Bimodal Displays Improve Speech Comprehension In Environments With Multiple Speakers

DARRELL S. RUDMANN

JASON S. McCARLEY

ARTHUR F. KRAMER

University of Illinois, Urbana-Champaign

Urbana, IL, USA

Requests for reprints: Darrell Rudmann, send e-mail to rudmann@uiuc.edu or write to 405 N. Mathews Ave., Beckman Institute, University of Illinois, Urbana, 61801, USA

Running head: Bimodal display augmentation with multiple speakers

ABSTRACT

Keywords: crossmodal speech perception, ventriloquism

Attending to a single voice when multiple voices are present is a challenging but common occurrence in military, industrial and aviation settings. An experiment was conducted to determine a) whether presenting a video display of the target speaker aided speech comprehension in an environment of two or more competing voices, and b) whether the "ventriloquism effect" (Jack & Thurlow, 1973) could be used to enhance speech comprehension, as found by Driver (1996), using more ecologically-valid stimuli than prior research. Participants listened for two target words from 12 two-minute videos of an actress reading from a novel while they simultaneously tried to ignore the voices of two to four different actresses reading from different portions of the same novel. Target-word detection declined as participants were required to ignore more distracting voices; however, this decline was reduced when participants could see the target speaker as compared to when they could not. Neither a signal-detection analysis of performance data nor a gaze-contingent analysis revealed a ventriloquism effect. Providing a video display of a speaker when competing voices are present improves speech comprehension, but the ventriloquism effect appears to be weak at best in naturalistic circumstances.

BIMODAL DISPLAYS IMPROVE SPEECH COMPREHENSION IN ENVIRONMENTS WITH MULTIPLE SPEAKERS

Bimodal speech perception

Some research examining the perception of speech has been unimodal; the visual and auditory aspects of speech comprehension are investigated separately. For example, researchers examining the visual aspects of speech perception, often referred to as "speechreading," explore what linguistic information an observer gains from a visual display of the speaker (MacDonald, Andersen, & Bachmann, 2000; Campbell & Massaro, 1997) and how the observer actively culls this information, examining the pattern of eye movements during speech (e.g., Lansing & McConkie, 1994). Auditory researchers have found that spatially segregating multiple speakers improves the speech comprehension of a target speaker, a phenomenon referred to as the "cocktail party effect." The spatial disparity of sound sources makes each source more distinguishable primarily by exploiting the human capacity for binaural hearing, which provides more auditory information than monaural hearing (Blauert, 1997, p. 393; also see review by Yost, 1997). The perceived location differences between speakers aids the listener's ability to segregate the voices into different streams, enhancing the listener's ability to attend to the desired voice and ignore distracting voices (Handel, 1989, p. 189), although increases in the number of competing speakers can still decrease comprehension, due to auditory load (Lee, 2001). The difference in the perceived locations of auditory streams primarily relies on a number of interaural differences in the acoustic wave, such as differences in sound-pressure levels and temporal cues at each ear (Middlebrooks & Green, 1991). A listener who is familiar with the speaker can also use knowledge of a speaker's speech habits and intonation as well as the message itself to help comprehension (Handel, 1989).

Bimodal research combines visual and auditory stimuli to investigate how the two perceptual systems influence each other. The increased complexity in design is often more similar to real-life speech perception. Simply providing a visual display of the speaker that is congruent with the display of speech improves comprehension, particularly as the level of background noise increases (e.g., MacLeod & Summerfield, 1987). The visual display complements the speech and adds an element of redundancy to the comprehension process, although precisely what the listener gains from watching the speaker is still unclear (MacDonald et al., 2000). In general, listeners tend to gaze mostly at the eyes and mouth of the speaker, and to focus on the lips more as the auditory display of the speech degrades (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Whether or not the benefit of providing a visual display of a target speaker remains when multiple, competing voices are present does not appear to have been tested prior to this study, despite the common real-world occurrence of listeners needing to comprehend one speaker among many.

Biases from intramodal incongruities

Researchers have also presented listeners with incongruities in the information between modalities to better understand speech perception. When shown a speaker whose lips do not match the syllables heard, for example, listeners report hearing syllables that are neither shown nor heard. Seeing a speaker's lips pronounce /pa-pa/ while hearing /na-na/ results in experiencing /ma-ma/, a phenomenon known as the "McGurk effect" (McGurk & McDonald, 1976).

Disparities in the locations of the source of the visual and auditory displays have been examined as well. When a voice is presented away from the visual image of a target speaker, listeners experience the voice as emanating from the image (Jack & Thurlow, 1973). The

phenomenological experience, known as the "ventriloquism effect," is one of experiencing the target speaker's voice emanating from a non-existent loudspeaker located near the display. The improved ability to comprehend the speaker appears to be the result of exploiting the comparative difference in acuity of the human visual and auditory systems by relying on the stronger spatial visual information to override auditory localization cues.

Speech comprehension can be increased by exploiting the ventriloquism effect when two speech streams are present. Participants in a study by Driver (1996) listened to two different syncopated streams of nonsense syllables from the same individual from one sound source. When they watched a video display of the individual from a spatial location that was spatially disparate from the source of the sound, their ability to hear the target nonsense syllables increased by about 15% over when the video and audio information spatially coincided. Thus tracking or shadowing the speech of a target speaker while ignoring another distracting speaker appears to be easier when the visual display of the target speaker is displayed some distance away from the sound source, rather than nearby.

Present study

Should such a spatial separation between the auditory and visual displays of speech increase comprehension, the ventriloquism effect could be applied in real-world situations, particularly in circumstances where a listener must filter a relevant stream of speech from among multiple, competing voices, a task often required of air traffic controllers and military personnel. But such a benefit needs to be evaluated beyond Driver (1996), which used somewhat artificial laboratory conditions to demonstrate the effect by presenting two tracks of timed nonsense syllables, both recorded from the same individual. Driver's (1996) approach illustrated that the effect can work under conditions of multiple speakers but has low ecological validity since one

person does not usually generate two streams of simultaneous speech. Additionally, nonsense syllables are normally much less intelligible than normal speech (Kryter, 1972).

This paper describes a multimodal examination of speech comprehension. First, we tested whether the improved intelligibility derived from being able to see a speaker is retained in the context of multiple, competing voices. Second, whether comprehension in such settings can be improved by exploiting the ventriloquism effect is tested in a setting that is more ecologically valid than prior research by using multiple voices with more than two competing at any one time and by using semantically coherent and syntactically correct speech.

METHOD

Participants

Twenty-four undergraduate students at the University of Illinois were paid for their participation. Their ages ranged from 18 to 38 years, with an average of 21.38. Sixty-seven percent were female. All reported normal hearing and vision abilities, and were native English speakers. All were able to hold normal conversations with the experimenter prior to and after data collection.

Materials

Six actresses from the University of Illinois' Theatre Department were videotaped reading different excerpts from a novel. All were native English speakers with training in enunciation. One actress was selected as the target speaker for the experiment; her recording was digitized into QuickTime format with 320 X 240 resolution and 16-bit color depth at 24 frames per second. This frame rate is over the level found to contain enough temporal information to allow for accurate shadowing of an auditory stimuli containing speech (Viktovitch & Barber, 1994). These digital videos represent blocks of trials.

The recordings of the other actresses were used as additional, distracting voices in the experiment. The audio for the target speaker and the other actresses was digitized in stereo at 16 Hz in AIFF. All digitized movies and audio files were two minutes in length. The amplitude of each digitized recording was adjusted to such that the loudness was subjectively equal, ranging from about 58 to 70 dBA.

Apparatus

An SGI Onyx controlled the experiment, back-projecting the videos onto a large screen to create a 0.91-meter wide by 0.61-meter tall image. Participants sat in a chair about three feet in front of the video screen. Audio was presented from one of two loudspeakers, one, the centered sound source, located directly in front of and below the video image, and the other, the displaced sound source, located about 0.91 meters or 30° of visual angle to the right or left of where participants sat. The loudspeakers were matched units. For half of the subjects the displaced loudspeaker was on the right while for the other half of the subjects the displaced loudspeaker was on the left. To nullify any unknown differences between the loudspeaker units themselves, they were exchanged in their role as centered or displaced loudspeaker from one participant to the next. The amplitude from either loudspeaker was approximately equal at the location of the chair. Button responses were collected by the SGI as well. Equipment in the room created a background noise level of 49 dBA.

Procedure

The participants were instructed to, during a block, watch and listen to the target speaker, and to press a button whenever they heard the target speaker say the word "and" and to press a second button whenever they heard the target speaker say the word "the." These words were chosen for their common usage in written and spoken language. The "and" target word was

present from zero to 10 times in each of the target speaker videos, with a mean of 2.7 times. The "the" target word was present from zero to 25 times in each of the target speaker videos, with a mean of 7.6 times. A total of about 100 words were presented in any one block by a speaker.

Two, three, or four distracting voices were heard simultaneously with the target speaker, following a five-second lead-in with only the target voice. When multiple voices were heard, the overall amplitude of word presentation was quite high. Participants were told to ignore the other distracting voices as much as possible, and to use the video image to help follow the target speaker. All audio was displayed from either the centered or displaced loudspeaker during a video. Participants were told that the audio might come from either or both of the loudspeakers at random and to simply focus on the word detection task. Accuracy was emphasized.

Prior to experimental testing, participants were given six practice blocks to familiarize themselves with the target speaker and task. Two blocks were shown for each level of distracting voices sequentially, one with the video on, the other with the video off. Each loudspeaker was used for half of the practice blocks.

Participants were given 12 experimental blocks showing the target speaker, with two, three, or four distracting voices presented simultaneously from the same loudspeaker, either centered or displaced. To aid the participants, the distracting voices were always presented after a five second delay from the start of a block. To establish an auditory-only performance baseline for comparison, half of the blocks were presented without a visual display of the target speaker. Video-on and video-off blocks were interleaved. Stimulus material (i.e. the 12 different two minute video/audio blocks) and experimental conditions were counterbalanced across subjects. A head-mounted eye-tracker monitored eye movements during test blocks for 21 of the subjects at a rate of 60 Hz.

Design

The variables create a 2 X 3 X 2 X 2 within-subjects design: Location of Sound Source {Centered, Displaced}, Number of Distracting Voices {2, 3, 4}, Video Display {On, Off}, and Target Word {"AND," THE"}. In all non-practice blocks, all conditions and videos were counterbalanced across participants in a Latin-square design.

RESULTS

A response was recorded as a correct response or hit if the button corresponding to the spoken target word was pressed within a three-second window from the beginning of the spoken word. A response was recorded as a false alarm if the button pressed did not correspond to a correct target word as spoken by the target speaker within a three-second window. Hit rates for individual blocks of trials ranged from zero to 1, with a mean of 0.39 (*SD* = 0.23). The participants had a mean hit rate of 0.387, (*SD* = 0.114). False alarm rates, the number of false alarms divided by the number of seconds during which a false alarm could have occurred for a given block, ranged from zero to 0.11, with a mean of 0.02 (*SD* = 0.02).

Az analysis of variance

A four-way within-subjects analysis of variance was performed to test for reliable differences between the conditions. This analysis was performed on *Az* scores, a nonparametric measure of sensitivity (Wickens, 2002) and ß, a measure of response bias. The results presented below for *Az* scores were also statistically reliable when the analysis was conducted using *A′* and *d′* scores.

The participants' response bias was on average fairly conservative, with an average ß of 1.489. A two-way interaction for the response bias was found between video display and the number of distracting voices, $F(2, 46) = 6.050$, $p = .005$, $MSE = .056$. The participants became

slightly less conservative with four distracting voices than three or two (M_{two} = 1.539, M_{three} = 1.481, M_{four} = 1.325), $T(285) = -0.155$, $p < .0001$, and two, $T(285) = -0.203$, $p = .002$.

The analysis of *Az* scores indicated that the participants tracked the target words better when they were able to see the target speaker than when they could not ($M_{video\ on}$ = 0.887, SD = .138; $M_{video\ off}$ = 0.817, SD = .206), $F(1, 23) = 24.831$, $p < .0001$, $MSE = .027$. As distracting voices were added to the task, the participants' ability to track the target words steadily worsened (M_{two} = 0.894, SD = .147; M_{three} = 0.864, SD = .162; M_{four} voices = 0.798; SD = .209), $F(2, 46) = 17.408$, $p < .0001$, $MSE = .024$.

Performance benefits due to video display. In a two-way interaction, the benefit of providing a video display of a target speaker increased with additional distracting voices, $F(2, 46) = 4.013$, $p = .025$, $MSE = .027$. As Figure 1 shows, the participants' ability to track target words declined regardless of whether a video display of the target speaker was provided or not; however, when the video display was not provided, performance declined at a faster rate. The mean *Az* score between video display on or off with two distracting voices was not reliably different, $T(190) = -1.544$, but was reliably higher with the video display on at three distracting voices, $T(190) = -2.150$, $p < .033$, and at four distracting voices, $T(190) = -4.318$, $p < .0001$.

Effects due to the location of sound source. The predicted pattern of improved performance when the target speaker could be seen and the location of the sound source was displaced (i.e. ventriloquism effect) was not apparent in the interaction between video display and sound source location, $F(1, 23) = 2.899$, $p = .102$, $MSE = .032$. When the video display was on, the location of the sound source did not cause a reliable change in performance, $T(286) = -0.487$.

Eye movement analysis

Because of the lack of evidence for a ventriloquism effect in the initial analysis, the eye movement data were examined to determine whether the participants who focused on the lips of the target speaker would show improved performance when the displaced sound source location was used. Perhaps the ventriloquism effect in a multi-speaker environment occurs only when participants focus on the lips of the target speaker: the task presented by Driver (1996) was so difficult that the participants had to focus on the lips of the speaker to be successful. In this experiment, the participants were allowed to look anywhere on the speaker's face and many did not exclusively focus on the speaker's lips. In the remaining analyses, only data from conditions when the video display was on are examined.

General eye movement patterns. The video display was segmented into three regions: the target speaker's lips, face, and the screen. This segmentation allowed an analysis of responses made while participants were focused on the lips of the speaker, as described in the next section. Six of the participants gazed in the area of the speaker's lips more than the other areas; fourteen other participants spent more time gazing around other areas of the speaker's face than the lips (the remaining participant's data shows the participant as looking at the screen but not at the speaker). Across all participants' data, 22.4% of recorded gaze time was on the speaker's lips, 34.7% was on the speaker's face but not lips, and 13.3% was on the screen but not the speaker's face (the remaining time was off the screen or eye blinks).

Gaze-contingent analysis. Because of the possibility that ventriloquism is only effective when a participant is observing lips, mean *Az* scores were analyzed based the participants' region of gaze. A hit was defined as described above, with a button press corresponding to the correct target word as spoken by the target speaker in a target video only within a 3-second window

from the beginning of the spoken word, but modified to include the additional criterion that the participant had spent at least half of a second looking in the region of the speaker's lips one second prior to or during the presentation of a target word. False alarms were redefined in the same manner.

A 2 X 3 X 2 within-subjects analysis of variance (Location of Sound Source {Centered, Displaced}, Number of Distracting Voices {2, 3, 4}, and Target Word {"AND," THE"}) showed a reliable degradation of performance when the number of distracting voices increased (M_{two} = .862; M_{three} = .817; M_{four} = .761), $F(2, 20) = 7.843$, $p = .003$, MSE = .016. Performance under the two distracting voices condition was not reliably better than the three distracting voices condition, but performance under three distracting voices was reliably better than with four distracting voices, $T(10) = 2.614$, $p = .026$. None of the other conditions or their interactions produced reliably different mean *Az* scores.

DISCUSSION

In an environment where an individual must selectively listen to only one voice out of many, presenting a video display of the selected speaker aids comprehension. This experiment demonstrated that as additional competing voices are presented to the listener, a video display of the desired speaker aids comprehension by lessening the degradation of comprehension (shown in Figure 1), a benefit that was apparent in the gaze-contingent analysis as well. The eye movement contingent analysis suggests that the performance enhancement from providing a video display of the speaker does not appear to require that the listener maintain a fixation on the lips of the speaker, but in the vicinity of the face.

The second goal of the study, an examination of whether the ventriloquism effect can be exploited to further improve comprehension in a realistic setting, was unsuccessful. A number of

possible explanations exist for this. Perhaps the strongest possibility is that the combination of three voices or more, each exhibiting their own natural vocal characteristics and each carrying a unique contextual message, was enough to nullify the ventriloquism bias found in prior research. The presence of multiple voices makes the location of the sound source clear to the listener very quickly, and coming to believe that one voice out of the set is emanating from the video display may not be natural in such a situation.

This explanation assumes that there is a cognitive component to the ventriloquism effect, as implied by Jack and Thurlow (1973) and Thurlow and Jack (1973), who tested the ventriloquism effect by covering the loudspeakers to enhance the likelihood that participants would mislocate the displaced sound towards the visual stimuli. In contrast, more recent research on the ventriloquism effect has assumed the effect to be preattentive (Driver, 1996; Bertleson & Radeau, 1976), and has frequently explored the effect using tones and flashes of light rather than speech. It is not clear whether the speech-based ventriloquism phenomenon described by earlier researchers is the same phenomenon investigated in preattentive, nonspeech ventriloquism research. If the same phenomenon is being investigated in both lines of research, then preattentive ventriloquism might have a limited spatial range within which it can operate, or might only work with simple stimuli in laboratory settings. If not, then for ventriloquism to have been achieved in this experiment, the loudspeakers would have needed to be covered and brought closer to the display, and the audio signal possibly reduced in amplitude, in order to enhance the likelihood of spatial confusion. Such changes would probably affect overall detection performance as well.

This experiment tested the potential benefit of providing a video display and of using ventriloquism under conditions of multiple, competing simultaneous speakers in a realistic and

ecologically valid fashion. For more realism, participants would be required to follow the message communicated by the target speaker, rather than tracking monosyllabic structural words contained in the message. Such an advance in design would more closely resemble the task of an air traffic controller or other military personnel trying to listen to one voice among many.

A number of interesting questions remain about bimodal interactions in speech comprehension. For example, can an artificial model or avatar of a speaker's face provide the same comprehension benefit? Three-dimensional modeling of a speaker's head combined with modern algorithms to move the model in conjunction with speech may provide the same benefit as a human talking head (Massaro, 1998). This seems possible giving the finding that spatially degrading the visual information from a speaker's face to fairly course levels still allows for the McGurk effect to occur reliably (MacDonald et al., 2000) and that point-light displays of the facial movements of a speaker help speech comprehension (Rosenblum & Saldaña, 1996). In addition, avatars can potentially be used to systematically examine the kinds of visual information that are most useful in enhancing comprehension, such as whether moving lips are sufficient or whether facial expressions are also necessary. While using avatars would degrade visual information available from a speaker, specially-created noise that is based on speech signals would test how visual information could improve a degraded auditory component of speech. It seems likely that a comparison of the two kinds of degradation, auditory and visual, would find the auditory component to be more critical to speech.

The level of benefit provided by visual displays of a speaker when the background noise level is higher or more realistic than tested here is also of interest. In this study the only distracting audio was in the form of competing voices, when in many real-world situations, numerous other kinds of sounds may be presented to a listener. Also, people frequently try to

accomplish some other task while listening as well, such as driving or flying. This study does not assess the presentation of video information of a speaker on dual-task performance, when the second task is not speech perception.

Applied Implications . The results of this experiment imply that when there are multiple, competing voices for a listener's attention, speech comprehension will be improved if a video display of the desired speaker can be provided. The need for a video display grows as more voices increase the difficulty of speech perception. While this study did not manipulate background noise or task difficulty, it is reasonable to assume that a video display will benefit listeners in a wide variety of noisy environments and tasks. Thus, providing a video display of a speaker may produce a comprehension benefit for cell phone users, pilots, air traffic controllers, and soldiers in situations where multiple voices are present. However, the ventriloquism effect does not appear to be an aid in comprehension in environments with a lot of auditory clutter.

REFERENCES

Bertelson, P., & Radeau, M. (1976). Ventriloquism, sensory interaction, and response bias: Remarks on the paper by Choe, Welch, Gilford and Juola. *Perception & Psychophysics, 19*, 531-535.

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization* (Revised ed.). Cambridge: The MIT Press.

Campbell, C. S., & Massaro, D. W. (1997). Perception of visible speech: Influence of spatial quantization. *Perception, 26,* 626-644.

Driver, J. (1996). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381,* 66-68.

Handel, S. (1989). *Listening: An introduction to the perception of auditory events.* Cambridge: MIT Press.

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual & Motor Skills, 37,* 967-979.

Kryter, K.D. (1972). Speech communications. In H.P. Van Cott & R.G. Kinkade (Eds.), *Human engineering guide to system design* , Washington, D.C.: U.S. Government Printing Office.

Lansing, C. R., & McConkie, G. (1994). A new method for speech-reading research: Tracking observer's eye movements. *Journal of the Academy of Rehabilative Audiology, 27,* 25-43.

Lee, M. D. (2001). Multichannel auditory search: Toward understanding control processes in polychotic auditory listening, *Human Factors, 43,* 328-342.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception, 29,* 1155-1168.

MacLeod, A., & Summerfield, Q. (1987). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology, 24,* 29-43.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 747-748.

Middlebrooks, J. C., & Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology , 42,* 135-159.

Proctor, R. W., & Van Zandt, T. (1994). *Human factors in simple and complex systems.* Boston: Allyn and Bacon.

Rosenblum, L. D, & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance, 22,* 318-331.

Thurlow, W. R., & Jack, C. E. (1973). Certain determinants of the "ventriloquism effect." *Perceptual & Motor Skills, 36,* 1171-1184.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceives during audiovisual speech perception. *Perception & Psychophysics, 60 ,* 926-940.

Viktovitch, M., & Barber, P. (1994). Effect of video frame rate on subjects' ability to shadow one of two competing verbal passages. *Journal of Speech & Hearing Research, 37,* 1204-1210.

Wickens, T. D. (2002). *Elementary signal detection theory.* Oxford.

Yost, W. A. (1997). The cocktail party problem: Forty years later. In R. H. Gilkey, & T. R. Anderson (Eds.), *Binaural and Spatial Hearing in Real and Virtual Environments* (pp. 329-348). Lawrence Erlbaum.

FIGURE CAPTIONS

Figure 1. Mean *Az* scores as a function of whether the video of the speaker was presented or not as a function of the number of distracting voices. The error bars represent twice the standard error of the plotted means.

Acknowledgments

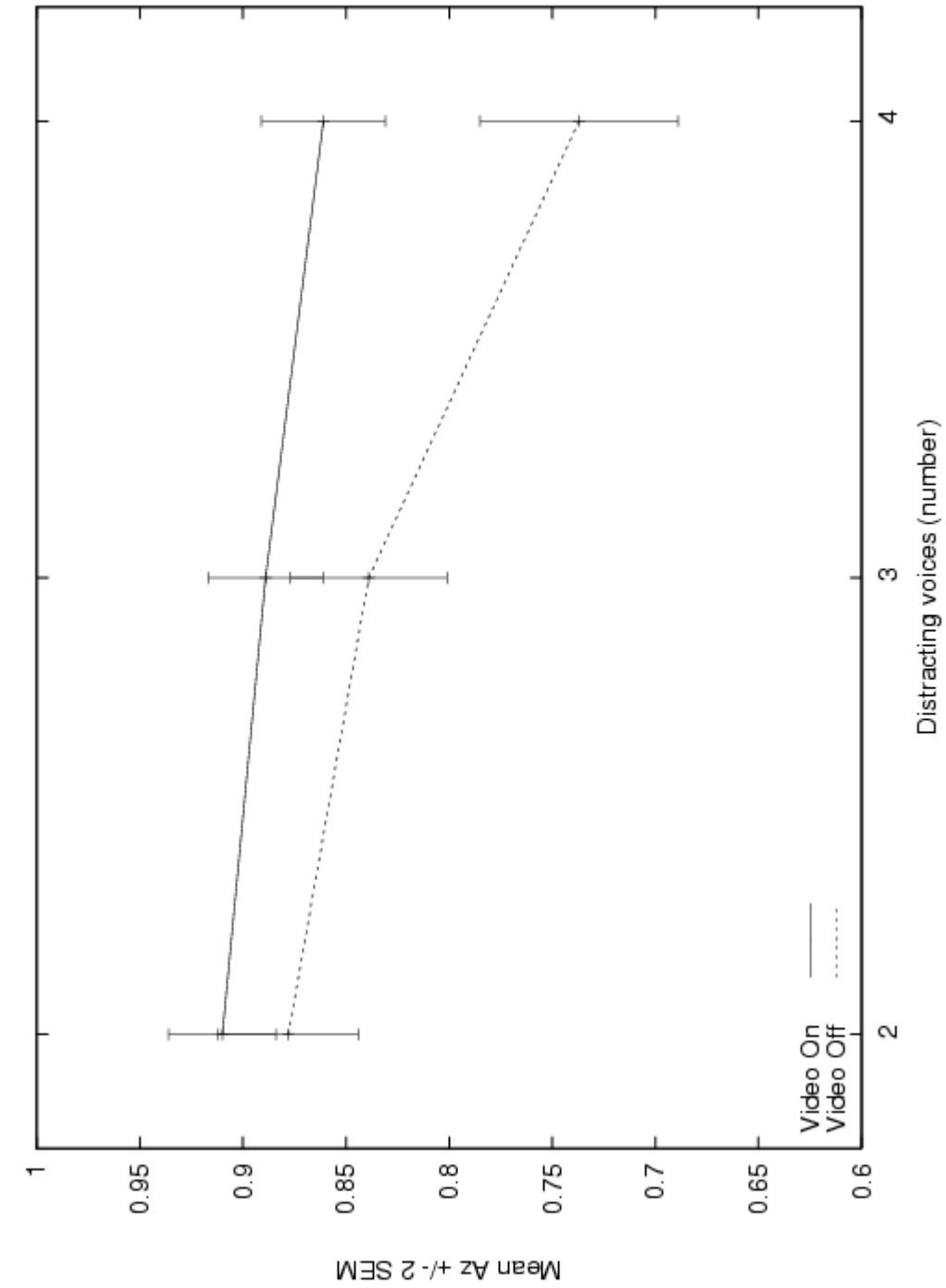
The research described in this manuscript was supported by grants from the Army Research Laboratory and the National Science Foundation.

BIOGRAPHIES

Darrell S. Rudmann, University of Illinois, M.A. (Psychology, 1997, California State University, Long Beach)

Jason S. McCarley, University of Illinois, Ph.D. (Experimental Psychology, 1997, University of Louisville)

Arthur F. Kramer, University of Illinois, Ph.D. (Psychology, 1984, University of Illinois)


Mean Az +/- 2 SEM
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
2
3
4
Distracting voices (number)
Video On
Video Off