

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 23 (1999)
Indiana University

**Effects of Multimodal Presentation and Lexical Density on
Immediate Memory Span for Spoken Words¹**

Lorin Lachs, Winston D. Goh,² and David B. Pisoni³

*Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by a grant from the NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University Bloomington. Special thanks go to Luis Hernández, Tyler Emley and Patrick Kelley for their invaluable assistance during the completion of this study.

² Also, Department of Social Work & Psychology, National University of Singapore.

³ Also DeVault Otologic Research Laboratory, Department of Otolaryngology-Head and Neck Surgery, Indiana University School of Medicine, Indianapolis, IN.

Effects of Multimodal Presentation and Lexical Density on Immediate Memory Span for Spoken Words

Abstract. Working memory span for spoken words was measured using a non-repeated sampling procedure. Stimuli were presented in audio-only (AO) or audiovisual (AV) conditions. In the AO condition, participants were presented with only the audio track of a talker's voice. In the AV condition, participants simultaneously heard and saw a talker speaking a word. The stimulus words differed in their neighborhood density, neighborhood frequency and word frequency. The results show that audiovisual presentation enhances memory span for words from sparse lexical neighborhoods, but has no effect on words from dense lexical neighborhoods. The findings are discussed with respect to the effects of perceptual distinctiveness on working memory.

Visual information about the articulation of spoken words has been shown to have large effects on speech perception. Conflicting information about speech in the visual and auditory modalities can lead to the illusory perception of speech sounds not included in either modality alone (commonly referred to as the “McGurk effect;” McGurk & MacDonald, 1976). The McGurk illusion is commonly elicited by presenting the syllable “ba” in the audio track and the syllable “ga” in the video track of a cross-dubbed movie clip. In the overwhelming majority of cases, participants report that the syllable presented was “da”. The very existence of this phenomenon demonstrates that some sort of integration of information in the auditory and visual domains occurs during the process of speech perception. The precise nature of this integration, however, remains in question.

The McGurk effect has generated a great deal of interest in the idea of audiovisual integration. Much of the research on the integration issue has concerned the effects of conflicting audio and visual cues on segmental identification. This work has produced a large body of knowledge concerning the ways in which acoustic and visual information influence each other. For example, it has been shown that the vowel context in which a segment is presented affects the degree to which visual information influences the McGurk illusion (Green, 1996; Green, Kuhl, & Meltzoff, 1988), that inverted faces reduce the effects of McGurk integration (Massaro & Cohen, 1996; Jordan & Bevan, 1997) and that temporal asynchrony between the audio and visual tracks has no effect on McGurk integration (Massaro, Cohen, & Smeele, 1996; Munhall, Gribble, Sacco, & Ward, 1996; Smeele, Sittig, & van Heuven, 1992). Even separating the spatial location of the auditory and visual aspects of the stimulus makes little difference on the extent of the illusion (Bertelson, Vroomen, Wiegendaal, & de Gelder, 1993; Fisher & Pylyshyn, 1994; Jones & Munhall, 1997).

Other studies have shown that the auditory and visual information in a bimodal stimulus is integrated during processing, such that the information in each channel is evaluated relative to the information present in the other channel. For example, Green and Kuhl (1989) showed that the perceived boundary in voice onset time along an /ibi/ to /ipi/ continuum was dependent on whether or not there was concurrent visual information. Presentation of the visual portion of the word caused the VOT boundary to shift as though it were along an /idi/ to /iti/ continuum, precisely the continuum specified by the McGurk effect. This finding demonstrated that low-level auditory cues are evaluated in the context of the *combined* audiovisual stimulus, suggesting that integration of the information in the various modalities happens before classification. A related study showed interference in the processing of one sensory modality based on variation in the other (Green & Kuhl, 1991), providing further support for the notion that the information in both channels is processed in tandem.

Most of this knowledge about audiovisual integration, however, has been accumulated using tasks that measure segmental identification. With the exception of Dekle, Fowler and Funnel's (1992) study of audiovisual integration in real words, the stimulus materials used in all of the McGurk effect experiments have been nonsense syllables. In order to fully understand how the phenomena associated with audiovisual speech stimuli are related to speech perception in general, it is necessary to expand investigations into more naturalistic settings. The work of Sumbly and Pollack (1954) showed that audiovisual speech effects are not confined to illusory percepts and segmental contexts. On the contrary, their study demonstrated that the intelligibility of words can be enhanced by as much as +15 dB in noisy environments, a gain that surpasses even the best hearing aid devices.

The use of audiovisual information in spoken word recognition is a relatively unexplored area, but some work has already been carried out. In spoken word recognition, it is hypothesized that bottom-up perceptual processes interact with multiple, higher sources of information during processing (e.g. Luce & Pisoni, 1998; McClelland & Elman, 1986). An idea central to the concept of spoken word recognition is the mental lexicon, an entity in long term memory that stores information about all the words ever encountered by a specific individual (see Lively, Pisoni, & Goldinger, 1994). The structure of lexical representations in long-term memory has profound effects on the process of spoken word recognition. For example, the frequency of occurrence in the language of a target word is directly related to its intelligibility, while the number of words similar to the target word ("neighborhood density") is inversely related to intelligibility (Luce & Pisoni, 1998). Thus, words are recognized in the context of other phonetically similar words in the lexicon.

Some evidence exists to support the notion that long-term memory representations for spoken words exist in a multimodal form. Lachs and Pisoni (submitted) found that the repetition of studied, dynamic visual information of a talker's face during test facilitated performance on a recognition memory task. The results were taken to imply that information about the dynamic aspects of visual articulation is stored in lexical memory for spoken words.

If audiovisual information is represented in the long-term memory representations of spoken words, and the structure of those representations is important during the process of spoken word recognition, then lexical structure should have an effect on audiovisual speech perception. Indeed, Brancazio (1999) found that varying the lexical properties of the auditory and visual parts of McGurk stimuli had effects on the extent to which those stimuli were susceptible to illusory percepts. Auer and Bernstein (1997) have investigated the computational properties of the mental lexicon when words are represented using visemes. Because visemes are sets of phonetic segments considered indistinguishable in visual-only environments, transcribing the lexicon in this way is presumed to be equivalent to collapsing across perceptual dimensions that are irrelevant to the process of spoken word recognition while speechreading. The structure left in these viseme-transcribed lexicons remains useful because many of the words in the lexicon remain unique, even when the number of segments is reduced by 75%. Although it has not yet been explored experimentally, this structure could be a source of additional information in audiovisual situations.

Cognitive process that may interact with speech perception processes is working memory. It is well known that serial recall of verbal material appears to be affected by the phonological properties of the material to be remembered, as well as other experimental manipulations on the acoustic environment of the experiment (cf. Baddeley, 1998). Phenomena such as the *phonological similarity effect*, (Baddeley, 1966; Conrad, 1964), the *word length effect* (Baddeley, Thomson, & Buchanan, 1975), the *unattended speech effect* (Salame & Baddeley, 1982), and the *articulatory suppression effect* (Baddeley et al., 1975) suggest that the representation of verbal material in working memory is in a phonological or speech-based code. Baddeley and Hitch's (1974) well-known model of working memory specifically posits a

phonological loop to handle the encoding and rehearsal of verbal material. The phonological loop comprises 2 subsystems, a phonological store and an articulatory rehearsal loop. The passive phonological store is a temporary, limited capacity buffer that can hold auditory memory traces. These traces begin to decay after about 2 seconds if they are not rehearsed by the articulatory loop, which serves to maintain the information in the store. However, the precise mechanisms underlying the phonological loop and the nature of the relationship to the phenomena mentioned above is still much debated (e.g., Bavelier & Potter, 1992; Cowan, Wood, Nugent, & Treisman, 1997; Nairne, Neath, & Serra, 1997).

Recent investigations have also indicated that the structure of LTM representation affects working memory (e.g., Bourassa & Besner, 1994; Goh & Pisoni, 1998; Hulme, Maughan, & Brown, 1991). These studies have begun to describe in detail the interaction between long-term properties of the lexicon and short-term memory processes (e.g., Gathercole, Frankish, Pickering, & Peaker, 1999; Hulme, Roodenrys, Schweickert, Brown, Martin, & Stuart, 1997; Schweickert, 1993). In general, this research has demonstrated that the information stored in the mental lexicon can facilitate the recall of verbal information in working memory. Memory span for nonwords or words of an unfamiliar language is lower than memory span for familiar words (Hulme et al., 1991).

Memory span is also affected by the semantic attributes of words (Bourassa & Besner, 1994), phonotactic properties (Gathercole et al., 1999), and word frequency (Hulme et al., 1997). Lexical status, phonotactic information, and word frequency are all assumed to be stored with the memory trace of the word in the mental lexicon. Using the framework of Schweickert's (1993) multinomial processing tree model of immediate recall, one can argue that such properties of the long-term traces of words can be used to aid in the reintegration of decayed traces in working memory, and thus facilitate their subsequent retrieval and recall. For example, items that do not have a lexical representation, such as nonwords, would not have the additional advantage of permanent traces to aid in the recall process, which thus translates to a lower memory span.

Goh and Pisoni (1998) recently investigated the effects of phonological neighborhoods on immediate memory span. One way to represent and quantify the organization of phonological information in the mental lexicon is to consider the notion of lexical neighborhoods defined by their phonological similarity (Landuaer & Streeter, 1973; Treisman, 1878; Luce, 1986; Luce & Pisoni, 1998). A similarity neighborhood is defined by the number of words that can be obtained by a single substitution, addition, or deletion of a phoneme. Using this metric, the neighbors of *cat* would include *hat*, *cut*, *cap*, *scat* and *at*, among others. Two factors have been used to characterize the structure of a lexical neighborhood (Luce & Pisoni, 1998). *Neighborhood density* refers to the number of words in a similarity neighborhood, whereas *neighborhood frequency* refers to the average frequency of occurrence of the words in a given similarity neighborhood. A third property, *word frequency*, refers to an individual word's frequency of occurrence in the language.

Based on these three variables, words can be classified into two dichotomous categories based on ease of recognition. *Easy words* are words that are higher in frequency relative to their neighbors, and reside in low density and low frequency neighborhoods. *Hard words* are words that are lower in frequency relative to their neighbors, and reside in high density and high frequency neighborhoods. Previous work has shown that these lexical properties influence spoken word recognition. Lexically easy words are recognized faster and more accurately than lexically hard words (Luce, 1986; Luce & Pisoni, 1998).

Goldinger, Pisoni, and Logan (1991) found lexical effects in serial recall of 10-word lists—easy words were recalled better than hard words. Goh and Pisoni (1998) extended the Goldinger et al. (1991) findings by replicating the lexical effects using an immediate memory span task. Furthermore, they

showed that the difference in immediate recall between easy and hard word spans was not related to participants' working memory capacity, as measured by the traditional digit-span task. This finding indicated that the locus of the effect was most likely from the long-term properties of the lexicon, and not from individual differences in working memory capacity. More interestingly, Goh and Pisoni found that the lexical effect emerges only when a non-repeated sampling procedure was used, i.e., each word was used only once on a particular list and trial, and never repeated in subsequent lists or trials. No lexical effect was observed when a repeated sampling procedure was used, in which words were resampled again across trials. Again, this pattern suggested a contribution from long-term storage, because repeated sampling would keep the working memory traces active and thus severely attenuate the effects from the permanent store. Collectively, these findings suggest that phonological neighborhood properties and their ramifications on the representational distinctiveness of a word's long-term memory trace affect the serial recall performance of these words. Lexically easy words, which are perceptually more distinctive, will have a greater probability of reintegration from partially decayed traces in working memory compared to lexically hard words, which are less distinctive (cf. Schweickert, 1993).

If the structure of spoken word representations in memory has an effect on immediate recall, and long-term memory representations for spoken words contain multimodal information, then audiovisual presentation of stimuli should also produce effects on working memory. The present study manipulated the lexical characteristics of the stimulus words used in the memory task, and examined memory capacity in auditory-only (AO) and audiovisual (AV) settings. Because lexical similarity is defined across inherently multimodal dimensions (phonemes), 'extra' visual information will not significantly add to the distinctiveness of target candidates in memory, if those targets come from an already dense neighborhood (hard words). If targets are easy, however, then the extra visual information will make differences between target candidates even *more* distinct, thereby improving memory capacity for AV presentation over AO presentation.

Method

Design

The present experiment measured memory span in a 2 x 2 repeated-measures factorial design. The two levels of the "Lexical Class" factor used lexically "easy" and lexically "hard" words. The two levels of the "Presentation Mode" factor presented stimuli in "Audiovisual (AV)" and "Audio-Only (AO)" conditions. All factors were administered within subjects. The order in which conditions were presented to participants was counterbalanced using a balanced Latin-square design.

In order to assess and control for individual differences in short-term memory, a control condition was administered first to all participants. This condition tested memory span using spoken digits as the items for study. This was presented using auditory information only.

Participants

Participants were 34 undergraduate students enrolled in an introductory psychology course who received partial credit for their participation. One participant's data was eliminated from the final analysis for failure to follow the instructions provided. All of the participants were native speakers of English who reported no hearing or speech disorders at the time of testing. In addition, all participants reported having normal or corrected-to-normal vision.

Materials

Two Apple Macintosh computers, each equipped with a 17" Sony Trinitron Monitor (0.26 dot pitch) and its own video processing board were used to present the stimuli to subjects. The video processing boards were each capable of handling clips digitized at 30fps with a size of 640 x 480 and 24-bit resolution. The auditory portion of each stimulus was presented over Beyer Dynamic DT100 headphones.

The word stimuli were a subset of the tokens contained in the Hoosier Audiovisual Multitalker Database [HAVMD] (Lachs & Hernández, 1998; Sheffert, Lachs, & Hernández, 1997). The video portion of the HAVMD tokens is digitized at 30 fps with 24-bit resolution with 640 x 480 pixel size. The audio portion of the HAVMD tokens is digitized at 22 kHz with 16-bit resolution. The digitized movie clips of one talker (F1) uttering 264 isolated words were used during this study. The talker F1 was chosen because previous intelligibility studies on the HAVMD showed that she is the most intelligible talker in the database in audiovisual (AV) and audio-only (AO) conditions. Half of the words chosen were classified as lexically "Easy" words, and the other half were classified as lexically "Hard" words.

The spoken digits (0 to 9) were tokens obtained from the Texas Instruments 46-Word (TI46) Speaker-Dependent Isolated Word Corpus (Texas Instruments, 1991). The original tokens on the CD-ROM were in 11,025 Hz 16-bit PCM-Motorola formatted files. These files were edited using a digital waveform editor to remove the silent portions from either side of the token and saved as 12,500 Hz, 16-bit, mono WAV files. The overall RMS (root-mean-square) amplitude levels for each digit token were digitally equated with the word tokens to ensure equal presentation levels. The tokens recorded by a female speaker were used for the digit-span task. This speaker was chosen because her mean intelligibility was 100%, as determined by a token identification task with ten volunteer participants who were not a part of the current study.

Procedure

In order to measure memory span, lists of increasing length were presented to participants for immediate recall. Two lists of each length were presented in each condition. The shortest length used was three items, while the longest length used was eight items. All participants received all the list lengths, in ascending order, in all the conditions. In all, 66 stimulus words were presented in each condition.

One response booklet for each condition was provided so that participants could write down the items in each list. Participants were instructed not to start writing down their answers until the end of each list. The response sheet for each list contained eight blank spaces. The response for each list was recorded on separate pages in the response booklet. Participants were told that they had to recall the items in the order they were presented. If they were unable to recall a particular word, they were instructed to leave a blank space in the position where it occurred. The next list was not presented until the participants indicated they were ready.

Prior to the presentation of the first stimulus, participants were instructed in the procedure and encouraged to ask questions if they were unclear on their task. Stimuli were randomly presented by computer.

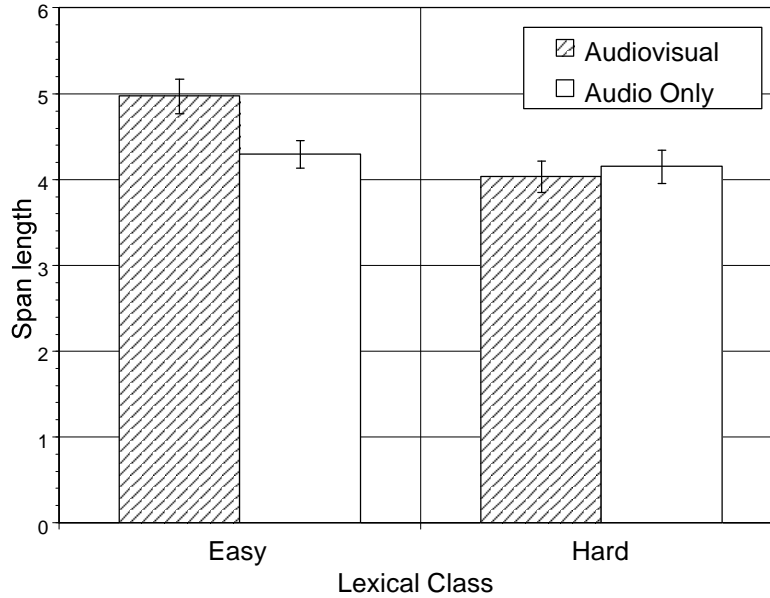


Figure 1. Average H/L score as a function of Lexical Class and Presentation Mode. Error bars denote standard errors.

Scoring

Each response list was scored independently. An item was scored as correct if, and only if, the item and serial order were both reproduced correctly. Phonetic transcriptions, homophones and obvious spelling errors (e.g., “cheif” for the target “chief”) were counted as correct.

Two measures of memory span were computed. The “Highest/Longest” (H/L) score was the longest length at which all items were scored correct *on both lists of that length*. The “Strict” (S) score was computed by taking one less than the minimum list length and adding 0.5 for every list that was completely correct. These two scoring methods are referred to as “length-based” measures, since they score in terms of list-length correct. These measures have been traditionally used for indexing working memory capacity (see La Pointe & Engle, 1990).

Results

Figure 1 shows the “Highest/Longest” (H/L) scores in each condition, averaged across participants. The left panel displays performance on lists that used “Easy” words, and the right set of columns displays performance on lists that used “Hard” words. Within each set of columns, the white bar represents performance on lists presented in the AO condition, while the striped bar represents performance on lists presented in the AV condition.

A 2 (Lexical Class) x 2 (Presentation Mode) repeated measures analysis of variance (ANOVA) on the H scores revealed a main effect of Lexical Class, $F(1, 33) = 15.32$, $MSE = 0.66$, $p < .001$, $h^2 = 0.32$. Memory span for lexically easy words ($M = 4.63$, $SD = 0.15$) was higher than memory span for lexically hard words ($M = 4.09$, $SD = 0.14$). There was no main effect of Presentation Mode.

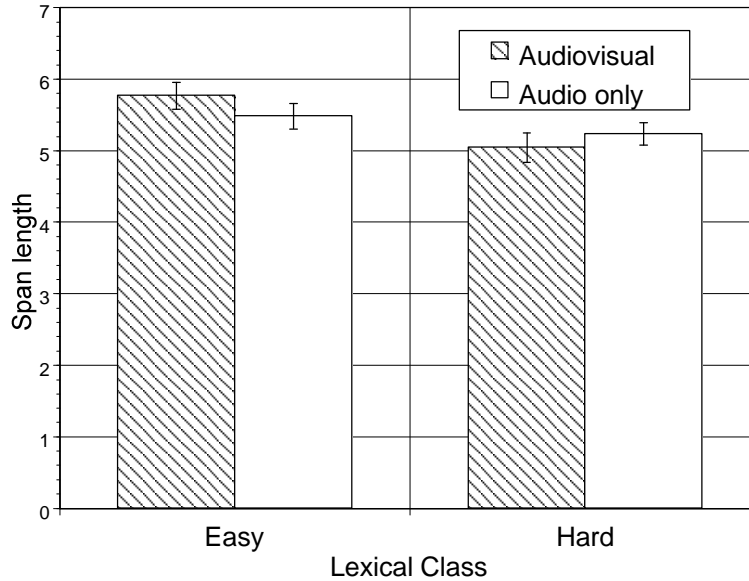


Figure 2. Average Strict (S) score as a function of Lexical Class and Presentation Mode. Error bars denote standard errors.

We also observed a significant interaction between Presentation Mode and Lexical Class, $F(1,33) = 5.30$, $MSE = 1.01$, $p < .05$, $h^2 = 0.14$. Simple effects analyses revealed that this interaction was supported by an effect of Presentation Mode within the “Easy” level of the Lexical Class factor, $F(1,33) = 9.98$, $MSE = 0.78$, $p < .01$. Performance in the Easy-AV condition was clearly better than performance in any of the other conditions. For lists made up of lexically easy words, memory span for lists presented audiovisually was better than memory span for lists presented audio-only. However, there were no differences in memory span as a function of Presentation Mode within the “Hard” level of Lexical Class, $F < 1$.

Figure 2 shows the “Strict” (S) scores for all conditions, averaged across participants. Clearly, the S scores show the same pattern of results evidenced by the H scores. A 2 (Lexical Class) x 2 (Presentation Mode) repeated measures ANOVA on the S scores again showed the main effect of Lexical Class, $F(1, 33) = 15.22$, $MSE = 0.53$, $p < .001$, $h^2 = 0.32$, with the memory span for lexically easy lists ($M = 5.63$, $SD = 0.16$) higher than those for hard lists ($M = 5.14$, $SD = 0.16$). Again, there was no main effect of Presentation Mode.

As with the H scores, we also observed a significant interaction between Presentation Mode and Lexical Class, $F(1,33) = 5.18$, $MSE = 0.36$, $p < .05$, $h^2 = 0.14$. Additional analyses revealed that this interaction was supported by a marginally significant effect of Presentation Mode within the Easy level of the Lexical Class factor, $F(1,33) = 2.86$, $MSE = 0.53$, $p = 0.10$. Again, audiovisual presentation facilitated memory for easy lists, relative to audio-only presentation. There was no effect of Presentation Mode for the Hard level of Lexical Class $F(1,33) = 1.11$, $MSE = 0.46$, n.s.). This means that for lists made up of lexically hard words, there was no difference in memory span due to differences in presentation mode.

Because it was important to determine whether our results replicated those of Goh and Pisoni (1998), the simple effect of Lexical Class within each level of Presentation Mode was analyzed for both the H and S scores. Our experiment would constitute a replication if we found an effect of Lexical Class in the Audio Only level of Lexical Class. The simple effect of Lexical Class within the Audio Only

condition is represented in both figures by comparing the white bar in the Easy panel with the white bar in the Hard panel. However, the simple effects analysis for H scores revealed that there was no effect of Lexical Class within the AO condition, $F < 1$. The analysis for S scores only revealed a marginal effect of Lexical Class within the AO condition, $F(1,33) = 3.57$, $MSE = 0.30$, $p < .07$. Thus, the present study did not replicate Goh and Pisoni (1998).

However, there were significant effects of Lexical Class in the AV condition (comparing the striped bars in each panel) for both sets of scores; H scores: $F(1,33) = 19.92$, $MSE = 0.76$, $p < .001$; S scores: $F(1,33) = 14.9$, $MSE = 0.59$, $p < .001$.

Discussion

While it is problematic that we did not replicate the results of Goh and Pisoni (1998), the marginal effect found using S scores indicates that running more subjects may reveal a difference in Lexical Class in the Audio-only condition. In fact, there are a couple of reasons to support the notion that increasing the N will actually lead to a difference. First, Goh and Pisoni used an N of 40 in their investigation. The present experiment only used an N of 33. Second, Goh and Pisoni only required participants to participate in two conditions. It may be that the additional two conditions in the current experiment served to lower memory spans in general, thereby reducing the size of differences between conditions.

Assuming that the lack of replication denotes a lack of power and not a fundamental flaw in the experiment, the results show that audiovisual presentation of spoken words improves memory span, but only when the words are lexically Easy. We propose that this advantage is due to differences in the ability to maintain the distinction between various list items in memory. Because Easy words are inherently more distinct from their phonological neighbors than Hard words, it is easier to maintain long lists of them in short-term memory (Goh & Pisoni, 1998). We speculate that this advantage may be due to less competition from similar sounding words. Retrieval cues for lists of easy words may be less affected by proactive interference effects from previous lists. Rehearsal processes may be more efficient because less neighbors would be activated during each rehearsal cycle compared to hard words.

Audiovisual presentation, however, interacts with the effects of lexical distinctiveness. Because a hard word has more neighbors, the chances of those neighbors possessing similar audio *and* visual qualities are very high. Thus, for hard words, “added” visual information does not add substantively to the distinctiveness of a target from its neighbors. However, for Easy words, the chance that the additional visual information will distinguish a target item from its neighbors is greater, since there is less of a chance that those neighbors will share *both* auditory and visual characteristics with the target item.

Previous research has shown an effect of audiovisual presentation on working memory. Pichora-Fuller (1996) showed that presenting stimuli at high signal-to-noise ratios actually decreases working memory capacity. However, presenting the words audiovisually counteracted this pattern. Memory spans for audiovisual items presented at high signal-to-noise ratios were roughly equivalent to memory spans for audio-only words presented at lower signal-to-noise ratios. Pichora-Fuller's findings were interpreted as evidence for the reallocation of resources from working memory to the lexically based processes involved in spoken word recognition. When AO items were presented in noise, resources are deallocated from working memory processes and sent to the processes involved in lexical access. AV presentation, it was reasoned, allows for more facile access to lexical representations, and thus, the deallocation of resources from working memory is unnecessary.

Because the task used by Pichora-Fuller was very different from the one used in this study, it is not possible to draw comparisons between her findings and the current ones. Similarly, it is not possible to distinguish between the interpretation of that study's findings with the interpretation offered here. Pichora-Fuller's findings are based on perceptual, "external" noise's effects on memory span, while ours are based on the "internal" noise generated by lexical competitors. It is entirely possible that these two sources of noise have completely different interactions with working memory processes. Our "internal" noise explanation relies on the fact that words from dense neighborhoods will be less likely to become distinct from their neighbors when visual information is included in the signal. In contrast, "external" noise may obscure the acoustic dimensions of spoken stimuli to such a degree that visual information becomes the only information in the signal that actually *can* make competitors distinct. An extension of the current study is currently planned which examines the differences between these two sources of noise and the ways in which they interact with working memory.

Other pilot experiments run in our own lab have also shown an effect of presentation mode on short-term memory tasks. Pisoni, Saldaña and Sheffert (1995) showed that audiovisual presentation of letters leads to a decrease in the number of items correctly recalled in lists of length 6 and 7. Again, this task and the stimuli used were extremely dissimilar from the ones presented in the current study, and so it is hard to draw direct comparisons between their results and our own. Still, the fact that multiple studies have found some effect of presentation mode on short-term memory capacity provides support for the proposal that subtle aspects of stimulus presentation have effects at higher levels of processing, especially on immediate memory span.

The present investigation indicates that visual information about a talker's utterances does interact with higher sources of information in short-term memory for spoken words. Further investigation into the nature of this interaction and its effects on spoken word recognition are necessary to explore more fully the ramifications of such a finding. However, these findings hint that the benefits of audiovisual speech may extend beyond the realm of perception, propagating up the processing system to affect encoding, rehearsal, and immediate memory span.

References

- Baddeley, A.D. (1966). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *Quarterly Journal of Experimental Psychology*, *18*, 363-365.
- Baddeley, A.D. (1998). *Human memory: Theory and practice* (revised ed.). Boston: Allyn & Bacon.
- Baddeley, A.D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *Recent advances in the psychology of learning and motivation* (Vol. VII, pp. 47-89). New York: Academic Press.
- Baddeley, A.D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*, 575-589.
- Bavelier, D., & Potter, M.C. (1990). Visual and phonological codes in repetition blindness. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 134-147.
- Bertelson, P., Vroomen, J., Wiegand, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. *Proceedings of 1994 International Conference on Spoken Language Processing*, *13*, 559-562.

- Bourassa, D.C., & Besner, D. (1994). Beyond the articulatory loop: A semantic contribution to serial order recall of subspan lists. *Psychonomic Bulletin & Review*, *1*, 122-125.
- Conrad, R. (1964). Acoustic confusions in immediate memory. *British Journal of Psychology*, *55*, 75-84.
- Cowan, N., Wood, N.L., Nugent, L.D., & Treisman, M. (1997). There are two word-length effects in verbal short-term memory: Opposed effects of duration and complexity. *Psychological Science*, *8*, 290-295.
- Dekle, D. J., Fowler, C. A., & Funnell, M. G. (1992). Audiovisual integration in perception of real words. *Perception and Psychophysics*, *51*, 355-362.
- Fisher, B. D., & Pylyshyn, Z. W., (1994). The cognitive architecture of bimodal event perception: a commentary and addendum to Radeau (1994). *Current Psychology of Cognition*, *13*, 92-96.
- Gathercole, S.E., Frankish, C.R., Pickering, S.J., & Peaker, S. (1999). Phonotactic influences on short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 84-95.
- Goh, W.D., & Pisoni, D.B. (1998). Effects of lexical neighborhoods on immediate memory span for spoken words: A first report. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 195-213). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 152-162.
- Green, K. P. (1996). The use of auditory and visual information in phonetic perception. In D. Stork & M.E. Hennecke, (Eds.), *Speechreading by humans and machines* (pp. 55-77). Springer-Verlag: Berlin.
- Green, K. P., Kuhl, K. P., & Meltzoff, N. A. (1988). Factors affecting the integration of auditory and visual information in speech: the effect of vowel environment. Paper presented at the meeting of the Acoustical Society of America, Honolulu.
- Hulme, C., Maughan, S., & Brown, G.D.A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*, 685-701.
- Hulme, C., Roodenrys, S., Schweickert, R., Brown, G.D.A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a reintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1217-1232.
- Jones, J. A., & Munhall, K. G. (1997). The effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*, 13-19.

- Jordan, T. R., & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology : Human Perception and Performance*, 23, 388-403.
- Lachs, L., & Hernández, L. R. (1998). Update: The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 22* (pp. 377-388). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Lachs, L., & Pisoni, D.B. (1999). *Effects of multi-modal speech cues on recognition memory for spoken words*. Manuscript submitted for review.
- Landauer, T.K., & Streeter, L.A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12, 119-131.
- La Pointe, L.B., & Engle, R.W. (1990). Simple and complex word spans as measures of working memory capacity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16, 1118-1133.
- Lively, S. E., Pisoni, D. B., & Goldinger, S. E. (1994). Spoken word recognition: Research and theory. In M. Gernsbacher (Ed.), *Handbook of Psycholinguistics*. New York: Academic Press.
- Luce, P.A. (1986). Neighborhoods of words in the mental lexicon *Research on Speech Perception Technical Report No. 6*. Bloomington, IN: Speech Research Laboratory, Indiana University.
- Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear & Hearing*, 19, 1-36.
- Massaro, D. W. & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics*, 58, 1047-1065.
- Massaro, D. W., Cohen, M. M. & Smeele, P. M. T. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, 23, 113-131.
- McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1-86.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- Munhall, K. G., Gribble, P., Sacco, L. & Ward, M. (1995). Temporal constraints on the perception of the McGurk effect. *Perception & Psychophysics*, 58, 351-362.
- Nairne, J.S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin & Review*, 4, 541-545.
- Pichora-Fuller, M. K. (1996). Working memory and speechreading. In D. Stork & M.E. Hennecke (Eds.), *Speechreading by humans and machines* (pp. 257 - 274). Springer-Verlag: Berlin.

- Salame, P., & Baddeley, A.D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, 21, 150-164.
- Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition*, 21, 168-175.
- Sheffert, S., Lachs, L. & Hernandez, L. R. (1996-1997). The Hoosier audiovisual multi-talker database. In *Research on Spoken Language Processing Progress Report No. 21* (pp. 578 - 583). Bloomington, IN: Speech Research Laboratory, Indiana University.
- Sumby, W. H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.
- Smeele, P. M. T., Sittig, A. C., & Van Heuven, V. J. (1994). Temporal organization of bimodal speech information. In *International Conference on Spoken Language Processing: Vol. 3* (pp. 1431-1434).
- Texas Instruments. (1991). TI 46-word speaker-dependent isolated word corpus (CD-ROM). Gaithersburg: NIST.
- Treisman, M. (1978). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, 17, 37-59.

