

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 20 (1995)
Indiana University

**Encoding of Visual Speaker Attributes and
Recognition Memory for Spoken Words¹**

Helena M. Saldaña,² Lynne C. Nygaard,³ and David B. Pisoni

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

¹This work supported in part by NIH-NIDCD Research Grant DC00111 and NIH-NIDCD Training Grant DC00012 to Indiana University–Bloomington.

²Now at House Ear Institute, Los Angeles, California.

³Now at Department of Psychology, Emory University, Atlanta, Georgia.

Encoding of Visual Attributes and Recognition Memory for Spoken Words

Abstract. An experiment was designed to assess the extent to which visual articulatory information is encoded in memory. In a recent study, Palmeri, Goldinger, and Pisoni (1993) reported that listeners were more accurate at recognizing previously presented words when they were presented in the same voice as at test than when they were presented in a different voice. This result suggests that detailed voice information is not stripped away by a normalization process during the early stages of spoken word recognition; instead information about the talker's voice is encoded into long-term memory and may later facilitate recognition of spoken words. The present investigation was designed to determine whether detailed cross-modal linguistic information is also retained in long-term memory. Subjects were presented with two audio-visual speakers producing lists of isolated words. The words were presented at three signal-to-noise ratios ranging from +5 dB to -5 dB. In the low signal-to-noise conditions, listeners are forced to attend more carefully to the visual information presented in order to extract the linguistic content. It is proposed that this manipulation will cause dynamic visual information to be encoded along with the spoken words in memory. If it is the case that detailed visual information about the talkers articulation is retained in long term memory, then we would expect that listeners will be better at recognizing "old" words when they are presented with the same visual speaker at test. No significant effect of visual articulation was found on recognition memory, however a consistent trend was observed. The present results have implications for current theories of spoken word recognition and the nature of the representation of words stored in the mental lexicon.

Introduction

Speaker Variability

The speech signal is highly variable across individual talkers. This is largely the result of differences in the shape and length of the vocal tract (Carrell, 1984; Fant 1973; Summerfield & Haggard 1973), glottal source function (Carrell, 1984), individual differences in articulation (Ladefoged, 1980), and dialect. Traditional theories of speech perception have often treated such variability as noise in the signal that must be filtered out by the listener (e.g., Blandon, Henton, & Pickering, 1984; Disner, 1980; Green, Kuhl, Meltzoff & Stevens, 1991). Many of these theories assume some kind of talker-normalization process, in which talker specific information is stripped away from the signal and the remaining phonetic cues are matched to idealized representations in memory. This view would predict that only an abstract symbolic linguistic code, and not its carrier are retained in long term memory.

Recent evidence suggests that detailed information about the speaker's voice is encoded into long-term memory. Craik & Kirsner (1974) utilized a continuous recognition paradigm and demonstrated that recognition memory was facilitated when the test item was presented in the same voice as the study item. Palmeri, Goldinger, & Pisoni (1993) extended this finding by presenting listeners with twenty different voices and increasing the lag between study and test items. They replicated Craik & Kirsner's finding by

showing a clear voice effect even when there were 64 intervening items between study and test. When a word was repeated in the same voice, subjects showed higher levels of recognition performance than when the same word was presented in a different voice.

In a similar study using visual stimuli, Kirsner (1973) found that recognition memory was better when test items were presented in the same typeface as study items. There is also evidence in the literature showing that surface information is retained when subjects read passages of text. Kolers & Ostry (1974) presented subjects with inverted passages of text and had them practice reading them. In a subsequent session, they had subjects read text in the same inverted form or in a different inverted form. They found that reading times were best when subjects were given the same inverted form at test. In fact, savings was found even one year after the original presentation. All of these findings indicate that surface features of auditory and visual stimuli are not discarded during the perceptual process, but rather are encoded in the memory trace and retained in long term memory.

Audio-Visual Speaker Variability

In a recent study, Sheffert & Fowler (1995) designed an experiment to determine whether visual information about the speaker was encoded along with the phonetic representation of the word. The question that they explored was the following: When listeners are able to view the speaker talking do they also retain in memory physical details about the speakers face along with the lexical representation? The authors utilized a continuous recognition procedure like the one used by Palmeri et al. with audio-visual tokens. They wanted to show that observers would be better at recognizing words when the same visual speaker was presented at study and test. In a series of four experiments, the authors consistently replicated the voice effects found by Palmeri et al., but they failed to show an implicit effect of face information on word recognition. The authors concluded that voice information shared a "privileged" status in relation to phonological information in memory for spoken words.

In this study, Sheffert & Fowler assumed that any information about the speaker should be encoded along with the phonetic representation. For instance, in one condition the authors presented listeners with the same speaker at test but changed an item of clothing on the speaker (i.e., hat or scarf). In all previous studies which demonstrated the encoding of surface features with the linguistic code, the surface features shared an integral relation with the spoken message. In other words, some processing of the surface features had to be performed to retrieve the linguistic message. It is clear that listeners under normal listening conditions do not require visual speaker information to retrieve linguistic content. Therefore, the lack of an implicit face effect in the Sheffert & Fowler study might be due to the fact that it was not necessary to process the face information in a mandatory way to perceive the word.

Current Investigation

In the present investigation, we manipulated the signal-to-noise ratio in a continuous recognition task, in an attempt to force listeners to process the visual articulatory information in a mandatory fashion. If the encoding of surface features is due to an integral relationship between surface features and the linguistic code, it is possible that this manipulation will cause visual surface features to be encoded in memory along with the lexical items. For half of the repeated items the visual speaker was presented in a dynamic display. For the other half of the repeated items, the visual speaker was presented in the form of a static image. Therefore, face information is available to the subject but no dynamic articulatory information is available for the subject to process. We propose that listeners will only show a benefit in recognition when the dynamic information from the same talker is available at both study and test.

Method

Subjects. One-hundred-nineteen subjects were paid \$5.00 an hour for participating in the experiment. All subjects were native speakers of English with normal hearing and normal or corrected vision.

Stimulus Materials. The stimuli were a list of isolated words spoken by a male and a female talker. Two different talkers were videotaped uttering three hundred monosyllabic words. Talkers were recorded in a sound attenuated studio. Each word was presented to the talker on a CRT screen and the talker was instructed to say each word while looking into the camera. The stimuli were digitized on a Macintosh Quadra 950. The sound was sampled at 20.5 kHz with 16 bit resolution. The video image was captured at 30 frames per second. The length of each word varied from 2 to 3 seconds. Some items were randomly selected to serve as audio-alone trials. These trials did include an image of the actors face, however, the actor was not articulating the word. A static image was constructed by placing the same frame in consecutive positions for the duration of the originally articulated utterance.

Noise-Embedded Conditions. A General Radio 1381 random-noise generator was attached to the Macintosh Quadra 950. The noise generator was controlled by a voice key. Any time a sound was detected from the sound board at the initiation of a word, white noise was output. When the sound ceased, the white noise was deactivated. Therefore, the noise was only present during the auditory presentation of each word.

The test list was constructed from a subset of two hundred words from each talker (the same words for each talker). Each word was presented and repeated once. The repetition of any given word occurred after a lag of 2, 8, 16, or 32 intervening items. Each lag value was used an equal number of times in each list. One hundred forty-four pairs of presented and repeated items were used in the list. All original items were articulated by each speaker, however, half of the repeated items consisted of a static image of the talker's face in place of the visually articulated utterance (72 items). Each subject was given an initial practice list of 15 words to become familiar with the task. None of these words were repeated in the experiment. The next 30 items were presented to establish a memory load and were not considered in the analysis. Twenty-one filler items were randomly distributed throughout the test portion of the list. All of the filler items consisted of a static visual presentation. The total number of words in the list equaled 354 items.

Each talker was presented an equal number of times on the list. On the initial presentation of the word, one of the talkers was randomly selected. The probability of the same talker or different talker repetitions was equal.

Design. The lag between the initial presentation and the repeated word was a within subjects variable (2,8,16,32), as was the talker who produced the repetition (same talker vs. different talker) and type of repetition (static vs. dynamic).

Procedure. Subjects were tested in groups of 3 or fewer. Stimulus presentation was controlled by a Macintosh Quadra 950. After hearing and seeing each word, the subject was allowed a maximum of four seconds to respond on a numbered response sheet. The subject's task was to simply circle the word OLD if they had heard the item previously or NEW if they had not. Subjects were told it was very important to watch as well as listen to each presentation. An experimenter was seated in the room to make sure subjects looked directly back at the monitor after each response.

Groups of subjects were randomly selected to serve in one of four conditions: The following number of subjects were assigned to each condition: Clear, N=29; SNR +5, N= 27 ; SNR 0, N=34; and SNR -5, N=29.

Results

All the data to be presented here are expressed in terms of hit rates (the number of items identified as "old" given that they were previously presented). We first examine the performance of the listeners in the clear. There was a significant effect of talker. Subjects were more likely to correctly identify an item as "old" if it was presented by the same speaker $F(1, 84) = 35.727, p < .001$. This was true regardless of whether the visual speaker was articulating the word or not. There was also a significant effect of lag ($3, 84) = 30.257, p < .001$. This result indicated that recognition memory performance decreased as the lag increased.

 Insert Figure 1 about here

An overall $3 \times 2 \times 2 \times 4$ ANOVA was conducted on the between subjects factor of signal-to-noise ratio (+5, 0, -5), visual-display (articulated, non-articulated), talker (same talker, different talker) and lag (2, 8, 16, 32). The main effect of signal-to-noise ratio was highly significant $F(2, 87) = 9.433, p < .001$, demonstrating that signal-to-noise ratio had an overall effect on the recognition of items. The effect of talkers was also highly significant, $F(1,87) = 360.53, p < .001$. Subjects were better at recognizing an item as "old" when it was presented by the same talker used in the original presentation. However, the interaction of articulation \times talker was not significant. This finding indicates that dynamic articulation did not differentially help recognition for same versus different talker.

Separate ANOVAs were conducted on each signal-to-noise ratio. In all of the conditions, we observed significant main effects for talker and stimulus lag. There was no main effect of articulation, again, suggesting that seeing a dynamic visual image of the talker's face during the repetition did not affect recognition performance.

 Insert Figure 2 about here

Discussion

The present findings replicate the implicit voice effect reported by Palmeri et al (1993). Subjects showed a clear advantage on item recognition when the repeated item was presented in the same voice as the original item. This implicit voice effect was not eliminated when words were embedded in noise.

We also observed an overall decrease in hit rate as SNR decreased. However, we did not observe a difference in performance between dynamic and static stimuli. It is well known that the presence of dynamic visual information about the talker's face increases intelligibility of degraded speech. Based on these results, we might have expected that subjects would do better when presented with a dynamic face regardless of voice (simply due to better intelligibility), however, this prediction was not observed. It is

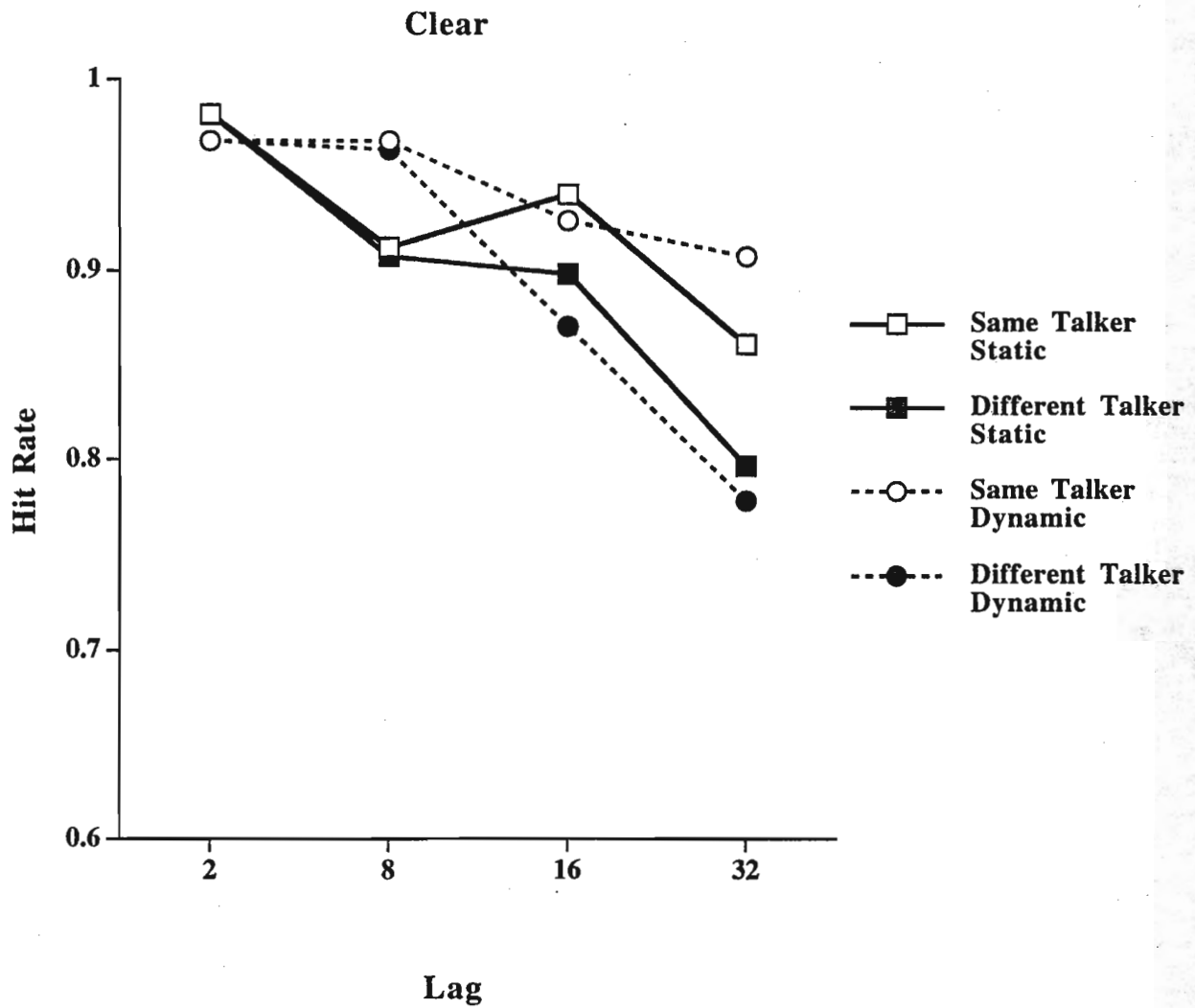


Figure 1: Hit rates for words in the clear speech condition

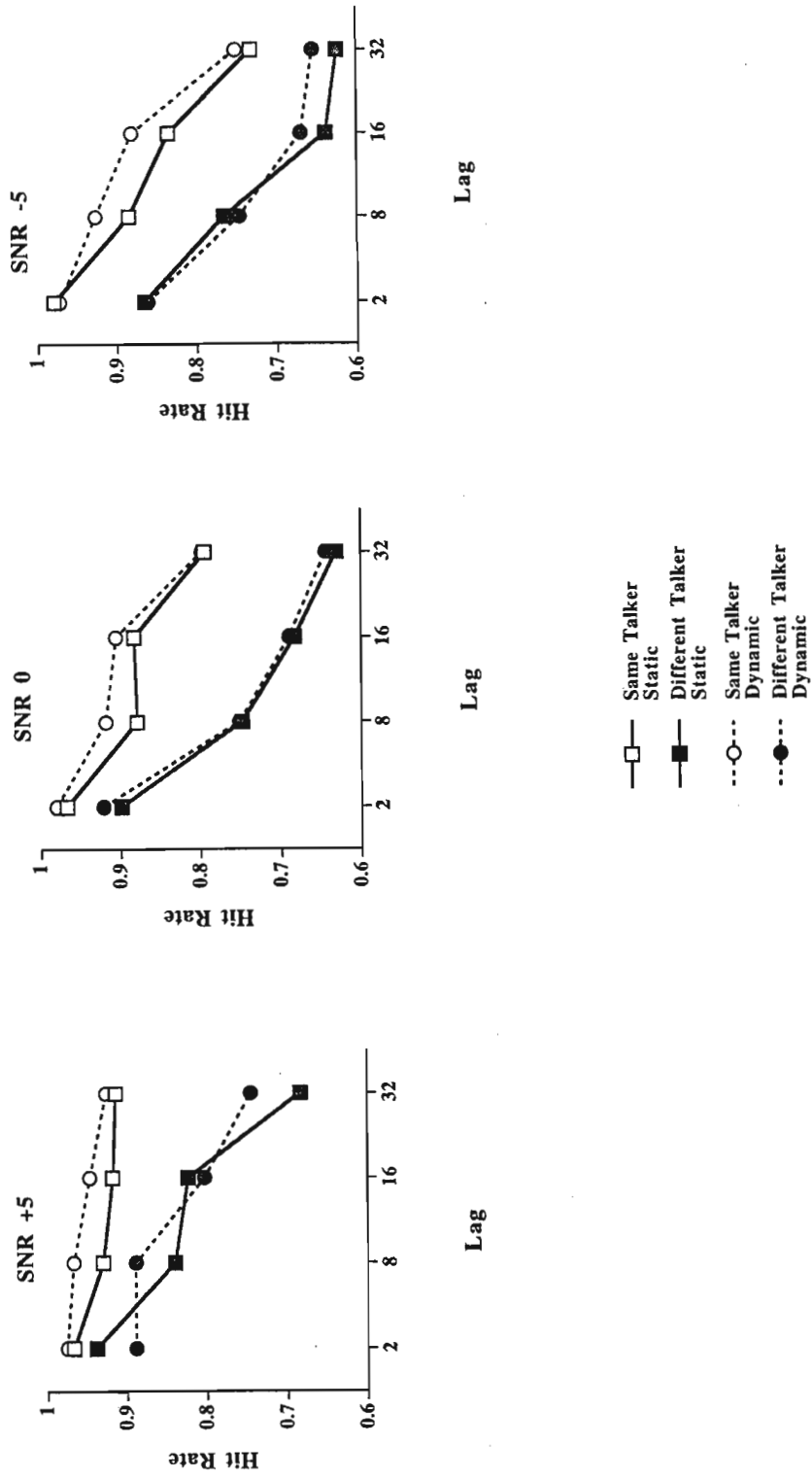


Figure 2: Hit rates for words in each of the three signal-to-noise conditions

possible that the noise levels were not high enough to force listeners to process the visual information in a mandatory way. It is also possible that the words were too distinctive therefore preventing confusion at the lower SNR ratios.

General Discussion

This experiment was designed to investigate whether visual surface features of spoken words are encoded in memory along with a symbolic linguistic code. We hypothesized that forcing listeners to use visual information to extract the linguistic message would cause them to encode the complementary visual information in a mandatory fashion. This hypothesis was not borne out. However, an interesting pattern emerged from the data. Across all signal-to-noise ratios, the function for the same talker-articulated (dynamic) visual stimuli was higher than the same talker-non articulated (static) visual stimuli. This observation suggests that visual articulatory variability might be encoded in long term memory. Recall that in the same talker condition the voice remained the same regardless of the visual display, the only difference between the same talker stimuli was the dynamic visual information. It is possible that the highly robust voice effect washed out the effects of dynamic visual information. In subsequent studies we plan to use a cross-splicing technique in order to completely disassociate the auditory and visual information and force observers to rely on only one source of information or the other in encoding the stimulus array.

Studies are also planned which compare the results from our normal listeners with performance obtained with hearing-impaired and good and poor speechreaders. It is possible that the absence of a robust effect in the present study is due to the fact that our normal-hearing listeners routinely rely on auditory information for speech communication. If so, we might observe a visual effect when we use a group of observers who rely heavily on visual information for speech communication.

In summary, this study was designed to investigate the nature of the representations encoded during spoken word recognition. It was demonstrated that voice information is encoded during word recognition and that dynamic visual talker information is also encoded. However, further research is required to determine to what extent the visual information is useful in subsequent recognition memory.

References

- Blandon, R.A.W., Henton, C.G., Pickering, J.B. (1984). Towards an auditory theory of speaker normalization. *Language and Communication*, 4, 59-69.
- Carrell, T.D. (1984). Contributions of fundamental frequency formant spacing and glottal waveform to talker identification. *Research on Speech Perception Technical Report No. 5*. Bloomington: Indiana University, Department of Psychology.
- Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, 26, 274-284.
- Disner, S.F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, 67, 253-261.
- Fant, G. (1973) *Speech Sounds and Features*. Cambridge, MA: MIT Press

- Green, K.P., Kuhl, P.K., Meltzoff, A.N., & Stevens, E.B. (1991). Integrating speech information across talkers, gender, and sensory modalities: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524-536.
- Kirsner, K. (1973). An analysis of the visual component in recognition memory for verbal stimuli. *Memory and Cognition*, **1**, 449-453.
- Kolers, P.A. & Ostry, D.J. (1974). Time course of loss of verbal information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior*, **13**, 599-612.
- Ladefoged, P. (1980). What are linguistic sounds made of? *Language*, **56**, 485-502.
- Palmeri, T.J., Goldinger, S.D. & Pisoni, D.B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory & Cognition*, **19**, 309-328.
- Sheffert, S. & Fowler, C.A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Learning & Memory*, **34**, 665-685/
- Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalization as demonstrated by reaction times. *Report in Progress in Speech Perception, Vol. 2.* (pp. 1-12). Belfast, UK: The Queen's University of Belfast, Department of Psychology.