Natural cross-modal mappings between visual and auditory features

Karla K. Evans

Brigham and Women's Hospital and Harvard Medical School, Cambridge, MA, USA

Anne Treisman

Psychology Department, Princeton University, Princeton, NJ, USA

The brain may combine information from different sense modalities to enhance the speed and accuracy of detection of objects and events, and the choice of appropriate responses. There is mounting evidence that perceptual experiences that appear to be modality-specific are also influenced by activity from other sensory modalities, even in the absence of awareness of this interaction. In a series of speeded classification tasks, we found spontaneous mappings between the auditory feature of pitch and the visual features of vertical location, size, and spatial frequency but not contrast. By dissociating the task variables from the features that were cross-modally related, we find that the interactions happen in an automatic fashion and are possibly located at the perceptual level.

Keywords: cross-modal integration, auditory, visual, perceptual level

Citation: Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision, 10*(1):6, 1–12, http://journalofvision.org/10/1/6/, doi:10.1167/10.1.6.

Introduction

Objects and events in the environment typically produce a correlated input to several sensory modalities at once. The perceptual system may disambiguate and maximize the incoming information by combining inputs from several modalities. To do this, it must determine which signals originate from the same source by detecting correspondences or mismatches between the different sensory streams. Spatial and temporal coincidence are the usual cues for integration, but the brain may also rely on a feature correspondence between the different sensory inputs (see Welch & Warren, 1986). The most dramatic examples occur in synesthetes, who often experience stimuli both in the appropriate modality and in another. For example, a letter may produce an additional sensation of color, or a visual shape may evoke a taste or a smell. In non-synesthetes, cross-modal metaphors are also common—for example a bitter smile, a loud tie, a heavy odor, suggesting that some mappings may be natural to most humans.

Over the past 40 years, studies in adults and children have investigated how both static and dynamic stimulations in one modality can modulate the perception of information in another modality. Pitch is involved in a number of cross-modal correspondences, and this is also the auditory feature that we investigated. In 1883, Stumpf concluded that all world languages apply the labels "high" and "low" to the pitch of tones. Bernstein and Edelstein (1971) were the first to measure the cross-modal congruence between pitch and visual vertical position, subsequently replicated by Melara and O'Brien (1987), Ben-Artzi and Marks (1995), and Patching and Quinlan (2002). Participants classified visual stimuli as high and low more quickly when the visual stimulus was accompanied by a tone that was congruent rather than incongruent (e.g., high pitch with high position rather than low). Even 6-month-old babies exhibit this cross-modal correspondence (Braaten, 1993), and young infants "match" visual arrows pointing up or down with tones sweeping up or down in frequency (Wagner, Winner, Cicchetti, & Gardner, 1981).

Auditory pitch is also matched to other visual attributes. One is the lightness or darkness of a surface, with participants showing faster responses to a higher pitch presented with a lighter surface and a lower pitch with a darker surface (Marks, 1987; Martino & Marks, 1999; Melara, 1989). Another is size: In a perceptual matching paradigm, 9 year-olds matched high pitch tones with small sizes and low pitch with large sizes (Marks, Hammeal, & Bornstein, 1987). Three year olds reliably matched high pitch tones with small bouncing balls and low pitch tones with large ones (Mondloch & Maurer, 2004), and adults judged the size of a variable visual stimulus relative to a standard more rapidly when the irrelevant sound frequency presented simultaneously was congruent with the size of the visual stimulus (e.g., high pitch tone with a small visual stimulus; Gallace & Spence, 2006). Finally, pitch can also generate cross-modal correspondences with visual shape: high pitch is matched to angular rather than rounded shapes (Marks, 1987). In addition to auditory pitch, loud sounds can improve the perception of bright

doi: 10.1167/10.1.6 Received February 2, 2009; published January 12, 2010 ISSN 1534-7362

lights and large objects, whereas soft sounds facilitate the perception of dim lights and small objects (Marks, 1987; Smith & Sera, 1992).

At what level might such cross-modal correspondences arise? Some cross-modal interactions may be based on similar neural codes, for example for dimensions like loudness and brightness that can be described in terms of more or less an increase in stimulation produces an increase in firing in each modality. Melara (1989) used multidimensional scaling to look for changes in the perceptual similarity relations between tones or lightness induced by concurrent stimuli in the other modality. He found none and concluded that the interactions result from failures of separability at the decision rather than the perceptual level. Some correspondences may arise through frequent associations in everyday experience. For example, larger objects resonate in lower pitch. The image of a cat is associated with the sound "meow" (Long, 1977). Pre-linguistic children as well as adults may show these learned perceptual associations (Braaten, 1993; Wagner et al., 1981). Verbal labels may also mediate cross-modal links. Melara and Marks (1990) presented the visual bigrams "LO" and "HI" simultaneously with sounds of different pitch and found effects that are unlikely to reflect inherent perceptual similarities. For stimuli such as these, the representations may share only a more abstract polarity (Garner, 1974, 1976; Martino & Marks, 1999). It seems then that cross-modal correspondences can arise at almost any level.

The speeded classification paradigm is often used to probe for interactions between different dimensions or modalities. The task requires the rapid identification of a stimulus in one modality (or on one dimension of a unimodal stimulus), while an irrelevant stimulus in another modality (or another dimension of a unimodal stimulus) is ignored. If the two modalities are automatically registered, it may be difficult to attend selectively to one of the pair and variation in the irrelevant modality may affect latencies to the relevant one. Any overall increase in latencies induced by variation on the irrelevant dimension is known as Garner interference, but if there is a correspondence between the polarities on the two dimensions, specific pairings may also show interference when the mapping is incongruent and facilitation when it is congruent.

In our studies, we selected four cross-modal mappings of auditory pitch to visual position, size, spatial frequency, and contrast. We used speeded classification to compare performance between unimodal visual and auditory presentations and bimodal simultaneous presentations in which the pairings were either congruent or incongruent. Having identified which pairs of visual dimensions show congruency effects with concurrent tones and can therefore be presumed to interact at some level, we used a new experimental paradigm to test whether these interactions arose automatically and possibly suggest an interaction at the perceptual level rather than being mediated by shared labels or responses. For each pair of dimensions, we used both a "direct" task in which, as in previously published experiments, the task was to discriminate the stimuli on the dimensions on which the hypothesized correspondence would be shown, and also an "indirect" task, in which the task was to discriminate the stimuli on some other orthogonal dimension (see Figure 1).

The direct tasks allow the corresponding cross-modal features to interact at any level between sensory registration and response. In the indirect tasks, the potentially corresponding dimensions are orthogonal both to the task and to the correct response keys. The congruent features are no more connected with one value than with the other on the task relevant dimension, so any congruence effects cannot be attributed to convergence at the decision or response level. There is still, however a possibility that irrelevant coding of both stimuli occurs in parallel with the relevant coding, for example in the language area, and that if there is any conflict between the two outcomes of this irrelevant coding, it could slow down the response to the relevant dimension. So if, for example, the high position and low tone are automatically labeled as such, the disagreement between these task-irrelevant verbal labels could conceivably slow down the response to the relevant orientation or instrument discriminations. This could apply to any condition in which verbal labels are automatically generated and are mismatched across modalities. Examples are pitch and position or pitch and contrast, which may share the labels "high" and "low". However, it is less plausible with pitch and size (where the natural labels are "small" and "large") and with spatial frequency, which most naive participants would label wide and narrow or thick and thin, not high and low. If we find any interference in these cases, it is likely to reflect a mismatch or conflict between the sensory representations.

Methods

Participants

After giving informed consent, eighty-five students (42 males) from Princeton University, aged 18 to 30 years, with normal or corrected-to-normal vision, participated in the experiments for course credit (8 in Experiments 1, 3, and 9; 12 in Experiment 2; 10 each in Experiments 4, 5, 6, and 7; and 9 in Experiment 8). Every aspect of this study was carried out in accordance with the regulations of Princeton University's Institutional Review Panel for Human Subjects.

Apparatus and stimuli

The stimuli were presented using MATLAB (Mathworks, Natick, MA) with Psychophysics Toolbox

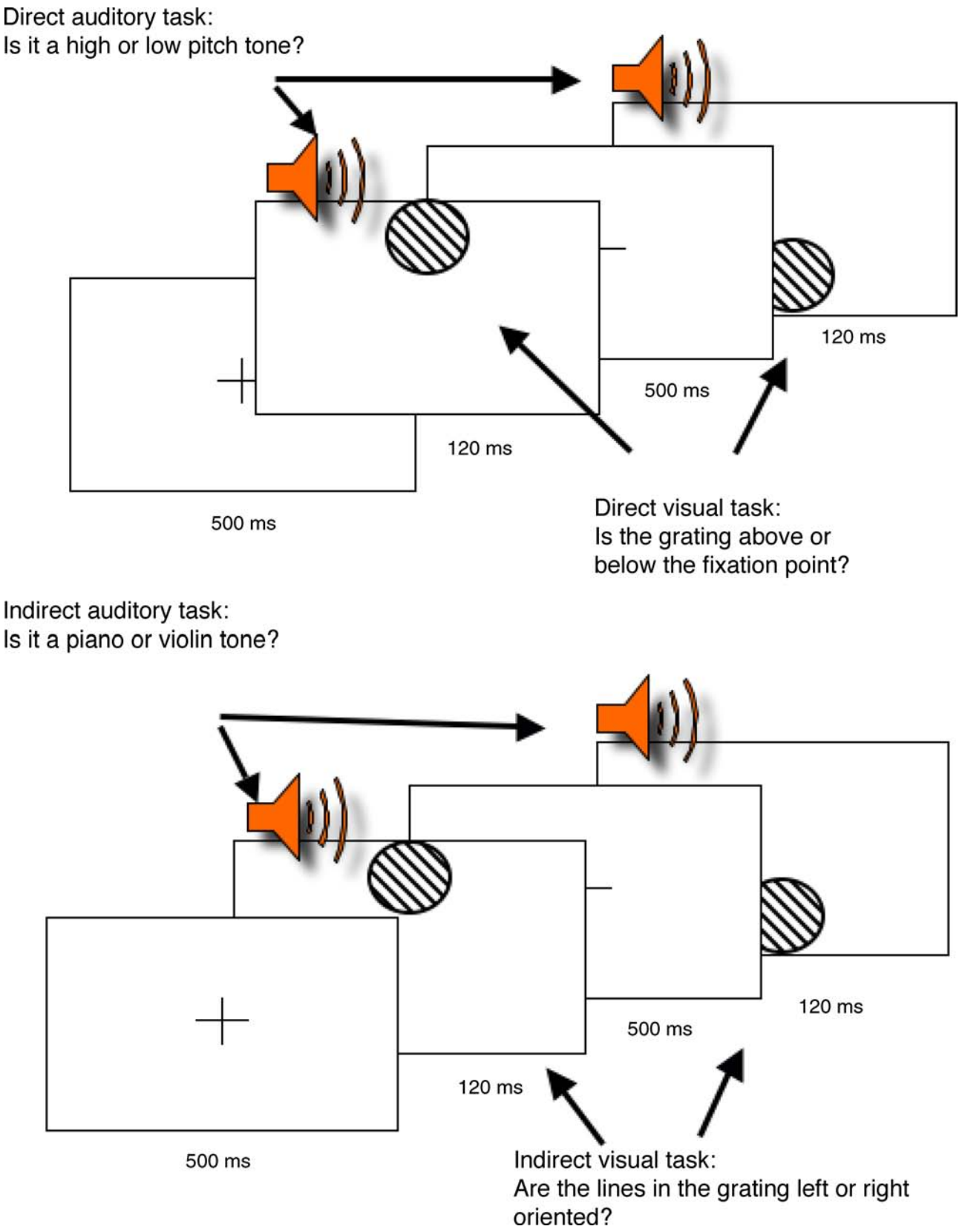

Figure 1. The timeline for (a) Experiment 1 and (b) Experiment 2. (a) and (b) represent the bimodal stimuli (both auditory and visual) and participant responses in Experiments 1 and 2. One-third of the trials in each experiment were presented as unimodal stimuli for comparison. Not drawn to scale.

extensions (Brainard, 1997; Pelli, 1997) running on a Macintosh G3. The visual displays were presented on a 17″ Apple screen at a viewing distance of 57 cm and monitor refresh rate of 75 Hz. The sound wave files were played through speakers positioned to the left and right of the computer screen and with center-to-center distance of 51 cm between the speakers.

The stimuli used in Experiments 1 to 8 were pairs with one auditory and one visual stimulus, played with the same onset and a shared duration of 120 ms. The auditory stimulus in the direct tasks (Experiments 1, 3, 5, and 7) was one of two pure tones of different pitches, 1500 Hz (high pitch) and 1000 Hz (low pitch) presented at an intensity of about 75 dB. The auditory stimuli in the indirect tasks (Experiments 2, 4, 6, and 8) were computer-generated tones simulating a violin or a piano, each with either high or low pitch (G2 or C2) presented at an intensity of 75 dB.

The visual stimulus in Experiments 1 to 8 consisted of a black and white sinusoidal grating (luminance of 8 for black and 320 cd/m^2 for white) presented on a gray background (80 cd/m^2). With the exceptions described below, the grating was positioned in the center of a gray screen. Its size was 3° visual angle, its frequency 4 cycles/deg, it was oriented to the left (45°), and it had high contrast. In Experiments 1 and 2, the grating varied in position: It was presented 4.5° either above or below a central fixation cross. In Experiments 3 and 4, it varied in size: It could be either small (subtending 3° visual angle) or large (subtending 6° visual angle). In Experiments 5 and 6, it varied in spatial frequency: It could be either 6 cycles/deg (high spatial frequency) or 2 cycles/deg (low spatial frequency). In Experiments 7 and 8, it varied in contrast. The bars of the grating could be either high contrast (0.95) or low contrast (0.56). In the Indirect tasks, in Experiments 2, 4, 6, and 8: on half the trials the grating had bars oriented at 45° and on half the trials at 135°. In Experiment 9, the two stimuli were in the same modality (vision). The grating was presented 4.5° above or below fixation, or at the center. It could be small or large (1.5° or 6°) or intermediate size (3°). It was high contrast and it was oriented to the left (45°).

Procedure

Participants performed the speeded classification task in 9 different experiments, separately judging the auditory or the visual stimuli (Experiments 1–8), or judging the visual stimuli on each of two dimensions (Experiment 9). Each experiment consisted of 3 conditions of which one was unimodal (i.e., stimuli were presented only in one modality, or in Experiment 9 only one dimension) and the other two were bimodal, one consisting of a congruent and the other of an incongruent combination of stimuli. In any block of the experiment, participants responded only to one modality. Stimuli in the other modality were irrelevant to the task.

Experiments 1, 3, 5, 7, and 9 employed the direct task (i.e., respond to the same features on which the correspondence was tested) and Experiments 2, 4, 6, and 8 employed the indirect task (i.e., respond to different features from those with the hypothesized cross-modal correspondence). In the direct tasks, participants discriminated either between the auditory stimuli (the two tone pitches) or the visual stimuli (two visual spatial positions, or two sizes, or two spatial frequencies, or two levels of contrast of the gratings). In the indirect tasks, participants discriminated either between the auditory stimuli (two different instruments—violin or piano) or the visual stimuli (the tilt of two differently oriented gratings, 45° or 135°). The auditory stimuli were still high or low in pitch and the visual gratings were high or low in visual position, large or small in size, high or low in spatial frequency, or high or low in contrast.

"Direct" tasks

In the auditory task, participants heard a pure tone (high or low pitch) played for 120 ms through the speakers. They responded by pressing the "s" key if they heard a high and the "k" key if they heard a low pitch tone. In the visual task, participants on each trial saw a fixation cross for 500 ms, followed by a grating with bars oriented 45 degrees to the left that appeared for 120 ms. In Experiment 1, it appeared either below or above the fixation cross, and participants made a choice response to the location, pressing "s" for below fixation and "k" for above (Figure 1a). By choosing the opposite response mapping in the two modalities, we hoped to avoid direct priming of the response keys in the congruent pairings.[1]

The procedure was the same for the other direct task experiments (3, 5, and 7) except that the visual task in Experiment 3 was to discriminate the size of the grating (large or small, see Figure 2a); in Experiment 5 it was to judge the width of the bars in the grating (wide or narrow, see Figure 2b), and in Experiment 7 it was to judge the contrast of the grating (high contrast or low, see Figure 2c). High visual position, small size, high spatial frequency, and high contrast gratings paired with high pitch were pairings that we predicted might be congruent, as were low visual position, large size, low spatial frequency, and low contrast gratings paired with low pitch. The converse pairings were considered incongruent.

"Indirect" tasks

Participants responded to different auditory and visual features from those expected to show a cross-modal correspondence. The auditory task was to indicate whether the tone was produced by a violin or a piano (pressing "s" for violin and "k" for piano). The visual task was to

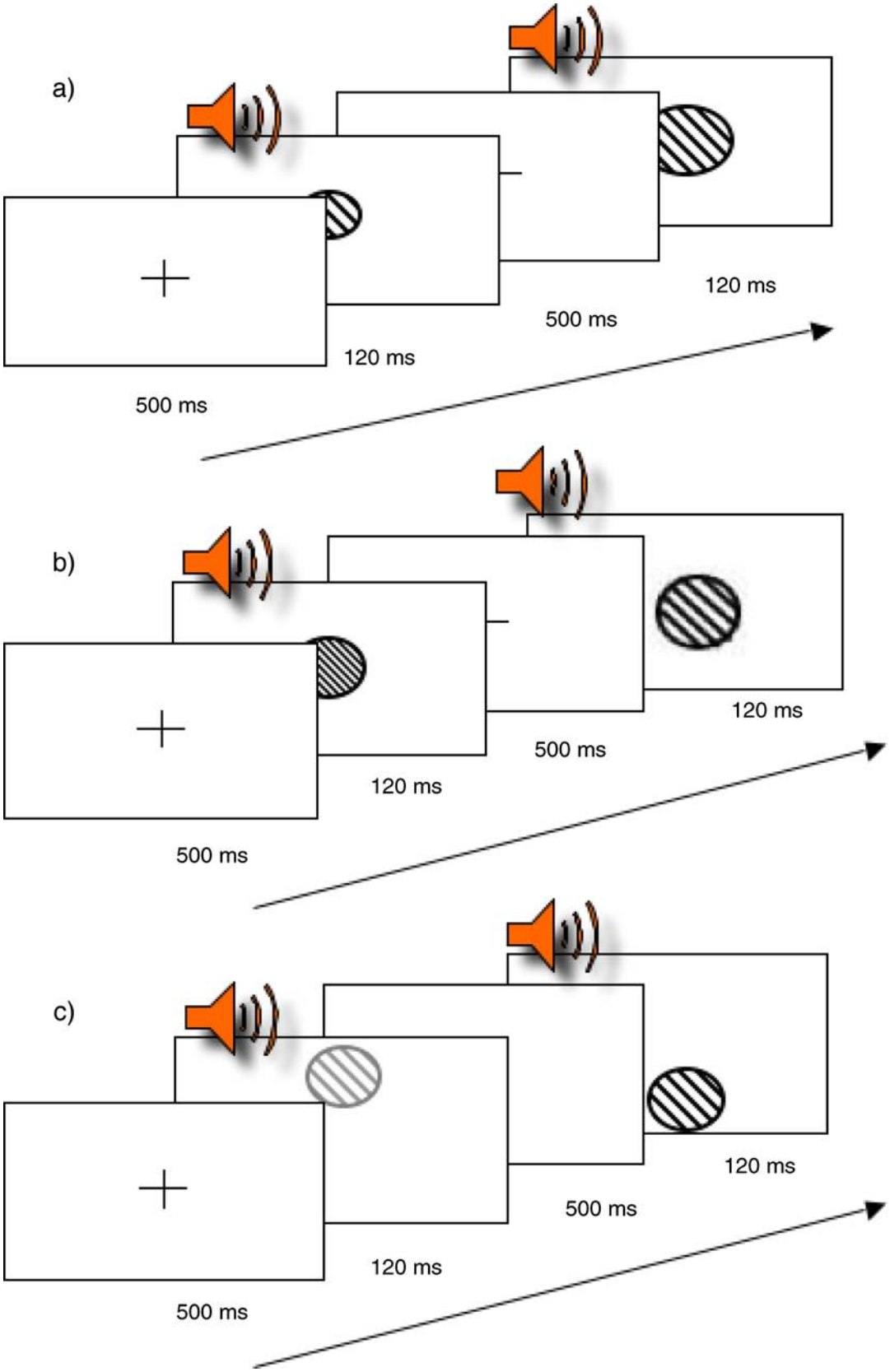

Figure 2. The timeline for (a) Experiments 3 (direct task) and 4 (indirect task) for pitch and size; (b) Experiments 5 (direct task) and 6 (indirect task) for pitch and spatial frequency; and (c) Experiments 7 (direct task) and 8 (indirect task) for pitch and contrast. Not drawn to scale.

indicate whether the grating they saw was left- or right-oriented (pressing "s" for left and "k" for right). The tones were still equally often high or low in pitch and the grating still appeared equally often above and below fixation (Figure 1b). The same pairings were considered congruent and incongruent as in the direct task.

In Experiment 9, we tested two dimensions within the visual modality. In the single-dimension position task, the grating was always 3° in size, and participants indicated whether it was presented above or below the fixation cross. In the single dimension size task, grating varied in both size and position. However, as with the bimodal conditions in previous experiments, participants only responded to a single dimension (size or position) and ignored the other dimension.

Results and discussion

The primary dependent measure in these studies was reaction time (RT), measured from stimulus onset, on correct trials only. We also recorded the accuracy of the responses. The mean RTs are shown for auditory pitch together with visual position, size, and spatial frequency in Figure 3. Note that the considerable baseline differences across experiments and for Direct versus Indirect tasks are for different groups of participants, whereas the differences between unimodal and bimodal congruent or incongruent, as well as the differences between visual and auditory tasks are within participants.

There was no evidence of a speed-accuracy trade-off: a significant congruency effect (faster RTs for bimodal congruent than incongruent) was always associated with either a significantly lower error rate for congruent pairs or no statistically significant difference. Significantly faster RTs for visual tasks were always associated either with significantly lower error rates for the visual task or no difference.

Figure 3 shows the mean response times and the accuracy in each of the six conditions in Experiments 1 to 6 separately for the Direct and Indirect tasks. Table 1 shows the results of four mixed model ANOVAs separately on the RTs with visual position (Experiments 1 and 2), size (Experiments 3 and 4), spatial frequency (Experiments 5 and 6), and contrast (Experiments 7 and 8). The within-participant factors were attended modality (visual or auditory) and condition (bimodal congruent, bimodal incongruent, and unimodal), and the between-participant factor was task (direct or indirect). The significant effects for each ANOVA are summarized in Table 1. The general pattern is similar across the three visual dimensions of position, size, and spatial frequency but not contrast, and there were also some informative differences between position, size, and spatial frequency.

Cross-modal congruence effects

Our main interest was in the effects of cross-modal congruence and how it depends on whether the task is Direct or Indirect. Table 2 shows the differences between the unimodal conditions and the congruent or incongruent conditions in Experiments 1 to 6, as well as the overall effect of congruence (Incongruent minus Congruent). We ran an ANOVA on the RT differences between Incongruent and Congruent RTs (i.e., the data in the 3rd and 6th columns of Table 2) with task and visual dimension as between-participant factors and modality and congruence as within-participant factors. The direct task (mean 19.7 ms) showed a larger overall congruence effect than the indirect task (mean 11.4 ms; $F(1, 52) = 13.6$, $p_{rep} = 0.99$, $\eta_p^2 = 0.21$). This is not too surprising, since two additional factors could contribute to the direct task effects: (1) the fact that the dimensions on which the mapping occurs are those that determine the responses and they therefore receive more attention; (2) the fact that, in the direct task, at least for the position dimension, there could be convergence on the shared verbal labels "high" and "low". These had to be remapped to opposite response keys for auditory and for visual stimuli, but they might nevertheless have had some effect at a level before the final response selection.

The important finding is that there is a significant effect of congruence in the indirect as well as the direct task on three visual dimensions (position, size, and spatial frequency, see Table 1). The cross-modal mapping between auditory and visual stimuli occurs automatically and affects performance even when it is completely irrelevant to the task.

Visual dimensions: Position, size, and spatial frequency

The differences in average congruence effects between the three visual dimensions were significant ($F(2, 52) = 13.2$, $p_{rep} = 0.99$, $\eta_p^2 = 0.34$) with a closer mapping between *spatial position* and pitch (in Experiments 1 and 2, mean 23.8 ms) than between either pitch and *size* (in Experiments 3 and 4, mean 11.3 ms) or pitch and *spatial frequency* (in Experiments 5 and 6, mean 11.7 ms). This could be due to a stronger contribution from the matching verbal labels, "high" and "low" for both position and pitch. The labels do not apply so naturally to size or to spatial frequency (where the obvious labels would be large/small and wide/narrow or thick/thin).

The interaction between visual dimension and direct versus indirect task was also significant ($F(2, 52) = 3.69$, $p_{rep} = 0.94$, $\eta_p^2 = 0.13$), reflecting the fact that the difference between direct and indirect tasks was considerably larger for spatial position (Experiments 1 and 2) than for size (Experiments 3 and 4), with the difference disappearing completely for spatial frequency (Experiments 5

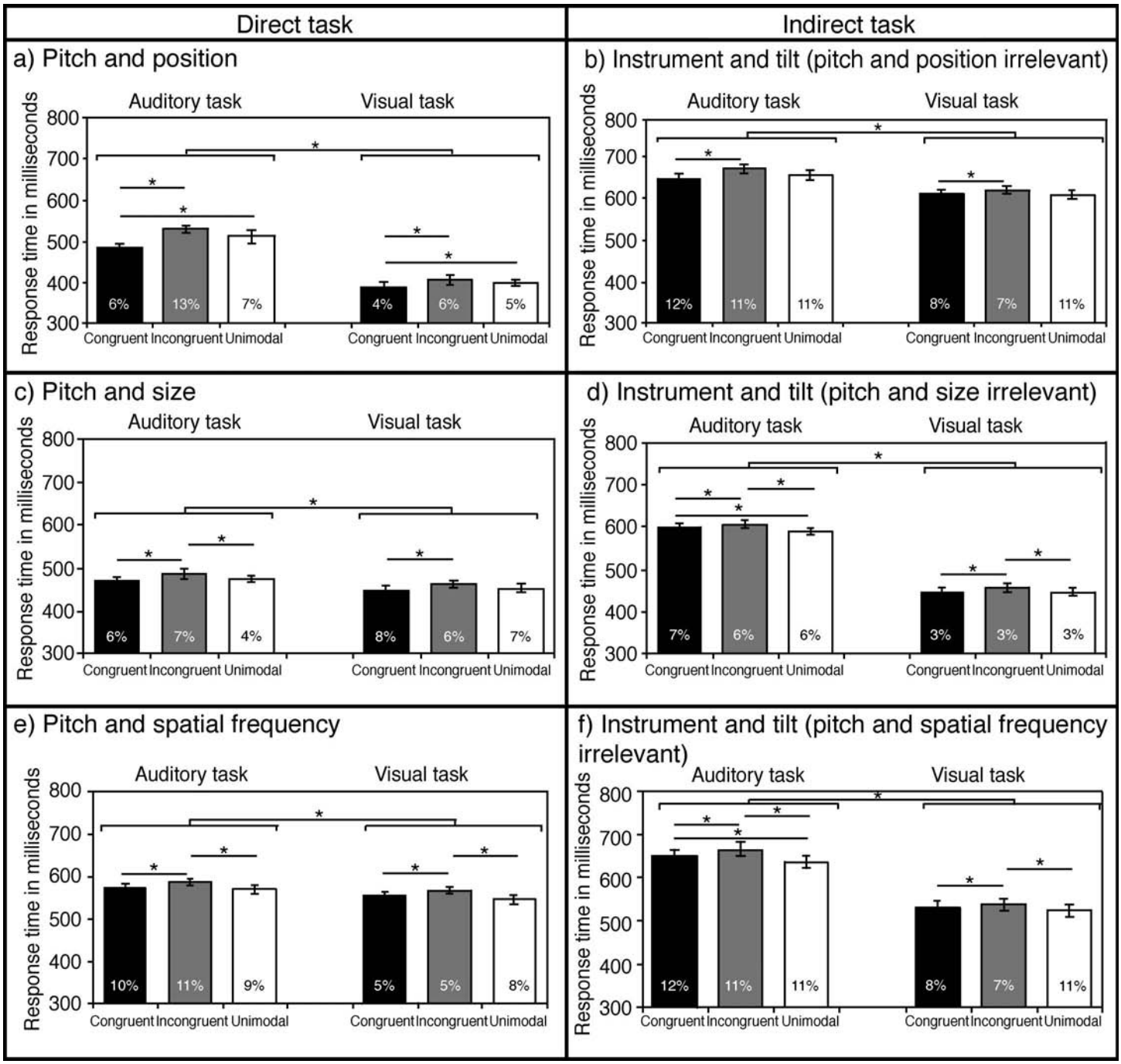

Figure 3. Mean response time in milliseconds with standard error of mean, averaged over participants in different conditions on different judgment tasks. (a, b) Pitch and position correspondence (Experiments 1 and 2); (c, d) pitch and size correspondence (Experiments 3 and 4); (e, f) pitch and spatial frequency correspondence (Experiments 5 and 6). The mean percentage of errors for each condition is shown at the foot of each column. Star indicates significant difference between the RTs.

and 6). Again the larger effect on the direct task for visual position could result from the verbal labeling high/low when those are the attended dimensions. It is striking that for spatial frequency the direct and the indirect tasks show the same effect of congruence whether participants are judging the spatial frequency of the gratings (the dimension on which the congruence with pitch is varied) or judging their orientation. The congruence relation seems to be registered automatically, as though the stimuli are taken in wholistically before being separately categorized.

Table 2 divides the congruence effects into facilitation (shown by the difference between unimodal and congruent conditions) and interference (shown by the difference between unimodal and incongruent conditions). In the ANOVAs shown in Table 1, the difference between unimodal and congruent conditions showed significant

Visual dimension	Experiments 1 and 2: Position	Experiments 3 and 4: Size	Experiments 5 and 6: Spatial frequency	Experiments 7 and 8: Contrast
Significant effects				
Modality (visual/auditory)	$F(1, 18) = 39.6$ $p_{rep} = 0.99$ $\eta_p^2 = 0.69$	$F(1, 16) = 61.5$ $p_{rep} = 0.99$ $\eta_p^2 = 0.79$	$F(1, 18) = 18$ $p_{rep} = 0.99$ $\eta_p^2 = 0.50$	$F(1, 17) = 48$ $p_{rep} = 0.99$ $\eta_p^2 = 0.74$
Task (direct/indirect)	$F(1, 18) = 33.8$ $p_{rep} = 0.99$ $\eta_p^2 = 0.65$	$F(1, 16) = 4.4$ $p_{rep} = 0.91$ $\eta_p^2 = 0.21$	$F < 1$	$F < 1$
Modality × task	$F(1, 18) = 6.4$ $p_{rep} = 0.94$ $\eta_p^2 = 0.26$	$F(1, 16) = 33.5$ $p_{rep} = 0.99$ $\eta_p^2 = 0.68$	$F(1, 18) = 8.5$ $p_{rep} = 0.97$ $\eta_p^2 = 0.32$	$F < 1$
Condition (congruent, incongruent, unimodal)	$F(2, 36) = 23.8$ $p_{rep} = 0.99$ $\eta_p^2 = 0.57$	$F(2, 32) = 18.2$ $p_{rep} = 0.99$ $\eta_p^2 = 0.53$	$F(2, 36) = 19.6$ $p_{rep} = 0.99$ $\eta_p^2 = 0.52$	$F < 1$
Congruent vs. incongruent	$F(1, 18) = 167.6$ $p_{rep} = 0.99$ $\eta_p^2 = 0.90$	$F(1, 16) = 32.1$ $p_{rep} = 0.99$ $\eta_p^2 = 0.67$	$F(1, 18) = 33.7$ $p_{rep} = 0.99$ $\eta_p^2 = 0.65$	$F < 1$
Congruent vs. unimodal	$F(1, 18) = 6.24$ $p_{rep} = 0.95$ $\eta_p^2 = 0.26$	$F < 1$	$F < 1$	$F < 1$
Incongruent vs. unimodal	$F(1, 18) = 7.4$ $p_{rep} = 0.96$ $\eta_p^2 = 0.29$	$F(1, 16) = 21.0$ $p_{rep} = 0.99$ $\eta_p^2 = 0.57$	$F(1, 18) = 32.8$ $p_{rep} = 0.99$ $\eta_p^2 = 0.65$	$F < 1$
Condition × task	$F(2, 36) = 3.1$ $p_{rep} = 0.92$ $\eta_p^2 = 0.15$	$F < 1$	$F < 1$	$F < 1$
Condition × modality	$F(2, 36) = 7.3$ $p_{rep} = 0.99$ $\eta_p^2 = 0.29$	$F < 1$	$F < 1$	$F < 1$

Table 1. Summary of ANOVAs on response times (from Figure 3) in Experiments 1 to 8.

	Direct task			Indirect task		
	Congruent minus unimodal	Incongruent minus unimodal	Incongruent minus congruent	Congruent minus unimodal	Incongruent minus unimodal	Incongruent minus congruent
Pitch–position						
Auditory	−27.0	18.1	45.1	−7.5	15.2	22.7
Visual	−13.1	5.5	18.6	3.7	12.4	8.7
Mean	−20.0	11.8	31.8	−1.9	13.8	15.7
Pitch–size						
Auditory	−2.1	13.5	15.6	10.3	17.6	7.3
Visual	−4.0	10.4	14.4	2.1	9.9	7.9
Mean	−3.0	12.0	15.0	6.2	13.8	7.6
Pitch–spatial frequency						
Auditory	3.7	18.3	14.6	15.9	30.7	14.8
Visual	9.4	19.7	10.3	6.6	13.7	7.1
Mean	6.6	19.0	12.4	11.2	22.2	11.0

Table 2. Effects of particular pairings in speeded classification: Differences in milliseconds between bimodal congruent, bimodal incongruent, and unimodal stimuli. Facilitation is shown by negative numbers and interference by positive numbers.

facilitation only for position and pitch, whereas the difference between unimodal and incongruent showed significant interference for all three visual dimensions. There may be some baseline interference simply from having a simultaneous stimulus in another modality, which is overridden by facilitation from congruence only when the latter is very strong. Earlier experiments on Garner interference have typically used as baseline a neutral stimulus on the irrelevant dimension rather than the absence of any irrelevant stimulus. We wanted to avoid any interaction between the modalities in the baseline condition. However, in relating our results to those of others, it may be more meaningful simply to compare congruent and incongruent pairings. To the extent that these differ we can conclude that there is a natural correspondence between polarities on the two modalities.

The two remaining conditions—pitch with visual contrast and visual position with visual size—were analyzed separately since they showed a different pattern of results.

Correspondence between pitch and contrast

Interestingly, there was no evidence for a congruency mapping between pitch and contrast, despite the fact that the verbal labels might apply better to contrast than to size or spatial frequency.

Within-modality correspondence between position and size (Experiment 9)

Using the "direct" task only, we probed the interaction between the two visual dimensions of position and size and found no congruence effect. Thus there was no interaction between position (congruent 366 ms, *SEM* 6 ms; incongruent 367 ms, *SEM* 7 ms) and size (congruent 492 ms, *SEM* 11 ms; incongruent 484 ms, *SEM*). The only significant effect was that of visual dimension (p_{rep} = 0.99, η_p^2 = 0.92) showing that participants were faster (367 ms, *SEM* 7 ms) and more accurate (3% error rate, *SEM* 1%) in classifying the stimuli by their position than by their size (487 ms, *SEM* 8 ms; 8% error rate, *SEM* 2%).

Auditory versus visual task

RTs were faster in the visual than in the auditory modality. This was not our main interest and we made no attempt to equate discriminability across the auditory and visual features. However, the faster responses to the visual stimuli here may have been due in part to the more natural stimulus-response mapping that we chose in this experiment for vision relative to audition. Participants have been shown to respond faster with the key on the left to lower locations and with the key on the right to higher locations, an effect sometimes called the "orthogonal Simon effect" (Lu & Proctor, 1995). We used the opposite mapping for audition in order to avoid cross-modal priming at the response level. Response compatibility may also have speeded responses to the visual stimuli in the indirect task, where we used the left key for left tilt of the Gabors and the right key for right tilt. There was no obvious compatible mapping between the auditory instruments, violin and piano and the left or right keys. Although these different stimulus-response compatibility effects were not our main interest, we followed up on their possible role in an experiment described in Appendix A.

Overall the auditory judgments (mean 20.0 ms) were also more strongly affected by congruence than the visual ones (mean 11.2 ms; $F(1,52) = 22.88$, $p_{rep} = 0.99$, $\eta_p^2 = 0.31$). The results suggest some bias in selective attention toward the visual modality, making it harder to ignore when it is irrelevant. The visual classifications were always made faster than the auditory ones, which may have made them less vulnerable to interference. If the auditory stimuli are discriminated more slowly, they may become available too late to affect the visual classifications. There was no modality by task interaction in the congruence effects, meaning that the difference between direct and indirect tasks was the same for both modalities.

General discussion

Our experiments have demonstrated a familiar cross-modal correspondence between auditory pitch and visual position and further explored a bidirectional correspondence between pitch and size that had been previously reported mainly in children. The findings replicate within our paradigm those described in the Introduction section (Ben-Artzi & Marks, 1995; Bernstein & Edelstein, 1971; Melara & O'Brien, 1987; Patching & Quinlan, 2002 for pitch with visual position, and Gallace & Spence, 2006; Marks et al., 1987 for pitch and size). They also provide the first experimental evidence of a cross-modal correspondence between pitch and spatial frequency and show no evidence of a correspondence between pitch and contrast. The absence of any effect with contrast is perhaps surprising given that Marks (1974, 1989) found interactions between pitch and brightness. Our manipulation of contrast left the average intensity constant whereas average brightness may be what was relevant in the experiments by Marks. There were some differences in the strength of the interactions that we observed with different dimensions. Pitch with spatial position gave the largest effects. Most of the correspondences appeared to be asymmetrical, with generally stronger effects on pitch than on the visual dimensions. The asymmetries may have been due in part to differences in the speed of responses.

Faster responses would allow more time for the congruency or incongruency to affect the slower responses.

In our studies, the indirect, orthogonal task sometimes showed effects that were as large as those in the direct task, despite the fact that attention was directed to the relevant dimensions in the direct case. For both pitch with visual size and pitch with spatial frequency, the effects were as large in the indirect as in the direct task. The larger effect in the direct task with visual position may reflect the fact that the verbal labels are shared and may actively conflict when there is a mismatch. As we argued earlier, it is not clear where the congruence effects could arise in the case of spatial frequency and size, since the verbal codes do not correspond for the visual and the auditory stimuli. They are certainly automatic and independent of attention. It appears that the stimuli are represented wholistically, with mismatches between the "expected" correspondences on irrelevant dimensions detracting from discriminations on the attended dimensions. Because no direct convergence between the auditory and the visual tasks exists in the indirect condition, it seems most plausible that they reflect an intrinsic correspondence at the perceptual level.

We tested for corresponding polarity matches within a single modality, between the visual dimensions of size and position. The finding that the corresponding polarity matches are not present is important because it suggests that each shares something different with auditory pitch, as if there are two or more independent mappings rather than a single representation at a more abstract level, which is shared between pitch and all the visual dimensions that have cross-modal correspondences.

Other papers have reported effects of an irrelevant stimulus dimension on the speed or accuracy of response and have used them as evidence for automatic processing of the irrelevant dimension but not as a tool to locate the effect in the system. For example, Rusconi, Kwan, Giordano, Umilta, and Butterworth (2006) showed that low tones are responded to more quickly and more accurately when the response key is low or to the left of space and high tones when the key is high or to the right. In a version of the experiment that seems initially to parallel our indirect task, they showed that the same stimulus-response mapping benefit also occurs when the response is made to a separate and supposedly irrelevant dimension—the instrument that played the tone (wind or percussion). In this case, the mapping was orthogonal to the pitch; for example, a wind instrument could be mapped to the left key and a percussion instrument to the right key. However, when the pitch of the tone happened to be low, pressing the left key to a wind instrument was faster than when it happened to be high. A similar finding was reported earlier by Dehaene, Bossini, and Giraux (1993) for numbers, which also appear to be mapped onto response space, with small numbers giving faster responses to a key on the left and high numbers to a key on the right. Again this spatial mapping effect was present even when the task was to judge the parity of the numbers rather than their magnitude. Thus the spatial correspondence seems to be detected automatically, but it still reflected priming of the response location by the stimulus even when the response was made to a different dimension of the stimulus.

Our indirect tasks differ from these in that there was no association between the potentially corresponding stimuli and the responses: both responses were made equally often to both stimuli. The only association was between the two stimuli in a simultaneously presented pair—for example, a high tone and a high position. The experiments of Rusconi et al. and Dehaene et al. showed priming only of responses whose locations were associated with the preferred mapping of pitch or number size. The critical difference in our experiments is not that the priming dimension is irrelevant to the task, nor that it is detected automatically rather than intentionally, although both are the case, but the fact that the responses were orthogonal to the stimuli and therefore cannot be the site of the interference or facilitation. The IAT (Greenwald, McGhee, & Schwartz, 1998) offers another example of priming from an irrelevant dimension. For example in one version of the test, whatever response key is associated with positive words is also associated with white faces and whatever response key is associated with negative words is also associated with black faces. The association here is between a particular semantic or emotional valence and a particular response key. Clearly the priming here is not at the perceptual level, but it does depend on a consistent mapping of stimuli to responses.

Our method also improves upon other designs previously used to understand possible levels of audio-visual congruence interactions. For example, Gallace and Spence (2006) designed an experiment in which they eliminated the possibility that the effects of alternating irrelevant sound on disk size discrimination could be attributed to a bias in response selection by having participants report on whether two disks matched in size or not. However, in their experiment, even though participants were not labeling the size of the object as smaller or larger, they were still judging the size of the object (i.e., the feature whose potential congruence interaction was being studied) rather than an orthogonal dimension. This leaves open the possibility of verbal or semantic mediation, which was ruled out with our indirect tasks.

Why should these cross-modal interactions exist? Some of the apparently arbitrary mappings seen in synesthesia have been attributed to cross-activation between brain areas that happen to be close together (such as the color-grapheme synesthesia in color area V4 and the number-grapheme area in the fusiform gyrus; Ramachandran & Hubbard, 2001). They attribute the many and varied correspondences found in individual synesthetes to multisensory convergence zones in the region of the temporal-parietal-occipital (TPO) junction, and might well locate our congruency effects between auditory pitch and the

various visual dimensions of position, lightness, brightness, shape, size, and spatial frequency to the same areas. Our data suggest that the natural mappings we observe between auditory pitch and the visual features of position, size, and spatial frequency may be the result of multisensory areas modulating modality-specific sensory systems' responses to heighten the perceptual salience of congruent pairs. There is some evidence of this type of multisensory interactions being successfully used by visual-to-auditory sensory substitution devices (e.g., vOICe, see www.seeingwithsound.com) that use pitch of the soundscape representing a visual object to render visible to the blind the position of that object in the vertical plane.

Appendix A

We conducted an additional experiment to make sure that stimulus-response compatibility did not interact with our main effect of congruency and in some way foster the results we observe. We reran Experiment 1 that shows both the largest congruence effects and the most likely potential interactions with location compatibility, counterbalancing both the response allocations, and the compatibility across modalities. In total, twenty-four participants took part completing 960 trials each. Response allocations were as given follows.

Group 1: Visual compatible and auditory incompatible. Left key for visual low and auditory high; right key for visual high and auditory low.

Group 2: Visual incompatible and auditory incompatible. Left key for visual high and auditory low and right key for visual low and auditory high.

Group 3: Visual and auditory compatible. Left key for visual and auditory low and right key for visual and auditory high.

Group 4: Visual and auditory incompatible. Left key for visual and auditory high and right key for visual and auditory low.

We imported the RT data into a mixed model ANOVA with modality (auditory and visual) and condition (bimodal congruent, bimodal incongruent, and unimodal) as within-participant factors, and response allocation (Group 1, Group 2, Group 3, Group 4) as a between-participant factor. We still find the main effects of modality and congruency ($F(1, 3) = 134.44$, $p < 0.01$ and $F(1, 3) = 35.27$, $p < 0.01$, respectively) but no main effect of response allocation (compatibility) nor any interactions of compatibility with either modality or condition.

Acknowledgments

This research was supported by NIH Grant 2004 2RO1 MH 058383-04A1 Visual Coding and the Deployment of Attention, by Grant # 1000274 from the Israeli Binational Science Foundation, and by NIH Grant 1RO1MH062331 Spatial Representations and Attention. We are grateful to Jeremy Wolfe for his helpful suggestions.

Commercial relationships: none.
Corresponding author: Karla K. Evans.
Email: kevans@search.bwh.harvard.edu.
Address: Visual Attention Lab, 64 Sidney Street, Suite 170, Cambridge, MA 02139, USA.

Footnote

[1]In case there were any interactions between the visual and the auditory stimuli in Experiment 1 and the spatial locations of the response keys, such as the "orthogonal Simon" effect (Lu & Proctor, 1995), we ran a supplementary experiment (see Appendix A) to check how the response allocations affected the congruency effects that were our main topic of interest.

References

Ben-Artzi, E., & Marks, L. E. (1995). Visual-auditory interaction in speeded classification: Role of stimulus difference. *Perception & Psychophysics, 57,* 1151–1162. [PubMed]

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *Journal of Experimental Psychology, 87,* 241–247. [PubMed]

Braaten, R. (1993). *Synesthetic correspondence between visual location and auditory pitch in infants.* Paper presented at the Annual Meeting of the Psychonomic Society.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10,* 433–436. [PubMed]

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology General, 122,* 371–371.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics, 68,* 1191–1203. [PubMed]

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: L. Erlbaum Associates; distributed by Halsted Press, New York.

Garner, W. R. (1976). Interaction of stimulus dimensions in concept and choice processes. *Cognitive Psychology, 8,* 98–123.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality, 74,* 1464–1480. [PubMed]

Long, J. (1977). Contextual assimilation and its effect on the division of attention between nonverbal signals. *Quarterly Journal of Experimental Psychology, 29,* 397–414.

Lu, C. H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin and Review, 2,* 174–174.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *American Journal of Psychology, 87,* 173–188. [PubMed]

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology and Human Perception Performance, 13,* 384–394. [PubMed]

Marks, L. E. (1989). On cross-modal similarity: The perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology and Human Perception Performance, 15,* 586–602. [PubMed]

Marks, L. E., Hammeal, R. J., & Bornstein, M. H. (1987). Perceiving similarity and comprehending metaphor. *Monogram Society Research Children Development, 52,* 1–102. [PubMed]

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception, 28,* 903–923. [PubMed]

Melara, R. D. (1989). Dimensional interaction between color and pitch. *Journal of Experimental Psychology and Human Perception Performance, 15,* 69–79. [PubMed]

Melara, R. D., & Marks, L. E. (1990). Processes underlying dimensional interactions: Correspondences between linguistic and nonlinguistic dimensions. *Memory Cognition, 18,* 477–495. [PubMed]

Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General, 116,* 323–336.

Mondloch, C. J., & Maurer, D. (2004). Do small white balls squeak? Pitch-object correspondences in young children. *Cognitive Affect Behavioral Neuroscience, 4,* 133–136. [PubMed]

Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology and Human Perception Perform, 28,* 755–775. [PubMed]

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10,* 437–442. [PubMed]

Ramachandran, V. S., & Hubbard, E. M. (2001). Psychophysical investigations into the neural basis of synaesthesia. *Proceedings of Biological Science, 268,* 979–983. [PubMed]

Rusconi, E., Kwan, B., Giordano, B. L., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect. *Cognition, 99,* 113–129. [PubMed]

Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognition Psychology, 24,* 99–142. [PubMed]

Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). "Metaphorical" mapping in human infants. *Child Development, 52,* 728–731.

Welch, R. B., & Warren, D. H. (1986). Intersensory interactions. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and performance* (vol. 1, pp. 251–256). New York: John Wiley and Sons.