# Combining vision with audition and touch, in adults and in children

David Burr*, Paola Binda[§^] & Monica Gori[^]

*Dipartimento di Psicologia, Università degli studi di Firenze

[§]Department of Psychology, Università Vita-Salute San Raffaele, Milano, Italy.

^Italian Institute of Technology – IIT Network, Research Unit of Molecular Neuroscience, Genova, Italy.

As the introductory chapter of this book clearly points out, efficient perceptual performance often requires integration of multiple sources of information, both within the senses and between them. Much experimental evidence suggests that the nervous system can and does combine information across senses, and does so in an optimal manner (see Chapter 1). In this chapter we discuss some studies from our own laboratory, showing how our perceptual systems combine visual with auditory and with tactile information in situations of conflict, both when introduced in the laboratory and when this occurs in natural conditions. We also explore how the capacity to integrate information develops in young children.

### *The ventriloquist effect*

Ventriloquism is the ancient art of making one's voice appear to come from elsewhere, exploited by the Greek and Roman oracles, and possibly earlier. We regularly experience the effect when watching television and movies, where the voices seem to emanate from the actors' lips rather than from the actual sound source. Originally ventriloquism was explained by performers *projecting* sound to their puppets by special techniques, and many still believe this (Connor, 2000). Of course modern physics tells us that it is impossible to "project sound", it has no option but to emanate from its actual source. More recently many suggested that ventriloquism was a perceptual phenomenon, with vision *capturing* sound, because of its inherent superiority (Caclin, Soto-Faraco, Kingstone, & Spence, 2002; Mateeff, Hohnsbein, & Noack, 1985; Pick, Warren, & Hay, 1969; Warren, Welch, & McCarthy, 1981), and this has become known as the "ventriloquist effect". Ventriloquists go to great lengths not to move their own lips, and to move their puppets' lips in rough synchrony with the voice: it is assumed that the visual movement of the lips "captures" the sound, so it appears to arise from the wrong source.

But why should vision "capture" sound? Why should vision set the "gold standard" when sight and sound are in discord? The answer is suggested by Eq. 1 of Chapter 1: the optimal way to combine information is to perform a weighted average, with the weights proportional to the *reliability* of the signal, where reliability is the inverse of the variance of the underlying noise distribution. So if auditory spatial localization were far worse than visual spatial localization, we would expect vision to dominate.

Alais and Burr (2004b) tested this idea empirically, under normal conditions and under conditions where visual localization was severely impaired. In order to estimate the theoretical weights that should be applied, they first measured visual and auditory localization performance and

various conditions of image degradation (following Ernst & Banks, 2002). They then measured localization for audio-visual stimuli in "conflict". The results were clear. When visual localization was good, with relatively sharply defined targets, vision does indeed dominate and "capture" sound. However, for severely blurred visual stimuli (that are poorly localized), the reverse holds: sound captures vision. For less blurred stimuli, neither sense dominates and perception follows the mean position. Precision of bimodal localization is usually better than either the visual or the auditory unimodal presentation. Thus the results are not well explained by one sense "capturing" the other, but by a simple model of optimal combination of visual and auditory information.

*Figure 1 about here*

Fig. 1A shows sample results for one observer, who was asked to report which of two presentations seemed to be displaced to the right of the other. Both stimuli were audio-visual, comprising a brief click presented together with a visual blob. In one presentation, the *probe*, the click and blob were presented to the same position that varied horizontally over the screen. In the other presentation (randomly first or second), the *conflict stimulus*, the visual stimulus was displaced $\Delta^\circ$ rightwards and the sound $\Delta^\circ$ leftwards (for the example of Fig. 1 $\Delta=5^\circ$). The abscissa shows the distance of the probe from the centre of the conflict stimulus (average of visual and auditory position). When the visual stimuli were small, hence well localizable in space, vision dominated: the psychometric functions are centered around the position of the visual standard, $+5^\circ$ (blue curve). However, when the blobs were heavily blurred, the centre of the psychometric function (the "point of subject equality" or PSE) was much closer to the auditory standard, at $-3^\circ$ (black curve). And for the intermediate blur, the PSE was close to 0, half way between the auditory and visual standards (red curve).

The maximum likelihood model outlined in Chapter 1 allows us to test quantitatively whether the results follow *ideal integration* predictions. On separate experiments we measured the ability of subjects to localize auditory and visual stimuli presented on their own. The standard deviations ($\sigma$) of those psychometric functions gave estimates of the expected *reliability* ($r$) of the three visual and one auditory stimuli ($r=\sigma^{-2}$), from which we could calculate the relative weights for the audio-visual stimuli (from Eq. 2 of Chapter 1). Eq. 1 of Chapter 1 then gives the predicted PSEs for the three levels of visual blur, for the various conflicts measured. These are shown on the abscissa of Fig. 1B, with the measured PSEs shown on the ordinate. Note that for all three observers

(different figure shapes) and blurs levels (different colors), the measured values followed closely the predictions (variance explained: $R^2$=0.87).

While these results support strongly the ideal-integrator model, they do not provide conclusive proof of integration: other strategies, like multiplexing could produce these types of results (although it would be unlikely to produce them so precisely). The strong proof of integration, perhaps the *signature* of integration, is the improvement of performance for observed for multimodal discriminations, because the reliabilities ($\sigma^{-2}_i$) should sum (Eq. 3 of Chapter 1). In practice the maximum likelihood model does not predict a large improvement in precision for only two modalities, at most $\sqrt{2}$ when the thresholds $\sigma_i$ are similar, far less when they differ (with the lower one dominating). Fig. 1C shows average thresholds for the clicks and 32˚ blobs (similar to each other), together with the measured and predicted audio-visual thresholds. Clearly, the bi-modal performance was much better than either uni-modal performance, and very close to the theoretical predictions.

Thus we conclude that the ventriloquist effect is a specific example of near-optimal combination of visual and auditory space cues, where each cue is weighted by an inverse estimate of noisiness, rather than one modality capturing the other. As visual localization is usually far superior to auditory location, vision normally dominates, apparently "capturing" the sound source and giving rise to the classic ventriloquist effect. However, if the visual estimate is corrupted sufficiently by blurring the visual target over a large region of space, the visual estimate can become worse than the auditory one, and optimal localization correctly predicts that sound will effectively capture sight. This is broadly consistent with other reports of integration of sensory information (Alais & Burr, 2004b; Battaglia, Jacobs, & Aslin, 2003; Clarke & Yuille, 1990; Ernst & Banks, 2002; Ghahramani, Wolpert, & Jordan, 1997; Jacobs, 1999). Note that in this study, for auditory localization to be superior to vision, the visual targets needed to be blurred extensively, over about 60°, enough to blur most scenes beyond recognition. However, we should recall that the location of the audio stimulus was defined by only one cue (interaural timing difference) and was not time varying, so auditory localisation was only about 1/6$^{th}$ as accurate as normal hearing (Mills, 1958; Perrott & Saberi, 1990). If the effect were to generalize to natural hearing conditions, then 10° blurring would probably be sufficient. This is still a gross visual distortion, explaining why the reverse ventriloquist effect is not often noticed for spatial events. There are cases, however, when it does become relevant, not so much for blurred as for ambiguous stimuli, such as when a teacher tries to make out which child in a large class was speaking.

In the next section we examine the ventriloquist effect in more natural situations, where vision is degraded not artificially by blurring, but by natural visual processes that occur during saccades.

### *Audio-visual integration at the time of saccades*

Saccades are rapid eye movements made frequently (2-3 times a second) to reposition our gaze across the visual field. While this is a clearly important strategy for extending the resolution of the fovea over the entire visual scene, it also poses serious challenges to the stability and continuity of visual perception – every time we make a saccade, visual images shift rapidly on the retina, generating spurious motion and displacement signals.

The visual system is believed to face these challenges by predictively compensating the consequences of eye movements. Before a saccade, a copy of the oculomotor signal (a "corollary discharge") is sent to visual areas, where it updates spatial representations according to the future post-saccadic gaze position (Ross, Morrone, Goldberg, & Burr, 2001; Wurtz, 2008) and may help suppressing the perception of spurious motion signals (Burr, Holt, Johnstone, & Ross, 1982; Diamond, Ross, & Morrone, 2000). While this compensation is effective in most conditions, it can lead to perceptual distortions under some conditions, such as when stimuli are briefly flashed just before or during the saccade. These stimuli are grossly mislocalized, systematically displaced in the direction of the end-point of the saccade (saccadic target), by as much as half the amplitude of the saccade (up to 10° for a 20° saccade, Matin & Pearce, 1965; Morrone, Ross, & Burr, 1997). As saccadic eye movements are unlikely to affect other modalities (such as audition), that rely on sensors that remain stable during gaze shifts, we took advantage of the saccade-induced selective disruption of visual perception to probe the integration of visual and acoustic information under more natural viewing conditions.

#### *Spatial ventriloquism during saccades*

This section describes how spatial information from vision and audition seems to be combined in a weighted fashion, following the maximum likelihood model of ideal cue-integration. We measured accuracy and precision of localization for visual, auditory and audio-visual stimuli briefly presented at the time of saccades (Binda, Bruno, Burr, & Morrone, 2007). Subjects sat before a wide hemispheric screen (covering nearly the whole visual field, see inset in Fig. 2), where visual stimuli (2° blue blobs flashed for one monitor frame, i.e. 15 ms) were front-projected. Auditory stimuli

(10 ms bursts of white noise) were played through one of 10 speakers mounted behind the screen. On each trial two stimuli were presented sequentially, a probe, displayed always at the same location, and a test, whose position was variable. In separate sessions, the stimuli were either unimodal (visual or auditory) or bimodal (a flash and a sound displayed to the same location). Subjects reported whether the test was to the left or right of the probe (two-interval forced choice alignment task). At about the time of the test presentation, a 32° rightward saccade was elicited. For each condition and stimulus type we measured a psychometric function, plotting the proportion of "test seen right of probe" responses against the position of the test relative to the probe (where 0° means that the two are aligned, Fig. 2). The median of the curves (PSE) estimates the bias of perceived test location (PSE < 0° implies that the test is mislocalized rightwards, in the direction of the saccade). The standard deviation $\sigma$ of the cumulative Gaussian fit gives the threshold or precision of localization.

*Figure 2 about here*

The uppermost panels of Fig. 2 compare the localization of unimodal visual (panel A) or auditory stimuli (panel B) presented perisaccadically (less than 25 ms before the saccadic onset) and during steady fixation. While auditory localization remains virtually identical in the two conditions, visual localization is dramatically affected by saccades. Not only is there a systematic tendency to perceive the perisaccadic test as mislocalized in the saccade direction (PSE << 0°), but also the precision of localization is about 10 times lower than in fixation conditions. During fixation, vision is far more precise than audition, and both modalities are accurate. Perisaccadically, visual precision is lowered to be similar to that of auditory localization, and becomes biased. As audition remains accurate, the two modalities provide conflicting cues to the location of a (physically congruent) bimodal test stimulus. The localization of such a bimodal stimulus is shown in panel C of Fig. 2 (for comparison, also visual and auditory perisaccadic curves are re-plotted). The bimodal PSE is intermediate between the two unimodal PSEs, implying that the two modalities are given approximately the same weight. The bimodal curve is slightly steeper than both unimodal curves, meaning that localization is more precise for the bimodal stimulus than for either of its unimodal components. Both these effects are consistent with an optimal (MLE) combination strategy of visual and auditory cues.

Since the exact time of saccadic onset was variable across trials, we were able to measure localization at various delays from the saccade. Fig. 3 (top and middle panels) shows the time-

courses of visual, auditory and bimodal localization, plotting bias and threshold values against the test presentation time (values referring to the steady fixation condition are reported as baselines). Visual stimuli (blue) were systematically mislocalized when they were presented less than ~50 ms before the saccade or during the first ~30 ms of its execution, the largest errors occurring at saccadic onset. The change of localization precision follows similar dynamics; in the perisaccadic interval, visual localization thresholds become similar or higher than auditory thresholds. The time-courses of bias and threshold for auditory localization (cyan) are nearly flat, as saccades do not affect them. The shape of bimodal time-courses (red) is similar to that of visual time-courses, but the magnitude of variations is reduced. There is a bias for perisaccadic stimuli, peaking at about the time of saccadic onset; localization errors, however, are about half the size of those observed for pure visual stimuli. Like visual thresholds, bimodal thresholds increased perisaccadically relative to the steady fixation condition, but they always remain lower than either unimodal threshold. The bimodal time-courses of both localization bias and threshold are well predicted by taking the MLE of the visual and auditory time-courses (black lines). The lower panels of Fig. 3 plot the variations of visual and auditory integration weights (Eq. 2 of Chapter 1) across the timecourse. The visual weight is progressively reduced as the test presentation approaches the saccadic onset and, perisaccadically, with the auditory weight increasing commensurately.

In summary, during steady fixation under normal conditions the perceived location of bimodal audio-visual stimuli is captured by the visual component. Saccades interfere with this phenomenon, decreasing the extent to which visual cues determine bimodal localization. The distortions induced by saccades on visual space perception are therefore partially rescued when spatial information from another modality (e.g. audition) is provided. Importantly, the localization of perisaccadic bimodal stimuli presented at various delays from a saccade can be quantitatively predicted by assuming an optimal (MLE) integration of auditory and visual cues, with the accuracy and precision of visual signals changing dynamically. This has two major implications.

First, it suggests that perisaccadic visual signals are already distorted when they are integrated with other sensory cues. They must be biased and imprecise (as suggested by the pure vision) when combined with auditory information, or the bimodal localization could not be predicted from the visual and auditory time-courses. Saccade-induced distortions of visual space should therefore occur before integration with multisensory spatial cues. Visual signals are initially encoded in a retinotopic frame of reference, where represented positions shift each time the eyes move. For visual perception to remain stable across eye movements, retinotopic representations need to be converted into gaze-invariant (e.g. craniotopic) maps. Such a transformation can be performed by taking into account the position of the eyes (e.g. relative to the head). We propose

that, in case of rapid gaze shifts, eye position information fails in accuracy and precision, resulting in the observed systematic localization errors and in the decrease of localization precision (Binda, Bruno, Burr, & Morrone, 2007). Audition, on the other hand, encodes stimuli in craniotopic coordinates, so the spatial cues from the two modalities need to be converted into a common format before integrations. A convenient format is craniotopic, stable across eye movements. Inaccurate and imprecise eye position signals will ultimately lead to a distorted representation of those visual signals that constitute the input to the process of multisensory integration. Deneve et al. (2001) demonstrated that a class of neural networks is able both to optimally integrate multisensory signals and to convert each signal into a new reference frame. In principle, such a network is able to simulate our findings in both the unimodal and the bimodal conditions, assuming that the output of the network is required to be craniotopic in all cases, and that the eye position input is inaccurate and imprecise. A detailed model of how this could occur is presented in Binda et al. (2007).

A second important implication of multisensory integration being near-optimal during saccades is that visual signals must be re-weighted dynamically, following the variations of visual precision. At the time of saccades, the neural representations of visual space undergo a rapid and continuous transformation. The coordinates of visual receptive fields in many visual areas transiently change due to the eye position signal (Duhamel, Colby, & Goldberg, 1992; Sommer & Wurtz, 2006), being a plausible substrate for the mislocalization of flashed stimuli observed behaviorally (Ross, Morrone, Goldberg, & Burr, 2001). Meanwhile, represented locations become progressively scattered, either because of an enlargement of receptive fields (at the single cell level: Kubischik, 2002) or due to changes at the level of the population code (Krekelberg, Kubischik, Hoffmann, & Bremmer, 2003), and this may result in the observed decrease of localization precision. For visual signals to be optimally weighted in the integration with other sensory cues, the mean and the variance of stimuli position in these rapidly changing maps needs to be estimated instantaneously, i.e. without averaging information across time (see Ma, Beck, & Pouget, 2008 and chapter X for a physiologically plausible implementation)[1].

Thus, our results suggest that an instantaneous estimate of visual precision is the optimal strategy for the localization of briefly presented stimuli, whose representation is transiently disrupted by an internal signal (the eye position signal, which we suggest is responsible for both the dynamic coordinate shift and the transient scattering of represented locations; see Binda, Bruno,

[1] Suggesting that visual weights are updated dynamically, we do not mean that the process of sensory re-weighting does not take time. Our results imply that, for each trial, visual precision is re-estimated. They also imply that the estimate is based on an instantaneous picture of visual representations, as these continually change in the perisaccadic interval. However, the estimation process itself may not be instantaneous – in principle, it could take all the time separating the stimulus presentation from the subject's response.

Burr, & Morrone, 2007 for more details). An interesting open question is whether such a strategy would be optimal in contexts other than brief perisaccadic stimulation. When stimuli remain continuously visible, for example, the averaging of information across time could be a preferable solution, under the reasonable assumption that the variance associated with the position (or with the size, the form, etc.) of an object remains constant. This assumption may fail when the internal status of the perceptual system changes, as it does during saccades.

*Perceived timing of audio-visual stimuli ventriloquism during saccades.*

While vision may be envisaged as an inherently spatial modality, with stimuli represented in a spatial format from the first processing stages (the retina), audition is best suited to process temporal information, as temporal intervals are explicitly encoded as early as in the cochlea. Given these structural constraints, we may expect vision to be more precise spatially than audition (as we have shown), but audition should be temporally more precise than vision. Indeed there is good evidence that this is the case, as shown by the phenomena of "auditory driving" (Berger, Martelli, & Pelli, 2003; Gebhard & Mowbray, 1959; Shams, Kamitani, & Shimojo, 2000; Shipley, 1964) and "temporal ventriloquism" (Aschersleben & Bertelson, 2003; Bertelson & Aschersleben, 2003; Hartcher-O'Brien & Alais, 2007; Morein-Zamir, Soto-Faraco, & Kingstone, 2003).

Following similar logic to the previous section, we asked whether saccades may interfere with this phenomenon, affecting the relative importance of visual and auditory temporal information (Morrone, Binda, & Burr, 2008). We measured the relative integration weights of visual and auditory temporal cues using a multisensory time-bisection task (similar to that of Hartcher-O'Brien & Alais, 2007). We asked subjects to compare the timing of a bimodal-conflicting test stimulus to two temporal markers. The test stimulus was a green vertical bar flashed at the centre of the screen, together with a 10 ms noise burst, presented before of after the flash (but was perceived as synchronous). The two markers were identical to the test, except that the visual and auditory components were synchronous (see inset in Fig. 4). Observers reported whether the test stimulus seemed temporally closer to the first or the second marker (two-alternative forced choice bisection task). The asynchrony between the auditory and visual components of the test was manipulated in a similar way to the spatial manipulation described previously. The time of presentation of the flash was advanced by $\Delta$ ms and that of the tone delayed by $\Delta$ ms, with $\Delta$ equal to $\pm$25 ms or +5 ms (yielding sound-flash separations of $\pm$50 ms and +10 ms). For each $\Delta$ we measured a psychometric curve, plotting the proportion of "closer to the first marker" responses against the test presentation time (defined as the average presentation time of its auditory and visual

components) relative to the two markers. The median of the curves (PSE) estimates the perceived time of the test stimulus (PSE > 0 ms implies that the test was systematically perceived as delayed).

PSE values are reported in Fig. 4 (panel A) as a function of Δ. Data points are adequately fitted by a linear function with positive slope (slope of the linear fit across PSEs for all subjects: $0.64 \pm 0.1$, implying an auditory weight[2] of 0.82). This implies that the perceived timing of the flash changed depending on the asynchrony between its visual and auditory component, and was determined for the most part (~80%) by the timing of the sound. Thus, in steady fixation conditions, auditory temporal information is given a stronger weight than visual information and sounds partially capture the perceived timing of bimodal audio-visual stimuli – the temporal ventriloquism effect.

*Figure 4 about here*

We repeated the experiment in a condition where subjects executed a 20° rightward saccade within ±25 ms from the time of the test presentation (the go-signal was given by the disappearance of the fixation dot and the appearance of a similar dot serving as saccade target). In this condition PSEs were also directly proportional to Δ (see Fig. 4, panel B); the linear regression of the PSEs from all subjects estimated a constant of proportionality higher than in fixation conditions ($0.88 \pm 0.2$, which implies an auditory weight of 0.94) – the difference was consistently observed in all subjects but, within subjects, it was not statistically significant (bootstrap t-test at 10000 repetitions). Note that the intercepts of the linear fits for perisaccadic and fixation data are also different (the intercept for perisaccadic data is $22.7 \pm 3.4$ versus $0.64 \pm 0.1$ for fixation data), suggesting that perisaccadic stimuli were perceived as delayed, irrespectively of the asynchrony between their visual and auditory components. This finding can be explained by assuming an effect of saccades exclusively on visual temporal information (more specifically, by assuming that perisaccadic flashes are perceived as strongly delayed, as we recently found to be the case (see Binda, Burr, & Morrone, 2007)).

These results show that the temporal ventriloquism effect is also observed at the time of saccades, with a tendency to be stronger than in steady fixation conditions. This is the opposite of what we observed in the spatial domain, where the magnitude of the spatial ventriloquism effect is strongly reduced during saccades. In principle, optimal cue combination is able to explain this

[2] Auditory weights ($w_a$) can be derived from the slope of the regression line ($\rho$) following the equation:
$w_a = 0.5\rho + 0.5$

difference (though we cannot make a quantitative prediction, since we did not measure unimodal performance). During steady fixation, visual spatial cues are more precise and hence dominate cues (Alais & Burr, 2004b); saccades reduce the precision of visual spatial information and consequently reduce its relative weight (see previous section). In the temporal domain, visual cues are weighted less than auditory cues even during normal fixation, resulting in the temporal ventriloquism effect (Morein-Zamir, Soto-Faraco, & Kingstone, 2003). Given an optimal integration rule, if saccades have an effect on visual temporal information (most plausibly, reducing its precision), it would leave audition as dominant, and even increase the dominance, as we observed.

The above findings imply that perisaccadically the perceived timing of a flash can be shifted backward of forward by the presentation of an asynchronous sound. Will this shift affect the dynamics of perisaccadic flash mislocalization? Given the sharply defined dynamics of perisaccadic visual mislocalization (see Fig. 3, blue lines), only flashes presented near saccadic onset should be mislocalized; those presented 50 ms before or afterwards are not. Now if we accompany flashes with sounds, leading or trailing by 50 ms, they should shift the flashes forward or backward in perceived time by some 47 ms (given the auditory weight of 0.94), changing completely the pattern of their mislocalization.

To test this prediction, we asked subjects to report the location of a green vertical bar (flashed at the centre of the screen) relative to a previously learned ruler while they made a 20° rightward saccade. The flashed bar was preceded or followed by a brief noise burst, with the same offsets used for the stimulus in the previous experiment. Fig. 4C plots the perceived location of the stimulus as a function of the time of flash presentation relative to the saccadic onset (average of four observers) for the various conditions (color-coded). If the temporal mislocalization of the flashes preceded the saccadic mislocalization, the time-courses of the +50 and –50 ms conditions should be separated by 94 ms relative to each other. The data are clearly inconsistent with this prediction. All the time-courses are strikingly similar, all peaking at saccadic onset. Not only the temporal dynamics of mislocalization, but also its magnitude is comparable across conditions; concurrent sounds did not reduce the perisaccadic localization errors. This was probably because in this setup the sounds provided no reliable spatial cues, as they were diffused by a speaker placed above the monitor screen.

We conclude that, while a spatially informative auditory presentation reduces the magnitude of perisaccadic mislocalization (Binda, Bruno, Burr, & Morrone, 2007), a temporally informative sound does not alter its dynamics. This is consistent with the hypothesis that the perisaccadic mislocalization of visual signals takes place prior to the integration of visual spatial cues with

information from the other modalities: the perceived timing of visual stimuli is altered by sounds, but multisensory integration operates on representations that have already been spatially distorted, following the characteristic dynamic of perisaccadic phenomena.

*Conclusion*

In the three experiments in this section, we tested the integration of audio-visual spatial or temporal information during saccades, by taking advantage of the selective disruption that eye movements produce on localization of visual stimuli briefly flashed just before or during a saccade. The results show that the perisaccadic distortions of visual space are greatly reduced when auditory spatial information is provided. Thus, while during steady fixation spatial vision dominates our perception of space (leading to the ventriloquist effect), perisaccadically other sensory modalities become dominant. This can be predicted by assuming an optimal strategy of cue combination (Chapter 1) because, during saccades, visual localization is far less precise than during fixation. Visual signals are therefore dynamically re-weighted in the perisaccadic interval, following the rapid variations of localization precision. Our findings also suggest that saccades interfere with visual space representations prior to their combination with other sensory cues – biased and imprecise visual spatial signals are provided as input to the multisensory integration mechanism. A similar conclusion can be drawn by the results of the other two experiments, which investigated the combination of visual and auditory temporal cues. Both during fixation and at the time of saccades, audition dominates over vision in determining our perception of time (the temporal ventriloquism effect). Perisaccadic flashes are therefore perceived as delayed or anticipated due to an auditory presentation. In spite of this, the dynamics of their mislocalization remains unaltered, suggesting that visual stimuli have already undergone to perisaccadic mislocalization before they are integrated with auditory cues.

The cues provided from the various modalities are initially encoded in disparately different frames of reference. For them to be combined, they need to be remapped into a common frame of reference. We suggest that saccades interfere with this remapping process because, during rapid "saccadic" gaze shifts, sensory representations cannot be fed with an accurate and precise representation eye position.

***Cross-modal facilitation of visual and tactile motion: common neural mechanisms?***

As many chapters (including this one) of the book show, it is now clear that the brain integrates information across the different senses in what is often a statistically optimal manner. However, it is not always clear that the integration is *functionally* optimal. For example, there is good evidence for

integration of vision and auditory motion, but the integration can be as strong for opposite as for matched directions of motion (Alais & Burr, 2004a; Meyer & Wuerger, 2001). It is hard to see how integration of opposite motion directions could be of any functional advantage for an organism. Other work shows that under what may be termed more "natural" conditions of object motion (Wuerger, Hofbauer, & Meyer, 2003), biological motion (Brooks et al., 2007) or even tap-dancing (Arrighi, Marrini, & Burr, 2009), the integration may occur in a more functional manner.

Another important issue is at what stage of neural processing does the integration occur. Does it involve common neural processing of sensory signals from the different senses, or is the information combined at a higher level, at the level of perceptual decisions? The previous sections of this chapter suggest that visuo-audio combination occurs after the site where saccades interact with visual signals to affect perceived position. However, other evidence points to an early site of combination of information. For example, evidence from fMRI studies suggests that tactile motion may activate similar cortical areas to those activated by vision, including area MT (Hagen et al., 2002; Ricciardi et al., 2004), implying that visual and tactile motion share neural mechanisms. However, given the coarse resolution of fMRI, it is conceivable that different neurons within the common neural structure where activated by visual and tactile motion.

We pursued the issue by studying visual and tactile motion perception psychophysically, by measuring minimal speed increment motion thresholds over a range of base-speeds (Burr, Sandini, & Gori, 2009). The stimuli were physical wheels etched with a sinewave profile of 10 c/deg (Fig. 5A). Subjects, seated at 57 cm, observed the front wheel with their index finger resting on the second (concealed from view). We first measured visual and tactile speed increment as a function of base speed: subjects responded in two alternative forced choice which of two presentations seemed faster, yielding the discrimination thresholds reported by red symbols in Fig. 5 B-C. Both visual (Fig. 5B) and tactile (Fig. 5C) motion produced the characteristic "dipper function", where the thresholds initially decrease with base speed to a minimum at base speeds around 0.08 cm/sec, then proceeded to rise, roughly in proportion to base speed (Weber's law). The visual and tactile motion curves are very similar in form, both in absolute sensitivity and position of the dip, suggesting similar mechanisms may be at work.

*Figure 5 about here*

We next repeated the study in a cross-modal condition, where the base-speed (*pedestal*) was presented in one modality, and the increment to be detected in the other. For example, in both

intervals the tactile motion could be, say, 1 cm/s (hence non-informative), and only in the *test* interval was there the visual motion to be detected. Blue symbols of Fig. 5 B-C show cross-modal visual and tactile thresholds over the range of base-speeds. Like the single-modality data, these data show a clear "dip", again at around 0.08 cm/s. However, the form of the curves differed from the others in that there was no rising limb with Weber-like behavior.

To study the facilitation more closely, for five naïve observers we measured visual and tactile speed thresholds with and without 0.08 cm/s pedestals of the same or different modality. Averaged results (normalized to base-threshold) are shown in Fig. 5 D-E. For both vision and touch, pedestals of the same (red bars) or different (blue) modality both reduced thresholds considerably, by more than a factor of two. In both cases the average effect of the pedestal was as strong for the cross-modal as for the intra-modal condition. To examine whether this may be due to reducing *temporal uncertainty*, we substituted the cross-modal motion-pedestal for a sound of matched duration (defining precisely the temporal interval of motion): but the concurrent sounds had no effect on base thresholds (green bars). We also measured facilitation with cross-modal pedestals moving in the opposite direction to the tests, informing observes that this was the case: but again, even though this condition contained as much information as the same-direction condition (in that observers knew the direction was *always* opposite), this pedestal had no effect (cyan bars).

The results of this study strongly suggest that visual and tactile motion share common neural mechanisms. Over a wide range of speeds, the sensitivity curves are very similar, both showing a dipper-like facilitation at around 0.08 cm/s. Most interestingly, at the speed of the dipper, there was cross-facilitation between the two modalities, vision facilitating touch and *vice versa*. Discrimination functions in many domains follow a "dipper function", including contrast discrimination (Nachmias & Sansbury, 1974; Pelli, 1985), blur (Watt & Morgan, 1983), visual motion (Simpson & Finsten, 1995) and temporal duration (Burr, Silva, Cicchini, Banks, & Morrone, 2009). Explanations for the dipper function generally involve non-linearities in the function that transduces physical events into neural signals. The function is assumed to accelerate positively at low speeds (effectively a thresholding mechanism), so a non-informative pedestal aids performance by shifting the operating range to the steepest part of the function.

Other explanations involve spatiotemporal uncertainty (Pelli, 1985), essentially suggesting that the pedestal reduces the time window – and its commensurate noise – that needs monitoring. However, the lack of facilitation by sound beeps speaks against this explanation for these results. That the pedestals of opposed direction did not facilitate threshold discrimination, even when observers knew the direction was inverted and "tried to take this into account", speaks against

cognitive explanations: this condition contained as much information as the same-direction (in that the inversion was totally predictable), but this information could not be used to facilitate thresholds. It seems for the facilitation to work, the motion must be of in the same direction, and within narrow bounds of speed, pointing to neural combination. If we accept that the dipper function reflects a non-linear, threshold-like transduction, then it would seem that the neural combination occurs before this thresholding. This conclusion is far stronger than the suggestion that multi-sensory combination is statistically optimal. The implication is that multi-sensory combination is not just statistically optimal, but is functionally important, and occurs at a moderately low level of sensory processing, before thresholding is applied. That is to say, visual and tactile motion are not processed separately, and their outputs (effectively decisions) combined, but they are processed through the same neural mechanisms, so a pedestal in one modality affects directly the neural threshold in the other. This is consistent with the fRMI studies that tactile motion activates similar cortical areas to those activated by vision, including area MT (Hagen et al., 2002; Ricciardi et al., 2004).

This result may seem to be at odds with the evidence of the previous section, showing that visual and auditory signals seem to combined after the influence of saccades on the visual signal. But perhaps this shows that saccades affect vision at a very precocious level. Certainly there is good evidence that saccadic suppression occurs early (Burr, Morrone, & Ross, 1994; Thilo, Santoro, Walsh, & Blakemore, 2003), and much evidence for the influence of saccades on receptive field position at very low early neural sites, including V1 (Nakamura & Colby, 2002). It would seem fundamental that spatial representations are mapped in non-retinotopic coordinates before combineing with non-visual cues. Interestingly, there is also evidence that the auditory influence on visual timing occurs early, even in V1 (Shams, Kamitani, Thompson, & Shimojo, 2001), not readily reconcilable to the saccadic results (Fig. 4). Perhaps the notion of a strict hierarchy is not the most appropriate model: analysis may proceed to a large part in parallel. However, it is becoming clear that the different sensory modalities share a good deal of neural processing. Indeed, the whole idea of sensory "modules" could be questioned (Burr, 1999).

### *Development of multi-modal integration*

Mammalian sensory systems are not mature at birth, but become increasingly refined as the animal develops. In humans, some properties, like visual contrast sensitivity, acuity, and binocular vision reach near-adult levels within a year of life (Atkinson, 1984), as do some basic tactile tasks (Streri, 2003), while other attributes, like form (Kovács, 1999), motion perception (Ellemberg, 2003) and visual or haptic recognition of a 3D object (Rentschler, 2004), continue to develop until 8-14 years

of age. There is also a difference in the developmental rate of different sensory systems, with touch developing first, followed by vestibular, chemical, auditory (all beginning to function prior to birth) and finally vision (Gottlieb, 1971). Some anatomical aspects (such as myelinization of the optic nerve and visual cortex development) continue to mature through to school age. Cross sensory integration could also pose particular challenges for maturing sensory systems that have constant need of *recalibration*, to take into account growing limbs, eye-length, inter-ocular distances etc. All these reasons may present an obstacle to optimal integration between modalities. So a natural question to ask is when do children develop the capacity to integrate information between senses.

Some studies suggest that young children and even infants possess a variety of multi-sensory abilities (Lewkowicz, 2000 for review). For example, 3-month-old children can match faces with voices on the basis of their synchrony (Bahrick, 2001), and 4-month-old babies can match visual and auditory motion (Lewkowicz, 1992). One recent psychophysical study (Neil, 2006) has shown that integration of visual and auditory orienting responses develops late in humans, after the unimodal orienting responses are well established. These results are also in accord with physiological studies in cat and monkey that show that while in adult animals, many neurons in the deep layers of superior colliculus show strong, super-linear integration of multimodal information (Stein, 1993), in young animals integration-enhanced response develops later, after the unimodal visual and auditory properties are completely mature (Stein, 1973; Wallace & Stein, 2001).

However, very few studies to date have investigated cross-sensory integration of spatial attributes, nor applied the MLE approach of Chapter 1 to examine whether the integration is statistically optimal. In our experiment (Gori, Del Viva, Sandini, & Burr, 2008) we measured visual-haptic integration of two aspects of form perception in young (5-10 year-old) children: size and orientation discrimination. We found that before 8 years of age there is little integration for either task. However, the pattern of results for the two tasks was quite different: for the size discrimination the haptic sense dominated in young children, while for orientation, vision dominated.

The size discrimination task (top left icon of Fig. 6) was a low-technology, child-friendly adaptation of Ernst and Banks' (2002) technique, where visual and haptic information are placed in conflict with each other to investigate which dominates perception under various degrees of visual degradation. For the size task the stimuli were physical blocks of variable height (48 to 62 mm, in 1 mm increments), displayed in front of an occluding screen for visual judgments, behind the screen for haptic judgments or both in front and behind for the bimodal judgments.

All trials involved a two-alternative forced-choice task where the subject judged whether a *standard* block seemed taller or shorter than a *probe* of variable height. For the single-modality trials, one stimulus (randomly first or second) was the *standard*, always 55 mm high, the other the *probe*, of variable height. The proportion of trials where the probe was judged taller than the standard was computed for each probe height, yielding psychometric functions. The crucial condition was the dual-modality condition, where visual and haptic sizes of the standard were in conflict, with the visual block 55+Δ mm and the haptic block 55–Δ mm (Δ = 0 or ±3 mm). The probe comprised congruent visual and haptic stimuli of variable height (48 – 62 mm). Despite the visuo-haptic conflict of the standard, the blocks appeared as one single stimulus to all adults and children tested.

We validated this technique with adults, demonstrating that optimal cross-modal integration did occur also under these conditions. Visual stimuli were degraded by blurring with a translucent screen positioned at variable distances from the stimulus, producing results very similar to those obtained by Ernst and Banks (2002): perceived size of visual-haptic stimuli followed closely the maximum likelihood estimate (MLE) predictions for all levels of visual blur and, most importantly, the thresholds for dual-modality presentation were lower than either visual or haptic thresholds.

*Figure 6 about here*

We then proceeded to measure haptic, visual and bimodal visuo-haptic size discrimination in 5-10 year-old children (in conditions with no visual blur). Fig. 6 shows for four children sample psychometric functions for the dual-modality measurements, fitted with cumulative Gaussian functions whose median estimates the point of subjective equality (PSE) between probe and standard, and standard deviation (σ) the discrimination threshold. The pattern of results for the 10 year-old (Fig. 6A) was very much like those for the adult: negative values of Δ caused the curves to shift leftwards, positive values caused them to shift rightwards. That is to say the curves followed the visual standard, suggesting that visual information was dominating the match, as the MLE model suggests it should, as the visual thresholds were lower than the haptic thresholds. This is consistent with the MLE model (indicated by color-coded arrows below abscissae): the visual judgment was more precise, and should therefore dominate. For the 5 year-old (Fig. 6 B), however, the results were dramatically different: the psychometric functions for the dual-modality presentation shifted in the direction opposite to that of the 10 year-old, following the bias of the haptic stimulus. The predictions (color-coded arrows under abscissa) are similar for both the 10 and

5 year-olds, as for both children visual thresholds were much lower than and haptic thresholds, so the visual stimuli should dominate: but for the 5 year old the reverse holds, with the haptic standard dominating the match.

*Figure 7 about here*

Fig. 7 reports PSEs for all children of all ages for the three conflict conditions, plotted as a function of the MLE predictions from single-modality discrimination-thresholds. If the MLE prediction held, the data should fall along the black dotted equality line (like Fig. 1C for the ventriloquist effect). For adults this was so, both size and orientation. However, at 5 years of age the story was quite different. For the size discriminations, not only do the measured PSEs not follow the MLE predictions, they varied inversely with Δ (following the haptic standard), lining up almost orthogonal to the equality line. The data for the six-year-olds similarly do not follow the prediction, but there is a tendency for the data to be more scattered rather than ordered orthogonal to the prediction line. By eight years of age the data begin to follow the prediction, and by ten falls along it well, similar to the adult pattern of results.

*Figure 8 about here*

In order to ascertain whether the haptic dominance was a general phenomenon, or specific to size judgments, we repeated the series of experiments with another spatial task, orientation discrimination; a very basic visual task which could in principle be computed by neural hardware of primary visual cortex (Hubel, 1968). The procedure was similar to the size discrimination task, again using a simple, low-technology technique (top right icon of Fig. 6). Subjects were required to discriminate which bar of a dual presentation (standard and probe) was rotated more counterclockwise. Rather than blurring the bar, visual discrimination was made more difficult by using an oblique rather than vertical standard, the so-called *oblique effect*, that also occurs in children (Appelle, 1972) but does not seem to affect haptic judgments (Appelle, 1986).

As with the size discriminations, we first measured thresholds in each separate modality, then visuo-haptically, by varying degrees of conflict (Δ=0 or ±4˚), with visual standards rotated clockwise by +Δ˚ and the haptic standards by –Δ˚. Fig. 6 C&D show sample psychometric functions for the dual-modality measurements for a 5 and 8 year-old child. As with the size

judgments, the pattern of results for the 8 year-old was very much like those for the adult, with the functions of the three different conflicts (fig. 6C) falling very much together, as predicted from the single modality thresholds by the MLE model (arrows under abscissae). Again, however, the pattern of results for the 5 year-old was quite different (Fig. 6D). Although the MLE model predicts similar curves for the three conflict conditions, the psychometric functions followed very closely the visual standards (indicated by the arrows above the graphs), the exact opposite pattern to that observed for size discrimination.

Fig. 7B plots measured against predicted PSEs for all children of all ages for the three conflict conditions. At 5 years of age the predictions (of very little scatter, given the similar weights) bore no relation to the measured PSEs, nearly proportional to $+\Delta$, following the visual stimulus. The data for the six-year-olds begin to become more scattered, but still do not follow predictions. By eight years of age the data begin to follow the prediction reasonably well, approaching the adult results where measured PSEs follow well the MLE predictions.

Fig. 8A-B show how the thresholds vary with age, for the various conditions. For both tasks, visual and haptic thresholds decreased steadily with age up until 10 years (orientation more so that size). The light-blue symbols show the thresholds predicted from the MLE model (Eq. 3, Chapter 1). For the adults, the predicted improvement was close to the best single-modality threshold, and indeed the dual-modality thresholds were never worse than the best single-modality threshold. A quite different pattern was observed for the five year-old children. For the size discrimination, the dual-modality thresholds were as high as the haptic thresholds, not only much higher than the MLE predictions, but twice the best single-modality (visual) thresholds. This result shows not only that integration is not optimal, it is not even a close approximation, like "winner take all". Indeed it shows a "*loser* take all" strategy. This reinforces the PSE data in showing that these young children do not integrate cross-modally in a way that benefits perceptual discrimination.

Fig. 8C-D tells a similar story, plotting the development of theoretical and observed visual and haptic weights: violet symbols show the theoretical MLE-predicted weights, and the black symbols the actual weights that were applied for the judgments, calculated from the PSE vs conflict functions. For both size and orientation judgments, the theoretical haptic weights were fairly constant over age, 0.2 – 0.3 for size and 0.3 – 0.4 for orientation. However, the haptic weights necessary to predict the 5 year-old PSE size data are 0.6 – 0.8, far, far greater than the prediction, implying that these young children give far more weight to touch for size judgments than is optimal. Similarly, the haptic weights necessary to predict the orientation judgments are around 0, far less

than the prediction, suggesting that these children base orientation judgments almost entirely on visual information. In neither case does anything like optimal cue combination occur.

Another recent study reinforces our evidence that children under the age of eight can take advantage of information from only one sense at a time, with optimal multisensory integration developing in middle childhood. Nardini, Jones, Bedford and Braddick (2008) studied the use of multiple spatial cues for the short-range navigation in children and adults. They provided subjects with two cues to navigation, visual landmarks and self-motion. While adults were able to take advantage and integrate both cues, increasing the precision of their navigation when both sources of information was available, children of 4-5 or 7-8 years showed no improvement in precision in the bimodal condition. In the conflictual condition, adults weighted each spatial cue with its reliability (Eq. 2, Chapter 1), as we have seen in many other examples in this book. But young children failed completely to integrate cues, alternating between them from trial to trial. These results suggest that the development of the two individual spatial representations occurs before they are integrated within a common unique reference frame. Taken together with our study, it would seem to be a general conclusion that optimal multisensory integration of spatial information only occurs after 8 years of age.

Our experiments showed that before 8 years of age children do not integrate information between senses, but one will dominate. Which sense dominates depends on the situation: for size judgments touch dominates, for orientation vision: neither seems to act as the "gold standard", as suggested by Berkeley's (1709) famous assertion that "touch educates vision". If this is true, it is only for size discrimination. Why should touch dominate size discrimination and vision orientation discrimination? It would appear that in both cases the most *direct* sense dominates. For example, size is not given directly to the visual system, but available only after computations involving both the extent of retinal image and the distance of the object: whereas proprioceptive units are linked closely to the motor system, providing a more direct estimate of size. On the other hand, neurons in primary visual cortex of primates are directly selective to orientation (Hubel, 1968; Tootell, 1998), while this would not be an immediate calculation for the haptic system. Indeed there does exist evidence that children younger that nine years of age have difficulty with size constancy (Granrud, 2006; Zeigler & Leibowitz, 1957), consistent with the fact that they rely less on visual than tactile information for size.

We believe that the reason children do not show cross-modal integration may have something to do with recalibration during development. As children develop, their limbs grow, their eyes grow and move further apart, their height, and hence viewpoint changes, all requiring constant

recalibration. We suggest that one aid to calibration is to use one sense to calibrate another. This idea is along the lines of Berkeley's "touch teaching vision", but does not always occur in that direction. The system would appear to be more flexible, where the modality with the most direct sensory information calibrates that with the less direct information. In other words, it is not the most precise modality that dominates, but the most *robust* one, the one for which the information can be derived directly from sense information with minimum calculations. And if one sensory modality is calibrating another, then information from the two can not be linked in an optimal manner.

### *Conclusions*

This chapter has presented several examples of how information from different senses is integrated within the human brain, yielding a more robust perception. Spatial information from vision and audition is combined in a linearly weighted fashion, with the weights proportional to the reliability of the signals, whether the signal reliability is degraded artificially, or "naturally" during saccadic eye movements. The weights seem to be updated dynamically, on the fly, taking into account the momentary precision of the visual localization, that varies predictably with time from a saccade. The audiovisual combination appears to occur after the saccade has influenced perceived position. However, the facilitation study suggests that visual and tactile motion information interact relatively early in sensory processing, probably in area MT. These results suggest that visual and tactile motion stimulate common neural mechanisms, so the combination is not just statistically optimal, but of fundamental functional importance. Finally, the developmental study showed that the capacity to integration visual and haptic spatial information is not present at birth, but develops at about 8 years of age. Before then one sense dominates, touch for size discrimination and vision for orientation discrimination. We suggest that the late development of integration may be connected with a need to keep maturing sensory systems calibrated, using one system to calibrate the other.

*Figure captions*

**Note that most of these will be rendered black and white (depending on our color quotient)**

**Figure 1.** The "ventriloquist" effect (Alais & Burr, 2004a). **A.** Example psychometric functions for one observer (LM), for one visuo-auditory conflict (Δ=5˚), for three different levels of blur (4˚: blue squares), 32˚ (red circles) and 64˚ (black squares). For the 4˚ stimuli, the probe followed the visual component in the conflict pair, producing a PSE (median of the curve) near the visual standard (indicated by icon). For the 64˚ stimulus the reverse held, with the PSE nearer the auditory standard. For 32˚ the PSE was in between the two unimodal standards. **B.** Results of all three observers of Alais and Burr's (2004a) experiment, for 5 different levels of conflict Δ (LM square, DM circle, SD triangle): color-coding for visual blur as for A. The results plot measured PSE (medians of psychometric functions like those of Fig. 1A) against the predictions from the unimodal threshold measurements (Eq. 1-2 of Chapter 1). The points differ little from the dashed diagonal equality line, with $R^2 = 0.87$. **C.** Average normalized thresholds (geometric means after normalizing with MLE predicted thresholds) for 5 observers, for discrimination of visual (32˚ stimuli: red bars), auditory (green), or cross modal (blue). The cross modal data is virtually identical to the MLE predictions, shown in light blue.

**Figure 2.** Visual, auditory and bimodal localization measured during steady fixation and perisaccadically. Visual stimuli appeared on a hemispheric screen, with 10 speakers mounted behind (grey symbols in inset). Subjects compared the locations of two successive presentations (SOA = 600 ms), a probe (fixed position, blue blob in the inset) and a test (variable horizontal location), both presented unimodally (visually or acoustically) or bimodally (each being composed by a flash and a sound exactly matched in position). The proportion of "test rightwards" responses was plotted against the relative test-probe position (panels A-C report data from one typical subject). Panels **A** and **B** compare the localization of unimodal stimuli presented while subjects maintained steady fixation (hollow symbols) or just before they executed a saccade from FP to ST (filled symbols; only trials in which the delay between the test presentation and the saccadic onset was < 25 ms were included). The localization performance for perisaccadic visual and auditory stimuli is re-plotted in panel **C**, together with the performance for bimodal stimuli. Unlike with steady fixation, visual and auditory localization have about the same precision. Since visual localization is biased while auditory localization remains veridical, the two modalities provide conflicting information of the location of a physically congruent bimodal stimulus. The localization curve for such a stimulus (red) lies in between the visual and the auditory curves, and it is steeper than either. Both these effects are consistent with an optimal (MLE) cue combination strategy.

**Figure 3.** Unimodal and bimodal localization measured at various delays from the start of a saccade. For each stimulus type (visual, auditory and bimodal), a psychometric function was fitted on a subset of trials, selected by a temporal window of 50 ms width that was iteratively displaced by 5 ms (note that the same data were reported in Figure 6 of (note that the same data were reported in Figure 6 of Binda et al. (2007), but the iterative selection of trials was performed with a different algorithm). The resulting bias and threshold values are plotted against the average time of stimulus presentation in each subset of trials. The leftmost panels report averages across PSE and threshold values estimated in three subjects; the behavior of one individual subject is shown in the rightmost panels. Bimodal localization (red) is always more precise than either visual (blue) or auditory localization (cyan); and the perisaccadic bias is less than observed for visual-only presentations. Both the bias and the threshold time-courses for bimodal stimuli are adequately predicted by taking the optimal combination of the visual and the auditory time-courses (black) – the prediction for each time point is computed by taking the MLE of the unimodal localization performances observed at that time point. Visual and auditory integration weights (calculated from Eq. 3 of Chapter 1) are reported in the bottom panels. Note that as the probe was always in a fixed and readily identifiable position, we assume in our calculations that noise associated with its localization was negligible compared with that of the variable test.

**Figure 4.** Perceived timing and location for temporally conflicting audio-visual stimuli. **A. & B.** Subjects compared the timing of a bimodal conflicting test stimulus to two temporal markers spanning it; each of the three stimuli comprised a vertical bar flashed at the centre of the screen and a sound played by a speaker over the screen. As shown in the inset, the flash and the sound of the test were presented asynchronously (with the flash preceding the tone by $2\Delta$ ms), while those of the markers were synchronous. The two markers were separated by 800 ms, with the test straddling them with a variable delay. The perceived duration of the test was estimated from the mean of fitted psychometric functions plotting the proportion of trials in which the test was reported as closer to the first marker against the test-markers relative timing, for various degrees of asynchrony between its visual and auditory components ($\Delta = 10$ or $\pm 25$ ms). PSEs from fixation and saccadic conditions (where the test was presented within $\pm 25$ ms from the onset of a 20° rightwards saccade) are reported respectively in Panels A and B (error bars report s.e. as estimated with a bootstrap re-sampling at 1000 repetitions, Efron & Tibshirani, 1994). PSEs from the three tested subjects were adequately fit by a linear function with positive slope (black lines; grey lines report 95% confidence limits), implying that perceived timing for the bimodal test stimulus was mainly determined by the timing of auditory component. **C.** Perceived location of the test stimulus as a function of the time of flash presentation relative to saccadic onset (averaged across four subjects; error bars report

bootstrap estimates of s.e. of the between subjects mean). Again, we tested three different degrees of asynchrony (2Δ) between the auditory and visual stimuli (identical to the test in the previous experiment), together with an additional condition in which the flash was presented alone. Apparent positions (reported relative to a remembered ruler) are plotted against the time of flash presentation relative to the saccadic onset. The time-courses for the four tested conditions (see legend) are all similar, implying that the apparent test location depends exclusively on the physical time of flash presentation.

**Figure 5.** Illustration of the stimuli and main results for intra-modal and cross-modal visuo-haptic facilitation. **A.** The stimuli: two physical wheels under independent computer control, etched with a sinewave profile of specific spatial frequency (10 c/deg). The front wheel was observed through a small aperture. **B. & C.** Incremental speed thresholds for visual (B) and tactile (C) motion, as a function of base speed, for two observers. Red circular symbols show uni-modal thresholds, where all signals are confined to the same modality, vision (B) or touch (C). Observers were required to choose the interval containing the faster speed ($V+\Delta V$ cm/s) from base-speed ($V$ cm/s). The blue symbols show thresholds for pedestals of different modality to the test, tactile (B) or visual (C). The pedestal moved at $V$ cm/s in both trial intervals, the test increment at $\Delta V$ cm/s only during the test interval. Test speed varied from trial to trial, following the adaptive algorithm QUEST (Watson & Pelli, 1983) that homed in on threshold. Thresholds were calculated by fitting a raised cumulative Gaussian the data (150 trials per point), with error bars on individual data points obtained by bootstrap (Efron & Tibshirani, 1994). No feedback was given in any condition. **D. & E.** Mean normalized thresholds of 5 naïve observers for visual (D) and tactile (E) speed increment discrimination. Individual thresholds were divided by thresholds for the no-pedestal condition, then averaged (geometric mean) across subjects. The dashed line at unity, indicates no pedestal effect. Red bars indicate thresholds for pedestals of the same modality, blue pedestals of different modality. The green bars show thresholds when the interval was marked by an auditory tone of 2450 Hz, and the cyan bar thresholds opposite directions of motion (observers were informed of the inversion). Error bars represent ± 1 SEM.

**Figure 6**. Example psychometric functions for discrimination tasks for four children, with various degrees of cross-modal conflict. **A. & B.** Size discriminations: SB age 10.2 (A); DV age 5.5 (B); **C. & D.** Orientation discrimination: AR age 8.7 (C); GF age 5.7 (D). The lower colour-coded arrows show the MLE predictions, calculated from threshold measurements (Eq. 1-2, Chapter 1). The black dashed horizontal lines show the 50% performance point, intersecting with the curves at their PSE (shown by short vertical bars). The upper colour-coded arrows indicate the size of the haptic standard in the size condition (A-B) and the orientation of visual standard in the orientation

condition (C-D). The older children generally follow the adult pattern, while the 5 year-olds were dominated by haptic information for the size task, and visual information for the orientation task.

**Figure 7.** Summary data showing PSEs for all subjects for all conflict conditions, plotted against the predictions, for size (**A**) and orientation (**B**) discriminations. Different colors refer to different subjects within each age group. The symbol shapes refer to the level of cross-sensory conflict (Δ): squares 3 mm or 4˚; circles –3 mm or –4˚; upright triangles 0; diamonds 2 mm; inverted triangles –2 mm. Closed symbols refer to the no-blur condition for the size judgments, and vertical orientation judgments; open symbols to modest blur (screen at 19 cm) or oblique orientations; cross in symbols to heavy blur (screen at 39 cm). For both size and orientation discriminations, the predictions are far from the measured results for children younger than 8. Error bars on individual data points obtained by bootstrap (Efron & Tibshirani, 1994).

**Figure 8. A. & B.** Average thresholds (geometric average) for haptic (red symbols), visual (green) and visuo-haptic (dark blue) size and orientation discrimination, together with the average MLE predictions (light blue), as a function of age. The predictions were calculated individually for each subject, then averaged. The tick labeled "blur" shows thresholds for visual stimuli blurred by a translucent screen 19 cm from the blocks. **C. & D.** Haptic and visual weights for the size and orientation discrimination, derived from thresholds via the MLE model or from PSE values. Weights were calculated individually for each subject, then averaged. After 8-10 years the two estimates converge, suggesting that the system begins to integrate visual and haptic information in a statistically optimal manner. Error bars represent ± 1 SEM.

***References***


Alais, D., & Burr, C. D. (2004a). No direction-specific bimodal facilitation for audiovisual motion detection. *Brain Res Cogn Brain Res, 19*(2), 185--194.

Alais, D., & Burr, C. D. (2004b). The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol, 14*(3), 257-262.

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the "oblique effect" in man and animals. *Psychol Bull, 78*, 266-278.

Appelle, S. C., M. (1986). Eliminating the haptic oblique effect: influence of scanning incongruity and prior knowledge of the standards. . *Perception, 15*, 325-329.

Arrighi, R., Marrini, F., & Burr, D. C. (2009). Meaningful auditory information enhances perception of visual biological motion. . *J. Vis.* , In press.

Aschersleben, G., & Bertelson, P. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension. 2. Evidence from sensorimotor synchronization. *Int J Psychophysiol, 50*(1-2), 157-163.

Atkinson, J. (1984). Human visual development over the first 6 months of life. A review and a hypothesis. *Hum Neurobiol., 3*(2), 61-74.

Bahrick, L. E. (2001). Increasing specificity in perceptual development: infants' detection of nested levels of multimodal stimulation. *J Exp Child Psychol, 79*(3), 253-270.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis, 20*(7), 1391-1397.

Berger, T. D., Martelli, M., & Pelli, D. G. (2003). Flicker flutter: is an illusory event as good as the real thing? *J Vis, 3*(6), 406-412.

Berkeley, G. (1709). *An essay towards a new theory of vision*. Indianapolis: 1963: Bobbs-Merril.

Bertelson, P., & Aschersleben, G. (2003). Temporal ventriloquism: crossmodal interaction on the time dimension. 1. Evidence from auditory-visual temporal order judgment. *Int J Psychophysiol, 50*(1-2), 147-155.

Binda, P., Bruno, A., Burr, D. C., & Morrone, M. C. (2007). Fusion of visual and auditory stimuli during saccades: a Bayesian explanation for perisaccadic distortions. *J Neurosci, 27*(32), 8525-8532.

Binda, P., Burr, D. C., & Morrone, M. C. (2007). *Spatio-temporal distortions of visual perception during saccades.* Paper presented at the 30th European Conference in Visual Perception, Arezzo, Italy.

Brooks, A., van der Zwan, R., Billard, A., Petreska, B., Clarke, S., & Blanke, O. (2007). Auditory motion affects visual biological motion processing. *Neuropsychologia, 45*(3), 523-530.

Burr, D. (1999). Vision: modular analysis--or not? *Curr Biol, 9*(3), R90-92.

Burr, D. C., Holt, J., Johnstone, J. R., & Ross, J. (1982). Selective depression of motion sensitivity during saccades. *J Physiol, 333*, 1-15.

Burr, D. C., Morrone, M. C., & Ross, J. (1994). Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature, 371*(6497), 511-513.

Burr, D. C., Sandini, G., & Gori, M. (2009). Cross-modal facilitation of visual and haptic motion. *J. Vis.*, Abstract, in press.

Burr, D. C., Silva, O., Cicchini, M., Banks, B. S., & Morrone, M. C. (2009). Temporal mechanisms of multi-modal binding. *Proc R Soc Lond B Biol Sci*, In press.

Caclin, A., Soto-Faraco, S., Kingstone, A., & Spence, C. (2002). Tactile "capture" of audition. *Percept Psychophys, 64*(4), 616-630.

Clarke, J. J., & Yuille, A. L. (1990). *Data fusion for sensory information processing.* Boston: Kluwer Academic.

Connor, S. (2000). *Dumbstruck: A Cultural History of Ventriloquism*. Oxford: OUP.

Deneve, S., Latham, P. E., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nat Neurosci, 4*(8), 826-831.

Diamond, M. R., Ross, J., & Morrone, M. C. (2000). Extraretinal control of saccadic suppression. *J Neurosci, 20*(9), 3449-3455.

Duhamel, J. R., Colby, C. L., & Goldberg, M. E. (1992). The updating of the representation of visual space in parietal cortex by intended eye movements. *Science, 255*(5040), 90-92.

Efron, B., & Tibshirani, R. (1994). *An introduction to bootstrap* (Vol. 57): Chapman & Hall).

Ellemberg, D. e. a. (2003). Comparison of sensitivity to first- and second-order local motion in 5-year-olds and adults. *Spat Vis, 16*, 419-428.

Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Gebhard, J. W., & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *Am J Psychol, 72*, 521-529.

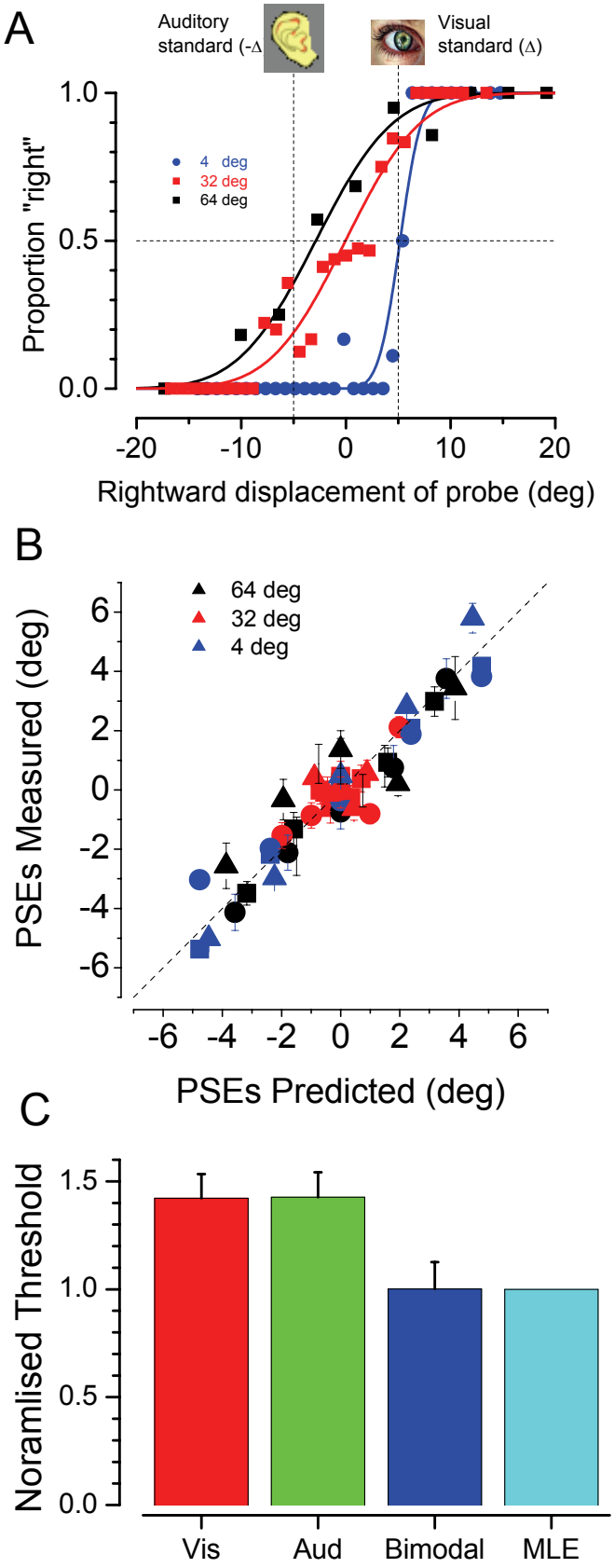
Ghahramani, Z., Wolpert, D. M., & Jordan, M. I. (1997). Computational models of sensorimotor integration. In P. G. Morasso & V. Sanguineti (Eds.), *Self-organization, computational maps and motor control.* (pp. 117-147). Amsterdam: Elsevier Science Publ.

Gori, M., Del Viva, M., Sandini, G., & Burr, D. C. (2008). Young children do not integrate visual and haptic form information. *Curr Biol, 18*(9), 694-698.

Gottlieb, G. (1971). *Development of species identification in birds: An inquiry into the prenatal*

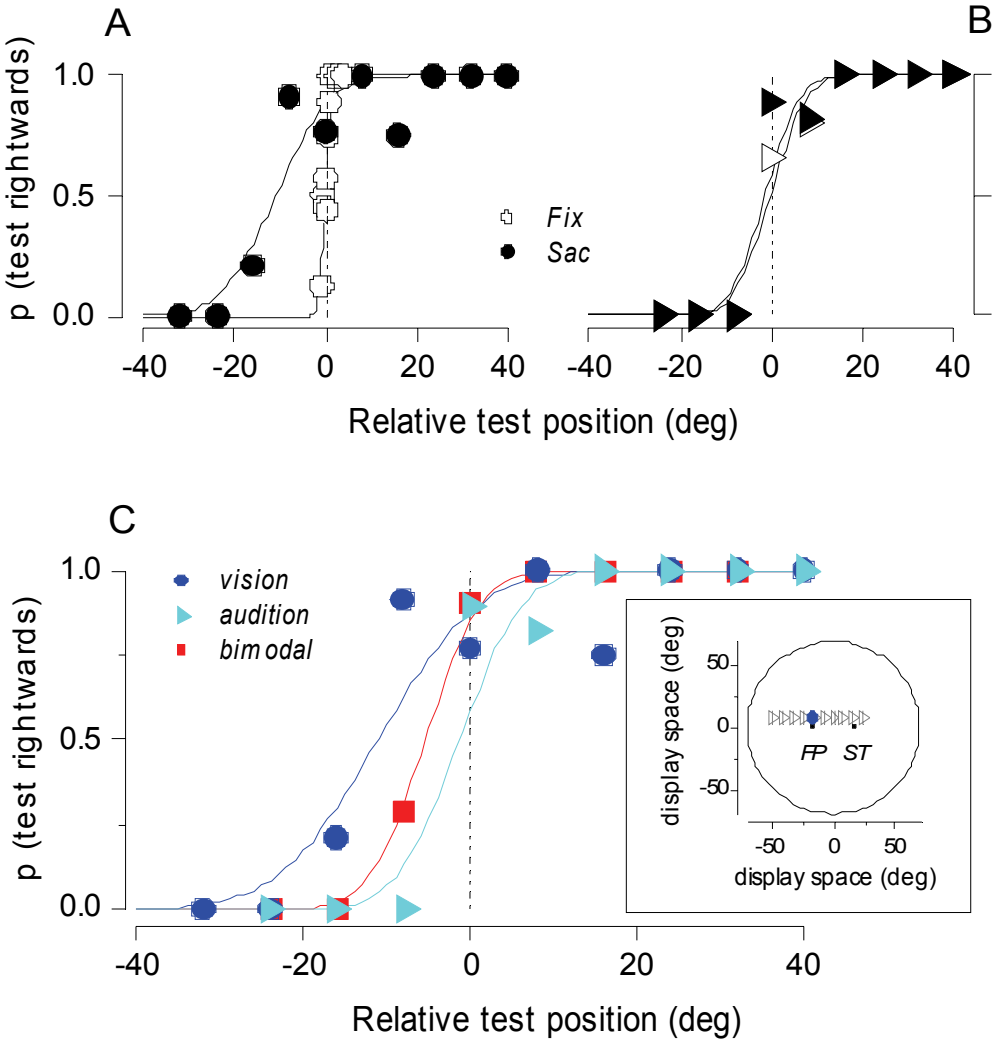
*determinants of perception.*: Chicago: University of Chicago Press.
Granrud, C. E. S., T. T. (2006). Development of size constancy in children: a test of the proximal mode sensitivity hypothesis. *Percept Psychophys, 68*, 1372-1381.
Hagen, M. C., Franzen, O., McGlone, F., Essick, G., Dancer, C., & Pardo, J. V. (2002). Tactile motion activates the human middle temporal/V5 (MT/V5) complex. *Eur J Neurosci, 16*(5), 957-964.
Hartcher-O'Brien, J., & Alais, D. (2007). *Temporal Ventriloquism: Perceptual shifts forwards and backwards in time predicted by the maximum likelihood model.* Paper presented at the 8th Annual Meeting of the International Multisensory Research Forum, University of Sydney, Australia.
Hubel, D. H. W., T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. (London), 195*, 215-243.
Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Res, 39*(21), 3621-3629.
Kovács, I., Kozma, P., Fehér, A., Benedek, G. (1999). Late maturation of visual spatial integration in humans. *Proc Natl Acad Sci U S A., 96*(21), 12204-12209.
Krekelberg, B., Kubischik, M., Hoffmann, K. P., & Bremmer, F. (2003). Neural correlates of visual localization and perisaccadic mislocalization. *Neuron, 37*(3), 537-545.
Kubischik, M. (2002). *Dynamic spatial representations during saccades in the macaque parietal cortex.* Ruhr-Universitaet Bochum), Bochum.
Lewkowicz, D. J. (1992). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Percept Psychophys, 52*(5), 519-528.
Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychol Bull, 126*(2), 281-308.
Ma, W. J., Beck, J. M., & Pouget, A. (2008). Spiking networks for Bayesian inference and choice. *Curr Opin Neurobiol, 18*(2), 217-222.
Mateeff, S., Hohnsbein, J., & Noack, T. (1985). Dynamic visual capture: apparent auditory motion induced by a moving visual target. *Perception, 14*(6), 721-727.
Matin, L., & Pearce, D. G. (1965). Visual perception of direction for stimuli flashed during voluntary saccadic eye movements. *Science*(148), 1485–1488.
Meyer, G. F., & Wuerger, S. M. (2001). Cross-modal integration of auditory and visual motion signals. *Neuroreport, 12*(11), 2557-2560.
Mills, A. (1958). On the minimum audible angle. *J. Acoust. Soc.. Am., 30*, 237-246.
Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: examining temporal ventriloquism. *Brain Res Cogn Brain Res, 17*(1), 154-163.
Morrone, M. C., Binda, P., & Burr, C. D. (2008). *Perception of space and time during saccade: a Bayesian explanation for perisaccadic distortions.* Paper presented at the Ann. Meeting of the Vision Science Society, Naples, Florida.
Morrone, M. C., Ross, J., & Burr, D. C. (1997). Apparent position of visual targets during real and simulated saccadic eye movements. *J Neurosci, 17*(20), 7941-7953.
Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: discrimination may be better than detection. *Vision Res., 14*, 1039-1042.
Nakamura, K., & Colby, C. L. (2002). Updating of the visual representation in monkey striate and extrastriate cortex during saccades. *Proc Natl Acad Sci U S A, 99*(6), 4026-4031.
Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Curr Biol, 18*(9), 689-693.
Neil, P. A., Chee-Ruiter, C., Scheier, C., Lewkowicz, D. J. & Shimojo, S. . (2006). Development of multisensory spatial integration and perception in humans. *Dev Sci, 9*, 454-464.
Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination. *J. Opt. Soc. Am., A2*, 1508-1532.

Perrott, D., & Saberi, K. (1990). Minimum audible angle thresholds for sources varying in both elevation and azimuth. *J. Accoust. Soc. Am., 87*, 1728-1731.
Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgements of spatial direction. *Percept Psychophys, 6*, 203-205.
Rentschler, I., Jüttner, M., Osman, E., Müller, A., Caelli, T.,. (2004). Development of configural 3D object recognition. *Behav Brain Res.. 149*(1), 107-111.
Ricciardi, E., Vanello, N., Dente, D., Sgambellur, i. N., Scilingo, E., Gentili, C., et al. (2004). *Perception of visual and tactile flow activates common cortical areas in the human brain. .* Paper presented at the Proceedings of EuroHaptics Munich, Germany.
Ross, J., Morrone, M. C., Goldberg, M. E., & Burr, D. C. (2001). Changes in visual perception at the time of saccades. *Trends Neurosci, 24*(2), 113-121.
Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408*(6814), 788.
Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport, 12*(17), 3849-3852.
Shipley, T. (1964). Auditory Flutter-Driving of Visual Flicker. *Science, 145*, 1328-1330.
Simpson, W. A., & Finsten, B. A. (1995). Pedestal effect in visual motion discrimination. *J Opt Soc Am A Opt Image Sci Vis, 12*(12), 2555-2563.
Sommer, M. A., & Wurtz, R. H. (2006). Influence of the thalamus on spatial visual processing in frontal cortex. *Nature, 444*(7117), 374-377.
Stein, B. E., Labos, E. & Kruger, L. (1973). Sequence of changes in properties of neurons of superior colliculus of the kitten during maturation. *J Neurophysiol, 36*, 667-679w.
Stein, B. E., Meredith, M. A. & Wallace, M. T. (1993). The visually responsive neuron and beyond: multisensory integration in cat and monkey. *Prog Brain Res, 95*, 79-90.
Streri, A. (2003). Cross-modal recognition of shape from hand to eyes in human newborns. *Somatosens Mot Res, 20*(1), 13-18.
Thilo, K. V., Santoro, L., Walsh, V., & Blakemore, C. (2003). The site of saccadaic suppression. *Nature Neurosci*.
Tootell, R. B. e. a. (1998). Functional analysis of primary visual cortex (V1) in humans. *Proc Natl Acad Sci U S A, 95*, 811-817.
Wallace, M. T., & Stein, B. E. (2001). Sensory and multisensory responses in the newborn monkey superior colliculus. *J Neurosci, 21*(22), 8886-8894.
Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: implications for transitivity among the spatial senses. *Percept Psychophys, 30*(6), 557-564.
Watson, A. B., & Pelli, D. G. (1983). Quest: a bayesian adaptive psychometric method. *Perception and Psychophysics*(33), 113 - 120.
Watt, R. J., & Morgan, M. J. (1983). The recognition and representation of edge blur: evidence for spatial primatives in human vision. *Vision Res., 23*, 1457-1147.
Wuerger, S. M., Hofbauer, M., & Meyer, G. F. (2003). The integration of auditory and visual motion signals at threshold. *Percept Psychophys, 65*(8), 1188-1196.
Wurtz, R. H. (2008). Neuronal mechanisms of visual stability. *Vision Res, 48*(20), 2070-2089.
Zeigler, H. P., & Leibowitz, H. (1957). Apparent visual size as a function of distance for children and adults. *Am J Psychol, 70*(1), 106-109.

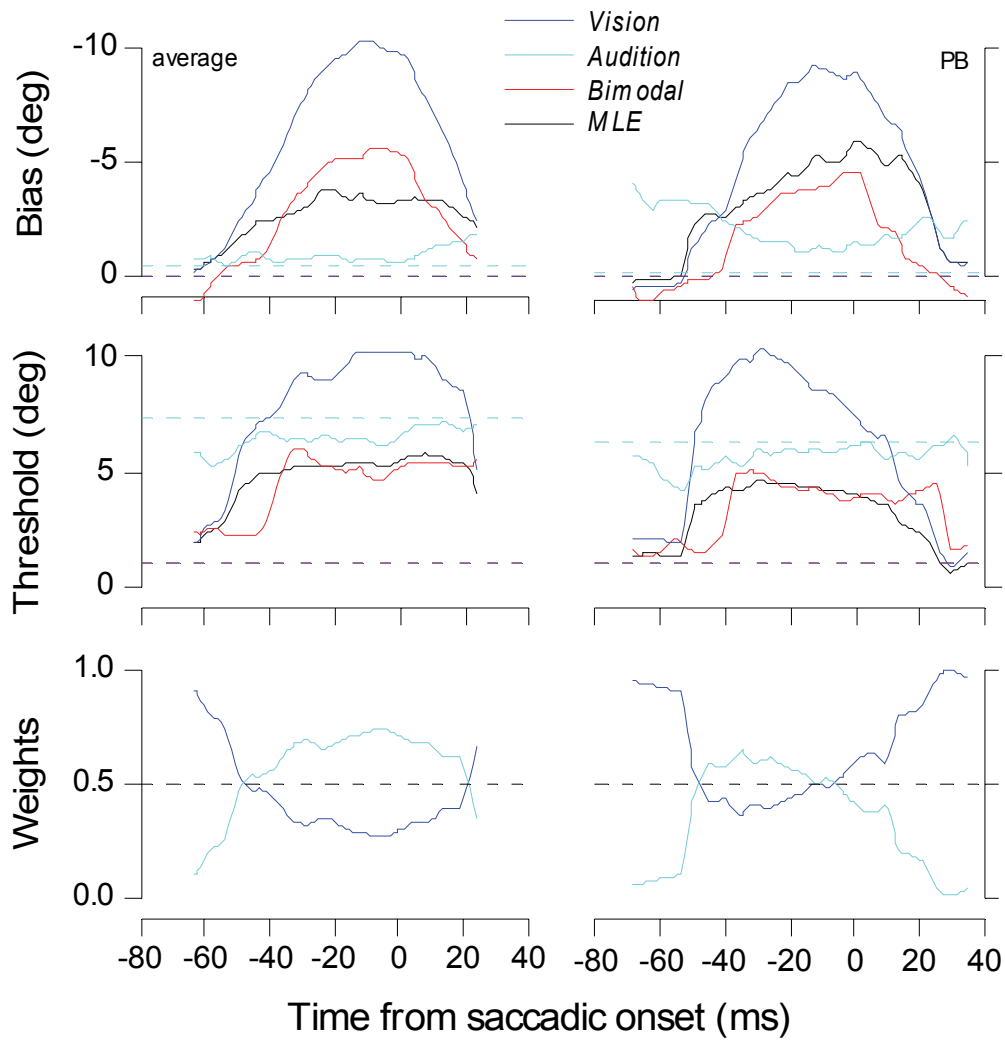

A
Auditory standard (-Δ
Visual standard (Δ)
1.0
0.5
0.0
Proportion "right"
4 deg
32 deg
64 deg
-20
-10
0
10
20
Rightward displacement of probe (deg)
B
64 deg
32 deg
4 deg
PSEs Measured (deg)
6
4
2
0
-2
-4
-6
-6 -4 -2 0 2 4 6
PSEs Predicted (deg)
C
Noramlised Threshold
1.5
1.0
0.5
0.0
Vis
Aud
Bimodal
MLE

# Figure 1

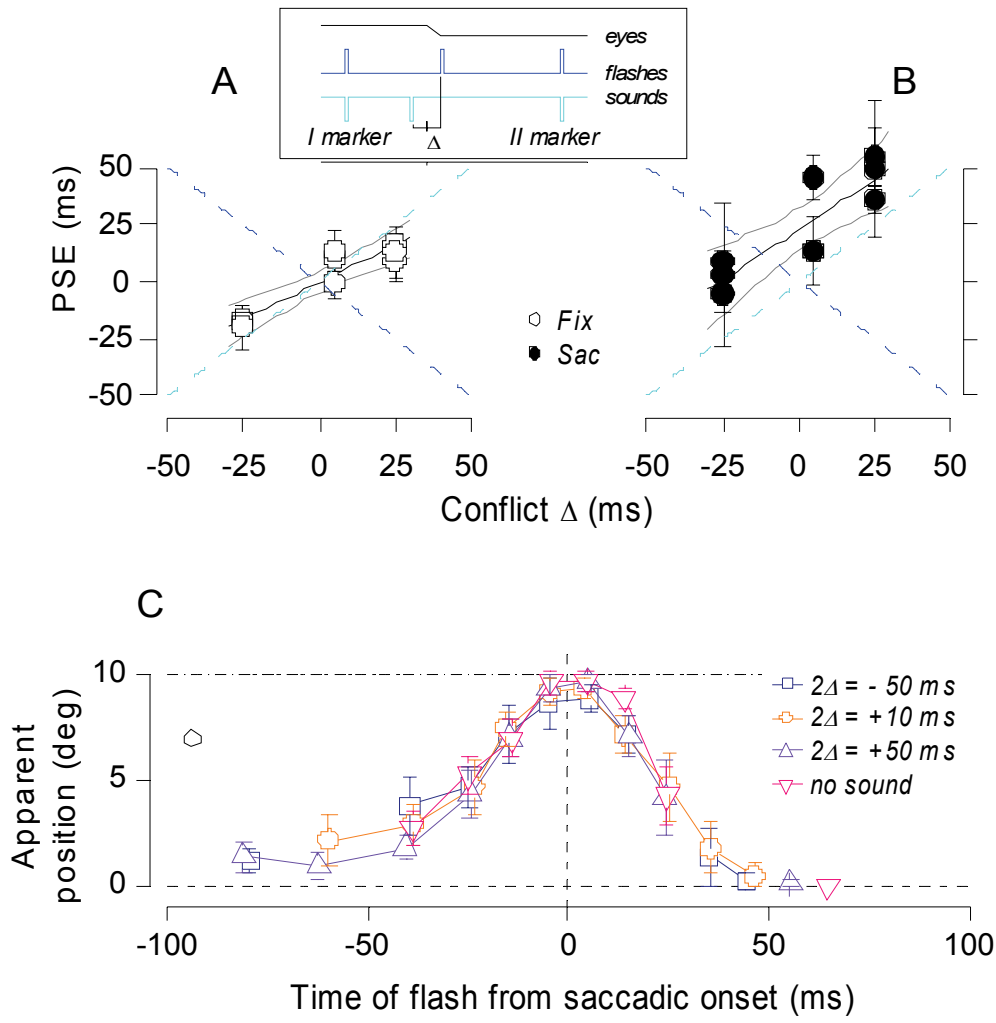
# Figure 2


A
B
C
p (test rightwards)
1.0
0.5
0.0
-40
-20
0
20
40
Relative test position (deg)
Fix
Sac
vision
audition
bimodal
display space (deg)
50
0
-50
FP
ST

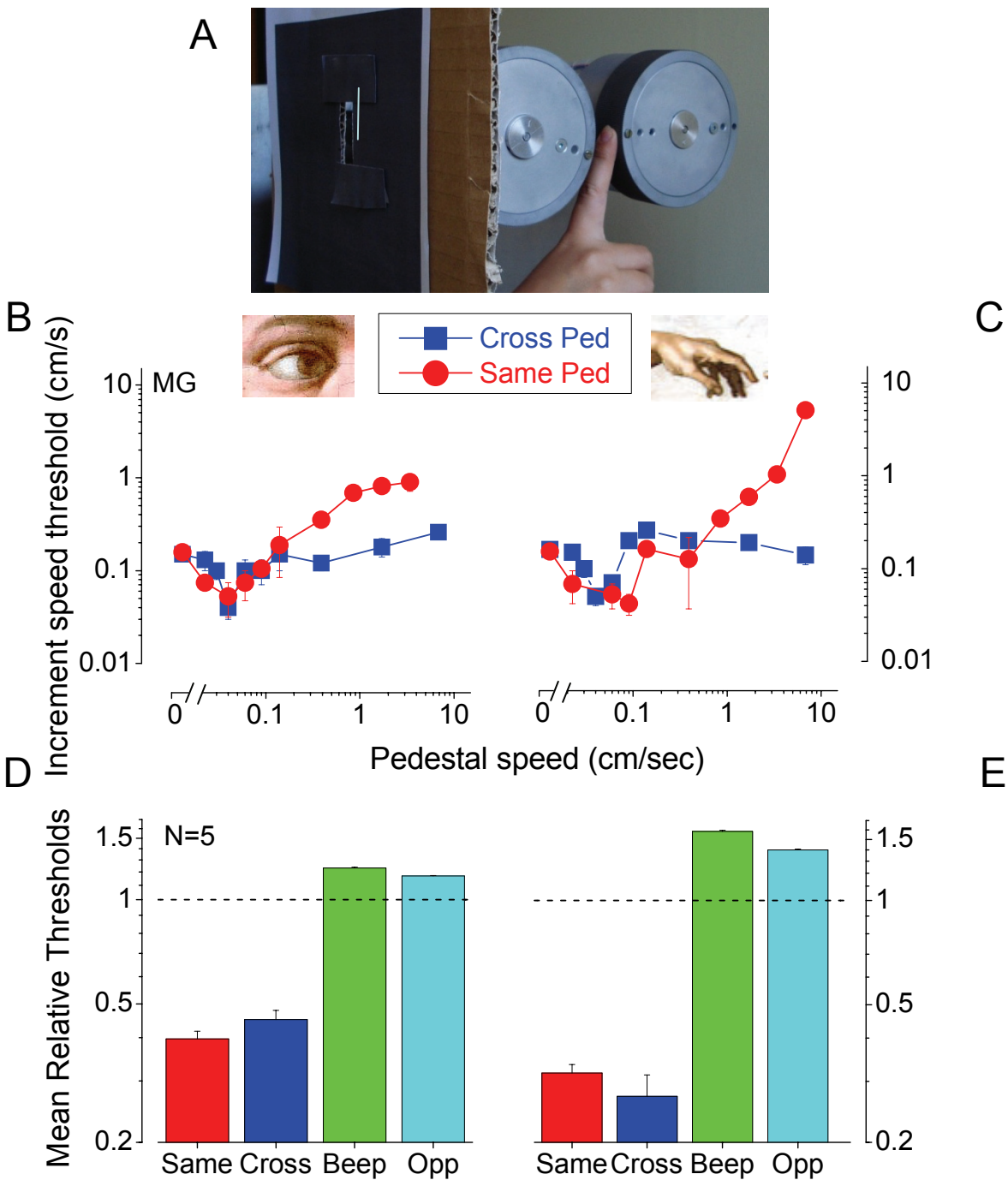
# Figure 3


Vision
Audition
Bimodal
MLE
average
PB
Bias (deg)
-10
-5
0
Threshold (deg)
10
5
0
Weights
1.0
0.5
0.0
-80 -60 -40 -20 0 20 40
-80 -60 -40 -20 0 20 40
Time from saccadic onset (ms)

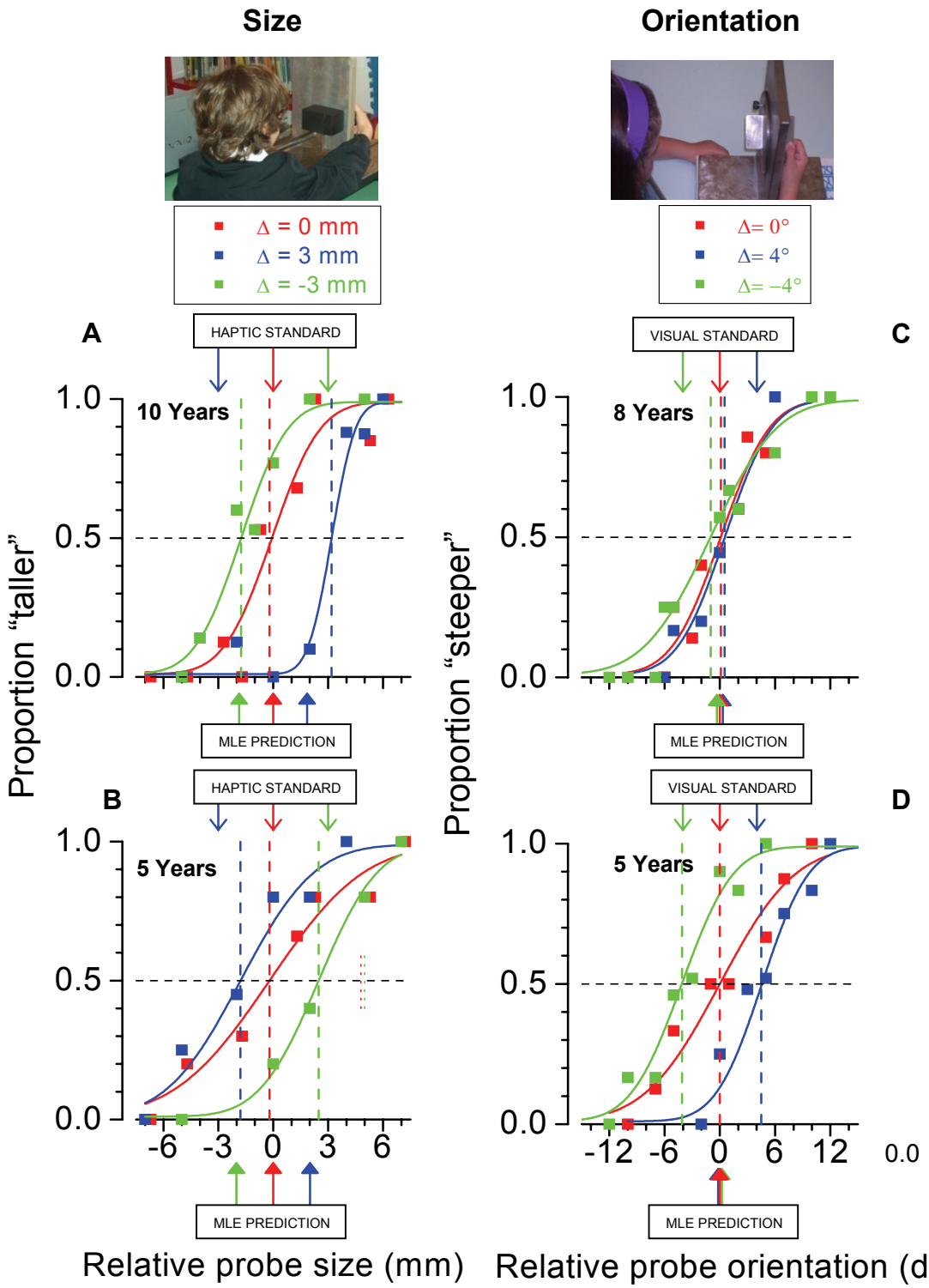
# Figure 4


A
B
eyes
flashes
sounds
I marker
Δ
II marker
PSE (ms)
50
25
0
-25
-50
Fix
Sac
-50
-25
0
25
50
Conflict Δ (ms)
C
Apparent position (deg)
10
5
0
2Δ = - 50 ms
2Δ = +10 ms
2Δ = +50 ms
no sound
-100
-50
0
50
100
Time of flash from saccadic onset (ms)

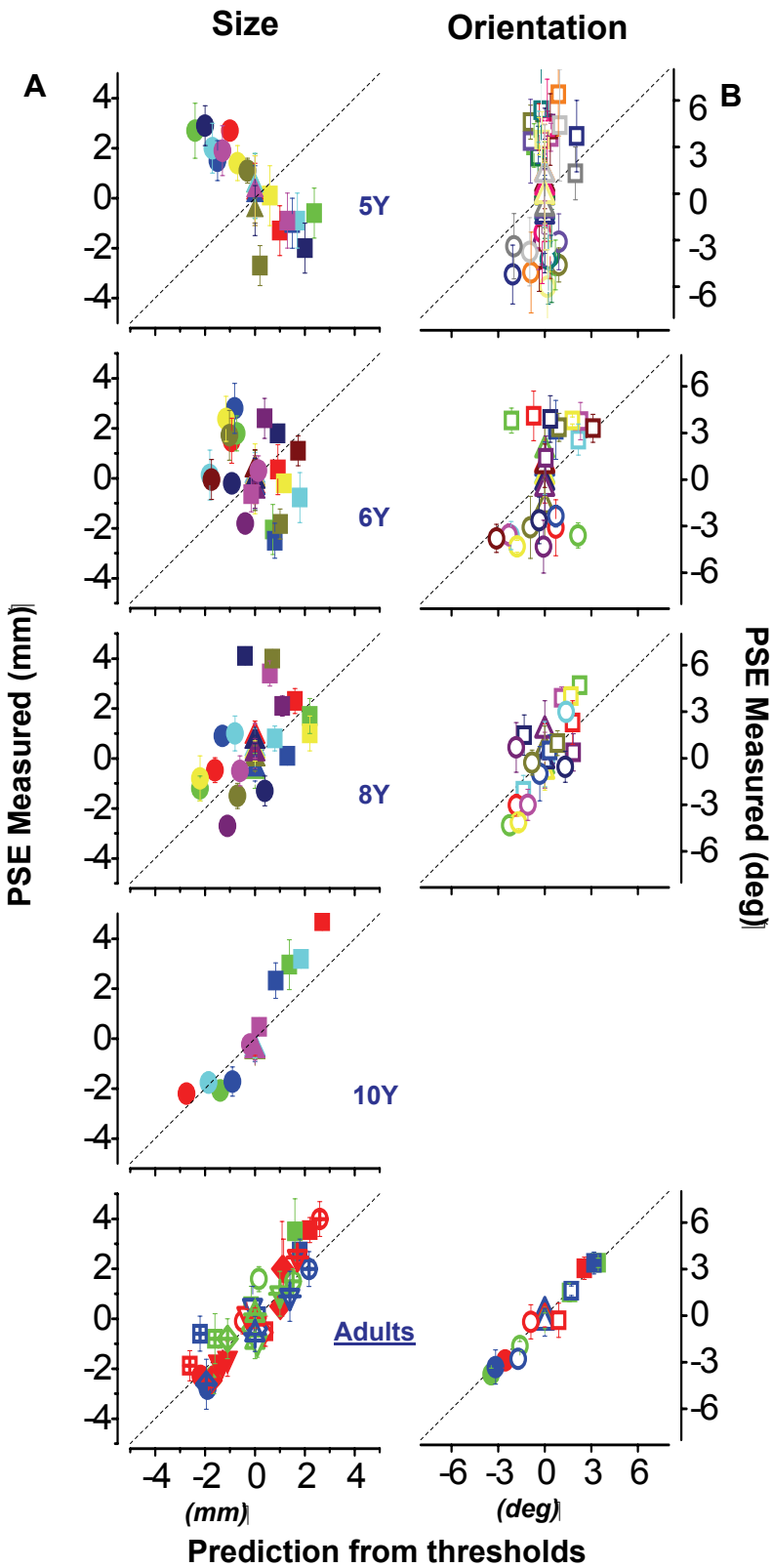
# Figure 5


A
B
C
D
E
MG
Cross Ped
Same Ped
Increment speed threshold (cm/s)
10
1
0.1
0.01
0
0.1
1
10
Pedestal speed (cm/sec)
Mean Relative Thresholds
N=5
1.5
1
0.5
0.2
Same
Cross
Beep
Opp

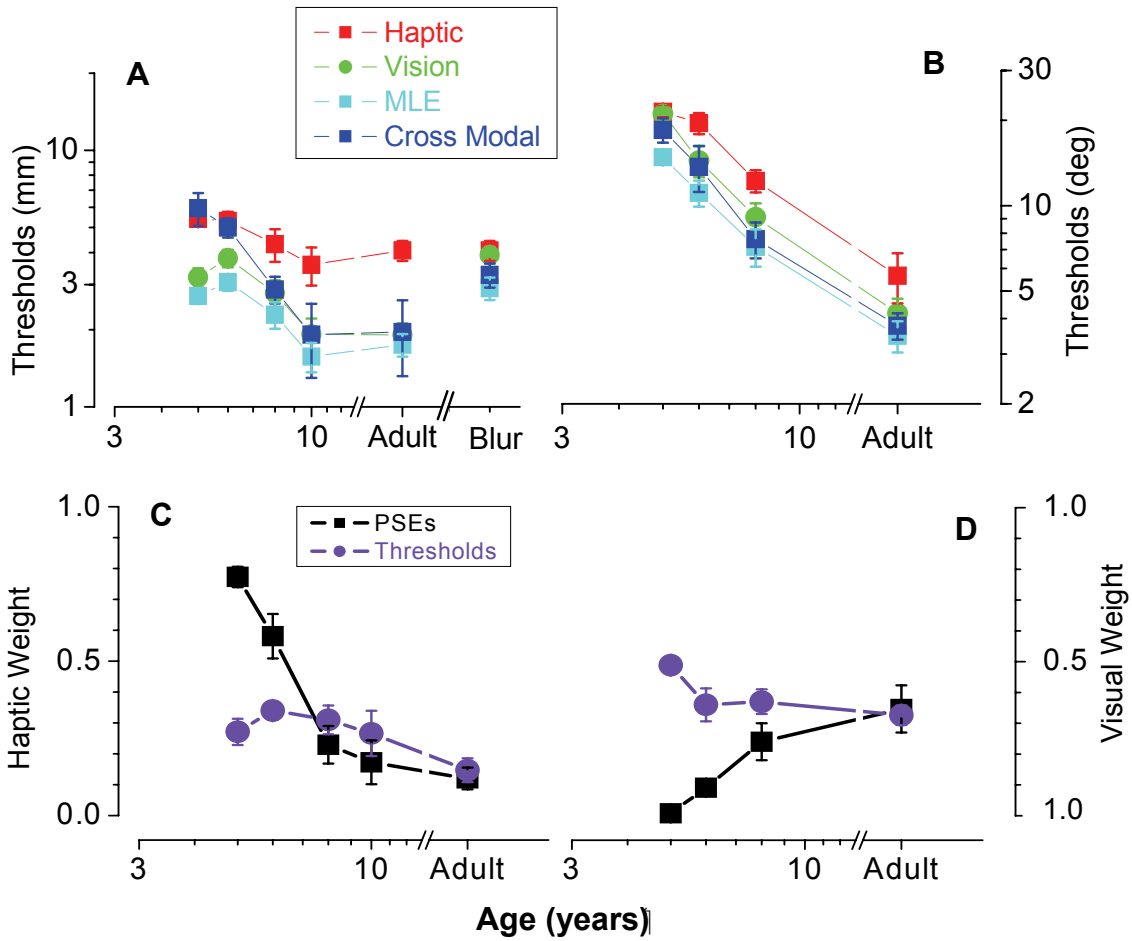
# Figure 6


Size
Orientation
Δ = 0 mm
Δ = 3 mm
Δ = -3 mm
Δ= 0°
Δ= 4°
Δ= –4°
A
B
C
D
HAPTIC STANDARD
VISUAL STANDARD
MLE PREDICTION
10 Years
8 Years
5 Years
Proportion "taller"
Proportion "steeper"
1.0
0.5
0.0
-6 -3 0 3 6
-12 -6 0 6 12
Relative probe size (mm)
Relative probe orientation (deg)

# Figure 7


Size
Orientation
A
B
5Y
6Y
8Y
10Y
Adults
PSE Measured (mm)
PSE Measured (deg)
(mm)
(deg)
Prediction from thresholds

# Figure 8


A
B
C
D
Haptic
Vision
MLE
Cross Modal
Thresholds (mm)
Thresholds (deg)
Haptic Weight
Visual Weight
PSEs
Thresholds
3
10
Adult
Blur
30
5
2
1.0
0.5
0.0
Age (years)