Sound Enhances Visual Perception: Cross-modal Effects of Auditory Organization on Vision

Jean Vroomen and Beatrice de Gelder

In Journal of Experimental Psychology: Human Perception and Performance, 26, 1583-1590

Jean Vroomen
Tilburg University, Dept of Psychology
Warandelaan 2
PO Box 90153, 5000 LE, Tilburg
The Netherlands
Phone: +31-13-4662394
FAX: +31-13-4662067
e-mail: j.vroomen@uvt

Abstract

Six experiments demonstrated cross-modal influences from the auditory on the visual modality at an early level of perceptual organization. Participants had to detect a visual target in a rapidly changing sequence of visual distractors. A high tone embedded in a sequence of low tones improved detection of a synchronously presented visual target (Exp. 1), but the effect disappeared when the high tone was presented before the target (Exp 2). Rhythmically-based or order-based anticipation was unlikely to account for the effect because the improvement was unaffected by whether there was jitter (Exp 3), or a random number of distractors between successive targets (Exp 4). The facilitatory effect was greatly reduced when the tone was less abrupt and part of a melody (Exps 5 and 6). These results show that perceptual organization in the auditory modality can have an effect on perceptibility in the visual modality.

Information arriving at the sense organs must be parsed into objects and events. In vision, scene analysis or object segregation succeeds despite partial occlusion of one object by the other, shadows that extend across object boundaries, and deformations of the retinal image produced by moving objects. Vision, though, is not the only modality in which object segregation occurs. Auditory object segregation has also been demonstrated (Bregman, 1990). This occurs, for instance, when a sequence of alternating high- and low-frequency tones is played at a certain rate. When the frequency difference between the tones is small, or when they are played at a slow rate, listeners are able to follow the entire sequence of tones, but at bigger frequency differences or higher rates, the sequence splits into two streams, one high and one low in pitch. While it is possible to shift attention between the two streams, it is difficult to report the order of the tones in the entire sequence. Auditory stream segregation appears to follow, like apparent motion in vision, AKorte=s third law@ (Korte, 1915). When the distance in frequency between the tones increases, stream segregation occurs at longer stimulus onset asynchronies.

Bregman (1990) has described a number of Gestalt principles for auditory scene analysis in which he stressed the resemblance between audition and vision, since principles of perceptual organization like similarity (in volume, timbre, spatial location), good continuation, and common fate seem to play a similar role in the two modalities. Such a correspondence between visual and auditory organization principles raises an interesting question: Can the perceptual system utilize information from one sensory modality to organize the perceptual array in the other modality? Or, in other words, is scene analysis a cross-modal phenomenon

There are, of course, well-known examples of cross-modal influences where one may assume that the perceptual system indeed attributes information from different sensory modalities to a unitary event. There is a whole literature showing that arbitrary combinations of intermodal stimulus features tend to heighten perceptual awareness and lower reaction time when compared to unimodal presentation of those features (e.g., Hershenson, 1962; Nickerson, 1973; Posner, Nissen & Klein, 1976; Simon & Craft, 1970). Cross-modal combinations of features not only enhance stimulus processing, but can also change the percept. The prime example is the McGurk effect (1976) where discrepant speech information from sound and vision is presented. When

listeners hear 'baba' and at the same time see a speaker articulating 'gaga', they tend to combine the information from the two sources into 'dada'. Cross-modal interactions in the spatial domain have also been found. For example, synchronized sounds and light flashes with a different spatial location tend to be localized closer together (the ventriloquist effect). The common finding is that there is a substantial effect of the light flashes on the location of the sound (e.g., Vroomen, 1999), but under the right conditions, one can also observe that the sound attracts the location of the light (Bertelson & Radeau, 1981). The spatial attraction thus occurs both ways, and is rather independent of where endogenous or exogenous spatial attention is located (Bertelson, Vroomen, de Gelder, Driver, in press; Vroomen, Bertelson, de Gelder, submitted). Cross-modal interactions have also been observed in the perception of emotions. Listeners having to judge the emotion in the voice are influenced by whether a face expresses the same emotion or a different one, and the converse effect, in which subjects have to judge a face while hearing a congruent or incongruent voice, has also been shown to occur (de Gelder & Vroomen, in press; Massaro & Egan, 1996).

Massaro (1997) has described an influential model accounting for a wide variety of such intersensory effects. The input to the so-called Fuzzy Logical Model of Perception (FLMP) are modality-specific features (e.g., a rising second formant and lip closure for /ba/; increased pitch level and inflected mouth corners for >happiness=, etcetera) which are evaluated against a prototype. The information from the different modalities is then integrated according to a multiplicative rule after which a decision process determines the relative goodness of match of the stimulus. An assumption of the FLMP is that these cross-modal effects occur relatively late, since features are first evaluated separately in each modality (for a critical comment, see Vroomen & de Gelder, 2000). An intriguing question, though, is whether cross-modal interactions can occur at a more primitive level of perception, a level at which features do not yet exist. In animal studies, intriguing neurophysiological evidence has been found indicating that cross-modal interactions can take place at very early stages of sensory processing. Probably one of the best known sites of multi-modal convergence and integration is the superior colliculus, a midbrain structure known to play a fundamental role in attentive and orientation behaviour (see Stein & Meredith, 1993 for review). In humans as well, neurophysiological evidence of very early cross-modal interactions has been found. For

example, Giard and Peronnet (in press) found that tones synchronized with a visual stimulus can generate new neural activities in visual areas as early as 40 ms post stimulus onset, and that a visual stimulus can modulate the typical N1 auditory waveform in the primary auditory cortex at around 90-110 ms. Another example of an early interaction between vision and audition is that an angry face when combined with a sad voice can modulate, at 178 ms, the electric brain response typical for auditory Mismatch Negativity (the MMN; de Gelder, et al., 1999).

Given that these cross-modal electrophysiological effects arise very early in time, it seems at least possible that intersensory interactions can occur at primitive levels of perceptual organization. There is, to our knowledge, only one behavioural study showing that, at the level of scene analysis, perceptual segmentation in one modality can influence the concomitant segmentation in another modality. O'Leary and Rhodes (1984) used a display of six dots, three high and three low. The dots were displayed one-by-one, alternating between the high and low positions and moving from left-to-right. At slow rates, a single dot appeared to move up and down, while at faster rates two dots were moving horizontally, one above the other. A sequence that was perceived as two dots caused a concurrent auditory sequence to be perceived as two tones as well at a rate that would yield a single perceptual object when the accompanying visual sequence was perceived as a single object. The number of objects seen, thus influenced the number of objects heard, and they also found the opposite influence from audition to vision.

At first sight, this seems to be an example of a cross-modal influence on perceptual organization. However, at this stage it is not clear whether the cross-modal effect was truly perceptual, or whether it occurred because participants deliberately changed the interpretation about the sounds and dots. It is well known that there is a broad range of rates/tones at which listeners can hear, at will, one or two streams (van Noorden 1975). O'Leary and Rhodes presented ambiguous sequences, and it may therefore be the case that a cross-modal influence was found because perceivers changed their interpretation about the sounds and dots, but the perception may have been the same. For example, participants having the impression to hear two streams instead of one, may infer that vision should also be two streams instead of one. A voluntary decision would then

account for the cross-modal influence, but not a direct perceptual link between audition and vision.

In the present study, we pursued this question and investigated a phenomenon that, to the best of out knowledge, has so far not been reported in the literature. It is an illusion that occurs when an abrupt sound is presented during a rapidly changing visual display. Phenomenally, it looks as if the sound is pinning the visual stimulus for a short moment so that the visual display 'freezes'. In the present study, we explored this freezing phenomenon. We first tried to determine whether the freezing of the display is a perceptually genuine effect or not. Previously, Stein, London, Wilkinson and Price (1996) had shown that a sound can enhance the perceived visual intensity of a stimulus. This seems to be close analogue of the freezing phenomenon we observed. However, Stein et al. used a rather artificial and indirect measure of visual intensity (a >visual analogue scale= in which participants judged the intensity of a light by rotating a dial), and they could not find an enhancement by a sound when the visual stimulus was presented subthreshold. It is therefore unclear whether their effect is truly perceptual rather than post-perceptual.

In our experiments, we tried to avoid this difficulty by using a more direct estimate of visual persistence by simply measuring maximum speeded performance on a detection task. Participants saw a four-by-four matrix of flickering dots that was created by rapidly presenting four different displays, each containing four dots in quasi-random positions (see Figure 1). Each display on its own was difficult to see, because it was shown only briefly and immediately followed by a mask. One of the four displays contained a target to be detected. The target consisted of four dots that made up a diamond in the upper-left, upper-right, lower-left, or lower-right corner of the matrix. The task of the participants was to detect the position of the diamond as fast and as accurately as possible. In Experiment 1, we investigated whether the detectability of the target could be improved by an abrupt sound presented together with the target. Participants in the experimental condition heard at the target display a high tone, but at the other 4-dots displays (the distractors) a low tone. In the control condition, participants heard only low tones. The idea was that the high tone in the sequence of tones was likely to segregate from the low tones, and that under these circumstances, it would increase the detectability of the target display.

EXPERIMENT 1

Method

Participants. Sixteen participants, all first year students from Tilburg University, received course credits for their participation. They all had normal, or corrected to normal vision.

Stimuli. *Visual display*. The visual display was a 4 x 4 matrix of quasi-randomly flickering small white dots presented on the dark background of a 15 inch computer screen (Olivetti DSM 60-510). The matrix measured 4.2 by 4.2 cm and was viewed from a distance of 55 cm. The size of each of the dots was 4 x 4 pixels. The flicker of the matrix was created by displaying successively four different displays at a high speed. Each display showed four unique dots of the matrix. When overlaid, the displays would make up the complete matrix. The third of the four 4-dots displays contained the to-be-detected diamond, either in the upper-left, upper-right, lower-left, or lower-right corner of the matrix. Each 4-dots display was shown for 97 ms (or 7 refresh cycles on a screen with a vertical retrace of 72 Hz), and was immediately followed by a mask which consisted of the full 4 x 4 matrix of dots. The duration of the mask was also 97 ms, and it was followed by a dark screen for 60 ms after which the next display was shown. One 4-dots display was thus shown for 97 ms every 254 ms, and within the sequence of four 4-dots displays, the target was visible for 97 ms every 1016 ms. The sequence was repeated continuously with no interruption until a response was given, or until a maximum of 10 cycles was reached.

Auditory sequence. Subjects either heard a sequence of four low tones of 1000 Hz (denoted as LLLL), or a LLHL-sequence (H = high tone) in which the H is a high tone of 4 ST (or 1259 Hz) above the L. Each tone was, like the visual 4-dots displays, 97 ms in duration with a 5 ms fade-in and fade-out to avoid clicks. All tones were presented in exact synchrony with the 4-dots displays, so the SOA between the tones was 254 ms. The high tone, if present, was always presented in synchrony with the target display.

Procedure and Design. It is well-known that auditory segregation requires time to build up (Bregman, 1990). Initially, subjects are able to follow an entire sequence of tones, but only after a short while, can they hear two streams, one high in pitch and one low in pitch. In order to allow segregation to build up, a random number of four to eight

sequences of four tones (LLLL or LLHL) was played before the actual 4-dots displays were shown. These >warm-up= sequences were presented together with the mask (i.e. the full matrix of sixteen dots shown for 194 ms), followed by a blank screen (60 ms). The warm-up sequences were then immediately followed by the same sequence of tones with the 4-dots displays. So at the time subjects saw a target, they could already have imposed an auditory organization on the tones.

There were 20 trials for each of the four positions of the diamond, and so the whole experiment consisted of 160 experimental trials; 80 for the LLLL-sequence, and 80 for the LLHL-sequence. All trials were pseudo-randomly mixed. Within a sequence of 16 consecutive trials, each of the eight possible trial-combinations (four positions of the diamond x two sound sequences) was presented twice. Before testing, participants were given 16 practice trials. The first eight practice items were presented at a slow rate (half the speed of the experimental trials), the others were presented at the same rate as the experimental trials. There was a short pause halfway. Testing lasted about 25 min.

Participants were tested individually in a dimly lit sound-shielded booth. They were instructed to detect as fast and accurately as possible the position of the diamond in the display by pressing, with their left or right, middle- or index-finger, one out of four spatially corresponding keys on a keyboard (e.g., left middle finger for a diamond in the upper-left corner, and the right index finger for a diamond in the lower-right corner). Participants were told about the two possible sound sequences (LLLL or LLHL), and they were also told that the high tone was synchronized with the target display.

Results and Discussion

For each subject and each condition, two response measures were determined: One was the percentage of correct responses, the other was the Number of Targets Shown (NTS) before a response was made. The NTS was determined for correct responses only, and, in all experiments, if the NTS deviated more than plus or minus 2 SDs from the individual grand average, it was removed from the analyses. These data were then submitted to an Analysis of Variance (ANOVA) with sequence of tones (LLLL vs. LLHL) as within-subjects variable.

All participants performed above chance level (i.e., above 25% with $p < .01$). The average proportion of correct responses was 55% with the LLLL-sequence, and 66%

with the LLHL-sequence, $F(1,15) = 7.84$, $p < .015$. Twelve out of sixteen participants performed better with the LLHL-sequence, one performed at the same level, and three participants performed worse, $Z = 2.06$, $p < .025$. Participants were not only more correct with the LLHL-sequence, but also required less NTS. The average NTS with the LLLL-sequence was 3.32, and with the LLHL-sequences it was 2.86, $F(1,15) = 4.88$, $p < .05$. Twelve out of sixteen subjects responded faster with the LLHL-sequence, and four slower. Participants were thus, on average, faster and more accurate when an H was presented with the visual target.

One possible interpretation of this result is that the H indeed enhanced the visibility of the target display. When participants were asked informally, most of them indeed confirmed that they had experienced the freezing phenomenon as described before. On the other hand, another interpretation of our result is that H acted as an attentional cue when to expect the target display. If indeed H is similar to a cross-modal attentional cue, one would expect that other cues that reduce uncertainty about target onset may also enhance performance. Our next experiment tested for this possibility.

EXPERIMENT 2

One possibility is that the H in Experiment 1 acted as a warning signal so that participants knew when to expect the target. As an example, it is well-known from the cross-modal attentional cuing literature that an auditory cue which precedes a visual target between 100-300 ms can enhance responding to a visual target (cf. Spence & Driver, 1997). If attentional cuing is at stake, one would expect that if H precedes the target by one display (i.e., 254 ms), performance should improve because uncertainty about target onset is reduced and because participants are allowed time to prepare for the upcoming target. On the other hand, if the freezing phenomenon is a perceptual phenomenon, one expects that synchrony between tone and visual display is of critical importance. In that case, one may expect that when H precedes the target and is synchronized with a distractor, it may freeze the distractor display so that performance may even deteriorate.

Method

Participants. Sixteen new participants drawn from the same population as in Experiment 1 were tested.

Stimuli and Design. The auditory and visual materials were exactly as in Experiment 1, except that the LLHL sequence was replaced with LHLL so that the H now preceded the target by one display (or 254 ms). The deviating tone thus now accompanied a distractor. Participants were informed about the temporal relation between H and the target, and they were told that they should use the deviant tone as a warning signal when to expect the target. As before, they were shown, in slow-motion, the relation between tone and target. All other aspects were exactly the same as in Experiment 1.

Results and Discussion

All participants performed above chance level. The average proportion of correct responses was 55% with the LLLL-sequence, but only 52% with the LHLL-sequence, $F(1,15) = 4.36$, $p = .05$. Eleven out of sixteen participants performed worse with the LHLL-sequence, one performed at the same level, and four performed better, $Z = 1.54$, NS. Participants required somewhat less NTS with the LHLL sequence, but this effect was not significant. The average NTS with the LLLL-sequence was 3.14, and with the LHLL-sequences it was 3.10, $F < 1$. Seven out of sixteen participants required more NTS with the LHLL-sequence, nine required less, $Z = 0.25$, NS.

The results of Experiment 2 thus show that participants made slightly more errors when H preceded the target display. This allows to exclude the possibility that a deviant tone simply acts as a warning signal because one would expect, then, that performance should improve because participants were given prior information about when to expect the target.

As an aside, an interesting observation was that a number of participants remarked that they were able to see the four random dots of the distractor display that was presented with the deviant tone. This was remarkable, because subjectively speaking, this seemed almost impossible when no abrupt sound was heard. This is at least suggestive in showing that the freezing phenomenon may even occur when participants are looking for a different display to appear at a different time.

However, so far we have not completely ruled out an attentional explanation. One possibility[1] is that there is rhythmically-based anticipation. It may be that the freezing

phenomenon is only observed when the target can be anticipated. If that is indeed the case, then jitter between tones should have a disruptive effect.

EXPERIMENT 3

Experiment 3 was similar to Experiment 1, except that there was an extra condition in which there was jitter between successive tones that disrupted the rhythm of the sequence. If rhythmically-based anticipation is at the heart of the freezing phenomenon, jitter should disrupt, or at least attenuate, the facilitatory effect of the high tone.

Method

Participants. Sixteen new participants were tested.

Stimuli and Design. The visual materials and the auditory tone sequences were the same as in Experiment 1. In the no-jitter condition, the Stimulus Onset Asynchrony (SOA) between successive tones and 4-dots display was, as before, 254 ms (i.e., 97ms for the 4-dots display, 97 ms for the mask, and 60 ms for the black screen). In the jitter condition, the SOA between successive tones and displays varied randomly from 204 ms to 304 ms in equally-likely steps of 25 ms. The visual 4-dots displays (97 ms) remained synchronized with the tones and were followed by a mask of the same duration (97 ms), but the duration of the black screen varied between 10ms and 110 ms depending on SOA. Successive SOA=s in the jitter condition were never the same, so rhythmically-based anticipation should have been very difficult.

The experiment comprised two blocks (jitter versus no-jitter) of 96 trials each. Within each block, there were 48 trials with the LLLL-sequences of tones, and 48 trials with the LLHL-sequences of tones, 12 for each position of the target. Jitter or no-jitter was blocked, and sequence of tones (LLLL versus LLHL) was randomized as before within a block. Half of the participants started with the jitter condition followed by the no-jitter condition, for the other half the order was reversed. Before each block was started, participants received 20 practice trials.

Results and Discussion

All participants performed above chance level (at $p < .01$). The average percentage of correct responses and the NTS is presented in Table 1, upper part. A two-way ANOVA

with jitter and sequence of tones as within-subjects factors was carried out on the percentage of correct responses and the NTS. In both analyses, there was a main effect of sequence of tones because target detection with the LLHL-sequence of tones was, on average, more correct, $F(1,15) = 8.38$, $p < .02$, and required less NTS, $F(1,15) = 5.32$, $p < .04$. The effect of jitter and the interaction between jitter and sequence of tones never even approached significance (All F=s < 1).

Experiment 3 thus replicated Experiment 1 in showing that H improved visual target detection. Moreover, the facilitatory effect did not seem to depend on the rhythmic regularity of the tones (or of the visual displays) because there was no hint that jitter disrupted the facilitatory effect. At first sight, this result rules out rhythmically-based anticipation as the primary reason for the facilitatory effect. However, two objections against this interpretation can be raised. First, one might argue that the variations in SOA as used in the present experiment were not substantial enough. Thus, although rhythmic regularity was disturbed, it might still be present and cause the facilitatory effect. Second, participants may anticipate the occurrence of a target by counting the number of distractor displays or tones. So far, targets were always followed by three distractor displays. Participants may count those displays (or their accompanying tones) and anticipate on the basis of serial order when the target is to appear. If indeed such order-based anticipation is at the basis of the freezing phenomenon, then varying the number of distractors should disrupt the effect.

EXPERIMENT 4

Experiment 4 was similar to the previous one, except that instead of jitter, a random number of distractor displays accompanied by low tones was presented between successive targets. The appearance of the target was thus much less predictable than in the case in which the number of distractors was fixed. If order-based anticipation is to account for the freezing phenomenon, then varying the number of distractors should disrupt the facilitatory effect of H.

Method

Participants. Sixteen new participants were tested.

Stimuli and Design. The visual materials and the auditory sequences of tones were the same as in Experiment 1. In the fixed-distractor condition, there were, as before, three distractor displays between successive targets. In the random-distractor condition, the number of distractors displays and their accompanying low tones varied between successive targets within a single trial from two to six. The number of distractors between successive target displays was thus never the same, so order- and/or rhythmic-based anticipation should have been extremely difficult in this condition.

The experiment comprised two blocks (fixed- versus random-number of distractors) of 96 trials each. Within each block, there were 48 trials with the LLLL-sequences of tones, and 48 trials with the LLHL-sequences of tones, 12 for each position of the target. Fixed- or random-number of distractors was blocked, and sequence of tones (LLLL versus LLHL) was randomized as before within a block. Half of the participants started with the fixed-distractor condition followed by the random-distractor condition, for the other half the order was reversed. Participants received 20 practice trials before each block.

Results and Discussion

All participants performed above chance level (at $p < .01$). The average percentage of correct responses and the NTS is presented in Table 1, lower part. A two-way ANOVA with number of distractors and sequence of tones as within-subjects factors was carried out on the percentage of correct responses and the NTS. In the analysis on accuracy, there was a main effect of sequence of tones because target detection with the LLHL-sequence was, as before, more correct, $F(1,15) = 10.81$, $p < .005$. There was no main effect of whether the number of distractors was fixed or varied, $F < 1$, and the interaction between number of distractors and sequence of tones was not significant, $F(1,15) = 2.33$, $p = .15$. Inspection of Table 1 suggests that, if anything, the facilitatory effect of H was bigger, and not smaller, when the number of distractors varied.

In the analysis of the NTS, no effect was significant (all F=s < 1). Although there was no effect of NTS, there was certainly no sign of a speed-accuracy trade-off because the average NTS was less with random-distractors than with fixed-distractors.

Experiment 4 thus replicated Experiments 1 and 3 in showing that the H improved visual target detection. Moreover, it appeared that the facilitatory effect of H did not hinge on whether the target could be anticipated or not. If anything, the enhancing effect

increased when target appearance was unpredictable (potentially because there was more room for improvement). This result therefore rules out order-based anticipation as the main reason for the facilitatory effect.

So far, then, we have shown that H improves detection of a synchronized visual target and that cross-modal attentional cuing is unlikely to account for the effect. However, thus far we have not shown that the effect depends on the auditory organization of the tones. When participants listened to a sequence of LLHL-tones, they may either have heard a single stream, or they may have heard two streams, one with low tones, and another with a high tone. There has, however, been no experimental demonstration of that, and it may be that whether or not a tone segregates is just epiphenomenal. In fact, it may well be the case that any tone that is different from other tones, whether it segregates or not, may cause the freezing phenomenon. In the next experiments, we therefore tested whether the auditory organization of the tones is essential.

EXPERIMENT 5

An obvious possibility to prevent segregation is to increase the duration between successive tones, or to decrease the frequency difference between the high and low tones. Listeners are then more likely to hear the sequence as a temporally coherent one. However, as shown by Noorden (1975), listeners can, at will, perceive a sequence either as temporally coherent or as what he called 'fission'. Fission, but not temporal coherence, can be heard quite easily, no matter what the size of the tone interval is. Listeners can thus quite easily segregate a high tone from a low tone, even when the difference between the tones is quite small. We therefore refrained from the obvious possibility of changing the duration or frequency difference between the tones, because participants may segregate the tones anyway in order to maximize performance on the task.

Instead, we presented participants a LMHL-sequence in which M stands for a middle tone with a 2 ST difference from L. There are two reasons why segregation of H in the LMHL-sequence is less likely to occur than in the LLHL-sequence. The first is that the H in LMHL is less abrupt than in LLHL, and it is well-known that abrupt sounds segregate more easily than less abrupt sounds (Bregman, 1990). The second reason why H is

unlikely to segregate in the LMHL-sequence is that we told participants that the LMHL-sequence is the beginning of the tune, 'Frère Jacques'. The notion is that when listeners perceive this sequence as a melody, then H is unlikely to segregate because it is captured as an indispensable part. For these reasons, we expected less segregation of H in the LMHL-sequence than in a LLHL-sequence. The crucial advantage of this procedure is that the H is the same sound in both sequences of tones. Thus, at the time the visual target is shown, exactly the same stimuli are heard and seen. The only difference between conditions is the tone preceding H. If segregation is critical for the facilitatory effect, one expects that target detection will be more difficult in the LMHL-sequence than the LLHL-sequence. On the other hand, if the sequential organization of the tones it not important, it should not matter whether H is part of a melody or not.

Method

Participants. Sixteen new participants, all first year students, were tested. As before, all had normal, or corrected to normal vision.

Stimuli and Design. The stimuli and design were exactly the same as in Experiment 1, except that participants heard instead of a LLLL-, a LMHL-sequence. There were thus two sequences of tones randomly mixed in a block: a LLHL- and a LMHL-sequence. The LMHL-sequence was introduced to listeners as the beginning of the tune, 'Frère Jacques', the other as a sequence of tones without reference to a melody. The M was a pure tone of 1122 Hz (2 ST above L), with a duration of 97 ms and with a 5 ms fade-in and fade-out.

Results and Discussion

All participants performed above chance level. The average proportion of correct responses was 52% with the Frère-Jacques tune, and 62% with the LLHL-sequence, $F(1,15) = 11.49$, $p < .004$. Fourteen out of sixteen participants performed better with the LLHL-sequence, and two performed at the same level, $Z = 3.47$, $p < .005$. Participants were not only more correct, but also required less NTS with the LLHL-sequence. The average NTS with the Frère-Jacques tune was 2.63, and with the LLHL-sequences it was 2.46, $F(1,15) = 13.24$, $p < .002$. Twelve out of sixteen participants required less

NTS with the LLHL-sequence, three required more, and one participant required equal amounts, $Z = 2.06$, $p < .025$.

These results thus show that the perceptual organization of the sequence of tones plays a critical role. When H was heard as part of a melody, the task was much harder than when exactly the same tone was not part of a melody. These results therefore show that the auditory organization of the sequence of tones is indeed of importance for observing the freezing phenomenon. Our next experiment explored this further.

EXPERIMENT 6

The results of Experiment 5 are crucial for the interpretation of the phenomenon. The LMHL-sequence made visual target detection more difficult because, we reasoned, it made segregation of H unlikely. Segregation was unlikely to occur for two reasons: one was that, compared to LLHL, the H in LMHL was less abrupt, the other was that the LMHL sequence was a tune. A potential problem with the tune explanation is that one runs the risk that when participants are told that they will hear a tune, they are actually performing two tasks at the same time: One is, indeed, trying to hear the sequence as a tune; the other is detecting the visual target. Trying to hear the sequence as a tune may then interfere with target detection because it requires a certain amount of limited processing resources.

To investigate whether this might be a potential difficulty, we replicated Experiment 5 and varied instructions. In one condition we stressed, as before, that the LMHL-sequence was the beginning of the tune, Frère-Jacques. But in the other condition we refrained from that and made no reference to the >tune-ness= of the LMHL-sequence. If the instructions caused the difference between the sequences of tones, it should disappear, or at least attenuate, when no mention to the tune-ness of the LMHL-sequence is made. On the other hand, if the abruptness of H is crucial, instructions should have no effect.

Moreover, we also explored a range of stimulus parameters under which the facilitatory effect of H can be observed. To do so, we varied the display times of the dots and the mask. One may expect that performance will improve when the display time of the target is increased and the display time of the mask is decreased. The question was whether the facilitatory effect of H critically depends on task difficulty.

Method

Participants. Two groups of 16 students each were tested. One group received the same instructions as in Experiment 5 in which the LMHL-sequence was introduced as the beginning of the tune, 'Frère Jacques'. In the other group, no reference to the tune was ever made.

Stimuli and Design. Participants heard, as in Experiment 5, a LLHL or a LMHL (Frère-Jacques) sequence of tones. These sequences were combined with three possible 4-dots/mask display times: a 97/97 ms 4-dots/mask display time as used in all previous experiments, and a 83/111 ms and 111/83 ms 4-dots/mask display time.

For each display time and sequence of tones, there were eight trials for each of the four positions of the diamond. The whole experiment therefore consisted of 192 experimental trials; 96 for the LMHL-sequence, and 96 for the LLHL-sequence. The trials were randomly mixed and within a block of 48 consecutive trials, each of the 24 different trials appeared twice. There was a short pause half way. Before actual testing began, a short practice session was given.

Results

The average proportion of correct responses and the NTS are presented in Table 2. As before, performance was better with the LLHL-sequence of tones than with the LMHL-sequence. Instructions and the different target/mask display times had no effect is this effect.

A three-way mixed ANOVA with instruction as between-subjects, and display time and sequence of tones as within-subject variables was carried out on the percentage of correct responses and the NTS. In the analysis on accuracy, there was no overall difference between the groups that received the different instructions, $F < 1$. As expected, there was main effect of display time, $F(2,60) = 83.53$, $p < .001$, because performance improved when targets were displayed for a longer duration. There was also a main effect of sequence of tones, $F(1,30) = 13.10$, $p < .001$, because performance was better with the LLHL-sequence than with the LMHL-sequence of tones. All other effects were non-significant. Thus, most importantly, instructions did not change the difference between the sequences of tones as indicated by a non-significant

interaction between instruction x sequence of tones, $F < 1$. Moreover, the interaction between display time and sequence of tones, $F(2,60) = 2.18$, $p = .12$, and the second-order interaction between instruction, display times, and sequence of tones were non-significant, $F(2,60) = 1.10$, $p = .31$.

The same pattern was found in the corresponding ANOVA on the NTS. The effect of display time was significant indicating that participants required less NTS when targets were shown for a longer duration, $F(2,60) = 13.15$, $p < .001$. The effect of sequence of tones was significant, $F(1,30) = 7.37$, $p < .011$, because less targets were seen when the LLHL-sequence was heard instead of the LMHL-sequence of tones, and all other effects were non-significant (all F=s < 1).

The present results replicate and extend those of Experiment 5. As before, target detection was more difficult when the high tone was part of the LMHL-sequence. Whether or not instructions specified that the LMHL-sequence was the beginning of 'Frère Jacques' had no effect on this. Varying the overall difficulty of the task had also no effect. This suggests that the abruptness of H, rather than the tune-ness of the LMHL-sequence is of crucial importance for the improvement of the detectability of the target.

General discussion

In the present study we demonstrated a new cross-modal phenomenon: the detectability of a visual stimulus could be enhanced by a synchronously presented abrupt tone. This so-called freezing phenomenon was closely related to the perceptual organization of the tone in the auditory modality: the effect was observed when the tone could easily segregate from a sequence of tones, but it was greatly attenuated (or even disappeared) when exactly the same tone was less abrupt or part of a melody. The phenomenon is unlikely to be accounted for by cross-modal attentional cuing because the effect disappeared when the abrupt sound preceded the target by an SOA at which one may expect a cross-modal attentional cuing effect, and it was unaffected by whether the onset of the target was predictable or not. Since our method allowed us to obtain a direct measure of visibility, these results strongly suggest that the freezing phenomenon is a perceptually genuine effect.

Our findings are similar to the observations made by Stein, London, Wilkinson and Price (1996) who reported that a sound can enhance the perceived visual intensity of a stimulus. Our study extends this observation because we used a different measure that relied on maximum speeded performance instead of subjective judgement. Moreover, we showed that the phenomenon was closely related to the perceptual organization of the sound in a sequence of tones. Consequently, we would predict that the results of Stein et al. can be modulated by the perceptual organization of the sound that is synchronized with the visual display.

Our results are also in line with those of O=Leary and Rhodes (1984) who reported that the perceptual organization of tones could influence the perceptual organization of moving dots. They found that when a sequence of high and low tones was heard as two streams, a dot that moved up and down was more likely to be seen as two streams of dots moving horizontally. Other examples of this cross-modal principle were recently demonstrated by Sekuler, Sekuler and Lau (1997). They found that two disks moving towards one another, coinciding, and then moving apart were perceived as >bouncing= when a sound was presented at the point of visual coincidence. When there was no sound, it appeared as if the disks continued in their original direction. Our results show that these cross-modal correspondences in perceptual organization have other profound consequences, namely, a tone that segregates from an auditory stream can segregate a synchronized visual stimulus from a visual stream.

At present, there is a wide variety of neurophysiological findings showing interactions between vision and audition in several cortical and subcortical areas. In humans, several studies have found neural sites of multi-sensory convergence. In synaesthete subjects (subjects making colour-word associations), a Positron Emission Tomography (PET) study has identified several cortical areas of interaction in the parieto-occipital junction and in the right prefrontal cortex (Paulesu et al, 1995). Magnetoencephalography (MEG) recordings have reported audio-visual interactions in the right parieto-temporal area (Sams & Imada, 1997). And with spatiotemporal analysis of event-related potentials (ERP), several distinct audio-visual interaction components have been identified in visual areas, auditory cortex, and right fronto-temporal areas (de Gelder et al., 1999; Giard & Peronnet, in press).

Animal studies have found polymodal cells which may provide a physiological basis for some of those cross-modal effects (Meredith & Stein, 1986). Multi-sensory neurons have been found in the deep layers of the superior colliculus in cat, monkey, and rat, but also in cortical areas (e.g., Wallace, Meredith & Stein, 1992). These cells not only respond to inputs from several modalities, but they also integrate information from different modalities by increasing the number of impulses in a multiplicative ratio when presented with multi-modal inputs (Wallace, Wilkinson & Stein, 1996).

One may speculate that cross-modal interactions in general, and the freezing phenomenon in particular, are consistent with a perceptual mechanism that makes coherent interpretations about auditory and visual information that originates from a single object or event. From an ecological point of view, it seems valid to assume that multi-sensory stimulation that covaries in place and time originates from a single object. Perceptual evaluation in one modality may then have consequences in other modalities so that coherence is maintained. A sound that segregates in the auditory modality may for that reason provoke segregation in the visual modality. Ventriloquism is another demonstration of this principle in the sense that discrepant information about the location of synchronised auditory and visual events is integrated into a coherent representation of the world. Future studies could demonstrate whether the freezing phenomenon can be observed in other modalities than the auditory and visual one (e.g., visual-haptic), and whether cross-modal effects can be found from vision on audition. For example, it may be the case that tone detection can be enhanced when an accompanying visual scene segregates.

References

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

Bertelson, P., Vroomen, J., de Gelder, B., & Driver, J. (in press). The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & Psychophysics.*

Bregman, A. S. (1990). Auditory scene analysis. Cambridge, MA: The MIT Press.

Gelder, B., Böcker, K. B. E., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from face and voice: early interaction revealed by human electric brain responses. Neuroscience Letters, 260, 133-136.

Gelder, B. de, & Vroomen, J. (in press). Emotions by ear and eye. Cognition and Emotion.

Giard, M. H., & Peronnet, F. (in press). Auditory-visual integration during multi-modal object recognition in humans: a behavioural and electrophysiological study. Journal of Cognitive Neuroscience.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. Journal of Experimental Psychology, 63, 289-293.

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

Massaro, D. W. (1997). Perceiving talking faces: From speech perception to a behavioral principle. The MIT Press.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. Psychonomic Bulletin & Review, 3, 215-221.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. Nature, 264, 746-748.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. Journal of Neurophysiology, 56, 640-662.

Noorden, L. P. A. S. van (1975). Temporal coherence in the perception of tone sequences. Unpublished Doctoral Dissertation, Technische Hogeschool Eindhoven, The Netherlands.

Nickerson, R. S. (1973). Intersensory facilitation of reaction time: Energy summation or preparation enhancement? Psychological Review, 80, 168-173.

O=Leary, A., & Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception. Perception and Psychophysics, 35, 565-569.

Paulesu, E., Harrison, J., Baron-Cohen, S., Watson, J. D. G., Goldstein, L., Heather, J., Frackowiak, R. S. J., & Frith, C. D. (1995). The physiology of coloured hearing. A PET activation study of colour-word synaesthesia. Brain, 118, 661-676.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. Psychological Review, 83, 157-171.

Sams, M., & Imada, T. (1997). Integration of auditory and visual information in the human brain: neuromagnetic evidence . Society for Neuroscience Abstracts, 23, 1305.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. Nature, 385, 308.

Simon, J. R., & Craft, J. L. (1970). Effects of an irrelevant auditory stimulus on visual choice reaction time. Journal of Experimental Psychology, 86, 272-274.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. Perception & Psychophysics, 59, 1-22.

Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. Journal of Cognitive Neuroscience, 8, 497-506.

Stein, B. E., & Meredith, M. A. (1993). The merging of the senses. Cambridge, MA: The MIT Press.

Vroomen, J. (1999). Ventriloquism and the nature of the unity assumption. In G. Aschersleben, T. Bachmann, and J. Müsseler (Eds.) Cognitive contributions to the perception of spatial and temporal events (pp. 388-394). North-Holland: Elsevier

Vroomen, J., Bertelson, P., & de Gelder, B. (1998). A visual influence in the discrimination of auditory location. Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP=98), (pp 131-135), Terrigal-Sydney.

Vroomen, J., Bertelson, P., & de Gelder, B. (submitted). Visual bias of auditory location and the role of exogenous automatic attention.

Vroomen, J., & de Gelder, B. (2000). Cross-modal integration: a good fit is no criterion. Trends in Cognitive Sciences, 4, 37-38.

Walace, M. T., Meredith, M. A., & Stein, B. E. (1992). Integration of multiple sensory modalities in cat cortex. Experimental Brain Research, 91, 484-488.

Walace, M. T., Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. Journal of Neurophysiology, 76, 1246-1266.

Table 1

Mean Percentage of Correct Responses and Number of Targets Shown (NTS) in Experiment 3 and Experiment 4

	No Jitter		Jitter	
Sequence of tones	%	NTS	%	NTS
LLLL	50	4.53	50	4.35
LLHL	58	4.18	60	4.25
Difference	8	0.35	10	0.11
	Fixed-number of distractors		Random-number of distractors	
LLLL	49	3.47	46	3.32
LLHL	53	3.41	56	3.23
Difference	4	0.06	10	0.10

Table 2

Mean Percentage of Correct Responses and Number of Targets Shown (NTS) as a Function of the Tone Sequence and Target/Mask Display Times in Experiment 6

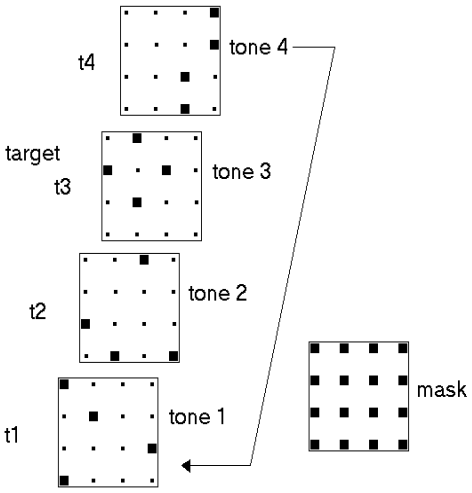
Sequence of tones	Target/Mask display times (in ms)					
	83/111		97/97		111/83	
	%	NTS	%	NTS	%	NTS
	Instructions specifying LMHL as Frère-Jacques					
LLHL	54	4.44	63	3.97	76	3.59
LMHL (Frère-Jacques)	46	4.71	62	4.39	68	4.07
Difference	8	0.26	1	0.42	8	0.48
	Instructions with no references of LMHL as Frère-Jacques					
LLHL	49	4.77	59	4.54	66	4.07
LMHL	42	5.1	57	4.94	65	4.44
Difference	7	0.32	3	0.39	1	0.37

Figure captions

Figure 1. A simplified representation of a stimulus sequence. Big squares represent the dots shown at time t; small squares were actually not seen, but are only there to show the position of the dots within the matrix. The 4-dots displays were shown for 97 ms each. Not shown in the figure is that each display was immediately followed by a mask (the full matrix of 16 dots) for 97 ms, followed by a dark blank screen for 60 ms. The target display (in this example the diamond in the upper-left corner) was presented at t3. The sequence of the four 4-dots displays was repeated without interruption until a response was given. Tones (97 ms in duration) were synchronized with the onset of the 4-dots displays. Also not shown in the figure is that four-to-eight tone sequences were presented before the 4-dots displays were seen. During this >warm-up= period, tones were synchronized with the mask (presented for 194 ms) followed by the blank screen (60 ms). Participants thus already had heard the sequence of tones several times before the 4-dots displays were shown.

Footnotes

Footnote 1. As suggested by one of the anonymous reviewers.


t4
tone 4
target
t3
tone 3
t2
tone 2
t1
tone 1
mask