A Stimulus-Computable Model for Audiovisual Perception and Spatial Orienting in Mammals

Cesare V. Parise, University of Liverpool

cesare.parise@liverpool.ac.uk

Abstract

Despite recent progress in multisensory research, the absence of stimulus-computable perceptual models fundamentally limits our understanding of how the brain extracts and combines task-relevant cues from the continuous flow of natural multisensory stimuli. In previous research, we demonstrated that a correlation detector initially proposed for insect motion vision can predict the temporal integration of minimalistic audiovisual signals. Here, we demonstrate how a population of such units can process natural audiovisual stimuli and accurately account for human, monkey, and rat behaviour, across simulations of 69 classic psychophysical, eye-tracking, and pharmacological experiments. Given only the raw audiovisual stimuli (i.e., real-life footage) as input, our population model could replicate observed responses with an average correlation exceeding 0.97. Despite relying on as few as 0 to 4 free parameters, our population model provides an end-to-end account of audiovisual integration in mammals—from individual pixels and audio samples to behavioural responses. Remarkably, the population response to natural audiovisual scenes generates saliency maps that predict spontaneous gaze direction, Bayesian causal inference, and a variety of previously reported multisensory illusions. This study demonstrates that the integration of audiovisual stimuli, regardless of their complexity, can be accounted for in terms of elementary joint analyses of luminance and sound level. Beyond advancing our understanding of the computational principles underlying multisensory integration in mammals, this model provides a bio-inspired, general-purpose solution for multimodal machine perception.

Introduction

Perception in natural environments is inherently multisensory. For example, during speech perception, the human brain integrates audiovisual information to enhance speech intelligibility, often beyond awareness. A compelling demonstration of this is the McGurk illusion[6], where the auditory perception of a syllable is altered by mismatched lip movements. Likewise, audiovisual integration plays a critical role in spatial localization, as illustrated by the ventriloquist illusion[8], where perceived sound location shifts toward a synchronous visual stimulus.

Extensive behavioural and neurophysiological findings demonstrate that audiovisual integration occurs when visual and auditory stimuli are presented in close spatiotemporal proximity (i.e., the spatial and temporal determinants of multisensory integration) [9,12]. When redundant multisensory information is integrated, the resulting percept is more reliable[13] and salient[16]. Various models have successfully described how audiovisual integration unfolds across time and space[3,10,17,18]–often within a Bayesian Causal Inference framework, where the system determines the probability that visual and auditory stimuli have a common cause and weigh the senses accordingly. This is the case for the detection of spatiotemporal discrepancies across the senses, or susceptibility to phenomena such as the McGurk or Ventriloquist illusions[10,17,20].

Prevailing theoretical models for multisensory integration, however, are not stimulus computable. That is, rather than directly analysing images and sounds, they operate on abstract and low-dimensional representations of the stimuli[3,10,17,18,20]. For instance, models of audiovisual time perception estimate subjective synchrony from a parameter of the stimuli (i.e., asynchrony in seconds[17,18]), not from the actual stimuli (i.e., pixels and audio samples). As a result, these models solve a different task than real observers, as the rely on explicit physical measures that are unavailable to the observers, while disregarding the information that observers actually process. Despite their obvious success in accounting for responses to simple audiovisual stimuli, these models are silent on how perceptual systems extract, process and combine task-relevant information from the continuous stream of audiovisual signals.

This omission is critical: audiovisual perception involves the continuous analysis of images and sounds, hence models that do not operate on the stimuli cannot provide a complete account of perception. Only a few models can process elementary audiovisual stimuli[22,23], none can tackle the complexity of natural audiovisual input. Currently, there are no stimulus-computable models for multisensory perception[24] that can take as input natural audiovisual data, like movies. This study addresses this gap by investigating how behaviour consistent with mammalian multisensory perception emerges from low-level analyses of natural auditory and visual signals.

Besides offering an account[25] for audiovisual integration in biological systems, an image and sound-computable model for multisensory perception would also provide a timely tool for computer vision. Animals consistently exhibit near-optimal integration of multisensory cues[13], a capability that current computer vision systems struggle to achieve when dealing with audiovisual signals[26]. Bio-inspired solutions are widely recognized as crucial to efficiently handle audiovisual data[26]. However, in the absence of stimulus-computable models of multisensory perception, inspiration from biology risks remaining superficial, driven more by intuition than by clearly defined theoretical principles. A validated, image- and sound-computable model for audiovisual perception would provide computer vision with a biologically-plausible framework for extracting and combining task-relevant cues from the continuous flow of multimodal sensory input[27].

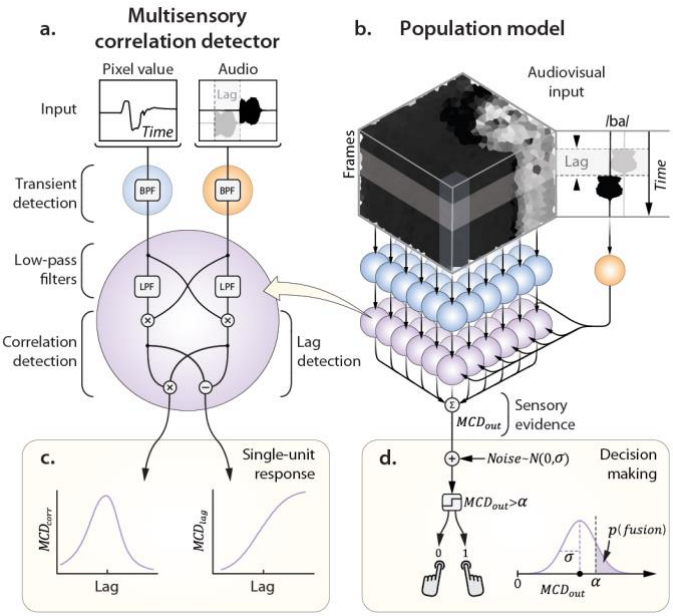


Figure 1. The MCD population model. Panel A: schematic representation of a single MCD unit. The input visual signal represents the intensity of a pixel (mouth area) over time, while the audio input is the soundtrack (the syllable /ba/). The grey soundtrack represents the experimental manipulation of AV lag, obtained by delaying one sense with respect to the other. BPF and LPF indicate band-pass and low-pass temporal filters, respectively. **Panel C** shows how single-unit responses vary as a function of crossmodal lag. **Panel B** represents the architecture of the MCD population model. Each visual unit (in blue) receives input from a single pixel, while the auditory unit receives as input the intensity envelope of the soundtrack (mono audio; see Figure 4A for a version of the model capable of receiving spatialized auditory input). Sensory evidence is then integrated over time and space for perceptual decision-making, a process in which the two model responses are weighted, summed, corrupted with additive Gaussian noise, and compared to a criterion to generate a forced-choice response (**Panel D**).

In an image- and sound-computable model, visual and auditory stimuli can be represented as patterns in a three-dimensional space, where *x* and *y* are the two spatial dimensions, and *t* the temporal dimension. An instance of such a three-dimensional diagram for the case of audiovisual speech is shown in Figure 1B (top): moving lips generate patterns of light that vary in synch with the sound. In such a representation, audiovisual correspondence can be detected by a local correlator (i.e., multiplier), that operates across space, time, and the senses[28]. In previous studies, we proposed a biologically plausible solution to detect temporal correlation across the senses (Figure 1A)[22,29-31]. Here, we will illustrate how a population of multisensory correlation detectors can take real-life footage as input and provide a comprehensive bottom-up account for multisensory integration in mammals, encompassing its temporal, spatial and attentional aspects.

The present approach posits the existence of elementary processing units, the Multisensory Correlation Detectors (MCD)[30], each integrating time-varying input from unimodal transient channels through a set of temporal filters and elementary operations (Figure 1A, see Methods). Each unit returns two outputs, representing the temporal correlation and order of incoming visual and auditory signals (Figure 1C). When arranged in a two-dimensional square lattice (Figure 1B), a population of MCD units is naturally suited to take movies (e.g., dynamic images and sounds) as input, hence capable to process any stimuli used in previous studies in audiovisual integration. Given that the aim of this study is to provide an account for multisensory integration in biological system, the benchmark of our model is to reproduce observers' behaviour in carefully-controlled psychophysical and eye-tracking experiments. Emphasis will be given to studies using natural stimuli which, despite their manifest ecological value, simply cannot be handled by alternative models. Among them, particular attention

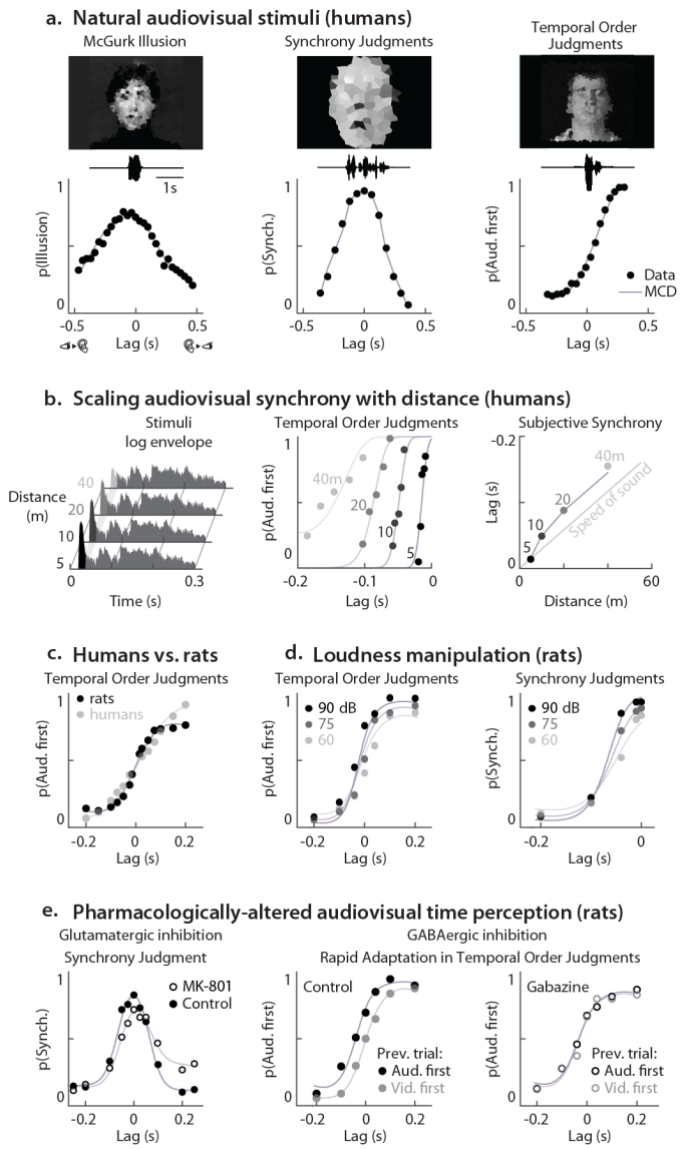


Figure 2. Natural audiovisual stimuli and psychophysical responses. Panel A stimuli (still frame and soundtrack) and psychometric functions for McGurk illusion[2], synchrony judgments[4], and temporal order judgments[5]. In all panels, dots correspond to empirical data, lines to MCD responses; negative lags represent vision first. **Panel B** left, envelopes of auditory stimuli (clicks) recorded at different distances in a reverberant environment (the Sydney Opera House)[7]. While the reverberant portion of the sound is identical across distances, the intensity of the direct sound (the onset) decreases with depth. As a result, the centre of mass of the envelopes shifts rightward with increasing distance. The central panel shows empirical and predicted psychometric functions for the various distances. The four curves were fitted using the same decision-making parameters, so that the separation between the curves results purely from the operation of the MCD. The lag at which sound and light appear synchronous (point of subjective synchrony) scales with distance at a rate approximately matching the speed of sound (right panel). The dots in the right panel display the point of subjective synchrony (estimated separately for each curve), while the jagged line is the model prediction. **Panel C**, shows temporal order judgments for clicks and flashes from both rats and human observers[11,12]. Rats outperform humans at short lag, and vice-versa. **Panel D**, rats' temporal order and synchrony judgments for flashes and clicks of varying intensity[14]. Note that the tree curves in each task were fitted using the same decision-making parameters, so that the MCD alone accounts for the separation between the curves. **Panel E**, pharmacologically-induced changes in rats' audiovisual time perception. Left: glutamatergic inhibition (MK-801 injection) leads to asymmetric broadening of the psychometric functions for simultaneity judgments[19]. Right: GABA inhibition (Gabazine injection) abolishes rapid temporal adaptation, so that psychometric curves did not change based on the lag of the previous trials (as they do in controls)[21]. All pharmacologically-induced changes in audiovisual time perception can be accounted for by changes in the decision-making process, with no need to postulate changes in low-level temporal processing.

will be dedicated to experiments involving speech, perhaps the most representative instance of audiovisual perception, and often claimed to be processed via dedicated mechanisms in the human brain.

Results

We tested the performance of our population model on three main aspects of audiovisual integration. The first concerns the temporal determinants of multisensory integration, primarily investigating how subjective audiovisual synchrony and integration depend on the physical lag across the senses. The second addresses the spatial determinants of audiovisual integration, focusing on the combination of visual and acoustic cues for spatial localization. The third one involves audiovisual attention and examines how gaze behaviour is spontaneously attracted to audiovisual stimuli even in the absence of explicit behavioural tasks. While most of the literature on audiovisual psychophysics involves human participants, in recent years monkeys and rats have also been trained to perform the same behavioural tasks. Therefore, to generalize our approach, whenever possible we simulated experiments involving all available animal models.

Temporal determinants of audiovisual integration in humans and rats. Classic experiments on the temporal determinants of audiovisual integration usually manipulate the lag between the senses and assess the perception of synchrony, temporal order, and audiovisual speech integration (as measured in humans with the McGurk illusion) through psychophysical forced-choice tasks[32,33]. Among them, we obtained both the audiovisual footage and the psychophysical data from 43 experiments in humans that used ecological audiovisual stimuli (real-life recordings of, e.g., speech and performing musicians, Figure 2A, Supplementary Figure 1 and Supplementary Table 1, for the inclusion criteria, see Methods): 27 experiments were simultaneity judgments[2,4,5,17,34-38], 10 temporal order judgments[5,39], 6 others assessed the McGurk effect[2,35,39].

For each of the experiments, we can feed the stimuli to the model (Figure 1B,D), and compare the output to the empirical psychometric functions (Equation 10, for details see Methods)[22,29-31]. Results demonstrate that a population of MCDs can broadly account for audiovisual temporal perception of ecological stimuli, and near-perfectly (rho=0.97) reproduces the empirical psychometric functions for simultaneity judgments, temporal order judgments and the McGurk effect (Figure 2A, Supplementary Figure 1). To quantify the impact of the low-level properties of the stimuli on the performance of the model, we ran a permutation test, where psychometric functions were predicted from mismatching stimuli (see Methods). The psychometric curves predicted from the matching stimuli provided significantly better fit than mismatching stimuli ($p<0.001$, see Supplementary Figure 1K). This demonstrates that our model captures the subtle effects of how individual features affect observed responses, and it highlights the role of low-level stimulus properties on multisensory integration. All analyses performed so far relied on psychometric functions averaged across observers; individual observer analyses are included in the Supplementary Information.

When estimating the perceived timing of audiovisual events, it is important to consider the different propagation speeds of light and sound, which introduce audio lags that are proportional to the observer's distance from the source (Figure 2B, right). Psychophysical temporal order judgments demonstrate that, to compensate for these lags, humans scale subjective audiovisual synchrony with distance (Figure 2B)[7]. This result has been interpreted as evidence that humans exploit auditory spatial cues, such as the direct-to-reverberant energy ratio (Figure 2B, left), to estimate the distance of the sound source and adjust subjective synchrony by scaling distance estimates by the speed of sound[7]. When presented with the same stimuli, our model also predicts the observed shifts in subjective simultaneity (Figure 2B, centre). However, rather than relying on explicit spatial representations and physics simulations, these shifts emerge from elementary audiovisual signal analyses of natural stimuli. Specifically, in reverberant environments, the intensity of the direct portion of a sound increases with source proximity, while the reverberant component remains constant. As a result, the envelopes of sounds originating close to the observers are more front-heavy than distant sounds (Figure 2B, left). These are low-level acoustic features that the lag detector of the MCD is especially sensitive to, thereby providing a computational shortcut to explicit physics simulations.

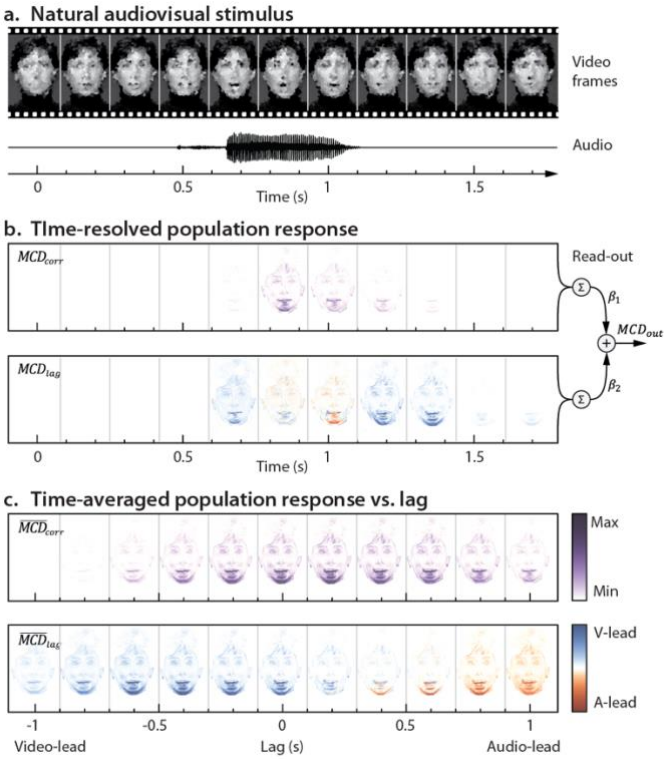


Figure 3. Ecological audiovisual stimuli and model responses. Panel A displays the frames and soundtrack of a dynamic audiovisual stimulus over time (in this example, video and audio tracks are synchronous, and the actress utters the syllable /ta/). **Panel B** shows how the dynamic population responses MCD_{corr} and MCD_{lag} vary across the frames of Panel A. Note how model responses highlight the pixels whose intensity changed with the soundtrack (i.e., the mouth area). The right side of Panel B represents the population read-out process, as implemented for the simulations in Figure 2: the population responses MCD_{corr} and MCD_{lag} are integrated over space (i.e., pixels) and time (i.e., frames), scaled and weighted by the gain parameters β_{corr}, and β_{lag} and summed to obtain a single decision variable that is fed to the decision-making stage (see Figure 1D). **Panel C** represents the time-averaged population responses $\overline{MCD_{corr}}$ and $\overline{MCD_{lag}}$ as a function of crossmodal lag (the central one corresponds to the time-averaged responses show in Panel B). Note how $\overline{MCD_{corr}}$ peaks at around zero lag and decreases with increasing lag (following the same trend shown in Figure 1C, left), while polarity of $\overline{MCD_{lag}}$ changes with the sign of the delay. The psychophysical data corresponding to the stimulus in this figure is shown in Supplementary Figure 1B. S

In recent years, audiovisual timing has been systematically studied also in rats[11,14,19,21,40,41], generally using minimalistic stimuli (such as clicks and flashes), and under a variety of manipulations of the stimuli (e.g., loudness) and pharmacological interventions (e.g., GABA and glutamatergic inhibition). Therefore, to further generalize our model to other species, we assessed whether it can also account for rats' behaviour in synchrony and temporal order judgments. Overall, we could tightly replicate rats' behaviour (rho=0.981; see Figure 2C-E, Supplementary Figure 2), including the effect of loudness on observed responses (Figure 2D). Interestingly, the unimodal temporal constants for rats were 4 times faster than for humans: such a different temporal tuning is reflected in higher sensitivity in rats for short lags (<0.1s), and in humans for longer lags (Figure 2C). This four-fold difference in temporal tuning between rats and humans closely mirrors analogous interspecies differences in physiological rhythms, such as heart rate (~4.7 times faster in rats) and breathing rate (~6.3 times faster in rats)[42]. While tuning the temporal constants of the model was necessary to account for the difference between humans and rats, this was not necessary to account for pharmacologically-induced changes in audiovisual time perception in rats (Figure 2E, Supplementary Figure 2F-G), which could all be accounted for by changes in the decision-making process (Equation 10). This latter finding suggests that these pharmacological interventions only affect perceptual decision-making, with negligible effects on low-level uni- and multisensory processing.

An asset of a low-level approach is that it allows one to inspect, at the level of individual pixels and frames, the features of the stimuli that determine the response of the model (i.e., the saliency maps). This is illustrated in Figure 3 for the case of audiovisual speech, where model responses cluster mostly around the mouth area and (to a lesser extent) the eyes (Figure 3B). These are the regions where pixels' luminance changes in synch with the audio track (Figure 3A). When the model responses for the same stimulus are averaged over time, it is possible to display how both population responses vary with audiovisual lag (Figure 3C).

Spatial determinants of audiovisual integration in humans and monkeys. Classic experiments on the spatial determinants of audiovisual integration usually require observers to localize the stimuli under systematic manipulations of the discrepancy and reliability (i.e., precision) of the spatial cues[3] (Figure 4B). This allows one to assess how unimodal cues are weighted and combined, to give rise to phenomena such as the ventriloquist illusion[8]. When the spatial discrepancy across the senses is low, observers' behaviour is well described by Maximum Likelihood Estimation (MLE)[3], where unimodal information is combined in a statistically optimal fashion, so as to maximize the precision (reliability) of the multimodal percept (see Equations 11-14, Methods).

Given that both the MLE and the MCD operate by multiplying unimodal inputs (see Methods), the time-averaged MCD population response (Equation 16) is equivalent to MLE (Figure 4A). This can be illustrated by simulating the experiment of Alais and Burr[3] using both models. In this experiment, observers had to report whether a probe audiovisual stimulus appeared left or right of a standard. To assess the weighing behaviour resulting from multisensory integration, they manipulated the spatial reliability of the visual stimuli and the disparity between the senses (Figure 4B). Figure 4C shows that the integrated percept predicted by the two models is statistically indistinguishable. As such, a population of MCDs (Equation 16) can jointly account for the observed bias and precision of the bimodal percept (Figure 4D-E), with zero parameters.

While fusing audiovisual cues is a sensible solution in the presence of minor spatial discrepancies across the senses, integration eventually breaks down with increasing disparity[43]–when the spatial (or temporal) conflict is too large, visual and auditory signals may well be unrelated. To account for the breakdown of multisensory integration in the presence of intersensory conflicts, Körding and colleagues proposed the influential Bayesian Causal Inference (BCI) model[10], where uni-

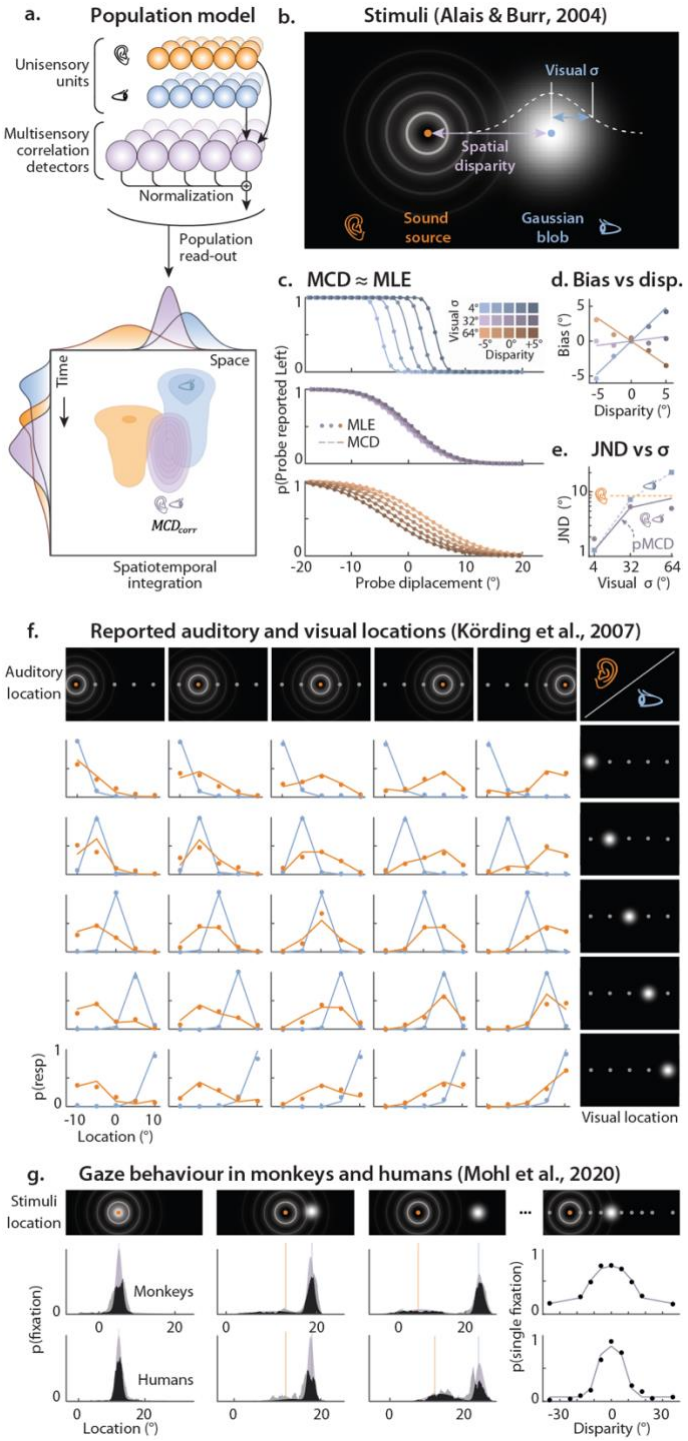


Figure 4. Audiovisual integration in space. Panel A top represents the MCD population model for spatialized audio. Visual and auditory input units receive input from corresponding spatial locations, and feed into spatially-tuned MCD units. The output of each MCD unit is eventually normalized by the total population output, so as to represent the probability distribution of stimulus location over space. The bottom part of Panel A represents the dynamic unimodal and bimodal population response over time and space (azimuth). and their marginals. When time is marginalized out, a population of MCDs implements integration as predicted by the MLE model. When space is marginalized, the output show the temporal response function of the model. In this example, visual and auditory stimuli were asynchronously presented from discrepant spatial locations (note how the blue and orange distributions are spatiotemporally offset). **Panel B** shows a schematic representation of the stimuli used to test the MLE model by Alais and Burr[3]. Stimuli were presented from different spatial positions, with a parametric manipulation of audiovisual spatial disparity and blob size (i.e., the standard deviation, σ of the blob). **Panel C** shows how the bimodal psychometric functions predicted by the MCD (lines, see Equation 16) and the MLE (dots) models fully overlap. **Panel D** shows how bimodal bias varies as a function of disparity and visual reliability (see legend on the left). The dots correspond to the empirical data from participant LM, while the lines are the predictions of the MCD model (compare to Figure 2A, of Alais and Burr[3]). **Panel E** shows how the just noticeable differences (JND, the random localization error) vary as a function of blob size. The blue squares represent the visual JNDs, the purple dots the bimodal JNDs, while the dashed orange line represents the auditory JND. The continuous line shows the JNDs predicted by the MCD population model (compare to Figure 2B, of Alais and Burr[3]). **Panel F** represents the stimuli and results of the experiment used by Körding and colleagues[10] to test the BCI model. Auditory and visual stimuli originate from one of five spatial locations, spanning a range of 20 deg. The plots show the perceived locations of visual (blue) and auditory (orange) stimuli for each combination of audiovisual spatial locations. The dots represent human data, while the lines represent the responses of the MCD population model. **Panel G** shows the stimuli and results of the experiment of Mohl and colleagues[15]. The plots on the right display the probability of a single (vs. double) fixation (top monkeys, bottom humans). The dots represent human data, while the lines represent the responses of the MCD population model. The remaining panels show the histogram of the fixated locations in bimodal trials: the jagged histograms are the empirical data, while the smooth ones are the model prediction (zero free parameters). The regions of overlap between empirical and predicted histograms are shown in black.

and bimodal location estimates are weighted based on the probability that the two modalities share a common cause (Equation 17). The BCI model was originally tested in an experiment in which sound and light were simultaneously presented from one of five random locations, and observers had to report the position of both modalities[10] (Figure 4F). Results demonstrate that visual and auditory stimuli preferentially bias each other when the discrepancy is low, with the bias progressively declining as the discrepancy increases.

Also a population of MCDs can compute the probability that auditory and visual stimuli share a common cause (Figure 1B, D; Equation 20), therefore we can test whether it also can implement BCI. For that, we simulated the experiment of Körding

and colleagues, and fed the stimuli to a population of MCDs (Equations 18-20) which near-perfectly replicated the empirical data (rho=0.99)–even slightly outperforming the BCI model (rho=0.97), while relying on fewer free parameters.

To test for the generalizability of these findings to different species and behavioural paradigms, we simulated an analogous experiment[15], where monkeys (Macaca Mulatta) and humans were instructed to direct their gaze towards spatially scattered audiovisual stimuli (Figure 4G). If observers perceive the stimuli as sharing a common cause, they should make a single fixation, otherwise two: one per modality. Results demonstrate that, in both humans and monkeys, the probability of a single fixation decreases with increasing disparity (Figure 4G, right). This pattern is fully captured by a population of MCDs, which could fit the probability of a single fixation (Equation 21) and from that closely predict gaze directions (Equation 20) in both species with zero free parameters (Figure 4G, see Methods).

Taken together, these simulations show that a behaviour consistent with BCI and MLE naturally emerges from a population of MCDs. Unlike BCI and MLE, however, a population of MCDs is both image- and sound-computable and it also makes explicit the temporal dynamic of the integration process (Figure 4A, bottom, Figure 3B and Figure 5B-C). On one hand, this extends the application of the MCD to complex, dynamic audiovisual stimuli (such as real-life videos, which cannot be handled by MLE or BCI); on the other, it allows for direct comparison of model response with time-varying neurophysiological measures[29].

Attention and audiovisual gaze behaviour. Multisensory stimuli are typically salient, and a vast body of literature demonstrates that spatial attention is commonly attracted to audiovisual stimuli[16]. This fundamental aspect of multisensory perception is naturally captured by a population of MCDs, whose dynamic response explicitly represents the regions in space

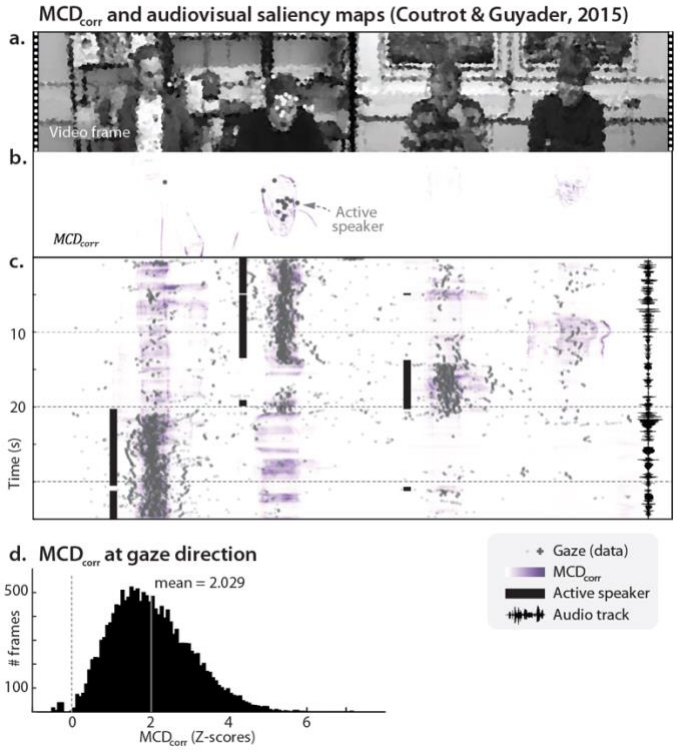


Figure 5. Audiovisual saliency maps. Panel A represents a still frame of Coutrot and Guyader's[1] stimuli. The white dots represent gaze direction of the various observers. **Panel B** represents the MCD population response for the frame in Panel A. The dots represent observed gaze direction (and correspond to the white dots of Panel A). **Panel C** represents how the MCD response varies over time and azimuth (with elevation marginalized-out). The black solid lines represent the active speaker, while the waveform on the right displays the soundtrack. Note how the MCD response was higher for the active speaker. **Panel D** shows the distribution of model response at gaze direction (see Panel B) across all frames and observers in the database. Model response was normalized for each frame (Z-scores). The y axis represents the number of frames. The vertical grey line represents the mean.

with the highest audiovisual correspondence for each point in time. Therefore, for a population of MCDs to provide a plausible account for audiovisual integration, such dynamic saliency maps should be able to predict human audiovisual gaze behaviour, in a purely bottom-up fashion and with no free parameters. Figure 5A shows the stimuli and eye-tracking data

from the experiment of Coutrot and Guyader[1], in which observers passively observed a video of four persons talking. Panel B shows the same eye-tracking data plotted over the corresponding MCD population response: across 20 observers, and 15 videos (for a total of over 16,000 frames), gaze was on average directed towards the locations (i.e., pixels) yielding the top 2% MCD response (Figure 5D, Equations 22-23). The tight correspondence of predicted and empirical salience is illustrated in Figure 5C: note how population responses peak based on the active speaker.

Discussion

This study demonstrates that elementary audiovisual analyses are sufficient to replicate behaviours consistent with multisensory perception in mammals. The proposed image- and sound-computable model, composed of a population of biologically-plausible elementary processing units, offers the first end-to-end account of multisensory perception. Starting exclusively from the raw stimuli (i.e., pixels and audio samples), our model quantitatively reproduced observed behaviour, including multisensory illusions, attention maps, and Bayesian Causal Inference with an average Pearson correlation above 0.97—as tested in a large-scale simulation of 69 classic audiovisual experiments with 7 different behavioural tasks, and involving 534 human observers, 110 rats, and 2 monkeys.

This novel framework has clear advantages over traditional, non-stimulus-computable perceptual models. By directly handling realistic audiovisual stimuli, our population model represents sensory input in a format that closely mirrors what animals encounter in the real world, rather than relying on abstracted or simplified representations that fail to capture the complexity of natural sensory inputs. This makes stimulus-computable models more consistent with how the brain represents and processes bimodal spatiotemporal information[24]. Moreover, by directly operating on the raw signals, our population model tailors its predictions to the exact stimuli and tasks used in each experiment, thereby capturing subtle effects of how individual features affect observed responses (see Supplementary Figure 1K).

The present approach naturally lends itself to be generalized and tested against a broad range of tasks, stimuli, and responses—as reflected by the unparalleled breadth of the experiments simulated here. Among the perceptual effects emerging from elementary signal processing, one notable example is the scaling of subjective audiovisual synchrony with sound source distance[7]. As sound travels slower than light, humans compensate for audio delays by adjusting subjective synchrony based on the source's distance scaled by the speed of sound. Although this phenomenon seems to rely on explicit physics modelling, our simulations demonstrate that auditory cues embedded in the envelope (Fig 2B, left) are sufficient to scale subjective audiovisual synchrony. In a similar fashion, our simulations show that phenomena such as the McGurk illusion, the subjective timing of natural audiovisual stimuli, and saliency detection, may emerge from elementary operations performed at pixel level, bypassing the need for more sophisticated analyses such as image segmentation, lip or face-tracking, 3D reconstruction, etc. Elementary, general-purpose operations on natural stimuli can drive complex behaviour, sometimes even in the absence of advanced perceptual and cognitive contributions. Indeed, it is intriguing that a population of MCDs, a computational architecture originally developed for motion vision in insects, can predict speech illusions in humans.

The fact that identical low-level analyses can account for all of the 69 experiments simulated here, directly addresses several open questions in multisensory perception. For instance, psychometric functions for speech and non-speech stimuli often differ significantly[44]. This has been interpreted as evidence that speech may be special and processed via dedicated

mechanisms[45]. However, identical low-level analyses are sufficient to account for all observed responses, regardless of the stimulus type (Figure 2, Supplementary Figures 1-2). This suggests that the differences in psychometric curves across classes of stimuli (e.g., speech vs. non-speech vs. clicks-&-flashes) are due to the low-level features of the stimuli themselves, not how the brain processes them. Similarly, experience and expertise also modulate multisensory perception. For example, audiovisual simultaneity judgments differ significantly between musicians and non-musicians[4] (see Supplementary Figure 1C). Likewise, the McGurk illusion[39] and subjective audiovisual timing[46] vary over the lifespan in humans, and following pharmacological interventions in rats[19,21] (see Supplementary Figure 1E,J and Supplementary 2F-G). Our simulations show that parametric adjustments in the decision-making process are sufficient to account for all these effects, without the need to assume structural or parametric differences in low-level perceptual analyses across observers or conditions.

Although the same model explains responses to multisensory stimuli in humans, rats, and monkeys, the temporal constants vary across species. For example, the model for rats is tuned to temporal frequencies over four times higher than those for humans. This not only explains the differential sensitivity of humans and rats to long and short audiovisual lags, but it also mirrors analogous interspecies differences in physiological rhythms, such as heart and breathing rates[42]. Previous research has shown that physiological arousal modulates perceptual rhythms within individuals[47]. It is an open question whether the same association between multisensory temporal tuning and physiological rhythms persists in other mammalian systems. Conversely, no major differences in the model's spatial tuning were found between humans and macaques, possibly reflecting the close phylogenetic link between the two species.

A compelling result from our simulations is the predictability of gaze direction for naturalistic audiovisual stimuli. Saliency, the property by which some elements in a display stand out and attract observer's attention and gaze direction, is a popular concept in both cognitive and computer sciences[48]. In computer vision, saliency models are usually complex and rely on advanced signal processing and semantic knowledge—typically with tens of millions of parameters[49]. Despite successfully predicting gaze behaviour, current audiovisual saliency models are often computationally expensive, and the resulting maps are hard to interpret and inevitably affected by the datasets used for training[50]. In contrast, our model detects saliency "out of the box", without any free parameters, and operating purely at the individual pixel level. The elementary nature of the operations performed by a population of MCDs returns saliency maps that are easy to interpret: salient points are those with high audiovisual correlation. By grounding multisensory integration and saliency detection in biologically plausible computations, our study offers a new tool for machine perception and robotics to handle multimodal inputs in a more human-like way, while also improving system accountability.

This framework also provides a solution for self-supervised and unsupervised audiovisual learning in multimodal machine perception. A key challenge when handling raw audiovisual data is solving the causal inference problem—determining whether signals from different modalities are causally related or not[10]. Models in machine perception often depend on large, labelled datasets for training. In this context, a biomimetic module that handles saliency maps, audiovisual correspondence detection, and multimodal fusion can drive self-supervised learning through simulated observers, thereby reducing the dependency on labelled data[26,27,51]. What is more, the simplicity of our population model offers a computationally lightweight solution for systems that require rapid sensory integration, such as real-time audiovisual processing, robotics, and AR/VR technologies.

Although a population of MCDs can explain when phenomena such as the McGurk Illusion occur, it does not explain the process of phoneme categorization that ultimately determines what syllable is perceived[20]. More generally, it is well known that cognitive and affective factors modulate our responses to multisensory stimuli[9]. While a purely low-level model does not directly address these issues, the modularity of our approach makes it possible to extend the system to include high-level perceptual, cognitive and affective factors. What is more, although this study focused on audiovisual integration in mammals, the same approach can be naturally extended to other instances of sensory integration (e.g., visuo- and audio-tactile) and animal classes.

Besides simulating behavioural responses, a stimulus-computable approach necessarily makes explicit all the intermediate steps of sensory information processing. This opens the system to inspection at all of its levels, thereby allowing for direct comparisons with neurophysiology[29]. In insect motion vision, this transparency made it possible for the Hassenstein-Reichardt detector to act as a searchlight to link computation, behavior, and physiology at the scale of individual cells[52], ultimately inspiring key advances in motion detection algorithms for modern computer vision. Being based on formally identical computational principles[15], the present approach holds the same potential for multisensory perception.

Methods

The MCD population model

The architecture of each MCD unit used here is the same as described in Parise and Ernst[30], however, units here receive time-varying visual and auditory input from spatiotopic receptive fields. The input stimuli (s) consist of luminance level and sound amplitude varying over space and time and are denoted as $s_m(x, y, t)$ – with x and y, representing the spatial coordinates along the horizontal and vertical axes, t is the temporal coordinate, and m is the modality (video and audio). When the input stimulus is a movie with mono audio, the visual input to each unit is a signal representing the luminance of a single pixel over time, while the auditory input is the amplitude envelope (later we will consider more complex scenarios where auditory stimuli are also spatialized).

Each unit operates independently and detects changes in unimodal signals over time by temporal filtering based on two biphasic impulse response functions that are 90 deg out of phase (i.e., a quadrature pair). A physiologically plausible implementation of this process has been proposed by Adelson and Bergen [53] and consists of linear filters of the form:

$$f_n(t) = \left(\frac{t}{\tau_{bp}}\right)^n \cdot e^{-\frac{t}{\tau_{bp}}} \cdot \left[\frac{1}{n!} - \frac{1}{(n+2)!} \cdot \left(\frac{t}{\tau_{bp}}\right)^2\right] \qquad \text{Equation 1}$$

The phase of the filter is determined by n, which based on Emerson et al.[54] takes the values of 6 for the fast filter, and 9 for the slow one. The temporal constant of the filters is determined by the parameter τ_{bp}; in humans, its best fitting value is 0.045 s for vision and 0.0367 s for audition. In rats, the fitted temporal constant for vision and audition are nearly identical and their value is 0.010 s.

Fast and slow filters are applied to each unimodal input signal and the two resulting signals are squared and then summed. After that, a compressive non-linearity (square-root) is applied to the output, so as to constrain it within a reasonable range[53]. Therefore, the output of each unimodal unit feeding into the correlation detector takes the following form

$$MCD_{mod}(x, y, t) = \sqrt{[s_m(x, y, t) * f_6(t)]^2 + [s_m(x, y, t) * f_9(t)]^2}, \qquad \text{Equation 2}$$

where $mod = vid, aud$ represents the sensory modality and * is the convolution operator.

As in the original version[22], each MCD consists of two sub-unit, in which the unimodal input is low-pass filtered and multiplied as follows

$$u_1(x, y, t) = MCD_{vid}(x, y, t) \cdot \left[MCD_{aud}(x, y, t) * f_{lp}(t)\right] \qquad \text{Equation 3}$$

$$u_2(x, y, t) = MCD_{aud}(x, y, t) \cdot \left[MCD_{vid}(x, y, t) * f_{lp}(t)\right] \qquad \text{Equation 4}$$

The impulse response of the low-pass filter of each sub-unit takes the form

$$f_{lp}(t) = \frac{t}{\tau_{lp}} \cdot e^{-\frac{t}{\tau_{lp}}}, \qquad \text{Equation 5}$$

where τ_{lp} represents the temporal constant, and its estimated value is 0.180 s for humans and 0.138 s for rats.

The response of the sub-units is eventually multiplied to obtain MCD_{corr}, which represents the local spatiotemporal audiovisual correlation, and subtracted to obtain MCD_{lag} which describes the relative temporal order of vision and audition

$$MCD_{corr}(x, y, t) = u_1(x, y, t) \cdot u_2(x, y, t) \qquad \text{Equation 6}$$

$$MCD_{lag}(x, y, t) = u_1(x, y, t) - u_2(x, y, t) \qquad \text{Equation 7}$$

The outputs $MCD_{corr}(x, y, t)$ and $MCD_{lag}(x, y, t)$ are the final product of the MCD and represent the outcome of early uni- and multisensory processing.

The **temporal constants** of the filters were fitted using the Bayesian Adaptive Direct Search (BADS) algorithm[55], set to maximize the correlation between the empirical and predicted psychometric functions of the temporal determinants of multisensory integration. For humans, that included all studies in Supplementary Figure 1, besides Patient PH; For rats, it included all non-pharmacological studies in Supplementary Figure 2. To minimize the effect of starting parameters, the fitting was performed 200 times using random starting values (from 0.001 to 1.5 s). The parameters estimated using BADS were further refined using the *fminsearch* algorithm in Matlab. Parameter estimation was considerably compute-intensive, hence the amount of data had to be reduced by rescaling the videos to 15% of the original size (without affecting the frame rate). Besides reducing run-time, this simulates the (Gaussian) spatial pooling occurring in early visual pathways.

Now that the MCD population is defined and its parameters are fully constrained, what remains to be explained is how to read out, from the dynamic population responses, the relevant information that is needed to generate a behavioral response, such as eye-movements, button-presses, nose- or lick-contacts etc. While the early sensory processing of each MCD unit is task-independent and operates in a purely bottom-up fashion, the exact nature of the read-out and decision-making process depends on the behavioral task, which ultimately determines how to weigh and combine the dynamic population responses. Given that in the present study we consider different types of behavioral experiments (investigating audiovisual integration trough temporal tasks, spatial tasks, and passive observation), the read-out process for each task will be described in separate sections.

Modeling the temporal determinants of multisensory integration

The experiments on the temporal constraints of audiovisual integration considered here, rely on psychophysical forced-choice tasks to assess the effects of crossmodal lags on perceived synchrony, temporal order, and the McGurk illusion. In humans, such experiments entail pressing one of two buttons; in rats, nose-poking or licking one of two spouts. In the case simultaneity judgments, on each trial observers reported whether visual and auditory stimuli appeared synchronous or not ($resp = \{yes, no\}$). In temporal order judgments, observers reported which modality came first ($resp =$

$\{vision\ first, audition\ first\}$). Finally, in the case of the McGurk, observers (humans) had to report what syllable they heard (e.g., "da", "ga" or "ba", usually recoded as "fused" and "non-fused percept"). When plotted against lag, the resulting empirical psychometric functions describe how audiovisual timing affect perceived synchrony, temporal order, and the McGurk illusion. Here, we model these tasks using the same read-out and decision-making process.

To account for observed responses, the dynamic, high-bandwidth populations responses must be transformed (compressed) into a single number, representing response probability for each lag. In line with standard procedures[56], this was achieved by integrating $MCD_{corr}(x, y, t)$ and $MCD_{lag}(x, y, t)$ over time and space, so as to obtain two summary decision variables

$$\overline{MCD_{corr}} = \sum_x \sum_y \sum_t MCD_{corr}(x, y, t) \qquad \text{Equation 8}$$

and

$$\overline{MCD_{lag}} = \sum_x \sum_y \sum_t MCD_{lag}(x, y, t) \qquad \text{Equation 9}$$

The temporal window for these analyses consisted of the duration of each stimulus, plus two seconds before and after (during which the audio was silent, and the video displayed a still frame); the exact extension of the temporal window used for the analyses had minimal effect on the results. $\overline{MCD_{corr}}$ and $\overline{MCD_{lag}}$ are eventually linearly weighted and transformed into response probabilities through a cumulative normal function as follows

$$p_{MCD}(resp) = \Phi(\beta_0 + \beta_{corr} \cdot \overline{MCD_{corr}} + \beta_{lag} \cdot \overline{MCD_{lag}}) \qquad \text{Equation 10}$$

Here, Φ represents the cumulative normal, β_{corr} and β_{lag} are linear coefficients that weigh and scale $\overline{MCD_{corr}}$ and $\overline{MCD_{lag}}$. β_0 is a bias term, which corresponds to the response criterion. Finally, $p_{MCD}(resp)$ is the probability of a response, which is $p_{MCD}(Synchronous)$ for the simultaneity judgment task, $p_{MCD}(Audio\ first)$ for the temporal order judgment task, and $p_{MCD}(Fusion)$ for the McGurk task. Note how Equation 9 entails that simultaneity judgments, temporal order judgments, and the McGurk illusions are all simulated using the very same model architecture.

With the population model fully constrained (see above), the only free parameters in the present simulations are the ones controlling the decision-making process: β_{crit}, β_{corr}, and β_{lag}. These were separately fitted for each experiment using *fitglm* Matlab (binomial distribution and probit link function). For the simulations of the effects of distance on audiovisual temporal order judgments in humans[7] (Figure 2B), and the manipulation of loudness in rats[14] (Figure 2D) β_{crit}, β_{corr}, and β_{lag} were constant across conditions (i.e., distance[7] or loudness[14]). This way, differences in the psychometric functions as a function of distance[7] or loudness[14] of the stimuli are fully explained by the MCD. Overall, the model provided an excellent fit to the empirical psychometric functions, and the average Pearson correlation between human and model response (weighted by sample size) is 0.981 for humans and 0.994 for rats. Naturally, model-data correlation varied across experiments, largely due to sample size. This can be appreciated when the MCD-data correlation for each experiment is plotted against the number of trials for each lag in a funnel plot (see Supplementary Figure 1I). The number of trials for each lag determine the binomial error for data points in Supplementary Figure 1 and 2; accordingly, the funnel plot shows that lower MCD-data correlation

is more commonly observed for curves based on smaller sample size. This shows that the upper limit in the MCD-data correlation is mostly constrained by the reliability of the dataset, rather than systematic errors in the model.

To assess the contribution of the low-level properties of the stimuli on model's performance, we ran a permutation test, where psychometric curves were generated (Equation 10) using stimuli from different experiments (but with the same manipulation of lag). If the low-level properties of the stimuli play a significant role, the correlation between data and model with permuted stimuli should be lower than with non-permuted stimuli. For this permutation we used the data from simultaneity judgment tasks on humans (Supplementary Figure 1), as with 27 individual experiments there are enough permutations to render the test meaningful. For that, we used the temporal constants of the MCD fitted before, so that each psychometric curve each permutation had 3 free parameters (β_{crit}, β_{corr}, and β_{lag}). The results from 200k permutations demonstrate that the goodness of fit obtained with the original stimuli is superior to that of permuted stimuli. Specifically, the permuted distribution of the mean Pearson correlation of predicted vs. empirical psychometric curves had a mean of 0.972 (σ =0.0036), while such a correlation rose to 0.989 when the MCD received the original stimuli.

Modeling the spatial determinants of multisensory integration

Most studies on audiovisual space perception only investigated the horizontal spatial dimension, hence the stimuli can be reduced to $s_m(x,t)$ instead of $s_m(x,y,t)$, as in the simulations above (see Equation 2). Additionally, based on preliminary observations, the output MCD_{lag} does not seem necessary to account for audiovisual integration in space, hence only the population response $MCD_{corr}(x,t)$ (see Equation 6) will be considered here.

MCD and MLE – simulation of Alais and Burr (2004). In its general form, the MLE model can be expressed probabilistically as $p_{MLE}(x) \propto p_{vid}(x) \cdot p_{aud}(x)$, where $p_{vid}(x)$ and $p_{aud}(x)$ represent the probability distribution of the unimodal location estimate (i.e., likelihood functions), and $p_{MLE}(x)$ is the bimodal distribution. When the $p_{vid}(x)$ and $p_{aud}(x)$ follow a Gaussian distribution, also $p_{MLE}(x)$ is Gaussian, with the variance (σ^2_{MLE}) equal to the product divided by the sum of the unimodal variances:

$$\sigma^2_{MLE} = \frac{\sigma^2_{vid} \cdot \sigma^2_{aud}}{\sigma^2_{vid} + \sigma^2_{aud}} \qquad \text{Equation 11}$$

and the mean consisting of a weighted average of the unimodal means

$$\mu_{MLE} = \omega_{vid} \cdot \mu_{vid} + \omega_{aud} \cdot \mu_{aud} \qquad \text{Equation 12}$$

where

$$\omega_{vid} = \frac{1/\sigma^2_{vid}}{1/\sigma^2_{vid} + 1/\sigma^2_{aud}} \quad \text{and} \qquad \text{Equation 13}$$

$$\omega_{aud} = 1 - \omega_{vid} \qquad \text{Equation 14}$$

To test whether a population of MCDs can replicate the study of Alais and Burr (2004), this simulation has two main goals. The first one is to compare the predictions of the MCD with that of the MLE model; the second one is to test whether the MCD model can predict observers' responses.

To simulate the study of Alais and Burr (2004), we first need to generate the stimuli. The visual stimuli consisted of a 1-D Gaussian luminance profiles, presented for 10ms. Their standard deviations, which determined visual spatial reliability, were defined as the standard deviation of the visual psychometric functions. Likewise, the auditory stimuli also consisted of a 1-D Gaussian sound intensity profile (with a standard deviation determined by the unimodal auditory psychometric function). Note that the spatial reliability in the MCD model jointly depends on the stimulus and the receptive field of the input units; however, teasing apart the differential effects induced by these two sources of spatial uncertainty is beyond the scope of the present study. Hence, for simplicity, here injected all spatial uncertainty into the stimulus (see also[22,29-31]). For this simulation, we only used the data from observer LM of Alais and Burr's[3], as it is the only participant for which the full dataset is publicly available (other observers, however, had similar results).

The stimuli are fed to the model to obtain the population response $MCD_{corr}(x,t)$ (see Equation 6), which is marginalized over time as follows

$$\overline{MCD_{corr}}(x) = \sum_t MCD_{corr}(x,t) \qquad \text{Equation 15}$$

This provides a distribution of the population response over the horizontal spatial dimension. Finally, a divisive normalization is performed to transform model response into a probability distribution

$$p_{MCD}(x) = \frac{\overline{MCD_{MCD}}(x)}{\sum_x \overline{MCD_{MCD}}(x)} \qquad \text{Equation 16}$$

It is important to note that Equations 15 and 16 have no free parameters, and all simulations are now performed with a fully constrained model. To test whether the MCD model can perform audiovisual integration according to the MLE model, we replicated the various conditions run by observer LM, and calculated the bimodal likelihoods distribution predicted by the MLE and MCD models (i.e., $p_{MLE}(x)$ and $p_{MCD}(x)$). These were statistically identical, when considering rounding errors. The results of these simulations are plotted in Figure 4C, and displayed as cumulative distributions as in Figure 1 of Alais and Burr[3].

Once demonstrated that $p_{MLE}(x) = p_{MCD}(x)$, it is clear that the MCD is equally capable of predicting observed responses. However, for completeness, we compared the prediction of the MCD to the empirical data: the results demonstrate that, just like the MLE model, a population of MCDs can predict both audiovisual bias (Figure 4D) and just noticeable differences (JND, Figure 4E).

MCD and BCI – simulation of Körding et a. (2007). To account for the spatial breakdown of multisensory integration, the BCI model operates in a hierarchical fashion: first it estimates the probability that audiovisual stimuli share a common cause ($p(C=1)$). Next, the model weigh and integrates the unimodal and bimodal information ($p_{mod}(x)$ and $p_{mle}(x)$) as follows:

$$p_{BCI,mod}(x) = p_{BCI}(C=1) \cdot p_{MLE}(x) + [1 - p_{BCI}(C=1)] \cdot p_{mod}(x) \qquad \text{Equation 17}$$

All the terms of Equation 17 have a homologous representation in an MCD population response: $p_{mle}(x)$ corresponds to $p_{MCD}(x)$ (see previous section). Likewise, the homologous of $p_{mod}(x)$ can be obtained by marginalizing over time and normalizing the output of the unimodal units as follows (see Equation 15)

$$\overline{MCD_{mod}}(x) = \frac{\sum_t MCD_{mod}(x,t)}{\sum_x \sum_t MCD_{mod}(x,t)} \qquad \text{Equation 18}$$

Finally, the MCD homologous of BCI's $p(C=1)$ can be read-out from the population response following the same logic as Equations 8 and 9:

$$p_{MCD}(C=1) = \Phi(\beta_0 + \beta_{corr} \log_{10} \sum_x \sum_t MCD_{corr}(x,t)) \qquad \text{Equation 19}$$

Here, we found that including a compressive non-linearity (logarithm of the total MCD_{corr} response) provided a tighter fit to the empirical data. With $p_{MCD}(C=1)$ representing the probability that vision and audition share a common cause, and $p_{MCD}(x)$ and $\overline{MCD_{mod}}(x)$ representing the bimodal and unimodal population responses, the MCD model can simulate observed responses as follows:

$$p_{MCD,mod}(x) = p_{MCD}(C=1) \cdot p_{MCD}(x) + [1 - p_{MCD}(C=1)] \cdot \overline{MCD_{mod}}(x) \qquad \text{Equation 20}$$

The similarity between Equation 20 and Equation 17 demonstrates the fundamental homology of the BCI and MCD models: what remains to be tested is whether the MCD can also account for the results of Körding and colleagues[10]. For that, just like the BCI model, also the MCD model relies on 4 free parameters: two are shared by both models, and represent the spatial uncertainty (i.e., the variance) of the unimodal input (i.e., σ_{vid}^2 and σ_{aud}^2). Additionally, the MCD model then needs two linear coefficients (slope β_{corr} and intercept β_0) to transform the dynamic population response into a probability of a common cause. Conversely, the remaining two parameters of the BCI model correspond to a prior for common cause and another for central location, neither of which are necessary to account for observed responses in the present framework. Parameters were fitted using BADS[55] set to maximize the Pearson correlation between model and human responses. Overall, the MCD provided an excellent fit to the empirical data (r=0.99), even slightly exceeding the performance of the BCI model (while relying on the same degrees of freedom). Given that the fitted value of the slope parameter (β_{corr}) approached 1, we repeated the fitting while removing β_{corr} from Equation 19: even with just 3 free parameters (one fewer than the BCI model), the MCD outperformed the BCI model and the correlation with the empirical data approached 0.99.

MCD and BCI – simulation of Mohl et al. (2020). Mohl and colleagues used eye movements to test whether humans and monkeys integrate audiovisual spatial cues according to BCI. The targets consisted of either unimodal or bimodal stimuli. Following the same logic as the simulation of Alais and Burr[3] (see above), the unimodal input consisted of impulses with a Gaussian spatial profile (Figure 4A-B), whose the variance (i.e., σ_{vid}^2 and σ_{aud}^2) was set equal to the variance of the fixations measured in the unimodal trials (averaged across observers). Although the probability of a single fixation decreases with increasing disparity (Figure 4G, right), observers sometimes failed to make a second fixation even when the stimuli were

noticeably far apart (lapses). This was especially true for monkeys, which had a lapse rate of 16%, indicative of low attention and compliance. To account for this, we can modify Equation 19 and obtain the probability of a single fixation as follows:

$$p_{MCD}(C=1) = p_{lapse} + \left(1 - 2p_{lapse}\right) \cdot \Phi(\beta_0 + \beta_{corr} \log_{10} \sum_x \sum_t MCD_{corr}(x,t)) \qquad \text{Equation 21.}$$

Here, p_{lapse} is a free parameter that represents the probability of making the incorrect number of fixations, irrespective of the discrepancy. As in Equation 19, β_{corr} and β_0 are free parameters that transform the dynamic population response into a probability of a common cause (i.e., single fixation). Equation 21 could tightly reproduce the observed probability of a single fixation (Figure 4G, right), and the Pearson correlation between the model and data was 0.995 for monkeys and 0.988 for humans.

With the parameters p_{lapse}, β_{corr} and β_0 fitted to the probability of a single fixation (i.e., the probability of a common cause), it is now possible to predict gaze direction (i.e., the perceived location of the stimuli) with zero free parameters. For that, we can use Equation 21 to get the probability of a common cause and predict gaze direction using Equation 20. The distribution of fixations predicted by the MCD closely follow the empirical histogram in both species (Figure 4G) and the correlation between the model and data was 0.9 for monkeys and 0.93 for humans. Note that figure 4G shows only a subset of the 20 conditions tested in the experiment (the same subset of conditions shown in the original paper).

MCD and audiovisual gaze behavior

To test whether the dynamic MCD population response can account for gaze behavior during passive observation of audiovisual footage, a simple solution is to measure whether observers preferentially looked at the MCD_{corr} responses is maximal. Such an analysis was separately performed for each frame. For that, gaze directions were first low pass filtered with a Gaussian kernel ($\sigma = 14\ pixels$) and normalized to probabilities $p_{gaze}(x,y)$. Next, we calculated the average MCD responses at gaze direction for each frame (t); this was done by weighing MCD_{corr} by p_{gaze} as follows

$$\overline{MCD_{gaze}}(t) = \sum_x \sum_y p_{gaze}(x,y,t) \cdot MCD_{corr}(x,y,t) \qquad \text{Equation 22}$$

To assess the MCD response at gaze ($\overline{MCD_{gaze}}(t)$) is larger than the frame average, we calculated the standardized mean difference (SMD) for each frame as follows

$$SMD(t) = \frac{\overline{MCD_{gaze}}(t) - \mu[MCD_{corr}(x,y,t)]}{\sigma[MCD_{corr}(x,y,t)]} \qquad \text{Equation 23}$$

Across the over 16,000 frames of the available dataset, the average SMD was 2.03 (Figure 5D). Given that the standardized mean difference serves as a metric for effect size, and that effect sizes surpassing 1.2 are deemed very large [57], it is remarkable that the MCD population model can so tightly account for human gaze behavior in a purely bottom-up fashion and without free parameters.

Code availability statement

A MATLAB script running the MCD population model will be included as Supplementary Material.

Pre-processing of ecological audiovisual footage

The diversity of the stimuli used in this dataset requires some preprocessing before the stimuli can be fed to the population model. First, all movies were converted to grayscale (scaled between 0 and 1) and the soundtrack was converted to rms envelope (scaled between 0 and 1), thereby removing chromatic and tonal information. Movies were then padded with two seconds of frozen frames at onset and offset to accommodate for the manipulation of the lag. Finally, the luminance of the first still frame was set as baseline and subtracted from all subsequent frames. Along with the padding, this helps minimizing transient artefacts induced by the onset of the video.

Video frames were to 15% of the original size, and the static background was cropped. On a practical side, this made the simulations much faster (which is crucial for parameter estimation); on a theoretical side, such a down sampling simulates the Gaussian spatial pooling of luminance across the visual field (unfortunately, the present datasets do not provide sufficient information to convert pixels into visual angles). In a similar fashion, we down sampled sound envelope to match the frame-rate of the video.

Datasets

To thoroughly compare observed and model behavior, this study requires a large and diverse dataset consisting of both the raw stimuli and observers' responses. For that, we adopted a convenience sampling and simulated the studies for which both stimuli and responses were available (either in public repositories, shared by the authors, or extracted from published figures). The inclusion criteria depend on what aspect of multisensory integration is being investigated, and they are described below.

For the **temporal determinants of multisensory integration in humans**, we only included studies that: (1) used real-life audiovisual footage, (2) performed a parametric manipulation of lag, and (3) engaged observers in a forced-choice behavioral task. Forty-three individual experiments met the inclusion criteria (Figure 2A-B and Supplementary Figure 1). These varied in terms of stimuli, observers, and tasks. In terms of stimuli, the dataset consists of responses from 105 unique real-life videos (see Supplementary Figure 1 and Supplementary Table 1). The majority of the videos represented audiovisual speech (possibly the most common stimulus in audiovisual research), but they varied in terms of content (i.e., syllables, words, full sentences, etc.), intelligibility (i.e., sine-wave speech, amplitude-modulated noise, blurred visuals, etc.), composition (i.e., full face, mouth-only, oval frame, etc.), speaker identity, etc. The remaining non-speech stimuli consist of footage of actors playing a piano or a flute. The study from Alais and Carlile[7] was included in the dataset because, even if the visual stimuli were minimalistic (blobs), the auditory stimuli consisted of ecological auditory depth cues depth cues recorded in a real reverberant environment (the Sydney Opera House, see Supplementary Information for details on the dataset). The dataset contains forced-choice responses to three different tasks: speech categorization (i.e., for the McGurk illusion), simultaneity judgments, and temporal order judgment. In terms of observers, besides the general population, the dataset consists of experimental groups varying in terms of age, musical expertise, and even includes a patient, PH, who reports hearing speech

before seeing mouth movements after a lesion in the pons and basal ganglia[39]. Taken together the dataset consists of ~1k individual psychometric functions, from 454 unique observers, for a total of ~300k trials; the psychometric curves for each experiment are shown in Figure 2A-B and Supplementary Figure 1. All these simulations are based on psychometric functions averaged across observers, for simulations of individual observers see Supplementary Information and Supplementary Figures 3-5.

For the **temporal determinants of multisensory integration in rats**, we included studies that performed a parametric manipulation of lag and engaged rats in simultaneity and temporal order judgment tasks. Sixteen individual experiments[11,14,19,21,40,41] met the inclusion criteria (Figure 2C-D and Supplementary Figure 2 and Supplementary Table 2): all of them used minimalistic audiovisual stimuli (clicks and flashes) and with a parametric manipulation of audiovisual lag. Overall, the dataset consists of ~190 individual psychometric functions, from 110 rats (and 10 humans), for a total of ~300k trials.

For the case of the **spatial determinants of multisensory integration**, to the best of our knowledge there are no available datasets with both stimuli and psychophysical responses. Fortunately, however, the spatial aspects of multisensory integration are often studied with minimalistic audiovisual stimuli (e.g., clicks and blobs), which can be simulated exactly. Audiovisual integration in space is commonly framed in terms of optimal statistical estimation, where the bimodal percept is modelled either through maximum likelihood estimation (MLE), or Bayesian causal inference (BCI). To provide a plausible account for audiovisual integration in space, a population of MCDs should also behave as a Bayesian-optimal estimator. This hypothesis was tested by comparing the population response to human data in the studies that originally tested the MLE and BCI models; hence we simulated the study of Alais and Burr[3] and Körding et al.[10]. Such simulations allow us to compare the data with our model, and our model with previous ones (MLE and BCI). Given that in these two experiments a population of MCDs behaves just like the MLE and BCI models (and with the same number of free parameters or less), the current approach can be easily extended to other instances of sensory cue integration previously modelled in terms of optimal statistical estimation. This was tested by simulating the study of Mohl and colleagues[15], who used eye movements to assess whether BCI can account for audiovisual integration in monkeys and humans.

Finally, we tested whether a population of MCDs can predict **audiovisual attention and gaze behaviour** during passive observation of ecological audiovisual stimuli. Coutrot and Guyader[1] run the ideal testbed for this hypothesis: much like our previous simulations (Figure 2A), they employed audiovisual speech stimuli, recorded indoor, with no camera shake. Specifically, they tracked eye movements from 20 observers who passively watched 15 videos of a lab meeting (see Figure 5A). Without fitting parameters, the population response tightly matched the empirical saliency maps (see Figure 5B-D).

References

1 Coutrot, A. & Guyader, N. in *2015 23rd European Signal Processing Conference (EUSIPCO)*. 1531-1535 (IEEE).
2 Van Wassenhove, V., Grant, K. W. & Poeppel, D. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* **45**, 598-607 (2007).
3 Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* **14**, 257-262 (2004).
4 Lee, H. & Noppeney, U. Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences* **108**, E1441-E1450 (2011).
5 Vroomen, J. & Stekelenburg, J. J. Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition* **118**, 75-83 (2011).
6 McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264** 746-748 (1976).
7 Alais, D. & Carlile, S. Synchronizing to real events: Subjective audiovisual alignment scales with perceived auditory depth and speed of sound. *Proceedings of the National Academy of Sciences* **102**, 2244-2247 (2005).
8 Stratton, G. Vision without inversion of the retinal image. *The Psychological Review* **4**, 463-481 (1897).
9 Stein, B. E. (MIT Press, Cambridge, MA, 2012).
10 Körding, K. P. *et al.* Causal inference in multisensory perception. *PLoS One* **2**, 943 (2007).
11 Mafi, F. *et al.* Temporal order judgment of multisensory stimuli in rat and human. *Frontiers in Behavioral Neuroscience* **16**, 1070452 (2023).
12 Stein, B. E. & Stanford, T. R. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* **9**, 255-266 (2008).
13 Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429-433 (2002).
14 Schormans, A. L. & Allman, B. L. Behavioral plasticity of audiovisual perception: Rapid recalibration of temporal sensitivity but not perceptual binding following adult-onset hearing loss. *Frontiers in Behavioral Neuroscience* **12**, 256 (2018).
15 Mohl, J. T., Pearson, J. M. & Groh, J. M. Monkeys and humans implement causal inference to simultaneously localize auditory and visual stimuli. *Journal of Neurophysiology* **124**, 715-727 (2020).
16 Talsma, D., Senkowski, D., Soto-Faraco, S. & Woldorff, M. G. The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences* **14**, 400-410 (2010).
17 Magnotti, J. F., Ma, W. J. & Beauchamp, M. S. Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology* **4**, 798 (2013).
18 Yarrow, K., Solomon, J. A., Arnold, D. H. & Roseboom, W. The best fitting of three contemporary observer models reveals how participants' strategy influences the window of subjective synchrony. *Journal of Experimental Psychology: Human Perception and Performance* (2023).
19 Al-Youzbaki, M. U., Schormans, A. L. & Allman, B. L. Past and present experience shifts audiovisual temporal perception in rats. *Frontiers in Behavioral Neuroscience* **17** (2023).
20 Magnotti, J. F. & Beauchamp, M. S. A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computational Biology* **13**, e1005229 (2017).
21 Schormans, A. L. & Allman, B. L. An imbalance of excitation and inhibition in the multisensory cortex impairs the temporal acuity of audiovisual processing and perception. *Cerebral Cortex* **33**, 9937-9953 (2023).
22 Parise, C. V. & Ernst, M. O. Correlation detection as a general mechanism for multisensory integration. *Nature Communications* **7:11543**, 1-9 (2016).
23 Cuppini, C., Shams, L., Magosso, E. & Ursino, M. A biologically inspired neurocomputational model for audiovisual integration and causal inference. *European Journal of Neuroscience* **46**, 2481-2498 (2017).
24 Burge, J. Image-computable ideal observers for tasks with natural stimuli. *Annual Review of Vision Science* **6**, 491-517 (2020).
25 Marr, D. *Vision: A computational investigation into the human representation and processing of visual information*. (MIT press, 1982).
26 Shahabaz, A. & Sarkar, S. Increasing Importance of Joint Analysis of Audio and Video in Computer Vision: A Survey. *IEEE Access* (2024).
27 Ngiam, J. *et al.* Multimodal deep learning. *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689-696 (2011).
28 Parise, C. V., Spence, C. & Ernst, M. O. When correlation implies causation in multisensory integration. *Current Biology* **22**, 46-49 (2012).
29 Pesnot Lerousseau, J., Parise, C. V., Ernst, M. O. & van Wassenhove, V. Multisensory correlation computations in the human brain identified by a time-resolved encoding model. *Nature Communications* **13**, 2489 (2022).
30 Parise, C. V. & Ernst, M. O. Multisensory integration operates on correlated input from unimodal transients channels. *eLife*, **RP90841** (2023).
31 Horsfall, R., Wuerger, S. & Meyer, G. Visual intensity-dependent response latencies predict perceived audio–visual simultaneity. *Journal of Mathematical Psychology* **100**, 102471 (2021).
32 Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E. & Hickok, G. Timing in audiovisual speech perception: A mini review and new psychophysical data. *Attention, Perception, & Psychophysics* **78**, 583-601 (2016).
33 Vroomen, J. & Keetels, M. Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics* **72**, 871-884 (2010).
34 Roseboom, W. & Arnold, D. H. Twice upon a time: multiple concurrent temporal recalibrations of audiovisual speech. *Psychological Science* **22**, 872-877 (2011).
35 Yuan, X., Bi, C., Yin, H., Li, B. & Huang, X. The recalibration patterns of perceptual synchrony and multisensory integration after exposure to asynchronous speech. *Neuroscience Letters* **569**, 148-152 (2014).
36 Ikeda, T. & Morishita, M. How Are Audiovisual Simultaneity Judgments Affected by Multisensory Complexity and Speech Specificity? *Multisensory Research* **34**, 49-68 (2020).
37 Van Laarhoven, T., Stekelenburg, J. J. & Vroomen, J. Increased sub-clinical levels of autistic traits are associated with reduced multisensory integration of audiovisual speech. *Scientific Reports* **9**, 9535 (2019).
38 Lee, H. & Noppeney, U. Temporal prediction errors in visual and auditory cortices. *Current Biology* **24**, R309-R310 (2014).
39 Freeman, E. D. *et al.* Sight and sound out of synch: Fragmentation and renormalisation of audiovisual integration and subjective timing. *cortex* **49**, 2875-2887 (2013).
40 Schormans, A. L. *et al.* Audiovisual temporal processing and synchrony perception in the rat. *Frontiers in Behavioral Neuroscience* **10**, 246 (2017).
41 Paulcan, S., Giersch, A., van Wassenhove, V. & Doyère, V. Temporal order processing in rats depends on the training protocol. *Journal of Experimental Psychology: Animal Learning and Cognition* **49**, 31 (2023).
42 Agoston, D. V. How to translate time? The temporal aspect of human and rodent biology. *Frontiers in neurology* **8**, 92 (2017).
43 Chen, L. & Vroomen, J. Intersensory binding across space and time: a tutorial review. *Attention, Perception, & Psychophysics* **75**, 790-811 (2013).
44 Vatakis, A., Ghazanfar, A. A. & Spence, C. Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision* **8**, 1-11 (2008).

45 Tuomainen, J., Andersen, T. S., Tiippana, K. & Sams, M. Audio–visual speech perception is special. *Cognition* **96**, B13-B22 (2005).
46 Petrini, K. *et al.* Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Experimental Brain Research* **198**, 339-352 (2009).
47 Legrand, F. D. *et al.* Brief aerobic exercise immediately enhances visual attentional control and perceptual speed. Testing the mediating role of feelings of energy. *Acta Psychologica* **191**, 25-31 (2018).
48 Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 1254-1259 (1998).
49 Chen, C., Song, M., Song, W., Guo, L. & Jian, M. A comprehensive survey on video saliency detection with auditory information: the audio-visual consistency perceptual is the key! *IEEE Transactions on Circuits and Systems for Video Technology* **33**, 457-477 (2023).
50 Adebayo, J. *et al.* Sanity checks for saliency maps. *Advances in Neural Information Processing Systems* **31** (2023).
51 Arandjelovic, R. & Zisserman, A. Look, listen and learn. *Proceedings of the IEEE International Conference on Computer Vision*, 609-617 (2017).
52 Serbe, E., Meier, M., Leonhardt, A. & Borst, A. Comprehensive characterization of the major presynaptic elements to the Drosophila OFF motion detector. *Neuron* **89**, 829-841 (2016).
53 Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America* **2**, 284-299 (1985).
54 Emerson, R. C., Bergen, J. R. & Adelson, E. H. Directionally selective complex cells and the computation of motion energy in cat visual cortex. *Vision Research* **32**, 203-218 (1992).
55 Acerbi, L. & Ma, W. J. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems* **30** (2017).
56 Parise, C. V. & Ernst, M. O. Noise, multisensory integration, and previous response in perceptual disambiguation. *PLoS Computational Biology* **13**, e1005546 (2017).
57 Sawilowsky, S. S. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* **8**, 26 (2009).

A Stimulus-Computable Model for Audiovisual Perception and Spatial Orienting in Mammals

Cesare V. Parise, University of Liverpool

cesare.parise@liverpool.ac.uk

Supplementary Information

Supplementary Materials and Methods

Supplementary Figures

Supplementary Tables

Supplementary Information

Pre-processing of ecological audiovisual footage

The diversity of the stimuli used in this dataset requires some preprocessing before the stimuli can be fed to the population model. First, all movies were converted to grayscale (scaled between 0 and 1) and the soundtrack was converted to rms envelope (scaled between 0 and 1), thereby removing chromatic and tonal information. Movies were then padded with two seconds of frozen frames at onset and offset to accommodate for the manipulation of the lag. Finally, the luminance of the first still frame was set as baseline and subtracted from all subsequent frames. Along with the padding, this helps minimizing transient artefacts induced by the onset of the video.

Video frames were to 15% of the original size, and the static background was cropped. On a practical side, this made the simulations much faster (which is crucial for parameter estimation); on a theoretical side, such a down sampling simulates the Gaussian spatial pooling of luminance across the visual field (unfortunately, the present datasets do not provide sufficient information to convert pixels into visual angles). In a similar fashion, we down sampled sound envelope to match the frame-rate of the video.

Datasets

To thoroughly compare observed and model behavior, this study requires a large and diverse dataset consisting of both the raw stimuli and observers' responses. For that, we adopted a convenience sampling and simulated the studies for which both stimuli and responses were available (either in public repositories, shared by the authors, or extracted from published figures). The inclusion criteria depend on what aspect of multisensory integration is being investigated, and they are described below.

For the **temporal determinants of multisensory integration in humans**, we only included studies that: (1) used real-life audiovisual footage, (2) performed a parametric manipulation of lag, and (3) engaged observers in a forced-choice behavioral task. Forty-three individual experiments met the inclusion criteria (Figure 2A-B and Supplementary Figure 1). These varied in terms of stimuli, observers, and tasks. In terms of stimuli, the dataset consists of responses from 105 unique real-life videos (see Supplementary Figure 1 and Supplementary Table 1). The majority of the videos represented audiovisual speech (possibly the most common stimulus in audiovisual research), but they varied in terms of content (i.e., syllables, words, full sentences, etc.), intelligibility (i.e., sine-wave speech, amplitude-modulated noise, blurred visuals, etc.), composition (i.e., full face, mouth-only, oval frame, etc.), speaker identity, etc. The remaining non-speech stimuli consist of footage of actors playing a piano or a flute. The study from Alais and Carlile(*33*) was included in the dataset because, even if the visual stimuli were minimalistic (blobs), the auditory stimuli consisted of ecological auditory depth cues depth cues recorded in a real reverberant environment (the Sydney Opera House, see Supplementary Information for details on the dataset). The dataset contains forced-choice responses to three different tasks: speech categorization (i.e., for the McGurk illusion), simultaneity judgments, and temporal order judgment. In terms of observers, besides the general population, the dataset consists of experimental groups varying in terms of age, musical expertise, and even includes a patient, PH, who reports hearing speech before seeing mouth movements after a lesion in the pons and basal ganglia(*32*). Taken together the dataset consists of ~1k individual psychometric functions, from 454 unique observers, for a total of ~300k trials; the psychometric curves for each

experiment are shown in Figure 2A-B and Supplementary Figure 1. All these simulations are based on psychometric functions averaged across observers, for simulations of individual observers see Supplementary Information and Supplementary Figures 3-5.

For the **temporal determinants of multisensory integration in rats**, we included studies that performed a parametric manipulation of lag and engaged rats in simultaneity and temporal order judgment tasks. Sixteen individual experiments(*34-39*) met the inclusion criteria (Figure 2C-D and Supplementary Figure 2 and Supplementary Table 2): all of them used minimalistic audiovisual stimuli (clicks and flashes) and with a parametric manipulation of audiovisual lag. Overall, the dataset consists of ~190 individual psychometric functions, from 110 rats (and 10 humans), for a total of ~300k trials.

For the case of the **spatial determinants of multisensory integration**, to the best of our knowledge there are no available datasets with both stimuli and psychophysical responses. Fortunately, however, the spatial aspects of multisensory integration are often studied with minimalistic audiovisual stimuli (e.g., clicks and blobs), which can be simulated exactly. Audiovisual integration in space is commonly framed in terms of optimal statistical estimation, where the bimodal percept is modelled either through maximum likelihood estimation (MLE), or Bayesian causal inference (BCI). To provide a plausible account for audiovisual integration in space, a population of MCDs should also behave as a Bayesian-optimal estimator. This hypothesis was tested by comparing the population response to human data in the studies that originally tested the MLE and BCI models; hence we simulated the study of Alais and Burr(*11*) and Körding et al.(*8*). Such simulations allow us to compare the data with our model, and our model with previous ones (MLE and BCI). Given that in these two experiments a population of MCDs behaves just like the MLE and BCI models (and with the same number of free parameters or less), the current approach can be easily extended to other instances of sensory cue integration previously modelled in terms of optimal statistical estimation. This was tested by simulating the study of Mohl and colleagues(*42*), who used eye movements to assess whether BCI can account for audiovisual integration in monkeys and humans.

Finally, we tested whether a population of MCDs can predict **audiovisual attention and gaze behaviour** during passive observation of ecological audiovisual stimuli. Coutrot and Guyader(*43*) run the ideal testbed for this hypothesis: much like our previous simulations (Figure 2A), they employed audiovisual speech stimuli, recorded indoor, with no camera shake. Specifically, they tracked eye movements from 20 observers who passively watched 15 videos of a lab meeting (see Figure 5A). Without fitting parameters, the population response tightly matched the empirical saliency maps (see Figure 5B-D).

Individual observers' analyses. Most of the simulations described so far rely on group-level data, where psychometric curves represent the average response across the pool of observers that took part in each experiment. Individual psychometric functions, however, sometimes vary dramatically across observers, hence one might wonder whether the MCD, besides predicting stimulus-driven variations in the psychometric functions, can also capture individual differences. A recent study by Yarrow and colleagues[1] directly addressed this question, and concluded that models of the Independent Channels family outperform the MCD at fitting responses individual difference.

Although it can be easily shown that such a conclusion was supported by an incomplete implementation of the MCD (which did not include the MCD_{lag} output), a closer look at the two models against the same datasets help us illustrate their fundamental difference and highlight a key drawback of perceptual models that take parameters as input. Therefore, we first simulated the impulse stimuli used by Yarrow and colleagues[1], fed them to the MCD and used Equation 10 to generate the individual psychometric curves. Given that their stimuli consisted of temporal impulses with no spatiotemporal manipulation, a single MCD unit is sufficient to run these simulations. Overall, the model provided an excellent fit to the original data and tightly captured individual differences: the average Pearson correlation between predicted and empirical psychometric functions across the 57 curves shown in Figure S2 is 0.98 (see Supplementary Figure S2). Importantly, for such simulations the MCD was fully constrained, and the only free parameters (3 in total) were the linear coefficients of Equation 10, which describe how the output of the MCD is used for perceptual decision-making. For comparison, also Independent Channels models achieved analogous goodness of fit, but they required at least 5 free parameters (depending on the exact implementation, see[1]).

To assess the generalizability of this finding, we additionally simulated the individual psychometric functions from the experiments that informed the architecture of the MCD units used here[2]. Specifically, Parise and Ernst[2] run two psychophysical studies using minimalistic stimuli that only varied over time. In the first one, auditory and visual stimuli consisted of step increments and/or decrements in intensity. Audiovisual lag was parametrically manipulated using the method of constant stimuli, and observers were required to perform both simultaneity and temporal order judgments. Following the logic described above, we fed the stimuli to the model, and used Equation 10 (with 3 free parameters) to simulate human responses. Results demonstrate that the MCD can account for individual differences regardless of the task: the average Pearson correlation between empirical and predicted psychometric curves was 0.97 for the simultaneity judgments, and 0.96 for the temporal order judgments (64 individual psychometric curves from 8 observers, for a total of 9600 trials). This generalizes the results of the previous simulation to a different type of stimuli (steps vs. impulses) and extends them to include a task, the temporal order judgment, which was not considered by Yarrow and colleagues[1] (whose model can only perform simultaneity judgments).

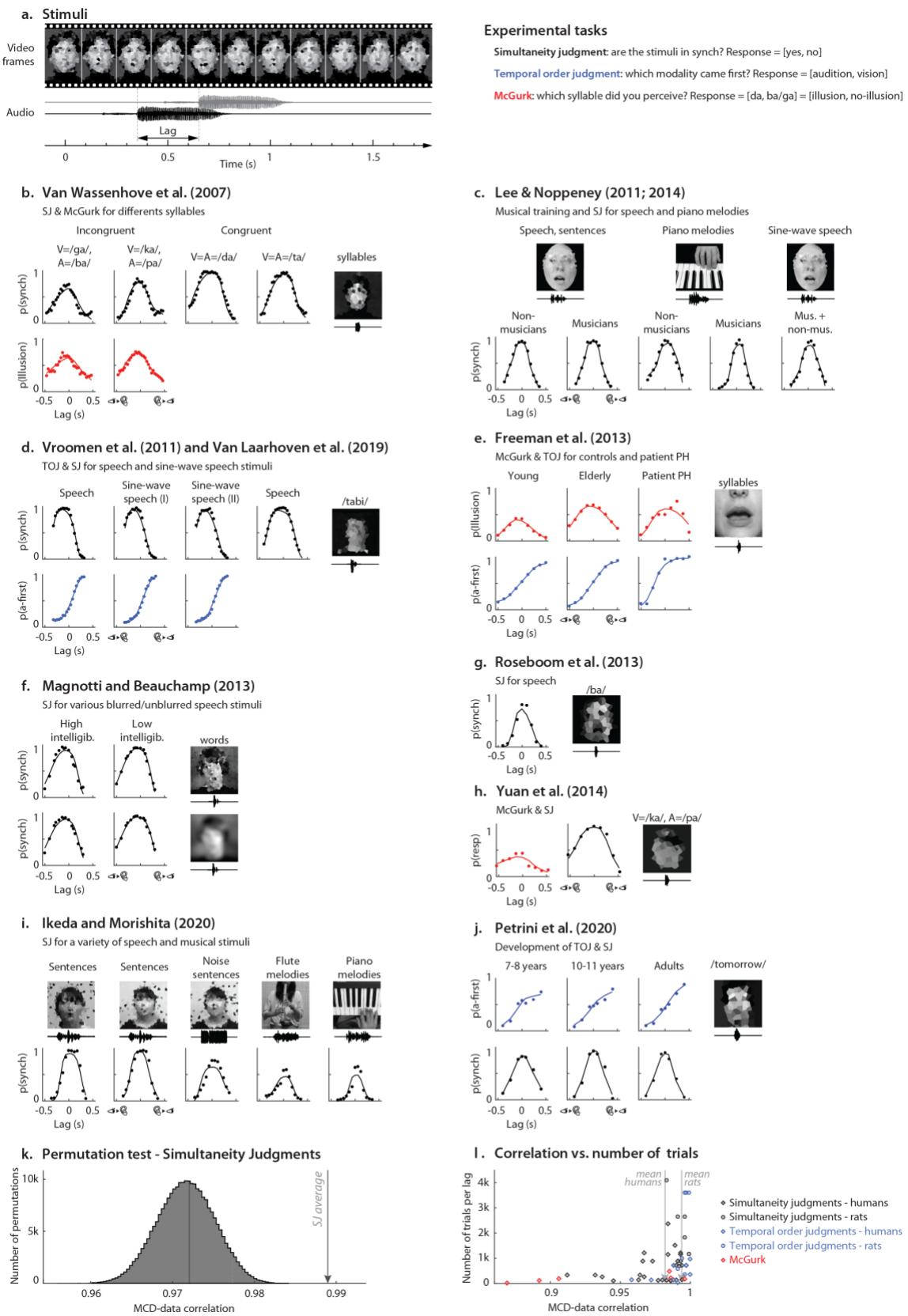
The second study of Parise and Ernst[2] consists of simultaneity judgments for periodic audiovisual stimuli defined by a square-wave intensity envelope (Figure S4A). Simultaneity judgments for this type of periodic stimuli are also periodic, with two complete oscillations in perceived simultaneity for each cycle of phase shifts between the senses (a phenomenon known as frequency doubling). Once again, using Equation 10 the MCD could account for individual differences in observed behavior (5 psychometric curves from 5 observers, for a total of 3000 trials) with an average Pearson correlation of 0.93, while relying on just 3 free parameters.

It is important to note that for these simulations, the same MCD model accurately predicted (in a purely bottom-up fashion) bell-shaped SJ curves for non-periodic stimuli, and sinusoidal curves for periodic stimuli. Alternative models of audiovisual simultaneity that directly take lag as input, enforce bell-shaped psychometric functions, where perceived synchrony monotonically decrease as we move away from the point of subjective simultaneity. As a result, in the absence of ad-hoc adjustments they all necessarily fail at replicating the results of Parise and Ernst[2], due to their inability to generate periodic psychometric functions. Conversely, the MCD is agnostic regarding the shape of the psychometric functions, hence the very

same model used to predict the simultaneity judgments of Yarrow and colleagues[1] can also predict the empirical psychometric functions of Parise and Ernst[2], including individual differences across observers (all while relying on just three free parameters).

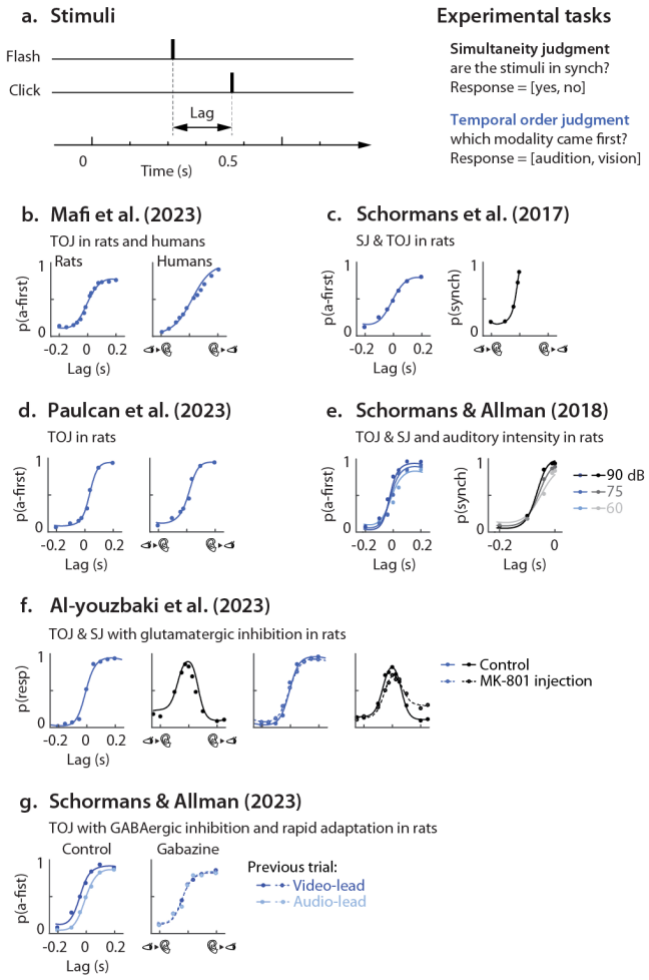
Simulation of Alais and Carlile (2005) For the simulations of Alais and Carlile[29], the envelope of the auditory stimuli was extracted from the waveforms shown in the figures of the original publication. That was done using WebPlotDigitizer to trace the profile of the waveforms; the digitized points were then interpolated and resampled at 1000Hz. To preserve the manipulation of the direct-to-reverberant waves, the section of the envelope with the reverberant signal was identical across the four conditions (i.e., distances), so that what varied across conditions was the initial portion of the signals (the direct waves). For the simulations of Alais and Carlile[29], all four psychometric functions were fitted simultaneously, so that the four psychometric functions all relied on just three free parameters: the ones related to the decision-making process .

Supplementary Figures

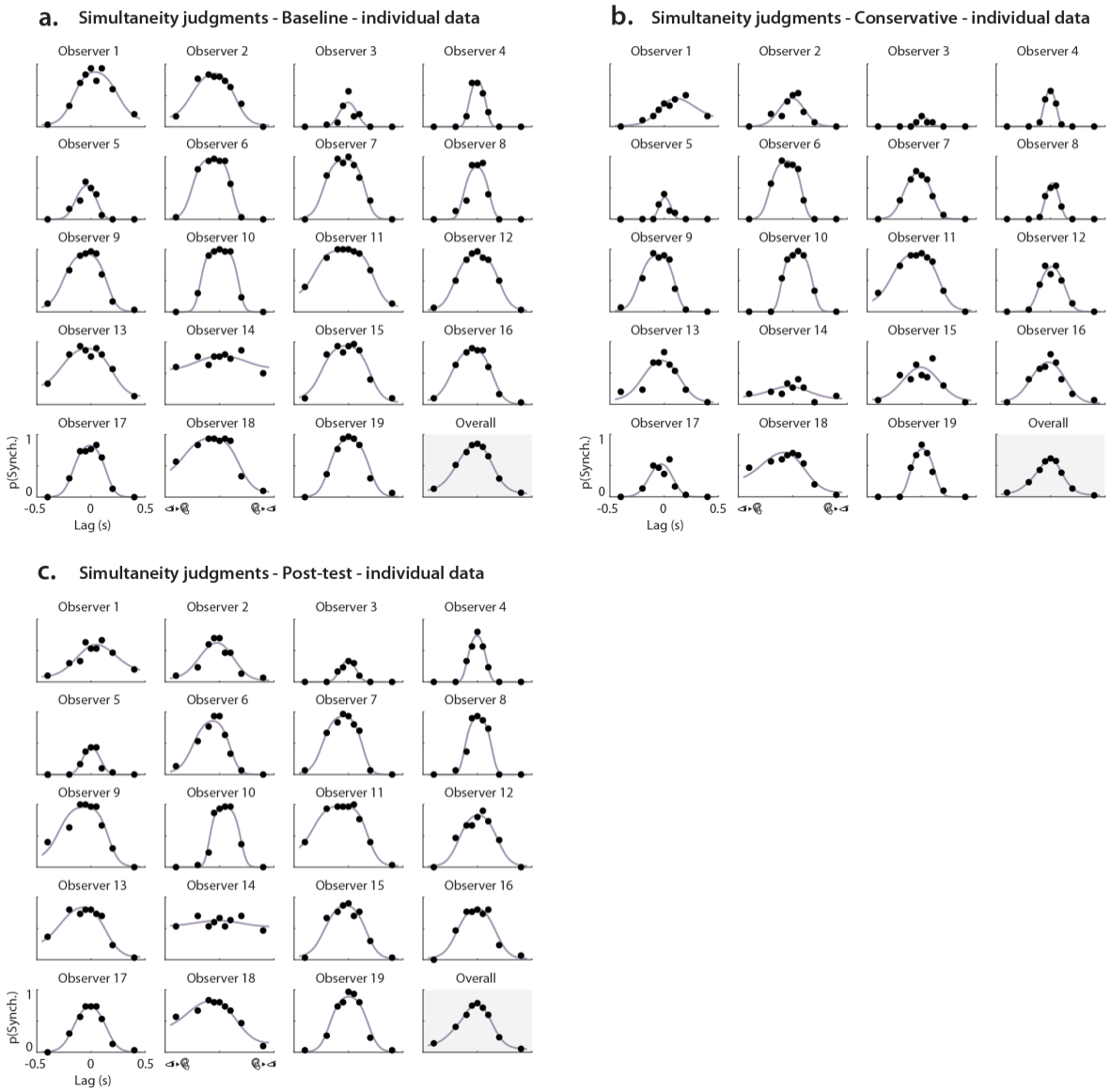


Supplementary Figure 1. Temporal determinants of multisensory integration in humans. Schematic representation of the stimuli with a manipulation of audiovisual lag (Panel A). Panels B-J shows a the stimuli, psychophysical data (dots)

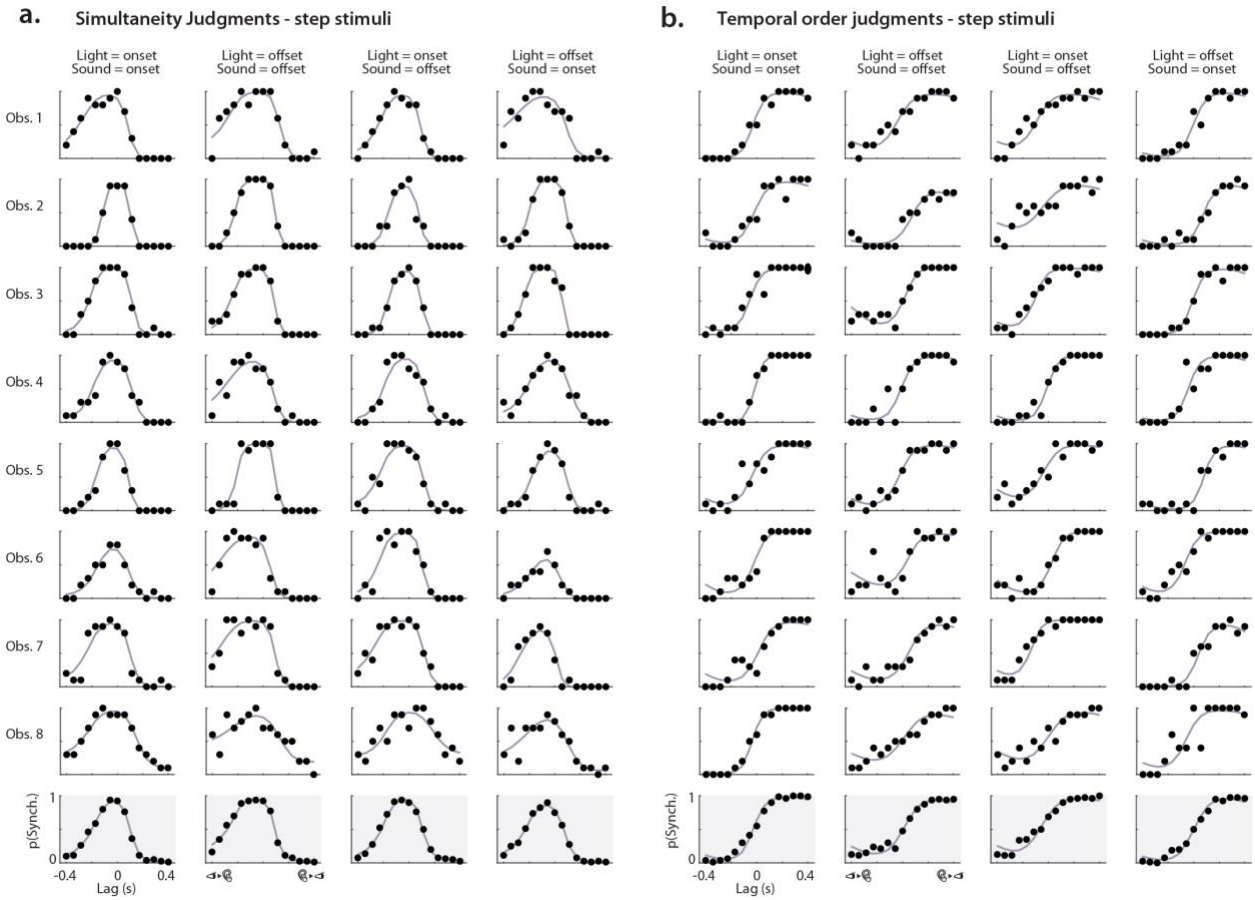
and model response (lines) of simulated studies investigating the temporal determinants of multisensory integration in humans using ecological audiovisual stimuli. The colour of the psychometric curves represents the psychophysical task (red=McGurk, black=simultaneity judgments, blue=temporal order judgments). Panel K shows the results of the permutation test: the histogram represents the permuted distribution of the correlation between data and model response. The arrow represents the observed correlation between model and data, and it is 4.7 standard deviation above the mean of the permuted distribution ($p<0.001$). Panel I represents the funnel plot of the Pearson correlation between model and data across all studies in Panels B-J, plotted against the number of trials used for each level of audiovisual lag. Red diamonds represent the McGurk task, blue diamonds and circles represent temporal order judgments, and black diamonds and circles represents simultaneity judgments. Diamonds represents human data, circles represent rat data. The dashed grey line represents the overall Pearson correlation between empirical and predicted psychometric curves, weighted by the number of trials of each curve. As expected, the correlation decreases with decreasing number of trials.



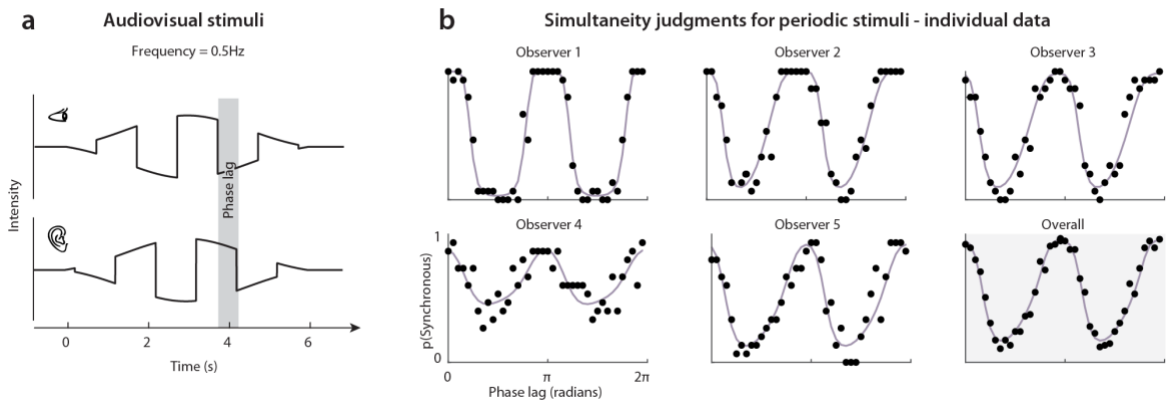
Supplementary Figure 2. Temporal determinants of multisensory integration in rats. Schematic representation of the stimuli (clicks and flashes) with a manipulation of audiovisual lag (Panel A). Panels B-J shows a the psychophysical data (dots) and model response (lines) of simulated studies. The colour of the psychometric curves represents the psychophysical task (black=simultaneity judgments, blue=temporal order judgments).



Supplementary Figure 3. Simultaneity judgment for impulse stimuli in humans, individual observer data (Yarrow et al. 2023). Panel A represents the baseline condition; Panel B represents the conservative condition (in which observers were instructed to respond "synchronous" only when absolutely sure); Panel C represents the post-test condition (in which observers were instructed to respond as they preferred, like in the baseline condition). Individual plots represent data from a single observer (dots) and model responses (lines), and consists of 135 trials each (total 7695 trials). The bottom-right plots represents the average empirical and predicted psychometric functions.



Supplementary Figure 4. Simultaneity and temporal order judgments for step stimuli (Parise & Ernst, 2023). Panel A represents the results of the simultaneity judgments; Panel B represents the data from the temporal order judgment. Each column represents a different combination of audiovisual step polarity (i.e., onset steps in both modalities, offset steps in both modalities, video onset and audio offset, video offset and audio onset). Different rows represent curves from different observers (dots) and model responses (lines). Each curve consists of 150 trials (total 9600 trials). The bottom row (light grey background) represents the average empirical and predicted psychometric functions.



Supplementary Figure 5. Simultaneity judgment for periodic stimuli (Parise & Ernst, 2023). Individual plots represent data from a single observer (dots) and model responses (lines). Each curve consists of 600 trials (total 3000 trials). The bottom-right curve represents the average empirical and predicted psychometric functions.

Supplementary Tables

Reference	Task	Stimuli	Congr.	Country	Observers	Num. obser.	Num lag	Num. trials	MCD-data correl.
Van Wassenhove et al. (2007)	McGurk	[V,A]=[/ga/,/ba/]	no	US	c.s.	18	29	5,580	906
Van Wassenhove et al. (2007)	McGurk	[V,A]=[/ka/,/pa/]	no	US	c.s.	21	29	6,510	.986
Van Wassenhove et al. (2007)	SJ	[V,A]=[/ga/,/ba/]	no	US	c.s.	18	29	3,132	.945
Van Wassenhove et al. (2007)	SJ	[V,A]=[/ka/,/pa/]	no	US	c.s.	21	29	3,654	.977
Van Wassenhove et al. (2007)	SJ	/da/	yes	US	c.s.	18	29	3,132	.986
Van Wassenhove et al. (2007)	SJ	/ta/	yes	US	c.s.	21	29	3,654	.986
Lee and Noppeney (2011)	SJ	Sentences (n=16)	yes	DE	c.s.	19	13	15,808	.993
Lee and Noppeney (2011)	SJ	Sentences (n=16)	yes	DE	musicians	18	13	14,976	.984
Lee and Noppeney (2011)	SJ	Piano melodies (n=16)	yes	DE	c.s.	19	13	15,808	.967
Lee and Noppeney (2011)	SJ	Piano melodies (n=16)	yes	DE	musicians	18	13	14,976	.993
Lee and Noppeney (2014)	SJ	Sentences (n=16)	sws	DE	c.s.+musicians	37	13	30,784	.984
Vroomen et al. (2011)	SJ	/tabi/	yes	NL	c.s.	30	17	12,240	.994
Vroomen et al. (2011)	SJ	/tabi/	sws	NL	c.s.	30	17	12,240	.99
Vroomen et al. (2011)	SJ	/tabi/ (disamb.)	sws	NL	c.s.	30	17	12,240	.988
Van Laarhoven et al. (2019)	SJ	/tabi/	yes	NL	c.s.	101	21	31,815	989
Vroomen et al. (2011)	TOJ	/tabi/	yes	NL	c.s.	30	17	12,240	.997
Vroomen et al. (2011)	TOJ	/tabi/	sws	NL	c.s.	30	17	12,240	.997
Vroomen et al. (2011)	TOJ	/tabi/ (disamb.)	sws	NL	c.s.	30	17	12,240	.988
Freeman et al. (2013)	McGurk	Syllables (n=3)	no	UK	young	27	9	4,374	.985
Freeman et al. (2013)	McGurk	Syllables (n=3)	no	UK	elderly	10	9	1,620	.996
Freeman et al. (2013)	McGurk	Syllables (n=3)	no	UK	Patient P.H.	1	9	162	.869
Freeman et al. (2013)	TOJ	Syllables (n=5)	no	UK	young	27	9	8,740	.999
Freeman et al. (2013)	TOJ	Syllables (n=5)	no	UK	elderly	10	9	3,240	.999
Freeman et al. (2013)	TOJ	Syllables (n=5)	no	UK	Patient P.H.	1	9	324	.991
Magnotti and Beauchamp (2013)	SJ	Words (n=4) HI	yes	US	c.s.	39	15	13,320	.966
Magnotti and Beauchamp (2013)	SJ	Words (n=4) LI	yes	US	c.s.	39	15	13,320	.991
Magnotti and Beauchamp (2013)	SJ	Words (n=4) HI, blur	yes	US	c.s.	39	15	13,320	.973
Magnotti and Beauchamp (2013)	SJ	Words (n=4) LI, blur	yes	US	c.s.	39	15	13,320	.993
Roseboom and Arnold (2011)	SJ	/ba/ (male/female)	yes	UK	c.s.	6	9	1,296	.963
Yuan et al. (2014)	McGurk	[V,A]=[/ga/,/ba/]	no	CN	c.s.	12	9	1,080	.892
Yuan et al. (2014)	SJ	[V,A]=[/ga/,/ba/]	no	CN	c.s.	13	9	1,170	.983
Ikeda and Morishita (2020)	SJ	Sentences (n=4)	yes	JP	c.s.	28	13	4,368	.964
Ikeda and Morishita (2020)	SJ	Sentences (n=6)	yes	JP	c.s.	22	13	3,432	.967
Ikeda and Morishita (2020)	SJ	Sentences (n=6)	noise	JP	c.s.	22	13	3,432	.937
Ikeda and Morishita (2020)	SJ	Flute melodies (n=4)	yes	JP	c.s.	28	13	4,368	.932
Ikeda and Morishita (2020)	SJ	Piano melodies (n=4)	yes	JP	c.s.	28	13	4,368	.912
Petrini et al. (2020)	TOJ	/tomorrow/	yes	UK	7-8 years old	15	7	1,050	.958
Petrini et al. (2020)	TOJ	/tomorrow/	yes	UK	10-11 years old	14	7	980	.972
Petrini et al. (2020)	TOJ	/tomorrow/	yes	UK	adults	14	7	980	.988
Petrini et al. (2020)	SJ	/tomorrow/	yes	UK	7-8 years old	15	7	1,050	.994
Petrini et al. (2020)	SJ	/tomorrow/	yes	UK	10-11 years old	14	7	980	.980
Petrini et al. (2020)	SJ	/tomorrow/	yes	UK	adults	14	7	980	.990
		105 unique stimuli		**6 unique countries**		**986 (individual curves) 454 (unique observers)**		**324,543 (total)**	**0.981 (mean)**

Supplementary Table 1. Summary of the experiments simulated in Supplementary Figure 1. The **first column** contains the reference of the study, the **second column** the task (McGurk, Simultaneity Judgment, and Temporal Order Judgment). The **third column** describes the stimuli: n represents the number of individual instances of the stimuli, "HI" and "LI" in

Magnotti and Beauchamp (2013) indicate speech stimuli with High and Low Intelligibility, respectively. "Blur" indicates that the videos were blurred. "Disamb" indicates that ambiguous speech stimuli (i.e., sine-wave speech) were disambiguated by informing the observers of the original speech sound. The **fourth column** indicates whether visual and acoustic stimuli were congruent. Here, incongruent stimuli refer to the mismatching speech stimuli used in the McGurk task. "SWS" indicates sine-wave speech; "noise" in Ikeda and Morishita (2020) indicates a stimulus similar to sine-wave speech but in which white noise was used instead of pure sinusoidal waves. The **fifth column** represents the country where the study was performed. The **sixth column** describes the observers included in the study: "c.s." indicates convenience sampling (usually undergraduate students) musicians in Lee and Noppeney (2011, 2014) were amateur piano players; Freeman et al. (2013) tested young observers (18-28 years old), a patient P.H. (67 years old) that after a lesion in the pons and basal ganglia reported hearing speech before seeing the lips move; and a group of age matched controls (59-74 years old). The **seventh column** reports the number of observers included in the study. Overall, the full dataset consisted of 986 individual psychometric curves; however, several observers participated in more than one experiment, so that the total number of unique observers was 454. The **eight column** reports the number of lags used in the method of constant stimuli. The **nineth column** reports the number of trials included in the study. The **tenth column** reports the correlation between empirical and predicted psychometric functions. The bottom row contains some descriptive statistics of the dataset.

Reference	Task	Stimuli	Observers	Num. obser.	Num lag	Num. trials	MCD-data correl.
Mafi et al. (2023)	TOJ	Click/flash	Male Wistar rats	8	14	28,784	.996
Mafi et al. (2023)	TOJ	Click/flash	Humans	10	14	12,964	.992
Schormans et al. (2017)	SJ	Click/flash	Male rats	7	5	7,000	.999
Schormans et al. (2017)	TOJ	Click/flash	Male rats	7	7	7,000	.994
Paulcan et al. (2023)	TOJ	Click/flash	male Sprague–Dawley rats	8	8	2,880	.996
Paulcan et al. (2023)	TOJ	Click/flash	male Sprague–Dawley rats	13	8	4,680	.991
Al-youzbaki et al. (2023)	TOJ	Click/flash	Sprague Dawley rats	10	9	32,400	.996
Al-youzbaki et al. (2023)	SJ	Click/flash	Sprague Dawley rats	12	11	47,520	.983
Al-youzbaki et al. (2023)	TOJ	Click/flash	Sprague Dawley rats / saline	10	9	32,400	.997
Al-youzbaki et al. (2023)	TOJ	Click/flash	Sprague Dawley rats / MK-801	10 (same)	9	32,400	.999
Al-youzbaki et al. (2023)	SJ	Click/flash	Sprague Dawley rats / saline	8	11	31,680	.996
Al-youzbaki et al. (2023)	SJ	Click/flash	Sprague Dawley rats / MK-801	8 (same)	11	31,680	.991
Schormans & Allman (2018)	TOJ	Click/flash Loudness	Sprague Dawley rats	10	7	n/a	.989
Schormans & Allman (2018)	SJ	Click/flash Loudness	Sprague Dawley rats	10	5	n/a	.996
Schormans & Allman (2023)	TOJ	Click/flash	Sprague Dawley rats	8	7	n/a	.995
Schormans & Allman (2023)	TOJ	Click/flash	Sprague Dawley rats / gabazine	8 (same)	7	n/a	.99
				193 (individual curves) 110 (rats)		**271,388 (total)**	**.994 (mean)**

Supplementary Table 2. Summary of the experiments simulated in Supplementary Figure 2. The **first column** contains the reference of the study, the **second column** the task (Simultaneity Judgment, and Temporal Order Judgment). The **third column** describes the stimuli. The **fourth column** indicates what rats were used as observers. The **fifth column** reports the number of rats in the study; 'same' means that the same rats took part in the experiment in the row above. The **sixth column** reports the number of lags used in the method of constant stimuli. The **seventh column** reports the number of trials included in the study (not available for all studies). The **eight column** reports the correlation between empirical and predicted psychometric functions. The bottom row contains some descriptive statistics of the dataset.

References

1 Yarrow, K., Solomon, J. A., Arnold, D. H. & Roseboom, W. The best fitting of three contemporary observer models reveals how participants' strategy influences the window of subjective synchrony. *Journal of Experimental Psychology: Human Perception and Performance* (2023).

2 Parise, C. V. & Ernst, M. O. Multisensory integration operates on correlated input from unimodal transients channels. *eLife*, **RP90841** (2023).