

RESEARCH ON SPOKEN LANGUAGE PROCESSING
Progress Report No. 27 (2005)
Indiana University

**Cross-modal Priming of Auditory and Visual Lexical Information:
A Pilot Study¹**

Adam B. Buchwald and Stephen J. Winters

*Speech Research Laboratory
Department of Psychological and Brain Sciences
Indiana University
Bloomington, Indiana 47405*

¹ This work was supported by NIH-NIDCD DC00012. The authors would like to thank David Pisoni for helpful advice and comments and Melissa Troyer for editing videos and running subjects.

Cross-modal Priming of Auditory and Visual Lexical Information: A Pilot Study

Abstract. This study assessed whether presenting visual-only stimuli prior to auditory stimuli facilitates the recognition of spoken words in noise. The results of the study indicate that this type of cross-modal priming does occur. Future directions for research in this domain are presented.

Introduction

Psycholinguistic studies often employ priming paradigms to address issues of whether and when certain representations are active in the course of language processing. In priming studies, researchers typically examine changes compared to a baseline level of performance in responding to a ‘target’ stimulus when the target is preceded by a ‘prime’ stimulus. The changes in participants’ performance on the task that result from presentation of the prime are argued to indicate something about the relationship between the target and prime stimuli in the cognitive processing required for the task. For this reason, primes are generally selected that share some – or all – features with the target stimuli. For instance, in phonological priming, a spoken target and a spoken prime typically share some subset of phonological features. In repetition (or identity), priming, the prime and target are identical and thus share all features with one another.

In this pilot study, we used cross-modal priming and presented spoken word primes in the visual domain only, followed by an auditory-only presentation of the same word, spoken in noise, as the target stimulus. We were interested in using this priming paradigm to find out whether there were enough shared features between sensory modalities in the auditory and visual representations of the spoken word for the visual prime to facilitate the recognition of the auditory target.

There exist two lines of evidence in the psycholinguistic literature that suggest the information in a visual-only stimulus will facilitate lexical processing. Dodd, Oerlemans, and Robinson (1989) found that lexical repetition priming is robust across different modalities of presentation. They presented research participants with lexical primes and targets in three different modalities: orthographic, auditory, and visual. On each critical trial, the lexical item presented as the target and the prime were identical. The research participants were required to make a speeded semantic categorization (‘animal’ or ‘plant’) on the target lexical item. Importantly, Dodd et al.’s results indicated that the presentation of visual-only stimuli facilitated processing in the semantic categorization task for all three target types. With respect to the present experiment, it is noteworthy that participants were faster on the semantic categorization task with auditory targets when there was a visual prime than when the priming stimulus was absent. This result suggests that the information present in the visual-only prime facilitated processing of the semantic lexical information present in the auditory target stimulus.

Another line of evidence suggests that observers can integrate information present in auditory and visual signals, even when those signals are presented in separate modalities. Lachs and Pisoni (2004) performed a series of ‘cross-modal matching’ tasks in which participants were asked to match visual-only stimuli with auditory-only stimuli. Using an XAB task, observers viewed a silent video of a speaker producing a word, followed by two auditory-only stimuli produced by two different speakers (of the same gender) saying the same word. The observers’ task was to identify which of the two auditory stimuli

came from the same speech event as the visual stimulus. Participants were able to match the appropriate auditory stimulus to the video at a rate significantly greater than chance. Importantly, when a still image of a speaker was presented as the visual stimulus instead of a dynamic video clip, participants performed at chance levels in matching the stimuli of the two modalities. Lachs and Pisoni argued that the information present in dynamic video clips and their corresponding auditory tracks were perceived as part of an integrated stimulus; that is, they were simply two sources of information about the same perceptual event in the external world.

Current Investigation

The present pilot study was performed to determine whether the presentation of a silent, dynamic video clip of a speaker would facilitate the recognition of spoken words, presented in only the auditory domain. This study tested whether this type of visual-audio cross-modal priming affects spoken word recognition, in addition to semantic categorization (as shown by Dodd et al., 1989). If so, the cross-modal priming paradigm could be used to explore a wide range of additional issues related to the operations and representations that underlie the processes required for spoken word recognition.

Method

Participants

Nine Indiana University undergraduate students, ages 18-20, participated in this study in partial fulfillment of course requirements for introductory psychology. All participants were native speakers of English with no speech or hearing disorders; they all had normal or corrected-to-normal vision at the time of testing.

Materials

All stimuli materials were drawn from the Hoosier multi-talker AV database (Sheffert, Lachs, & Hernandez, 1997). Only monosyllabic, CVC words produced by one female speaker in the database were selected for use in this study. Of the 96 different word tokens included in this study, half were “Easy” words – high frequency words with sparse phonological neighborhood densities (e.g., “fool,” “noise”), while the other half were “Hard” words – low frequency lexical items with high neighborhood densities (e.g., “hag,” “mum”; Luce and Pisoni, 1998).

Auditory Stimuli. In this experiment, we used envelope-shaped or ‘random bit flip’ noise (Horii, House, & Hughes, 1971) to reduce performance on the spoken word recognition task. Presenting the auditory stimulus in noise is necessary to detect the potential facilitatory effects of priming in the spoken word recognition task; performance must be below ceiling for any effects to be detected. To create these stimuli, the acoustic track of each video recording was first saved to an independent .AIFF file, using QuickTime Pro software. These files were then processed through a MATLAB program which randomly changed the sign bit of the amplitude level of 30% of the spectral chunks in the acoustic waveform.

Visual Stimuli. Two kinds of visual primes were used in this study: static and dynamic. Dynamic primes consisted of the original, unedited video clips associated with each target word from the Hoosier Audio-Visual Multi-Talker database. The video track of the static primes, on the other hand, consisted of only a still shot of the first frame of the video associated with the target word in the Audio-Visual database. The duration of the static primes was identical to that of their counterparts in the dynamic prime condition.

Procedure

Participants were tested in groups of four or fewer in a quiet room. During testing, each participant wore Beyer Dynamic DT-100 headphones while sitting in front of a Power Mac G4. A customized SuperCard (version 4.1.1) stack, running on the PowerMac G4, presented the stimuli to each participant. The instructions for the experiment were presented to the participants on the computer screen prior to the first experimental trial and are repeated below:

In this experiment, you will attempt to identify a series of words that you hear. The words will be difficult to understand. After you hear each word, you should attempt to identify the word that was spoken. You can respond by typing into the computer.

Before each word, you will see either a still image of a speaker or a movie of a person saying a word. Regardless of what you see, your task is to identify the word that you hear. Even if you do not think you understood the word, please make your best attempt to identify what you heard.

On each trial during the experiment, the SuperCard stack first presented participants with either a Static or a Dynamic visual prime. All videos had a 640x480 aspect ratio and filled the entire monitor screen when they were presented to the participants. The sound output from the computer was muted while the videos were presented to the participants. Five hundred milliseconds after the presentation of the visual prime, the participants then heard the auditory target word over the headphones. Following the auditory stimulus, a prompt appeared on the screen asking the participant to type in the word they heard. The prompt remained until the participant pressed the “Enter” key on the keyboard at which point a “Next Trial” prompt appeared. The next trial began after the participant used the mouse to click the “Next Trial” prompt.

Words were presented to participants in random order with Dynamic and Static primes randomly interleaved. Each participant responded to 48 words in each priming condition.

Results

The data were analyzed using a repeated measures Analysis of Variance (ANOVA) with prime condition (Dynamic vs. Static) and target type (Easy vs. Hard) as independent variables and word identification accuracy as the dependent variable. Correct responses were counted for all typographical matches between stimulus and response, as well as homophones (e.g., “peace” and “piece”) and obvious typos (e.g., “cheif” for “chief”). The percentage of correct responses in each priming condition, for both target types, is represented graphically in Figure 1. The ANOVA revealed a main effect of prime condition [$F(1,8) = 33.71, p < 0.001$]. Participants were significantly more accurate on trials with Dynamic primes ($\bar{x} = 68.1\%$, $\sigma = 7.0\%$) than on trials with Static primes ($\bar{x} = 52.5\%$, $\sigma = 5.4\%$). A main effect of target type was also significant, with participants performing better on Easy targets ($\bar{x} = 68.1\%$, $\sigma = 12.9\%$) than Hard targets ($\bar{x} = 52.1\%$; $\sigma = 8.8\%$; $F(1,8) = 18.14, p < .01$). There was no interaction between prime type and target type, although there was a trend, $F(1,8) = 3.46, p < .11$, with a larger effect of prime condition for Easy words than for Hard words.

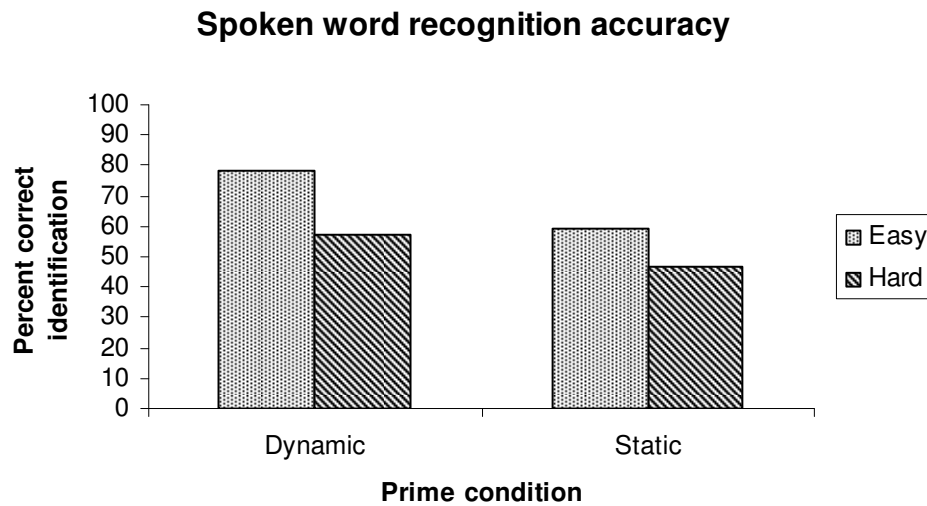


Figure 1. Percentage of words correctly identified as a function of prime condition and target type.

Discussion and Future Directions

The results of this initial study indicate that presenting a visual-only version of a word prior to presenting that word auditorily in noise facilitates the correct identification of the auditory target. This preliminary result has important implications for our understanding of spoken word recognition. In particular, the accuracy benefit for visually primed words suggests that the information that observers receive from the visual presentation of words aids the successful lexical identification of the auditory signals.

Future research building on this pilot study will involve attempting to determine what type of cognitive processing mechanism underlies this pattern of results. Two such proposals are considered here. First, Lachs and Pisoni (2004) have argued that success on the cross-modal matching task was due to observers' integration of the auditory and visual components of the speech act, as each provides information about the same event in the physical world (following the framework of Gibson, 1966). With respect to the study reported here, on trials when observers view the dynamic video clip, they are being presented with information about the speech act in the visual modality and this information allows the observers an additional channel of information with which to directly perceive the speech act. Speech acts are multi-modal by definition, and observers know that the visual-only prime has lawful consequences for the possible speech acts, thus providing information that may be absent in the noise-filtered auditory target.

A second possible explanation holds that the critical property of the relationship between the visual prime and the auditory target is that they share linguistic (or lexical) information, and not that they are from the same physical event in the world. Under this view, the visual prime activates sublexical representations which may also be activated by the auditory target. When these two stimuli contain the same information, they activate representations that enable accurate identification of the auditory signal. Here we are agnostic as to whether these representations are either linked representations of modality-specific information or whether an amodal representational level is activated. Crucially, this second hypothetical mechanism holds that the priming benefit should be maintained even when the visual prime

and the auditory signal come from different speakers, whereas the claim that the priming effect comes from the integration of auditory and visual information does not predict a priming benefit when the auditory and visual information has different sources.

In future work, we plan on replicating this result with a larger population of participants. Additionally, we will use the cross-modal priming paradigm to address these two hypotheses discussed above. In particular, we will investigate the extent to which the word identification accuracy benefit from the visual-only prime comes from the match in lexical information in the visual prime and auditory target as opposed to the match in the entire audiovisual event despite the separation of these two components along a temporal dimension. This issue will be explored by presenting observers with different speakers in the two modalities: speaker A will be seen producing a word and speaker B will be heard in the auditory stimulus component of the trial. If the priming effects observed here are due entirely to the integration of audio-visual information, no priming benefit should be seen in this condition. On the other hand, if the priming effects observed in this pilot study arise solely from the match in lexical information in the two stimulus events, the priming effect should be replicated when the auditory and visual stimuli are produced by different speakers. A third possibility is that a priming effect will be observed, but that the magnitude of the effect will be attenuated when the visual and auditory stimuli are produced by different speakers. This result would suggest that the priming effect observed here relies on both lexical identity and the integration of audio-visual information, such that removing the latter factor diminishes the effect but does not make it disappear altogether.

A second research question will explore the nature of the visual stimuli that can engender this priming effect. The pilot study reported here employs full-face visual stimuli of a speaker producing the given lexical item. In future work, we plan to replace these full-face stimuli with point-light displays that present a relatively impoverished depiction of the speaking event, to determine whether the priming benefit observed here is also obtained when the prime stimulus is a degraded dynamic visual stimulus.

Summary

This study was carried out to determine whether presenting visual-only stimuli prior to auditory stimuli facilitates the recognition of spoken words in noise. The results of the study indicate that this type of cross-modal priming does occur, and may be a useful tool for investigating issues related to spoken word recognition in future work.

References

- Dodd, B., Oerlemans, M., & Robinson, R. (1989). Cross-modal effects in repetition priming: A comparison of lip-read graphic and heard stimuli. *Visible Language*, 22, 59-77.
- Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston, MA: Houghton Mifflin.
- Horii, Y., House, A.S., & Hughes, G.W. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, 49, 1849-1856.
- Lachs, L., & Pisoni, D.B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology*, 16, 159-187.
- Luce, P.A., & Pisoni, D.B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19, 1-36.
- Sheffert, S., Lachs, L., & Hernandez, L.R. (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.