Chapter 2
Cue Combination Within a Bayesian Framework

David Alais and David Burr

Abstract To interact effectively with the world, the brain must optimize its perception of the objects and events in the environment, many of which are signaled by more than one sense. Optimal perception requires the brain to integrate redundant cues from the different senses as efficiently as possible. One effective model of cue combination is maximum likelihood estimation (MLE), a Bayesian model that deals with the fundamental uncertainty and noise associated with sensory signals and provides a statistically optimal way to integrate them. MLE achieves this through a weighted linear sum of two or more cues in which each cue is weighted inversely to its variance or "uncertainty." This produces an integrated sensory estimate with minimal uncertainty and thus maximized perceptual precision. Many studies show that adults integrate redundant sensory information consistent with MLE predictions. When the MLE model is tested in school-aged children, it is found that predictions for multisensory integration are confirmed in older children (>10 years) but not in younger children. Younger children show unisensory dominance and do not exploit the statistical benefits of multisensory integration, even when their dominant sense is far less precise than the other. This curious finding may result from each sensory system having an inherent specialization, with each specialist sense tuning the other senses, such as vision calibrating audition for space (or audition calibrating vision for time). This cross-sensory tuning would preclude useful combination of two senses until calibration is complete, after which MLE integration provides an excellent model of multisensory cue combination.

D. Alais (✉)
School of Psychology, The University of Sydney, Sydney, NSW, Australia
e-mail: David.Alais@sydney.edu.au

D. Burr
Neuroscience Institute, National Research Council, Pisa, Italy

Department of Neuroscience, University of Florence, Florence, Italy
e-mail: Dave@in.cnr.it

A. K. C. Lee et al. (eds.), *Multisensory Processes*, Springer Handbook of Auditory Research 68, https://doi.org/10.1007/978-3-030-10461-0_2

Keywords Bayesian · Calibration · Cross-modal · Cue combination · Development · Maximum likelihood model · Optimal integration · Sensory deficit · Sensory integration · Sensory noise · Space perception · Time perception · Uncertainty reduction · Weighted sum

2.1 Multisensory Integration and the Problem of Cue Combination

The years since the turn of the twenty-first century have witnessed an explosion of research activity in multisensory processing. Prior to this, most sensory research, whether cognitive or neurophysiological, focused on each modality separately and did not seek to understand multisensory integration (Jones and Powell 1970; Benevento et al. 1977). This reflected the prevailing view of cortical organization that each sensory modality initially processed information independently and that sensory integration or "binding" only occurred at later stages of processing in polysensory association areas of the brain. On this view, the emphasis on unisensory research was sensible and provided a tractable starting point for sensory research when relatively little was known about cortical processing. However, recent findings show the brain's neural architecture contains more connectivity between early unisensory areas than was previously known (Kayser et al. 2008; Murray et al. 2015). The early interaction between unisensory cortices probably reflects the fact that many of the stimuli in the environment are fundamentally multisensory in nature and activate multiple senses, each one encoding a complementary aspect of the stimulus, with the multiple representations also providing redundancy (e.g., of spatial location, timing, intensity). It is appropriate that these sensory signals be combined early so that external stimuli are coherently represented as multisensory objects and events as early as possible, but that raises the question of how to combine different sensory signals efficiently and effectively. This chapter reviews a Bayesian approach to multisensory cue combination known as maximum likelihood estimation (MLE) that provides a statistically optimal model for cue combination and provides a good account of many instances of multisensory integration.

Humans live in a multisensory world where an event in the environment often produces signals in several senses. These multiple signals provide redundant and complementary information about the event, and when they are spatially and temporally correlated (typically, the case for signals originating from a common event), the brain exploits these properties by combining responses across sensory modalities. This is a sensible strategy that brings considerable benefits. First, the redundancy of a multisensory representation provides great flexibility, preventing catastrophic failures of perception if one sense is permanently lost or if environmental conditions render one sense temporarily ineffective (e.g., vision is impaired at night; a critical sound is masked by background noise). Second, the statistical advantages of having two samples of the same stimulus leads to important perceptual benefits, seen in faster reactions times and better discrimination of multisensory stimuli (Alais et al. 2010). This latter aspect of multisensory perception has received a good deal of attention in the last couple of decades, and it is clear that the key to

robust and coherent perception is the efficient combination of multiple sources of sensory information (Ernst and Bulthoff 2004). Although the perceptual benefits of multisensory integration are well established, understanding how the brain achieves this integration remains a challenging question in sensory and cognitive neuroscience. Moreover, it is not a trivial problem for the brain to solve because the information to be combined arrives in different primary cortices, is often offset in time, and is mapped (at least initially) in different coordinate systems.

2.2 Cue Combination in a Bayesian Framework

Despite the challenges in doing so, the human perceptual system has an impressive ability to seamlessly integrate the senses into a coherent and reliable perception of the external world. It achieves this despite working with neural signals that are inherently noisy and variable. This variability means that perception is intrinsically a probabilistic process (Fig. 2.1A and B), making interpretations and inferences about the likely nature of external stimuli in a process known as "unconscious inference," as von Helmholtz (1925) termed it. Prior knowledge acquired through experience of the world plays a role in guiding these perceptual inferences, as do the incoming sensory signals. A Bayesian framework (Kersten et al. 2004; Knill and Pouget 2004; Pouget et al. 2013) is perfectly suited to modeling perceptual inference for two reasons. First, it is a mathematical model based on probabilities. Second, its two components, called the prior probability and the likelihood, map perfectly onto the two sources of information for perceptual inference: acquired knowledge of the sensory world (the prior) and incoming noisy sensory signals (the likelihood).

Bayes' theorem states that the posterior probability is proportional to the product of the prior probability and the likelihood (Fig. 2.1C). The prior describes the probability of a stimulus before any stimulus information is received and thus reflects, for example, learning, knowledge, and expectations. The likelihood is the probability of the stimulus given its possible states. As applied to perception and behavior, the prior is thought of as an internal model of the statistics of the environment and the likelihood represents an incoming noisy sensory signal. In the case of multisensory stimuli, there will be signals in two or more modalities and a corresponding likelihood for each component. In the audiovisual example shown in Eq. 2.1, the likelihoods for the auditory and visual stimuli are the first two terms of the numerator and the prior is the third term. Multiplicatively combining these three terms (or four terms for a trimodal stimulus) satisfies Bayes' theorem. If this product is then normalized by the product of the simple probabilities for each component, we obtain Bayes' equality. Equation 2.1 shows Bayes' equality for combining two estimates (here, estimates of spatial location [*S*] from auditory and visual cues, with *P* indicating probability) into an integrated multisensory estimate

$$P\left(S_{AV} \mid S_A,\ S_V\right) = \frac{P\left(S_A \mid S_{AV}\right) P\left(S_V \mid S_{AV}\right) * P\left(S_{AV}\right)}{P\left(S_A\right) P\left(S_V\right)} \tag{2.1}$$

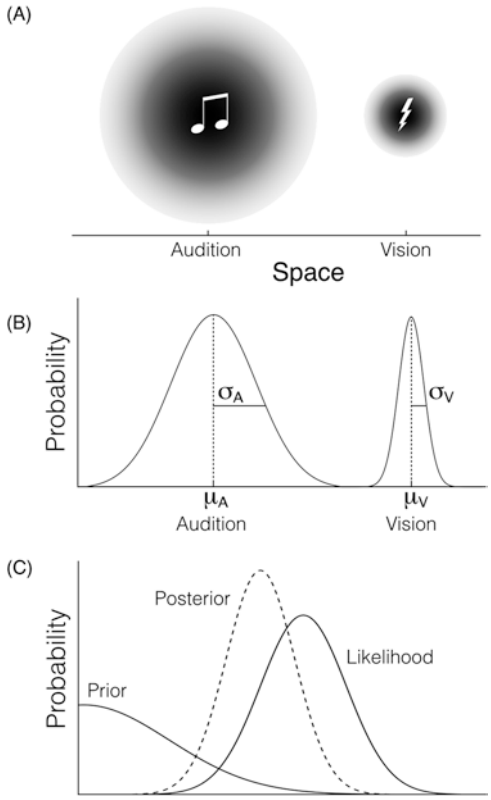

Fig. 2.1 (**A**) The world is crudely sampled through receptive fields of various sizes generating noisy neural signals. Together, these factors degrade the precision of perception. Here the example of spatial location is illustrated, an attribute much more precisely coded in vision than audition. (**B**) The noise accompanying a signal can be modeled by a Gaussian distribution described by two parameters, the mean (μ) and the standard deviation (σ). For spatial location, an auditory estimate is less precise (i.e., higher standard deviation) than a visual one. (**C**) Bayesian theory, being based on probability distributions, provides a convenient way to model the combination of noisy information. Its two components are the prior distribution and the likelihood distribution. Incoming sensory information constitutes the "likelihood," whereas acquired knowledge of the world and its statistics are embodied in the "prior." These can be combined (multiplied) to produce the posterior distribution, an optimal fusion of stored knowledge and current sensory information

One model that has been successful in accounting for many instances of multisensory cue combination is the maximum likelihood estimation (MLE) model. MLE is a simplified Bayesian model that only takes account of the likelihood (it has no prior component, the final term in the numerator of Eq. 2.1). MLE describes how noisy sensory information can be combined from two or more independent sources (e.g., auditory and visual signals). It takes account of the variability of each signal and combines them in a statistically optimal fashion that maximizes the likelihood that the combined response will truly reflect the external stimulus (Fig. 2.2A).

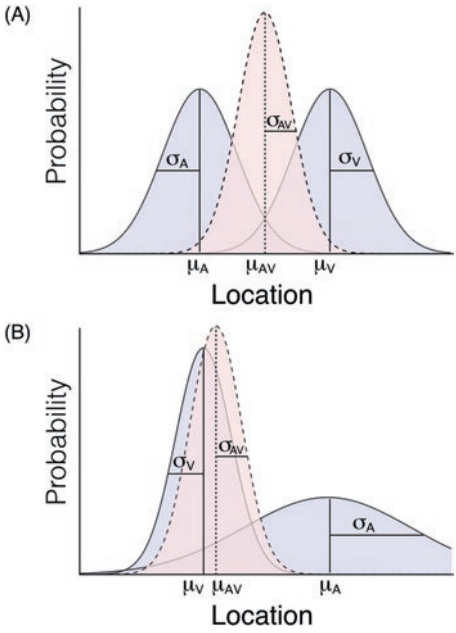

Fig. 2.2 In many cases, perceptual judgments require no access to the stored information and expectations represented by the prior and the Bayesian model simplifies to the likelihood. In multisensory contexts (such as audiovisual localization), each signal will produce a likelihood and combining them produces a product distribution with the highest possible probability, known as the maximum likelihood. Maximizing the likelihood is desirable because it will minimize the distribution's variance, corresponding to maximal perceptual resolution. (**A**) Here the two likelihoods have identical standard deviations but different means. From Eq. 2.3, equal widths (σ) lead to equal component weights, and thus the combined "maximum likelihood" distribution is located at the mean position (see Eq. 2.2), with a narrower width (see Eq. 2.4). (**B**) If component distributions have different widths, their weighting in the product distribution will differ, as per Eqs. 2.2 and 2.3. In effect, the product distribution's location is drawn toward the narrower, and thus the perceptually more reliable, component. Regardless of the relative widths of the component distributions, the product distribution will always be the solution providing the maximum possible probability and thus the minimal standard deviation

In maximizing the likelihood, it minimizes stimulus uncertainty. In essence, MLE is a weighted linear sum that combines two or more signals, each weighted by its reliability. Reliable signals receive a high weight, whereas unreliable signals receive a low weight (Fig. 2.2B). The combination rule is considered statistically optimal because it always provides the result that is most reliable, where "most reliable" means the most probable or least variable solution. In producing the least variable combination, the MLE model effectively minimizes stimulus uncertainty arising from noise in the component signals.

2.3 The Maximum Likelihood Estimation Model

The MLE model is best demonstrated by working through an example of multisensory perception. One of the best-known examples of how the perceptual system deals with redundant spatial signals is the ventriloquist effect (Howard and Templeton 1966). In this effect, provided the auditory and visual stimuli are aligned in time to be synchronous or nearly so (Slutsky and Recanzone 2001), displacing the visual stimulus over modest distances will usually cause the auditory stimulus to be "captured" by the visual event (i.e., perceived as colocalized with the visual stimulus). Being simultaneous and roughly collocated, the signals satisfy the conditions for audiovisual fusion, but how best to fuse them? MLE assumes that the signal in each sensory modality provides an independent estimate about a particular stimulus attribute (here, estimated spatial location, $\hat{s}$) and has a Gaussian-distributed uncertainty. The estimate and its uncertainty are represented by the mean and variance, respectively, of a Gaussian probability distribution. MLE combines the auditory and visual estimates in a weighted linear sum to obtain the estimated bimodal spatial location

$$\hat{s}_{AV} = w_A \hat{s}_A + w_V \hat{s}_V \tag{2.2}$$

where w_A and w_V are the weights allocated to the component modalities. The weights are determined by the relative reliability of each modality's estimate of the stimulus attribute where variance (σ^2) and reliability are inversely related

$$w_A = \frac{1/\sigma^2{}_A}{1/\sigma^2{}_A + 1/\sigma^2{}_V} = \frac{\sigma^2{}_V}{\sigma^2{}_A + \sigma^2{}_V} \tag{2.3}$$

Equation 2.3 shows that the auditory weight and the visual weight are easily obtained by changing the subscript of the numerator. As should be clear from Eq. 2.3, each modality accounts for a proportion of total variance and thus the component weights are relative weights and sum to 1. In short, the more variable a modality is in contributing to the perceptual estimate, the less reliable it is and the less it is weighted in the bimodal percept. The MLE solution is optimal because it provides the combined estimate with the lowest variance, given the available information, and thus provides maximal stimulus precision. Indeed, the combined variance can never be larger than either of the components because of the following relationship

$$\sigma^2{}_{AV} = \frac{\sigma^2{}_A \sigma^2{}_V}{\sigma^2{}_A + \sigma^2{}_V} \tag{2.4}$$

From Eq. 2.4, the combined estimate must always have a lower variance than either of the components. The reduction in combined variance (and consequent gain in precision) is maximal when the component variances are equal, reducing variance in that case by a factor of $\sqrt{2}$ (Fig. 2.2A). This benefit reduces if the compo-

nent variances diverge, and in the limit, very different component variances produce a combined variance that approaches the value of the smaller of the component variances (Fig. 2.2B).

The MLE integration rule therefore makes two key predictions when two signals are combined because it specifies both the mean value of the combined estimate ($\hat{s}_{AV}$) and its variance (σ^2_{AV}). These predictions have been tested and confirmed in a range of different multisensory contexts, showing that multisensory integration closely approximates the MLE model (Clarke and Yuille 1990; Ghahramani and Wolpert 1997; Landy et al. 2011). Examples include audiovisual spatial localization (Alais and Burr 2004) and visual-tactile size estimation (Ernst and Banks 2002). MLE has even been demonstrated in trimodal contexts (Wozny et al. 2008), but it may also occur within a single modality between independent cues (Hillis et al. 2002). The available evidence suggests that MLE integration occurs automatically and does not require that attention to be directed to the component stimuli (Helbig and Ernst 2008). In multisensory contexts, there is evidence that the perceptual estimate of each modality's component cue are not lost when MLE integration occurs, although this appears not to be the case for cues within a single modality where MLE integration is obligatory and the component information is lost (Hillis et al. 2002).

2.4 Maximum Likelihood Estimation: A Flexible Cue Combination Model

The MLE model allows a useful reinterpretation of some earlier ideas in the multisensory literature. One prevalent idea was the "modality appropriateness hypothesis" that stated that conflicts between the modalities were resolved in favor of the most relevant modality (Welch and Warren 1980). In an audiovisual context, the most appropriate modality would be vision for a spatial task and audition for a temporal task. The MLE model supersedes the modality appropriateness hypothesis without resorting to arbitrary notions such as "appropriateness." MLE predicts a dominance of vision over audition for spatial judgments (such as in ventriloquism) because spatial resolution is higher in the visual domain, which means less uncertainty and a higher weighting for vision relative to audition. Conversely, MLE predicts that audition should dominate vision for temporal tasks, such as in auditory driving (Shipley 1964; Recanzone 2003) or for the "double flash" illusion (Shams et al. 2000) because the auditory modality is specialized for temporal processing. Of course, modality appropriateness predicts the same dominances, but it does so within an arbitrary and rigid framework, whereas MLE is flexible and will weight the components in favor of the incoming stimulus with the higher certainty. This flexibility was shown clearly in Alais and Burr's (2004) ventriloquism study (Fig. 2.3) where they demonstrated both conventional ventriloquism and reverse ventriloquism (i.e., auditory capture of visual locations). The reverse ventriloquism

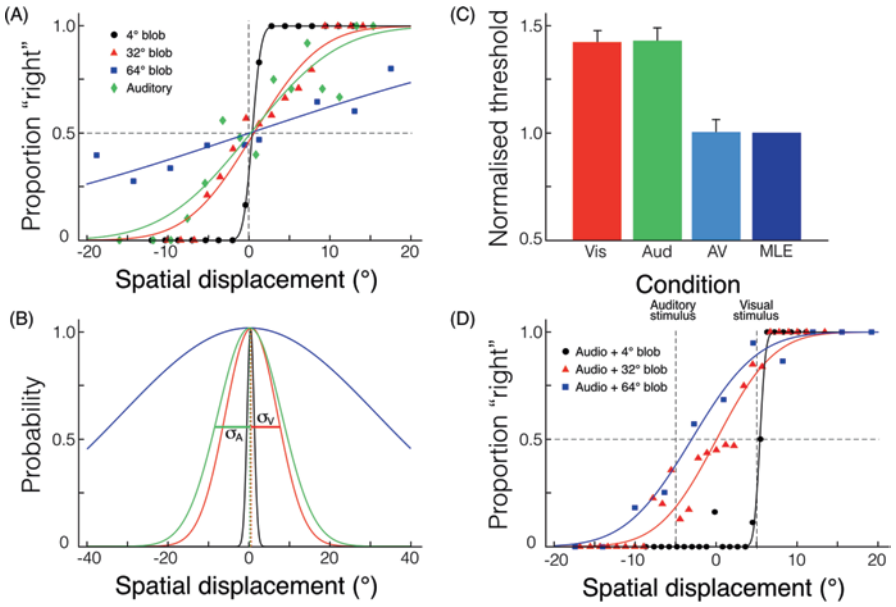

Fig. 2.3 Applying the maximum likelihood estimation model to psychophysics (adapted from Alais and Burr 2004). (**A**) Cumulative Gaussian psychometric functions for localizing an auditory click or Gaussian blobs of various widths ($2\sigma = 4$, 32, or 64°). Functions all pass through ≈0° (all stimuli accurately localized on average) but varied systematically in width. The width is given by the σ term in the cumulative Gaussian equation and defines the discrimination threshold. (**B**) Functions from (**A**) replotted as probability densities to highlight their standard deviations (i.e., σ). The auditory and middle-sized visual stimuli have similar widths and should produce a near-maximal reduction in the combined distribution's width (close to the maximum $\sqrt{2}$ reduction for components of equal widths). (**C**) Audiovisual localization precision (normalized to 1.0) for the collocated auditory and middle-sized visual stimuli was better than for each component separately, indicating increased perceptual precision, and closely matched the maximum likelihood estimation (MLE) prediction. (**D**) Localization of the audiovisual stimulus when the components were separated by ±5° also followed MLE predictions. When the visual component was much better localized than the auditory one (*squares, black curve*), the mean audiovisual position shifted to the visual location (as in Fig. 2.2B). MLE thus accounts for the classic ventriloquist effect. When the auditory stimulus was paired with the poorly localized visual stimulus, audiovisual location was drawn to the (better localized) auditory component (reverse ventriloquism), as MLE predicts

occurred because the visual stimulus was blurred to the point that the auditory signal was more precisely localized (Fig. 2.3A). MLE correctly predicted auditory capture of vision when vision was blurred (Fig. 2.3D), whereas modality appropriateness adheres to a rigid dichotomy of visual spatial dominance and auditory temporal dominance.

MLE is not only a flexible combination rule rather than a rigid assumption of sensory dominances but also takes into account of all the available information. It has been clear since early multisensory studies (Rock and Victor 1964) that one sensory modality rarely dominates completely over another: there is always a residual contribution from the dominated modality. MLE captures this in that the estimate

from the less reliable modality is always summed into the combined estimate but is simply downweighted if it has low reliability. It therefore contributes to the combined estimate but with a reduced influence. In this way, MLE provides an intrinsically multisensory estimate, whereas modality appropriateness chooses the most appropriate single modality. The MLE model therefore provides a flexible, quantitative, and principled alternative to the modality appropriateness hypothesis and provides a convenient mathematical framework for combining sensory estimates with their inherent noise and uncertainty.

The MLE model also lends itself readily to psychophysical investigation of multisensory phenomena. This is because simple experiments in which the subject discriminates each of the unisensory components (e.g., which interval is louder, brighter, bigger, more rightward) provide psychometric data that can be modeled with a cumulative Gaussian function to obtain estimates of the mean and its variance, the two parameters needed for Eqs. 2.2–2.4 (see Fig. 2.3A and B). This was the approach adopted in Ernst and Banks's (2002) study of visual-tactile size perception, subsequently applied in Alais and Burr's (2004) audiovisual study of ventriloquism. Both studies found strong evidence for MLE integration. A number of other studies have adopted similar approaches to quantify the variability of sensory estimates and have found results consistent with MLE integration (van Beers et al. 1999; Knill and Saunders 2003; Hillis et al. 2004). The data in Fig. 2.3A show position discrimination results from Alais and Burr (2004). In a two-interval, forced-choice procedure, either a sound source or a Gaussian luminance blob varied in location along the horizon in front of the observer. The subjects had to judge in which interval the stimulus was located further to the right. All stimuli were accurately localized at 0° (directly ahead of the observer) but with a degree of precision that varied with the size of the visual stimuli. As blob size increased over three levels, precision declined. In an analogous experiment varying the location of a sound source, position discrimination data were comparable with the middle-sized visual stimulus. Cumulative Gaussian functions were fitted to the data, which are plotted in Fig. 2.3B as probability densities to highlight the differences in variance.

The MLE model makes the strong prediction that pairing the sound and the middle-sized blob should produce audiovisual discrimination data with significantly higher precision. This is because these two stimuli have roughly equivalent variances and thus should produce an increase in precision that is close to the ideal maximum of $\sqrt{2}$. From Eq. 2.4, the predicted reduction in variance can be calculated and compared against empirical data for discriminating the audiovisual stimulus. As shown by the variances plotted in Fig. 2.3C, discrimination precision for the audiovisual stimulus was indeed significantly lower than for each of the component stimuli and was very close to the value predicted by the MLE model. The test of variance reduction is critical because it provides strong evidence that information was integrated across two sources to produce increased discrimination precision. It rules out alterative possibilities, such as switching between independent information sources, because this would produce a worse performance than the best of the components. It also argues against a probability summation account because this

may lead to improved discrimination but by less than a factor or $\sqrt{2}$ (making it imperative to closely match the component variances to distinguish between MLE and probability summation predictions).

The other prediction made by the MLE model concerns the mean of the combined distribution. When the component distributions are centered at different locations, the position of the combined distribution is not simply the average of the two but is a weighted average based on the variability of the components. As shown in Fig. 2.2, the product distribution is drawn to the component with the smaller variance, as predicted by Eqs. 2.2 and 2.3. This aspect of the MLE model is very relevant to multisensory processing because redundant stimulus estimates to be combined across different modalities will often be discrepant despite signaling the same event. This can happen in the temporal domain due to latency differences between the senses or in the spatial domain due to misaligned spatial maps. Provided the signals are roughly spatiotemporally aligned, the brain will attempt to integrate them, but where should the fused stimulus be located? As illustrated in Fig. 2.1A, auditory stimuli will normally be localized with less precision than visual stimuli, meaning that the fused estimate should be drawn toward the (more precise) visual location, according to the MLE model, as shown in Fig. 2.2B. This is the well-known ventriloquist effect. Note that based on the weights in Eq. 2.2 (determined by Eq. 2.3), the MLE model makes a specific quantitative prediction concerning by how much the lesser weighted spatial location should be drawn to the higher weighted location in the fused percept. In this way, it differs from a simple bias to favor one stimulus over the other and from the binary selectivity of the modality appropriateness hypothesis (Welch and Warren 1980) that holds that the most appropriate sense (vision, for a spatial task) will determine perception.

To test if the MLE model could provide an account of the ventriloquist effect, Alais and Burr (2004) compared two conditions. In one, the auditory stimulus and the well-localized visual stimulus (see Fig. 2.3A) were presented simultaneously at horizontally displaced locations and their perceived location was discriminated against the same stimuli both presented at 0° (directly in front of the observer). Location discrimination in this audiovisual condition was compared with another that paired the auditory stimulus with the poorly localized visual stimulus. In the first condition, the spatial discrepancy between the components was resolved by the audiovisual stimulus being localized very near to the visual location. This is the classic ventriloquist effect and is explicable in terms of competing accounts such as simple visual "capture" of auditory location and the modality appropriateness hypothesis (Welch and Warren 1980). However, only the MLE model could account for the second condition. In this case, where the visual stimulus was less reliable than the auditory stimulus, it was the auditory stimulus that dominated audiovisual localization (Fig. 2.3D). This result had never been reported before and is effectively a case of reverse ventriloquism because the location of the visual stimulus was drawn to the location of the auditory stimulus. Importantly, accounts such as modality appropriateness cannot explain such a result, but MLE can; simply, reverse

ventriloquism will occur whenever the auditory stimulus is better localized than the visual stimulus (as predicted by Eqs. 2.2 and 2.3).

More recently, interest has turned to the MLE model at the neural level (Rowland et al. 2007; Gu et al. 2008). The study by Gu et al. examined the MLE model using single-neuron recordings from a macaque monkey trained to make heading discriminations in a two-alternative forced-choice task. They measured heading discrimination for vision alone using optic flow stimuli and for vestibular signals alone using a moving platform. The critical condition was the visual-vestibular condition, where conflicting heading directions were introduced from each cue and, as predicted, mean heading direction was drawn to the more reliable component. Confirming the other key prediction of the MLE model, discrimination was better in the visual-vestibular condition (i.e., psychometric functions were steeper, illustrating reduced variance in the bimodal estimate). To bolster the evidence for MLE integration, the authors manipulated the reliability of the visual cue by adding noise to reduce its motion coherence and found that heading discrimination was drawn away from the visual direction toward the vestibular direction, in accordance with MLE predictions of a downweighted visual estimate. Their neural data, recorded while the monkeys performed the behavioral task, showed that spiking rates in single neurons from the dorsal region of the medial superior temporal area were consistent with optimal integration of visual and vestibular cues in heading discrimination.

The evidence for MLE is strong as far as integration of low-level sensory cues is concerned, although to provide an effective explanation for multisensory integration of higher order information, such as speech and semantic information, it may need to be expanded. At this level, other factors exert an influence on multisensory interactions, such as knowledge, expectations, and learning. However, as noted in Sect. 2.2, the MLE model is a simplified Bayesian model in that it does not include a prior, yet these other influences on multisensory integration can be accommodated easily within a Bayesian framework by using a prior probability distribution to account for them. The danger of this approach is that unlike applying the MLE model to low-level sensory cues, which is well constrained and can be well described by psychophysical experiments, priors can be difficult to characterize empirically and there is a risk of invoking them in a post hoc manner to account for unexpected results. Although there is no dispute at a conceptual level about priors embodying learning, experience, and knowledge in a probability distribution that could, in theory, be combined with the likelihood arising from the incoming sensory cues, quantifying and experimentally manipulating priors remains an empirical challenge. Several studies have shown evidence of Bayesian integration involving likelihoods and priors in visual-motor tasks (Kording and Wolpert 2004; Kwon and Knill 2013) and in unisensory tasks involving multiple visual cues (Knill 2007) as well as in the time domain with reproduction of temporal intervals (Jazayeri and Shadlen 2010; Cicchini et al. 2012).

2.5 Maximum Likelihood Estimation Cue Combination in the Time Domain

Multisensory studies finding evidence of MLE integration have used a variety of tasks, including spatial tasks such as judgments of size (Ernst and Banks 2002) or location (Alais and Burr 2004) and visual-motor (Kording and Wolpert 2004) and visual-vestibular (Angelaki et al. 2009) tasks. However, multisensory research assessing MLE in time perception has produced mixed results, some showing that perceptual estimates of elapsed time from a marker event do not obey MLE (Ulrich et al. 2006; Burr et al. 2009; Hartcher-O'Brien and Alais 2011), whereas another report finds that it does (Hartcher-O'Brien et al. 2014). Duration provides a curious case for two reasons. First, elapsed time is not necessarily a sensory representation and may be encoded by central accumulators at a supramodal level. Second, duration estimates cannot be made until the sensory stimulus has ceased so the perceptual judgment must be made on a stored representation, and it may be that these factors preclude MLE integration. Alternatively, there may be procedural differences between these studies that account for the discrepant findings. Studies that failed to find MLE integration in time perception defined elapsed time with brief marker stimuli that could be auditory, visual, or audiovisual at onset and offset of the temporal interval. By using empty intervals (i.e., an interval defined only by onset and offset stimuli), it is not clear whether multisensory integration is expected for the markers or for the elapsed time (which is effectively amodal). Using filled intervals overcomes this problem, and duration perception under these conditions is found to exhibit MLE integration.

In the study of duration discrimination using filled temporal intervals (Hartcher-O'Brien et al. 2014), a sequential two-interval, forced-choice procedure was used to compare a standard and a variable interval, with the intervals both defined by audio, visual, or audiovisual signals. In the audiovisual trials, audio signals with three levels of noise were combined with visual signals with a small temporal conflict to test if the duration mismatch was resolved according to MLE using unisensory signal weights. The finding was that audiovisual duration estimates did exhibit the MLE-predicted weighted average of unisensory estimates with component weights proportional to their reliabilities. This shows that MLE integration is possible on stored duration estimates and suggests that both signal durations and their associated variances needed for appropriate weighting are both available from a stored representation of elapsed time. For further evidence of Bayesian inference in duration perception, the reader is referred to Shi and Burr (2015).

2.6 Changes in Maximum Likelihood Estimation Cue Weightings Over Development

Multisensory perception is often thought to reflect the inherent dominance of one specialized modality over another. Even though recent work inspired by the MLE model shows that the precise weightings of unisensory components can vary flexibly

depending on the noise in the incoming signal (e.g., producing reverse ventriloquism when vision is degraded; Alais and Burr 2004), it remains true that in most circumstances vision will dominate for multisensory spatial tasks and audition for temporal tasks. In spatial localization, these multisensory interactions appear to be automatic. For example, when observers need only to localize the auditory component of a pair of simultaneous but spatially displaced audiovisual signals, their judgments still show a bias toward the visual location (Bertelson and Radeau 1981). Other studies too have suggested that ventriloquism occurs automatically (Vroomen et al. 2001), and the same conclusion has been drawn for spatial interactions between touch and vision (Bresciani et al. 2006; Helbig and Ernst 2008) and between touch and audition (Caclin et al. 2002; Guest et al. 2002).

As shown in Sect. 2.5, the MLE model of cue combination accounts well for how information from different senses are combined. However, it is not clear whether this is inherently the case or whether these dominances arise gradually over the span of development. The bulk of the evidence supporting MLE in multisensory integration has been done with adult subjects and does not address the developmental perspective. Sensory systems are not mature at birth but become increasingly refined during development. The brain must take these changes into account and continuously update its mapping between sensory and motor correspondence over the time course of development. This protracted process requires neural reorganization and entails cognitive changes lasting well into early adolescence (Paus 2005). Complicating the matter is that the senses develop at different rates: first touch, followed by vestibular, chemical, and auditory (all beginning to function before birth), and finally vision (Gottlieb 1990). Even though auditory development generally proceeds faster than visual development, perceptual skills within audition continue to develop at different rates, with auditory frequency discrimination (Olsho 1984; Olsho et al. 1988) and temporal discrimination (Trehub et al. 1995) all improving during infancy (Jusczyk et al. 1998). Vision in general develops later than audition, especially visual acuity and contrast sensitivity that continue to improve up until 5–6 years of age (Brown et al. 1987).

These differences in developmental sequences within and between modalities are all potential obstacles for the development of cue integration. Some multisensory processes, such as cross-modal facilitation, cross-modal transfer, and multisensory matching are present to some degree at an early age (Lewkowicz 2000; Streri 2003). Young infants can match signals between different sensory modalities (Dodd 1979; Lewkowicz and Turkewitz 1981) and detect equivalence in the amodal properties of objects across the senses (Rose 1981; Patterson and Werker 2002). For example, they can match faces with voices (Bahrick and Lickliter 2004) and visual and auditory motion signals (Lewkowicz 1992) on the basis of their synchrony. However, the varied time course of sensory development suggests that not all forms of multisensory interaction develop early. For example, multisensory facilitation during a simple audiovisual detection task does not occur until 8 years of age in most children (Barutchu et al. 2009, 2010). Few studies have investigated multisensory integration in school-age children and those that have point to unimodal dominance rather than optimal MLE integration across the senses (McGurk and Power 1980; Hatwell 1987; Misceo et al. 1999).

One study that did test for optimal integration across the senses in school-age children examined visual-haptic integration (Gori et al. 2008). This study tested visual-haptic integration in children aged 5, 6, 8, and 10 years of age and compared their size discrimination against an adult sample. In one experiment, they examined size perception in a paradigm that was essentially the same as that used by Ernst and Banks (2002). Size discrimination thresholds were measured for touch and vision separately to obtain measures of mean and variance for each modality and then the stimuli from both modalities were combined with a small-size conflict to see if the integrated estimate reflected the weights of the individual components. Their results showed that prior to 8 years of age, integration of visual and haptic spatial information was far from optimal. Indeed, directly contradicting the MLE model, in young observers (Fig. 2.4A), they observed that the sense that dominated the multisensory percept was the less precise one: the haptic modality. Haptic information was found to dominate perceived size and its discrimination threshold. Interestingly, however, the weighting of the component signals evolved progressively over development and by 8–10 years of age, visual-haptic integration became statistically optimal and followed MLE predictions, as observed in adults.

In a second analogous experiment, Gori et al. (2008) measured visual-haptic discrimination of orientation in the same age groups. This is another basic spatial task that should favor the visual modality with its specialized orientation-selective neurons in the primary visual cortex (Hubel and Wiesel 1968). Subjects discriminated which one of two bars was rotated more counterclockwise. As with the size discrimination task, thresholds were first measured separately for the visual and haptic modalities and then in a visual-haptic condition with an orientation conflict between the modalities. As with visual-haptic size judgments, the data for 8 year olds were much like the adult data and followed predictions from the MLE model based on the single-modality thresholds. Again, however, the pattern of results for the 5-year-old group was quite different; against the MLE model's predictions, orientation discrimination followed very closely the visual percept rather than incorporating haptic information (Fig. 2.4B). Although both experiments involved visual-haptic spatial tasks, the visual dominance for perceived orientation is the exact opposite to the haptic dominance observed for size discrimination.

In another study, the same group investigated audiovisual integration in both space and time perception across a developmental span covering four age ranges (5–7, 8–9, 10–11, and 13–14 years of age) and compared it to audiovisual integration in adults (Gori et al. 2012). Their goal was to examine the roles of the visual and auditory systems in the development of spatial and temporal audiovisual integration. They used similar tasks to study spatial and temporal perception in which subjects were required to bisect a temporal or a spatial interval. For the temporal bisection task, MLE integration was not observed at all in either in the adult group or any of the four children's age groups. This agrees with another study (Tomassini et al. 2011) showing that multisensory integration is suboptimal for a visual-tactile time reproduction task and with other temporal studies showing auditory dominance over vision rather than optimal integration in adults (Shams et al. 2000;

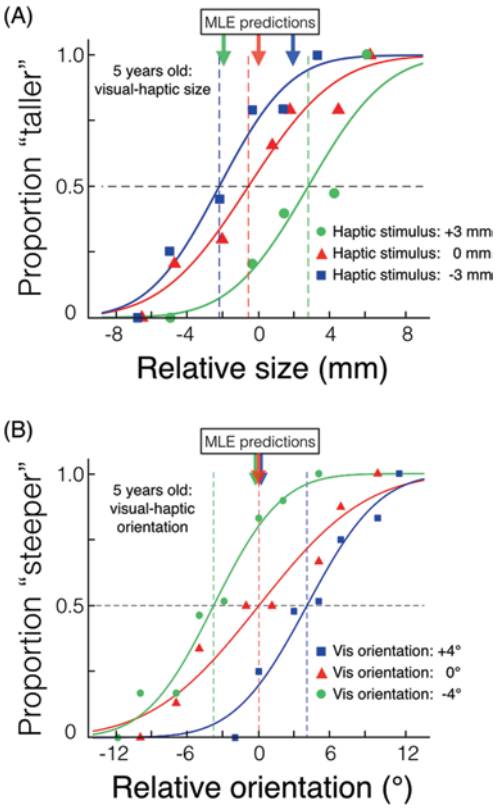

Fig. 2.4 Data showing the absence of MLE integration in children discriminating visual-haptic size and orientation where the haptic stimulus conflicted with the visual one (adapted from Gori et al. 2008). (**A**) Visual-haptic size discrimination: children did not use available visual information to optimize their discrimination and were very strongly dominated by haptic information. This is seen in the locations of the psychometric functions, which were centered at +3 mm when the haptic stimulus was 3 mm larger than vision (*right-hand function, circular data symbols*) and at −3 mm when the haptic stimulus was 3 mm smaller (*left-hand function, square data symbols*). Tellingly, the order of the psychometric functions (*squares, triangles, circles*) was the inverse of the MLE predictions (*indicated by the arrows*). (**B**) Visual-haptic orientation discrimination: children were dominated by visual information for the orientation task and did not use the available haptic information. Showing complete visual dominance, the psychometric functions shifted to +4° when the visual stimulus was 4° clockwise of the haptic stimulus and to −4° when it was 4° counterclockwise of the haptic stimulus

Burr et al. 2009). Alternatively, the lack of optimal temporal integration may have been due to the use of markers to define the start/end of the interval rather than filled intervals (discussed further in Sect. 2.7). For the spatial bisection task, MLE integration was observed only in the adult group, showing that optimal adult-like MLE integration emerges quite late in development for audiovisual temporal tasks, as it does for visual-haptic integration (Gori et al. 2008).

2.7 Cross-Modal Calibration During Development

These studies of multisensory integration over the developmental span (Gori et al. 2008, 2012) show that young children exhibit strong unisensory dominance and that the time course for the development of optimal multisensory integration is rather slow (Fig. 2.5). This is supported by other developmental studies in other sensory domains (Nardini et al. 2008, 2010). With visual-haptic stimuli, haptic information dominates size perception and vision dominates orientation perception. With audiovisual stimuli, audition dominates time perception and vision dominates space perception. The authors account for this developmental change in terms of

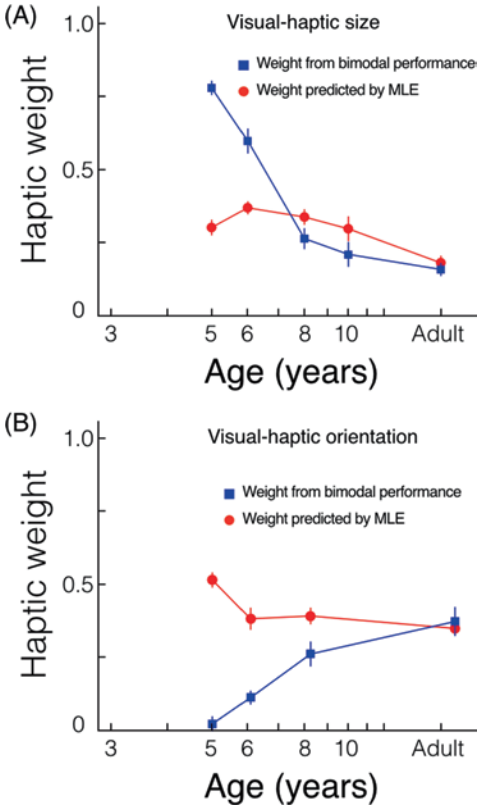

Fig. 2.5 Developmental time course of MLE integration (adapted from Gori et al. 2012). *Circular data symbols* show MLE predictions for the haptic weights when visual and haptic stimuli are combined in a bimodal stimulus. The visual weight is given by the complement (one minus the haptic weight), and both weights are predicted based on Eq. 2.3 using discrimination thresholds (σ) obtained in unimodal experiment (see Fig. 2.3A and B). *Square data symbols* show actual bimodal performance. In both visual-haptic size discrimination (**A**) and visual-haptic orientation discrimination (**B**), there is a large discrepancy at 5 and 6 years of age between bimodal performance and MLE predictions, yet both clearly converge well in adulthood. From about 10 years of age, bimodal visual-haptic performance approximates the statistically optimal MLE performance seen in adults

cross-sensory calibration. The idea is that perceptual systems must be "tuned up" and calibrated during development and that comparing signals across the senses is essential in this process and the more precise modality guides the less specialized one. While one sense is calibrating the other, the sensory signals in those two modalities cannot be usefully combined. The visual dominance for space and the auditory dominance for time could reflect the dominant modality overriding the other modality while it is still developing. This proposal is a reasonable one given that vision is fundamentally spatially specialized and audition is temporally specialized. Many studies in adults show that vision usually dominates audition when spatial locations are in conflict (Warren et al. 1981) and the greater precision of audition (Burr et al. 2009) ensures that it dominates in multisensory temporal tasks (Gebhard and Mowbray 1959; Shams et al. 2000).

Given these established modality specialities, it is reasonable that one particular modality should take the lead in calibrating and "tuning up" the other nonspecialized modalities, with vision tuning both tactile and auditory modalities for spatial tasks and audition tuning vision for temporal tasks. In agreement with this cross-modal calibration proposal, many studies in adults show that the visual system is the most influential in determining the apparent spatial position of auditory stimuli (Pick et al. 1969; Alais and Burr 2004). Only after 12 years of age does visual-auditory integration seem to occur in this spatial task, suggesting a very late development. Audiovisual space integration seems to mature later than visual-haptic spatial integration (which develops after 8–10 years of age; Gori et al. 2008) and visual-auditory temporal integration. This could be related to the time of maturation of the individual sensory systems. Indeed, previous work (Gori et al. 2008) suggested that multisensory integration occurs after the maturation of each unisensory system. The unisensory thresholds for both vision and audition continue to improve over the school years, particularly for the spatial task. For the spatial bisection task, the unisensory thresholds are still not mature at 12 years of age nor is integration optimal at this age. For the temporal task, unisensory thresholds become adult-like after 8–9 years of age, and at this age, the auditory dominance appears. Thus the delay in the development of unisensory systems seems to be related to the delay in the development of optimal sensory integration typically seen in adults.

2.8 Cross-Modal Calibration and Sensory Deficits

The hypothesis that unisensory dominance seen in the early years of development occurs while the dominant modality calibrates other modalities is a generalization of an idea originating with Berkeley's (1709/1963) proposition that touch calibrates vision. More generally, the notion is that the more robust and accurate sense for a particular perceptual task should calibrate the other. This idea raises interesting questions. In particular, what would happen to the nondominant modality if the dominant "calibrating" modality were impaired? A deficit in the more accurate calibrating sense should be detrimental to the system it calibrates. How would visual

time perception be impaired in subjects with auditory disabilities? If early unisensory dominance really occurs because cross-modal calibration of the nondominant modality has yet to occur or is incomplete, subjects with visual disabilities should show deficits in auditory spatial tasks because the calibration of space in audition by the visual system will be diminished by the visual impairment. Conversely, subjects with auditory disabilities should show temporal deficits in visual temporal tasks because of the impaired ability of audition to calibrate vision.

Gori et al. (2010, 2014) tested these predictions using stimuli and procedures similar to those used in their other multisensory studies. They established that congenitally blind subjects show severe but selective impairments in haptic discrimination tasks for orientation but not for size discrimination (Gori et al. 2010). Congenitally blind subjects also showed a severe impairment in a task requiring auditory spatial representation, namely auditory space bisection, consistent with the notion that vision is fundamental for space perception (King 2009). On the other hand, thresholds for congenitally blind subjects for simple auditory tasks such as pointing, minimal angle acuity, and temporal bisection were similar to those in control subjects. These findings illustrate the importance of visual spatial representations in establishing and calibrating auditory spatial representations. In another group, it was found that haptically impaired patients showed poor visual size discrimination but not orientation discrimination (Gori et al. 2014). An interesting observation was that in both cases the results were quite different for patients with acquired deficits rather than congenital disabilities, suggesting that cross-sensory calibration at an early age is essential. In addition, blind subjects were not uniformly bad at all auditory tasks but only in the particular spatial bisection task that was designed to tap into a sophisticated map of Euclidean relationships that would require a well-calibrated spatial sense in audition.

In other work pointing to a similar conclusion, Schorr et al. (2005) used the McGurk effect where a visual and an auditory speech signal become perceptually fused into a new phoneme to study bimodal fusion in children born deaf but whose hearing was restored by cochlear implants. Among the group who had implants at an early age (before 30 months), a similar proportion perceived the fused phoneme as normal controls, suggesting that bimodal fusion was occurring. For those who had late implants, however, only one subject showed cross-modal fusion and all the others showed visual dominance. Together, these results highlight the importance of adequate sensory input during early life for the development of multisensory interactions and show that cross-modal fusion is not innate and needs to be learned.

2.9 Summary

To perceive a coherent world, it is necessary to combine signals from the five sensory systems, signals that can be complementary or redundant. In adults, redundant signals from various sensory systems—vision, audition, and touch—are often integrated in an optimal manner following MLE integration and thus lead to an

improvement in the bimodal precision relative to the individual unimodal estimates. While much of this work was originally done in adult subjects and showed strong evidence for optimal MLE integration, more recent studies have investigated when and how optimal integration develops in children. A number of studies have shown that multisensory integration is not present at birth but develops over time and optimal integration for some tasks is not attained until about 8 years of age. One of the reasons for this may be that sensory specializations (temporal processing in audition, spatial processing in vision) need to be taught to other nonspecialized senses in a calibration process. Moreover, the continual anatomical and physiological changes occurring during development, such as growing limbs, eye length, and head circumference, mean that a recurrent updating or "recalibration" needs to take place. Until the recalibration process is complete, the two senses cannot be meaningfully combined and the default position is to rely on the specialized sense until optimal integration is possible. This calibration process may occur in different directions between senses, such as touch educating vision for size but vision educating touch for orientation, but in general, the more robust sense for a particular task calibrates the other. Once cross-modal calibration is complete, MLE integration provides an excellent model of multisensory cue combination.

Although this chapter has focused on Bayesian integration of multisensory cues, the principles are general and apply equally to combination of auditory cues. Although less research has been done on Bayesian cue combination in audition than in vision or in cross-modal contexts, a useful overview of Bayesian applications in acoustics has recently appeared (Xiang and Fackler 2015). There are many fundamental research questions remaining to be addressed in Bayesian modeling of auditory processing and psychoacoustics. Among these are, When two cues define a signal, are they combined according to the MLE model or do priors also play a role? How does the variance associated with a given cue get encoded so that cue weightings can be established? Where priors contribute to the Bayesian solution, are they stable internal models of acoustic signal statistics or are they malleable and adaptable? When fusion of two cues takes place, is access to the component cues lost, as occurs in fusion of visual cues (Hillis et al. 2002)? The Bayesian approach has been very effective in modeling visual and multisensory perception and has the potential to provide many insights into auditory perception and psychoacoustics.

Compliance with Ethics Requirements David Alais declares that he has no conflict of interest.

David Burr declares that he has no conflict of interest.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14*(3), 257–262.

Alais, D., Newell, F. N., & Mamassian, P. (2010). Multisensory processing in review: From physiology to behaviour. *Seeing and Perceiving, 23*(1), 3–38.

Angelaki, D. E., Gu, Y., & DeAngelis, G. C. (2009). Multisensory integration: Psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology, 19*(4), 452–458.

Bahrick, L. E., & Lickliter, R. (2004). Infants' perception of rhythm and tempo in unimodal and multimodal stimulation: A developmental test of the intersensory redundancy hypothesis. *Cognitive, Affective, & Behavioral Neuroscience, 4*(2), 137–147.

Barutchu, A., Crewther, D. P., & Crewther, S. G. (2009). The race that precedes coactivation: Development of multisensory facilitation in children. *Developmental Science, 12*(3), 464–473.

Barutchu, A., Danaher, J., Crewther, S. G., Innes-Brown, H., Shivdasani, M. N., & Paolini, A. G. (2010). Audiovisual integration in noise by children and adults. *Journal of Experimental Child Psychology, 105*(1–2), 38–50.

Benevento, L. A., Fallon, J., Davis, B. J., & Rezak, M. (1977). Auditory-visual interaction in single cells in the cortex of the superior temporal sulcus and the orbital frontal cortex of the macaque monkey. *Experimental Neurology, 57*(3), 849–872.

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

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics, 29*(6), 578–584.

Bresciani, J. P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision, 6*(5), 554–564.

Brown, A. M., Dobson, V., & Maier, J. (1987). Visual acuity of human infants at scotopic, mesopic and photopic luminances. *Vision Research, 27*(10), 1845–1858.

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research, 198*(1), 49–57.

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

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *Journal of Neuroscience, 32*(3), 1056–1060.

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

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology, 11*(4), 478–484.

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

Ernst, M. O., & Bulthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences, 8*(4), 162–169.

Gebhard, J. W., & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology, 72*, 521–529.

Ghahramani, Z., & Wolpert, D. M. (1997). Modular decomposition in visuomotor learning. *Nature, 386*(6623), 392–395.

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

Gori, M., Sandini, G., Martinoli, C., & Burr, D. (2010). Poor haptic orientation discrimination in nonsighted children may reflect disruption of cross-sensory calibration. *Current Biology, 20*(3), 223–225.

Gori, M., Sandini, G., & Burr, D. (2012). Development of visuo-auditory integration in space and time. *Frontiers in Integrative Neuroscience, 6*, 77.

Gori, M., Sandini, G., Martinoli, C., & Burr, D. (2014). Impairment of auditory spatial localization in congenitally blind human subjects. *Brain, 20*, 288–293.

Gottlieb, G. (1990). *Development of species identification in birds: An inquiry into the prenatal determinants of perception*. Chicago: University of Chicago Press.

Gu, Y., Angelaki, D. E., & Deangelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience, 11*(10), 1201–1210.

Guest, S., Catmur, C., Lloyd, D., & Spence, C. (2002). Audiotactile interactions in roughness perception. *Experimental Brain Research, 146*(2), 161–171.

Hartcher-O'Brien, J., & Alais, D. (2011). Temporal ventriloquism in a purely temporal context. *Journal of Experimental Psychology: Human Perception and Performance, 37*(5), 1383–1395.

Hartcher-O'Brien, J., Di Luca, M., & Ernst, M. O. (2014). The duration of uncertain times: Audiovisual information about intervals is integrated in a statistically optimal fashion. *PLoS One, 9*(3), e89339.

Hatwell, Y. (1987). Motor and cognitive functions of the hand in infancy and childhood. *International Journal of Behavioural Development, 10*, 509–526.

Helbig, H. B., & Ernst, M. O. (2008). Visual-haptic cue weighting is independent of modality-specific attention. *Journal of Vision, 8*(1), 21.1–21.16.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298*(5598), 1627–1630.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of Vision, 4*(12), 967–992.

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. New York: Wiley.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology, 195*(1), 215–243.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020–1026.

Jones, E. G., & Powell, T. P. (1970). An anatomical study of converging sensory pathways within the cerebral cortex of the monkey. *Brain, 93*(4), 793–820.

Jusczyk, P., Houston, D., & Goodman, M. (1998). Speech perception during the first year. In A. Slater (Ed.), *Perceptual development: Visual, auditory, and speech perception in infancy* (pp. 357–388). Hove: Psychology Press.

Kayser, C., Petkov, C. I., & Logothetis, N. K. (2008). Visual modulation of neurons in auditory cortex. *Cerebral Cortex, 18*(7), 1560–1574.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55*, 271–304.

King, A. J. (2009). Visual influences on auditory spatial learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*(1515), 331–339.

Knill, D. C. (2007). Learning Bayesian priors for depth perception. *Journal of Vision, 7*(8), 13.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*(12), 712–719.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research, 43*(24), 2539–2558.

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature, 427*(6971), 244–247.

Kwon, O. S., & Knill, D. C. (2013). The brain uses adaptive internal models of scene statistics for sensorimotor estimation and planning. *Proceedings of the National Academy of Sciences of the United States of America, 110*(11), E1064–E1073.

Landy, M. S., Banks, M. S., & Knill, D. C. (2011). Ideal-observer models of cue integration. In K. Tromershauser, K. Körding, & M. S. Landy (Eds.), *Book of sensory cue integration* (pp. 5–30). New York: Oxford University Press.

Lewkowicz, D. J. (1992). Infants' responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Perception & Psychophysics, 52*(5), 519–528.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: An epigenetic systems/limitations view. *Psychological Bulletin, 126*(2), 281–308.

Lewkowicz, D. J., & Turkewitz, G. (1981). Intersensory interaction in newborns: Modification of visual preferences following exposure to sound. *Child Development, 52*(3), 827–832.

McGurk, H., & Power, R. P. (1980). Intermodal coordination in young children: Vision and touch. *Developmental Psychology, 16*, 679–680.

Misceo, G. F., Hershberger, W. A., & Mancini, R. L. (1999). Haptic estimates of discordant visual-haptic size vary developmentally. *Perception & Psychophysics, 61*(4), 608–614.

Murray, M. M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., & Matusz, P. J. (2015). The multisensory function of the human primary visual cortex. *Neuropsychologia, 83*, 161–169.

Nardini, M., Jones, P., Bedford, R., & Braddick, O. (2008). Development of cue integration in human navigation. *Current Biology, 18*(9), 689–693.

Nardini, M., Bedford, R., & Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences of the United States of America, 107*(39), 17041–17046.

Olsho, L. W. (1984). Infant frequency discrimination as a function of frequency. *Infant Behavior and Development, 7*, 27–35.

Olsho, L. W., Koch, E. G., Carter, E. A., Halpin, C. F., & Spetner, N. B. (1988). Pure-tone sensitivity of human infants. *The Journal of the Acoustical Society of America, 84*(4), 1316–1324.

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*(1), 93–115.

Paus, T. (2005). Mapping brain development and aggression. *Journal of the Canadian Academy of Child and Adolescent Psychiatry, 14*(1), 10–15.

Pick, H. L., Warren, D. H., & Hay, J. (1969). Sensory conflict in judgements of spatial direction. *Perception & Psychophysics, 6*, 203–205.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: Known and unknowns. *Nature Neuroscience, 16*, 1170–1178.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89*(2), 1078–1093.

Rock, I., & Victor, J. (1964). Vision and touch: An experimentally created conflict between the two senses. *Science, 143*, 594–596.

Rose, S. A. (1981). Developmental changes in infants' retention of visual stimuli. *Child Development, 52*(1), 227–233.

Rowland, B., Stanford, T., & Stein, B. (2007). A Bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. *Experimental Brain Research, 180*(1), 153–161.

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 18748–18750.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408*(6814), 788.

Shi, Z., & Burr, D. (2015). Predictive coding of multisensory timing. *Current Opinion in Behavioral Sciences, 8*, 200–206.

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science, 145*, 1328–1330.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *Neuroreport, 12*(1), 7–10.

Streri, A. (2003). Cross-modal recognition of shape from hand to eyes in human newborns. *Somatosensory and Motor Research, 20*(1), 13–18.

Tomassini, A., Gori, M., Burr, D., Sandini, G., & Morrone, M. C. (2011). Perceived duration of visual and tactile stimuli depends on perceived speed. *Frontiers in Integrative Neuroscience, 5*, 51.

Trehub, S. E., Schneider, B. A., & Henderson, J. L. (1995). Gap detection in infants, children, and adults. *The Journal of the Acoustical Society of America, 98*(5), 2532–2541.

Ulrich, R., Nitschke, J., & Rammsayer, T. (2006). Crossmodal temporal discrimination: Assessing the predictions of a general pacemaker-counter model. *Perception & Psychophysics, 68*(7), 1140–1152.

van Beers, R. J., Sittig, A. C., & Gon, J. J. (1999). Integration of proprioceptive and visual position-information: An experimentally supported model. *Journal of Neurophysiology, 81*(3), 1355–1364.

von Helmholtz, H. (1925). *Treatise on physiological optics* (Vol. 3). New York: Dover.

Vroomen, J., Bertelson, P., & de Gelder, B. (2001). The ventriloquist effect does not depend on the direction of automatic visual attention. *Perception & Psychophysics, 63*(4), 651–659.

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. *Perception & Psychophysics, 30*, 557–564.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*(3), 638–667.

Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision, 8*(3), 24.1–24.11.

Xiang, N., & Fackler, C. (2015). Objective Bayesian analysis in acoustics. *Acoustics Today, 11*(2), 54–61.