Influence of Auditory Closure and Working Memory on Audio-Visual Perception of Speech

[1]Deepashree & [2]Sandeep Maruthy

Abstract

The present study investigated the influence of the speech perception in noise and the working memory on audio-visual integration in speech perception. Sixty six adult native speakers of Kannada in the age range of 18 to 70 years, having normal hearing and normal or corrected vision of 6/6, participated in the study. The participants were assessed for their speech perception in noise, working memory capacity and audio-visual integration. In the assessment of audio-visual integration there were three conditions in quiet (A, V, AV) and eight conditions in the presence of noise (A, AV in +5dB, 0dB, -5dB & -10 dB SNR). Results showed that both SPIN and working memory have positive effect on speech perception through AV-mode. The influence was particularly significant at lower SNRs and the relationship was direct. That is, AV speech perception was better in individuals who had better SPIN and/or working memory. Thus, it can be concluded that internal redundancy controlled by central auditory mechanism/s (as evident through SPIN) and the cognitive abilities (represented as working memory, a cognitive factor) has a direct relationship with the AV speech perception and hence demands attention while interpreting on AV speech perception.

Keywords:*Speech perception, working memory, auditory enhancement, visual enhancement.*

Introduction

Speech perception is influenced by visual cues even though it is primarily an auditory process. In instances where auditory cues are compromised (such as in noisy environments or hearing impairment), visual input has been reported to significantly improve speech intelligibility by supplementing the missing auditory cues (Tye-Murray, Sommers & Spehar, 2007; Munhall, Kroos, Jozan & Vatikiotis-Bateson, 2004; MacLeod & Summerfield, 1987). The benefits provided by bimodal presentation of a speech signal are larger when auditory stimuli are degraded than when the speech signal is clear (Sumby & Pollack, 1954; O'Neil, 1954; Neely, 1956; Erber, 1969; Grant & Seitz, 2000; Rudmann, Mc Carley & Kramer, 2003; Bernstein, Auer & Takayanagi, 2004; Ross, Saint-Amour, Leavitt, Javitt & Foxe, 2007). But studies on relationship between the amount of redundancy and audio-visual (AV) speech perception show equivocal results (Sumby & Pollack, 1954; Erber, 1969; Summerfield, 1979; Middelweerd & Plomp, 1984; Shannon, Zeng & Wygonski, 1998; Anderson, 2006; Huffman, 2007).

Sumby and Pollack (1954) demonstrated that with increase in difficulty of auditory-only perception, the benefits obtained by combining the auditory and visual speech information also increased. Erber (1969) reported an improvement of 60% in word recognition scores from auditory-only condition to AV condition at -10 dB SNR for young adults.Similar results were reported using sentence materials (Middelweerd & Plomp, 1984; Summerfield, 1979). Whereas Anderson (2006) and Huffman (2007) reported that the amount of AV integration did not vary across different auditory signal manipulations and hence, systematically removing information from the auditory stimulus does not necessarily affect the degree of integration benefit.

The studies on effects of age are equivocal. The study by Tye-Murray, Sommers and Spehar (2007) compared the dependency on visual cues between older adults with and without hearing loss and suggested that older adults with hearing loss may rely much more on the visual cues than those with normal hearing. The degree of hearing loss was also shown to affect the integration of auditory and visual syllables as reported by Grant, Walden and Seitz (1998). Further, the older adults are reported to be less successful in combining information across two or more sensory modalities (Shoop & Binnie,1979; Middelweerd & Plomp, 1984; Plude & Hoyer, 1985). But in contrast, Spehar, Tye-Murray and Sommers (2008) suggested that the intermodal and intra-modal integration abilities are largely resistant to changes with age and hearing loss. The researchers have argued that age related changes in cognitive or central auditory processing abilities play a limited role in the poor recognition of speech (Sommers, 1997; Schneider, Danema, & Pichora-Fuller, 2002).

It is shown in one of the earlier studies that older adults have decreased processing resources available to them and hence perform poorer than young adults on memory tasks (Craik & Byrd, 1982). It is suggested that there is an interdependence of aging sensory systems and cognitive functions (Li & Lindenberger, 2002). The construct of working memory is considered important in the study of cognitive aging even though there are other measures of processing resources such as perceptual speed and reasoning. Studies by Park and colleagues

[1]Email: nimmadeepa@gmail.com,
[2]Reader in Audiology, Email: msandeepa@gmail.com

(1996, 2002) measured the working memory capacity in adults across life span and it was found that it was greatest in young adulthood and then decreases across life span. Martin and James (2005) proposed that age-related deficits in processing of inter-hemispheric information may bring about some of the listening problems in older adults. Hence, it is suspected that age-related changes in working memory may provide basis for decreased age-related performance on a range of cognitive tasks.

It has been explained that speech understanding needs a constant encoding of information into and out of the working memory and manipulation of the information stored in memory (Pichora-Fuller, Schreider & Daneman, 1995). Hence, loss of working memory hinders speech communication and may contribute to the age-related declines in the comprehension of speech which may be evident even in favorable listening conditions. In adverse listening conditions such as in the presence of noise, this may degrade further (Cohen, 1979, 1981; Light et al., 1982; Pichora-Fuller, Schreider & Daneman,1995). On the contrary, some of the earlier investigators have argued that age related changes in cognitive or central auditory processing abilities play a limited role in the poor recognition of speech (Sommers, 1997; Schneider, Danema, & Pichora-Fuller, 2002).

Thus, it is clear that the dependency on the visual information increase with degradation of the auditory signal, even though there are mixed opinions about the effect of the amount of redundancy on integration. This means that subjects depend on the information in other modalities when external redundancy of the auditory modality is cut down. For participants with hearing loss and processing disorder, there are equivocal studies about the effect of internal redundancy on audio-visual speech perception. The studies on aging effects have been inconclusive due to poor control of the variables like hearing loss and visual problem. Further, there was no systematic study to understand the relationship between working memory and AV speech perception. Hence, the present study was taken up to scientifically study the relationship among working memory, speech perception in noise and AV speech perception.

Method

In the present study, quasi-experimental research design was used to test the null hypothesis that there is no effect of speech perception in noise and working memory on audio-visual speech perception. The following method was used to test the hypothesis.

Participants

Sixty six normal hearing adults in the age range of 18 to 70 years participated in the study. The selected participants had pure tone thresholds within 15 dBHL at octave frequencies between 250 Hz and 8 kHz (ANSI, 1996), and normal or corrected vision of 6/6. All the selected participants were native speakers of Kannada.

All the participants were assessed for their speech perception in noise and working memory (details available later in this chapter). Based on their performance in speech in noise and working memory tests, they were divided into 4 groups. Confidence intervals (95%) of speech in noise scores and the working memory scores of the 66 participants were used as cut-offs. The scores less than lower boundary were grouped as 'poor' and the scores higher than the upper boundary were grouped as 'good'. As a result, 4 groups were formed.

Group I, named as Low speech in noise (LowSPIN) group had 20 participants with poor speech perception in noise i.e., speech identification scores of less than or equal to 72% (lower boundary score of confidence interval) at 0 dB SNR.

Group II, named as High speech in noise (HighSPIN) group had 36 participants with good speech perception in noise i.e., speech identification scores of more than 80% (upper boundary score of confidence interval) at 0 dB SNR.

Group III, named as Low working memory (LowWM) group had 14 participants with poor working memory i.e., working memory scores of less than or equal to 69% (lower boundary score of confidence interval).

Group IV, named as High working memory (HighWM) group had 21 participants with good working memory i.e., working memory scores of more than 76% (upper boundary score of confidence interval).

A written consent was obtained from each participant prior to their inclusion in the study.

Test Stimulus

Six lists of bi-syllabic Kannada (Dravidian language spoken in the state of Karnataka, India) words having ten words in each list were developed specifically for the purpose of the present study.

To begin with, 300 bi-syllabic Kannada words, frequently spoken by native speakers were collected from recorded speech samples, news papers and media interviews. From this list, 23 words with clusters were omitted leaving 277 words in the list. These 277 words were then given to 15 native speakers of Kannada to rate them according to familiarity on a 3-point scale (unfamiliar- if they were unaware of the word, familiar, & very familiar- if the word occurred very frequently in conversation). Out of these, 169 words, which were rated 'very familiar' by more than 12 participants (80%) were considered for the next level.

The selected 169 words were audio recorded using a

computer with adobe audition (version 1.5) software at a sampling frequency of 44,100 Hz and 16 bit digitization using a unidirectional microphone, in an acoustically treated room. An adult female native speaker of Kannada, who was a professional voice user, uttered the words. The stimuli were further edited for removal of noise and, hiss reduction and a gap of three seconds was introduced between consecutive words, using the same software. Root Mean Square (RMS) normalization was done for all words in order to minimize differences in the stimulus amplitude.

The variability with respect to audibility was reduced and in turn the homogeneity across spondaic words was increased using the standard procedure (Hirsh, Reynolds, and Joseph, 1954). Ten speech and hearing professionals who had Kannada as their native language and with a minimum of three years training were selected for this procedure. To begin with, speech recognition threshold (SRT) was found out for all ten subjects using the procedure by Tillman and Olsen (1973), a descending method for SRT measurement. After obtaining SRT, the paired-words were presented at +4, +2, 0, -2, -4 and -6 dBSL (ref: SRT), and the participants were asked to orally repeat the words. The whole list of 169 words was presented at each sensation level (SL) and it was randomized during each presentation. The responses obtained from the ten listeners were analyzed to identify 'Easy' and 'Hard' words. The words which were missed once or less by all listeners when the list was presented at +4, +2, 0, -2, -4 and -6 dBSL were considered as 'Easy words'. Whereas the words which were missed five or more times by all listeners when presented at +4, +2, 0, -2, -4 and -6 dBSL were considered as 'Hard words' (based on procedure by Hirsh et al., 1952). All the 'Easy' and 'Hard' words were eliminated to increase the homogeneity with respect to audibility. This finally resulted in sixty bisyllabic Kannada words.

The frequency of occurrence of all the speech sounds in the list of sixty bisyllabic Kannada words were calculated in order to find out whether the list of bisyllabic words was phonetically balanced. Although this was not mandatory for testing the objectives of the current study, the investigator felt that a phonetically balanced list would have been an advantage while generalizing the results of the present study. Hence, the frequency of occurrence of sounds in the list was tested using the method used by Ramakrishna, Nair, Chiplunkar, Ramachandran and Subramanian (1962). The probability of any speech sound was obtained by taking ratio of number of occurrence of speech sound in the list to the total number of speech sounds in it. The relative frequencies of speech sounds obtained in the present study were compared with relative frequencies of speech sounds obtained by Ramakrishna et al., 1962. Both the relative frequencies were similar and hence the present list was phonetically balanced. Finally, from the list of sixty words, six lists having ten words in each list were obtained. The total number of vowels and consonants in the list which were obtained from the Ramakrishna et al.'s method were distributed equally across the six lists. Thus, it was attempted to keep similar frequency of occurrence of vowels and consonants in all the six lists of bisyllabic words.

Audio and Video Recording

Five adult, female, native speakers of Kannada, who were speech and hearing professionals with clear articulation, were selected and the video of final six lists were recorded by a professional videographer using a digital video camera. A white screen was used as a background and the participants were instructed to avoid bright clothing to avoid distractions in the visual stimuli. The participants were also instructed to produce the words clearly without exaggerating the articulators, reduce eye blinks and avoid head movements, while recording. The words which were articulated unclearly were recorded twice.

Simultaneously, along with the video recording, the audio of the speech stimuli was recorded digitally for all the five participants. The audio recording was done using a collar microphone, at a sampling frequency of 44,100 Hz and 16 bit digitization, using Praat Software (version 5.1.31). Out of the 5 samples (recordings of 5 individuals) of stimuli, best sample was selected based on clarity of stimuli visually as well as auditorily. The audio and video recordings were done in a sound treated room.

The auditory stimuli were edited using adobe audition (version 1.5) software for noise and hiss reduction and a gap of five seconds was introduced between the words. All the words were normalized to a constant scaling factor. To test in the degraded conditions, all the six lists were further superimposed with speech noise at +5 dB, 0 dB, -5 dB and -10 dB SNRs.

The visual stimuli of the selected subject were edited using Windows movie maker software, to introduce a gap of five seconds between the words. The original audio of the video sample was muted and the auditory stimuli which was recorded using Pratt software was overlapped with the visual stimuli. This was necessary as the original audio was distorted and had high background noise.

To test for homogeneity in auditory and visual stimuli among the lists, all the lists were presented to five Kannada native speakers in three modalities; only Auditory (A), only Visual (V) and Auditory-visual (AV). The pilot comparison showed that the scores obtained for all the lists were similar.

Instrumentation and Test Environment

In the present study, a calibrated Madsen Orbiter-922 type I diagnostic audiometer with TDH-39 headphones was used to estimate the air-conduction thresholds and administer speech audiometry. A laptop computer with windows movie maker was used for video editing. The Pratt and Adobe Audition softwares were used for recording, editing and presenting the test stimuli. A digital video camera was used to record visual stimuli and a Sony MX78 omni-directional collar microphone was used to record auditory stimuli. Headphones were used to present auditory stimuli in 'A' mode and 'AV' mode.

All tests were administered in an acoustically treated room with noise levels at permissible limits (ANSI S3.1, 1991).

Test Procedure

The procedure started with preliminary evaluations which included case history, puretone audiometry and speech audiometry. For all the participants, the puretone air conduction thresholds (0.25, 0.5, 1, 2, 4, and 8 kHz) and speech recognition thresholds were obtained monaurally for both the ears. Only the individuals who fulfilled all the subject selection criteria were chosen for the study. After preliminary evaluation, the procedure included assessment of speech perception in noise, working memory and audio-visual integration.

Speech perception in noise (SPIN) was binaurally tested at 0 dB signal to noise ratio (SNR). The phonemically balanced (PB) word list developed by Yathiraj and Vijayalakshmi (2005) was used as signal and was presented along with the speech noise. The participants were asked to repeat the words and total number of words repeated correctly was noted down. The percentage of correct responses was calculated by dividing number of correctly repeated words by the total number of words and multiplying this factor by 100.

The procedure used to measure working memory capacity was adapted from versions of the operation span task used by Kane, Hambrick, Tuholski, Wilhelm, Payne and Engle (2004). Guidelines recommended by Conway, Kane, Bunting, Wilhelm and Engle (2005) were followed during administration and scoring. The operation span task consisted of 5 items and 20 elements. The number of elements per item varied from 2 to 6. Each element consisted of a mathematical operation which included addition and division, and a Kannada word (E.g. (6/3) + 7 = 8? 'Ka:ge'). The participant's task was to read the mathematical problem aloud, then say 'yes' or 'no' to indicate whether the given answer was correct or wrong, and then say the word. After all the elements in an item were presented, the participants were required to repeat all the words in that item. The difficulty of the items was randomized such that the number of elements was unpredictable at the outset of an item. The scoring for mathematical problem and words were done separately. The participants who scored less than seventeen out of twenty (80%) in mathematical problem were not considered for analysis as results of those subjects are not valid. For each correct item, one mark was given, only if all the elements were repeated correctly. If some of the elements were incorrect, the number of correct elements was divided by total number of elements. Finally, the scores of all five items were added and divided by five to obtain the final score for working memory. The scores for working memory ranged between 0 and 1.

The testing for the assessment of audio-visual integration was conducted in quiet as well as in the presence of noise. In quiet, stimuli were presented in 3 modalities (auditory only, visual only & auditory-visual). In the presence of noise, the stimuli were presented in 2 modalities (auditory only, & auditory-visual) at four different signal to noise ratios (+5dB, 0dB, -5dB and -10 dB SNR). The audio stimuli were presented binaurally through headphones at most comfortable level (MCL). The visual stimuli were presented from wide screen laptop with fifteen inches and the distance between the participant and laptop was maintained at fifteen inches. The audio of the video stimuli was muted while testing in V mode. Finally there were three conditions in quiet (A, V, AV) and eight conditions in the presence of noise (A, AV in +5dB, 0dB, -5dB & -10 dB SNR). In each condition, only one list of 10 bi-syllabic words out of six lists was presented and the lists were randomized when they were needed to be presented for the second time. The mode of presentation was randomly chosen. The subjects responded by repeating the words and the words repeated were noted down by the clinician.

Response Analyses

In each condition (quiet, -10 dB, -5 dB, 0 dB, & +5 dB SNR) the total number of words repeated correctly out of ten words was noted separately for A, V, and AV mode. The AV speech perception was quantified by calculating visual enhancement (VE, the benefit obtained from adding a visual signal to an auditory stimulus) and auditory enhancement (AE, the benefit obtained from adding an auditory signal to a visual-only stimulus) scores. The VE and AE were calculated according to the equations 1 and 2 respectively. The scores for VE and AE ranged between -1 to +1.

$$VE = \frac{AV - A}{1 - A} \quad (1)$$

$$AE = \frac{AV - V}{1 - V} \quad (2)$$

The data obtained from the participants were subjected to the statistical analysis to test the objectives of the study.

Results

The primary objective of the study was to analyze whether the SPIN and WM influence AV perception of speech. The independent variables were SPIN and WM, while the dependent variable was AV speech perception. SPSS (version 16) was used for the statistical analysis of data. Descriptive statistics, Repeated measures ANOVA, correlation, independent t-test and one-way ANOVA were the statistical tests used for the purpose.

Effect of Modality on Speech Perception Scores

To begin with, the speech perception scores of 66 participants in the three modalities were analyzed to obtain the mean and standard deviation scores. This was done separately for each stimulus condition (-10 dB, -5 dB, 0 dB, +5 dB SNR, & quiet). The mean (raw score out of 10) and standard deviation of speech perception scores in the 3 modalities at different signal-to-noise ratios are given in Table 1.

In general, the mean speech perception scores were higher in AV-mode compared to A and V-mode. The scores were least in the V-mode. This was true in quiet as well as in all the SNR conditions. The mean scores also decreased with decrease in SNR in the A and AV modality. In the visual modality, only the scores obtained in quiet was used for modality-wise comparison in all the stimulus conditions.

The difference in the mean scores was tested for the statistical significance on Repeated measures ANOVA. The results of Repeated measures ANOVA are given in Table 2

Table 1: Mean and standard deviation (S.D) speech perception scores in the visual, auditory, and auditory-visual modes in quiet, -10 dB, -5 dB, 0 dB, 5 dB SNR

Stimulus Condition	V-mode Mean (S.D)	A-mode Mean (S.D)	AV-mode Mean (S.D)
-10 dB SNR	CNT	1.12 (0.87)	3.65 (1.96)
-5 dB SNR	CNT	3.58 (1.61)	6.82 (1.66)
0 dB SNR	CNT	7.74 (1.25)	9.21 (0.94)
5 dB SNR	CNT	9.71 (0.70)	9.85 (0.40)
Quiet	1.09 (1.17)	9.95 (0.21)	10.0 (0.00)

Note: CNT- could not be tested as presentation of auditory stimulus was inevitable in these conditions.

Table 2: Results of repeated measures ANOVA comparing speech perception scores across the 3 modalities in the 5 stimulus conditions

Stimulus Condition	F	df(error)
-10 dB SNR	108.881*	2(130)
-5 dB SNR	363.733*	2(130)
0 dB SNR	1718.676*	2(130)
5 dB SNR	3216.879*	2(130)
Quiet	3762.289*	2(130)

*Note: * - $p < 0.01$*

The results of repeated measures ANOVA showed a significant ($p<0.01$) main effect of modality on speech perception scores in all the 5 stimulus conditions. As there was a significant main effect, the pair-wise comparison was tested on Bonniferroni post-hoc test. The results of post-hoc test showed significant difference across all three modalities (A-V, A-AV & V-AV) at 0, -5 and -10 dB SNR. However, in quiet and +5 dB SNR the significant difference was found only between A-V and V-AV modalities.

Effect of Stimulus Condition on AV Speech Perceptions

In order to test the effects of independent variables on AV speech perception, the scores were first converted as ***visual enhancement*** and ***auditory enhancement*** scores. The visual enhancement VE score was a quantity of the benefit obtained from adding a visual signal to an auditory stimulus score, whereas the auditory enhancement AE score was a quantity of the benefit obtained from adding an auditory signal to a visual-only stimulus. The raw scores obtained in the 3 modalities were used to compute VE and AE scores, separately for each individual. This was done for all the 5 stimulus conditions. The mean and standard deviation of VE and AE scores in the 5 stimulus conditions are shown in Figure 1.

The inspection of the Figure 1 shows that there are mean differences among the 5 conditions in both AE and VE scores. The mean VE scores were maximum at 0 dB SNR and decreased with either increase or decrease in SNR. The AE scores on the other hand increased with increase in SNR. The mean differences observed were statistically tested for the effect of condition on repeated measures ANOVA. The results showed a significant main effect of condition on VE [$F(4,260) = 46.907$, $p<0.001$] and AE scores [$F(4,260) = 491.085$, $p<0.001$]. Consequently, pair-wise comparison was tested using Bonferroni post-hoc test and the results showed a significant difference in the mean scores across all conditions in AE. However in VE, there was no significant difference between quiet and +5 dB SNR, and, 0 and +5 dB SNR.

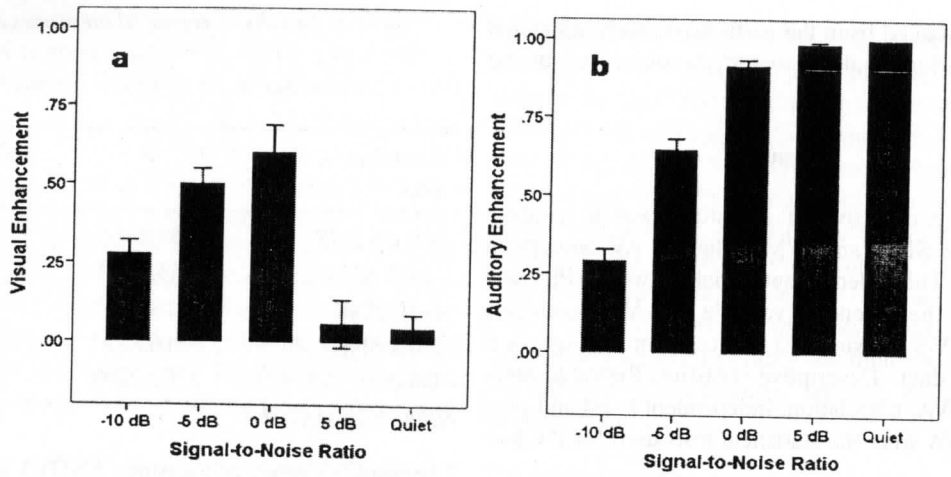

Figure 1: Mean and standard deviation of visual enhancement (a) and auditory enhancement (b) scores in quiet, -10 dB, -5 dB, 0 dB, and +5 dB SNR.

Relation between Speech Perception in Noise and AV Speech Perception

To understand the relation between SPIN and AV speech perception, the correlation was obtained for both VE and AE with SPIN scores in different stimulus conditions. The data were statistically tested for correlation between two variables on Pearson product moment correlation and the results are given in Table 3 and Table 4. The results of the correlation showed a significant low positive correlation at -10 dB and -5 dB SNR but a significant low negative correlation in quiet condition between mean scores of SPIN and VE. On the contrary, correlation of mean differences between SPIN and AE showed that there was a significant low positive correlation at -10 dB SNR but a significant low negative correlation in quiet and 5 dB SNR conditions. There was no significant correlation in other conditions for both VE and AE with SPIN scores.

The influence of SPIN on AV speech perception was further tested using a second method wherein the participants were divided into two groups based on their SPIN scores (good SPIN & poor SPIN groups) and compared with each other for their AV speech perception. The grouping was done based on the confidence interval (at 95%) determined for the SPIN scores of 66 participants of the study. The lower-boundary and the upper-boundary of the interval for the SPIN scores thus obtained were found to be 72% and 82% respectively. The participants with scores lesser than the lower boundary (72%) were grouped as 'poor SPIN' and those with scores higher than the upper boundary (82%) were grouped as 'good SPIN'. The participants with scores within the interval (72-82%) were not considered for any comparisons.

The VE and AE scores of the resultant two groups were then compared to investigate whether there was any difference in the AV speech perception between the two groups. It was hypothesized that the presence of significant difference in VE scores would indicate that SPIN influences AV speech perception. The mean and the standard deviation of the VE and AE scores of the two SPIN groups (good & poor), in each stimulus condition (quiet, -10 dB, -5 dB, 0dB & 5 dB SNR) are shown in Figure 2.

The mean VE scores were higher in the good SPIN group compared to poor SPIN group except in +5 dB SNR and quiet, wherein the case was reverse. In contrast, the mean AE scores of good SPIN group was

Table 3: The results of correlation test correlating visual enhancement (VE) with SPIN in quiet, -10 dB, -5 dB, 0dB, and +5 dB SNR

Stimulus Condition	SPIN and VE	
	Pearson Correlation	Significance
-10 dB SNR	0.496**	0.000
-5 dB SNR	0.253*	0.041
0 dB SNR	0.003	0.979
5 dB SNR	-0.166	0.184
Quiet	- 0.407**	0.001

*Note: * - $p < 0.01$*

Table 4: The results of correlation test correlating auditory enhancement (AE) with SPIN in quiet, -10 dB, -5 dB, 0dB, and +5 dB SNR

Stimulus Condition	SPIN and VE	
	Pearson Correlation	Significance
-10 dB SNR	0.512**	0.000
-5 dB SNR	0.335**	0.006
0 dB SNR	0.243*	0.049
5 dB SNR	0.207	0.095
Quiet	CNT	CNT

*Note: * - $p < 0.01$*

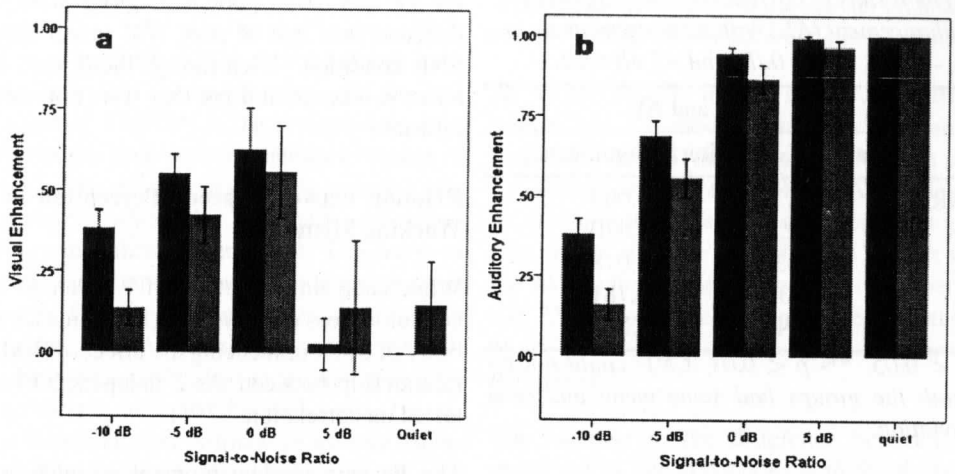

Figure 2: Mean and standard deviation of visual enhancement (a) and auditory enhancement (b) scores in quiet, -10 dB, -5 dB, 0 dB, and +5 dB SNR. Note: Visual and auditory enhancement scores would vary between -1 and 1 and WM scores ranged between 0 and 1.

Table 5: The results of the t-test comparing the good SPIN and poor SPIN groups for their visual enhancement (VE) and auditory enhancement (AE) scores in quiet, -10 dB, -5 dB, 0dB and +5 dB SNR

Stimulus Condition	VE(df= 54) t	AE (df= 54) t
-10 dB SNR	4.755**	4.967**
-5 dB SNR	2.042**	2.814**
0 dB SNR	0.557*	3.119**
5 dB SNR	-1.003	2.585*
Quiet	-2.475*	CNT

*Note: *- $p < 0.05$, **- $p < 0.01$, CNT- could not be tested as both the groups had same mean and no standard deviation.*

higher than that of poor SPIN group in all the stimulus conditions except in quiet, where the mean scores were equal. The mean scores of the two groups were statistically compared using independent samples t-test. The results of the t-test for VE and AE scores are given in Table 5.

The results showed that the scores for VE of good SPIN group were significantly higher in -10 dB and -5 dB SNR conditions and significantly lower in quiet, than the poor SPIN group. The mean differences in other stimulus conditions (0 dB & +5dB SNR) were not significant. On the contrary, the difference in the mean AE scores of the two groups was statistically significant in all stimulus conditions except in quiet

Relation between Working Memory (WM) and AV Speech Perception

The second independent variable of the study was working memory. Similar to effect of SPIN, the effect of working memory on AV speech perception was examined. The relation between WM and AV perception of speech across different stimulus conditions (quiet, -10 dB, -5 dB, 0dB, & +5 dB SNR) was examined by obtaining the correlation. The relationship between VE and AE with WM scores were statistically tested on Pearson product moment correlation test for different stimulus conditions (quiet, -10 dB, -5 dB, 0dB, and +5 dB SNR). The results of correlation test are given in Table 6 and Table 7.

From Table 6, we can witness clearly that, there was a significant moderate positive correlation at -10 dB SNR and a significant low negative correlation in quiet and 0 dB SNR conditions between WM and VE. Whereas, correlation between WM and AE showed a significant moderate positive correlation at -10 dB SNR and a significant low positive correlation at 0 dB SNR conditions as shown in Table 7.

Comparison between Poor and Good WM Groups

In a second method of testing the relationship between WM and AV speech perception, the participants were divided into good WM and poor WM groups to test whether there was any difference between the two groups. The confidence intervals (95%) for WM scores

Table 6: The results of correlation test comparing visual enhancement (VE) with working memory in quiet, -10 dB, -5 dB, 0dB, and +5 dB SNR

Stimulus Condition	WM and VE	
	Pearson Correlation	Significance
-10 dB SNR	0.572**	0.000
-5 dB SNR	0.192	0.178
0 dB SNR	-0.307*	0.028
5 dB SNR	-0.208	0.143
Quiet	- 0.349*	0.012

*Note: *- $p < 0.05$, **- $p < 0.01$, CNT- could not be tested as both the groups had same mean and zero standard deviation.*

Table 7: The results of correlation test comparing auditory enhancement (AE) with working memory in quiet, -10 dB, -5 dB, 0dB, and +5 dB SNR

Stimulus Condition	WM and AE	
	Pearson Correlation	Significance
-10 dB SNR	0.517**	0.000
-5 dB SNR	0.256	0.070
0 dB SNR	-0.330*	0.018
5 dB SNR	-0.040	0.781
Quiet	CNT	CNT

*Note: *- $p < 0.05$, **- $p < 0.01$, CNT- could not be tested as both the groups had same mean and zero standard deviation.*

were found out first. Then the participants with scores lesser than the lower boundary (70%) were grouped as 'poor WM' and those with scores higher than the upper boundary (76%) were grouped as 'good WM'. The participants with scores within the interval (70-76%) were excluded. The mean and standard deviation of VE and AE scores obtained for the 2 groups, across 5 conditions, are represented in Figure 3.

Figure 3 shows that the mean VE scores for the good WM group were better than those of poor WM at -10 dB, -5 dB and 0 dB SNR. But at +5 dB SNR and in quiet, poor WM group had better scores. Whereas, the mean AE scores of good WM group was higher than that of poor WM group in all the stimulus conditions except in quiet, where the mean scores were equal.

To test for significance in the mean differences between the two groups, independent samples t-test was administered. The results of the t-test for VE and AE scores across 5 conditions are given in Table 8. The results of the independent samples t-test showed that the scores for VE and AE of good WM group were significantly different than that of poor WM group only in -10 dB SNR condition. Even though there were mean differences in other conditions, they were not statistically significant.

Relation between Speech Perception in Noise and Working Memory

While analyzing the effect of WM on AV speech perception, across different SNR conditions, there is possibility of SPIN influencing the effect of WM. Hence, the relationship between the 2 independent variables was tested on correlation.

The Pearson product moment correlation was found out to test for correlation between SPIN and WM groups. The results showed that there was a significant ($p<0.001$) moderate positive correlation ($r=0.508$) between the two independent variables.

Table 8: The results of the t-test comparing the visual enhancement and auditory enhancement scores for good WM and poor WM groups in quiet, -10 dB, -5 dB, 0dB and +5 dB SNR

Condition	AE (df: 33)t	VE(df: 33)t
-10 dB SNR	2.380*	3.421**
-5 dB SNR	2.007	1.986
0 dB SNR	1.312	-0.986
+5 dB SNR	1.338	-0.337
Quiet	CNT	-1.817

*Note: *- $p < 0.05$, **- $p< 0.01$, CNT- could not be tested as both the groups had same mean and zero standard deviation.*

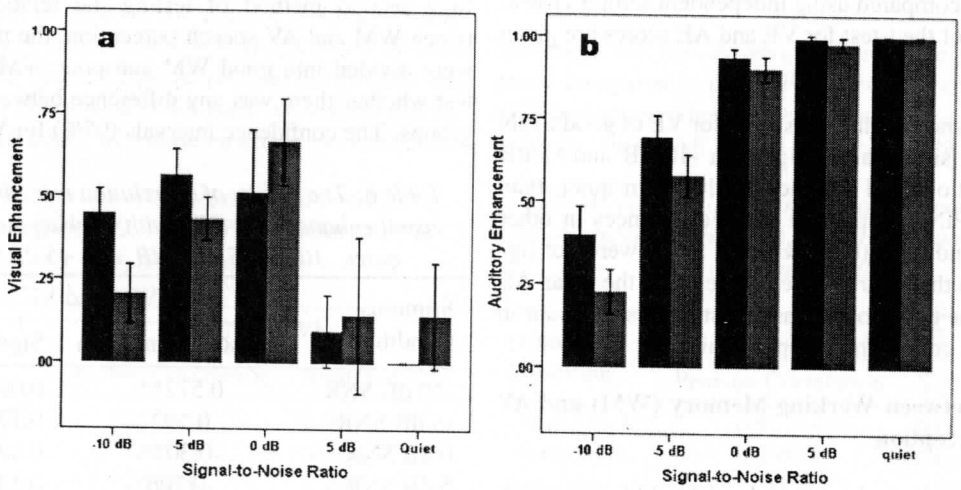

Figure 3: Mean, standard deviation of visual enhancement (Left panel) and auditory enhancement (Right panel) scores for good working memory (Dark bars) and poor working memory (Light bars) groups in the 5 different stimulus conditions. Note: Visual and auditory enhancement scores would vary between -1 and 1 and WM scores ranged between 0 and 1.

Discussion

The perception of speech could be through either auditory (A-mode) or visual (V-mode) mode. Of the two modalities, auditory modality is the primary one as it provides information on place, manner and voicing of speech sounds. The dependency on the visual mode is only in instances where the information provided through auditory modality is insufficient. This happens due to reduction either in external redundancy or internal redundancy. The negative influences of reduction in both external and internal redundancy on speech perception have been well established in the literature. Although it is universally accepted that audio-visual mode (AV-mode) is beneficial over A-mode in adverse listening conditions, the underlying mechanisms of AV perception are not clearly understood.

The present study was an attempt to probe into the underlying mechanisms of AV perception in sensory and cognitive domain. In the sensory domain auditory closure was taken as an independent variable, while in cognitive domain working memory was the independent variable. The present experiment to analyze the effects of these two independent variables on AV perception, on 66 participants, showed some interesting findings.

Effect of Modality on Speech Perception

Evidence from previous research explains that speech intelligibility improves when listeners receive information from both auditory and visual modes in instances where the auditory cues are degraded by reducing external redundancy (Sumby & Pollack, 1954; Anderson, 2006; Huffman, 2007). The present findings are in consonance with these earlier reports. Overall performance of the participants was best in the AV-mode compared to auditory-alone or visual-alone modes. The performance was least in the V-mode. Auditory mode provides cues pertaining to the place, manner and voicing of a speech sound while in the visual mode one gets cues of only place of articulation, that too not completely (Anderson, 2006). Hence, it is logical to expect better performance in the A-mode compared to V-mode. Further, as the redundancy of the cues is more in the bimodal presentation, AV-mode showed better performance than that in A-mode or V-mode, which were unimodal presentations.

The results showed that the performance in the A-mode was comparable to that in AV-mode in quiet and at +5 dB SNR. This means that the necessary cues for an ideal performance are not cut in quiet and +5 dB SNR. However, the performance reduced with further reduction in the SNR both in A and AV-modes. As the reduction of redundancy was only in the auditory signal without distorting the visual signal, the reduction in the scores of A and AV-modes can be attributed solely to the reduction in the SNR of the auditory signal. The Figure 4 represents the deterioration in the speech identification in A and AV-modes across different stimulus conditions.

As we can observe in the Figure 4, the performance of both A and AV- modes is same in quiet and decreases with decreasing SNR. But the decrease in performance for A-mode was more than that of AV-mode which is evident by the steepness of the curves. That is, the steepness of the A-mode (blue line) is greater than that of the AV-mode (red line). The difference between the A and AV-mode is due to complementary cues provided by the addition of visual cues, which however functions nonlinearly.

The increase in the difference between the performances in A and AV-modes with the decrease in SNR evidences, greater dependency on visual cues at lower SNRs. The finding is in agreement with the earlier studies (Sumby & Pollack, 1954; O'Neil, 1954; Neely, 1956; Erber, 1969; Grant & Seitz, 2000; Rudmann, Mc Carley & Kramer, 2003; Bernstein, Auer & Takayanagi, 2004; Ross, Saint-Amour, Leavitt, Javitt & Foxe, 2007) which showed larger benefits of bimodal presentations when the auditory signal was degraded. In the study by Sumby and Pollack (1954), it was demonstrated that the addition of visual cues improved speech perception by an amount equivalent to a 5 to 18 dB increase in the SNR which accounts for improvement of up to 60% in word recognition. They also demonstrated that with increase in difficulty of auditory-only perception, the benefits obtained by combining the auditory and visual speech information also increased. Although the present results showed a similar trend, the quantity of improvement at the worst SNR (-10 dB SNR) was only about 25%. The difference among the studies in the extent of improvement may be attributed to the different test materials used.

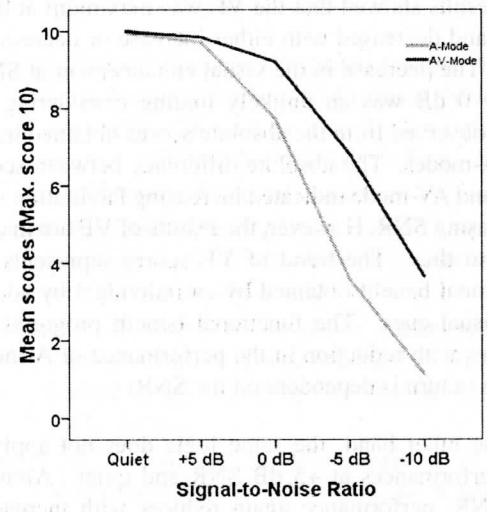

Figure 4: Mean identification scores in auditory mode (lower line) and audio-visual mode (upper line) in quiet, -10 dB, -5 dB, 0 dB, and +5 dB SNR conditions.

Effect of Stimulus Condition on AV Speech Perception

In the previous section, it was learnt that AV-mode was better than A-mode only at lower SNRs. That is, although the redundancy of visual cues was not varied, it nonlinearly facilitated speech perception. Hence the absolute difference between the scores of AV and A-modes would not have given a clear picture of the VE. For example, the importance of 10% facilitation by addition of V-mode will be different for 3 individuals who score 20%, 50% and 90% in the A-mode. An absolute difference between A and AV-mode would be 10% in all these cases, but logically the weightage of relative enhancement should be maximum for one who had 90% score in the A-mode followed by one with 50% and 20% score. This was achieved by calculating visual enhancement scores using the following formula in equation 3.

For the above example, the visual enhancement will be 0.12, 0.2, and 1 for the individuals with 10%, 50% and 90% score in the A-mode, respectively. With the similar logic the AE scores were derived from the performances in AV and V-mode using the equation 4. The AE score is a quantity of the benefit obtained from adding an auditory signal to a visual-only stimulus.

$$VE = \frac{AV - A}{1 - A} \quad (3)$$

$$AE = \frac{AV - V}{1 - V} \quad (4)$$

Effect of Stimulus Condition on Visual Enhancement (VE)

The results showed that the VE was maximum at 0 dB SNR and decreased with either increase or decrease in SNR. The decrease in the visual enhancement at SNRs below 0 dB was an unlikely finding considering the trend observed from the absolute scores obtained in AV and A-modes. The absolute difference between scores of A and AV-mode indicated increasing facilitation with decreasing SNR. However, the results of VE are in contrary to this. The trend of VE scores represents the functional benefit obtained by an individual by adding the visual cues. The functional benefit progressively reduces with reduction in the performance of A-mode, which in turn is dependent on the SNR.

On the other hand, the same logic does not apply to the performances at +5 dB SNR and quiet. Above 0 dB SNR, performance again reduces with increasing SNR. This is attributed to the ceiling effect. Because the scores in the A-mode were already 100%, the resultant VE scores would become 0. The number of such individuals would increase with increase in SNR, resulting in progressively decreasing mean VE performance. Based on these findings, it may be inferred that any study intending to investigate visual enhancement shall take SNRs of 0 dB or lesser and not take the higher SNRs, in order to get a true picture of VE.

Table 9: The Pearson correlation co-efficient correlating auditory enhancement scores and auditory mode scores in quiet, -10 dB, -5 dB, 0 dB, and +5 dB SNR conditions

Stimulus conditions	Correlation co-efficient
-10 dB SNR	0.352**
-5 dB SNR	0.423**
0 dB SNR	0.536**
5 dB SNR	0.488**
Quiet	CNT

*Note: **- $p<0.001$, CNT- could not be tested as all the participants had the same AE score of one (1).*

Effect of Stimulus Condition on Auditory Enhancement (AE)

The AE was maximum in quiet (approximately 40 dB SNR) and decreased with decreasing SNR. This trend is logical as the enhancement shall increase with increasing redundancy of the auditory cues. To strengthen the notion, the correlation between A scores and AE scores was done in different SNRs. It was expected to get high positive correlation between the two variables in all the stimulus conditions if the stated logic was true. Table 9 gives the Pearson correlation co-efficients in the 5 stimulus conditions.

The results showed that the correlation was maximum at 0 dB SNR and reduced at other conditions. Hence the above stated logic was defeated. The trend in Table 9 in turn shows the influence of VE as a primary variable. In the formula used for the calculation of AE, AV and V scores are the two parameters considered. As the V scores remained constant in all the stimulus conditions, the differences in the AE had to be because of AV which in turn was related to VE.

As the trend of VE (Figure 1) across conditions is same as that of trend of correlation observed in Table 9, it can be inferred that it is this influence of VE over AV scores that varied AE across stimulus conditions. This was supported by the results of correlation of VE scores and AE scores at 0, -5 and -10 dB SNR. The data of +5 dB SNR and quiet were not considered as VE in these conditions was erroneous due to ceiling effect. The results of correlation showed a moderate positive co-efficient ($r=0.632$, $p<0.001$) between AE and VE scores.

Effect of SPIN on AV Speech Perception

From the previous two sections, it is clear that AV speech perception is influenced by external redundancy

of the auditory signal. Earlier studies have shown negative effects of advancing age on AV speech perception (Shoop & Binnie, 1979; Middelweerd & Plomp, 1987; CHABA, 1988; Walden, Busacco, & Montgomery, 1993; Cienkowski, 1999; Cienkowski & Carney, 2002; 2004). This age related decline could be either partially or completely due to reduced internal redundancy secondary to central auditory processing deficits. Hence it was of interest to study the relationship between internal redundancy and speech perception in noise. The relationship was tested using two methods; one, by correlating SPIN and AV speech perception. The other, by comparing the AV speech perception of good and poor SPIN groups.

The results of the correlation showed that below 0 dB SNR both VE and AE correlated with the SPIN scores. That is, VE and AE increases with increase in SPIN scores. Additionally, AE positively correlated with SPIN at 0 dB. A negative correlation between VE and SPIN is erroneous as most of the individuals had 0 VE at +5 dB SNR and quiet. Further, the comparison of AV speech perception among the good and poor SPIN groups also showed that the SPIN scores have a positive effect on speech perception through AV-mode.

From these findings it can be concluded that internal redundancy controlled by central auditory mechanisms, as evident through SPIN, is directly related to the AV speech perception. Hence, one should take SPIN into consideration while commenting on the benefit provided by AV speech perception.

Effect of Working Memory (WM) on AV Speech Perception

It has been explained that speech understanding needs a constant encoding of information into and out of the working memory and manipulation of the information stored in memory (Pichora-Fuller, Schreider & Daneman, 1995).

Similar to SPIN, the relationship between working memory and AV speech perception was analyzed by correlating the two parameters and by comparing good and poor working memory groups. This was done to decipher the underlying mechanisms of speech perception under the cognitive domain. Considering that the integration of information from two different modalities is a complex higher level task, cognitive factors such as working memory were expected to influence the AV speech perception.

The results of both correlation and group comparison showed evidence for working memory influencing AV speech perception in noise. The influence was particularly significant at lower SNRs and the relationship was direct. That is, AV speech perception was better in individuals who had better working memory.

The results of the present study are in consonance with the earlier report (Pichora-Fuller, Schreider & Daneman, 1995) which stated that the loss of working memory hinders speech communication and may contribute to the age-related declines in the comprehension of speech which may be evident even in favorable listening conditions.

From these findings it can be concluded that working memory, a cognitive factor, has a direct relationship with the AV speech perception and hence demands attention while interpreting on AV speech perception.

Relationship between SPIN and WM

As both working memory and speech perception in noise were influencing AV speech perception in a similar way, it was of interest to know how these two independent variables related to each other. The results of correlation showed that the speech perception in noise was better in individuals with good working memory and vice versa. However, it was only a moderate correlation which shows that they are not controlled by same physiological mechanism.

Conclusions

The present study has important implications in rehabilitative audiology. In audiological clinics, AV-mode of speech perception is recommended for individuals with poor speech identification scores. The present findings showed that the WM and SPIN directly influence AV speech perception. Hence, one should consider WM and SPIN, assess for them, estimate the probable benefit with AV speech perception and accordingly recommend the same.

References

American National Standards Institute. (1991). *American National Standard Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms*. ANSI S3.1- (1991). New York: American National Standards Institute.

Anderson, E. (2006). *Speech Perception with Degraded Auditory Cues*. Undergraduate honors thesis, The Ohio State University.

Bernstein, L. E., Auer, E. T., Jr., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication, 44*(1), 5-18.

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.

CHABA, Committee on Hearing and Bioacoustics, Working Group on Speech Understanding and Aging. (1988). Speech understanding and aging. *Journal of the Acoustical Society of America, 83,* 859?895.

Cienkowski, K. M., & Carney, A. E. (2002). Auditory-visual speech perception and aging. *Ear and Hearing, 23,* 439?449.

Cienkowski, K. M., & Carney, A. E. (2004). The Integration of Auditory-Visual Information for Speech in Older Adults. *Journal of Speech-Language Pathology and Audiology, 28*(4), 169-172.

Cohen, G. (1979). Language comprehension in old age. *Cognitive Psychology, 11,* 412-429.

Cohen, G. (1981). Internal reasoning in old age. *Cognition, 9,* 59-72.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R.W. (2005). Working memory span tasks: A methodological review and user?s guide. *Psychonomic Bulletin and Review, 12*(5), 769-786.

Craik, F. I. M., & Byrd, M. (1982). Aging and cognitive deficits: The role of attentional resources. In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes*, (pp. 191-211), New York: Plenum.

Erber, N. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423?425.

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, 4*(4), 2438-2449.

Grant, K. W., & Seitz, P. F. (2000). The recognition of isolated words and words in sentences: Individual variability in the use of sentence context. *Journal of the Acoustical Society of America, 107,* 1000?1011.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103,* 2677?2690.

Hirsh, I. J., Reynolds, E. G., & Joseph, M. (1954). Intelligibility of different speech materials. *Journal of the Acoustical Society of America, 26,* 530?537.

Huffman, C. (2007). *The Role of Auditory Information in Audiovisual Speech Integration.* Undergraduate honors thesis, The Ohio State University.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working-memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189-217.

Li, K. Z. H., & Lindenberger, U. (2002). Relations between aging sensory/ sensorimotor and cognitive functions. *Neuroscience and Biobehavioral Reviews, 26,* 777?783.

MacLeod, A., & Summerfield, Q. (1987). "Quantifying the contribution of vision to speech perception in noise". *British Journal of Audiology, 21,* 131-141.

Martin, J. S. & Jerger, J. F. (2005). Some effects of aging on central auditory processing. *Journal of Rehabilitation Research and Development, 42,* 25?44.

Middelweerd, M., & Plomp, R. (1984). The effect of speech reading on the speech reception threshold of sentences in noise. *Journal of the Acoustical Society of America, 82,* 2145?2147.

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

Neely, K. K. (1956). Effects of visual factors on intelligibility of speech. *Journal of the Acoustical Society of America, 28,* 1276-1277.

O'Neill, J. J. (1954). Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders, 19,* 429-439.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17,* 299-320.

Park, D., Smith, A., Lautenschlager, G., & Earles, J. (1996). Mediators of long-term memory performance across the life span. *Psychology and Aging, 11,* 621-637.

Pichora-Fuller, M. F., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America, 97,* 593-608.

Plude, D., & Hoyer, W. (1985). Attention and performance: Identifying and localizing age deficits. In N. Charness (Ed.), *Aging and Human Performance*, (pp. 47?99). Chichester, England: Wiley.

Ramakrishna, B. S., Nair, K. K., Chiplunkar, V. N., Ramachandran, V., & Subramanian, R. (1962). *Some aspects of the relative efficiencies of Indian languages.* Ranchi, India: Catholic press.

Ross, L. A., Saint-Amour, D., Leavitt, V., Javitt, D. C. & Foxe J. J. (2007). Do you see what I?m saying? Optimal Visual Enhancement of Speech

Comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147-53.

Rudmann, D. S., McCarley, J. S., & Kramer, A. F. (2003). Bimodal display augmentation for improved speech comprehension. *Human Factors, 45,* 329-336.

Schneider, B. A., Daneman, M., & Pichora-Fuller, M. K. (2002). Listening in aging adults: from discourse comprehension to psychoacoustics. *Canadian Journal of Experimental Psychology, 56*(3), 139?52.

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.

Shoop, C., & Binnie, C. A. (1979). The effects of age upon the visual perception of speech. *Scandinavian Audiology, 8,* 3?8.

Sommers, M. S. (1997). Speech perception in older adults: The importance of speech-specific cognitive abilities. *Geriatric Bioscience, 45,* 633-637.

Spehar, B., Tye-Murray, N., & Sommers, M. S. (2008). Intra- versus intermodal integration in young and older adults. *Journal of the Acoustical Society of America, 123*(5), 2858?2866.

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

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

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lip-reading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*(5), 656-668.

Walden, B. E., Busacco, D. A., & Montgomery, A. A. (1993). Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons. *Journal of Speech and Hearing Research, 36,* 431?436.

Yathiraj, A., and Vijayalakshmi, C. S. (2005). *Phonemically balanced word list in Kannada.* A test developed at Department of Audiology, AIISH, Mysore.