Perception & Psychophysics
1984, 35 (6), 565-569

Cross-modal effects on visual and auditory object perception

ANN O'LEARY and GILLIAN RHODES
Stanford University, Stanford, California

Cross-modal influences on perceptual organization were demonstrated using a display that combined a stimulus for auditory stream segregation with its visual apparent movement analogue. Both phenomena give rise to the perception of either one or two objects, depending on the rate of presentation of the stimuli. At slower rates, one object is perceived, while two are perceived at faster rates. Subjects indicated the stimulus onset asynchrony (SOA) between successive stimuli at which the perceptual shift occurred in each modality. Then visual and auditory stimuli were presented concurrently and subjects responded to the "target" modality sequence. Two intergroup separations for the nontarget stimuli were used. Distances were chosen, based on the subject's calibration data, which represented one and two objects, respectively, at the stream segregation point for the target sequence. Segregation occurred at larger SOAs when the nontarget stimulation indicated two objects than when it represented one. This was true for both visual and auditory target sequences.

One of the most important functions of perception is to parse the sensory array into objects and to inform the organism about the behavior of these objects. Visual object segregation succeeds despite partial occlusion of one object by another, two-dimensional deformation of the retinal image produced by moving objects, shadows that extend across object boundaries, and texture and illumination changes across object surfaces.

Vision is not the only modality in which object segmentation occurs. Auditory object perception, called "primary auditory stream segregation," or PASS, has also been demonstrated (Bregman, 1978, 1981; Bregman & Campbell, 1971; Bregman & Rudnicky, 1975). This occurs when a sequence of alternating high- and low-frequency tones is played. At slow presentation rates, subjects are able to follow the entire sequence of tones, but at higher rates the sequence splits into two streams, one high and one low in pitch. While it is possible to shift attention back and forth between the two streams, it is difficult to report the order of tones in the entire sequence. Auditory stream segregation, like apparent motion in vision, appears to follow "Korte's third law" (Korte, 1915); as the distance in frequency between the subgroups of tones decreases, stream segregation occurs at shorter stimulus onset asynchronies (SOAs), that is, at slower rates of alternation.

This work was supported by NSF Grant BNS 80-05517 to Roger Shepard. We thank Edward Kessler for his assistance with programming and equipment. We also thank Roger Shepard and Steven Pinker for helpful discussions. The authors' mailing address is: Department of Psychology, Jordan Hall, Bldg. 420, Stanford University, Stanford, CA 94305-2099.

Auditory stimulation typically corresponds to objects undergoing change over time, to object events, so to speak. The PASS phenomenon reflects the tendency of the perceptual system to decompose the ambient auditory stimulation so that it is likely to correspond to discrete sound sources. Bregman and his co-workers have described the operation of a number of analogues to Gestalt organizational principles in the perception of "sound objects." They include similarity, for example, in volume and timbre (McAdams & Bregman, 1979), good continuation (Dannenbring, 1976), and common fate (McAdams, 1977). (See also Bregman, 1981, and Julesz & Hirsh, 1972.)

There is a visual analogue of PASS, which its investigators call "visual stream segregation" (VISS) (Achim & Bregman, 1973). It is an apparent movement phenomenon which occurs when a series of light spots, two high and two low in the visual field, are presented in an alternating H1-L1-H2-L2 pattern. At low presentation rates, a single dot undergoes apparent movement up and down (H1-L1-H2-L2), while at higher rates two dots are seen bobbing on their own shorter paths, one above the other (H1-H2 and L1-L2).

These findings raise an interesting question: Can the perceptual system utilize converging information from *more* than one sensory modality to organize the perceptual array in situations in which multimodal information is available? There is evidence that, under certain conditions, the perceptual system attributes covarying stimulation in vision and audition to a unitary object event. The so-called "ventriloquism effect" is an example of this perceptual tendency: sounds that covary with visual stimulation originat-

ing at a different spatial location are localized at, or closer to, the visual display (Hay, Pick, & Ikeda, 1965; Jackson, 1953; Thomas, 1941; Witkin, Wapner, & Leventhal, 1952). Precise covariation of visual and auditory stimulation is critical for the effect (Jack & Thurlow, 1973). Radeau and Bertelson (1976) and Bertelson and Radeau (1981) have also demonstrated "auditory capture," in which the position of an auditory source affected the localization of a simple synchronous visual stimulus.

The present study was designed to directly test the hypothesis that perceptual segmentation in one modality can be influenced by concomitant, covariant stimulation in another. Specifically, it was proposed that the *number* of objects perceived might be subject to cross-modal influence, when the number of objects indicated is ambiguous in one (our "target") modality and unambiguous in the other. The study employed the PASS and VISS paradigms in combination. Analogous and synchronized visual and auditory stimulus arrays were used. In the calibration phase of the study, data were collected separately for each sensory modality. The functions relating separation in space and frequency to the SOA demarcating percepts of one versus two objects were found for each subject. This SOA will be called the "stream segregation point." Subsequently, in the experimental phase, a stimulus with a selected intergroup separation was paired in turn with a "one-object" sequence and a "two-object" sequence in the other modality. These "one-object" and "two-object" sequences were selected on the basis of the subject's calibration data in such a way that the sequence was perceived as one object or two, respectively, when presented at the rate corresponding to the stream segregation point in the target modality. It was expected that the stream segregation point of the target modality would depend on whether a "one-object" or a "two-object" sequence was present in the nontarget modality. We predicted that the stream segregation point would occur at larger SOAs when the covariant, nontarget stimulus indicated two objects than when it indicated one. This would be a demonstration of cross-modal influences on perceptual organization.

METHOD

Subjects

Eight subjects, six women and two men, were paid for their participation in the study. Three potential subjects were rejected following collection of calibration data—two because their auditory data were extremely variable and the third because calibration data in the two modalities failed to overlap, making the experimental phase impossible.

Stimuli

The stimuli were generated by an Apple II PLUS microcomputer using low-density graphics and the Mountain Computer Music System Synthesizer card. In the calibration phase of the experiment, the subjects reported the stream segregation points for stimuli in each modality separately. In the experimental phase, visual and auditory stimuli were presented simultaneously.

Visual stimuli. The visual display was generated in the Apple's low-density graphics mode on a Sanyo video monitor. It consisted of six white rectangles whose dimensions were 4×6 mm (vertical × horizontal). These "dots" were presented one at a time and were arranged vertically. The three high positions alternated with the three low ones in an H1-L1-H2-L2-H3-L3 pattern (see Figure 1), which was continuously repeated. At large SOAs, smooth apparent motion of a single dot was seen over time, while at smaller SOAs (higher rates of alternation) apparent motion over the central gap broke down and two visual objects were experienced, one high and one low, each undergoing apparent motion over a shorter path. The distance between adjacent dots within subgroups was always 4 mm (measured between adjacent edges), and four separations between subgroups were used: 8, 24, 48, and 72 mm. The subject was seated 120 cm from the display. The visual angle between the highest and lowest dots thus ranged from 3.8 to 9.9 deg. SOA was varied by changing the durations of presentation of the dots; there was a constant interstimulus interval of 40 msec. All six dots were of equal duration. The display monitor was set to maximum overall luminance. The light emitted by the screen was the only source of illumination in the room.

Auditory stimuli. Each auditory stimulus sequence consisted of a set of six tones, presented in a pattern analogous to that for the visual display, where height in pitch corresponds to height in the visual field. The order of stimulus "positions" was the same as that for the dots. As with the visual stimuli, the sequence was repeated continuously. The waveform of the tones was a close approximation to a sawtooth wave. The amplitude envelope consisted of attack and decay times of 26 msec, with the period of maximum amplitude varying to produce the range of SOAs. A constant interstimulus interval of 40 msec separated temporally adjacent tones. The maximum amplitude was set to comfortable listening level. The PASS phenomenon gave rise to the perception of either one or two "objects" in a fashion analogous to the visual organization described above. At larger SOAs a single stream was perceived; at smaller ones two streams, one high in frequency and the other low, were reported. The stream segregation point was a function of the separation in frequency between the two subgroups. Separation in pitch between adjacent (in pitch) members

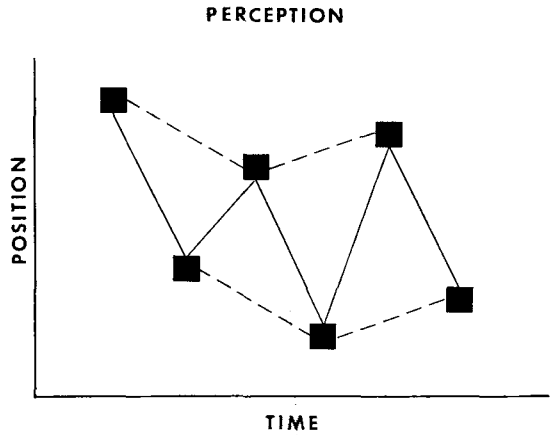

Figure 1. The displays used in the experiment. "Position" refers to position in space for the visual display and frequency for the auditory display. The abscissa represents time; the visual stimuli were vertically aligned. The solid line represents the "one-object" percept experienced with slow presentation rates; the dotted line represents the "two-object" percept experienced with faster presentation rates.

of the subgroups was kept perceptually constant by the maintenance of a constant frequency ratio (1.06) between them. Four intergroup separations were used during the calibration phase; they were 40, 80, 160, and 320 Hz. The middle frequency was always 600 Hz, and the frequencies used ranged from 390.4 to 856.5 Hz.

Combined visual and auditory stimuli. In the experimental phase of the experiment, the repeating sequences of six auditory and six visual stimuli were paired so that the tone highest in frequency was presented simultaneously with the dot highest in the field, the second highest tone with the second highest dot, and so on. SOAs in the two modalities were equal, so that the onset and termination of dot and tone presentations coincided.

Calibration

For each subject, the first four 1-h sessions were devoted to obtaining the functions relating stream segregation point SOA to distance (in space and frequency, respectively) in each modality separately. On each trial, a stimulus with one of the previously listed intergroup separations was presented with an initial SOA of 336 msec. At 5-sec intervals, the experimenter increased the rate of stimulus presentation by decreasing the SOA by 8 msec. This was effected by depressing a microswitch, which produced a clicking sound clearly audible to the subject. The subject was instructed to attend to the stimuli and to attempt to see or hear them as a single "dot" or "sequence." Free scanning of the visual display was recommended. When a spontaneous shift occurred, so that there seemed to be two dots or sequences, the subject was to continue watching or listening and, if the stimuli remained segmented for a full 5-sec interval, to respond. The experimenter then pressed a second button to record the current SOA. If another spontaneous shift to the original perceptual organization occurred before the 5 sec elapsed, the subject was instructed to await a second splitting and to respond when the more stable, 5-sec-plus duration was obtained. This procedure had been determined in pilot testing to be superior, because of the lability of the PASS phenomenon, to one relegating control of SOA changes to the subject. Spontaneous vacillation of auditory organization occurs over a wide range of frequency/SOA combinations (van Noorden, 1975), and this results in high variance in data collected with a less controlled procedure.

Calibration data were collected in one sensory modality at a time, in counterbalanced order. Ten random permutations of the intergroup separation values were administered; the average of the 10 values at a separation became the subject's stream segregation point score for that separation. The experimenter demonstrated the stream segregation phenomena at the beginning of the first calibration session for each modality. The auditory intergroup separation of 160 Hz and the visual separation of 24 mm were used for this demonstration. Auditory calibration was preceded by 16 practice trials (four random permutations of the frequency separations) and visual calibration, by 8 practice trials (two permutations).

Procedure

For the experiment proper, a separation value was chosen in each modality whose stream segregation point for the subject fell most centrally between the stream segregation points for the largest and smallest separations in the other modality. This value was paired in turn with the two extreme separations in the other modality, which were experienced as "one-object" or "two-object" sequences. The two visual separations were simply the two extreme separations used during calibration, 8 and 72 mm. The auditory separations were 20 and 320 Hz. A 20-Hz separation was used to ensure that the nontarget auditory stimulus would indicate a single object. Stream segregation was impossible with the 20-Hz separation because the intergroup frequency separation was slightly smaller than the intragroup separation. Thus, four combinations of visual and auditory separations were selected: a central visual value to be used on trials in which vision was to be the target modality, paired with a large and a small auditory separation; and a central auditory target value, to be paired with a large and a small visual separation. Thus, the target stimulus was presented sometimes in conjunction with an other-modality stimulus whose organization could be expected to represent a single object at the target's stream segregation point (small separation) and sometimes with one representing two objects (large separation).

Ten random permutations of the four stimulus combinations constituted the experimental phase of the study for each subject. The 10 SOAs obtained in each condition were averaged to give the subject's SOA for that condition. The 40 experimental trials were administered in two 45-min sessions. Prior to each experimental trial, the subject was informed which was to be the target modality, that is, whether "dots" or "tones" were to be responded to. The format of the trial remained identical to that of the calibration trials, except that for auditory target trials the subject was instructed to look at the visual display.

Upon completion of the experiment, subjects were requested to write brief essays describing any hypotheses they had developed about the experiment. They were asked to state when each hypothesis had occurred to them, and how it might have affected their responses, if at all. No one generated the experimental hypothesis. Most subjects had supposed that some interference or facilitation of the stream segregation should accrue across the board to the combination of the visual and auditory stimuli, although they had had no expectations concerning which, interference or facilitation, should occur. Two subjects were not naive; however, on average their data resemble those of the other subjects. These nonnaive subjects reported that they had employed no conscious response strategies.

RESULTS

Calibration Data

The mean stream segregation points (and standard deviations, in parentheses) in audition were as follows (values are in milliseconds): 40 Hz, 158.5 (36.1); 80 Hz, 193.7 (41.4); 160 Hz, 245.2 (27.2); and 320 Hz, 275.9 (30.7). In vision the values were: 8 mm, 128.4 (37.3); 24 mm, 172.2 (46.5); 48 mm, 217.4 (48.9); and 72 mm, 257.1 (41.5).

Calibration data followed Korte's third law. The segregation point (y msec) was a linear function of the intergroup separation (x mm or Hz) for both modalities. In the visual modality, the function is $y = 31.6x + 474.4$ mm, and it accounts for 99% of the variance. In the auditory modality, the function is $y = 0.4x + 158.5$ Hz, and it accounts for 89% of the variance. Adherence to Korte's third law is evident for every subject in both modalities.[1] Visual apparent motion was always reported at the beginning SOA of 336 msec, and continued until the reported streaming point was reached.

Experimental Results

The results of the experimental phase of the study are summarized in Table 1. The hypothesis that the number of objects perceived can be influenced by the number perceived in another concomitant, covariant sense modality was supported. Stream segregation in both modalities occurred at larger SOAs when the nontarget stimulation indicated two objects than

Table 1
Mean Stream Segregation SOAs (in Milliseconds) are Shown for Auditory Influences on a Visual Target Sequence and for Visual Influences on an Auditory Target Sequence

Target Modality			
Vision		Audition	
	SOA		SOA
Auditory–one object	211	Visual–one object	182
Auditory–two objects	221	Visual–two objects	199
Difference	10	Difference	17
$t(7) = 2.23$ ($p < .05$)		$t(7) = 4.29$ ($p < .001$)	

when it indicated one. The effect of visual stimulation upon auditory stream segregation resulted in a difference of 17 msec [$t(7) = 4.29$, $p < .001$]. This corresponds to a difference in duration of the six-stimulus sequence of 102 msec and is 8% of the mean stream segregation SOA for audition in the experimental phase. Seven of the eight subjects showed this effect. The magnitude of the effect of auditory organization on vision amounted to an SOA difference between successive dot presentations of 10 msec [$t(7) = 2.23$, $p < .05$]. This corresponds to a sequence difference of 60 msec and is 5% of the mean stream segregation SOA for vision in the experimental phase. Six of the eight subjects showed the effect. The effect of visual organization upon audition was larger than that of auditory organization upon vision; however, this difference was not significant [$t(7) = .83$, $p > .1$].

DISCUSSION

We have shown that perceptual organization in one modality can influence organization in another. In particular, the presence of a visual sequence that is perceived as two moving objects (dots), causes a concurrent auditory sequence to be perceived as two objects (tones) at an SOA that yields a perception of a single auditory object when the accompanying visual sequence is perceived as a single object. Thus, the number of objects seen influences the number of objects heard in an analogous, in-phase, covarying auditory sequence. This influence also occurs in the converse direction, with the number of objects heard affecting the number seen. It remains to be seen what degree of covariation is necessary for cross-modal influences to occur. Our results show that people are sensitive to covariation between dynamic patterns in different modalities. Moreover, they are consistent with a perceptual mechanism that interprets covarying auditory and visual information as originating from a single-source object.

Recent work supports the view that our perceptual apparatus has evolved to mirror constraints on the behavior of objects in the world (Shepard, 1981). One such constraint is that temporally covariant stimulation in two sense modalities typically originates from a single source. The behavior of young infants is consistent with the operation of this constraint (Spelke, 1979; Spelke & Cortelyou, 1981). In addition, animal studies have demonstrated the existence of polymodal cortical cells (Rosensweig & Leiman, 1982) and spatially aligned visual and auditory cortical representations (Knudsen, 1982); these may provide a physiological basis for cross-modal effects.

Further research should be directed toward defining the precise conditions under which cross-modal effects occur. For example: Are the results the same when height in the visual field corresponds to its inverse in frequency (that is, where high dots are paired with low tones)? Or is the correspondence between pitch height and height in space a fundamental one? Does one obtain larger cross-modal effects when the auditory and visual display items covary in more than one dimension (for example, when lightness of the visual display items covaries with loudness of the tones)? The paradigm used in this experiment may be instrumental in defining the determinants of cross-modal perceptual organization.

REFERENCES

Achim, A., & Bregman, A. S. (1973). Visual stream segregation. *Perception & Psychophysics*, **13**, 451-454.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, **29**, 578-584.

Bregman, A. S. (1978). The formation of auditory streams. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, NJ: Erlbaum.

Bregman, A. S. (1981). Asking the what-for question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, **89**, 244-249.

Bregman,, A. S., & Rudnicky, A. I. (1975). Auditory segregation: Stream or streams? *Journal of Experimental Psychology: Human Perception and Performance*, **1**, 263-267.

Dannenbring, G. L. (1976). Perceived auditory continuity with alternately rising and falling frequency transitions. *Canadian Journal of Psychology*, **30**, 99-114.

Hay, J. C., Pick, H. L., & Ikeda, K. (1965). Visual capture produced by prism spectacles. *Psychonomic Science*, **2**, 215-216.

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.

Jackson, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology*, **5**, 52-65.

Julesz, B., & Hirsh, I. J. (1972). Visual and auditory perception—an essay of comparison. In E. E. David, Jr. & P. B. Denes (Eds.), *Human communication: A unified view*. New York: McGraw-Hill.

Knudsen, E. I. (1982). Auditory and visual maps of space in the optic tectum of the owl. *The Journal of Neuroscience*, **2**, 1177-1194.

Korte, A. (1915). Kinematoscopische Untersuchungen. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, **72**, 193-296.

McAdams, S. E. (1977). *The effect of quality on auditory stream*

segregation. Undergraduate Honors Thesis, Department of Psychology, McGill University, Montreal, Quebec.

McAdams, S. E., & Bregman, A. S. (1979). Hearing musical streams. *Computer Music Journal*, **3**, 26-43.

Radeau, M., & Bertelson, P. (1976). The effect of a textured visual field on modality dominance in a ventriloquism situation. *Perception & Psychophysics*, **20**, 227-235.

Rosenzweig, M. R., & Leiman, A. L. (1982). *Physiological psychology*. Lexington, MA: Heath.

Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization*. Hillsdale, NJ: Erlbaum.

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, **15**, 626-636.

Spelke, E. S., & Cortelyou, A. (1981). Perceptual aspects of social knowing: Looking and listening in infancy. In M. E. Lamb & L. R. Sherrod (Eds.), *Infant social cognition*. Hillsdale, NJ: Erlbaum.

Thomas, G. J. (1941). Experimental study of the influence of vision on sound localization. *Journal of Experimental Psychology*, **28**, 167-177.

van Noorden, L. P. A. S. (1975). *Temporal coherence in the perception of tone sequences*. Ph.D. Dissertation, Tech. Hogeschool, Eindhoven, The Netherlands (published by the Institute of Perception Research, Eindhoven, The Netherlands).

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

NOTE

1. Since the demonstration of Korte's third law was not the main purpose of the study, the calibration was not conducted in such a way as to yield precise functions for each subject. For one thing, trials were performed with the SOA always decreasing rather than with a combination of ascending and descending approaches. Secondly, the initial SOA value remained constant throughout the experiment, so that trials involving sequences with smaller intergroup separations lasted longer than those utilizing larger ones. Thirdly, the loudnesses of the tones were not equalized precisely for each subject, and stream segregation may have been facilitated by the tendency for the lower subgroups to be of greater volume than the higher, a tendency that was exaggerated at greater frequency separations.

(Manuscript received January 16, 1984;
revision accepted for publication April 26, 1984.)