Bayesian networks and information theory for audio-visual perception modeling

Patricia Besson, Jonas Richiardi, Christophe Bourdin, Lionel Bringoux, Daniel R. Mestre, Jean-Louis Vercher

Biol Cybern (2010) 103:213–226
DOI 10.1007/s00422-010-0392-8

ORIGINAL PAPER

Biological Cybernetics

Bayesian networks and information theory for audio-visual perception modeling

Patricia Besson · Jonas Richiardi · Christophe Bourdin · Lionel Bringoux · Daniel R. Mestre · Jean-Louis Vercher

Received: 18 November 2009 / Accepted: 12 April 2010 / Published online: 26 May 2010

Abstract Thanks to their different senses, human observers acquire multiple information coming from their environment. Complex cross-modal interactions occur during this perceptual process. This article proposes a framework to analyze and model these interactions through a rigorous and systematic data-driven process. This requires considering the general relationships between the physical events or factors involved in the process, not only in quantitative terms, but also in term of the influence of one factor on another. We use tools from information theory and probabilistic reasoning to derive relationships between the random variables of interest, where the central notion is that of conditional independence. Using mutual information analysis to guide the model elicitation process, a probabilistic causal model encoded as a Bayesian network is obtained. We exemplify the method by using data collected in an audio-visual localization task for human subjects, and we show that it yields a well-motivated model with good predictive ability. The model elicitation process offers new prospects for the investigation of the cognitive mechanisms of multisensory perception.

P. Besson (✉) · C. Bourdin · L. Bringoux · D. R. Mestre · J.-L. Vercher
Institute of Movement Sciences, CNRS & Université de la Méditerranée, Marseille, France
e-mail: patricia.besson@univmed.fr

J. Richiardi
Medical Image Processing Laboratory, EPFL, Lausanne, Switzerland
e-mail: jonas.richiardi@epfl.ch

J. Richiardi
University of Geneva, Geneva, Switzerland

Keywords Graphical model · Information theory · Mutual information · Causal Bayesian networks · Model elicitation · Decision process

1 Introduction

Human beings experience their environment through their different senses. The convergence of this multisensory (MS) information enables them to cope with a complex world. This is only true, however, because of an efficient brain processing which turns these multiple and sometimes contradictory cues into a coherent percept (Knill and Richards 1996). Thus, at the neural level, MS enhancement is observed (see e.g., the model of Anastasio and Patton 2003). At the behavioral level, a percept impacted by the different MS stimuli arises if specific conditions, yielding cross-modal interactions, are met: the information are integrated. A well-known example of integration is the ventriloquism effect, where the perceived source of an auditory signal is shifted toward an unrelated visual source (e.g., a puppet moving mouth). The factors determining the cross-modal bias strength have been widely studied (see for example Welch and Warren 1980; Spence 2007; Lewald and Guski 2003; Andersen et al. 2004; Heron et al. 2004). The spatio-temporal proximity of the MS stimuli together with their relative perceptual reliability (determined by both the stimulus reliability and the modality appropriateness; Welch and Warren 1980; Andersen et al. 2004) are predominant.

Thus, integration can be understood as a statistically optimal mechanism where the brain takes advantage of the assumed redundancy in the MS information to increase the sensory estimate reliability (Ernst 2006). Different models supporting this view have been proposed, where the final percept is shown to follow a Maximum Likelihood Estimation (MLE) principle. This principle reduces—because of

normality and conditional independence assumptions—to a linear combination of the multiple available cues weighted according to their inversed variance (Ernst and Banks 2002; Battaglia et al. 2003; Alais and Burr 2004; Deneve and Pouget 2004; Andersen et al. 2005). The MLE models correctly predict how a stimulus stemming from one modality impacts the perception of another modality stimulus in the integrative case. However, they are only partial model of MS perception. Indeed, they fail to explain situations in which signals are not integrated. In these segregative situations, the impact of one stimulus on the perception of the other is null or very weak. Recently, some non-linear models accounting for both the integration and segregation have been proposed (Ernst and Bülthoff 2004; Shams et al. 2005; Roach et al. 2006; Körding et al. 2007; Sato and Toyoizumi 2007; Wozny et al. 2008). These models rely on a Bayesian probabilistic framework, but introduce either a non-uniform prior on the joint distribution of the two perceived stimuli, or a binary random variable (rv) weighting models for either a single or two perceived sources (Sato and Toyoizumi 2007; Körding et al. 2007). Generally speaking, these approaches build a model of the process leading to the final percept (referred to as *decision process* in the following) using a top-down scheme. A structure modeling the relationships between events such as the emission and perception of a stimulus is a priori hypothesized, based on expertise knowledge. Its validity is then tested on data collected through a dedicated experiment.

The singularity of our approach is to put the analysis of the *structure* of the decision process at the core, in a *data-driven* scheme. No a priori hypotheses are made about the structure of the model (i.e., about the variable relationships). The latter emerges from the data, as they are systematically analyzed through an information theoretical framework we propose. This framework takes place in the general theoretical framework of graphical models (see e.g., Lauritzen 1996) (though not tied to it) which affords a very expressive language for interpreting the phenomena under study. In a graphical model, the physical events are modeled through rvs, represented by nodes in the graph. The edges between nodes represent association. A model is completely defined once the graph structure—made of the nodes and edges—has been elicited and the probabilistic relationship it describes has been learned from some collected data. The present work precisely addresses the problem of determining the graph structure, i.e., the qualitative relationships between events such as the stimulus emission and perception, as a first objective. Our second objective is to propose a model of MS perception in its more general manifestation, i.e., accounting for both the integration and segregation *effects*, using this theoretical framework. Thus, we aim at going beyond the modeling task and to show how the systematic elaboration of the model, together with the use of this information theoretical framework we propose, afford for drawing robust conclusions about MS perception.

Wozny et al. (2008) recently modeled MS perception using graphical networks as well. However, the question of learning the model structure, which is the main point of our article, was not addressed. In particular, they were asking the subjects to report three judgments per trial but did not investigate the potential dependence between these three judgments. Hospedales and Vijayakumar (2008) present also a very interesting work where they tackle the problem of structure inference for MS perception in a temporal context. However, the framework they present is mainly designed to solve multi-target data association problems (speaker tracking). Instead, we intend to investigate MS perception and to draw robust conclusions about this phenomenon. The systematic and mathematical analysis of the MS perception task we propose allows putting in evidence the qualitative and quantitative relationships that exist between observable events and that underly the hidden perceptual stage. The appealing properties of the approach appear as we end up with robust evidences that subjects are actually reporting a combined MS percept and undergoing either integration or segregation effects. These effects arise despite our experimental framework is unconstrained. Indeed, we wanted to avoid any a priori hypothesis about the cognitive process behind MS perception but to investigate the phenomenon in its broader form (including both the integrative and segregative situations). Therefore, the subjects are asked for a single unisensory answer per trial, reported using a continuous value range.

This article starts with a presentation of some theoretical concepts attached to Bayesian networks (BNs), a family subset of graphical models. In this same section (Sect. 2), the information theoretical framework we propose to both build the model and get new insights about the data is presented. The first stage of the proposed approach consists in collecting data specific to the problem at hand. The experimental protocol we set up for audiovisual perception is described in Sect. 3, together with its results. The data are then investigated using the proposed theoretical framework and the graph structure of a consistent BN model is learned. This structure states the dependencies between rvs modeling the events (such as the emitted and perceived stimulus locations). This stage is presented in Sect. 4. Once the BN topology is known, the corresponding probabilistic relationships are learned and the complete model is used to perform inference (Sect. 5). Finally, the approach is discussed and perspectives are drawn in Sect. 6.

2 Theoretical approach

2.1 Bayesian network models in a nutshell

Multisensory perception can be understood as a particular example of causality induction, where the observer has to

take a decision about the cause (source) of the observations (perceived stimuli). A complete model should then consider how people learn the underlying structure of the causal decision process.

We propose an information theoretical approach, which takes place in the general framework provided by Bayesian networks, to build the model structure and to investigate MS perception through an unconstrained experimental setup (in order to observe MS perception in its more general form).

Bayesian networks (Pearl 1988) are directed acyclic graphs representing joint probability density functions (pdfs) over a set of rvs $\mathcal{V}$, which are related to the domain under study (domain variables). Nodes in the graph generally have a one-to-one correspondence to domain variables, and the graph edges encode conditional dependence between these variables. Local pdfs are attached to each variable in the network. They quantify the strength of the relationships depicted in the BN through its topology.

Building a BN can be done by a three-step procedure. The first step consists in eliciting the domain variables, i.e., in identifying the events involved in the process of interest and representing them by suitable rvs. In the second step, the topology of the BN is determined by analyzing the conditional independences between the rvs. This stage is at the core of the modeling process presented in this article. The mathematical analysis of the data, which aims at establishing the statistical relationships between the rvs, can bring out important data properties. In the present case, the causal structure of the decision process attached to MS perception will be established, asserting that the subjects' answer patterns are typical of either integration or segregation effects. Assuming the BN structure is fixed, the third step consists in determining its associated joint pdf. This can be done in a training phase where pdfs are learned for each node. They can then be used to perform inference, i.e., to compute the effects of observing certain variables on other nodes (Murphy 2002). These points will be detailed in Sect. 5. For now, let us introduce the information theoretical framework we propose to perform a mathematical analysis of the data and to cope with the BN structure learning stage.

2.2 From correlation to mutual information

In this modeling task, we focus on bringing to the fore the structure of the statistical relationships between the variables involved in the decision process, using a data-driven approach. The commonly used Pearson correlation coefficient would be ineffective since it only finds out linear dependence between variables. Instead, we advocate the use of mutual information (MI), as it is a measure of general dependence. The mutual information, *I(X;Y)* appraises the amount of information shared by two random variables X and Y:

$$I(X;Y) = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

$p(x)$, respectively $p(y)$, denotes the probability associated to the outcome x, respectively y, taken from the rv sample space Ω_X, respectively Ω_Y. By definition, $I(X;Y) = 0$ if and only if X is independent from Y ($X \perp\!\!\!\perp Y$). For easier interpretation and comparisons, a normalized version of $I(X;Y)$ that ranges from 0 ($X \perp\!\!\!\perp Y$) to 1 ($X \not\perp\!\!\!\perp Y$, where $\not\perp\!\!\!\perp$ denotes that X is non-independent from Y, in the limit of $Y = X$) is desirable. Strehl and Ghosh (2002) suggested to normalize the MI by the geometric mean:

$$NI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}, \tag{2}$$

where $H(X)$ and $H(Y)$ are Shannon's marginal entropies of X and Y, i.e., the average amount of uncertainty about X and Y. More details about mutual information and entropy can be found in (Cover and Thomas 1991; MacKay 2003). Once the pdfs associated to X and Y are known, MI can be straightly estimated from Eq. 1. We thereafter use non-parametric estimation methods like histogramming to estimate these pdfs. Therefore, some approximations are introduced and the MI between two independent rvs might not be strictly equal to zero. Thus, we will deem X and Y independent if $I(X;Y) \leq \epsilon$, with $\epsilon \in \mathbb{R}_+$ and close to zero (Richiardi 2007). This adjustable definition of independence is used in practical learning cases, for example the level of significance of a z-test can be adjusted (Druzdzel and Glymour 1995). In order to get the value to be set up for ϵ, we propose to build two independent rvs by generating random uniform pdfs on the histogram ranges for X and Y. The distance between the MI values of the true rvs and those obtained with these artificially generated independent variables will tell us whether the rvs X and Y can be reasonably considered to be independent or not.

2.3 Accounting for conditional relationships

Finding a dependence between two rvs is not enough to conclude about a cause–effect relationship. Some further analysis is required before stating causality between the variables, including a coherence with a possible temporal order (an effect cannot occur before the cause). Also, a third rv (a cause in this case) can explain the dependence observed between two rvs, or can make two independent rvs to become dependent. We will examine this possibility using conditional mutual information (CMI). The CMI assesses the independence relationships between the random variables X and Y given Z. It is defined as:

$$I(X;Y|Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z), \tag{3}$$

$$= I(X;Y) - I(X;Y;Z). \tag{4}$$

$H(X|Z)$, $H(Y|Z)$, and $H(X, Y|Z)$ denote the conditional entropies, i.e., the remaining uncertainty about X, Y, and $X \cup Y$ once Z has been observed. It must be noticed that the three term mutual information showing up in Eq. 4, $I(X; Y; Z)$, can be negative. Generally speaking, small values of $I(X; Y|Z)$ indicate that the knowledge of Z decreases the dependence between X and Y. At the limit, $I(X; Y|Z) = 0$ indicates that $X \perp\!\!\!\perp Y|Z$. Put in practice, CMI are compared to ϵ thresholds determined as for MI. We propose to normalize the CMI by the geometric mean to enable easier comparison. For three random variables X, Y, Z, such that $X \neq Z$ and $Y \neq Z$, this normalized conditional mutual information is defined as (Richiardi 2007)[1]:

$$NI(X; Y|Z) = \frac{I(X; Y|Z)}{\sqrt{H(X|Z)H(Y|Z)}}. \quad (5)$$

$NI(X; Y|Z)$ ranges from 0 ($X \perp\!\!\!\perp Y|Z$) to 1 ($X \not\perp\!\!\!\perp Y|Z$ in the limit of $X = Y$).

Establishing the independence and conditional independence relationships between the domain variables yields a dependency model. In turn, a *faithful* directed acyclic graph, which satisfies the independence assertions in the dependency model, and only those (Verma and Pearl 1992), may be created, but it is not always feasible. To this end, it is more informative to compare CMI with regard to the MI. Indeed, a CMI below the ϵ threshold may correspond to different independence relationships between the three involved rvs. In order to facilitate the interpretation of the results, we introduce the normalized difference ΔI_{XYZ} between the normalized MI and CMI of rvs X, Y, Z, as follows:

$$\Delta I_{XYZ} = [NI(X; Y) - NI(X; Y|Z)]/NI(X; Y). \quad (6)$$

2.4 Causal Bayesian networks

The dependency model can only identify network topologies up to Markov equivalence, the class of Markov-equivalent BNs being uniquely represented by a chain graph called an *essential graph* (Andersson et al. 1997). The number of candidate networks can be further reduced provided some of them form Markov chains. Indeed, the *data processing inequality* tells us that $I(X; Y) > I(X; Z)$ if the three rvs X, Y, Z form a Markov chain in that order $X \rightarrow Y \rightarrow Z$ (that is the conditional distribution of Z depends only on Y and is conditionally independent of X) (Cover and Thomas 1991). If the conditions for applying this theorem are not met, or if more than one candidate network topology remains afterward, we will resort to causal Bayesian networks. In this way, graphs that are not consistent with a causal interpretation can be removed from the candidate set.

[1] The restriction on X and Y avoid dividing by zero.

If observational data are available, it is possible to recover a causal graph if the faithfulness condition holds, sufficient data are available, and a certain minimum number of variables is present (Verma and Pearl 1992; Spirtes et al. 2001; Neapolitan 2004), but some counter that domain knowledge is needed (Robins and Wasserman 1999). In the remainder of this section we will follow the approach of Spirtes et al. (2001) and Neapolitan (2004), and effect a semantic change of the meaning of an edge between rvs. By putting an edge from rv X to rv Y only if X is a direct cause of Y with respect to the domain variables, the resulting directed acyclic graph becomes a causal directed acyclic graph or causal network (Neapolitan 2004).

Since domain knowledge is available in the present work, graphs that satisfy the dependency model but are not plausible causally because of the experimental protocol can be ruled out. An important caveat to mention is that we assume the domain variables are all properly identified, and no hidden variable exists which is the common cause of two domain variables. In this case, the domain variables are said to be causally sufficient (Neapolitan 2004). Domain knowledge is not always available or sufficient to resolve remaining ambiguities. In this case, or if we cannot preclude the existence of unobserved variables outside the set of the elicited domain variables, we must relax the causal faithfulness assumption. One way to do this is to assume *embedded* faithfulness (Neapolitan 2004). We assume the set of domain variables $\mathcal{V} \subseteq \mathcal{W}$, where $\mathcal{W}$ can contain unobserved variables. By using the embedded faithfulness assumption, it is often possible to recover graphs that are faithful to a probability distribution over a dataset where some data are unobserved.

2.5 Learning model parameters

The BN topology is learned by the procedure outlined above which, as such, forms the core framework of this work. The validity of the elicited causal structure must then be appraised, by learning the parameters of the associated conditional probability distributions and by testing the ability of the resulting model to perform inference.

The first step of the process is to choose a parametric form for the distribution. For discrete data, multinomial distributions are often used. Then, if the graph structure is known and the data are fully observed, a maximum likelihood approach is used to find the pdf parameters θ_{ijk} (Murphy 2002), where θ_{ijk} is the probability of observing the value k at node X_i given its parents Pa have value j ($\theta_{ijk} = P(X_i = k|Pa(X_i) = j)$). A K-fold cross-validation scheme is followed to learn the parameters and to perform inference (Theodoridis and Koutroumbas 2006). This way, no overlaps exist with the training set and over-fitting is avoided.

The proposed information theoretical framework will be used for robustly investigating the particular case of MS perception. First, data specific to the problem at hand have to be collected. The experiment we set up for this purpose is now described.

3 Data collection and analysis

3.1 Experimental protocol

The primary purpose of the experiment was to collect the data allowing for a robust mathematical analysis and modeling of audiovisual perception. The precise effect we wanted to explore was the impact a visual stimulus might have on the perception of a temporally coincident but possibly spatially discrepant acoustic stimulus, and vice versa. Since our objective was to investigate MS perception in its more general conditions, we designed an experimental framework as unconstrained as possible, i.e., where the subjects would not be compelled to use a predefine cognitive strategy to cope with the MS stimuli.

A group of ten subjects (seven men, three women) were exposed to audiovisual stimuli (buzzer beeps and diode flashes) following the protocol described hereafter. All the subjects had normal or corrected-to-normal vision and no auditory deficits. The subjects were seated in complete darkness, the eyes at the center of a curved screen with a radius of curvature of 57 cm. On this screen, nine equally distant red LEDs were aligned in the azimuthal eye plane, ranging from $-20°$ to $+20°$ by $5°$ steps. A buzzer was located just above this diode trail. The experimenter, located behind the subject, could move the buzzer circularly from $-20°$ to $+20°$ using a rotating device whose rotation axis was vertically centered on the subject eyes (see Fig. 1 for a schematic representation of this apparatus). The subject could not see the buzzer, being in complete darkness. No prior inspection of the visual nor the auditory setup was made available to the subjects.

The experiment was conducted in two phases, the acoustic and visual perception parts, in a switching order for the two subject halves. In the acoustic perception task, a 35-ms long acoustic stimulus (*primary* or *main stimulus*) was emitted at each trial, coming sometimes together with a visual stimulus (*secondary stimulus*), sometimes alone. The subject was asked to report which direction she heard the sound from. In the visual perception task, the primary and secondary stimuli were the visual and acoustic ones, respectively. The subject was asked to report where she had seen the flash. The main stimulus occurred randomly at $\pm 10°$, $\pm 5°$, or $0°$, and the possible secondary stimulus at $0°$, $\pm 5°$, or $\pm 10°$ from the main stimulus position. These spatial mismatches were chosen based on the experiments of Körding et al. (2007), where both the integration and segregation effects had been

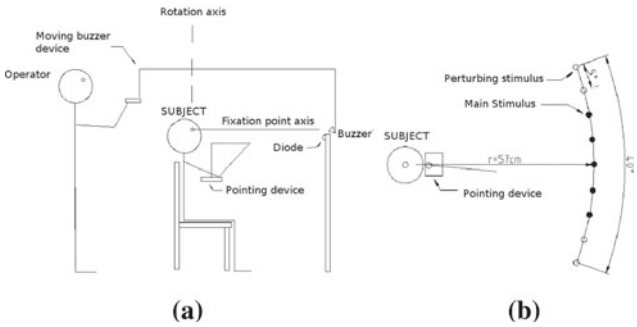

Fig. 1 Schematic views of the experimental design. **a** Side view, **b** Top view

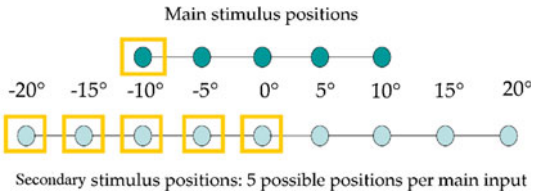

Fig. 2 Schematic representation of the stimulus positions with an example of possible bimodal stimuli

observed. In each task, 450 stimuli were presented to the subjects (15 occurrences of each possible combinations).

In order to report the perceived location of the main stimulus, the subject used a 43.5-cm long rotating pointer connected to a calibrated potentiometer. The pointer rotation axis was placed in front of the subject's mouth (so as to be very close to her eyes and ears). The subject held the tip end and moved it from a neutral position located at $40°$—the right stop position of the pointer—to the chosen position. She remained about 1 s in this position before coming back to the neutral position. The derivative of the report curve gave the subject's answer (last zero value before the second maximum of the derivative absolute value, i.e., before the subject's return to the neutral position). Schematic views of the experimental apparatus are displayed on Fig. 1, whereas Fig. 2 shows a possible bimodal input configuration.

The precise instructions given to the subjects were to localize the sound in the acoustic perception task, and the light in the visual perception task. They were informed that the acoustic stimulus might come with a visual stimulus in the acoustic perception task, and vice-versa in the visual perception task. Thus, the instructions clearly asked them to focus on the primary modality, but let them free to do whatever they wanted as far as the secondary modality was concerned. These instructions, together with the possibility of using continuous values to report their judgment, were given to ensure the so-called unconstrained feature of our experimental framework. Indeed, the subjects were not driven to use a pre-defined cognitive strategy to deal with the MS stimuli. Therefore, the insurance that subjects are undergoing MS

experiences is not provided by our experimental framework itself, but by the mathematical analysis we are performing later on (see Sect. 4).

3.2 Statistical statements

We are following a statistical approach to audiovisual perception modeling. Thus, it is necessary to cast the problem in a statistical framework. The acoustic and visual inputs of the system are modeled by two rvs A_τ and V_τ. $\tau \in \mathbb{N}$ is an indexing parameter that ranges from 1 to T, T being the number of trials (for any subject, $T = 450$ for each perception task). A_τ and V_τ are the spatial positions of the stimuli, drawn from discrete uniform distributions defined on the finite sets $\{0, \pm5, \pm10\}$ or $\{0, \pm5, \pm10, \pm15, \pm20, q\}$, depending on whether they stand for the primary or the secondary stimulus. q is a theoretical value which has to be assigned to the secondary stimulus in the unimodal case. In order to be true to life, we set up q to 180° in the visual perception task (behind the subject, thus not visible), and to −90° in the acoustic perception task.

The acoustic and visual outputs of the system (the values pointed by the subject) are modeled by two rvs $\hat{A}_\tau$ and $\hat{V}_\tau$, where a given value of either $\hat{A}_\tau$ or $\hat{V}_\tau$ models an answer to the current acoustic or visual trial input. These are spatial values defined on the finite range $[-40, 40]$ (these limits corresponding to the physical limitations of the experimental device). Notice that in the context of a causality analysis, the chronological distinction between the inputs and the outputs of the system is a crucial point to be taken into account. For the sake of the explanation, we drop off the indexing term in the notation from now on.

Since the pointer is taken back by the subject at a "neutral" position between two trials, we can assume the process to be stationary (the input or the output values of the system at a trial τ can be assumed to be independent from the system state at the trial $\tau - 1$). Finally, the problem domain is modeled by the set $\chi = \{A, V, \hat{A}, \hat{V}\}$, where each value taken on by one of the rvs in χ corresponds to a spatial position in degrees, at a trial τ, $\tau = 1, \ldots, T$.

A data instantiation vector $\mathbf{d}$ is a vector in which all the variables in χ have been assigned a value (Kontkanen et al. 1998). $D = (\mathbf{d}_1, \ldots, \mathbf{d}_T)$ is a set of T independently identically distributed (i.i.d.) data instantiations where each $\mathbf{d}_t$ is assumed to be sampled from the joint distribution $P(A, V, \hat{A}, \hat{V})$ of the variables in χ. If we consider the system to be subject-independent, i.e., the system outputs to be drawn from a same joint distribution $P(A, V, \hat{A}, \hat{V})$ whatever the subject $\{s_i\}_{i=1,\ldots,S}$, then a random sample is defined as $D = (\mathbf{d}_1, \ldots, \mathbf{d}_N)$, $N = T \times S$ i.i.d. data instantiation vectors. For the system's subject-independence assumption to hold, the inter-subject variability must be smoothed. A normalization step is performed where the subject's mean answers to unimodal inputs are removed. Thus, the subject's answers to unimodal inputs are viewed as a reference to which the multisensory experiment results are to be compared.

3.3 Distribution estimations

The data cannot be safely considered as normally distributed (Shapiro–Wilk two-sided test, significance level set to 0.05), advocating a non-parametric approach to the problem of estimating the pdfs. A straightforward and widespread non-parametric method is histogram estimation[2]. The bin width to estimate the pdfs of A and V, the input signals, is set to five and one bin is centered on each possible value of the ground truth (so that there are five bins in total for the primary stimuli, and ten bins for the secondary ones). This way, the ground truth pdfs are uniform. Moreover, the possible inaccuracy pertaining to the experimental design is taken into account. The output data range are covered by 15 bins. Thirteen bins of width 5 are centered on $\{0, \pm5, \pm10, .., \pm30\}$ and two larger bins are covering the bounding ranges $[-40, -32.5]$ and $[32.5, +40]$, where the data are very sparse (thus a trade-off is maintained between pdf estimate accuracy and overfitting).

3.4 Results analysis

After these preliminary developments, let us focus now on the data results per se. The mean and standard deviations of the system outputs (values pointed by the subjects) are shown in Fig. 3.

Visual dominance for spatial localization tasks is well known (Warren 1979). The results indeed show that the subjects' answers are less variable (i.e., more precise) when localizing visual stimuli than acoustic ones. Mean values of standard deviations[3] are equal to 7.5° in the acoustic case, and to 2.8° only in the visual one. The pointed value means are equal to zero in the unimodal case due to the normalization (see Sect. 3.2). The variability difference is also found when bimodal coinciding and non-coinciding stimuli are emitted. For the acoustic perception task, the answer standard deviations are equal to 5.8° and 9.5° for coincident and non-coincident inputs respectively, whereas they are equal to 3.2° and 3.0°, respectively, in the visual perception task. In the case of bimodal coincident inputs, comparable mean answer values are observed in the acoustic and visual perception tasks (−0.8° and −0.4°, respectively). In the case of non-coincident inputs, the subjects' answer accuracy decreases much

[2] Histogram-based estimation is equivalent to multinomial distribution parameter estimation as long as the pdf estimated by histogramming is the true but unknown distribution generating the sample (He and Meeden 1997; Scott and Sain 2005).

[3] Standard deviations and pointed value means are computed for the subjects' answers to each of the five primary stimuli positions before being averaged.

more in the acoustic than in the visual perception task (mean values equal to $-1.1°$ and $-0.2°$, respectively).

In the visual perception task, subjects segregate the acoustic and visual information instead of fusing it (i.e., no cross-modal bias is observed). Indeed, in the bimodal case, whatever the position of the irrelevant acoustic stimulus, it has no real impact on the subjects' localization of the primary visual stimulus.

In the acoustic perception task (see Fig. 3a), subjects' answers clearly undergo the effect of the visual stimuli. Indeed, the mean answer values follow the position of the irrelevant visual stimuli, though the subject is supposed to point toward the perceived acoustic stimulus location. Since the visual spatial localization is more precise than the acoustic one, the capture of audition by vision leads to a smaller dispersion and a better accuracy in bimodal coincident stimuli localization. The subjects then integrate the bimodal information in the acoustic perception case. It should be noticed, however, that, despite its decrease with visual coincident stimuli, the answer variance is still larger than in the visual perception task. This suggests that subjects are not simply ignoring the instructions and *localizing* directly the visual stimuli, as could be suspected due to our unconstrained experimental framework. A further mathematical analysis performed in Sect. 4 will robustly argue about this specific point.

This experiment has achieved its goal since we could observe and collect representative data about possible ways humans handle MS information. Strong cross-modal biases corresponding to integration and segregation effects are observed. These phenomena occur in either the acoustic or visual perception tasks, though we expected them to be present in both of them. Indeed, we used spatial mismatches similar to Körding et al. (2007). However, our experimental protocol is deliberately less constrained (unisensory judgment per trial, using a continuum of values) which certainly explain the weaker coupling found in the visual task, compared to Körding et al. (2007). Nevertheless, it will be interesting to apply our information theoretical framework to these data. We will thus be able to investigate the causal relationships associated to different strengths of cross-modal biases to MS perception, which was our primary goal. In particular, the acoustic perception task will challenge the model since unimodal or bimodal inputs yield different subjects' answer patterns.

4 Causal models of audiovisual perception

4.1 Mutual information analysis of the data

The analysis of the subjects' answers performed in Sect. 3.4 revealed different cross-modal bias patterns for the acoustic and visual perception tasks. These data will now be investigated through the mathematical framework presented in Sect. 2, in order to elicit the causal BN model structures associated to these cross-modal effects.

For an easier understanding, the following notation is used: the MI are labeled with a *"a"* or a *"v"* exponent depending whether they refer to the acoustic or the visual perception tasks. They also come with a subscript "*u*", "*c*", "*nc*", or "*a*" meaning that the data correspond to the unimodal, coincident, or non-coincident cases, or all three.

A fine analysis of the input signal normalized MI values is very interesting to start with. It nicely illustrates how MI gives cues about the dependence between two rvs. For unimodal signals (subject exposed to an acoustic or lone visual stimulus), the input rvs are obviously independent and we find logically $NI^a_{\mathrm{u}}(A,V) = NI^v_{\mathrm{u}}(V,A) = 0$. In the coincident cases, $NI^a_{\mathrm{c}}(A,V) = NI^v_{\mathrm{c}}(A,V) = 1$, that is, A and V are totally dependent. Indeed, A and V follow there the same uniform distribution since $A = V$. In the non-coincident case, we observe that $NI^a_{\mathrm{nc}}(A,V) = NI^v_{\mathrm{nc}}(V,A) = 0.39$. Actually, there exists in this case a certain dependence between the two input stimuli: having fixed the main stimulus position, the secondary stimulus is restricted to a given set of values ($\pm 10°$ or $\pm 5°$ away from the primary stimulus).

Let us now consider the cases involving the system outputs, i.e., the subjects' answers. The MI values indicate the tested rvs to be relatively dependent (values are all above their related ϵ thresholds, which are equal to 0.02 at most and estimated using independent rvs artificially generated as proposed in Sect. 2).

In the visual perception case, the high values of the MI values, especially in the unimodal case ($NI^v_{\mathrm{u}}(\hat{V};V)$ is about three times higher than $NI^a_{\mathrm{u}}(\hat{A};A)$), bear out the appropriateness of the sight sense in spatial localization tasks. This is a robust confirmation of what had been observed from the data variances in Sect. 3.4. Indeed, the normalized MI metric, restrained to the finite $[\epsilon, 1]$ range, allows for more effective interpretations and comparisons than the unbounded variance values. For the three unimodal, coinciding, and non-coinciding cases, the values of $NI^v(\hat{V};V)$ are very close (between 0.51 and 0.59), stressing the limited impact of the secondary acoustic stimuli on the subjects' answers. Actually, the latter are only weakly dependent upon the acoustic stimuli, $NI^v_{\mathrm{nc}}(\hat{V};A)$ and $NI^v_{\mathrm{a}}(\hat{V};A)$ being quite lower (0.21 and 0.15) than the MI between $\hat{V}$ and V (above 0.5).

In the acoustic perception case, the MI values are noticeably low, reflecting the inappropriateness of the hearing sense for spatially localizing targets. Visual stimuli spatially coincident with the acoustic stimuli lead to a 1.5 times higher MI between $\hat{A}$ and A than the value observed in the unimodal case[4] (0.34 for $NI^a_{\mathrm{c}}(\hat{A};A)$ instead of 0.21 for

[4] Notice that, of course, $NI^a_{\mathrm{c}}(\hat{A};A) = NI^a_{\mathrm{c}}(\hat{A};V)$

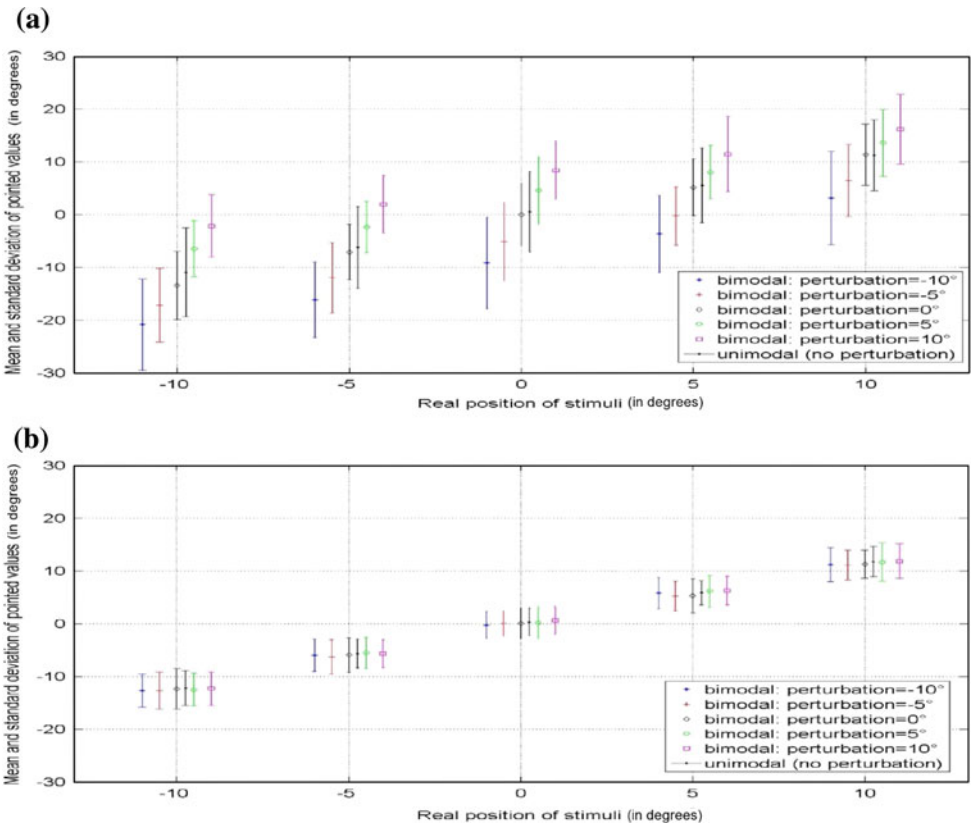

Fig. 3 Means and standard deviations of the values pointed by the subjects when localizing the acoustic stimuli (**a**) and the visual stimuli (**b**) in the unimodal, coinciding and non-coinciding cases. Values are grouped by six, a group standing for the unimodal plus the five bimodal inputs at a main stimulus position (X-axis). The values of the possible secondary stimuli mentioned in the legend have to be read as distances from the primary stimulus positions (i.e., the "0°" secondary stimulus is the coincident one)

$NI^a_{\mathrm{u}}(\hat{A}; A)$). On the contrary, non-coincident visual stimuli reduce this MI ($NI^a_{nc}(\hat{A}; A) = 0.14$). Meanwhile, the normalized MI between $\hat{A}$ and V is notably greater than $NI^a_{\mathrm{nc}}(\hat{A}; A)$ ($NI^a_{nc}(\hat{A}; V) = 0.29$). That is, the subjects' answer exhibit a stronger dependence with the secondary input than with the primary one in the non-coincident case (*vision captures sound*). The subject's answers $\hat{A}$ always show stronger dependence with the visual inputs V than with the acoustic stimuli A. However, it is essential to notice that, in all the cases, this dependence is weaker than between the subjects' answers in the visual perception task, $\hat{V}$, and the visual inputs V. This attests that, if the subjects are attracted by the visual stimuli, they do not *localize* it. Instead, they do report a combined percept, bearing out that they are undergoing MS experiences.

These results mathematically establish that secondary visual inputs impact the subjects' perception of the primary acoustic stimulus locations. That is, it ascertains that subjects integrate the MS information in their final percept, in the acoustic perception task. The MI analysis performed in the visual perception task establishes that subjects are experiencing MS perception in this task too, though the secondary acoustic stimuli weakly impact their judgment. Indeed, the simultaneous occurrence of a sound with a primary visual input *decreases* the dependence between this visual input and the output $\hat{V}$,($NI^v_{\mathrm{u}}(\hat{V}, V) = 0.59$, $NI^v_{\mathrm{nc}}(\hat{V}, V) = 0.54$, $NI^v_{\mathrm{c}}(\hat{V}, V) = 0.51$). It shows that, like in the acoustic perception case, the subjects do not report the same percept in the unimodal and in the bimodal cases.

Since dependences exist between the subjects' answers $\hat{A}$ and both the inputs A and V ($NI^a(\hat{A}, A)$ and $NI^a(\hat{A}, V)$ always greater than ϵ), we cannot decide at this stage whether the capture of the sound by vision is total or not. Similarly, the subjects' answers $\hat{V}$ and the secondary acoustic inputs A cannot be deemed independent as the MI values are all above the ϵ thresholds (and close to the values found in the acoustic perception task). A further CMI analysis is required to find out potential third variable effect (see Sect. 2.3) and to be able to robustly qualify these cross-modal interactions as either total integration or segregation phenomena. From a modeling point of view, this means that no edge can be removed between A, V, and $\hat{A}$ or $\hat{V}$ at this stage of the modeling process.

4.2 Conditional mutual information analysis of the data

The CMI asks the question of whether the knowledge of a third random variable Z makes two variables X and Y more or less dependent (see Sect. 2.3). This CMI analysis should alleviate the uncertainties pointed out in Sect. 4.1, i.e., gives means for determining whether the stressed cross-modal interactions stand for total integration or segregation effects in some cases.

Table 1 Normalized differences between MI and CMI, as defined by Eq. 6

Stimuli	Acoustic perception	Visual perception
	$\Delta I_{\hat{A}AV}(\%)$	$\Delta I_{\hat{V}VA}(\%)$
Unimodal	0	0
Coinciding	–	–
Non-coinciding	79	13
All	56	11
	$\Delta I_{\hat{A}VA}(\%)$	$\Delta I_{\hat{V}AV}(\%)$
Unimodal	–	–
Coinciding	–	–
Non-coinciding	17	95
All	29	93
	$\Delta I_{AV\hat{A}}(\%)$	$\Delta I_{VA\hat{V}}(\%)$
Unimodal	–	–
Coinciding	0	0
Non-coinciding	10	26
All	38	52

Missing results correspond to undefined normalized version of either MI or CMI

The normalized CMI are estimated using Eq. 5 and the ϵ thresholds are estimated as proposed in Sect. 2 (range of order of 0.05) . Some values are not defined since a condition for Eq. 5 to hold is that the variables of interest, X and Y, are not equal to the conditioning variable Z.

As stated in Sect. 2, it is more informative to compare the CMI with regard to the MI. In order to facilitate the interpretation of the results, we use the normalized difference defined in Eq. 6. The results are presented in Table 1. Broadly speaking, conditioning by a third rv decreases the dependence between the two former rvs whatever the considered perception task (acoustic or visual) and kind of inputs. In two cases only the dependence relationship remains unchanged: for coinciding stimuli, when focusing on the input rvs A and V, and for unimodal inputs when the rvs modeling the primary stimulus and the subject's answer are analyzed with or without conditioning on the secondary stimulus ($\Delta^{co}_{AV\hat{A}}$, $\Delta^{co}_{VA\hat{V}}$, $\Delta^{u}_{\hat{A}AV}$ and $\Delta^{u}_{\hat{V}VA}$ equal to 0%). Besides these two cases, conditioning reduces the dependence, with a decrease varying from case to case.

In the acoustic perception task, conditioning on the perturbing stimulus V makes the dependence between the primary stimulus A and the subject's answer $\hat{A}$ to drop off drastically in the non-coincident and in the "all stimuli" cases. For the latter, it decreases by half (56%), from 0.16 for $NI^a_a(\hat{A};A)$ to 0.07 for $NI^a_a(\hat{A};A|V)$. In the non-coinciding case, the decrease is of more than half the MI amount (79%), and leads to a CMI below the dependence threshold indicated by the theoretical values ($NI^a_{nc}(\hat{A};A|V) = 0.03 < \epsilon$, whereas $NI^a_{nc}(\hat{A};A) = 0.14 > \epsilon$).

In the visual perception case, the MI analysis revealed a weak yet non null (above the associated ϵ threshold) dependence between the subjects' judgments $\hat{V}$ and the secondary acoustic stimuli A. The CMI values point out that this dependence between $\hat{V}$ and A can be totally explained by V, the primary visual stimuli. At least, the potential information shared by A and $\hat{V}$ and unaccounted for by V cannot be distinguished from noise: $NI^v(\hat{V};A|V) < \epsilon$.

Therefore, the CMI analysis round out the MI analysis carried out in Sect. 4.1. This analysis mathematically established that subjects were not localizing directly the visual stimuli but an audiovisual percept in both the acoustic and visual perception tasks (since $NI^a(\hat{A};V) < NI^v_u(\hat{V};V)$ and $NI^v(\hat{V};V)$ being weaker for bimodal than for unimodal inputs). Now, the CMI analysis ascertains that the information are either partially or totally (when inputs are non-coincident) integrated in the acoustic perception task. It also bears out the segregation of the visual information from the acoustic one in the visual perception task.

4.3 Bayesian network structure learning

A global audiovisual perception model cannot be straightforwardly obtained from the current data. This would require a study of the $\hat{A}$ and $\hat{V}$ relationships but the co-occurrence of $\hat{A}$ and $\hat{V}$ at a single trial never happens since the subjects'answers are unisensory. Thus, it cannot be investigated with this experimental framework.

In order to uniquely orient the edges and obtain a causal model, we have to analyze both the conditional and causal dependence relationships between the variables. Thus, not all the edges remain after the CMI analysis contrary to the MI analysis: in Sect. 4.2, it has been shown that the dependence between A and $\hat{V}$ given V is below the noise level. Thus, accounting for a direct relationship between these two rvs would amount to overfit the data. The direct $A \rightarrow \hat{V}$ link must be removed. The independence statements entailed by the data lead to Markov equivalent graphs. Some of them such as $V \rightarrow A \rightarrow \hat{V}$ can be discarded since they do not hold in regards with the data processing inequality[5] (see Sect. 2). However, in the the visual perception condition for example, we are still left with the Markov equivalent graphs $A \leftarrow V \rightarrow \hat{V}$, $A \rightarrow V \rightarrow \hat{V}$, and $A \leftarrow V \leftarrow \hat{V}$: they encode the same conditional independence relationship $A \perp\!\!\!\perp \hat{V}|V$. Applying domain knowledge about experimental conditions, the temporal difference between the system inputs and outputs indicates that the last chain can be removed from the set of possible structures. Furthermore, the input variables $\{A, V\}$ are manipulated (subject to intervention due to the

[5] Notice that the unnormalized MI must be taken into account when talking about the data processing inequality.

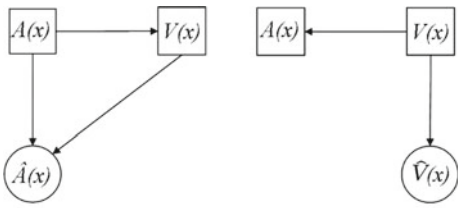

Fig. 4 Graphical models of the acoustic (**a**) and visual (**b**) perception conditions. The two chain graphs labeled respectively $\mathscr{M}_{\mathrm{a}}$ and $\mathscr{M}_{\mathrm{v}}$, represent Markov-equivalent PDAGs elicited from conditional independence assertions, except for the edge between the inputs which have been directed consistently with the experimental protocol. The pictorial convention given by Murphy (2002) is used, where discrete rvs are represented as rectangular nodes, and continuous rvs as round nodes. The potential information shared by A and $\hat{V}$ and unaccounted for by V is below the noise threshold ϵ, thus the direct $A \rightarrow \hat{V}$ link has been removed in $\mathscr{M}_{\mathrm{v}}$

experimental protocol) and do affect the distribution of $\hat{V}$, while the converse is not true.

Both the $A \rightarrow V$ and $V \rightarrow A$ subgraphs are now consistent with a causal interpretation. The model could be further refined by resorting to the embedded faithfulness assumption. Indeed, there is a superset of variables $\mathcal{W} = \{\mathcal{V}, P\}$ which satisfies it. The hidden P variable represents the experimental protocol, which dictates which values of V or A should be associated for the different subject tasks. Thus, the P variable could be input as common parent of both the A and V, resulting in a model $A \leftarrow P \rightarrow V \rightarrow \hat{V}$. However, it is of no particular interest to learn the probability distributions associated with the experimental protocol. It can also be argued that, in a visual perception task, V dictates A, while in an acoustic perception task, A dictates V. Thus, our choice is to avoid the embedded faithfulness assumption and orient the A–V edge from the rv modeling the primary stimulus to the rv modeling the secondary stimulus. Finally, the BNs shown on Figs. 4a, b and denoted respectively $\mathscr{M}_{\mathrm{a}}$ and $\mathscr{M}_{\mathrm{v}}$ in the remainder of this article, are proposed as topologies for the acoustic and visual perception.

5 Inference and performance

5.1 Model parameters learning

The main objective of this work was to use a data-driven method we proposed to mathematically investigate the cross-modal interactions inherent to MS perception, and in particular, the integration and segregation effects. As a result, the causal structures of the associated decision processes have been learned and modeled using BNs. We still have to demonstrate empirically that these BN topologies correspond to realistic models of acoustic and visual perception (as observed within our experimental framework). That is, we have to learn the joint distribution compactly expressed by these BNs third step of the model building process detailed in Sect. 2.1). Then, the goodness-of-fit of the model will be assessable by inference using the complete model.

The joint distributions $P_{\mathscr{M}_{\mathrm{a}}}$ and $P_{\mathscr{M}_{\mathrm{v}}}$ associated to the acoustic and the visual models $\mathscr{M}_{\mathrm{a}}$ and $\mathscr{M}_{\mathrm{v}}$ shown on Figs. 4a, b are given by Eqs. 9 and 12, respectively.

$$P_{\mathscr{M}_{\mathrm{a}}} = P(\hat{A}, A, V), \tag{7}$$
$$= P(\hat{A}|A, V)P(A, V), \tag{8}$$
$$= P(\hat{A}|A, V)P(V|A)P(A). \tag{9}$$
$$P_{\mathscr{M}_{\mathrm{v}}} = P(\hat{V}, V, A), \tag{10}$$
$$= P(\hat{V}|V, A)P(V, A), \tag{11}$$
$$= P(\hat{V}|V)P(A|V)P(V). \tag{12}$$

No simplifications can be made on the joint probability associated to the Bayesian network $\mathscr{M}_{\mathrm{a}}$, as its graph is complete (all variables are connected along an undirected cycle Margaritis 2003). On the contrary, the joint probability associated to $\mathscr{M}_{\mathrm{v}}$ can be simplified since $\hat{V} \perp\!\!\!\perp A|V$. This illustrates how BNs may easily lead to the simplification of joint pdf expressions.

Since the rvs do not follow a normal distribution, multinomial distributions are assumed ($\hat{A}$ and $\hat{V}$ are rounded to the nearest integer to make them discrete).

The answers from $N_T = 9$ subjects randomly picked up define the training sample, from which the parameters of the pdfs are learned (as described in Sect. 2.5). The answers of the left out subject make up the testing set.

5.2 Model performances

Once the parameters of the pdfs have been learned, inference can be performed. That is, we can compute the posterior distribution of the system outputs (subjects' answers) given the inputs. Inference is done on the testing set using a Maximum A posteriori (MAP) approach. MAP is defined as:

$$\hat{A}^* = \arg\max_{\hat{A}} P(\hat{A}|A, V)P(A, V), \tag{13}$$
$$\hat{V}^* = \arg\max_{\hat{V}} P(\hat{V}|V, A)P(V, A). \tag{14}$$

$\hat{A}^*$ and $\hat{V}^*$ are the optimal acoustic and visual predicates. Both the learning and inference stages have been implemented using the Bayes Net Toolbox for Matlab (Murphy 2002).

Examples of inference results are shown on Figs. 5 and 6, where the mean values of the Third subject's answers (test set) are plotted together with the MAP values predicted by the models $\mathscr{M}_a$ and $\mathscr{M}_{\mathrm{v}}$. The latter are trained on $N_T = 9$ subjects (all of the 10 subjects but the third one). The models predict well the integration and segregation phenomena

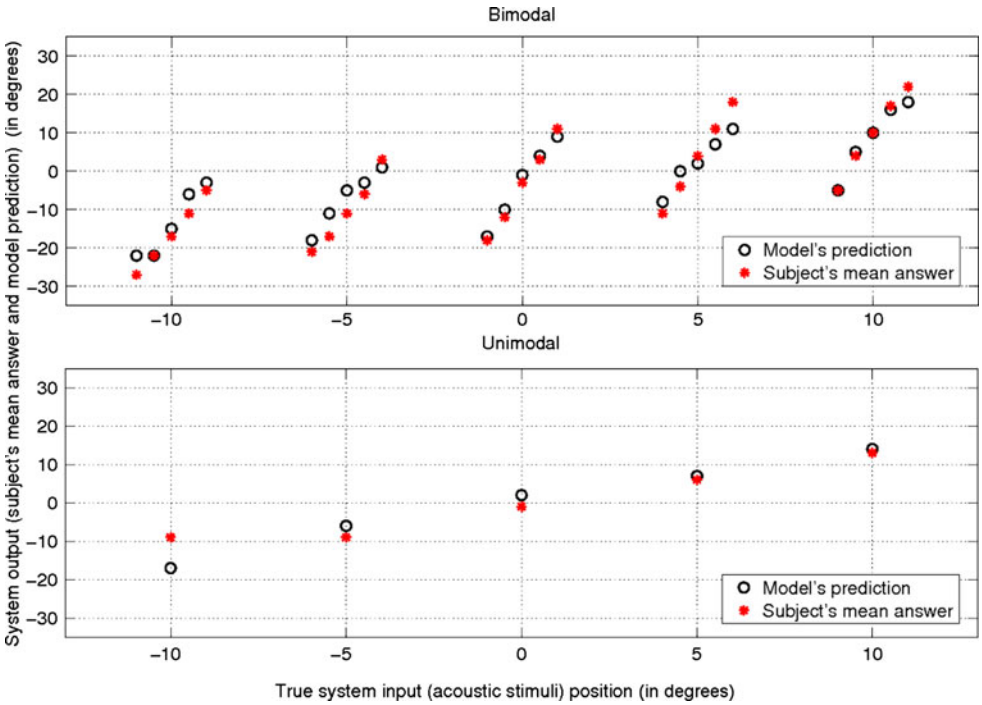

Fig. 5 Inference results for the acoustic perception task, in the case of bimodal inputs (*top*) and unimodal inputs (*bottom*). The figures show the mean values of the 3rd subject's answer (*test set*) together with the MAP values predicted by the $\mathscr{M}$ model trained on $N_T = 9$ subjects (all of the 10 subjects but the third one)

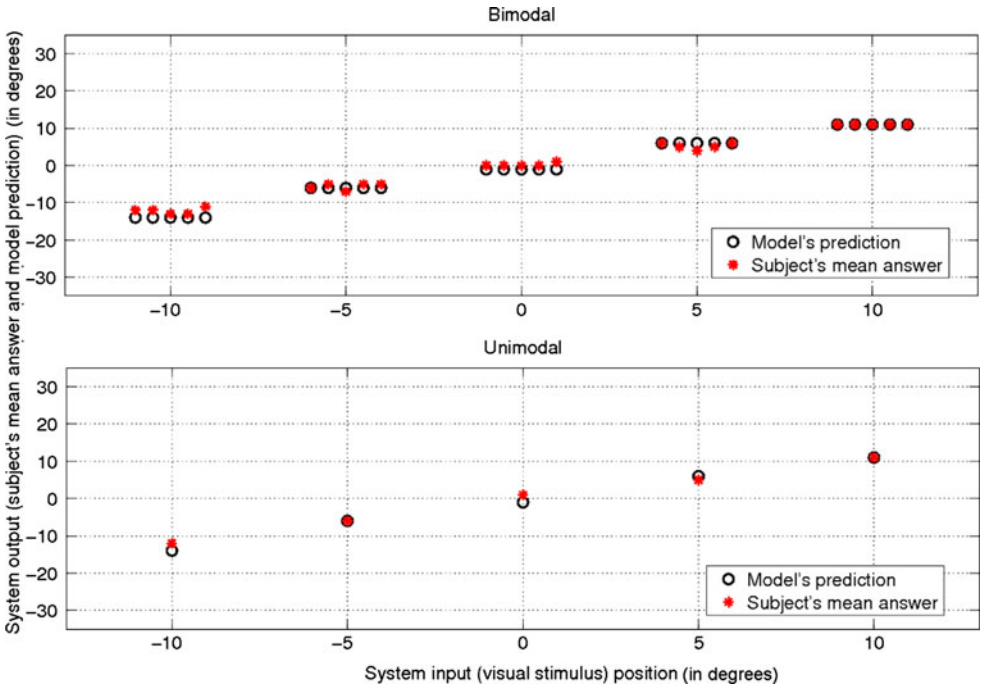

Fig. 6 Inference results for the visual perception task, in the case of bimodal inputs (*top*) and unimodal inputs *bottom*). The figures show the mean values of the 3^{rd} subject's answer (*test set*) together with the MAP values predicted by the $\mathscr{M}_{\mathrm{v}}$ model trained on $N_T = 9$ subjects (all of the 10 subjects but the third one)

observed in reality (see Sect. 3.4). Similarly to the true data, they are attracted by the visual inputs in the bimodal acoustic perception case, whereas they keep on following the visual inputs in the bimodal visual perception case. The outputs are aligned on the inputs in the unimodal acoustic and visual perception cases, as observed in reality. Thus, the acoustic perception model is able to predict both the integration and segregation effects, as was required.

The mean coefficient of determination r^2 (Sheskin 2004) (over the 10 time test procedure) between the MAP output by

the model and the mean subject's answer per main stimulus positions are equal to 0.87 and 0.98 in the acoustic and visual perception cases, respectively[6]. Obviously, acoustic perception is a more complex phenomenon, thus more challenging for the model, than is the visual one. These values stress the goodness-of-fit of the model, which is able to correctly predict unseen subject's answer to bimodal and unimodal inputs.

In order to complete the analysis of the model performances, the Akaike's information criterion (AIC; Akaike, 1974) is computed for the three possible—and sensible—model structures, in both the acoustic and visual perception cases:

- *structure 1*: fully connected model;
- *structure 2*: the link between the secondary input and the subject's answer variables is removed from the fully-connected model;
- *structure 3*: the link between the primary input and the subject's answer variables is removed from the fully-connected model.

The Akaike's criterion ranks competing models according to their performance and their complexity. Since we used multinomial distribution estimations, the number of model parameters is large (4000 parameters for the fully connected model) compared to the training set size. Therefore, the corrected AIC (AICc) is computed, as advocated by Burnham and Anderson (2002a). We used 40500 data, sampled from the learned joint pdfs of each possible model. The scores obtained for the three structures, in both the acoustic and visual perception cases, are presented in Table 2, as a difference with the best (i.e., the minimal) obtained AICc value. Burnham and Anderson (2002b) point out that models whose AICc scores exceed the minimal AICc value by at least 20 can be robustly rejected. According to these score differences, the models with structures 1 (for the acoustic perception task) and 2 (for the visual perception task) lead to the best results in term of performance versus complexity. Therefore, these results confirm the information theoretical analysis performed in Sect. 4 and the elicited models $\mathscr{M}_{\mathrm{a}}$ and $\mathscr{M}_{\mathrm{v}}$.

To be exhaustive, our model evaluation should comprise a comparison to previous models, specifically those where a full integration of the multisensory information is not assumed (Shams et al. 2005; Roach et al. 2006; Sato and Toyoizumi 2007; Körding et al. 2007; Wozny et al. 2008). However, we have already stressed that our approach was especially addressing the problem of eliciting the MS perception process structure in the context of an unconstrained experimental setup (to afford for an observation of the MS perception phenomenon in its more general manifestation). As a result, we do not end up with the same model variables: we have one output only and our input variables do not model hidden percepts but the inputted audiovisual stimuli. Also, we are using the observable events to investigate the structure of the cognitive decision process leading to the final MS percept, whereas the previously mentioned models intend to model the hidden perceptual stage. This major difference between these approaches and ours makes irrelevant any straight quantitative performance comparisons.

[6] There is one MAP prediction for each of the 30 possible combinations of primary and secondary stimuli, to be compared to the mean of the 5 subject's answers to each of these 30 possible inputs. The models are made of 400, respectively 4000, parameters in the visual, respectively acoustic, perception case.

6 Discussion and conclusion

Multimodal sensory signals provide human beings with information about their environment and own body. This multisensory perception can be understood as a *decision process*, whose implicit *structure* should be analyzed for better grasping the phenomenon. It is the singularity of this work to put the analysis of the structure of the decision process at the core: no a priori hypotheses are made about the model causal structure. Instead, the latter is learned from a systematic data analysis. Thus, the first and main objective of this work was to propose a theoretical data-driven framework to investigate cognitive tasks such as MS perception and to learn the causal structure of the associated decision process. The second objective was to propose a model of MS perception in its more general manifestation, i.e., accounting for both the integration and segregation *effects*, using this theoretical framework.

In our approach, the systematic analysis of the implicit MS perception process starts from the explicit physical events. In order to acquire data representative of MS perception in its more general manifestation, an unconstrained experimental setup was built where the subjects were not compelled to use a pre-oriented cognitive strategy when coping with the MS stimuli. The subjects were asked to report the location of the primary stimulus location, using a continuous range of values. Notice that the single unisensory judgment per trial and the release of the forced choice certainly explain the difference between our results and those of Körding et al. (2007), though similar spatial mismatches between stimuli were used. An informal analysis of the subjects' answers is performed in a first time and points out different cross-modal bias strengths in either the acoustic or the visual perception tasks, corresponding to integration or segregation answer patterns. The information theoretical framework we propose is then applied to the data, with a twofold objective. First, it intends to establish that the subjects are reporting a MS percept (since our unconstrained experimental framework gives us no direct support about this point). Second, it aims at formally learn the causal structures associated to these different

Table 2 AICc differences with the best (i.e., the minimal) AICc score, for each of the three sensible candidate structures associated to both the acoustic and visual perception tasks

	Structure 1	Structure 2	Structure 3
Acoustic perception task	0	30242	9486
Visual perception task	2853	0	59539

Structure 1 fully connected model; *Structure 2* link between the secondary input and the subject's answer variables removed from the fully connected model; *Structure 3* link between the primary input and the subject's answer variables removed from the fully connected model

cross-modal biases, for MS perception models accounting for both the integration and segregation effects to be built.

The MI analysis carried out on the data proves that, in any of the bimodal experimental situations, subjects are reporting a combined MS percept. Of course, visual inputs strongly impact the subjects' percepts, as could be expected in a localization task. However, the MI values between the subjects' answers $\hat{A}$ and $\hat{V}$ and the visual input V (standing as primary or secondary stimulus) is never as high in case of bimodal inputs as it is in the case of unimodal visual inputs (even in the stronger "capture" situation, we still have $NI^a_{\text{nc}}(\hat{A};V) < NI^v_{\text{u}}(\hat{V};V)$). In other words, the dependence—or impact—between the visual stimuli and the subjects' answers is weaker in case of bimodal inputs, especially when subjects are asked to localize an acoustic stimulus. This establishes that the subjects are not simply localizing the visual stimuli, whatever the instructions they received, but are indeed reporting a *multisensory* percept.

The CMI analysis we propose allows for building and modeling the causal relationship structures associated to the cross-modal biases observed in the visual and acoustic perception tasks. As a result, the segregative and integrative nature of these interactions is formally established. Except in the obvious case of unimodal acoustic inputs, the visual information is shown to fully explain the dependence observed between the subjects' answers and the acoustic inputs (vision captures sound). This point is not in contradiction with the previous assertion, which simply states that the subjects were experiencing a MS stimulation, without concluding about the effect of this MS stimulus nature on the final percept. Stated differently, the MI analysis establishes that the subject's perception takes into account the MS nature of the stimuli, whereas the CMI analysis evaluate the strength of each stimulus on the perception. Hence, these two complementary assertions validate the two widely observed *effects* inherent to *multisensory* perception, namely, *integration* and *segregation*, and establish the associated model causal structures.

Finally, the models, built in a systematic, data-driven way using the mathematical approach we propose, are shown to properly predict the integrative and segregative effects. In particular, the acoustic perception model is able to predict the two different answer patterns associated to unimodal acoustic inputs and to non-coincident audiovisual inputs, where the subjects' answers are driven either by the acoustic or by the visual inputs.

This work constitutes a basis for further exploration of MS perception. In the future, we intend to use the proposed framework to investigate the impact of different factors on MS perception, such as temporal delays for example Welch and Warren (1980) and Andersen et al. (2004). Equalizing the noise levels in both modalities, as is done in Roach et al. (2006), should also lead to very interesting results. In particular, we would expect to observe segregation not only in the visual perception case but in the acoustic perception case as well.

Generally speaking, low to high-level models can be developed using BN's. Hence, a further step should be to make explicit other factors in future models, such as the noise level for example, or a hidden "subject" rv to investigate possible subject-specific effects. Finally, it would be worth designing new experiments to finely investigate the causal structure of the MS perception process by manipulating the perceived relationship between visual and auditory stimuli.

Acknowledgments The authors would like to thank F. Buloup for the electronic support, G. Gauthier, E. Daucet, and C. Besson for fruitful discussions and A. Donneaud for his help in building the experimental setup. Thanks also to all the subjects who kindly accepted to take part in the experiment.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723

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

Anastasio TJ, Patton PE (2003) A two-stage unsupervised learning algorithm reproduces multisensory enhancement in a neural network model of the corticotectal system. J Neurosci 23(17):6713–6727

Andersen TS, Tiippana K, Sams M (2004) Factors influencing audiovisual fission and fusion illusions. Cogn Brain Res 21:301–308

Andersen TS, Tiippana K, Mikko S (2005) Maximum likelihood integration of rapid flashes and beeps. Neurosci Lett 380:155–160

Andersson SA, Madigan D, Perlman MD (1997) A characterization of Markov equivalence classes for acyclic digraphs. Ann Stat 25(2):505–541

Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A 20(7):1391–1397

Burnham KP, Anderson DR (2002a) Model selection and multimodel inference: a practical information-theoretic approach, Chap 2.4, 2nd edn. Springer, New York, pp 66–67

Burnham KP, Anderson DR (2002b) Model selection and multimodel inference: a practical information-theoretic approach, Chap 2.6, 2nd edn. Springer, New York, pp 70–72

Cover TM, Thomas JA (1991) Elements of information theory. Wiley, New York, USA

Deneve S, Pouget A (2004) Bayesian multisensory integration and cross-modal spatial links. J Physiol Paris 98(1-3):249–258

Druzdzel MJ, Glymour C (1995) What do college ranking data tell us about student retention: causal discovery in action. In: Proceedings of 4th workshop on intelligent information systems. IPI PAN Press, Augustow, Poland, pp 1–10

Ernst MO (2006) A Bayesian view on multimodal cue integration. In: Human body perception from the inside out, Chap 6. Oxford University Press, New York, pp 105–131

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415(6870):429–433

Ernst MO, Bülthoff HH (2004) Merging the senses into a robust percept. TRENDS Cogn Sci 8(4):162–169

He K, Meeden G (1997) Selecting the number of bins in a histogram: a decision theoretic approach. J Stat Plan Inference 61(1):49–59

Heron J, Whitake D, McGraw PV (2004) Sensory uncertainty governs the extent of audio-visual interaction. Vis Res 44:2875–2884

Hospedales TM, Vijayakumar S (2008) Structure inference for bayesian multisensory scene understanding. IEEE Trans Pattern Anal Mach Intell (PAMI) 30(12):2140–2157

Knill DC, Richards W (eds) (1996) Perception as Bayesian inference. Cambridge University Press, New York

Kontkanen P, Myllymäki P, Silander T, Tirri H, Grünwald P (1998) Bayesian and information-theoretic priors for bayesian network parameters. In: Proceedings on 10th European conference on machine learning, vol 1398. Springer-Verlag, Chemnitz, Germany, pp 89–94

Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS ONE 2(9):e943

Lauritzen SL (1996) Graphical models. Oxford University Press, New York,USA

Lewald J, Guski R (2003) Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Cogn Brain Res 16:468–478

MacKay DJ (2003) Information theory, inference, and learning algorithms. Cambridge University Press, New York, USA

Margaritis D (2003) Learning bayesian network model structure from data. Phd thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Murphy KP (2002) Dynamic bayesian networks: representation, inference and learning. Phd thesis, University of California, Berkeley, USA

Neapolitan RE (2004) Learning Bayesian networks. Prentice Hall, Upper Saddle River, NJ

Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan kaufmann, San Francisco, CA

Richiardi J (2007) Probabilistic models for multi-classifier biometric authentication using quality measures. These no. 3954, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Roach NW, Heron J, McGraw PV (2006) Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. In: Proceedings of the Royal Society B: biological sciences, vol 273. Royal Society of London, London, UK, pp 2159–2168

Robins JM, Wasserman L (1999) On the impossibility of inferring causation from association without background knowledge. In: Computation, causation, and discovery. MIT Press, Cambridge, MA, USA, pp 305–321

Sato Y, Toyoizumi T (2007) Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural Comput 19:3335–3355

Scott DW, Sain SR (2005) Multi-dimensional density estimation. In: Data mining and computational statistics, Handbook of statistics, Chap 9, vol 23. Elsevier, Amsterdam, pp 229–262

Shams L, Ma WJ, Beierholm U (2005) Sound-induced flash illusion as an optimal percept. NeuroReport 16(17):1923–1927

Sheskin DJ (2004) Handbook of parametric and nonparametric statistical procedures, 3rd edn. CRC press, New York, NY, USA

Spence C (2007) Audiovisual multisensory integration. Acoust Sci Technol 28(2):61–70

Spirtes P, Glymour C, Scheines R (2001) Causation, prediction, and search. MIT Press, Cambridge, MA, USA

Strehl A, Ghosh J (2002) Cluster ensembles—a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617

Theodoridis S, Koutroumbas K (2006) Pattern recognition. Academic Press, Orlando, FL, USA

Verma T, Pearl J (1992) An algorithm for deciding if a set of observed independencies has a causal explanation. In: Proceedings of 8th annual conference on uncertainty in artificial intelligence (UAI-92). Morgan Kaufmann, San Mateo, CA, pp 323–333

Warren DH (1979) Spatial localization under conflict conditions: is there a single explanation?. Perception 8(3):323–337

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88(3):638–667

Wozny DR, Beierholm UR, Shams L (2008) Human trimodal perception follows optimal statistical inference. J Vis 8(3):1–11