Spatial Synergies between Auditory and Visual Attention

Jon Driver

Charles J. Spence

University of Cambridge

Revision of chapter submitted for publication in C. Umilta & M. Moscovitch (Eds.), *Attention and Performance: Vol. XV*. Hillsdale, NJ: Erlbaum.

Correspondence to: Jon Driver, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

Abstract

Three experiments examined spatial links between endogenous attention in vision and audition. In Experiment 1 subjects were presented auditorily with two verbal messages, one from either side of their midline. They had to repeat one message, for which they were also given lip-read information in some conditions. The lip-read information produced a larger improvement in performance when presented on the same side as the target sounds rather than the opposite side, suggesting a difficulty in attending to different locations in the two modalities. This difficulty cannot be attributed to the direction of gaze, as there was no effect of passively fixating meaningless lip-movements on the same versus opposite side as the target sounds. A similar (albeit reduced) difficulty in attending to different locations in the two modalities was observed in Experiment 2, where the auditory and visual tasks were unrelated. The auditory task was shadowing as before, while the visual task was monitoring for a specified target in a stream of visual characters appearing successively at a single location. A decrement in shadowing was observed when the monitored visual events were in a different location to the target sounds. Again, there was no effect of the direction of passive fixation. A final experiment found that the cost of attending to different locations in the two modalities can be observed in the conventional dual-task situation, i.e. with no distractor stimuli. These results demonstrate synergetic links between endogenous spatial attention in vision and audition.

Introduction

A substantial research effort has gone into examining the mechanisms of selective attention in audition and vision, as previous volumes of 'Attention and Performance' attest. The classic work of the 1950s concentrated on auditory attention, using the selective shadowing paradigm (e.g. Cherry, 1953; Broadbent, 1958). Many such studies found that subjects could select one of two verbal messages for report quite readily, provided that there was a physical distinction between them, with differences in location leading to particularly efficient selection. In the subsequent years, research has shifted towards investigations of the mechanisms of visual attention (see Kahneman & Treisman, 1984, and Neumann, van der Heijden & Allport, 1986, for historical perspectives), and many authors have proposed that space has special status in the control of visual attention as well as for auditory attention (e.g. Posner, 1980; Treisman & Gelade, 1980; Tsal & Lavie, 1988).

Previous research has tended to focus on selection within a single modaltity at a time. However, in the real world we are typically faced with multi-modal stimulation, and thus require selective mechanisms to operate simultaneously in several modalities. To take the textbook example of selective attention (listening to one conversation at a noisy cocktail party), the relevant information can be selected both auditorily *and* visually, by lip-read information. Indeed, shadowing performance is improved when the actors speaking the relevant and irrelevant messages can be seen by the subject (Reisberg, 1978). Assuming that the purpose of selective attention is primarily to select relevant *objects* from cluttered and noisy scenes for the selective control of action (Allport, 1987; Driver & Baylis, 1989; Duncan, 1984; Tipper, Driver & Weaver, 1991), these crossmodal considerations lead to an immediate computational problem. Selection has to be coordinated across the modalities, so that attention is directed to sights and sounds emanating from the same distal object. However, this is potentially problematic, given the existence of modality-specific sensory systems which do not have corresponding spatiotopic organisations at input. For instance, there is no auditory equivalent to the inherent retinotopic organization in vision and therefore no direct spatial correspondence at input between sights and sounds coming from the same object.

Two fundamental questions arise; to what extent can attentional mechanisms operate independently in distinct unimodal systems, and how is attention coordinated across modalities despite their different input organizations? In the present series of studies, we approach these questions by examining whether there are spatial links between auditory and visual attention, such that it is relatively difficult to attend to one location visually while attending to another auditorily.

Existing research on crossmodal links in attention has primarily examined *overt* orienting responses, involving shifts in the direction of peripheral receptors. The primate visual system tends to rapidly foveate the spatial location of unexpected or salient auditory events (e.g. Thompson & Masterton, 1978; Whittington, Hepp-Raymond & Flood, 1981) suggesting that overt orienting mechanisms have strong audiovisual links. Does this also apply for *covert* orienting, i.e. shifts in attention without corresponding shifts in peripheral receptors? Covert orienting has been widely studied in the now familiar cueing paradigm (e.g. Posner, 1980). Although most of this work has examined vision, the distinction between overt and covert orienting applies to audition as well (Spence & Driver, in prep.) since auditory attention can be shifted with or without corresponding shifts of the head or pinna. In the case of vision, it is well established that visual detection can be facilitated if the location of a target is precued even if no eye-movement is made towards the cued location (e.g. Muller & Humphreys, 1991). This benefit is attributed to covert orienting, and can be found following cues such as uninformative peripheral flashes at the target location which do not predict the likely target locus, or following informative cues such as an arrow at fixation pointing in the direction of the likely target locus (Posner, 1980). Uninformative peripheral cues are thought to engage different covert orienting systems to informative symbolic cues. The former produce exogenous orienting (i.e. bottom-up control of attention by salient peripheral events), whereas informative cues produce endogeneous orienting (i.e. strategic, voluntary shifts as a result of the subject's expectancies). Several qualitative differences between these forms of covert orienting have now been demonstrated (e.g. Jonides, 1981; Muller & Rabbitt, 1989; Rafal, Henik, & Smith, 1992), suggesting the involvement of distinct mechanisms.

Such findings imply that the nature of crossmodal links in covert orienting should be examined separately for exogeneous and endogenous mechanisms. Previous work using the cueing paradigm has demonstrated clear crossmodal links in the exogenous case. Cueing effects are found when uninformative peripheral auditory events are presented prior to visual targets (Klein, 1987; Farah, Wong, Monheit, & Morrow,1989), suggesting either that exogenous orienting systems are polymodal, or that when auditory attention is exogenously drawn to a location, a visual orienting system also shifts towards that locus.

Current evidence is less clear on the nature of crossmodal links in *endogenous* orienting. Crossmodal cueing effects for targets in one modality following informative peripheral cues in another modality have been demonstrated (e.g. Buchtel & Butter, 1988, between auditory and visual stimuli; Butter, Buchtel & Santucci, 1989, between tactile and visual stimuli) but these findings are difficult to interpret. Since the cues were informative, endogenous orienting mechanisms should have been involved. However, the critical issue is whether endogenous orienting in one modality has effects on another modality, and unfortunately crossmodal effects from informative cues do not establish that this is the case. They simply show that a signal in one modality (e.g. a tone) can instruct subjects about the likely locus of a target in another modality (e.g. a flash), in a manner analogous to the interpretation of a central visual arrow prior to a visual target. After such an interpretation, the endogenous orienting may take place entirely *within* the target modality (vision in the chosen example with an informative tone as the cue). For this reason, the cueing paradigm is unsatisfactory for examining crossmodal links in endogenous attention, i.e. when the direction of attention is strategic rather than the automatic outcome of salient peripheral events.

Accordingly, we devised a novel technique for examining crossmodal links in endogenous attention. In essence, we presented subjects with visual events on either side of their midline, and also with auditory events on either side of their midline. We required them to attend to one side while ignoring the other side in each modality. The attended side could be the same or different for the two modalities. If there are crossmodal links in the endogenous direction of attention, it may be difficult to attend to different sides in the two modalities. The specified target and distractor sides were quite arbitrary (i.e. the target stimuli were no more salient in bottom-up terms than the distractor stimuli). This ensured that endogenous, strategic mechanisms of selection were required to fulfill the task requirements, rather than exogenous, stimulus-driven mechanisms.

Experiment 1

Our subjects heard two verbal messages simultaneously, one presented from a loudspeaker to their left, the other to their right. One message had to be shadowed, the other ignored. The side of the target message was arbitrarily specified by the experimenter, so that endogenous selection mechanisms were involved. Each message was composed of a series of triplets of unrelated words spoken in rapid succession. Both messages were spoken by the same person. The task was to repeat each triplet from the target auditory location in the interval between successive triplets. Subjects were also presented with visual information. In the active visual tasks, this comprised synchronous lip-read information for the target verbal message, which had to be fixated. In the passive visual tasks, meaningless lip-movements (chewing) were presented for fixation. The main manipulation was whether the visual information was presented on the same side as the target sounds, or on the side of the distractor sounds. If there is a difficulty in attending to different locations in vision and audition simultaneously, shadowing performance should be worse when the lip-read information was presented on the opposite side to the target sounds. If any such difficulty simply reflects the direction of fixation, it should also be found when subjects fixated meaningless lip-movements on the opposite side to the target sounds.

Subjects

The subjects were 16 unpaid volunteers; 13 were undergraduates, and 3 were research staff in the Psychology department. All were naive as to the purpose of the experiment. The mean age was 22 years, and the range 19-35 years. All reported normal hearing and normal or corrected vision. Each participated for around 75 minutes.

Apparatus and Materials

Two monitors were placed 170 cm apart (centre-to-centre), symmetrically to the right or left in front of the subject. Each was 155 cm from the subject, who was seated at a table. Thus, the angle between the monitors was 57^{o}. One monitor was a Sony PVM201OQM, the other a Panasonic BTD2000PSN. Immediately below each monitor was a KEF SP1146 loudspeaker, and at the midpoint between these two loudspeakers (directly in front of the subject) there was a third loudspeaker which produced continuous white-noise. There were four videoplayers.

The first 6 subjects used a chinrest and were verbally reminded several times during the experiment not to turn their heads towards either monitor or loudspeaker. Although they appeared to comply with this request, the remaining subjects used a chin rest which also clamped their temples so that head-movements were completely precluded. Omitting the initial six subjects does not affect the pattern of results. Subjects' eyes were level with the loudspeaker cones. The direction of fixation was monitored by the experimenter in the first 14 subjects; for the final 2 subjects, eye-movements were also recorded by video-cameras, one above each monitor.

The materials were video-recordings of the second author reading two-syllable words presented to him off-camera. A full-frontal view of his head was shown continuously. During the recording, he attempted to pronounce the words at a regular rate by means of computerized clicks played to him over headphones. The 774 different words were all two-syllable nouns, with stress on the initial syllable. Two tapes of the second author reading 387 words were recorded, with no repetitions within or between tapes. On each tape, the words were grouped into lists, which themselves consisted of several triplets of words. The triplets were read aloud at at rate of approximately 1.3s per triplet. There were 8 experimental lists on each tape, each consisting of 14 triplets, with 4 s intervals between successive triplets. The interval between each list was 60 s, with a longer interval of 240 s after four experimental lists. Each video began with two practice lists, the first containing 9 triplets, and the second 8 triplets.

These two tapes were played back simultaneously to provide concurrent pairs of different audiovisual words. Each tape had a visual frame-counter at the top left. Synchrony of the tapes was achieved by starting the tapes simultaneously with an editor at predetermined frames which were judged to maximise phenomenal simultaneity when used as starting positions. Thus, while the synchrony of concurrent word pairs was approximate, and fluctuated with the second author's skill in maintaining a regular reading rate during the video-recording, it was the same for each subject. Concurrent pairs of words were matched as far as possible on Brown (1984) spoken frequency, and on number of letters. In addition there were two different 25 minute video-recordings which both showed a full-frontal view of the second author vigorously chewing gum without opening his mouth. These tapes had no soundtrack. While their use may seem somewhat bizarre, they were intended to provide transient visual information within the same face as for the word tapes, but without yielding any phonological information via lip-reading.

Each of the four videotapes remained in one videoplayer throughout the experiment. The four players were connected to the two monitors such that the four tapes could be run continuously and simultaneously, with switches determining which of the four visual channels appeared on any monitor. Each of the two spoken soundtracks went separately to a different lateralized loudspeaker and were presented from this loudspeaker throughout the session. The spoken words were presented through either lateralized loudspeaker at approximately 60 dB(A) on average, with a peak of around 64 dB, as measured from the subject's position. White-noise was presented at about 56 dB centrally. The same level was used for all subjects. It was intended to bring performance in naming the target words below ceiling, so that any benefit from the additional presence of relevant lip-read information was likely to be observed. The talking and chewing heads subtended about 11.1^{o} of visual angle vertically and about 6.6^{o} horizontally on the monitors. The lips appeared about 8.1^{o} above the centre of the loudspeaker positioned below the monitor.

Design

There were three within-subject factors. The first was whether the auditory target words appeared on the left or right. The second was whether meaningful lip-read information was provided or not (Speaking-lips versus Chewing-lips). In the Speaking-lips conditions, the visual channels from both tapes of the spoken words were shown on different monitors. In the Chewing-lips conditions, the two tapes of chewing were presented visually. The third within-subjects factor was whether the visual information which subjects had to fixate was presented on the Same side as the auditory words they had to repeat or on the Opposite side. The Speaking/Chewing and Same/Opposite factors were crossed to yield the four conditions described below, which were then duplicated for auditory target words appearing on the left versus right:

Same Speaking-lips; subjects had to name the words coming from one loudspeaker while fixating the monitor on the same side which showed lip-read information for the target words. The visual and auditory channels for the distractor words were both presented on the opposite side.

Opposite Speaking-lips; subjects again had to name the words coming from one loudspeaker (e.g. the left). However, the lip-read information for these words now appeared on the monitor at the opposite side (e.g. the right), which subjects fixated. The audio channel for the distractor words came from this opposite side (e.g. the right), while the visual channel for the distractor words came from the same side as the target sounds (e.g. the left).

Same Chewing-lips; subjects again had to name the words coming from one loudspeaker. Distractor words were presented from the other loudspeaker, and the two monitors showed different videos of chewing. Subjects had to fixate the chewing lips on the side of the *target* sounds.

Opposite Chewing-lips; as for Same Chewing-lips, except that subjects had to fixate the chewing lips on the side of the *distractor* sounds.

If there is any difficulty in attending to different locations in vision and audition, shadowing performance should be worse in the Opposite conditions. This could arise either because of a difficulty in *selecting* information from different locations in distinct modalities, or because of a difficulty in *rejecting* distractor information in one modality from a location which is relevant for another modality, (or as a result of both difficulties). Any effect which is due to the direction of fixation alone should be found equally for the

Speaking-lips and Chewing-lips conditions, provided subjects adhere to fixation instructions.

The contrast between Speaking-lips and Chewing-lips conditions also provides a measure of any benefit provided by having visual lip-read information in addition to auditory information for the target words.

Procedure

The experiment was conducted under dim illumination to preclude reflections on the monitors. The experimenter started the two chewing tapes, and then initiated the two word-tapes simultaneously from their predetermined starting frames. The four tapes then ran continuously (except where stated) so that words were presented at the rates given in the Materials section. The subject had to name as many words as possible from each triplet on the specified channel in the interval between successive triplets. A session began with practice, comprising 102 pairs of concurrent words. The practice materials were the first 17 triplets of words from each tape, run through twice. The first triplet was used for demonstration, and then each subject had blocked practice giving equal experience in the four conditions. The remaining 8 lists of 14 triplets were then run through twice to provide the experimental data. The condition was changed after every 14 triplets. The order of conditions was different for each of the first 8 subjects, with every condition appearing equally often in each serial position overall. Whether the target sounds were on the left or right alternated every 14 triplets. Across subjects, each word was equally likely to appear as a target or distractor in each condition overall. On the second run through the experimental materials, each subject had to repeat the words that had previously been irrelevant for them.

The last 8 subjects were run as for the first 8, except that all equipment on the left or right was swapped, i.e. monitors, loudspeakers and videoplayers. The videotapes stayed in the exchanged videoplayers so that their soundtracks changed side of presentation.

Between each list of 14 triplets, the experimenter explained the task for the next list. For example, in the Opposite Speaking-lips condition he might explain that the subject had to name the words from the left loudspeaker, but would see the lip-movements for these words on the right monitor, and had to fixate those lips.

The experimenter recorded subjects' responses, which were scored as correct, incorrect or intrusions (i.e. reports of the distractor words) without regard to the order of report for words within each triplet. Close approximations (e.g. derivational errors) were scored as incorrect.

Results and Discussion

Monitoring by the experimenter (and by means of video in the case of the final two subjects) suggested that the fixation instructions were followed. No differences in this respect were detected between the Speaking- and Chewing-lips conditions. While the eye-movement monitoring was admittedly somewhat crude, it should certainly be sufficient to detect whether subjects were fixating towards the appropriate monitor, given the substantial distances involved.

The percentage of correctly repeated target words for each subject in each condition were subjected to a three-way within-subject analysis of variance (ANOVA) with these factors: Left or Right auditory targets, Speaking or Chewing lips, and fixation at the Same or Opposite side as the target sounds. The ANOVA found a main effect of Left versus Right target sounds ($F(1,15)=14.8$, $p<.002$), with more accurate performance when auditory targets came from the subject's left (mean of 52.2%) rather than right (mean of 44.6%). This result is the opposite of the right-sided advantage normally found in selective listening for words (e.g. Kimura, 1961; Morais & Bertelson, 1975). Since the present effect may simply reflect the acoustics of the testing room, and there is no hint of any interaction with the other factors, we shall not consider the Left/Right factor further (neither in the subsequent analyses, nor in our subsequent experiments, where no main effects or interactions involving side of the auditory targets were observed).

Figure 1 about here

Pooling across the side of the auditory targets, the mean percentage of words named correctly for the remaining conditions are shown in Figure 1A (out of 168 words per subject for each condition). This figure suggests that Speaking-lips led to better word recognition, and that they produced a greater benefit when on the same side as the target sounds. This was confirmed by a two-way within-subject ANOVA, with Speaking- versus Chewing-lips as one factor, and the relation of fixation to the auditory target words as the other (Same versus Opposite side). There was a powerful effect of Speaking- versus Chewing-lips ($F(1,15)=321.2$, $p<.0001$) and an effect of Same versus Opposite side ($F(1,15)=49.2$, $p<.0001$). These factors interacted ($F(1,15)=58.5$, $p<.0001$). Wilcoxon matched-pairs signed-rank tests confirmed that while Same performance exceeded Opposite performance with Speaking-lips ($T=0$, $p<.001$) there was no effect of the direction of fixation with chewing-lips ($T=37.5$, n.s.), producing the observed interaction.

The substantial drop in performance for the Opposite Speaking-lips condition relative to the Same Speaking-lips condition suggests a difficulty in actively attending to visual lip-read information on one side while selecting auditory information from the other side. Note, however, that performance benefited from the presence of visual lip-read information relative to the corresponding Chewing-lips condition in both the Same and Opposite cases ($T=0$, $p<.001$ for each comparison). This suggests that while it was difficult to select one side visually and the other auditorily, it was nevertheless possible to some extent; otherwise no benefit from relevant lip-read information should be observed for the Opposite conditions.

Assuming that fixation patterns were indeed the same in the Chewing-lips and Speaking-lips conditions, the results suggest that the difficulty in attending to one location visually and another auditorily only applies for *active* endogenous visual attention, i.e. when the fixated visual events contribute to task performance. There was no detectable difficulty in passively fixating meaningless events on one side while attending auditorily to the other side in the Chewing-lips conditions.

Accurate reports of words from the *irrelevant* channel were occasionally observed, although these intrusion-error data were lost for one subject due to experimenter error. Pooling over the remaining 15 subjects, there were 27 such errors in total for Same Chewing-lips, 31 for Opposite Chewing-lips, 11 for Same Speaking-lips, and 45 for Opposite Speaking-lips. These data are consistent with the suggestion that it was difficult to ignore distractor auditory words when attending to visual lip-read information on their side in the Opposite Speaking-lips condition. Despite the small number of intrusions overall, a two-way ANOVA on the intrusion-error data showed an interaction between Same versus Opposite side and Chewing- versus Speaking-lips ($F(1,14)=8.3$, $p=.01$). There were more intrusion errors in the Opposite than the Same condition for Speaking-lips ($T=0$, $p<.01$), but not for Chewing-lips ($T=58.5$, n.s.), producing the interaction.

In summary, performance was poorer when the relevant lip-read information was on the opposite side to the target auditory information rather than on the same side. Intrusions from auditory distractors on the same side as the relevant lip-read information contributed to this decrement, but were not frequent enough to account for the whole effect. Despite the observed difficulty in attending to different locations in vision and audition, there was some benefit from lip-read information even in the Opposite side conditions. Shadowing performance was unaffected by the direction of passive fixation, suggesting that the observed links between vision and audition only apply when the fixated visual events are relevant to the ongoing task.

However, it is possible that these findings do not reflect attentional mechanisms, but rather spatial limitations on the mechanisms that integrate visual and auditory information in lip-reading (e.g. Jack & Thurlow, 1973). In other words, the cost in the Opposite Speaking-Lips may simply arise because the target auditory and visual speech events that must be integrated are further apart, and integration falls off with increasing distance. There are at least two counters to this suggestion. First, the intrusion data suggest an attentional effect, whereby it is difficult to ignore auditory distractors at the current locus of visual attention. Second, it may be that that spatial limitations on crossmodal integration result from difficulties in attending to distinct locations simultaneously, rather than vice-versa. Nevertheless, we ran a further experiment to examine whether we have simply rediscovered a spatial restriction on cross-modal integration. In this follow-up, the subjects performed *unrelated* auditory and visual tasks on stimuli presented in the same or different locations, with no requirement for crossmodal integration. Thus, subjects had to divide their attention among unrelated streams of auditory and visual events, rather than selectively attending to related audiovisual events that had to be perceptually integrated. If the results of Experiment 1 depend on the requirement for integration, they should not obtain in the follow-up study. On they other hand, if they reflect a general difficulty in endogenously attending to different locations in distinct modalities, they should still be observed.

Another rationale for this follow-up study was that lip-reading may be a special (albeit interesting) case. In the real world, speech originates from the lips of the person making the utterance. Through experience of this contingency, auditory and visual locations might come to have an especially powerful synergy in the case of lip-reading.

On the other hand, from the armchair one could equally argue that it may be particularly *easy* to attend to different locations in two modalities while lip-reading, since the target visual and auditory information will be linked by temporal and phonemic correspondences that subjects may employ to circumvent the mismatch in locations. This might explain our observation in Experiment 1 that while it was relatively difficult to select lip-read information from one side and corresponding auditory information from the other side, it must have been possible to some extent (since a lip-read benefit was observed even in this Opposite-side case). Given the equivocality of these speculations about the possible special status of crossmodal integration for lip-reading, we examined whether it is difficult to attend endogenously to different locations when performing completely independent auditory and visual tasks.

Experiment 2

The auditory stimuli were as for Experiment 1, and subjects again had to repeat triplets of rapidly presented words from one side while ignoring concurrent auditory words presented on the opposite side. However, the visual tasks were now unrelated to the auditory materials in every case. A series of visual characters were presented in front of one or other loudspeaker, all the characters appearing in the same location in rapid serial visual presentation (RSVP). They were either presented at the loudspeaker which produced the target words, or at the loudspeaker which produced the distractor words. In the *passive* visual tasks, subjects simply fixated the stream of characters which were otherwise irrelevant to the ongoing task, analogously to the Chewing-lips conditions from Experiment 1. In the *active* visual tasks (intended to be analogous to the previous Speaking-lips conditions) subjects continually monitored the RSVP stream for a prespecified character which occasionally appeared. If the visual target appeared, they had to stop shadowing and verbally report that they had detected it.

The question was whether the pattern of results observed in Experiment 1 would be replicated when the auditory and visual tasks were unrelated, with no requirement for crossmodal integration. If so, performance should be worse when the RSVP stream appears on the opposite side to the target words. However, if the spatial synergy between active visual and auditory attention observed in Experiment 1 is specific to lip-reading, or to tasks requiring crossmodal integration, it should not be found with the present, unrelated active visual task. Given the results of Experiment 1, we would expect that passively fixating one side versus the other would have no effect in the passive visual tasks.

Subjects

The 24 new subjects were unpaid volunteers, all undergraduates. The mean age was 21 years and the range 18-23 years. All were naive as to the purpose of the study, which took about 65 minutes.

Apparatus and Materials

The apparatus and spatial layout was as for Experiment 1, except that the two auditory messages were transferred from video onto audiotape which was played back on a Revox machine, and an IBM microcomputer was used to generate the visual stimuli on one of the monitors used previously. These visual stimuli were presented to the subject immediately in front of one of the loudspeaker-cones on a small mirror via another mirror from the monitor. The stimuli were only visible to the subject on the mirror in front of either loudspeaker, where they subtended a visual angle of 1.6^{o} in width and 3.3^{o} in height on average. The background stimuli for the visual monitoring task were characters from the set <, >, !, (,), *, {, }, [,], -, =, /, :, &, ^, |, and + was the visual target. The visual stimuli for the passive fixation task were the characters Z and O. The sequence of characters was random in the monitoring task, except for the insertion of visual targets.

The characters alternated in the passive fixation task. The auditory stimuli were as in Experiment 1.

Design

There were again two within-subjects factors of interest. One factor was as before, i.e. fixation was on the Same side as the target auditory information or on the Opposite side. The other factor was analogous to the previous Speaking- versus Chewing-lips comparison; the visual task was either active monitoring or passive fixation. These two factors were crossed to yield four conditions: Same Active, Opposite Active, Same Passive and Opposite Passive.

Procedure

This was identical to the previous study, except for the change in visual tasks. Passive conditions were substituted for Chewing-lips and Active conditions for Speaking-lips in the sequences used previously.

The RSVP stimuli were presented at a rate of 1 character every 56 ms and ran continuously during each auditory list. The visual target appeared as the next item in the RSVP sequence if the experimenter pressed a remote button during the monitoring conditions. His button-pushes were neither visible nor audible to the subject. Visual targets occurred equally often in the three positions within spoken triplets. The targets appeared pseudorandomly with the constraint that there were equally likely to appear during 2, 3, 4 or 5 triplets from each list of 14 triplets. They appeared 14 times during each of the two Active conditions. Subjects were not informed about the probabilities of visual target occurrence, but were told that they would only occur during auditory words in the Active conditions.

The accuracy of naming for the target auditory words was recorded as before, and also each hit, miss or false alarm for visual targets. Naming was not scored for triplets when a target was presented. Thus, performance on 126 words was scored for each condition, rather than the full 168 as in Experiment 1. Fixation was again monitored by the experimenter, and additionally by video-recording for the final two subjects. No fixation differences were detected between the Active and Passive conditions.

Results

The mean percentage of words named correctly (out of 126 words per subject in each condition) is shown in Figure 1B, which indicates that the requirement to perform concurrent visual monitoring disrupted shadowing performance less when the visual events were on the same side as the target sounds. A two-way within-subject ANOVA was conducted, with visual task as one factor, and side of fixation as the other. There was an effect of visual task ($F(1,23)=29.7$, $p<.0001$) with slightly poorer shadowing during the active visual task, i.e. a conventional dual-task cost was observed. There was also a main effect of Same versus Opposite side of fixation ($F(1,23)=16.8$, $p<.0001$). The interaction was significant ($F(1,23)=6.2$, $p=.02$). Wilcoxon tests confirmed that while Same performance exceeded Opposite performance in the Active tasks ($T=25.5$, $p<.01$) there was no effect of the direction of fixation in the passive tasks ($T=92$ with 2 ties, n.s.), producing the observed interaction. Thus, as for Experiment 1, a cost in shadowing performance was observed when subjects had to *actively* attend to one side visually and another auditorily, but there was no effect of the direction of passive fixation.

Pooling over the 24 subjects, there were 24 intrusion errors in total for the Same Passive condition, 18 for Opposite Passive, 32 for Same Active, and 45 for Opposite Active. A two-way ANOVA on these relatively scarce errors found a main effect of Active versus Passive visual task ($F(1,23)=7.4$, $p<.01$), but the effect of Same versus Opposite side did not reach significance ($F(1,23)=.6$) and likewise for the interaction ($F(1,23)=2.24$), although note that the trend is towards the pattern found in the accuracy data, i.e. more of an Opposite versus Same cost in the Active tasks. The mean number of correct visual target detections out of a possible 14 was 11.2 for the Same Active task, and 10.8 for the Opposite Active task, with no reliable difference ($F(1,23)=1.6$).

Thus, a decrement in shadowing performance was found when subjects had to endogenously attend to one location visually and another auditorily. As in Experiment 1, there was no effect of the direction of passive endogenous fixation. Since the shadowing and visual monitoring tasks were unrelated, these results confirm a spatial synergy between active visual and auditory attention which is neither specific to lip-reading, nor to a difficulty in crossmodal integration across space, as might have been proposed to account for Experiment 1.

At least two possible explanations for the cost of endogenously attending to distinct locations remain, because our Same- versus Opposite-side manipulation confounded two factors. When the target visual information was presented on the Opposite side to the target sounds, it was both further from the target sounds *and* closer to the distractor sounds. Either factor might be responsible for the performance decrement. It might be difficult to *select* information from different locations in distinct modalities, or difficult to *reject* distractor information in one modality from a location that is relevant for another modality (or both difficulties may apply). Intrusion errors presumably reflect a difficulty in rejection , and these errors were only partly responsible for our effects, suggesting that both difficulties may be involved. To examine the issue further, we conducted a final study which was like Experiment 2 except with *no* distracting information; in other words, a conventional dual-task situation. Any different-location cost in this situation could only be attributed to a difficulty in selecting from different locations, rather than a difficulty in selecting one modality at a particular location while rejecting another at the same location.

Experiment 3

Subjects

The 8 new subjects were again unpaid volunteers. All were undergraduates reporting normal hearing and normal or corrected vision. They were naive as to the purpose of the study, which took about 65 minutes.

Method

The method was as for Experiment 2 except that only one auditory message was presented at any time. Its volume was reduced to bring shadowing performance down to a level comparable with the earlier studies. No fixation differences were detected between the Active and Passive conditions.

Results and Discussion

The mean percentage of words named correctly out of 126 words per subject are shown in Figure 1C. A two-way within-subject ANOVA was conducted as before, with visual task as one factor, and Same versus Opposite side of fixation as the other. There was no effect of visual task on this occasion ($F(1,7)=.02$). If this null effect can be relied upon, one might speculate that the dual-task decrement found in Experiment 2 with the active visual task is eliminated when the auditory task is data-limited by low signal-to-noise ratio, rather than resource-limited as a result of competition from distracting auditory messages. There was a main effect of Same versus Opposite side of fixation ($F(1,7)=9.0$, $p=.02$) and a significant interaction ($F(1,7)=8.6$, $p=.02$). Wilcoxon tests confirmed that while Same performance exceeded Opposite performance in the Active tasks ($T=0$, $p<.01$) there was no effect of the direction of fixation in the passive tasks ($T=10.5$ with 1 tie, n.s.), producing the observed interaction. Thus, as for Experiments 1 and 2 , a cost in shadowing performance was observed when subjects had to *actively* attend endogenously to one side visually and another auditorily, but there was no effect of passively fixating the same or opposite side. The mean number of correct visual target detections out of a possible 14 was 10.3 for the Same Active task, and 9.1 for the Opposite Active task, with no reliable difference ($F(1,7)=1.3$).

Thus, the Active Opposite-side decrement found in Experiment 2 was replicated when no distracting stimuli were presented, suggesting that the previous results did not solely reflect a difficulty in rejecting information from one modality at a position that is attended in another modality. Instead, we appear to have identified a difficulty in combining auditory and visual tasks when the stimuli come from different locations. As in the previous experiments, this difficulty cannot be attributed to the direction of gaze, since it was not found in the Passive conditions.

The null effect of the direction of passive fixation in all three of our studies is somewhat surprising since Reisberg, Scheiber & Potemken (1981) reported that selective listening is enhanced when subjects are allowed to fixate the loudspeaker source for the target auditory message, extending the earlier work of Gopher (1973). Similarly, Larmande, Elghori, Sintes, Bigot & Autret (1983) found that the ear advantage in dichotic listening to nonverbal sounds can be reversed by directing gaze in the opposite direction. Honore, Bordeaud'hui & Sparrow (1989) and Pierson, Bradshaw, Meyer, Howard, and Bradshaw (1991) observed that the direction of passive gaze can also affect response to tactile stimuli, reducing cutaneous reaction time in the fixated direction. Finally, Larmande & Cambier (1981) report that tactile extinction of the contralateral event on double simultaneous stimulation was reduced in right hemisphere patients when they were required to fixate towards the contralateral side. These data all suggest that the direction of gaze can affect attention in other modalities, in contrast with the present null findings concerning passive fixation. There are many procedural differences between these prior studies and our own experiments which might account for the discrepancy. However, Wolters and Schiano (1989) failed to replicate Reisberg et al.'s (1981) effect of fixation using their exact procedure. Thus, the precise circumstances under which passive fixation modulates selective listening, or selection in other modalites, remain unclear. However, our studies demonstrate that any effects of passive fixation on audition must be smaller than those produced by the direction of *active* endogenous visual attention towards visual events which must be processed to perform the ongoing task.

General Discussion

We began with two fundamental questions; whether attentional mechanisms operate independently in distinct unimodal systems, and how attention comes to be coordinated across the modalities despite their different input organizations. The present results illustrate that endogenous attention does not operate indepedently for vision and audition. Three experiments found evidence for a spatial synergy between auditory and visual endogenous attention, such that it is difficult to attend to one side of the midline visually and the other auditorily. This opposite-side decrement was found for tasks requiring crossmodal integration (listening while lip-reading in Experiment 1) and for unrelated auditory and visual tasks (listening while RSVP monitoring in Experiments 2 and 3), although the cost was larger in the integrative case. Experiment 3 found the opposite-side decrement in a conventional dual-task situation (i.e. when no distracting information was presented) suggesting that the spatial synergy may be a quite general limit on concurrent task performance. In contrast to the significant decrement found when endogenously attending to different locations, there was no observable decrement for passively fixating one location while attending auditorily to another.

The spatial synergy we have observed makes considerable functional sense from the perspective that attentional mechanisms serve to select relevant *objects* for the control of action (e.g. Allport, 1987; Driver & Baylis, 1989; Duncan, 1984). Sights and sounds emanating from the same distal object are likely to come from approximately the same location. Spatial links between visual and auditory attention might therefore provide a rough solution to the problem raised by our second question, namely how attention is coordinated across modalities so that the same object is attended to visually and auditorily.

One extreme solution to the coordination problem would be to direct attention in a supramodal representation of the scene following crossmodal integration (Butter et al., 1989), rather than within each modality. Can the present data be taken as evidence for such a supramodal attentional system? We think there is currently no unequivocal evidence in favour of such a system. First, our data are as readily accommodated by the proposal of spatial links between otherwise distinct modality-specific systems. Second, Experiment 1 found that while it was difficult to endogenously attend to different locations in vision and audition, it was nevertheless possible to some extent, as a lip-read benefit was observed even in the Opposite-side condition. This is problematic for a purely supramodal account of spatial selectivity. Third, clinical evidence from attentional deficits following brain damage argues strongly for the existence of modality-specific attentional systems, since the severity of neglect or extinction in audition can be unrelated to the severity in vision (Barbieri & De Renzi, 1989).

In contrast to these dissociations between modalities, crossmodal interactions can also arise in attentional deficits. For example, extinction on double simultaneous stimulation can be observed in unilateral parietal patients between vision and audition (Farah et al., 1989) if not between touch and vision (Inhoff, Rafal & Posner, 1992). These findings tempt one to postulate both modality-specific attentional mechanisms at lower levels of representation (responsible for the dissociations), and supramodal attentional mechanisms at higher levels (producing the crossmodal interactions). Posner (1990) proposes exactly this kind of hierarchical attentional model, with modality-specific systems feeding into a supramodal attentional system. The strength of such models is that they can accommodate both dependence and independence between modalities. However, for this very reason they are difficult to falsify, unless couched in more precise neuroanatomical terms.

While we consider that the case for a supramodal attentional system remains unproven, there is now clear evidence that endogenous selection can take place *after* crossmodal integration in some cases. Driver (in prep.) examined this issue by exploiting an established audiovisual interaction; sounds tend to be mislocated towards the position of their apparent visual source (e.g. cinema audiences typically mislocate the spoken soundtrack towards the actors appearing on the screen). This 'ventriloquism' effect is particularly powerful for visual lip-read information and matching speech-sounds (Witkin et al., 1952). Driver presented his subjects with two concurrent messages spoken by the same person, as in the present studies. However, they were now both played-back from the same loudspeaker in mono. Subjects had to repeat one message, for which they were also given lipread information on video, as in Experiment 1. This visual information was either presented in the same location as the mono sound-source, or slightly displaced towards a dummy loudspeaker. Performance was improved in the displaced case. Driver's (in prep.) account is that the target sounds were mislocated towards the matching lips via the ventriloquism effect, while the distractor sounds were correctly located at the mono sound-source because they did not match the lips and were therefore not subject to ventriloquism. This produced an illusion of auditory stereo when the lips were displaced from the mono sound source, which improved selective listening as for true auditory stereo. In other words, an illusion (mislocation of the target sounds via ventriloquism) actually enhanced veridical perception (report of the target words). The implication is that crossmodal matching must have taken place *before* endogenous attentional selection, as the improved performance with displaced lips requires that only the sounds which matched the lips were mislocated towards them, allowing enhanced segregation from the mismatching sounds.

On the face of it, these results appear to provide a counterexample to the current finding that attending to different physical locations in distinct modalities carries a performance cost, since performance was enhanced in Driver's (in prep.) study when the visual lip-read information was slightly displaced *away* from the target (and distractor) sounds. Presumably, however, the target sights and sounds in Driver's study were *coded* as emanating from the same (or similar) location even in the displaced condition, as a result of ventriloquism. In other words, the difficulty in attending to distinct locations may only apply for locations which are *perceived* as different following crossmodal integration, rather than applying for any locations which are sufficently distinct physically to be discriminated within a modality.

We do not suggest that Driver's (in prep.) data establish a strong case for supramodal mechanisms of selection. They can be accommodated if selection proceeds within individual modalites, provided one allows crossmodal influences on the coding of location for each modality. However, the results clearly rule out an extreme modality-specific account, on which each sensory modality is an encapsulated module prior to the allocation of endogenous attention.

Returning to the present experiments, the spatial synergy we have observed could be used to examine the representation of space by different modalities. Each modality has a distinct spatial organization at receptor input. For example, there is no auditory equivalent to the retinotopic organization that provides the initial coordinate system for vision. This point raises many questions about the mappings that relate space in one modality to space in another to produce crossmodal spatial interactions, and to allow the crossmodal coordination of attention. By systematically varying the relative locations of events in distinct modalities, and examining any cost of attending to these locations simultaneously, we should be able to identify what counts as the 'same' or 'different' location across modalities at the processing stages where endogenous attention operates. In the present experiments, stimuli were either in the same or opposite hemifields, so further studies are required to examine whether similar principles apply for distinct locations within a hemifield. Despite this uncertainty, our present data cannot be accounted for in terms of differential hemisphere activation in the various conditions (Kinsbourne, 1987) . The activation produced by the requirement to fixate one side or the other should be identical for our Passive and Active conditions, yet differential results were observed. Equally, the visual events in the Same- and Opposite-side conditions should produce equivalent hemispheric activation since they were always at fixation, and yet differential results were again observed.

It remains for future research to establish the range of task- and modality-combinations for which the observed spatial synergy in active endogenous attention applies. The fact that we found this synergy for *unrelated* auditory and visual tasks suggests that it may be quite general. If so, the spatial synergy would place a major restriction on our ability to perform two tasks in different modalities simultaneously. It remains controversial whether previous studies of dual-task performance have succeeded in identifying general limits on performance, or limits which are entirely specific to the pair of tasks in question (e.g. Allport, 1980, 1992 ; Broadbent, 1982; Shallice, McLeod & Lewis, 1985). The least controversial factors considered to improve dual-task performance include separation of input modality (Treisman & Davies, 1973) and separation of response modality (e.g. McLeod, 1977), leading to the widely held rule-of-thumb that dissimilar tasks are easier to combine than similar tasks. We appear to have identified one factor which violates this principle; similarity in the location of stimuli for tasks in different modalities helps rather than hinders task combination.

Acknowledgements

We thank Alan Allport for his encouragement and advice on these studies. The research was supported by a grant to the first author from the Medical Research Council (UK).

Allport, D.A. (1980). Attention and performance. In G. Claxton (ed.), *Cognitive psychology: new directions*. London: Routledge & Kegan Paul.

Allport, D.A. (1987). Selection-for-action: Some behavioral and neurophysiological consideration sof attention and action. In H. Heuer & A.F. Sanders (Eds.), *Perspectives on perception and action* (pp. 395-419). Hillsdale, NJ: Erlbaum.

Allport, D.A. (1992, in press). Selection and control: A critical review of 25 years. In S. Kornblum & D.E. Meyer (Eds.), *Attention and Performance XIV: A silver Jubilee*. Hillsdale, NJ: Erlbaum.

Barbierei, C., & De Renzi, E. (1989). Patterns of neglect dissociation. *Behavioural Neurology*, **2**, 13-24.

Broadbent, D.E. (1958). *Perception and communication.* London: Pergamon.

Broadbent, D.E. (1982). Task combination and selective intake of information. *Acta Psychologica*, **50**, 253-290.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behaviour Research Methods, Instrumentation and Computers*, **16**, 502-532.

Buchtel, H.A., & Butter, C.M. (1988). Spatial attention shifts: Implications for the role of polysensory mechanisms. *Neuropsychologia*, **26**, 499-509.

Butter, C.M., Buchtel, H.A., & Santucci, R. (1989). Spatial attentional shifts: Further evidence for the role of polysensory mechanism using visual and tactile stimuli. *Neuropsychologia*, 27, 1231-1240.

Cherry, E.C. (1953). Some experiments upon the recognition of speech, with one and two ears. *Journal of the Acoustical Society of America*, **25**, 975-979.

Driver, J. (in prep.). Selective listening and lip-reading: crossmodal integration can occur before attentional selection.

Driver, J., & Baylis, G.C. (1989). Movement and visual attention: the spotlight metaphor breaks down. *Journal of Experimental Psychology: Human Perception and Performance*, **15**, 448-456.

Duncan, J. (1984). Selective Attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*, 501-517.

Farah, M.J., Wong, A.B., Monheit, M.A., Morrow, L.A. (1989). Parietal mechanisms of spatial attention: Modality-specific or supramodal? *Neuropsychologia*, **27**, 461-470.

Gopher, D. (1973). Eye-movement patterns in selective listening tasks of focussed attention. *Perception & Psychophysics*, **14**, 259-264.

Honore, J., Bourdeaud'hui, M., Sparrow, L. (1989). Reduction of cutaneous reaction time by directing eyes towards the source of stimulation. *Neuropsychologia*, **27**, 367-371.

Inhoff, A.W., Rafal, R.D., & Posner, M.I. (1992). Bimodal extinction without crossmodal extinction. *Journal of Neurology, Neurosurgery and Psychiatry*, **55**, 36-39.

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

Jonides, J. (1981). Voluntary versus automatic control over the mind's eye's movement. In J. Long & A. Baddeley (Eds.), *Attention and Performance: Vol. IX*. Hillsdale, NJ: Erlbaum.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D.R. Davies (Eds.), *Varieties of attention*. New York: Academic.

Kimura, D. (1961). Cerebral dominance and the perception of verbal stimuli. *Canadian Journal of Psychology*, **15**, 166-171.

Kinsbourne, M. (1987). Mechanisms of unilateral neglect. In M. Jeannerod (Ed.), Neurophsyiological and neuropsychological aspects of spatial neglect. (pp. 69-86). North-Holland: Elsevier.

Klein, R. (1987). Covert cross-modality orienting of attention. Presented to the Psychonomic Society.

Larmande, F., Elghori, D., Sintes, J., Bigot, T., & Autret, A. (1983). Test d'ecoute dichotique verbal et non verbal chez le suject normal: influence de l'etat d'activation hemishperiue. *Revue Neurologique*, **139**, 65-69.

Larmande, P., & Cambier, J. (1981). Influence de l'etat d'activation hemispherique sur le phenomene d'extinction sensitive chez 10 patients atteints de lesions hemispheriques droites. *Revue Neurologique*, **137**, 285-290.

McLeod, P.D. (1977). A dual-task response modality effect: Support for multiprocessor models of attention. *Quarterly Journal of Experimental Psychology*, 29, 651-667.

Morais, J., & Bertelson, P. (1975). Spatial position versus ear of entry as determinant of the auditory laterality effects: A stereophonic test. *Journal of Experimental Psychology: Human Perception and Performance*, 1, 253-262.

Muller, H.J., & Humphreys, G.W. (1991). Luminance-increment detection: Capacity-limited or not? *Journal of Experimental Psychology: Human Perception and Performance*, 17, 107-124.

Muller, H.J., & Rabbitt, P.M.A. (1989). Reflexive and voluntary orienting of visual attention: Time course of activation and resistance to interruption. *Journal of Experimental Psychology; Human perception and Performance*, 15, 315-330.

Neumann, O., van der Heijden, A.H.C., & Allport, D.A. (1986). Visual selective attention: Introductory remarks. *Psychological Research*, **48**, 185-188.

Pierson, J.M., Bradshaw, J.L., Meyer, T.F., Howard, M.J., & Bradshaw, J.A. (1991). Direction of gaze during vibrotactile choice reaction time tasks. *Neuropsychologia*, **29**. 925-928.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, **32**, 3-25.

Posner, M.I. (1990). Hierarchical distirbuted networks in the neuropsychology of selective attention. In A. Caramazza, Ed., *Cognitive neuropsychology and neurolinguistics: Advances in models of cognitive function and impairment* (pp. 187-210). Hillsdale, NJ: Erlbaum:

Rafal, R.D., Henik, A., & Smith, J. (1992). Extrageniculate contribution to reflex visual orienting in normal humans: A temporal hemifield advantage. *Journal of Cognitive Neuroscience*, **3**, 322-328..

Reisberg, D. (1978). Looking where you listen: Visual cues and auditory attention. *Acta Psychologica*, **42**, 331-341.

Reisberg, D., Scheiber, R., and Potemken, L.(1981). Eye position and the control of auditory attention. *Journal of Experimental Psychology: Human Perception and Performance*, **7**, 318-323.

Shallice, T., McLeod, P., & Lewis, K. (1985). Isolating cognitive modules with the dual task paradigm: Are speech perception and production separate processes. *Quarterly Journal of Experimental Psychology*, **37A**, 507-532.

Spence, C.J., & Driver, J. (in prep.). Covert spatial orienting in audition: Exogenous and endogenous mechanisms facilitate sound localization.

Thompson, G.C., & Masterton, R.B. (1978). Brain stem auditory pathways involved in reflexive head orientation to sound. *Journal of Neurophysiology*, **41**, 1183-1202.

Tipper, S.P., Driver, J., & Weaver (1991). Object-centred inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology*, **43A**, 289-298.

Treisman, A.M., & Davies, A. (1973). Divided attention to ear and eye. In S. Kornblum (ed.) *Attention and performance: vol IV*. London: Academic.

Treisman, A.M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, **12**, 97-136.

Tsal, Y., & Lavie, N. (1988). Attending to color and shape: The special role of location in visual processing. *Perception & Psychophysics*, **44**, 15-21.

Whittington, D.A., Hepp-Reymond, M.C., & Flood, W. (1981). Eye and head movements to auditory targets. *Experimental Brain Research*, **41**, 358-363.

Witkin, H.A., Wapner, S., & Leventhal, T. (1952). Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology*, **43**, 58-67.

Wolters, N.C.W., & Schiano, D.J. (1989). On listening where we look: the fragility of a phenomenon. *Perception & Psychophysics*, **45**, 184-186.

Figure legends

Figure 1

(A) The mean percentage of correct repetitions of the target words in Experiment 1, averaged over subjects for each condition. The same-side conditions are indicated with hatched columns, the opposite-side conditions with solid columns.

(B) The mean percentage of correct repetitions of the auditory target words in Experiment 2, averaged over subjects for each condition. The same-side conditions are indicated with hatched columns, the opposite-side conditions with solid columns.

(C) The mean percentage of correct repetitions of the auditory target words in Experiment 3, averaged over subjects for each condition. The same-side conditions are indicated with hatched columns, the opposite-side conditions with solid columns.