Spatiotemporal models for multisensory integration

Cesare V. Parise, University of Liverpool

cesare.parise@liverpool.ac.uk

Abstract

Multisensory integration is a process of redundancy exploitation, in which our brains combine information across the senses to obtain more reliable perceptual estimates. While the high-level computational principles of multisensory integration are well understood, little is knowns as to how the low-level properties of the signals ultimately determine the integrated percept. This study demonstrates that a bottom-up approach, based on luminance- and sound-level analyses, is sufficient to jointly explain the spatiotemporal determinants of audiovisual integration and crossmodal attention. When implemented using an architecture analogous to the motion detectors found in the insect brain, such low-level analyses can broadly reproduce human behaviour–as tested in a large-scale simulation of 42 classic experiments on the spatial, temporal and attentional aspects of multisensory integration.

Introduction

Our experiences are inherently multisensory. For instance, during speech perception, our brains integrate audiovisual information to enhance speech intelligibility, often beyond conscious awareness. This becomes evident in the McGurk illusion[1], where the auditory perception of a spoken syllable is influenced by the visual presentation of a mismatched syllable. Similarly, our brains integrate audiovisual cues for spatial localization, as illustrated by the ventriloquist illusion[2], where perceived sound location shifts toward a synchronous visual stimulus.

In recent decades, numerous studies have explored the interplay between the senses[3]: consistent behavioural and neurophysiological findings reveal 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)[4]. When redundant multisensory information is integrated, the resulting percept is more accurate and reliable[5]. Several theories have been successfully proposed to account for audiovisual integration across time and space[6-9]–often framed in terms of Bayesian Causal Inference (BCI), 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[6,7,10].

Models of multisensory integration usually operate at a high level of abstraction, and predict behavioural or neurophysiological responses based on parameters of the stimuli, not the actual stimuli [6-10]. For example, in models of audiovisual time perception, crossmodal lag (expressed in seconds) is used to explain perceived synchrony[7,8]. However, observers lack direct access to lag information: that must be estimated from continuous, dynamic patterns of light and sound. Only a few models can take simple stimuli as input[11,12], and none of them can handle complex real-life audiovisual stimuli. In essence, we lack a biologically-plausible model capable of taking audiovisual footage as input and replicate basic phenomena in multisensory perception, such as detecting crossmodal synchrony, bimodal localization, and attentional capture.

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. A section 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. In previous studies, we proposed a biologically plausible solution to detect temporal correlation across the senses (Figure 1A)[11,13-15]. 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, encompassing its temporal, spatial and attentional aspects.

The present approach posits the existence of elementary processing units, the Multisensory Correlation Detectors (MCD)[14], each integrating time-varying input from unimodal transient channels through a set of temporal filters and elementary operations (Figure 1A, see Supplementary Material). 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 as input, hence capable to process any stimuli used in previous studies in audiovisual integration. This makes it possible to directly compare its behaviour to human observers'. Emphasis will be given to studies using real-life stimuli which, despite their manifest ecological value, simply cannot be handled by alternative models. However, the aim of this study is not to champion a specific model but rather to discuss a general approach to multisensory perception, based on biologically-plausible low-level analyses of visual and auditory information (see[16]).

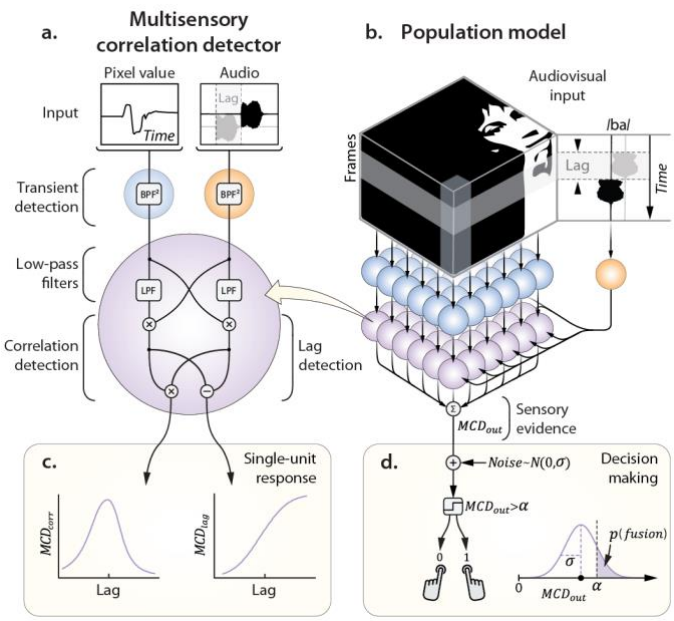

Figure 1. The MCD population model. Panel A: schematic representation of a single MCD unit. The input visual signal represents the intensity over time of a pixel in the mouth area of one of the movie stimuli used by van Wassenhove et al. (2007), while the audio input is the movie soundtrack (the syllable /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, corrupted with additive Gaussian noise, and compared to a criterion to generate a forced-choice response (**Panel D**).

Temporal determinants of audiovisual integration

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 with the McGurk illusion) through psychophysical forced-choice tasks[17,18]. Among them, we obtained both the audiovisual footage and the psychophysical data from 36 different experiments that used ecological audiovisual stimuli (i.e., speech and performing musicians, Figure 2A, Table 1, for the inclusion criteria, see Supplemental Information): 24 experiments were simultaneity judgments[7,19-26], 6 temporal order judgments[24,27], 6 others assessed the McGurk effect[19,21,27]. All used the same manipulation of audiovisual lags (i.e., method of constant stimuli).

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 Supplemental Information)[11,13-15]. Results demonstrate that a population of MCDs can broadly account for audiovisual temporal perception of ecological stimuli, and near-perfectly (rho=0.99) reproduces the empirical psychometric functions for simultaneity judgments, temporal order judgments and the McGurk effect (Figure 2C). 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 Supplementary Material). In no case the psychometric curves predicted from permuted stimuli led to a better fit than the original sample (p=0, see Figure 2B): this demonstrates the significant role of low-level stimulus properties on multisensory integration.

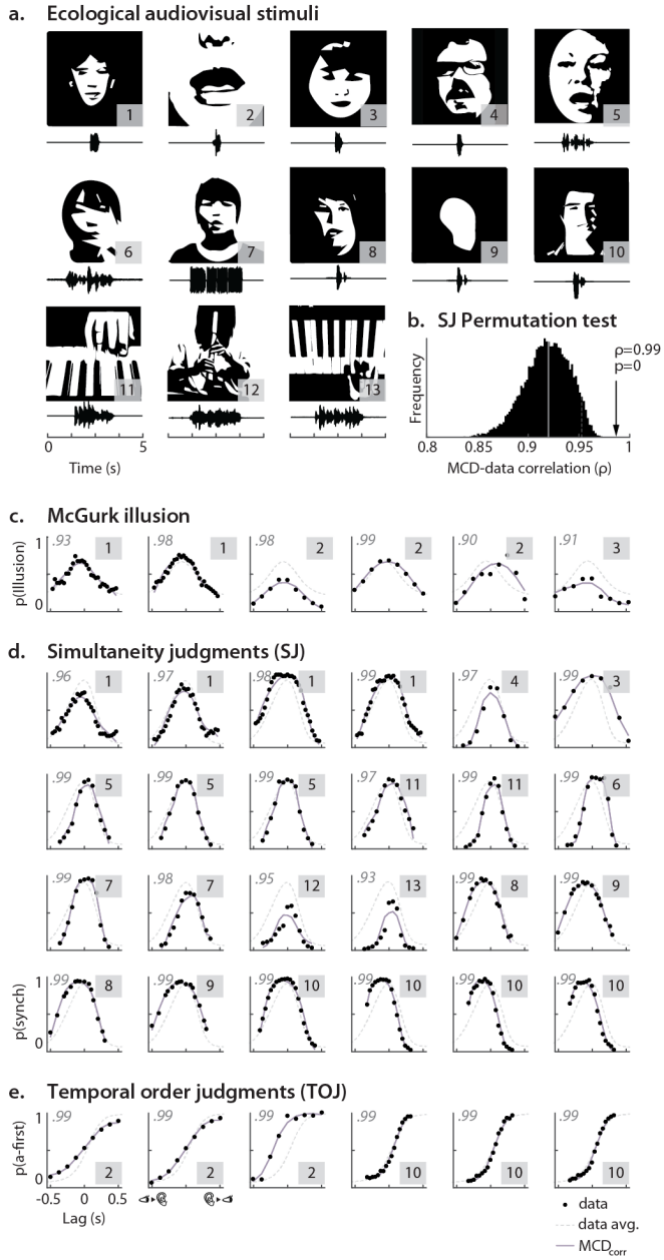

Figure 2. Ecological stimuli and psychophysical responses. Panel A shows a still frame and the soundtrack of the ecological stimuli used to investigate the temporal constraints of audiovisual integration. Note that this is only a representative sample of all the stimuli used in the various experiments, which consists of 104 unique stimuli. The numbers in each frame correspond to the numbers in Panels C-E. In compliance with biorXiv guidelines, all video frames in this and the other figures have been distorted and converted to black and white to conceal the actors' identities. **Panel B** displays the results of the permutation test: the horizontal axis represents the correlation between the model and the data of the simultaneity judgments. Note how none of the permutations lead to a Pearson correlation above that obtained with the original stimuli (rho=0.99, p=0), as shown in Panel D. **Panels C-D** illustrate how crossmodal lag affects the McGurk illusion, perceived simultaneity and probability of perceiving audio before vision, respectively. The dots correspond to human responses, while the continuous lines represent the responses of the model. Dashed thin lines represent the average empirical psychometric curves for the three tasks to highlight the idiosyncrasies of the individual experiments. Taken together the full dataset consists of 36 experiments (comprising 900 individual psychometric functions), run across 6 different countries on 411 participants, for a total of 318,523 trials. Details for each dataset and stimuli are reported in Supplementary Table 1.

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).

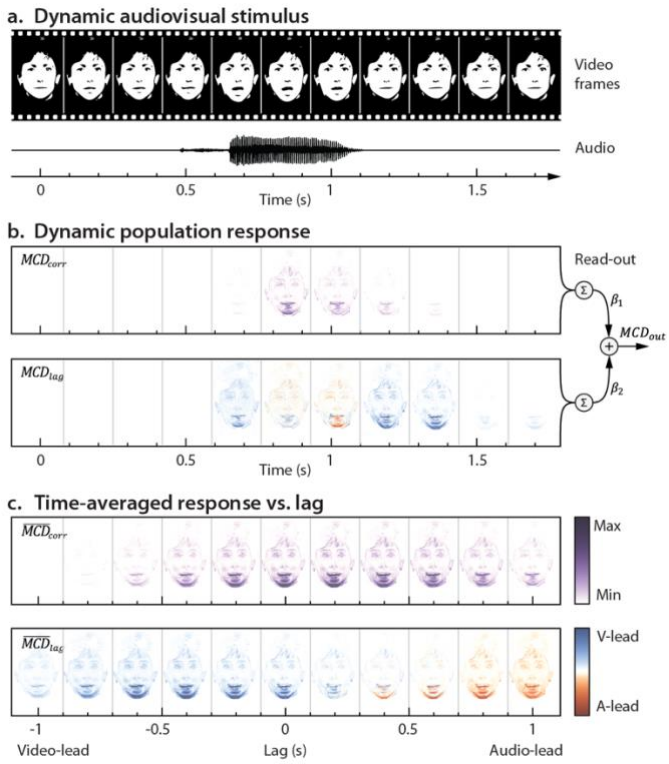

Figure 3. Ecological audiovisual stimuli and model responses. Panel A displays the frames and soundtrack of a dynamic audiovisual stimulus over time (in this example, video and audio tracks are synchronous, and the actress utters the syllable /da/). **Panel B** shows how the dynamic population responses 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 Figure 2D, third box from left.

Spatial determinants of audiovisual integration

Classic experiments on the spatial determinants of audiovisual integration usually require participants to localize the stimuli under systematic manipulations of the discrepancy and reliability (i.e., precision) of the spatial cues[9] (Figure 4B). This allows one to assess how unimodal cues are weighted and combined, to give rise to phenomena such as the ventriloquist illusion[2]. When the spatial discrepancy across the senses is low, human responses are well described by Maximum Likelihood Estimation (MLE)[9], where unimodal information is combined in a statistically optimal fashion, so as to maximize the reliability of the multimodal percept (see Equations 11-14, Supplementary Material).

Given that MCD units are multipliers, the time-averaged MCD population output is equivalent to that of the MLE model (Figure 4A, see Supplementary Information). This can be illustrated by simulating the experiment of Alais and Burr[9] using both models. In this experiment, participants 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, without fitting parameters, a population of MCDs (Equations 15-16) can jointly account for the bias and precision of the bimodal percept (Figure 4D-E). The MCD model, however, goes beyond the MLE model and also returns the

temporal profile of both unimodal and bimodal responses (which can be used in future research to test the model against neurophysiological data).

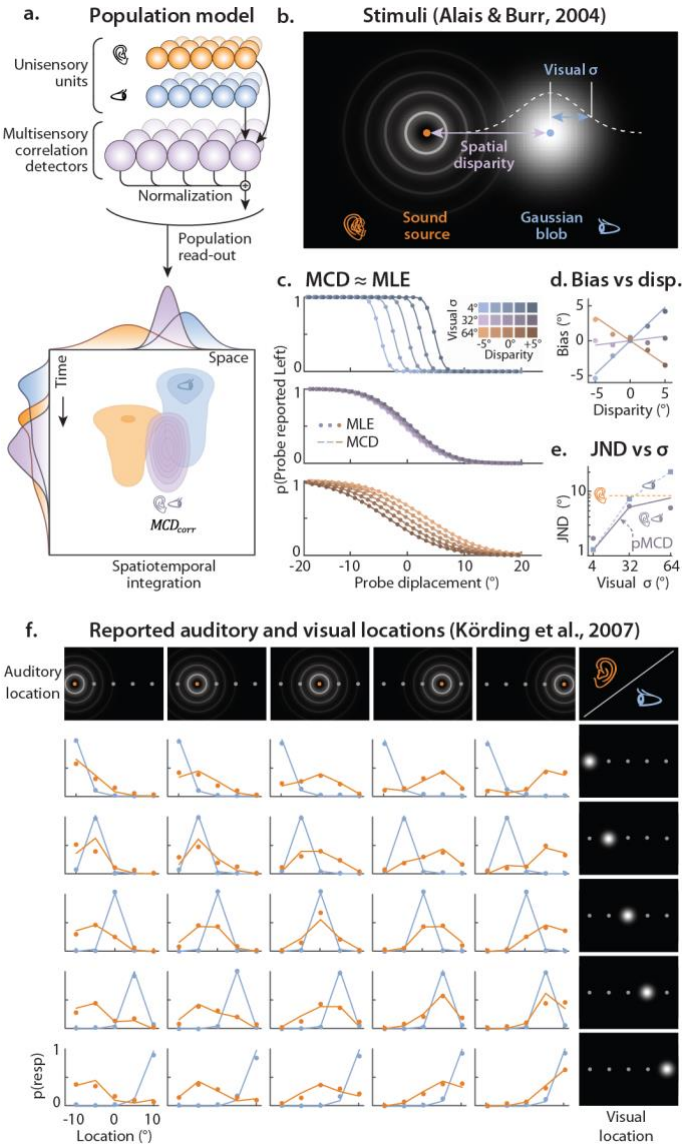

Figure 4. Audiovisual integration in space. Panel A top represents the MCD population model for spatialized audio. Visual and auditory input units receive input from corresponding spatial locations, and feed into spatially-tuned MCD units. The output of each MCD unit is eventually normalized by the total 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 (2004). 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, 2004). **Panel E** shows how the just noticeable differences (JND, the random localization error) varies 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, 2004). **Panel F** represents the stimuli and results of the experiment used by Körding et al (2007) 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 combinations of audiovisual spatial locations. The dots represent human data, while the lines represent the responses of the MCD population model.

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[28]—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[6], where uni- and bimodal location estimates are weighted based on the probability that the two modalities share a common cause (Equation 17). BCI model was originally tested in an experiment in which a sound and a light were simultaneously presented from one of five locations, and participants had to report the spatial location of both modalities[6] (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 (see Supplementary Material) which near-perfectly replicated the empirical data (rho=0.99)–even slightly outperforming the original model (rho=0.97), while relying on fewer free parameters. Together, these simulations show that a behaviour consistent with BCI and MLE naturally emerges from a population of MCDs.

Attention and audiovisual gaze behaviour

Multisensory stimuli are typically salient, and a vast body of literature shows that spatial attention is usually pulled towards audiovisual stimuli in a bottom-up fashion[29]. This fundamental aspect of multisensory perception is naturally captured by a population of MCDs, whose dynamic response explicitly represents the regions in space 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, with no free parameters and in a purely bottom-up fashion.

Figure 5A shows the stimuli and eye-tracking data from the experiment of Coutrot and Guyader[30], in which participants 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 21-22). The tight correspondence of predicted and empirical salience is illustrated in Figure 5C 5C: note how population responses peak based on the active speaker.

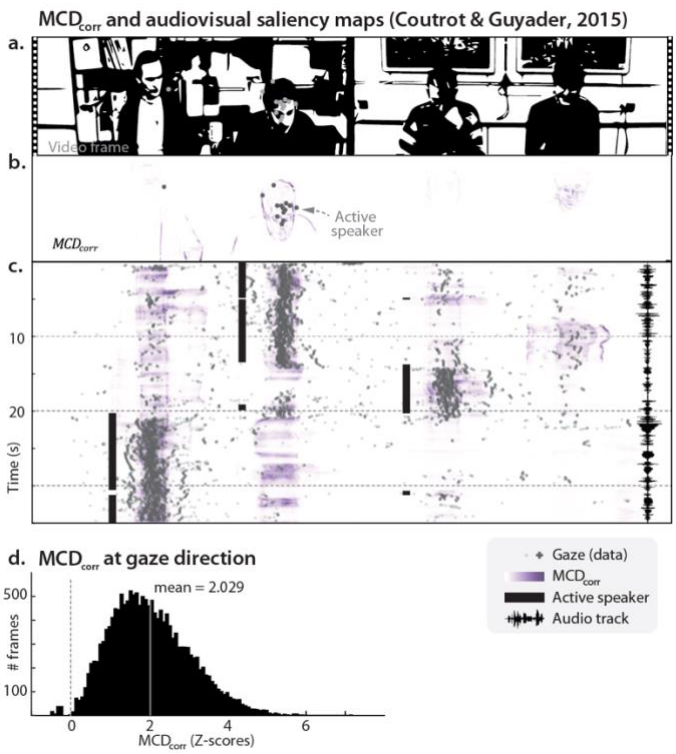

Figure 5. Audiovisual saliency maps. Panel A represents a still frame of Coutrot and Guyader's (2015) stimuli (Video 1). The white dots represent gaze direction of the various observers. **Panel B** represents the MCD population response for the frame in Panel A (see model schematics on the right). The gray dots represent gaze direction of the various observers (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.

Discussion

This study highlights the value of a low-level approach to the problem of multisensory integration. In a purely bottom-up fashion, a population of processing units built using elements of known biological plausibility, is sufficient to account for the spatiotemporal determinants of multisensory integration. Being capable of receiving audiovisual input that vary over time and space, a population of MCD is flexible enough to process any possible audiovisual stimuli. This provides unprecedented freedom in simulating prior research: as long as both stimuli and responses are available, the population read-out process can be adapted to quantitatively predict (or fit) observed responses. Across a large-scale simulation of 42 classic experiments (450 participants, ~1000 individual psychometric functions, ~120 unique stimuli), a population of MCDs proved capable to quantitatively reproduce the spatial, temporal and attentional aspects of human audiovisual integration, with an average Pearson correlation above 0.97. All while keeping the population model constant–both structurally and parametrically–and only allowing the decision making processes to vary across the six different behavioral tasks.

The fact that identical low-level analyses can account for all the various experiments considered here, directly addresses several open questions in multisensory research. For instance, psychometric functions for speech and non-speech stimuli are often significantly different[31]: this has been considered evidence that speech may be special and processed via dedicated mechanisms[32]. However, the same low-level analyses are sufficient to account for the variability of observed responses (Figure 2D). This implies that what determines the difference in psychometric curves across stimuli, are ultimately the low-level properties of the stimuli themselves. Similarly, expertise and experience strongly modulate certain aspects of multisensory perception[22]: reported audiovisual simultaneity illusions significantly vary between musicians and non-musicians (see Figure 2D, second row, compare fourth and fifth curves). Likewise, the McGurk illusion varies between the young and aging population (Figure 2C, compare third and fourth curves)[27]. Our simulations show that differences in the decision-making process are sufficient to account for such variations, with no need to postulate structural or parametric differences in low-level perceptual analyses across the groups. More generally, it is intriguing that a population of MCDs, a computational architecture originally introduced for insect motion vision, can explain the occurrence of speech illusions in humans.

The present approach has its limits. For instance, while a population of MCDs can explain *when* the McGurk Illusion occurs, it does not explain the process of phoneme categorization that ultimately determines *what* syllable is perceived[10]. More generally, it is well known that cognitive and affective factors modulate our responses to multisensory stimuli[3]. While a purely bottom-up approach cannot provide a complete explanation for the problem of multisensory perception, the modularity of a population model makes it easy to extend the system to include high-level factors. Indeed, a population of MCDs may well sit as a biologically-inspired front-end of a multimodal neural network, with deeper layers representing progressively higher-levels of cognition. Although this study focused on human audiovisual integration, the same approach can be naturally extended to other instances of sensory integration and animal species, or leveraged in machine learning[33].

Besides simulating observed behavior, a low-level 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[13]. 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[34]. Being based on formally identical computational principles[15], the present approach holds the same potential for multisensory research.

Supplementary Material

The MCD population model

The architecture of each MCD unit used here is the same as described in Parise and Ernst[14], 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 [16] 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.[35] 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}; for simplicity, here we constrained τ_{bp} to be identical for both modalities, and its value is 0.034 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[16]. 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[11], 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 0.123 s.

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

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

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

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

The **temporal constants** of the filters were fitted using the Bayesian Adaptive Direct Search (BADS) algorithm[36], set to maximize the correlation between the empirical and predicted psychometric functions of the temporal determinants of multisensory integration shown in Figure 2 (weighted by sample size). 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.

The temporal constants used in the simulations of the temporal determinants of multisensory integration (see below) were estimated with a leave-one out cross-validation. The data left-out consisted of all experiments that used the same stimuli, were run by the same team, or both. The estimated parameters, $(\tau_{bp}, \tau_{lp}) = (0.034\ , 0.123)s$, were stable across the iterations of the cross-validation; this was expected given the size and diversity of the total dataset. Parameter estimation was considerably compute-intensive, hence the amount of data had to be reduced by rescaling the videos to one-tenth 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 a button-press in a forced-choice task. 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.

Temporal determinants of multisensory integration

The experiments on the temporal constraints of audiovisual integration considered here, all rely on psychophysical forced-choice tasks to assess the effects of crossmodal lags on perceived synchrony, temporal order, and the McGurk illusion. In the case simultaneity judgments, on each trial participants reported whether visual and auditory stimuli appeared synchronous or not ($resp = \{yes, no\}$). In temporal order judgments, on each trial participants had to respond which modality came first ($resp = \{vision\ first, audition\ first\}$). Finally, in the case of the McGurk, participants 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 describes how audiovisual timing affect perceived synchrony, temporal order, and the McGurk illusion. Here we will model all these three tasks using the same read-out and decision-making process.

To account for human 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[37], 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 function, β_{corr} and β_{lag} are linear coefficients representing the scaling and weighing of $\overline{MCD_{corr}}$ 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

psychometric functions using *fitglm* Matlab (binomial distribution and probit link function). Overall, the model provided an excellent fit to all 36 empirical psychometric functions, and the average Pearson correlation between human and model response (weighted by sample size) is r=0.986. Naturally, model-data correlation varied across the 36 experiments, largely due to sample size. This can be appreciated when the MCD-data correlation for each experiment is plotted against the number of trial for each lag in a funnel plot (see Supplementary Figure S1). The number of trials for each lag determine the binomial error for data point in Figure 2, and accordingly the funnel plot shows that lower MCD-data correlation are more commonly observed for curves based on smaller sample size. This shows that the upper limit in the MCD-data correlation is constrained by the reliability of the dataset rather than systematic errors in the model.

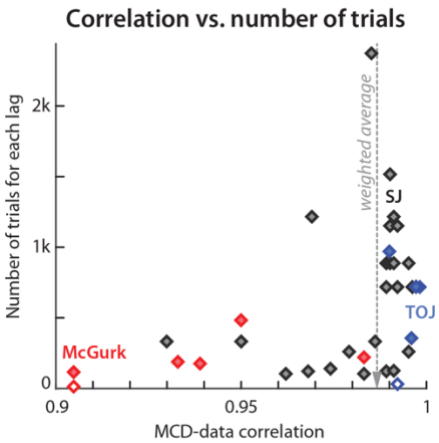

Figure S1. Funnel plot of the simulations of the temporal determinants of multisensory integration shown in Figure 2. The Pearson correlation between empirical and predicted psychometric functions is plotted against the number of trials used for each level of audiovisual lag. Red diamonds represent the McGurk task, blue diamonds represent temporal order judgments, and black diamonds represents simultaneity judgments. Empty diamonds represents data from patient P.H. The dashed grey line represents the overall Pearson correlation between empirical and predicted psychometric curves, weighted by the number of trials of each curve.

To test to what extent the MCD model could fit the psychometric functions, regardless of the exact input stimuli used to predict each curve, we performed a permutation test, where psychometric curves were predicted using stimuli from different experiments (but with the same manipulation of lag). For this simulation we only used the data from simultaneity judgment tasks, as with 24 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 20,000 permutations, shows that the goodness of fit obtained with the original stimuli was always superior to that obtained with permuted stimuli. Specifically, the null distribution of the mean Pearson correlation of predicted vs. empirical psychometric curves had a mean of 0.92 (sigma=0.025), while such a c13orrelation rose to 0.986 when the MCD received the original stimuli.

Spatial determinants of multisensory integration

Most behavioral studies on AV perception of space only manipulated the horizontal spatial dimension, hence the stimuli can be reduced to $s_m(x,t)$. Therefore, also the population outputs vary over the same dimensions (x,t). Additionally, based on preliminary simulations, the output MCD_{lag} does not seem necessary to account for audiovisual integration in space, hence only the population response $MCD_{corr}(x,t)$ 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 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 used in the present simulations 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 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 in the stimulus (see also[11,13-15]). For this simulation, we only used the data from Alais and Burr's[9] subject LM, as it is the only participant for which the full dataset is publicly available (other participants, 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, I replicated the various conditions run by observer LM (conditions slightly varied across observers), 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 in line with Figure 1 of Alais and Burr (2004)

Once shown that $p_{MLE}(x) = p_{MCD}(x)$, it is clear that the MCD is equally capable of predicting human responses. However, for completeness, I 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 and just noticeable differences (JND).

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:

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

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 human responses as follows:

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

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[6]. 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 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. Parameters were fitted using BADS[36] 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 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 participants preferentially looked at the pixels with the highest MCDcorr responses. 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 assessed the average MCD responses at gaze direction for each frame (t); this was done by weighing MCD_{corr} by p_{gaze} as follows

$$\overline{MCD_{gaze}}(t) = \sum_x \sum_y p_{gaze}(x,y,t) \cdot MCD_{corr}(x,y,t) \quad \text{Equation 21}$$

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

$$SMD(t) = \frac{\overline{MCD_{gaze}}(t) - \mu[MCD_{corr}(x,y,t)]}{\sigma[MCD_{corr}(x,y,t)]} \quad \text{Equation 22}$$

Across the over 16,000 frames of the available dataset, the average SMD was 2.03. Given that the Standardized Mean Deviation (SMD) serves as a metric for effect size, and effect sizes surpassing 1.2 are deemed very large [38], it is remarkable that, with zero free parameters, the MCD model can so tightly account for human gaze behavior in a purely bottom-up fashion.

Pre-processing of ecological audiovisual footage

The diversity of the stimuli used in this dataset requires some preprocessing before feeding them to the population model. The 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 delay. 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 induced by the onset of the video.

Video frames were spatially down sampled by a factor of 10, and the static background was cropped. On a practical side, this made the simulations much lighter (which is crucial for parameter estimation); on a theoretical side, such a down sampling can represents the Gaussian spatial pooling of luminance across the visual field (unfortunately, the present datasets do not always 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 human and model behavior, this study requires a large and diverse dataset consisting of both the raw stimuli and human responses. For that, we adopted a convenience sampling and simulated the studies for which both datasets 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. For the **temporal determinants of multisensory integration**, 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. Thirty-six individual experiments met the inclusion criteria. These varied in terms of stimuli, participants, and tasks. In terms of stimuli, the dataset in Figure 2 consists of responses from 104 unique real-life videos. 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, camera shake, etc. The remaining non-speech stimuli consist of footage of actors playing a piano or a flute. 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 participants, besides the general population of six different countries, the dataset in Figure 2 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[27]. Taken together the 36 curves shown in Figure 2 represent the average from 900 individual psychometric functions for a total of 318,523 trials, from 411 unique observers (some of them run more than one experiment).

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 therefore 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[9] and Körding et al.[6]. Such simulations allow us to compare the data with our model, and–perhaps more importantly–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.

The temporal constants of the MCD were fitted to experiments that used mostly audiovisual speech stimuli, recorded indoor, with artificial lights, and with little to no camera shake. Therefore, given comparable speech stimuli a population of MCDs should account for human **audiovisual attention and gaze behaviour**. Coutrot and Guyader[30] run the ideal testbed for this hypothesis: they recorded eye movements from 20 participants 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 5 and Movie S2).

Reference	Task	Stimuli	Congr.	Country	Observers	Num. obser.	Num lag	Num. trials	MCD-data correl.
Van Wassenhove et al. (2007)	McGurk	[V$_v$A]=[/ga/$_a$/ba/]	no	US	c.s.	18	29	5,580	.933
Van Wassenhove et al. (2007)	McGurk	[V$_v$A]=[/ka/$_a$/pa/]	no	US	c.s.	21	29	6,510	.983
Freeman et al. (2013)	McGurk	Syllables (n=3)	no	UK	young	27	9	4,374	.95
Freeman et al. (2013)	McGurk	Syllables (n=3)	no	UK	elderly	10	9	1,620	.939
Freeman et al. (2013)	McGurk	Syllables (n=3)	no	UK	Patient P.H.	1	9	162	.905
Yuan et al. (2014)	McGurk	[V$_v$A]=[/ga/$_a$/ba/]	no	CN	c.s.	12	9	1,080	.905
Van Wassenhove et al. (2007)	SJ	[V$_v$A]=[/ga/$_a$/ba/]	no	US	c.s.	18	29	3,132	.962
Van Wassenhove et al. (2007)	SJ	[V$_v$A]=[/ka/$_a$/pa/]	no	US	c.s.	21	29	3,654	.968
Van Wassenhove et al. (2007)	SJ	/da/	yes	US	c.s.	18	29	3,132	.983
Van Wassenhove et al. (2007)	SJ	/ta/	yes	US	c.s.	21	29	3,654	.989
Roseboom and Arnold (2011)	SJ	/ba/ (male/female)	yes	UK	c.s.	6	9	1,296	.974
Yuan et al. (2014)	SJ	[V$_v$A]=[/ga/$_a$/ba/]	no	CN	c.s.	13	9	1,170	.991
Lee and Noppeney (2011)	SJ	Sentences (n=16)	yes	DE	c.s.	19	13	15,808	.991
Lee and Noppeney (2011)	SJ	Sentences (n=16)	yes	DE	musicians	18	13	14,976	.99
Lee and Noppeney (2014)	SJ	Sentences (n=16)	sws	DE	c.s.+musicians	37	13	30,784	.985
Lee and Noppeney (2011)	SJ	Piano melodies (n=16)	yes	DE	c.s.	19	13	15,808	.969
Lee and Noppeney (2011)	SJ	Piano melodies (n=16)	yes	DE	musicians	18	13	14,976	.992
Ikeda and Morishita (2020)	SJ	Sentences (n=4)	yes	JP	c.s.	28	13	4,368	.986
Ikeda and Morishita (2020)	SJ	Sentences (n=6)	yes	JP	c.s.	22	13	3,432	.995
Ikeda and Morishita (2020)	SJ	Sentences (n=6)	noise	JP	c.s.	22	13	3,432	.979
Ikeda and Morishita (2020)	SJ	Flute melodies (n=4)	yes	JP	c.s.	28	13	4,368	.95
Ikeda and Morishita (2020)	SJ	Piano melodies (n=4)	yes	JP	c.s.	28	13	4,368	.93
Magnotti and Beauchamp (2013)	SJ	Words (n=4) HI	yes	US	c.s.	39	15	13,320	.989
Magnotti and Beauchamp (2013)	SJ	Words (n=4) HI, blur	yes	US	c.s.	39	15	13,320	.995
Magnotti and Beauchamp (2013)	SJ	Words (n=4) LI	yes	US	c.s.	39	15	13,320	.99
Magnotti and Beauchamp (2013)	SJ	Words (n=4) LI, blur	yes	US	c.s.	39	15	13,320	.991
Van Laarhoven et al. (2019)	SJ	/tabi/	yes	NL	c.s.	101	21	31,815	.99
Vroomen et al. (2011)	SJ	/tabi/	yes	NL	c.s.	30	17	12,240	.996
Vroomen et al. (2011)	SJ	/tabi/	sws	NL	c.s.	30	17	12,240	.992
Vroomen et al. (2011)	SJ	/tabi/ (disamb.)	sws	NL	c.s.	30	17	12,240	.989
Freeman et al. (2013)	TOJ	Syllables (n=5)	no	UK	young	27	9	8,740	.99
Freeman et al. (2013)	TOJ	Syllables (n=5)	no	UK	elderly	10	9	3,240	.996
Freeman et al. (2013)	TOJ	Syllables (n=5)	no	UK	Patient P.H.	1	9	324	.992
Vroomen et al. (2011)	TOJ	/tabi/	yes	NL	c.s.	30	17	12,240	.998
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	.997
		104 unique stimuli		**6 unique countries**		**900 (individual curves) 411 (unique observers)**	**13 (median)**	**317,071 (total)**	**0.986 (mean)**

Table S1. Summary of the experiments simulated in Figure 2. 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 participants 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 colum** 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 participants (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 participants included in the study. Overall, the full dataset consisted of 900 individual psychometric curves; however, several observers participated in more than one experiment, so that the total number of unique observers was 411. 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.

References

1 McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264** 746-748 (1976).
2 Stratton, G. Vision without inversion of the retinal image. *The Psychological Review* **4**, 463-481 (1897).
3 Stein, B. E. (MIT Press, Cambridge, MA, 2012).
4 Stein, B. E. & Stanford, T. R. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience* **9**, 255-266 (2008).
5 Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429-433 (2002).
6 Körding, K. P. *et al.* Causal inference in multisensory perception. *PLoS One* **2**, 943 (2007).
7 Magnotti, J. F., Ma, W. J. & Beauchamp, M. S. Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology* **4**, 798 (2013).
8 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).
9 Alais, D. & Burr, D. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology* **14**, 257-262 (2004).
10 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).
11 Parise, C. V. & Ernst, M. O. Correlation detection as a general mechanism for multisensory integration. *Nature Communications* **7:11543**, 1-9 (2016).
12 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).
13 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).
14 Parise, C. V. & Ernst, M. O. Multisensory integration operates on correlated input from unimodal transients channels. *eLife*, **RP90841** (2023).
15 Horsfall, R., Wuerger, S. & Meyer, G. Visual intensity-dependent response latencies predict perceived audio–visual simultaneity. *Journal of Mathematical Psychology* **100**, 102471 (2021).
16 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).
17 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).
18 Vroomen, J. & Keetels, M. Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics* **72**, 871-884 (2010).
19 Van Wassenhove, V., Grant, K. W. & Poeppel, D. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia* **45**, 598-607 (2007).
20 Roseboom, W. & Arnold, D. H. Twice upon a time: multiple concurrent temporal recalibrations of audiovisual speech. *Psychological Science* **22**, 872-877 (2011).
21 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).
22 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).
23 Ikeda, T. & Morishita, M. How Are Audiovisual Simultaneity Judgments Affected by Multisensory Complexity and Speech Specificity? *Multisensory Research* **34**, 49-68 (2020).
24 Vroomen, J. & Stekelenburg, J. J. Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition* **118**, 75-83 (2011).
25 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).
26 Lee, H. & Noppeney, U. Temporal prediction errors in visual and auditory cortices. *Current Biology* **24**, R309-R310 (2014).
27 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).
28 Chen, L. & Vroomen, J. Intersensory binding across space and time: a tutorial review. *Attention, Perception, & Psychophysics* **75**, 790-811 (2013).
29 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).
30 Coutrot, A. & Guyader, N. in *2015 23rd European Signal Processing Conference (EUSIPCO)*. 1531-1535 (IEEE).
31 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).
32 Tuomainen, J., Andersen, T. S., Tiippana, K. & Sams, M. Audio–visual speech perception is special. *Cognition* **96**, B13-B22 (2005).
33 Baltrusaitis, T., Ahuja, C. & Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 423-443 (2018).
34 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).
35 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).
36 Acerbi, L. & Ma, W. J. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems* **30** (2017).
37 Parise, C. V. & Ernst, M. O. Noise, multisensory integration, and previous response in perceptual disambiguation. *PLoS Computational Biology* **13**, e1005546 (2017).
38 Sawilowsky, S. S. New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* **8**, 26 (2009).