The Role of Auditory Information in Audiovisual Speech Integration

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by

Crystal Huffman

The Ohio State University
June 2007

Project Advisor: Dr. Janet M. Weisenberger, Department of Speech and Hearing Science

Abstract

Communication between two people involves collecting and integrating information from different senses. An example in speech perception is when a listener relies on auditory inputs to hear spoken words and on visual input to read lips, making it easier to communicate in a noisy environment. Listeners are able to make use of visual cues to fill in missing auditory information when the auditory signal has been compromised in some way (e.g., hearing loss or noisy environment). Interestingly, listeners integrate auditory and visual information during the perception of speech, even when one of those senses proves to be more than sufficient.

Grant and Seitz (1998) found a great deal of variability in the performance of listeners on perception tasks of auditory-visual speech. These discoveries have posed a number of questions about why and how multi-sensory integration occurs. Research in "optimal integration" suggests the possibility that listener, talker, or acoustic characteristics may influence auditory-visual integration.

The present study focused on characteristics of the auditory signal that might promote auditory-visual integration, specifically looking at whether removal of information from the signal would produce greater use of the visual input and thus greater integration. CVC syllables from 5 talkers were degraded by selectively removing spectral fine-structure but maintaining temporal envelope characteristics of the waveform. The resulting stimuli were output through 2-.4-, 6-, and 8-channel bandpass filters. Results for 10 normal-hearing listeners showed auditory-visual integration for all conditions, but the amount of integration did not vary across different auditory signal manipulations. In addition, substantial across-talker differences were observed in

auditory intelligibility in the 2-channel condition. Interestingly, the degree of audiovisual integration produced by different talkers was unrelated to auditory intelligibility. Implications of these results for our understanding of the processes underlying auditory-visual integration are discussed.

Acknowledgments

I would like to thank my advisor, Dr. Janet M. Weisenberger, for providing me with the opportunity to work alongside her on this thesis. I was able to grow personally and professionally through her support, guidance, insight and experience. I would like to show my gratitude to Tom Street for being very supportive, patient and encouraging throughout this process. I would also like to thank my subjects for being flexible and devoting their time to help out in the lab. Finally, I am grateful to Natalie Feleppelle for her time and assistance with all aspects of this project.

This project was supported by an ASC Undergraduate Scholarship and an SBS Undergraduate Research Scholarship.

Table of Contents

Chapter 1: Introduction and Literature Review

Speech perception between two people involves the transmission of information through the auditory and visual sensory modalities. Typically speech perception is perceived as a predominantly auditory function, while visual cues serve as an aid to people in audibly compromising situations, such as in noisy environments, or in the presence of hearing impairment. There is a degree of truth to this perception; visual cues can significantly improve speech intelligibility when auditory cues are distorted. However, research by McGurk and MacDonald (1976) found that audiovisual speech integration occurs even in situations where the auditory signal is not impaired in any way.

The McGurk and MacDonald study was conducted by pairing auditory stimuli with conflicting visual stimuli, such as the audio /ba/ with the visual /ga/ (b/g), the audio /pa/ with the visual /ka/ (p/k), the audio /ma/ with the visual /da/ (m/d), and the audio /va/ with the visual /da/ (v/d) (Grant and Seitz, 1998). The purpose of pairing these different phonemes together was to determine how the participants were integrating the different audio and visual inputs and whether the audio input would dominate the perceptual response. Participants were asked to identify speech sounds manipulated under auditory only and auditory plus visual conditions. The results of the study showed that integration of stimuli such as the visual /ga/ and the audio /ba/ produced a fusion response of /da/. This means that the glottal /g/ that was visually represented was fused with the auditory bilabial /b/ sound to form /da/, which is articulated in an area between these two points. Combination responses were also observed from the combining of the visual /ba/ with the auditory /ga/, resulting in /bga/. In this case, the bilabial visual stimulus is very salient and is not fused with the auditory input. These two types of results indicate integration of

the audio and visual stimuli. McGurk and MacDonald concluded that a person cannot ignore inputs, and use all senses available to them when comprehending speech, even if only utilizing one sensory modality proves to be more than sufficient.

This discovery has produced many questions about why and how this integration occurs, one of which is the stimulus conditions that promote optimal integration. Finding out what aspects of the integration process prove to be critical components has become a focal point of current research. One question regards the circumstances that promote optimal integration: is integration facilitated by clear, highly intelligible speech, or by ambiguity in the speech signal? The normal auditory speech signal contains far more information than is necessary to identify the sound, as shown in work by Shannon and his colleagues.

Auditory Cues for Speech Perception

Shannon conducted research concerning auditory speech recognition that focused on degrading selected aspects of the speech waveform in a manner similar to that employed in cochlear implant processors. This waveform consists of spectral and temporal envelopes that provide information relating to place, manner, and voicing of a speech sound. The idea that the speech waveform is highly redundant, containing more information than necessary to identify the presented sound, was an underlying premise in Shannon's study. He started by reducing the spectral information in the speech sound, but preserving the temporal envelope from recorded speech tokens. Fine-structure spectral information was replaced with a band-limited noise, which added ambiguity to the signal (Shannon *et al*., 1995). The results were dramatic: although identification by the subjects

improved as the number of noise bands modulated by the speech temporal envelope increased, high levels of speech recognition performance could be achieved with only three bands of modulated noise. This finding supports the idea that non-distorted auditory speech signals contain a substantial degree of redundant information for identification and even the smallest amount will aid in speech recognition.

Shannon further assessed the effect of temporal envelope cues under reduced spectral conditions for the recognition of consonants, vowels, and sentences (Shannon, Zeng, & Wygonski, 1998). This study consisted of the following four experiments: spacing of the cutoff frequencies, warping the spectral distribution of envelope cues, frequency shifting envelope cues, and spectral smearing. These experiments produced overall results that showed, for four frequency bands, that the frequency alignment of the analysis bands is critical for good performance (Shannon *et al*., 1998). Experiments II and III demonstrated results showing that warping the spectral distribution of envelope cues renders speech completely unintelligible, and that a tonotopic shift of the envelope pattern resulted in poor intelligibility, even when the relative cochlear distribution of envelope cues was preserved (Shannon *et al*., 1998). Experiments I and IV produced results that show the frequency divisions and overlap in carrier bands are not as critical because the exact cutoff frequencies which define the four bands as well as the selectivity of the envelope carrier bands were not critical for speech recognition; performance deteriorated only when the bands were broadly overlapping, smearing the tonotopic specificity of envelope cues (Shannon *et al*., 1998). The study also confirmed that consonant recognition is not affected as much as vowel recognition when spectral cues are distorted (Shannon *et al*., 1998).

Visual Cues for Speech Perception

The studies previously reviewed by McGurk and McDonald and Shannon focused on auditory cues for articulatory features such as place, manner and voicing to produce critical evidence supporting factors related to intelligible speech perception. Research focusing on the visual component of audiovisual integration has also been done in efforts to identify the significant mechanisms and cues provided by the visual input. Unlike auditory signals that contain multiple articulatory features, visual signals only provide information regarding place of articulation. The visible cues that a talker displays may consist of movement of the talker's eyes, mouth, and head and can provide significant information regarding speech perception (Munhall, Kroos, Jozan, & Vatikiotis-Bateson, 2004).

When relying on visual cues alone, a problem arises when sounds have similar visual characteristics and cannot be distinguished from one another, such as the phonemes /p, b, m/. This is a prevalent problem since it has been estimated that around sixty percent of English phonemes are not readily visible, thus making speechreading in visual only conditions extremely difficult (Woodward & Barber, 1960). A large focus of the visual-only aspect of audiovisual integration has dealt with the grouping of similar sounds or phonemes in regards to their visual movement. These similar sounds are known as visemes, a term coined by Fisher (1968) to indicate the distinguishable visual characteristics of speech sounds. In other words, these visemes only allow speechreaders to distinguish between groups of sounds, rather than distinguishing individual sounds within the group (Jackson, 1988).

Examining differences between consonants and vowels in regard to their distinctive visual features has also been a focal point. The number of distinctive visual features of vowels and the place of articulation for consonants is reduced to the shape of the mouth for vowels and the place of articulation for consonants that play a role in the categorization of visemes (Schow & Nerbonne, 2007). In a study on consonant confusion in consonant-vowel syllables in a variety of conditions, it was indicated that in the visual-only conditions the strongest feature for speech perception was place of articiulation, (Binnie, Jackson, Montgomery, 1976). Further evaluation of classification systems provided researchers with evidence that the consonants /p,b,m/, /f,v/, and /θ/, are commonly grouped as visemes, most likely due to visible movements that are universal (Jackson, 1988). The production of vowels can provide visual cues that are beneficial for identifying speech sounds; although every vowel has a distinct shape, they can still be classified into visemes. Despite the fact that visual components are useful for speechreaders, differences among talkers can create confusion among viseme categories for vowels.

Speechreading can be a difficult task in situations that limit a person to using visual cues alone to identify and distinguish words. Nitchie (as cited in Jackson, 1988) provided the term homophenous to describe speech sounds that appeared alike, but noted that visual cues alone could not provide speechreaders with the necessary information to make a distinction. Because groups of consonants are produced at the same points of articulation, the phonemes within these groups cannot be differentiated visually without grammatical, phonetic, or lexical information, therefore labeling these visually confusable units as speech homophenes (Schow & Nerbonne, 2007). Homophenous

words look alike when spoken despite sounding different and having unrelated spellings, such as: pet, bed, and men; tip, limb, and dip; and cough and golf (Schow & Nerbonne, 2007).

Auditory-Visual Integration Theories

Researchers have developed several models in efforts to describe the process of integration across modalities for optimal speech perception of auditory-visual stimuli. The Fuzzy Logical Model of Perception (FLMP) is a theory used for explaining auditory-visual integration. Massaro (1998) suggests that incoming auditory, visual and auditory-visual information is independently evaluated by listeners to extract summary descriptions of the incoming sensory information. These summary descriptions are then compared to known descriptions in the memory to determine the degree to which the cues from a given source match learned information. The summary descriptions are integrated together using guidelines of memory descriptions and perceptual alternatives are formed. Perceptual decisions are made based on the degree of support for each perceptual alternative. According to Massaro (1987) the multiplicative integration rule used to determine auditory-visual speech perception performance in the FLMP is an optimal decision rule and is applied to minimize the differences between obtained and predicted scores, and therefore may be considered more of a fit to obtained bimodal scores rather than a prediction of optimal bimodal speech performance. According to Grant (2002) two consistent aspects are demonstrated by the FLMP: first, it seeks to apply the multiplicative integration to unimodal confusion data (i.e., probabilities of responding *y* given *x*) to obtain a bimodal prediction, and second, human receivers often do better at

recognizing consonants than the FLMP predicts. With the FLMP, the assumption is made that the model predicts optimal integration (Massaro, 1987; Massaro and Cohen, 2000). In other words, poor multimodal performance results from poor unimodal inputs, rather than from poor integration abilities.

In contrast, the prelabeling (PRE) model of integration seeks not to optimally fit observed auditory-visual data, but rather seeks to "label" incoming bimodal stimuli based on an optimal combination of mutual information collected from separate fits to auditory-alone and visual-alone performance (Braida, 1991). The PRE model first derives an estimate of unimodal information and then predicts how an unbiased receiver with no interference across modalities might do given the particular unimodal information available (Grant, 2002). This model allows for the possibility that integration ability may be suboptimal. When applied to the Grant and Seitz (1998) study, the PRE model seemed better suited to estimate integration efficiency in accounting for individual differences seen in the speech perception of hearing-impaired listeners. However, it is important to keep in mind that auditory-visual integration efficiency is a presumed skill employed by subjects independently from their ability to extract information from auditory and visual speech inputs and that the validity of these derived estimates of integration efficiency cannot be based solely on the accuracy of model fits (Grant, 2002).

The term "auditory-visual integration" is used to denote the processes employed by individual receivers to *combine* the information extracted from auditory and visual sources (Grant, 2002). This process is distinctly different from the ability to *extract* auditory and visual cues and higher-order language processing of the information received by the two senses (Massaro, 1998). Auditory-visual speech recognition by

human receivers has been repeatedly shown to be incredibly robust and greatly resistant to environmental and biological sources of signal distortion; however, some individuals still demonstrate significant problems understanding speech in these communication settings (Grant, 2002). Overall, listeners perform better in difficult listening situations with the addition of visual cues, however, there is a great deal of variability in auditory-visual speech perception performance.

These models were used by Grant and Seitz (1998) to determine whether individual receivers integrate auditory and visual cues with varying degrees of efficiency. This study offered possible explanations for differences observed across subjects on auditory-visual speech recognition tests when more obvious factors such as hearing loss, visual acuity, vocabulary, and language competence were accounted for (Grant and Seitz, 1998). Grant and Seitz (1998) found that by establishing the validity of integration efficiency as an independent process and examining potential differences in integration processing as a function of speech-processing demands, individual hearing-impaired subjects differ with respect to integration efficiency on a variety of measures. Differences in individual integration efficiency explains a substantial portion of the variance observed across individuals in auditory-visual speech recognition.

The Role of Auditory Information in Audiovisual Speech Integration

What exactly do we mean when we say that individuals differ in integration efficiency? How does the process of auditory-visual speech integration occur? What other factors might play a role in auditory-visual speech integration? One approach to understanding the differences in auditory-visual integration abilities is to examine

responses to specific stimuli to determine whether integration ability and the characteristics of integration change for different types of inputs and situations. The present study examined how altering characteristics of the auditory signal impacted auditory-visual integration. By isolating and removing progressively greater amounts of information from the auditory signal, we can study the features that are extracted from the auditory signal during multimodal speech perception. Varying the degree to which information is removed provides information regarding how integration processes change as a result of the available information found in the auditory signal.

Auditory stimuli were degraded using a method similar to Shannon *et al*. (1995); auditory syllables were reduced to a waveform composed of a broadband noise fine structure that is modulated by the temporal envelope of the original speech stimulus recording. Each degraded speech stimulus was then filtered into two, four, six, or eight spectral bands. This method of degrading removes the fine structure and discrete frequency information found in the speech signal, effectively reducing the redundancy of the auditory stimulus to varying degrees based on the number of spectral channels. Auditory and visual cues provide temporal and spectral information in the speech stimulus; however, the auditory signal is highly redundant and the visual signal is rather ambiguous. The present study explored whether acoustic redundancy or ambiguity better facilitates optimal auditory-visual integration.

Chapter 2: Method

Participants

Participants included five talkers and ten observers. The talkers consisted of three female and two male participants with ages ranging from 20 to 23, who produced a set of eight single syllable stimuli that were recorded by a video camera. All the talkers were undergraduate/graduate university students and reported having normal hearing and normal or corrected vision. The observers consisted of eight female and two male participants with ages ranging from 17 to 22. Three of the ten observers were undergraduate university students in the Speech and Hearing Sciences major. All ten observers self-reported to have normal or corrected vision and underwent audiometric testing to verify normal hearing. None of the participants reported knowing about the McGurk effect. Eight of the ten observers received eighty dollars in payment for their time, while the other two observers received academic credit.

Interfaces for Stimulus Presentation

Visual stimuli were presented on a 20 inch video monitor, while the auditory stimuli were presented via TDH 39-Audiologic headphones.

Stimuli Selection

A limited set of eight CVC syllables were used as the stimuli for this study, chosen for their ability to satisfy the following conditions:

1. Pairs of the stimuli were minimal pairs, differing only by the initial consonant
2. All stimuli were accompanied by the vowel /ae/, which does not exhibit lip

rounding or lip extension.

3. Multiple stimuli were used in each category of articulation, consisting of: place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).
4. All stimuli were presented without a carrier phrase (citation-style)
5. Stimuli were known to elicit McGurk-like responses

Stimuli:

For each of the conditions the same stimuli were administered which consisted of the following signle-syllable stimuli and dual-syllable stimuli.

Signal-syllable stimuli

Bilabial: mat, bat, pat

Alveolar: sat, zat, tat

Velar: gat, cat

Dual-syllable stimuli

1. bat-gat
2. gat-bat
3. cat-tat
4. tat-cat

Stimulus Presentation

Audio Signal Degrading:

The software program Video Explosion Deluxe was used to record each of the five talkers producing a set of eight monosyllabic stimuli, words five times each. They were recorded through a microphone directly into a computer, which permitted the files

to be stored in .wav format. These auditory files were then input to a subroutine created by Bertrand Delgutte in MATLAB 5.3. The subroutine (chimeras) begins with two stimuli, the input speech waveform and a broadband noise. The program then exchanges the amplitude envelope waveform and fine structure of the two stimuli, keeping the waveform containing noise fine structure and speech envelope characteristics, and discarding the other waveform. Each speech signal was then filtered into four broad spectral bands where the bandwidths of the four channels are chosen to provide equal spacing in basilar membrane distance. The upper cutoff frequencies for the four spectral bands were: 504 HZ, 1,794 Hz, 5,716 Hz, and 17,640 Hz. Auditory syllables are thus reduced to a waveform composed of a broadband noise fine structure that is modulated by the temporal envelope of the original speech stimulus recording; this is similar to those created by Shannon et al. (1998), as described in the Introduction.

Digital Video Editing:

Visual stimuli for the study were created by recording two male and three female talkers with a digital video camera while repeating the list of eight stimulus words a total of five times each. The auditory and visual stimuli were then downloaded to the program Video Explosion Deluxe, where editing of the clips took place. This program allows for any auditory clips, which were degraded into "chimeras" earlier in the MATLAB 5.3 program, to be dubbed onto any visual clip. This made it possible to create stimuli that featured different auditory and visual dual-syllable (incongruent) stimuli, creating the possibility for a McGurk-type integration effect. This combination process of the auditory and visual clips also made it possible to create auditory-visual stimuli that featured both degraded and normal auditory and visual components. This present study

paired visual stimuli produced by a talker to auditory stimuli produced by the same talker. Randomized lists were made for each talker in the four different conditions with the purpose of reducing the possibility of effects that can occur from order of stimulus presentation. From these lists a compilation of videos featuring sixty stimulus clips were created with the use of the program Video Explosion Deluxe. Sonic MY DVD was the software program used to convert individual videos created on Video Explosion Deluxe to be converted and made into DVDs. Each talker had three randomized lists for each of the four conditions, resulting in the production of a total of sixty DVDs.

Procedure

Testing Setup:

This present study tested all observers in the basement lab room of Pressey Hall, part of The Ohio State University's Speech and Hearing Department. This room provided an environment conducive to testing: quiet environment, well-lit with its fluorescent lighting, sound-attenuating booths, and the necessary digital equipment. The sound-attenuating booths contained a chair positioned along the back wall such that the observer faced a double-glass window. When seated in the chair the observer was approximately 4 feet from the 50 cm video monitor, located on the outside of the double-glass window of the booth. TDH 39-Audiolgoic Headphones were used to transmit the auditory stimuli. An intercom system was placed inside the booth to allow observer responses to be transmitted to the examiner outside of the booth.

Testing Presentation:

Before each observer was initially tested they were given a set of instructions to read over as well as a verbal explanation of the instructions by the examiner. The instructions explained that the observers would be tested under three randomized conditions: degraded auditory-alone, where the observer would just be listening to the headphones with no visual aid, visual-alone, where the observer would be watching the television screen with no auditory aid, and degraded auditory plus visual, where the observer would watch the television screen while listening to the auditory stimuli. During each of these conditions the observers were instructed that 60 stimulus words would be presented and a response needed to be given after each one. (The auditory-only and visual-only conditions consisted of 60 single-syllable stimuli, while the auditory-visual condition used 30 single-syllable stimuli to collect percent correct responses while 30 dual-syllable stimuli were used to elicit McGurk type responses). It was also explained that these 60 stimuli were phonemes that all ended in "at." However, any initial consonant or cluster of consonants could be provided as a response (open response set). This means that there could be any consonant or combinations of consonants to form these combinations that may or may not exist in the English language. Three conditions (auditory-only, visual-only, and auditory-visual) were tested for each condition: 2-channel, 4-channel, 6-channel, and 8-channel.

Testing Procedure:

Each observer was tested with all three presentation conditions. Sixty trials were presented via prerecorded DVDs for each of the five talkers in the four different channel conditions. The presentation order of each condition was randomly varied across

participants. The examiner recorded all observer responses for every trial and condition. Each observer was tested for approximately ten hours over multiple sessions that lasted two hours or less. Rest periods were encouraged to minimize fatigue.

Chapter 3: Results and Discussion

Results for two types of stimuli were analyzed. First, performance was evaluated for single-syllable (congruent) presentations, in which all modalities tested (degraded auditory only, visual only, degraded auditory plus visual) received the same stimulus and the percent correct performance was measured. Integration can be assessed by determining the degree to which degraded auditory plus visual performance was better than performance in the degraded auditory only or visual only conditions.

Second, performance was assessed for dual syllable (incongruent) presentation, where the auditory stimuli differ from the visual stimuli. There is no "correct" response for incongruent phonemes, but the responses are categorized into one of three groupings: auditory, where the response is identical to the auditory stimulus used in the incongruent pairing; visual, where the response is identical to the visual stimulus used in the incongruent pairing; or other, where the response matches neither the auditory nor visual stimulus used in the incongruent pairing.

Percent Correct Performance

Figure 1 shows the percent correct identification for all three testing conditions (auditory only, visual only, auditory plus visual) in each of the four conditions (2-channel, 4-channel, 6-channel, 8-channel) by displaying the averaged results across all subjects and all talkers. There are several things worth noting from this figure. The consistency across the graph for visual only performance makes sense, because it is not affected by the presence or absence of any particular auditory stimuli. Also, the figure shows that auditory-only performance systematically increases as the number of channels

is increased. This implies that individuals take advantage of an increase in the number of output channels and use the additional information in the stimulus. The addition of information improved the subjects' ability to recognize the presented stimuli by 12% from 2-channel to 4-channel, 6% from 4-channel to 6-channel, and 8% from 6-channel to 8-channel.

It is also useful to determine the percentage improvement provided by auditory plus visual input over auditory alone, which is an indication of auditory-visual integration. Comparing the percentage results for auditory-only in a 2-channel condition to auditory plus visual in a 2-channel condition shows an improvement of 15%. In the 4-channel condition there is a 13% improvement, and in the 6-channel condition there is a 16% improvement; these percentages suggest that a similar amount of integration is ocurring in each condition. These results tell us that in the first three conditions adding the visual stimulus provides some new information over what is available in the auditory-only conditions. A smaller percentage of improvement in the 8-channel conditions suggests that the additional auditory information may be more redundant with the visual stimulus.

Figure 2 shows the percent correct for visual only presentation for each of the five talkers used in the study. The data show only small variability in the percent correct averages across talkers. However, Figure 3 shows a different pattern in the percent correct in auditory-only 2-channel conditions by talker. These data include substantial differences across talkers, with talker LG considerably more intelligible, while talker KS is considerably less intelligible. Figure 4 represents the percent correct in auditory-only 4-channel conditions by talker; again, LG is most intelligible, but far less variability is

shown across talkers as compared to the 2-channel results. This is consistent with the data from Figure 5 for percent correct for the auditory-only 6-channel condition. Figure 6 shows percent correct for auditory-only in the 8-channel condition. Here, all talkers yield similar performance and show high levels of intelligibility.

McGurk Type Integration

Figure 7 shows response patterns for incongruent inputs, by the number of channels in the auditory condition. The results show the lowest percentage consistently across all channels for auditory responses, while the visual and other responses were similar to each other across all channels. The lack of difference in the percentage of auditory responses across channels is quite surprising given the high intelligibility of these talkers in the 6-channel and 8-channel conditions (see Figures 5 & 6). Figure 8 analyzes the "other" responses in Figure 7 to assess potential integration for incongruent inputs. Combination McGurk type responses (where the subject combines the first consonants of each stimulus presented and produces a response) had the lowest level of responses. This is not unexpected when compared to previous research, where the rate of combination responses is relatively low due to the fact that these types of consonant clusters are foreign to Standard American English. Fusion responses also were at a relatively low level for all auditory stimuli. This finding was unexpected in comparison to past studies in the laboratory, in which fusion integration percentages were near 50%-60%. In the present study, fusion percentages were 19% for the 2-channel, 20% for the 4-channel, 29% for the 6-channel, and 26% for the 8-channel; this substantial loss of fusion responses across all channels of degraded stimuli is puzzling. The fact that we see

less fusion integration suggests that removing any information from the auditory stimuli can be harmful. However, another concern in the present study was a surprisingly large percentage response of "hat" by all subjects, across all stimuli. /h/ was not classified as a fusion response due to the location of production (glottal), which is not between the bilabial /b/ and the velar /g/.

Figure 9 further investigates the low levels of fusion responses by analyzing responses to different talkers. This Figure shows a fairly substantial difference in the 2-channel auditory plus visual condition between talker LG, who produces the greatest percentage of fusion responses, and talker KS, whose percentage is quite low. Figure 10 shows fusion responses across all channels for individual listeners, revealing differences in response patterns across listeners; listeners TM and KC have higher levels of integration responses compared to listeners AC and AV, who show low levels of fusion responses.

Confusion Matrices

Consideration of all the results from the present study raises several questions. According to Figures 3 and 9, better talkers, like LG, produce more fusion responses. Clearly, there are noticeable differences across talkers. Two confusion matrices were created to take a closer look at percent correct listener responses for talkers LG and KS in the 2-channel condition. Results from these matrices showed that for talker LG, pat, mat, gat, and cat were all very intelligible, showing that these phonemes all provided good place information, while mat had 100% intelligibility, providing a nasality cue. Performance for LG's productions of sat and tat were low; listeners often confused these

phonemes for fat and cat, showing the loss of some high frequency information. KS' results showed that voicing for several phonemes was lost; the stimulus bat was mistaken for pat, pat for that, sat for bat, and sat for that. Also, for KS there was a loss of manner information, as indicated by the stimulus bat being mistaken for mat, while zat was mistaken for bat or gat. Although these confusion matrices show large differences in performance between the talkers LG and KS in the auditory-only 2-channel condition these talkers display relatively similar percent correct responses in the 6-channel and 8-channel condition (see Figure 5 & 6). This information is very surprising and suggests that further analysis of the auditory signal in each of the four conditions should be performed.

Statistical Analysis

Statistical analysis (ANOVA) revealed several significant findings. A two-factor, within subject, ANOVA was performed to determine the significance of differences across channels and presentation conditions. First, there was a significant main effect of number of channels, $F(3, 147) = 49.85$, $p<.0001$, $\eta^2 = .50$. Follow-up Pairwise Comparisons indicated significant differences between all pairs of channels, except the 6-channel to 8-channel comparisons. Second, there was a significant main effect of presentation condition, $F(2, 98) = 235.58$, $p<.0001$, $\eta^2 = .83$. Follow-up Pairwise Comparisons indicated significant differences between all presentation conditions. Third, there was a significant channel by presentation interaction; $F(6, 294) = 25.03$, $P< .0001$, $\eta^2 = .34$.

A second analysis was preformed to investigate differences across the five talkers in the two-channel condition. There was a significant main effect of talkers, $F(4, 36) = 17.52$, $p < .0001$, $\eta^2 = .66$. Follow-up Pairwise Comparisons indicated that talkers LG and JK performed differently from most of the other talkers. In this analysis, there was also a significant main effect of presentation condition, $F(2, 18) = 41.54$, $p < .0001$, $\eta^2 = .82$. Lastly, there was a significant talker by presentation condition interaction, $F(8, 72) = 7.76$, $p < .0001$, $\eta^2 = .46$.

Chapter 4: Summary and Conclusion

Results of this study indicate that listeners perform increasingly better with auditory stimuli when more spectral information is available (less ambiguity is present), to a certain point. However, systematically removing information from the auditory stimulus does not necessarily affect the degree of integration benefit.

In addition, talker differences were an important factor in regard to percent correct responses from listeners. Differences across talkers in the degree of benefit provided in the audiovisual condition were examined by the construction of confusion matrices for the best talker, LG, and the worst talker, KS. These matrices showed that for KS there was a substantial loss of both manner and voicing cues. This finding argues for more in-depth analysis of specific productions by these talkers. Differences across listerns, as shown in Figure 10, suggest that individual listener characteristics are also an important component of the integration process. This finding supports earlier work by Grant and Seitz (1998).

Results from this study suggest that systematically removing information from the auditory stimulus does not necessarily affect the degree of integration benefit. Overall, further study is required to determine how the degree of benefit varies across talkers and auditory manipulations. Additional analyses will evaluate the acoustic characteristics of specific syllables produced by these talkers to search for specific features that might better explain across-talker performance and possibly address the very low levels of McGurk-type integration observed here.

References

Binnie, C.A., Jackson, P., & Montgomery, A. (1976). Visual intelligibility of consonants: A lipreading screening test with implications for aural rehabilitation. *Journal of Speech and Hearing Disorders*, 41, 530-539.

Braida, L.D. (1991). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. 43A (3), 647-677.

Fisher, C.G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 12, 796-804.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America*, 112 (1), 30-33.

Grant, K.W. & Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, 04, (4), 2438-2449.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*, 90 (5), 99-114.

Massaro, D.M. (1987). Speech perception by ear and eye: A paradigm for psychology

inquiry. Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D. (1998) *Illusions and issues in bimodal speech perception*. in Auditory-Visual Speech Processing Conference. Terrigal, Sydney, Australia. p. 21-26

Massaro, D.W., and Cohen, M.M. (2000). Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception, J. Acoust. Soc. Am., 108, 784-789.

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

Munhall, K.G., Kroos, C., Jozan, C., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perceptions & Psychophysics*, 66 (4), 574-583.

Schow, R.L., & Nerbonne, M.A. (2007). Introduction to audiologic rehabilitation [rev. ed.] Boston, MA: Pearson Education, Inc.

Shannon, R.V, Zeng, F.G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

Shannon, R.V., Zeng, F.G., Wygonski, J. (1998). Speech recognition with altered

spectral distribution of envelope cues. *The Journal of the Acoustical Society of America*, 104 (4), 2467-2475.

Woodward, M.F., & Barber, C.G. (1960). Phoneme perception in lipreading. *Journal of Speech and Hearing Research*, 3, 212-222.

Tables

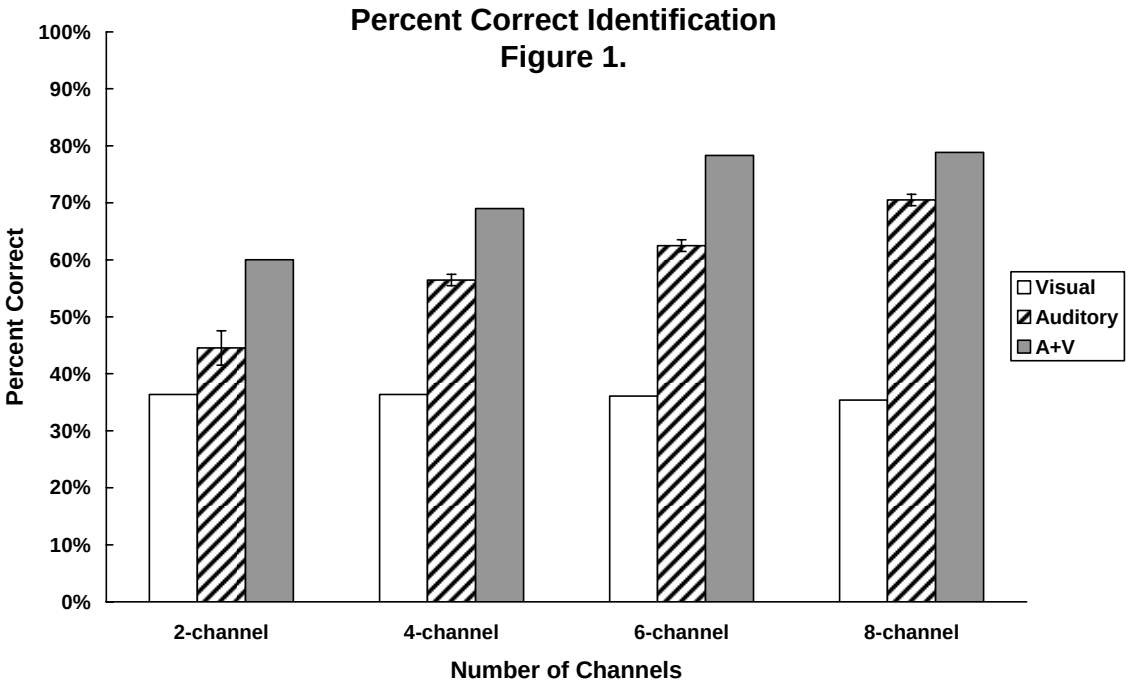
Table 1. Confusion matrix of talker LG in the 2-channel auditory-only condition

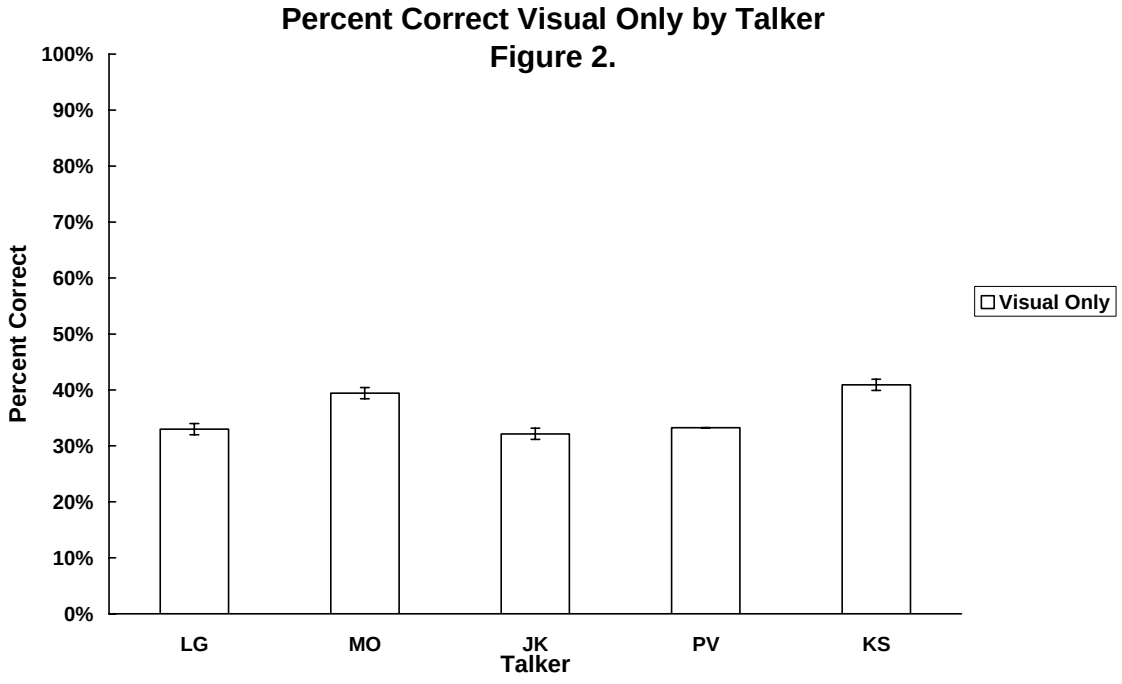
Stimulus \ Response	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	FAT	VAT	THAT
BAT	.66		.28	.07							
PAT	.03	.88			.09						
MAT			1.00								
GAT				.97	.03						
CAT		.19			.81						
ZAT	.42		.17	.08						.08	.25
TAT		.17			.33		.50				
SAT	.17					.17		.08	.58		

Table 2. Confusion matrix of talker KS in the 2-channel auditory-only condition

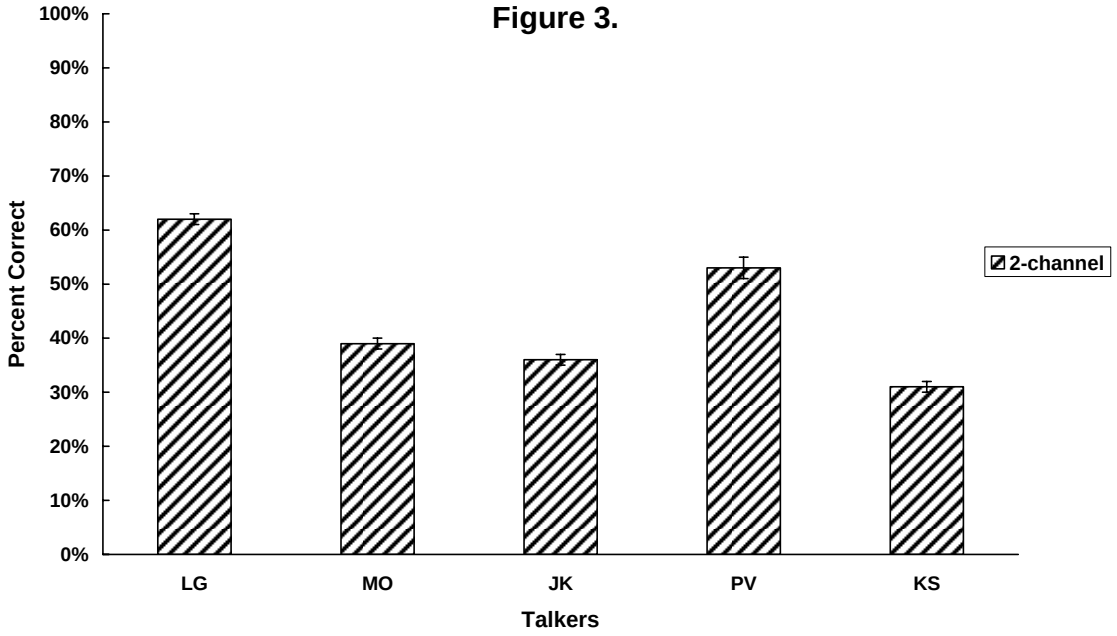
Stimulus \ Response	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	FAT	THAT
BAT	.37	.26	.16							.21
PAT	.04	.42	.21	.04	.04			.04		.21
MAT	.12		.88							
GAT	.39	.09	.39	.09		.04				
CAT	.04	.27		.45	.09				.09	.04
ZAT	.38			.25		.12		.12		.12
TAT	.12		.12	.12	.25		.12	.12		.12
SAT	.63		.12							.25

List of Figures

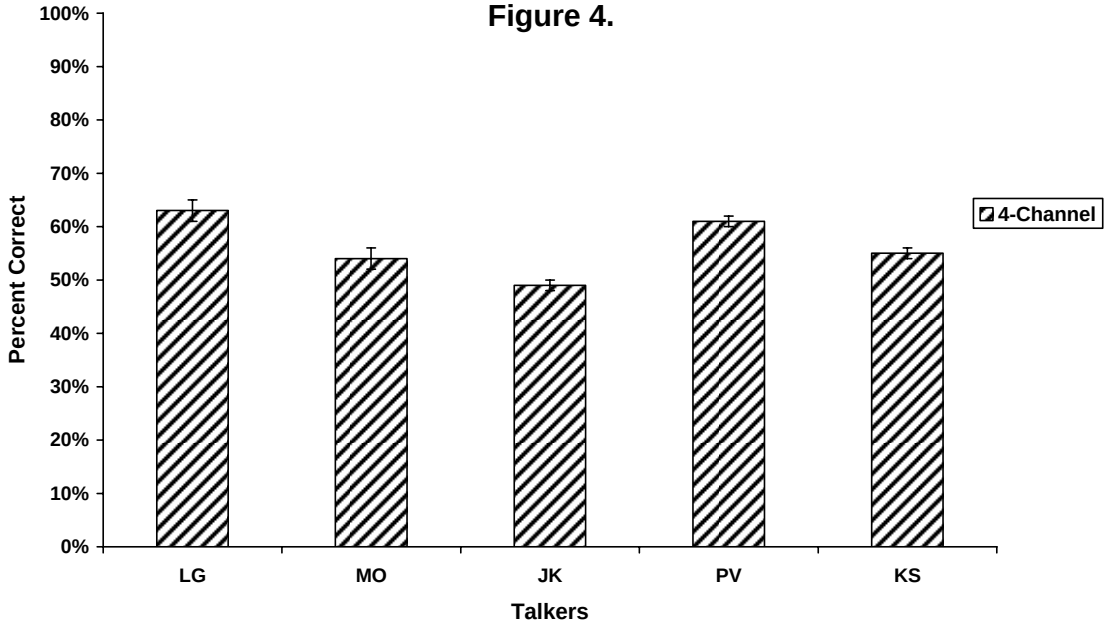

Percent Correct Identification
Figure 1.
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent Correct
2-channel
4-channel
6-channel
8-channel
Number of Channels
Visual
Auditory
A+V


Percent Correct Visual Only by Talker
Figure 2.
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Percent Correct
LG
MO
JK
PV
KS
Talker
Visual Only

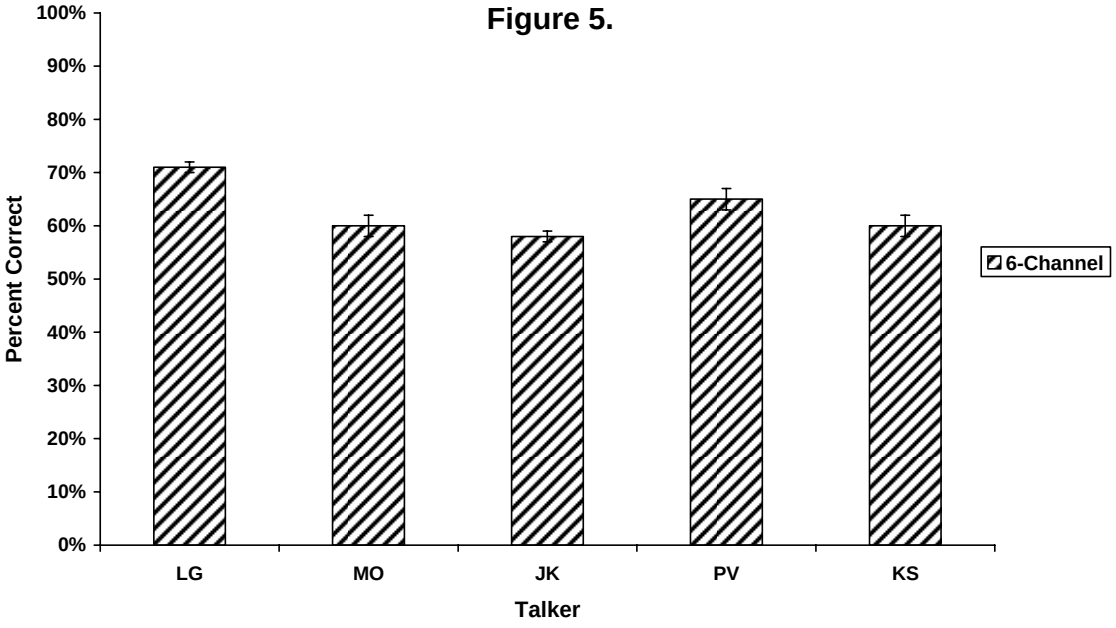
Percent Correct Auditory 2-Channel by Talker
Figure 3.



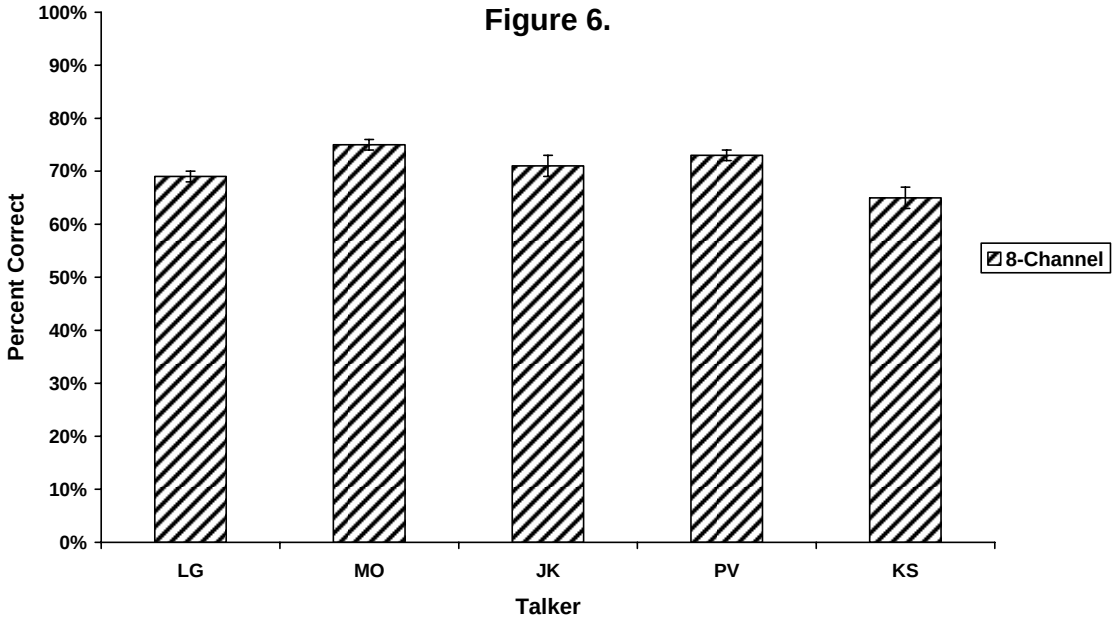
Percent Correct Auditory 4-Channel by Talker
Figure 4.



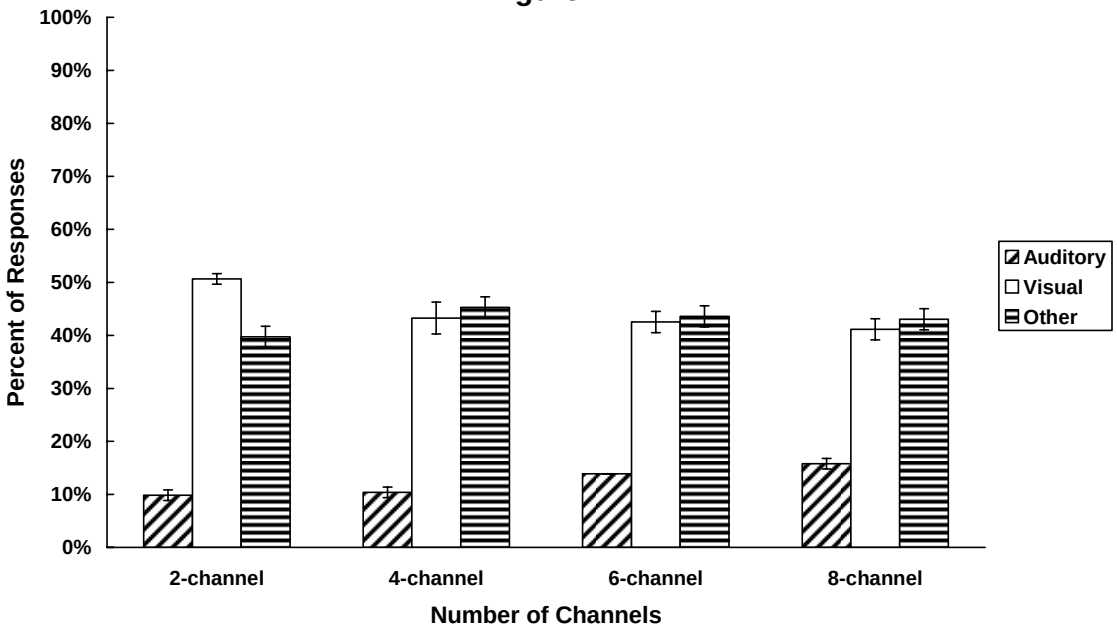
Percent Correct Auditory 6-Channel by Talker
Figure 5.



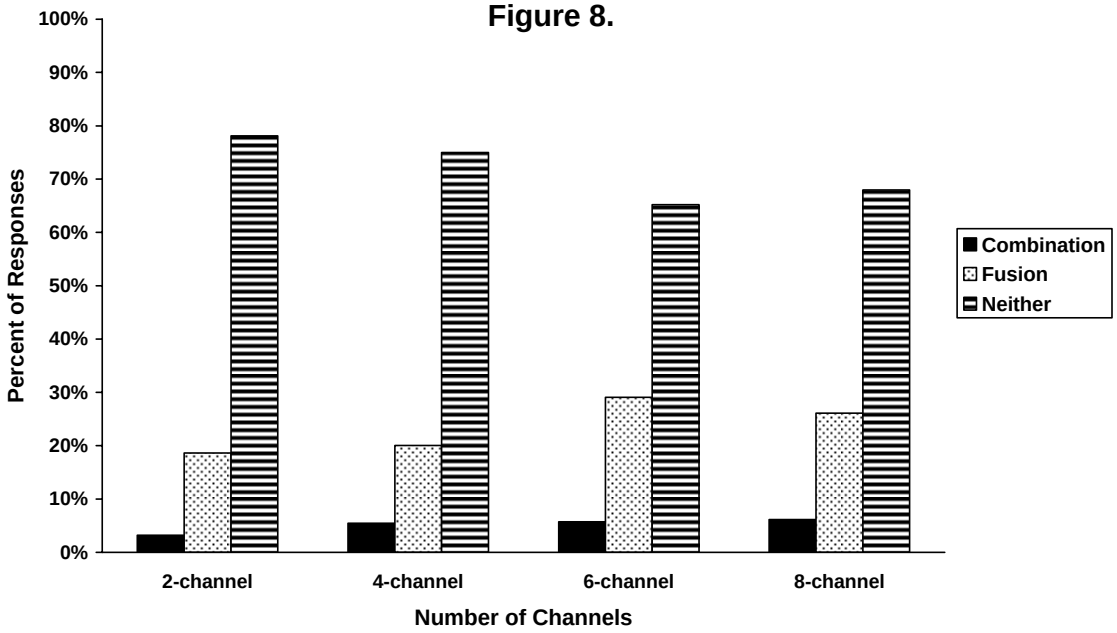
Percent Correct Auditory 8-Channel by Talker
Figure 6.



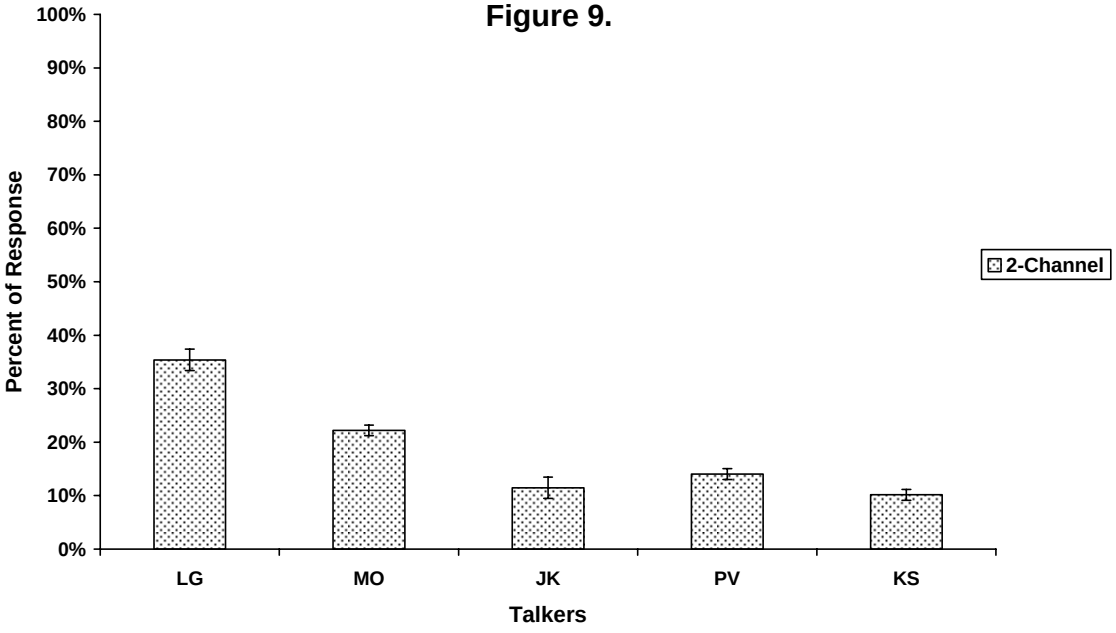
Percent of Auditory+Visual Integration Responses for Incongruent Inputs
Figure 7.



Auditory+Visual Integration for Incongruent Inputs
Figure 8.



Percent of Fusion Responses 2-Channel by Talker
Figure 9.



Fusion Responses Across All Channels by Subject
Figure 10.