responses could be the result of errors in auditory perception. However, because the auditory stimuli were the same for all conditions, any systematic differences would not be the result of errors in auditory perception. Thus, the number of /b/s reported by subjects is taken as an index of the strength of the McGurk effect; smaller number of /b/s in comparison to the control condition indicating a McGurk effect.