Effects of training on intelligibility and integration of sine-wave speech

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

Megan Gariety

The Ohio State University
June 2009

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

Abstract

Although the auditory signal usually provides sufficient information for speech perception, visual cues become important when the auditory signal is compromised, as in the case of a hearing loss. However, research has shown that visual cues are used even when the auditory signal is completely intelligible (McGurk & MacDonald, 1976).

Subsequent studies have investigated the impact of reducing the quality of auditory information on the integration process. Grant and Seitz (1998) studied audiovisual integration by hearing-impaired subjects, and reported that even when the auditory input is poor, speech perception can improve with the aid of visual cues.

Integration for artificially reduced auditory inputs has also been investigated. One form of reduction was that used by Remez et al. (1981), who reduced speech signals to three time-varying sinusoids following the formants of the speech ("sine-wave speech"). Remez et al. showed that sine-wave speech can be highly intelligible in sentences, but in studies of audio-visual perception of individual syllables, sine-wave reductions have yielded poor performance (e.g., Anderson, 2007). It is possible that the relative unfamiliarity of this form of speech led to the poor results in integration studies.

This question was addressed by Exner (2008), who evaluated the effects of increased exposure to sine-wave syllables on auditory only and audio-visual perception. She found that even two hours of training produced significant improvement in performance. However, it was not clear in her study was whether performance had reached asymptotic levels.

The present study investigated the effects of longer training periods to determine if further improvements to sine-wave performance can be achieved. Five listeners received ten hours of auditory only training with eight syllables spoken by three talkers. Results showed significant improvement across training sessions, but the amount of audio-visual integration did not change. This supports the argument of Grant & Seitz (1998) that integration is a process independent of auditory or visual processing.

Acknowledgements

I would like to thank my advisor, Dr. Janet M. Weisenberger, for providing me with this wonderful opportunity to work with her on this thesis. Through her dedication, experience, and support I was able to grow academically and professionally. I would like to thank Michelle Hungerford for her time, patience, and assistance she has given me throughout this process. Furthermore, I would like to thank my subjects for being flexible and devoting their time while assisting me on my thesis.

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

Table of Contents

CHAPTER 1: Introduction and Literature Review

It is commonly thought that the auditory input is the dominant modality for understanding speech, when in fact both auditory and visual inputs are used simultaneously. Although the auditory signal usually provides enough information to understand speech, visual cues become important when the auditory signal is being degraded in some way (e.g., listening in background noise, reverberation, or with a hearing loss.) People rely on visual cues in these situations to assist them in understanding speech. Even with normal hearing and a perfect auditory signal, listeners still use visual and auditory inputs at the same time to improve their speech perception. This can be demonstrated in what is known as the McGurk Effect.

It is known that the visual stimulus can actually change the perception of an auditory sound. McGurk and McDonald (1976) studied audio-visual integration, in an experiment which the presentation of an auditory syllable such as a bilabial /ba/ shown with a visual, velar consonant /ga/ resulted in the perception of the alveolar consonant sound /da,/ a fusion of /ba./ and /ga./ When reversing this process, a visual /ba/ paired with an auditory /ga/ a combination response, /baga/ was reported. This kind of response is often reported because the strong visual stimulus /ba/ cannot be ignored. The McGurk and MacDonald study indicates that speech is not a purely auditory percept; rather, it is influenced by the visual input, even when the auditory signal is perfect. Today, it is well documented that audio-visual integration occurs automatically, without the listener’s conscious awareness. Evaluating both the auditory and visual speech signals people receive is important for a full understanding of audio-visual integration.

Auditory Cues for Speech Perception

Normally, auditory cues alone provide enough information for the auditory speech signal to be intelligible. There are three main ways to characterize a consonant auditory signal. These are place of articulation, manner of articulation, and voicing. The place of articulation is the point of contact in the mouth where the sound is produced (location). These include bilabials, labiodentals, interdentals, alveolars, palatal-alveolars, palatals and velars. The manner of articulation describes how the speech organs are involved in making a sound (how the sound is produced.) These include stops, fricatives, affricates, liquids and glides. Voicing refers to the presence or absence of vocal folds vibrations. When the vocal folds are vibrating the sound is voiced and when not vibrating the sound is voiceless. This information can be found in the spectral and temporal envelopes of a speech waveform (Ladefoged 2006).

A number of studies have investigated specific ways of degrading the auditory signal to determine how reduced auditory information impacts the integration process. One method of degrading the auditory stimulus that has been studied involves reducing the speech sound to a series of sine-waves that follow the formant structure of that speech. Remez et al. (1981) created “a three-tone sinusoid replica of a naturally produced utterance” commonly referred to as sine-wave speech. The sinusoid components followed the formant structure of the original utterance. The study employed three different conditions. Listeners in the first group, condition A, were told nothing about the nature of the sound. Listeners in the second group, condition B, were informed they would hear a sentence produced by a computer. In the third group, condition C, listeners were told exactly what they would hear. All groups were asked to report what they heard to the best of their ability. Although listeners in the third group reported that almost all the words presented had an unnatural voice quality, they were found to be intelligible. The results of this study showed that speech is still intelligible even when speech is drastically reduced to just three time-varying sinusoids.

Other studies, one in particular by Shannon et al. (1995), focused on degrading selected aspects of the speech waveform in a manner similar to the processing used in cochlear implants. Shannon et al. showed high intelligibility of reduced-information signals, suggesting that the normal speech signal is highly redundant in that it contains more information than is required for identification. In Shannon's study, the fine-structure information from speech syllables was replaced with band-limited noise, but the temporal envelope of the speech signal was retained. Results indicated that speech recognition was possible for these signals even with only three or four filter bands, indicating that minimal temporal cues alone can provide effective information for the identification of speech sounds.

Visual Cues for Speech Perception

Although visual cues are very important for identifying speech sounds, the visual signal provides less information than does the normal auditory signal. One can identify primarily the place of articulation by using visual input alone, and that itself can be ambiguous (Jackson, 1988). There is some information on manner of articulation, but absolutely no voicing information in the visual signal. Therefore one cannot always identify a sound correctly from visual information alone.

For example: /p, b, m/ constitute a viseme group. Visemes are visual phonemes that have more then one speech sound, but are all produced with similar facial and articulatory movements

(Jackson, 1988). The phonemes /p, b, m/ all have the same place of articulation, but differ in the manner of articulation and voicing. This can make it very difficult to identify speech sounds when there is no auditory signal. The difficulty level for speechreaders varies depending on the characteristics of different talkers. Jackson found that the talkers who created more viseme categories were more intelligible than the talkers who created fewer. Other visual cues also contributed to the difficulty level of the talker such as eye movements, facial hair, gestures and head movements. All of these cues transmit or interfere with information from the talker to the listener.

Audio-Visual Integration Theories

Researchers have introduced several models to describe the process of integration between the two modalities for optimal speech perception of auditory-visual stimuli. Two of these that have received considerable attention are The Pre-Labeling Model of Integration (PRE) and the Fuzzy Logical Model of Perception (FLMP). The Pre-Labeling Model of Integration suggests that the prediction of auditory–visual (AV) recognition involves a combination of information gained from auditory-only and visual-only performance (Braida, 1991). The PRE model suggests that an optimum processor is used to take information from both unimodal conditions, and preserve this information in the multi-modal case, with no biasing from other modalities. Thus, the PRE model predicts that AV speech recognition should be equal to or better than the auditory or visual recognition alone obtained from observers. According to Grant and Seitz (1998), the PRE model seemed better to estimate integration efficiency. They measured how far a listener's actual performance deviated from the predicted audio-visual recognition score. The closer an individual was to the predicted score, the better the individual was integrating. Because an individual's integration efficiency could not be predicted by knowing his or her auditory-alone or visual-alone performance, Grant argued that auditory – visual integration is independent of a person's ability to extract auditory and visual information from speech. Further integration occurs early, prior to actual phoneme identification (Grant, 2002).

In contrast, the Fuzzy Logical Model of Perception (FLMP) suggests that integration occurs very late, after identification of the auditory and visual inputs (Grant, 2002). The reasoning behind this comes from Massaro (1998). The preferred method for applying the FLMP is to fit the model to the data from auditory, visual, and auditory-visual conditions. These are evaluated independently. Massaro argues that information from the auditory and visual inputs is then compared to prototypes in memory to determine the number of response alternatives and the support for each response alternative. Finally, the identification responses are determined from the response alternatives. The FLMP then uses data from stimulus/response confusion matrices (similar to the PRE model) to determine the probability that a particular stimulus presentation will elicit a particular response alternative (Massaro, 2000). This model employs confusion data from auditory, visual, and auditory-visual conditions, while the PRE model only considers auditory and visual matrices.

Current Audio-Visual Perception Studies

Several recent studies in our laboratory have investigated audio-visual perception of degraded speech stimuli. Andrews (2007) and Huffman (2007) both degraded audio stimuli in a manner similar to that of Shannon et al. (1995), and found good levels of identification. However, Tamosuinas (2007) used sine-wave speech (containing the first three formants, F1+ F2

+ F3), similar to the work of Remez et al. (1981) and found poor performance. One possible explanation for this difference was that Tamosuinas's participants had no prior exposure to sine-wave speech. Previous studies in our lab have shown that increasing exposure to unfamiliar stimuli can improve performance. In 2008, Exner evaluated the effects of increased exposure to sine-wave syllables on audio-visual perception. She found that even two hours of training produced a significant improvement in performance. However, what was not clear in her study is whether the performance had reached asymptotic levels. The present study addressed this by increasing the amount of training so the participants had more exposure to sine-wave speech. Participants were tested prior to training in auditory-only, visual-only, and auditory + visual conditions. After five hours of training under auditory-only conditions, a mid-test was administered. Finally a post-test was administered after the completion of all the training hours to determine whether intelligibility increased after exposure to sine-wave stimuli and whether audio-visual integration was impacted by auditory-only training.

It was anticipated that increased exposure to sine-wave speech would produce additional improvements in performance. It was also of interest to assess any potential effects on integration performance. Some previous work has suggested that a more ambiguous auditory stimulus will facilitate auditory-visual integration. However, some level of intelligibility of the auditory input is clearly required for an auditory contribution to the percept.

The present study specifically investigated whether additional training in the auditory alone condition improved auditory-only performance. In addition, the present study evaluated whether auditory-only training improved auditory-visual integration. It was expected that integration performance would not improve for two reasons: 1) Training should reduce ambiguity in the auditory signal. 2) Grant & Seitz (1998) have argued for the independence of the integration process.

The results of this investigation should provide insights into the intelligibility of highly reduced speech sounds that may be helpful in the design of aural rehabilitation programs for hearing impaired persons.

CHAPTER 2: Methods

Participants

Participants in the present study included five listeners. Three females and two males, ages 19-22, participated. All five listeners reported having normal hearing and normal or corrected vision. None of the listeners reported any training in linguistics or phonetics. None had heard of the McGurk effect. Participants were compensated for their involvement in the study.

Each talker produced each stimulus syllable five times, to create a set of tokens for stimulus creation.

Interfaces for Stimulus Presentation

Each participant was presented with three conditions: 1) visual only, 2) degraded auditory only, and 3) degraded auditory and visual. For the visual portions of the stimulus, a 50 cm video monitor was positioned directly in front of the participants outside the window of the sound booth. The monitor was positioned at eye level, approximately four feet away from the participant’s face when seated. Stimuli were presented on recorded DVDs using a DVD player for each condition. The sound was turned off for visual only presentations.

The degraded auditory stimuli were presented from the headphone output to the video monitor. TDH-39 headphones were used. The video monitor was turned off for auditory only presentations.

Stimulus Selection

A limited set of eight CVC syllables were presented as stimuli for the study. Previous digital recordings from three talkers were used in this study, including one female and two males between the ages of 20-23. All three talkers were native English speakers.

The syllables satisfied the following conditions:

1) The pairs of stimuli were minimal pairs, the initial consonant being the only difference.

2) All stimuli contained the vowel /ae/, because it does not involve lip extension or lip rounding.

3) Multiple stimuli were used in each category of articulation, including: place (bilabial, alveolar, velar), manner (stop, fricative, nasal), and voicing (voiced and voiceless).

4) All stimuli were presented without a carrier phrase.

5) Stimuli were known to elicit McGurk- like responses.

Stimuli

For each condition, the same single-syllable stimuli were administered:

Bilabial:	mat, bat, pat
Alveolar:	sat, zat, tat
Velar:	gat, cat

The four following dual-syllable stimuli were used in the degraded auditory and visual conditions. The first word indicates the visual stimulus; the second word indicates the auditory stimulus.

1) bat-gat

2) gat-bat

3) pat-cat

4) cat-pat

Visual and Audio Digital Video Recording

The syllables from the three talkers were converted into degraded speech stimuli using a PRATT script created by Chris Darwin. This program produces sine wave speech by reducing the stimuli into three sine waves representing the first three formants of the original signal, similar to the procedure described by Remez et al. (1981).

Video Explosion Deluxe video editing software was used to dub the degraded auditory stimuli onto visual representations of the talker. DVDs were burned and consisted of sixty randomized stimuli to minimize memorization among the participants. Four testing DVDs were created for each of the three talkers containing auditory and visual stimuli. Ten training DVDs were created for each of the three talkers containing auditory-only stimuli.

Testing Procedure

Testing for this study was completed in The Ohio State University's Speech and Hearing Department. Participants were instructed to read a set of instructions. The instructions explained the following to each participant: “You will be hearing and/ or seeing different talkers. The talkers will be saying different syllables (words), all ending in “at.” The syllables might be an English word such as, “mat,” or might not be an English word such as, “zat” or “gat.” The speech syllables that you will hear have been degraded, or have been “messed up.” Concentrate on what you hear and see, and tell me what syllable you think the person said. It is important that you take a guess for each syllable presentation.” The participants were given a written closed set of responses including; bat, pat, mat, dat, tat, nat, sat, zat, gat, cat, bgat, pcat, ptat, and bdat. This was given to them on a piece of paper that they could look at during testing. They were told there would be three testing conditions, auditory alone, visual alone, and auditory + visual and to respond verbally to what they perceived.

Participants were tested individually in a sound attenuating booth, with the door closed to create a quiet environment. A chair was positioned in the booth so that each participant could see the video monitor directly in front of them. The auditory stimulus was presented to the headphones worn by the participant.

Listeners were given a ‘pre-test’ under auditory-only (A), visual-only (V) and auditory-visual (AV) conditions, feedback not provided. For the ‘pre-test,’ listeners were presented with 60 A-only trials from each talker, 60 V-only trials from each talker, and 120 AV trials from each talker. Listeners were then given five training sessions, each one hour in length, with stimuli under A conditions only, feedback provided. A ‘mid-test’ was presented, in the same format of the ‘pretest.’ After this, another five training sessions were presented, each one hour in length, with stimuli under A conditions only, feedback provided. Finally, the listeners were given a ‘post-test,’ again following the format of the ‘pre-test’ and ‘mid-test.’ Total testing and training for each participant took approximately fifteen hours and was broken up into eight sessions of 1 1/2 – 2 hours. Within each session, breaks were encouraged to minimize fatigue.

CHAPTER 3: Results and Discussion

The pre-test, mid-test and post-test data were analyzed to determine whether training improved a subject's ability to correctly identify speech sounds. Results for two different types of stimuli were analyzed. First, percent correct identification was evaluated for the congruent stimuli (same auditory and visual syllables). This was evaluated for all three testing conditions, auditory (A), visual (V), and auditory-visual (A+V). Second, the percent response of the discrepant stimuli (different visual and auditory stimulus), which are known as the McGurk type responses, were evaluated. These responses were not recorded as percent correct because there is no "correct" response among the different stimuli. These responses were categorized into three different categories: auditory (the response exactly the same as the auditory stimulus), visual (the response was in the same viseme group as the visual stimulus), and "other" (the response differed from both the auditory and visual stimulus). The "other" responses were then categorized into three categories: combination (a response containing the first consonant of a stimulus of both the visual and auditory stimuli, e.g., bgat), fusion (a response in which the initial consonant is at a place intermediate to the visual and auditory stimuli, e.g., dat), and neither (the response differed from both the combination and fusion stimuli).

Percent Correct Performance

Figure 1 shows the percent correct identification for all three testing conditions, auditory (A), visual (V), and auditory-visual (A+V) for the pre-test mid-test, and post-test. Results shown are averaged across all listeners and talkers. There are several things worth noting from this figure. First, there was a significant improvement in performance of listeners in the A and A+V conditions from pre-test to post-test for all talkers. A two factor repeated measures ANOVA indicated a significant main effect of test (pre, mid, post), [F(2,8) = 24.7, p= .002]. In addition, a significant main effect of modality was observed (A, V, AV), [F(2,8) = 19.4, p= .005]. Means comparisons indicated significant improvement from pre-test to mid-test, but no significant improvement from mid-test to post-test. These results indicate that the listeners benefited more in the first five hours of training than in the last five hours of training. In addition, no significant interaction effect was found: [F (4,16) = 2.87, p= .08].

Figures 2, 3, and 4 show the percent correct identification for auditory-only, visual- only, and auditory-visual presentation for each of the three talkers, averaged across listeners, in the pre-test and post-test conditions. All talkers showed improved intelligibility from the pre-test to post-test in A, V, and A+V conditions. Although the intelligibility of all talkers did improve, the amount of improvement was variable across talkers. The variation across talkers supports the fact that individual talker characteristics can impact intelligibility of degraded speech signals (e.g., Andrews, 2007). Another thing worth noting among these graphs is shown in Figure 2. Figure 2 shows the percent correct identification for the auditory-only condition. A larger improvement following training was seen for talkers DA and JK than for talker KS. It is possible that the sex of talker KS (female) with a correspondingly higher pitched voice made it more difficult to identify her productions.

Training Sessions

Figure 5 shows the overall percent correct identification for auditory-only training for each listener, averaged across talkers. For most listeners, performance improved across sessions, as can be seen by comparing scores for training session one versus training session ten. Training data for individual listeners are shown in the Appendix.

Integration

Figure 6 shows the amount of AV integration exhibited across talkers and listeners for the pre-test, mid-test, and post-test. Averaged across talkers and listeners. The amount of AV integration was defined as the difference between audio-visual performance and the better single modality, A or V). Results showed no increase in AV integration from pre-test to mid-test and post-test. Again, this supports the idea that training improves performance only in the auditory condition, and did not impact integration ability.

McGurk Type Integration

As noted previously, there is no “correct” response for the discrepant stimuli. Therefore, Figure 7 shows the percent response for discrepant stimuli, as visual, auditory, and other responses (which consist of fusion, combination, or neither) for pre-test, mid-test and post-test, averaged across talkers and listeners. Listeners showed slightly more reliance on the auditory modality after training, with a corresponding reduction in “other” responses. This suggests that after the listeners were trained in the auditory alone condition, they changed their processing strategy for discrepant stimuli.

Figure 8 analyzes the “other” responses (which consist of fusion, combination, and neither responses) from Figure 7. Figure 8 shows percent McGurk-type integration for dual syllables, averaged across talkers and listeners. Listeners showed a larger percentage of combination responses (e.g.,bgat) prior to training, and more fusion responses (e.g., dat) after training. This result is not expected given previous studies in the laboratory, in which fusion integration percentages are typically high and combination responses are generally low. The confusion responses were likely lower for two different reasons: 1) The subjects were never presented with a combination stimuli such as bgat or pcat in training. During training in the auditory alone condition they were presented with eight single syllables; bat, mat, pat, cat, tat, sat, and zat. As a result the listeners may have altered their response strategies. 2) The combination response, possibilities, e.g., bgat, are typically not permissible consonant clusters in American English.

Confusion Matrices

Tables 1 and 2 show confusion matrices for auditory-only performance, in the pre-test and post-test conditions, averaged across listeners and talkers. The confusion matrices are shown to indicate how stimuli were perceived prior to training and after training. Results show an increase in the percent correct for every stimulus from pre-test to post-test.

The largest percent correct increase from pre-test to post-test were from the stimuli mat and gat. Mat is a nasal consonant (manner of articulation) that is characterized by a relatively low frequency resonance. Results indicate that with training the listeners were able to hear the nasalized sound. Gat has a velar place of articulation characterized by a steeply sloping F2 transition. Listeners apparently became better at comparing the steeply falling F2 transition with other stimuli. Future research might address why these two syllables were much more intelligible after training.

Tables 3 and 4 show confusion matrices for auditory + visual performance, in the pre-test and post-test conditions. Table averaged across all listeners and talkers. Results show an increase in the percent correct for every stimulus from pre-test to post-test. When listeners did pick the incorrect stimuli, a large percent of the incorrect responses were in the appropriate viseme categories, suggesting the use of visual information. Pat, bat, and mat were very easily confused due to the fact that they are all visually salient. /P/, /b/, and /m/ are all bilabials (same place of articulation) and only differ in manner of articulation and voicing, thus making it very hard to distinguish among these three syllables. Gat and cat were also easily confused because they were in the same viseme group. Gat and cat are both velar (place of articulation) stops (manner of articulation) that only differ by voicing. Zat and sat were also easily confused. Zat and sat are both alveolar (place of articulation), fricative (manner of articulation) consonants and only differ by voicing. In the pre-test tat was mostly confused with other alveolars, but in the post-test it was mainly confused with cat. Cat and tat are both voiceless stops and only differ by place of articulation. These results indicate that the listeners showed an improvement in the ability to detect voicing after training.

Comparison with Previous Studies

The findings in this study are consistent with those of Exner (2008). Exner evaluated the effects of increased exposure to sine-wave syllables on auditory-only and audio-visual perception. She found that even two hours of training produced significant improvements in performance. What was not clear in her study is whether the listeners had reached asymptotic levels. This study increased the number of training hours the subjects received from two training hours to ten training hours to see if audio-visual performance would show further improvement. The present study suggests that additional training does improve performance with this stimulus reduction, at least up to five hours of exposure.

Chapter 4: Summary and Conclusion

Overall, results of testing indicated an improvement in performance in the A and A+V conditions from the pre-test to post-test, with a significant improvement from pre-test to mid-test, but not from mid-test to post-test. This suggests that the listeners benefited from the first five hours of training more than the last five hours. Also, the amount of integration did not a change as a function of training, suggesting that training in an auditory alone condition improved only listeners' auditory performance. This finding argues that different types of training methods may be necessary in order for integration to occur. For example, training listeners in all three modalities, A-only, V-only, and A+V conditions might be needed. Interestingly, the degree of benefit from training varied across listeners, as well as across the talkers. Some listeners improved more than others. Also, some listeners improved substantially less with talker KS (female), and more with talkers JK and DA (male). Further analysis must be done to determine how the sex of the talker influences intelligibility for this type of auditory stimulus reduction.

Listeners in this study were trained for a longer period than the previous study by Exner (2008), who provided only two hours of training time. This study improvements were made from pre-test to post-test, but with only a significant improvement from pre-test to mid-test. Therefore, training periods in excess of five hours are not necessary for this stimulus set. Training is important, but how one is trained is ultimately more important.

Overall, results from this study indicate that with training a listener's performance can improve considerably when listening to highly degraded auditory signals. This is important in gauging the amount of information that is actually available in such signals. Generalization of this skill improvement to other stimulus sets is a question that should be addressed in future work.

In addition, specific attention should be given to the potential benefits of training in A+V conditions, to determine if integration skills can be developed. The present study suggests that training in all three conditions, auditory-only, visual-only, and auditory + visual is needed to improve one's integration ability. Results of this study can be used for the design of aural rehabilitation programs for hearing impaired persons. Training for patients with a hearing loss in dual modalities may help an individual make use of what residual hearing he or she has, consistent with the argument of Grant and Seitz (1998) that integration is a process independent of auditory or visual processing.

Chapter 5: References

Anderson. C. (2007). *Auditory and Visual Characteristics of Individual Talkers in Multimodal Speech Perception.* The Ohio State University Department of Speech and Hearing Science. Unpublished Honors Thesis, Project Advisor: Janet M. Weisenberger, Ph.D.

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

Andrews, B. (2007). *Auditory and visual information facilitation speech integration*. Senior Unpublished Honors Thesis, The Ohio State University.

Exner, M. (2008). Training *Effects in Audio-Visual Integration of Sine Wave Speech*. Senior Unpublished Honors Thesis, The Ohio State University.

Grant, K.W. (2002). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, 104 (4), 2438-2450.

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*. 104 (4). 2438-2450.

Huffman, C. (2007). *The role of auditory information in audiovisual speech integration*. Unpublished Honors Thesis, The Ohio State Univeristy.

Jackson, P.L. (1998). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*, 90 (5), 99-114.

Ladefoged, P. (2006). *A Course in Phonetics-Fifth Edition*. Boston : Wadsworth.

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). Test 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

Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrel, T.D. (1981). Speech Perception without traditional speech cues. Science, 212 (4497). 947-950.

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

Tamosiunas, M. (2007). *Auditory-visual integration of sine-wave speech.* Unpublished Honors Thesis, The Ohio State Univeristy

List of Tables and Figures

Figure 6: Amount of integration in the pre-test, mid-test, and post-test, averaged across listeners and talkers.

Figure 7: Percent response dual-syllable stimuli for auditory-only, visual-only, and auditory + visual conditions in the pre-test, mid-test, and post-test, averaged across listeners and talkers.

Figure 8: McGurk type integration for combination, fusion, and “other” responses in the pre-test, mid-test, and post-test, averaged across listeners and talkers.

Table 1
Pre-test Confusion Matrix, Auditory-only
Average of all listeners and talkers

response

		bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	hat	ptat	bgat	pcat	bdat
stimuli	**bat**	**11.4**	10.86	8.23	2.53	3.16	17.72	4.43	20.25	5.7	12.66			1.27	1.27	0.63
	pat	6.06	**12.12**	8.48	6.06	11.5	8.48	7.88	12.7	12.12	6.06		2.42	2.42	0.61	3.03
	mat	15.5	3.45	**62.07**					1.72	5.17	3.45			3.45	1.72	1.72
	gat	3.49	4.07	3.49	**35.47**	6.98	5.81	1.16	8.14	4.65	13.37		2.33	8.14	0.58	2.33
	cat	2.94	3.53	7.65	13.53	**28.24**	6.47	3.53	6.47	4.71	0.59		1.76	8.82	10	0.59
	zat	1.69		15.25	5.08	3.39	**37.29**		13.56	8.47	5.08		1.69	3.39	1.69	3.39
	tat	5.08	6.78	10.17	3.39	25.42	5.08	**10.17**	6.78	8.47	6.78			6.78	5.08	
	sat	5.08	3.39	1.69	11.86	1.69	27.12	5.08	**25.42**	5.08	8.47			1.69		3.39

Table 2
Post-test Confusion Matrix, Auditory-only
Average of all listeners and talkers

response

		bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	hat	ptat	bgat	pcat	bdat
stimuli	**bat**	**39.49**	13.38	6.38	3.18	5.73	10.19	5.73	12.1		1.91					0.64
	pat	15.43	**27.78**	8.64	6.79	4.32	2.47	26.54	6.79	0.62			0.62			
	mat	5.08		**83.05**	3.39	3.39	3.39	1.69								
	gat	5.14	8.57	1.71	**65.71**	6.86	4	3.43	4					0.57		
	cat	5.85	17.54	2.92	12.28	**42.69**	2.34	11.11	2.92				0.58		1.75	
	zat	3.45	8.62	5.17	1.72	4.6	**56.9**	5.17	10.34		1.72					
	tat	3.39	11.86	5.08	12.07	35.59	6.78	**23.73**					1.7			
	sat	10.17	5.08		8.47	8.47	23.73	6.78	**33.9**		1.7			1.7		

Table 3
Pre-test Confusion Matrix, Auditory + Visual
Average of all listeners and talkers

response

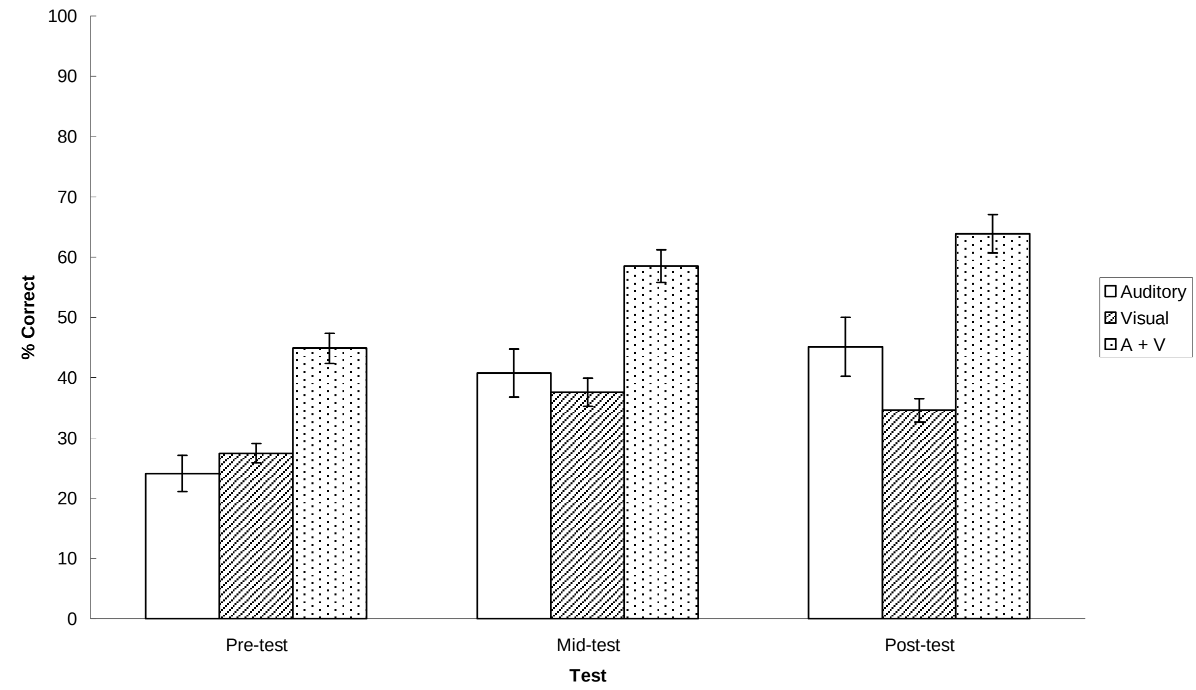
		bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	hat	ptat	bgat	pcat	bdat
	bat	**53.9**	26.47	8.82	0.98		3.9					0.98		3.92	0.98	
	pat	27.45	**37.25**	19.61	0.98	1.96		1.96		0.98	0.98		2.94	0.98	1.96	0.98
	mat	18.64	6.78	**70.34**					0.85	0.85			0.85			
stimuli	**gat**	0.88	0.88		**54.87**	12.39	6.19	1.77	9.73	4.42	5.31			2.65		0.88
	cat	1.85	2.78	1.85	23.15	**47.22**	1.85	1.85	2.78	2.78	0.93		1.85	4.63	9.26	0.93
	zat				1.78	4.42	**61.95**	0.88	15.93	.88	12.39			1.77		
	tat		1.78	7.08	2.65	22.12	19.47	**16.81**	15.93	2.65	7.08		0.88		3.54	
	sat		0.85	0.85	3.42	0.85	41.03	0.85	**35.9**	0.85	11.97					1.71

Table 4
Post-test Confusion Matrix, Auditory + Visual
Average of all listeners and talkers

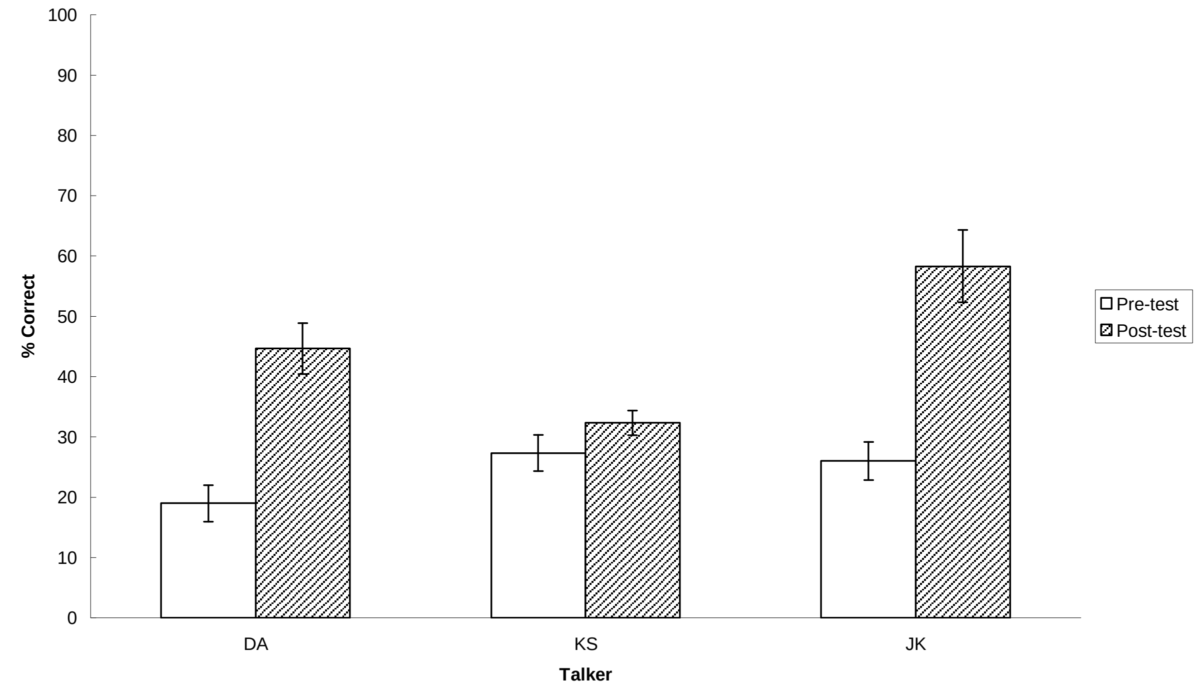
response

		bat	pat	mat	gat	cat	zat	tat	sat	nat	dat	hat	ptat	bgat	pcat	bdat
	bat	**75.49**	13.73	3.92	0.98		0.98	1.96	2.94							
	pat	28.85	**60.58**	6.73				2.88			0.96					
	mat	9.4	4.23	**82.91**			3.41									
stimuli	**gat**	0.89	0.89	0.89	**77.68**	11.61	2.69	0.79	2.68					0.89		
	cat		5.5		12.8	**66.06**	0.92	3.67	2.75	0.92	0.92		1.83	1.83	1.83	0.92
	zat		0.85	0.85	4.24	0.85	**71.19**	2.54	13.56		5.08					0.85
	tat		0.86		1.72	37.07	6.03	**40.51**	10.34	1.72	0.86		0.86			
	sat		1.71		9.4	0.85	38.46	4.27	**39.32**		5.13					0.85

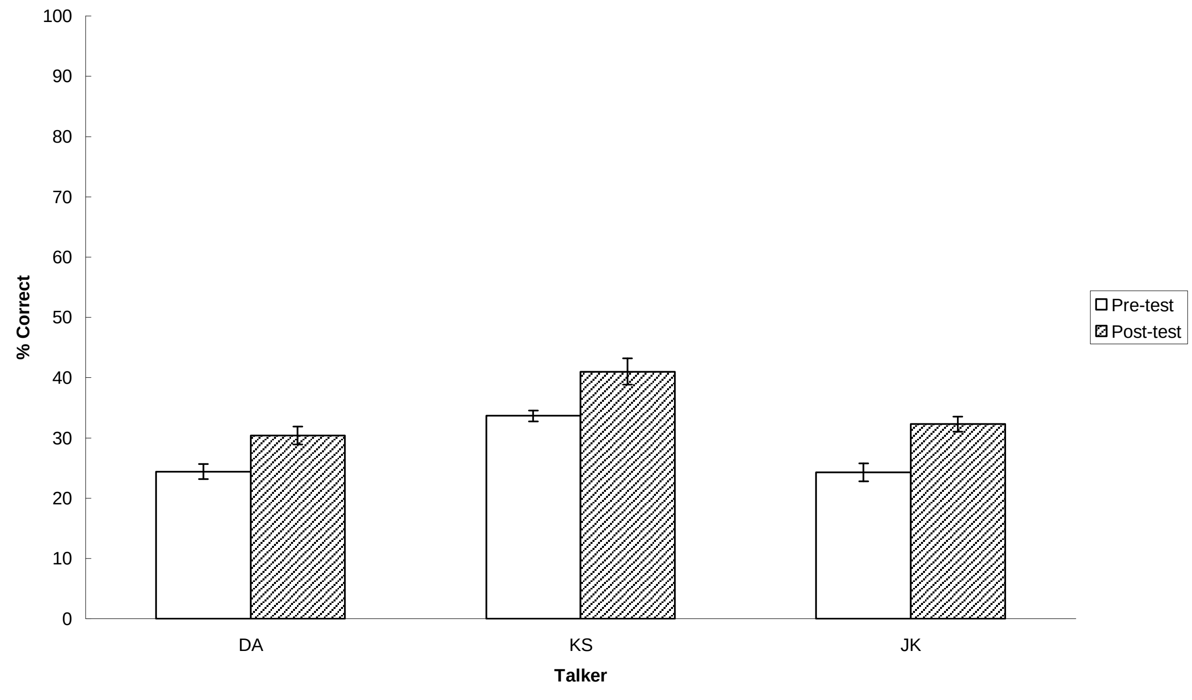
Overall Percent Correct
Figure 1



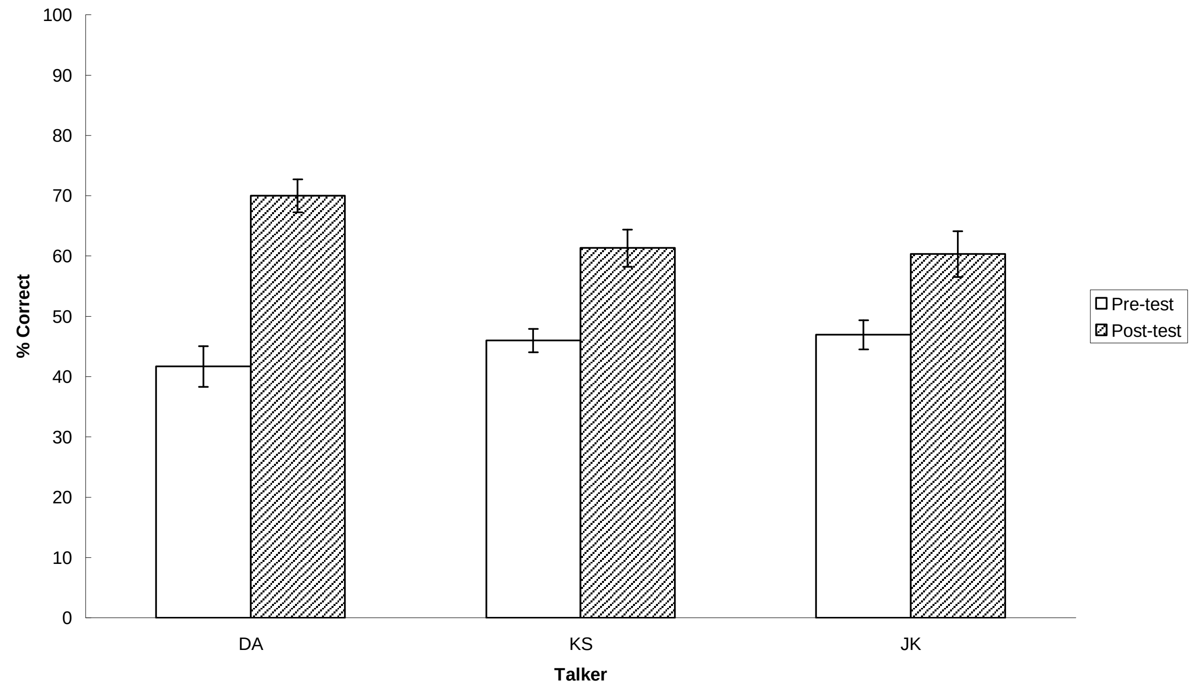
Percent Correct Identification Auditory Only by Talker
Figure 2

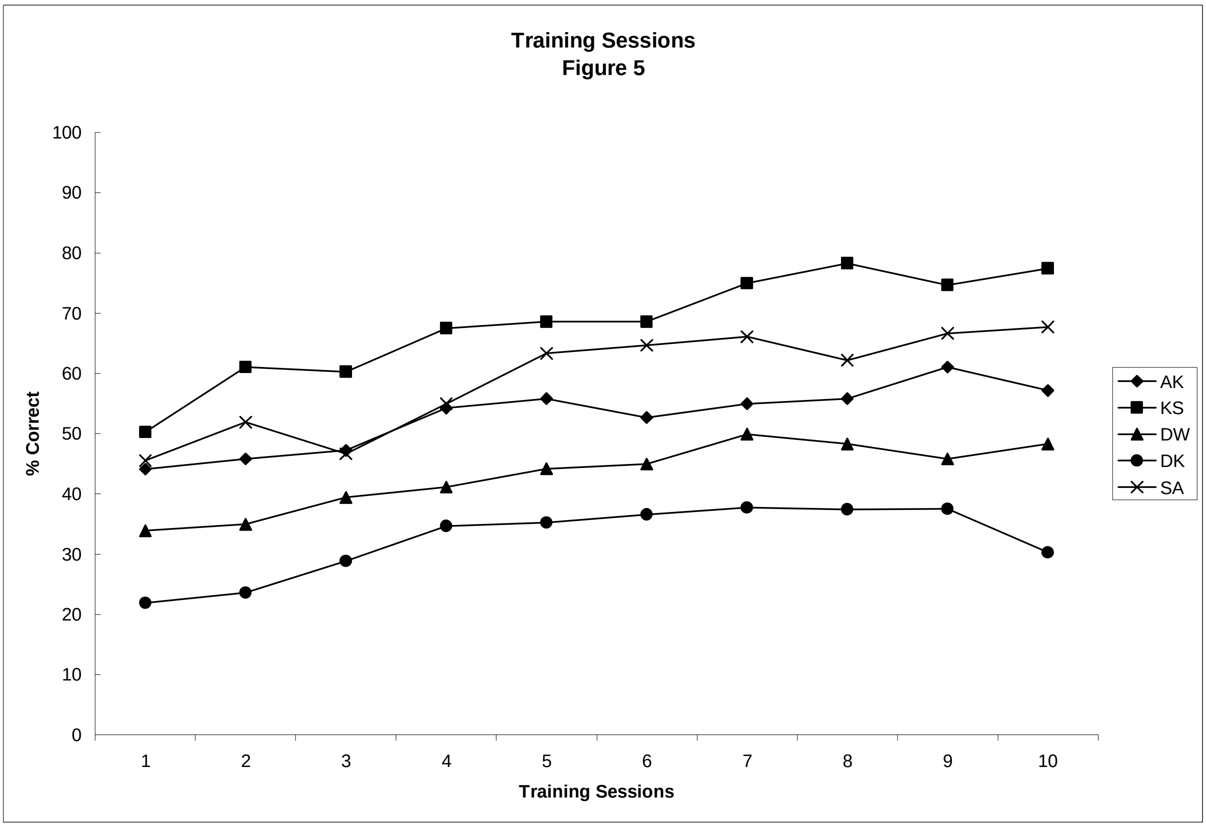


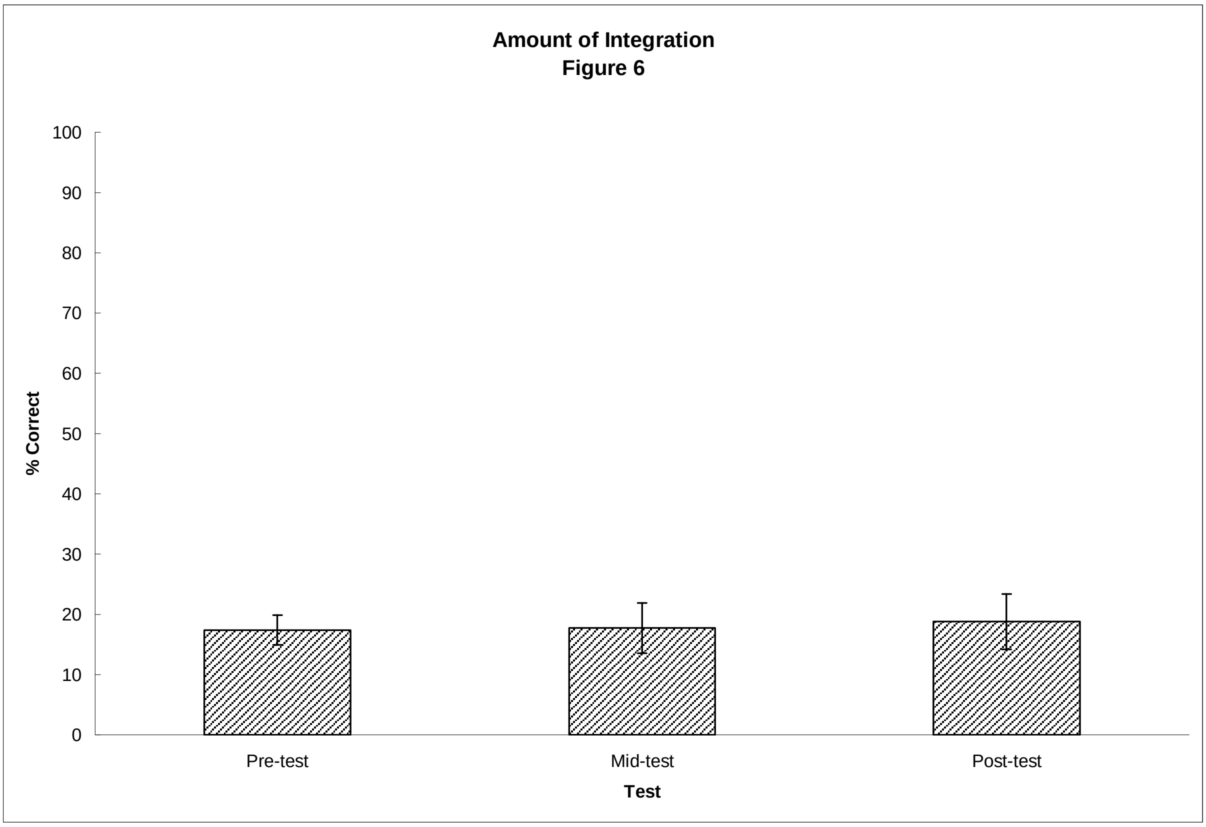
Percent Correct Identification Visual Only by Talker
Figure 3

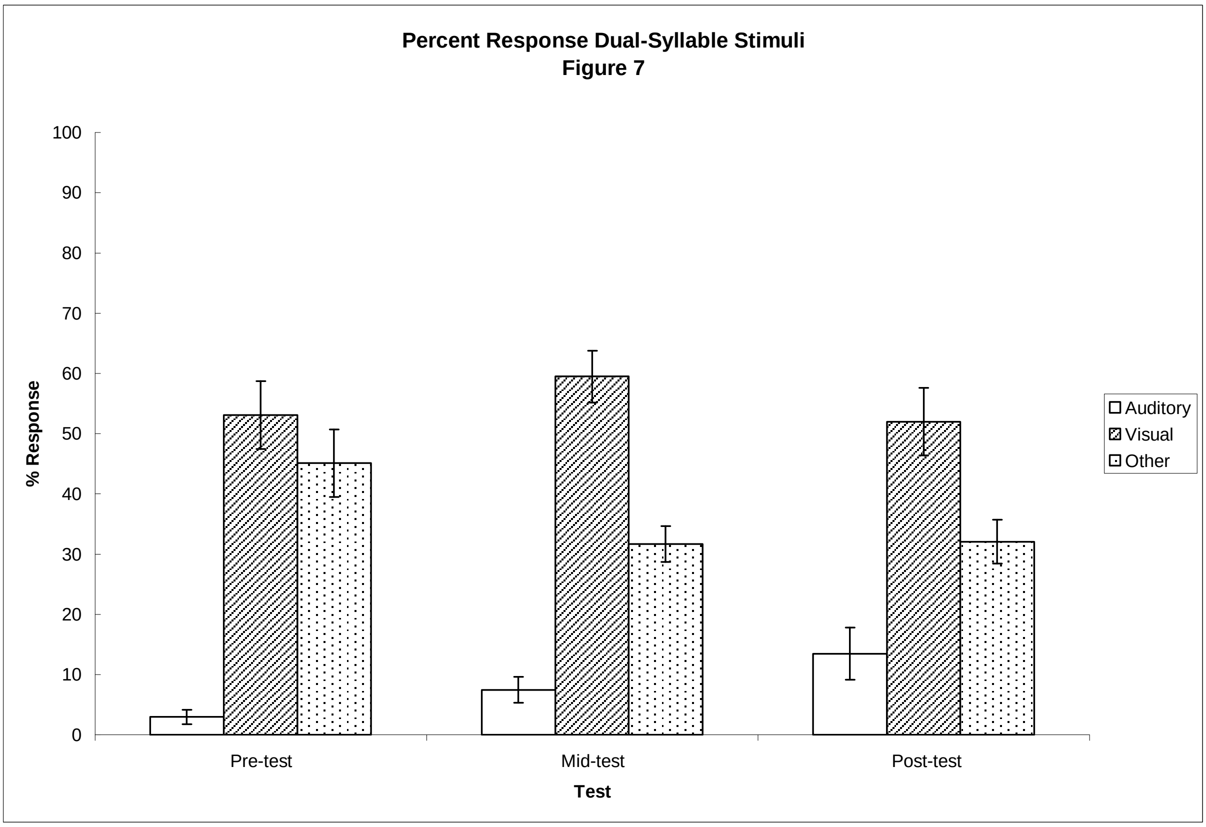


Percent Correct Identification Auditory and Visual by Talker
Figure 4

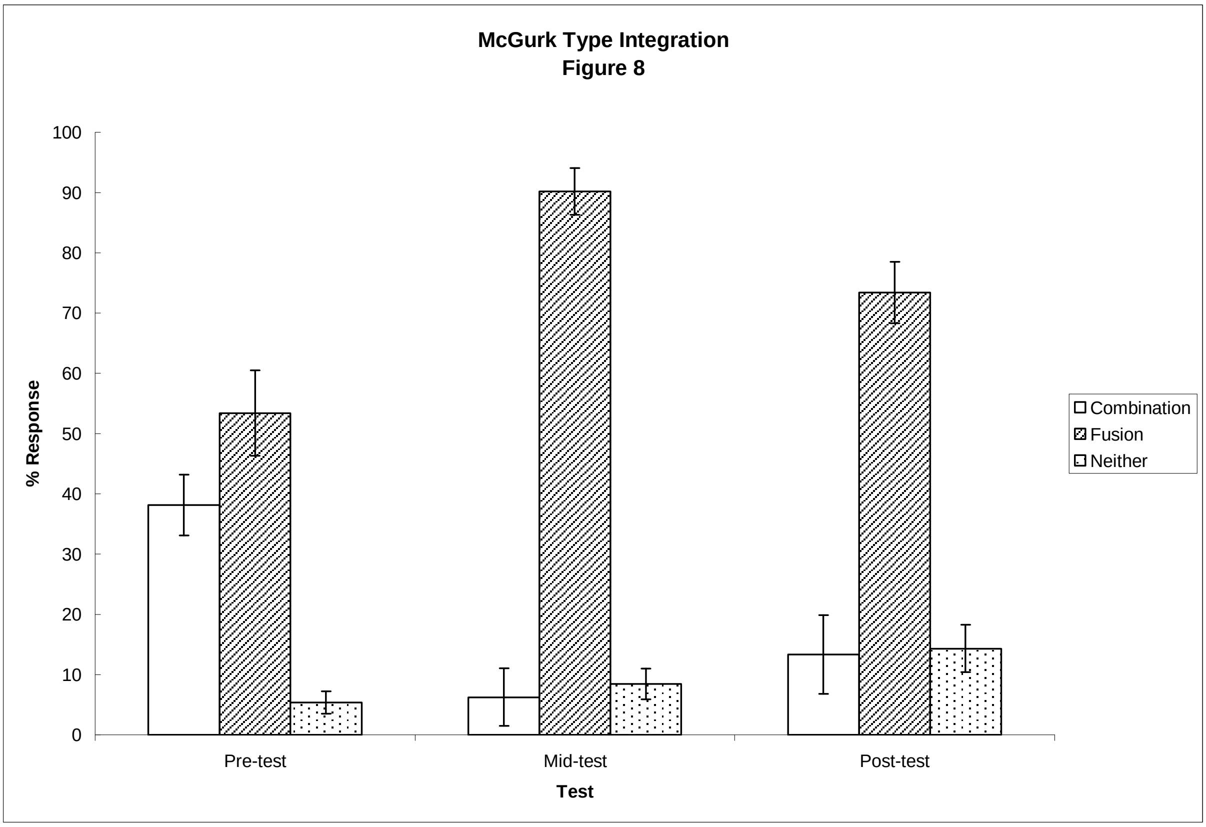



Training Sessions
Figure 5
% Correct
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
Training Sessions
AK
KS
DW
DK
SA


Amount of Integration
Figure 6
% Correct
100
90
80
70
60
50
40
30
20
10
0
Pre-test
Mid-test
Post-test
Test



Percent Response Dual-Syllable Stimuli
Figure 7

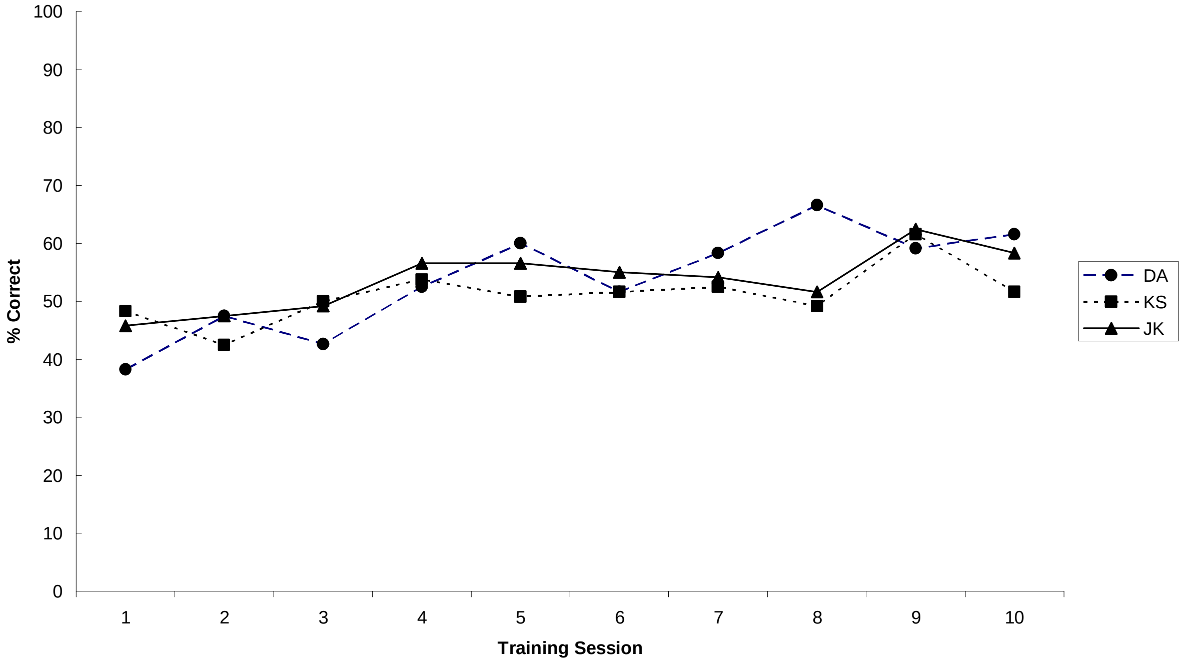


McGurk Type Integration
Figure 8

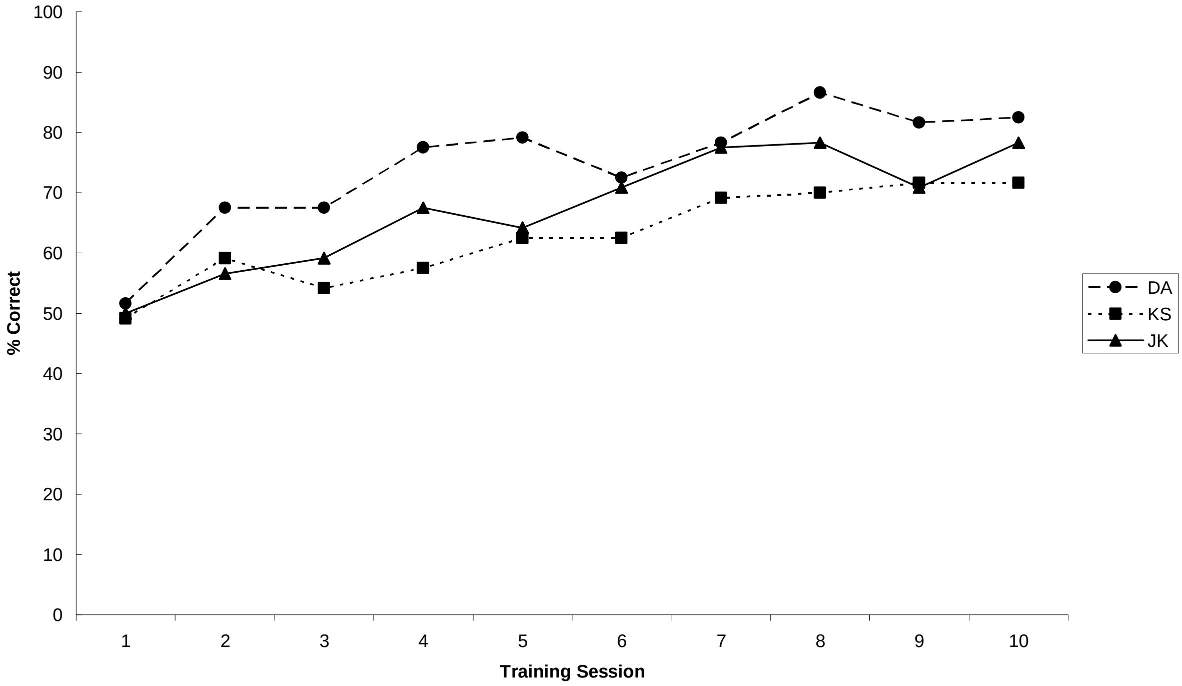
Appendix

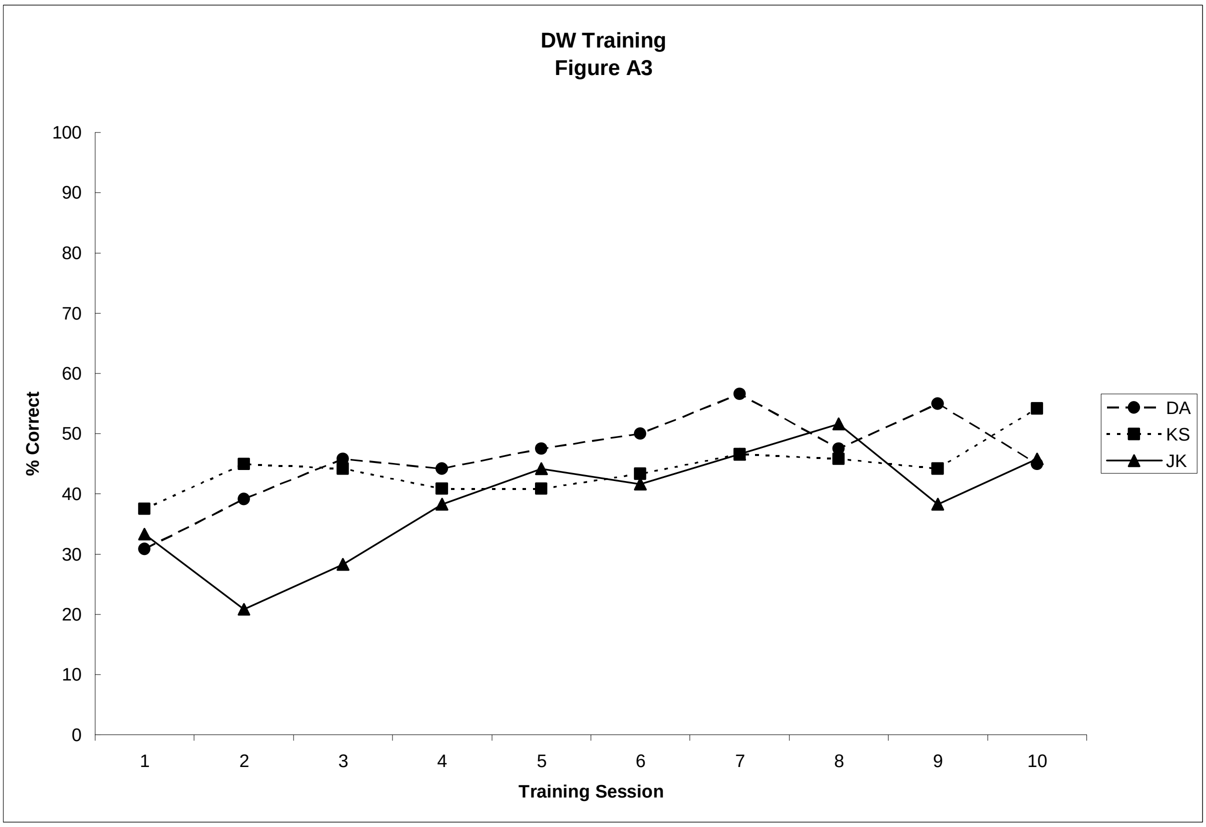
List of Figures

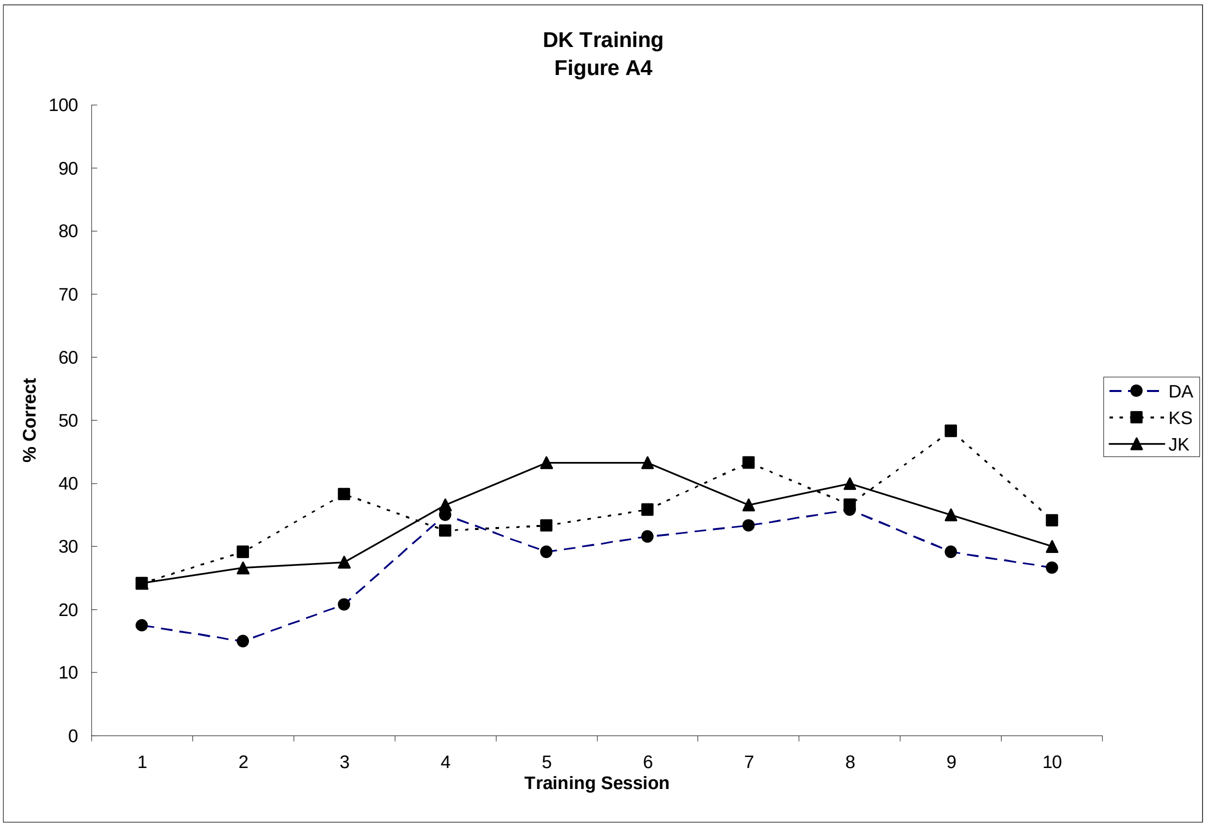
AK Training
Figure A1

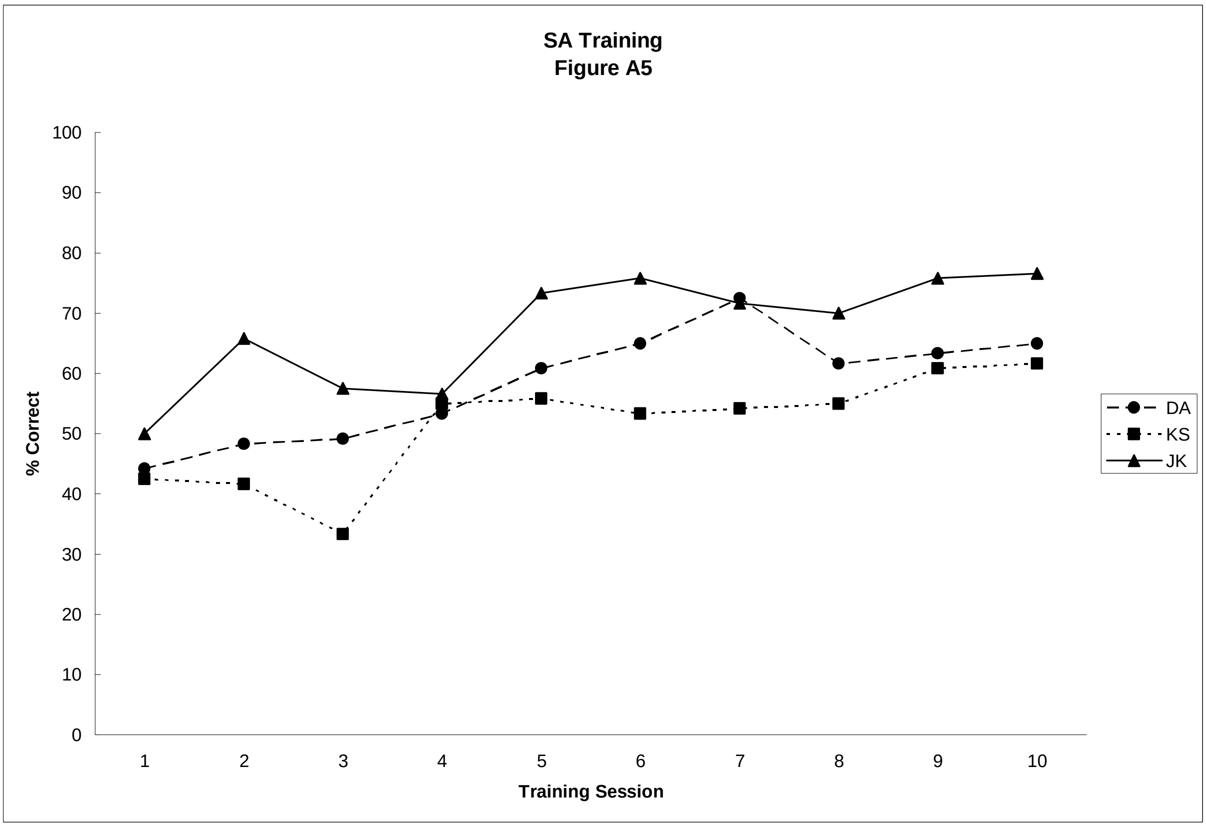


KS Training
Figure A2




DW Training
Figure A3
% Correct
0
10
20
30
40
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10
Training Session
DA
KS
JK


DK Training
Figure A4
% Correct
0
10
20
30
40
50
60
70
80
90
100
1
2
3
4
5
6
7
8
9
10
Training Session
DA
KS
JK


SA Training
Figure A5
100
90
80
70
60
50
40
30
20
10
0
% Correct
1
2
3
4
5
6
7
8
9
10
Training Session
DA
KS
JK