MULTI-MODAL ENCODING OF SPEECH IN MEMORY: A FIRST REPORT

David B. Pisoni, Helena M. Saldaña, & Sonya M. Sheffert

Speech Research Laboratory
Indiana University
Bloomington, IN 47405

ABSTRACT

Why do people like to watch videos on TV? Why is there now increased interest in video telephones and multi-media technologies that were developed back in the 1960's? Obviously, the availability of new digital technology has played an enormous role in this transition. But, we also believe this is in part due to the same operating principle that encourages listeners in noisy environments to orient toward a talker's face. A multi-modal speech signal is extremely robust and informative and provides information that perceivers are able to exploit during perceptual analysis. In this paper, we present results from two experiments that examined performance in immediate memory and serial recall tasks with normal-hearing listeners using unimodal (auditory-only) and multi-modal (auditory+visual) presentation. Our findings suggest that the addition of visual information in the stimulus display about the speakers' articulation affects the efficiency of initial encoding operations at the time of perception and also results in more detailed and robust representations of the stimulus events in memory. These results have implications for current theories of speech perception and spoken language processing.

1. INTRODUCTION

1.1. Stimulus Variability in Speech Perception

In recent years, we have been studying the problem of stimulus variability in speech perception and spoken word recognition. Our research has shown that the neural mechanisms used in speech perception encode and store fine details of the stimulus input and that many sources of information are not lost or discarded during the perceptual process. These findings have encouraged us to reassess our views about several long-standing theoretical issues in speech perception such as acoustic-phonetic invariance and perceptual normalization, as well as traditional assumptions about the architecture of the mental lexicon (see Pisoni, 1996 for a review).

Mullennix et al (1989) reported that intelligibility for isolated words in noise is influenced by the number of talkers that are presented in the experimental condition. Listeners were presented with lists spoken by a single talker or lists spoken by 15 different talkers. Results showed that identification performance was always better for words presented in the single-talker lists than the multiple-talker lists. These findings suggest that under high talker-variability conditions, listener's engage in some form of on-line "recalibration" each time a new voice is encountered in a set of test trials. These initial results along with other follow-up studies using naming (Mullennix et al., 1989) and speeded classification (Mullennix & Pisoni, 1990) suggest that spoken word recognition is related to the processing of the talker's voice.

1.2. Perceptual Learning

Subsequent perceptual learning experiments demonstrated that detailed properties of a talker's voice are encoded into memory and can be used to facilitate word recognition in noise (Nygaard et al., 1994). These experiments suggest that listeners incidentally encode information about the vocal source attributes when listening to different speakers. It was argued that listeners retained a "procedural memory" for a talker's voice, in addition to specific details about a linguistic event. Thus, the neural representation of spoken words may encompass both an abstract phonetic description of the utterance as well as detailed information about the structural description of the talker's vocal tract.

1.3. Multi-Modal Speech

Taken together, the results of these lines of research suggest that listeners are tracking and encoding many detailed changes in their perceptual environment. In the present report, we describe the results of two studies that have extended our research on stimulus variability in speech perception to the case of multimodal speech perception. The first study reports results on immediate memory span; the second describes findings on serial recall of lists of isolated words. Both studies compared unimodal and multimodal presentation formats to assess how these sources of information interact and influence the representation of spoken words in memory.

It has been known since the early 1950's that listeners show substantial increases in intelligibility of speech when they attend to the talker's face (Sumby & Pollack, 1954). Other studies have demonstrated that visual articulatory information can override or fuse with an auditory speech signal causing a listener to report hearing a combination of the auditory and visual signal (McGurk & McDonald 1976). Some theorists have proposed that articulatory events can be conveyed to the listener through several sensory modalities and that multiple sources of information are used by the perceptual system to recognize speech (Fowler & Rosenblum, 1991). Based on several

findings in the area of multi-modal speech perception, Summerfield (1981) argued for a theory of speech processing which takes into account the integration of various sources of information (auditory, visual, tactile). We believe that methodologies utilized by our lab to look at the issue of stimulus variability might prove enlightening in this area. In particular, we are interested in the issue of whether visual information about a speech event is encoded in memory, and, if so, whether this process requires additional resources from the perceptual system.

Previous research has demonstrated an effect of talker variability on immediate memory span (Saldaña, 1995). This study revealed that listeners' resources are taxed by the use of multiple voices in a traditional memory span experiment. These findings have important implications for current conceptions of working memory. Perhaps one of the most influential theories in working memory to date has been proposed by Baddeley & Hitch (1974) who posit an "articulatory loop" as a mechanism for working memory span. According to Baddeley, memory span is constrained by how many items a subject can repeat or that can be "refreshed" by a set of articulatory control processes in approximately a two second duration. Our previous results showed that working memory span is also affected by the *amount* of information that is contained in each representation. Research on the nature of working memory span is particularly important in light of the role that the concept of working memory plays (either implicitly or explicitly) in almost all current theories of language processing. One goal of our current research is to determine the effect that visual speech information has on working memory.

We are also interested in the effect that visual speech information has in secondary (long-term) memory. Although previous research from our lab has consistently shown that stimulus variability influences perceptual processing, we have also found a *benefit* in recall due to stimulus variability (Goldinger et al, 1991). This research demonstrated that listeners are actually better at recalling items from multiple-talker lists than single-talker lists, however, this effect is only found in the primacy portion of the list and is only observed when listeners are given sufficient time for rehearsing the word lists. Goldinger et al (1991) proposed that the addition of multiple voices aids in the elaboration and transfer of information into long term memory. Following from this result, we were interested in the effect of visual speaker information on serial recall.

2. IMMEDIATE MEMORY SPAN

2.1. Method

Subjects. Twenty-one subjects participated in the experiment as partial fulfillment of a class requirement in Introductory Psychology. Four subjects were discarded from the final analysis for failing to follow instructions. All subjects were native speakers of English with normal hearing and normal or corrected vision. The experiment lasted approximately 45 minutes.

Stimulus Materials. The stimuli consisted of a list of letters spoken by a female actor. The actor was videotaped with a camcorder which was patched into a professional video recorder. A microphone was positioned near the actor's mouth and out of view of the camera. The actor produced 26 letters in a sound attenuated recording studio. Each item was presented to the actor in random order on a teleprompter and the actor was instructed to look directly into the camera and say the item clearly. The video clip was digitized using a Macintosh Quadra 950 at 30 frames per second. The audio signal was sampled at 22 kHz with 16-bit resolution. Individual clips of each utterance were made by cutting the clip when the mouth was in a neutral position prior to an utterance and then in a neutral position after the utterance.

A preliminary intelligibility test demonstrated that all of the stimuli were intelligible at 100%. A subset of the stimulus items were then chosen for the memory span experiment. The items selected were: B, D, F, H, J, K, L, M, Q, R, Z. The lists were presented in a staircase fashion, with the list increasing or decreasing by one item on each trial. Each subject started with an easy list of four items. The list then increased in length by one item on each trial until subjects were presented with a nine item list. Then the length of the list decreased by one item on each trial until subjects were presented with a four item list. For each half of the experiment, the subject was presented with each list length six times. The first six trials of each part of the experiment served as practice trials and were not included in the final analysis. Two presentation conditions were used: unimodal and multimodal. The presentation of the unimodal and multimodal conditions were blocked and counterbalanced.

Procedure. Subjects were presented auditory items over the headphones. The first item was always a 1000 Hz 500 ms tone accompanied by a visual display that said "get ready". The ready signal was followed by a list of test items, which was then followed by a second tone. The subjects were instructed to wait for the second tone and then repeat the letters that they heard in the exact order that they were presented. The responses were recorded by hand by an experimenter in the next booth. Subjects were told to keep their gaze focused on the monitor throughout the entire experiment.

2.2. Results

A list was scored as correct if, and only if, all items were recalled in the proper serial order (Saldaña, 1995). An overall analysis of variance was conducted with Mode (2), and List Length (6), as within subject variables and order (2), as a between subjects variable. The analyses revealed an overall effect of Mode $F(1,15)=16.17$, $p<.01$. Subjects' memory span was longer for unimodal presentations compared to multimodal presentations. There was also an overall effect of List Length $F(5,75)=160.23$, $p<.01$. Subjects were better at shorter lists than longer lists. There

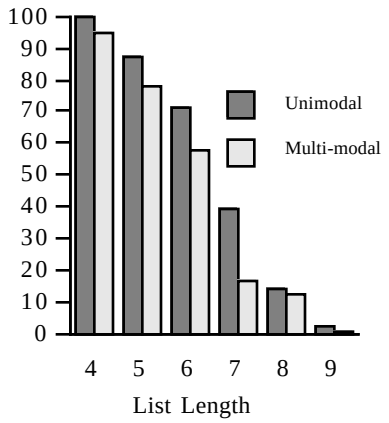

Figure1: Percentage of correct list lengths for the unimodal and multi-modal conditions in the letter span experiment.

was no effect of order. Post-hoc comparisons revealed that the presentation modality was significant at List Length 6 and 7, $F(1,75)= 5.53$, $p<.05$; $F(1,75)=13.881$, $p<.05$, respectively. Unimodal memory span was higher than multimodal span.

2.3. Discussion

The results show that the addition of visual information to auditory speech results in a shorter working memory span. This finding is consistent with previous talker variability results which demonstrate that working memory is constrained by the amount and quality of information that is being processed by the system, not just the absolute duration of items. However, in contrast to previous talker variability results, this result can be accounted for within Baddeley's framework of working memory. The theory proposes three mechanisms for working memory: the articulatory loop, the visual spatial sketch pad, and the central executive system. The articulatory loop is responsible for maintaining phonological codes, while the visual spatial sketch pad is responsible for rehearsing spatial location as well as spatial movement. It is assumed that working memory is a limited capacity system, therefore, the ability to process information in the articulatory loop is affected by any simultaneous processing of information in the visual spatial sketch pad. According to this view, it might be expected that the auditory and the visual information are being processed separately at the level of working memory, and are making demands on a common set of resources.

The present experiment was conducted under very favorable signal-to-noise ratios and as a consequence it is an unusual example of audio-visual speech perception. For normal listeners under good listening conditions, visual information is not necessary for recognition. It is possible that our results were not due to the additional processing of visual information but rather a consequence of the distracting quality of the audio-visual condition. One way to investigate this issue is to determine whether visual information has been transferred to long-term memory store. Our next experiment addressed this issue using a serial recall procedure.

3. Serial Memory

3.1 Method

Subjects. Forty subjects participated in the experiment as partial fulfillment of a class requirement in Introductory Psychology. All subjects were native speakers of English with normal hearing and normal or corrected vision.

Stimulus Materials. The stimuli consisted of thirty lists of 10 monosyllabic English words that were taken from the Johns Hopkins Laser Disk Corpus (Bernstein & Eberhardt, 1986). All of the words were produced by a male talker.

Procedure. The subjects were tested individually. The subjects were randomly selected to participate in either the multimodal or the unimodal condition. On each trial, subjects were presented with a list of ten words with an SOA of 2 seconds. A 1000 Hz 500 ms tone was sounded prior to and following the presentation of each word list. After the second tone, the subjects were instructed to write down the items presented to them. They were allowed to output their responses in any order. However, they were instructed to write each item in the space on their answer sheet which corresponded to the order in which the words were presented.

3.2. Results

The items were only scored as correct if they were written down in the space corresponding to the order of presentation. An analysis of variance revealed a main effect of position $F(9,342)=77.71$, $p<.01$, as well as a marginally significant interaction of position and condition $F(9,242)=1.78$, $p=.07$. Post-hoc analyses showed significant differences between the audio and the audio-visual conditions for serial positions 1 and 2 $F(1,38)$ 14.05, $p<.001$, $F(1,38)$ 357.72, $p<.001$, respectively.

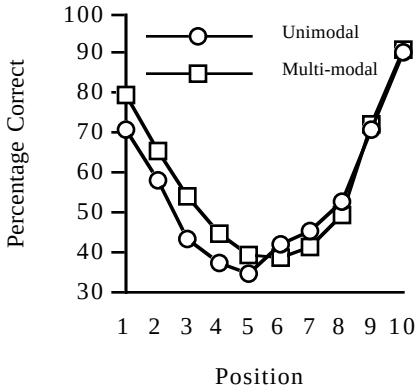

Figure 2: Percentage correct for unimodal and multi-modal conditions in the serial memory experiment.

3.3. Discussion

The present findings are consistent with previous findings from our lab which demonstrated a benefit in recall for multiple-talker lists in the primacy portion of the serial position curve. In earlier papers, we argued that this effect was due to elaborative encoding and transfer of information into a long-term memory store (Goldinger et al, 1991). We believe that the same explanation is appropriate for the present set of results. This conclusion is supported by the finding that the benefit of visual information is only evident in the primacy portion of the list which is usually believed to reflect recall of items from long-term memory. This account is also supported by our previous immediate memory span results, which indicate that the limited capacity working memory system is taxed by the perceptually rich multi-modal presentations.

4. GENERAL DISCUSSION

The present set of findings on immediate memory span and serial recall for multi-modal presentations add to our earlier results which demonstrate that specific details about the form of the speech signal are processed and encoded by the perceptual system. We suggest that attributes of the talkers face are encoded in working memory and transferred to representations in long-term memory. In addition, the present findings on immediate memory span raise several important questions about the "articulatory loop" hypothesis of working memory, specifically, questions about the nature of information that working memory has access to, as well as the properties and operations of the rehearsal mechanism in working memory. It is now clear that immediate memory span is affected by attributes of both the talker's voice (Saldaña, 1995), and the talker's face. The present results suggest that this information may not be perceptually integrated at the time of initial encoding. Instead the visual information in the talker's face may use resources from the visual-spatial sketch pad which places additional demands on a limited capacity working memory system.

The results obtained in the serial recall experiment under multi-modal presentations suggest that the additional visual information about the talkers face *is* retained and used in subsequent recall. We propose that the additional information facilitates the rehearsal process and/ or the transfer of items to long-term memory. The findings from this experiment are similar to the results obtained several years ago using multiple voices (Goldinger, et al 1991). The presence of additional information about an item, such as the talkers voice, appears to provide the perceptual system with the ability to build a more detailed or robust representation. The presence of additional stimulus dimensions about a talkers voice may aid retrieval mechanisms which use discriminability and distinctiveness to recover items from memory. Apparently, multi-modal presentation of speech also helps this elaboration process to work more efficiently. However, this elaboration is very selective in nature, showing up, as in our earlier recall experiments, only in the primary portion of the serial position curve.

5. ACKNOWLEDGMENTS

This research was supported by NIDCD Research Grant DC-00111-18 to Indiana University in Bloomington, IN. We gratefully acknowledge the assistance of Luis Hernandez, Bob Bernacki, and Lisa Burgin.

6. REFERENCES

1. Baddeley, A. D., & Hitch, G. (1974)."Working memory". In G. A. Bower (Ed.), *Recent advances in learning and motivation*, Vol. 8. New York: Academic Press.

2. Bernstein, L.E. & Eberhardt, S.P. (1986). Johns Hopkins Lipreading Corpus I-II: Disc 1. Baltimore, MD: Johns Hopkins University.

3. Goldinger, S. D., Pisoni D.B., & Logan J.S. (1991). "On the nature of talker variability effects on serial recall of spoken word lists". *JEP: Learn., Mem., and Cog., 17*, 152-162.

4. Fowler, C. A. & Rosenblum, L. D. (1991). "Perception of the phonetic gesture". In I. G. Mattingly and M. Studdert-Kennedy (Eds.), *Modularity and the motor theory* (pp., 33-59). Hillsdale, NJ: Erlbaum.

5. Johnson, K., & Mullenix, J. (1996). *"Talker variability in speech processing"*. New York: Academic Press.

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

7. Mullennix J.W., Pisoni D.B., & Martin C.S. (1989). "Some effects of talker variability on spoken word recognition". J. *Acoust. Soc. Am. 85*, 365-378.

8. Mullennix, J.W. & Pisoni, D. B. (1990). "Stimulus variability and processing dependencies in speech perception". *Perc. & Psych., 47*,379-390.

9. Nygaard, L. C., M. S. Sommers, & Pisoni, D. B. (1994). "Speech perception as a talker-contingent process." *Psych. Sci.* **5**, 42-46.

10. Saldaña, H. M. (1995). The effects of talker-specific information on immediate memory span. *Paper presented at the 129th meeting of the Acous. Soc. Am.* Washington D.C. , May 30- June 3.

11. Sumby W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soci. Am., 26*, 212-215.

12. Summerfield, A. Q. (1981). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip reading* (pp. 3-51). London: Erlbaum.