scientific reports

Check for updates

OPEN

Directed attention influences optimality of top-down and bottom-up multi-modal perceptual integration

Clement Abbatecola[1]✉, Henry Kennedy[2] & Kenneth Knoblauch[2]✉

Multimodal sensory integration is a ubiquitous neural process that can be modeled as optimal cue combination, incorporating both top-down, attention-like signals and bottom-up evidence that impact the precision of response variables. Accordingly, reducing attention or adding noise to one modality is expected to decrease proportionally its contribution while increasing that of the other modality. We tested this prediction using a gender-comparison task employing stimuli for which the face and voice were independently morphed between average male and female exemplars. Top-down influences were manipulated by having observers judge the stimuli with respect to either one or both modalities. Bottom-up influences were manipulated by introducing independent and varying amounts of visual and auditory noise. The contributions of each modality were estimated by maximum likelihood within a signal detection model of the decision process. As expected, if the attended modality was degraded by noise, the contribution of the unattended modality increased in compensation. Contrary to prediction, however, noise in the unattended modality had no impact on the attended modality. The results signal a departure from an optimal cue combination rule and are relevant to theories of predictive processes and observations in bimodal learning in modality-specific agnosia (prosopagnosia, phonagnosia).

Keywords Face and voice gender perception, Attention, Precision, Predictive coding, Prosopagnosia, Phonagnosia

Multimodal sensory integration is a ubiquitous feature of brain function[1] that is thought to play a predictive role in perception[2]. Categorizing and explaining bottom-up and top-down effects is a major issue in the field of multimodal integration[3,4] that can be modeled as multimodal cue combination incorporating both top-down, precision or attention-like signals and bottom-up evidence[5–7]. As a particular instantiation, predictive coding models of cognition incorporate focused attention as an internal precision signal[8].

Face-voice interactions in perception have provided a particularly fertile paradigm for studying such interactions[9]. For example, effects of noise and context have been shown to influence the Ventriloquist and McGurk Effects, respectively[10,11]. While face-voice properties can be varied and integrated in the perception of a number of features, such as identity or emotion, gender has proven ideal because it can be defined along a single perceptual dimension (varying between masculine and feminine), simplifying analyses of responses. Technically, it is possible to generate a continuous physical variation between male and female exemplars via morphing of auditory and visual stimuli[12,13]. Previous studies have found that incongruency in the gender of face and voice cues can interfere with identification of the voice gender[14] or both[13]. We previously reported a 40-fold variation in the relative weights of visual and auditory modalities to judging the gender of audio-visual stimuli with independent morphing of the face and voice gender as the top-down demands of the task were varied by focusing on the visual, auditory or both modalities[15].

In this study, we examine the influence of noise, a bottom-up source of variation, on the relative contributions of face and voice cues to gender comparisons as the top-down demands are varied, using Maximum Likelihood Conjoint Measurement (MLCM)[15–18]. The method is based on paired comparisons between stimuli varying independently along two or more dimensions, here face and voice gender (Fig. 1a). On each trial, observers

[1]Centre for Cognitive Neuroimaging, School of Psychology and Neuroscience, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QB, UK. [2]Stem Cell and Brain Research Institute, Inserm U1208, University of Lyon, Université Claude Bernard Lyon 1, Lyon, France. ✉email: clement.abbatecola@glasgow.ac.uk; ken.knoblauch@inserm.fr

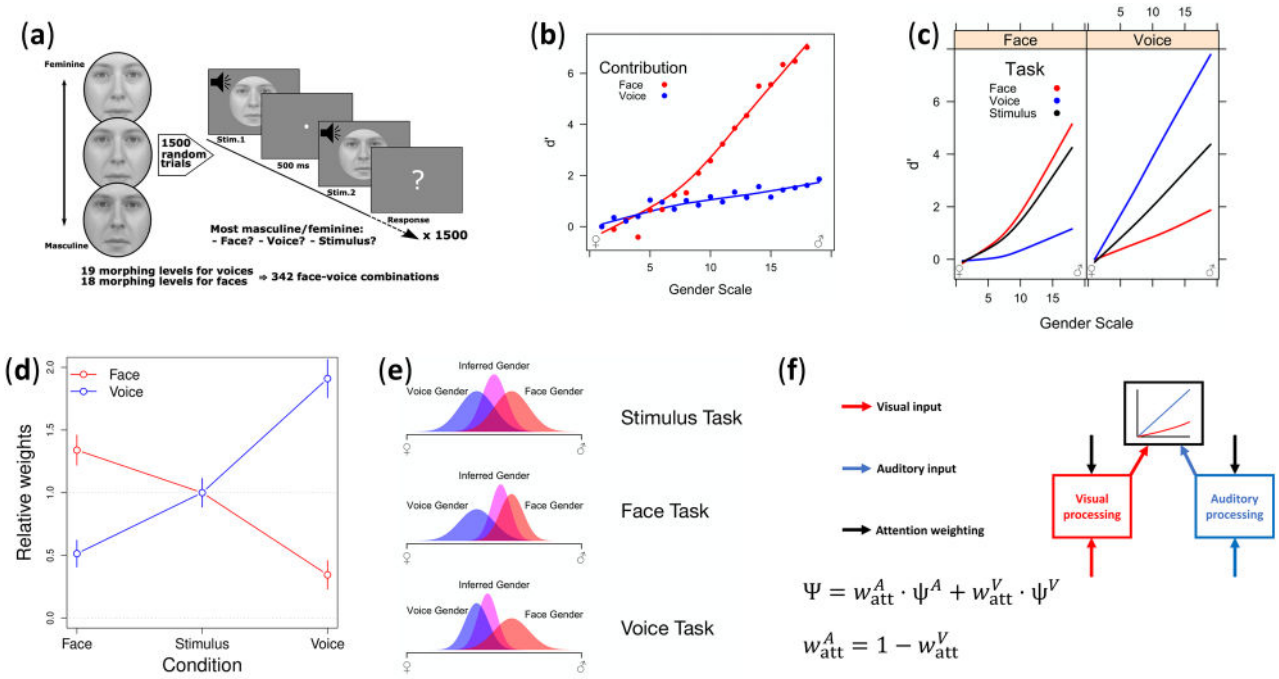

Fig. 1. MLCM task description, typical results and optimal cue combination model. (**a**) Conjoint measurement protocol. Pairs of face-voice video sequences with independently varying levels of face (3 examples shown on the left) and voice gender morphing were judged by observers according to: (1) face gender, (2) voice gender, or (3) stimulus gender. (**b**) Additive contributions of the face (red) and voice (blue) as a function of gender level for each modality when comparing stimuli face gender for one observer. The points are the estimates from a generalized linear model, the curves from a generalized additive model. (**c**) Average contributions of the face (left) and voice (right) for each of three tasks, replotted from Abbatecola et al.[15]. (**d**) The relative change in amplitude of the contribution of face (red) and voice (blue) for each of the three tasks with 95% confidence intervals. (**e**) Schematic illustration of the distribution of internal responses to face (red) and voice (blue) cues and the optimally combined response (purple) for each of the three tasks. (**f**) Feedback model of combination of precision estimates to influence weighted cue combinations of gender judgements. Feedback (black arrows) influences the weights in the contributions of visual and auditory signals in the decision process.

readily judge which of the pair is more masculine (or feminine) using the face, voice or both cues conjointly[15]. Pairs are chosen from a large set of stimuli where the apparent gender of the face and voice are independently modulated through morphing of combinations of average male and female exemplars. Repeated presentations allow estimation of the contributions of each modality to the subject's judgements[19] and the accurate recovery of the functions that encoded the stimuli[20].

For example, Fig. 1b shows the estimated contributions from each modality to the judgements obtained from one subject when the task was to choose the more masculine face. Intuitively, for a fixed gender difference between the faces of the first and second stimuli, if the probability of choosing the first face as more masculine is unaffected as the voice gender is independently varied, then we can conclude that the voice cues do not influence the face judgements, and vice versa for voice judgements. On the contrary, if the probability does change, then the face judgements are influenced by the voice cues. The contributions of both cues to the judgements that best predict the ensemble of judgements are estimated by maximum likelihood within a signal detection model of the decision process that occurs on a supramodal gender scale, (Eqs. (1) and (2) in Methods)[17]. In Fig. 1b, an additive model of the cue contributions (Eq. (3)) shows that the face judgements were dominated by a contribution from the face (red) with a smaller but significant contribution of the voice (blue). When the task was changed to judge the stimuli based on the voice or both modalities, the shape of the face and voice curves remained similar but the relative contributions of the two modalities to the judgements covaried systematically (Fig. 1c)[15]. Thus, we can assess the relative change in contributions by simply tracking the change in amplitude of each curve across tasks (Fig. 1d).

These results are best described by a simple cue-combination model (Fig. 1e) where the internal responses from the visual and auditory modalities are distributed along a supramodal gender response continuum, varying from female to male. The uncertainty of response from each modality is represented as a distribution along the continuum. In the situation depicted, there is a conflict between the gender of the face (red) and voice (blue) stimuli. The gender is inferred by an optimal combination rule (purple)[8,21] under which the distribution with higher precision pulls the inferred gender in its direction and is interpreted as a change in weighting induced by feedback connectivity (Eqs. (5) and (6)), Fig. 1f).

This paradigm provides a powerful method to investigate perceptual mechanisms involving the integration of top-down and bottom-up signals. The precision of each distribution can be manipulated in a graded fashion by adding noise to one or both modalities. For example, when judging the face, adding visual noise (Fig. 2a) increases uncertainty and a change in the top-down precision signal of the face response distribution (Fig. 2b right noisy attended modality, black solid curves) but has little effect on the voice distribution, resulting in a shift of the inferred distribution toward the unattended voice distribution (dashed curve). If, instead, noise is added to the unattended voice modality, the precision of the voice distribution is reduced while the face distribution is unchanged, resulting in a shift of the inferred distribution toward the attended face distribution (Fig. 2b left noisy unattended modality). More generally, according to optimal combination theory[5,6], graded addition of noise in the attended or unattended modalities is predicted to result in relative changes in the weights attached to each modality in opposite directions. Addition of noise to both modalities reduces the weights for both modalities (Fig. 2c for an illustration of the model and possible cognitive mechanisms).

Here, we test the hypothesis that top-down, focused attention is exchangeable in its influence on bottom-up manipulation of precision, i.e., the weighting model where focusing attention to the face has an equivalent effect to increasing precision in (or removing noise from) the face signal or decreasing precision in (or adding noise to) the voice signal. Although, in this model, the influences are independent, it is conceivable that they interact. If this is the case, what are the experimental conditions that control such an interaction?

To the extent that modality specific perceptual deficits such as prospagnosia and phonagnosia reflect imbalances in the contributions of visual and auditory channels[22], our manipulation of bottom-up signal quality may reproduce some of the features of this perceptual experience in a general population, leading to a better understanding of the compensation mechanisms involved in these deficits (such as redirecting attention to one modality when the other is unreliable) and potential therapeutic strategies.

Results

Simulated observers results

In simulations of the MLCM task, noise in the attended modality (Fig. 3a) reduced the contribution of that modality to judgements (shift from red dashed to solid curve) and increased the contribution of the unattended modality (shift from blue dashed to solid curve) ; the reverse effect occurred when adding noise to the unattended modality (Fig. 3b, Eq. (7)).

Similar to the effect of shifting attention, if we assume response curves of invariant shape[15], we can model the relative change in contributions induced by progressively adding noise in either or both modalities by simply tracking the change in amplitude of each curve across tasks. As expected, adding noise to the attended modality,

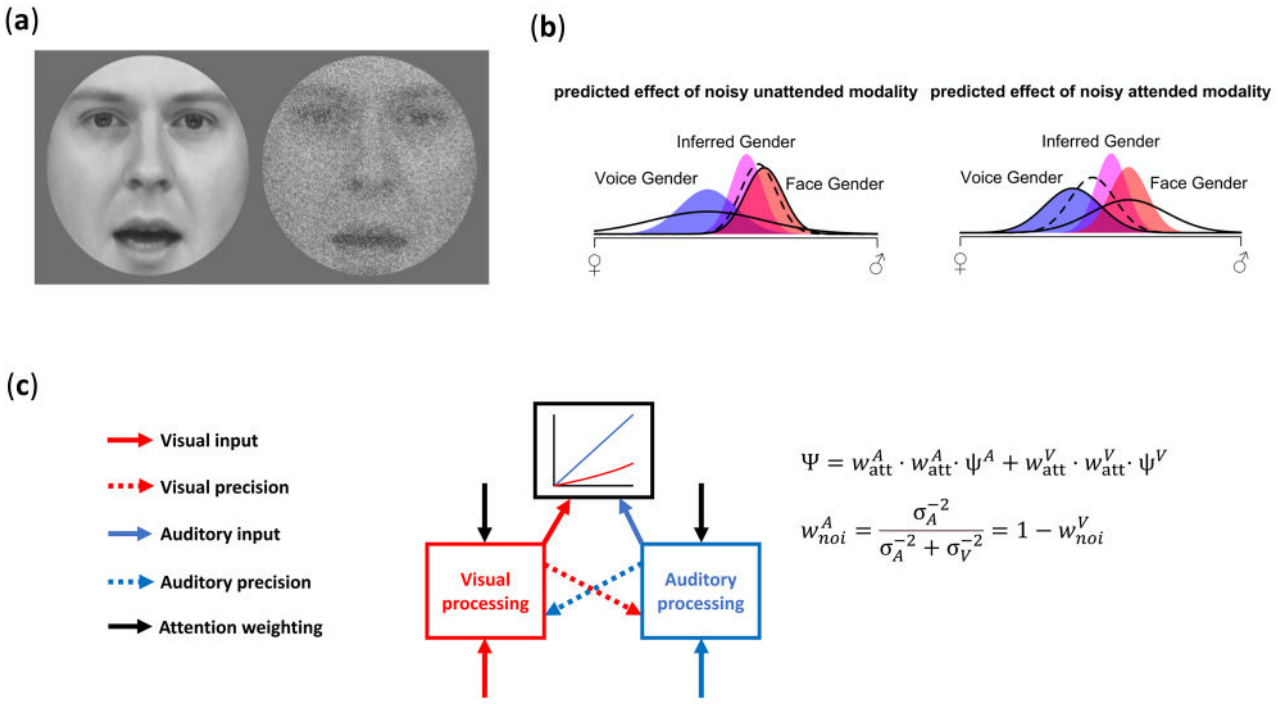

Fig. 2. Expected effects of noise on MLCM. (**a**) Example of Face stimuli without (left) and with (right) added noise. (**b**) Schematic illustration of the distribution of internal responses to face (red) and voice (blue) cues when noise is added to the unattended (left) and attended (right) modality (dashed curve) and without noise (purple) on the optimally combined responses for the Face task. (**c**) Feedback model of combination of precision estimates to influence weighted cue combinations of gender judgements. Noise (dotted arrows) and feedback from attention (black arrows) both influence the weights in the contributions of visual and auditory signals in the decision process.

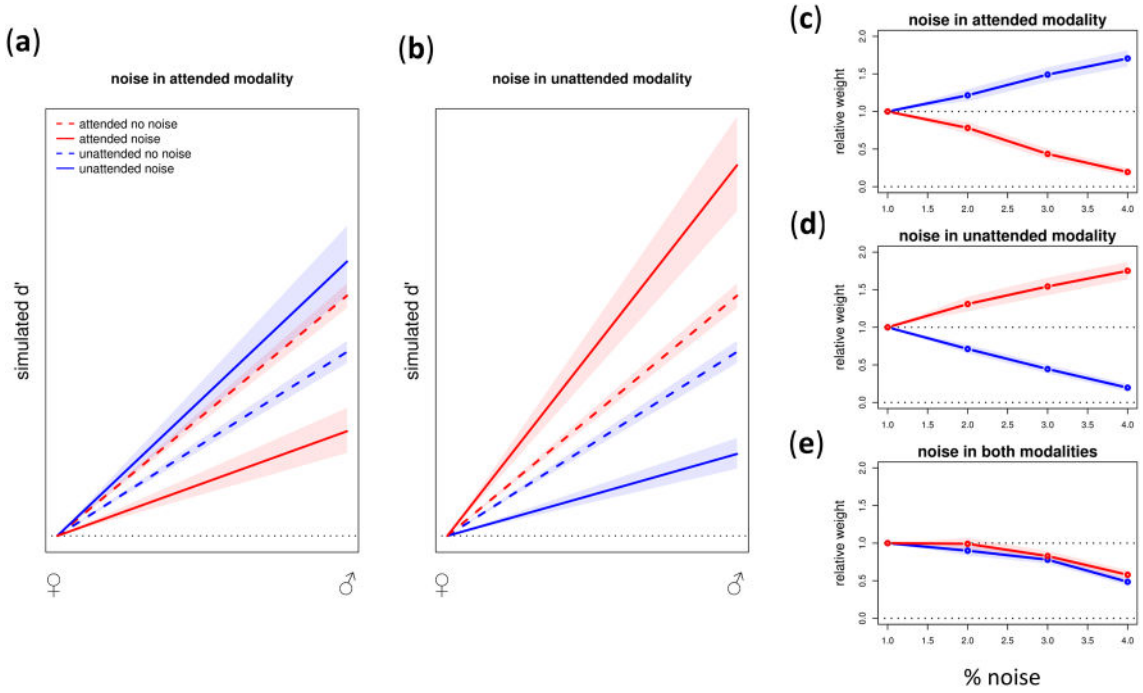

Fig. 3. Results from simulations (**a**, **b**) Model predictions for change of contribution of attended (red) and unattended (blue) modalities when adding noise to the attended (**a**) vs unattended (**b**) modality (no noise = dotted lines vs noise = solid lines). (**c**, **d**, **e**) Model predictions for relative change of contribution of attended (red) and unattended (blue) modalities as compared to low noise when progressively adding noise to the attended (**c**) unattended (**d**) or both modalities (**e**). Shaded envelopes indicate 95% confidence intervals over 25 simulations.

reduced its contribution to judgements with a concomitant increase in the unattended contribution (Fig. 3c). In symmetric fashion, adding noise to the unattended modality decreased its contribution to judgements while the attended contribution increased (Fig. 3d). Finally, adding noise to both modalities resulted in a similar decrease in contribution of both modalities to the judgements (Fig. 3e).

Experimental results

We acquired data across attention and noise conditions in which the attention of the participants could be focused on either or both modalities, and noise could be applied to either or both modalities. All noise/attention combinations were tested.

The conventional tool for visualizing raw data from a conjoint measurement experiment is the conjoint proportions plot (CPP)[16,17,23,24]. CPPs define an upper triangular heatmap in which the coordinates of each pixel represent a combination of visual and auditory gender levels for a pair of stimuli, i.e., a possible trial. Grey levels indicate the proportions of choices the participants made for each trial. Grouped CPPs for each task and individual plots can be found in Supplementary section 2. Figure 4 illustrates the results of adding increasing amounts of auditory noise (Fig. 4c, d, e, f) and visual noise (Fig. 4c, g, h, i) during the Face task as CPPs, compared to simulated ideal observers who made choices simply on the basis of the ordering of the visual (Fig. 4a) or the auditory gender levels (Fig. 4b).

The low noise plot (Fig. 4c) resembles most closely the graph for visually based choices (Fig. 4a), as expected during the Face task. The differences from the ideal case indicate evidence for contributions of the auditory information to the judgements. Adding visual noise progressively modifies this pattern (Fig. 4g, h, i), and responses become more consistent with auditory based choices (Fig. 4b), in line with a compensation mechanism during cue combination. Adding varying levels of auditory noise, however, (Fig. 4d, e, f) does not seem to produce much change in the results. The complementary pattern of results is obtained with the auditory task with the roles of the visual and auditory noise reversed (Supplementary section 2). The same trends are evident in the CPPs from each individual (Supplementary section 2) except that the data are noisier since based on fewer total trials.

The trends illustrated in the CPP plots are summarized quantitatively in Fig. 5a, b, c, d, e, f, g, h and i, which show the average relative change in weight of the contribution of each modality as a function of the task (rows) and the modality of the noise (columns). Individual results from which the averages were calculated are presented in Supplementary section 3. As in the simulations, when noise is added to the attended modality, its weights decreased with noise level and the weights of the non-attended modality increased (Fig. 5a, e). In contrast with the simulations, when noise is added to the non-attended modality, its weights decreased and the weights of the attended modality were unaffected (Fig. 5b, d). In summary, there was an interaction between

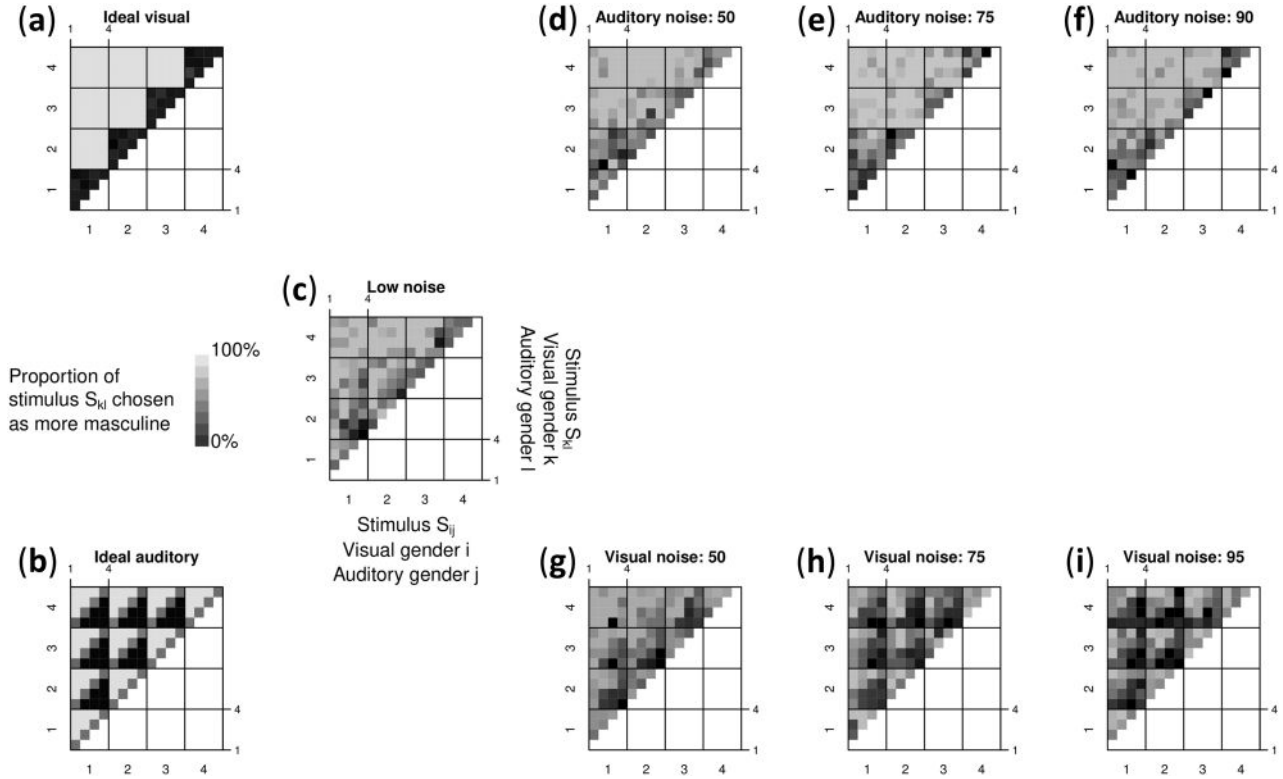

Fig. 4. Conjoint proportions plots for ideal (**a**, **b**) and grouped data (n = 6) for the face task (**c**, **d**, **e**, **f**, **g**, **h**, **i**). The conjoint proportion plot shows the proportion of responses for choosing one stimulus over another for pairs of stimuli with visual *v* and auditory *a* gender levels S_{va}. On each graph, the outer axes indicate the visual gender levels *i*, *k* and within each grid box, the inner axes represent the auditory gender levels *j*, *l* corresponding to all possible paired comparison trials between S_{ij} and S_{kl}. The pixel grey levels indicate the proportion of responses on which the participants on average chose S_{kl} over S_{ij} based on the instruction to choose the stimulus with the most masculine face. (**a**, **b**) The ideal observer graphs indicate the expected pattern of results if the choices were made only on the basis of one modality ((**a**) visual, (**b**) auditory). (**c**) Participants responses in the low noise condition for both modalities that serves as a control level. (**d**, **e**, **f**) Participants responses under increasing auditory noise. (**g**, **h**, **i**) Participants responses under increasing visual noise. Empirical plots correspond to the first column and row of the first graph in Supplementary section 2, where the graphs for the other two tasks and for each individual can also be found.

noise and attention, since no compensation occurred in the attended modality when the noisy modality was unattended. These results are task dependent, because when the subject was asked to attend to both modalities, the contributions of the two channels changed as when the noise was added only to the attended modality (Fig. 5g, h). Adding noise to both modalities decreased the contributions of both modalities independently of the task (Fig. 5c, f, i).

Discussion

Our results demonstrate an asymmetry in the combination of top-down and bottom-up influences in multimodal judgements in a gender perception task. When noise is added to an attended modality, its contribution to gender judgements decreases systematically with the level of noise while the contribution of the unattended modality increases in a contrary fashion. In contrast, when the noise is added to the unattended modality, its contribution decreases while the contribution of the attended modality remains unaffected. It seems therefore that attention plays a role in excluding signals considered as irrelevant from the usual cue combination weighting process.

One possible explanation for this asymmetry is that the resources allocated to the attended modality are at 100%, so that while adding noise to the unattended modality decreases its contribution, the attended modality is already at its maximum weight. However, control experiments with no noise added (Fig. 1b and c, and our previous study[15], see Supplementary section 1 for a direct comparison with our lowest noise condition), show that even though the task requires judgements on the basis of only one modality, the other modality continues to make a significant contribution. Therefore, it seems unlikely that the top-down attentional resources allocated to one modality are at maximum value.

We can account for this interaction effect by including a conditional factor that nullifies the noise compensation mechanism under certain conditions. This could be implemented in the brain as a top-down signal that, when attention is strongly directed to one modality, precision information from the other modality is inhibited. Such an operation makes it impossible to compute the weight of noise compensation (Fig. 5j),

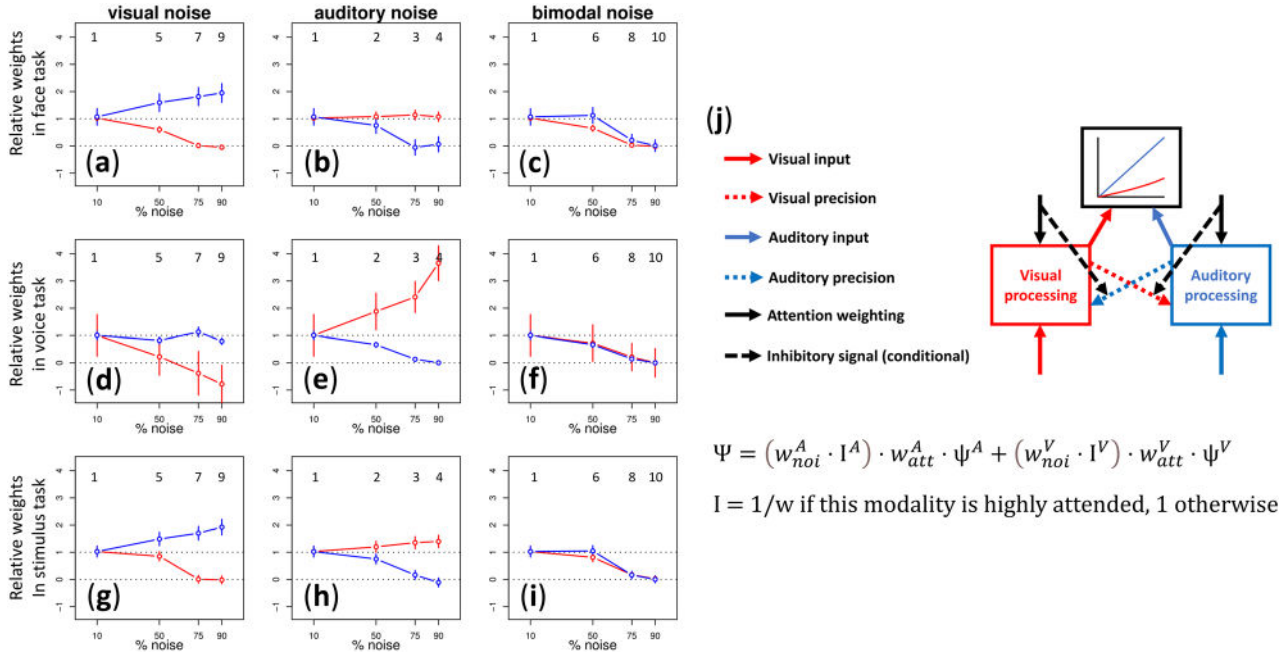

Fig. 5. Experimental results. (**a**, **b**, **c**, **d**, **e**, **f**, **g**, **h**, **i**) Average relative change of contributions of visual (red) and auditory (blue) modalities as compared to lowest bimodal noise condition when progressively adding noise to the visual (left column), auditory (middle column) or both modalities (right column), while participants were performing the face task (top row), voice task (middle row) or stimulus task (bottom row). Each panel shows the average of 6 observers. Error bars display 95% confidence intervals. The numbers at the top of each plot indicate combinations of visual and auditory noise as referenced in Table 1. (**j**) Feedback model of combination of precision estimates to influence weighted cue combinations of gender judgements. Feedback from attention (black arrows) and noise (dotted arrows) both influence the weights in the contributions of visual and auditory signals in the decision process. There is additionally an inhibitory signal conditional on attention (dotted black arrows) when attention is strongly focused on one modality, negating noise weighting from the other modality.

and the result is a decreased contribution for the noisy unattended modality as a direct effect of noise, but no compensating increase in the non-noisy strongly attended modality. When both modalities are equally attended, the attention threshold for this mechanism is not attained by the information in either modality.

Given the ubiquity of multimodal sensory integration[1], these observations are challenging for optimal cue combination models[5,6,16] and, in particular, theories of predictive coding[25,26]. Juni et al.[27] cite several factors that limit ideal cue combination. For example, they note that near optimal cue combination performance is attained in naturalistic tasks similar to our face-voice gender comparisons rather than tasks involving cognitive decisions and learning. The gender choice might be seen as a cognitive decision but no learning is involved in our study. Also, having the cues simultaneously present as here with the face and voice may aid in optimal combination, though the comparisons between stimuli here were successive. Having a large number of cues present may also lead to sub-optimal combination because of memory limitations. Our results may depend upon the nature of the task since visual cues appear to predominate auditory cues in spatial localization, even when the visual cues are severely degraded through blur[28].

Our results are also relevant to certain brain disorders such as bimodal learning in prosopagnosia and phonagnosia. For example, contrary to control subjects, voice recognition in prosopagnosics does not benefit from bimodal learning (reviewed by Maguinness and von Kriegstein[22]). Under the assumption that channels processing face information in prosopagnosics behave similarly to an unattended modality, the relevant auditory modality does not compensate for a decrease in signal from the irrelevant visual modality. In contrast, voice recognition in phonagnosics does benefit from bimodal learning. Even though they experience difficulty in recognizing voices, when phonagnosics attend to voices, information from the visual channel aids in compensating for their deficit. Given the pattern of results in our data, these observations lead to the prediction that face recognition in prospagnosics would benefit from auditory information, but that face recognition in phonagnosics would not benefit from voice information.

Two particularities of our work compared to other cue combination studies are that subjects were instructed to attend to either one or both modalities, thus selectively influencing the precision from one or both channels, and that noise was added in varying amounts to both modalities. Under these conditions, our results demonstrate violations of optimal (Bayesian) inference in multimodal cue combination.

Recently, Smeets and Brenner[29] advanced a framework to explain task dependent inconsistent judgements about object properties, for example, arising from local versus global cues in a visual illusion for which a Bayesian analysis of the parts would lead to a non-Bayesian perception of the whole object. Thus, this violates the view of a Bayesian observer as someone "who, vaguely expecting a horse and catching a glimpse of a donkey, strongly

	%Auditory noise			
%Visual noise	10	50	75	90
10	1	2	3	4
50	5	6		
75	7		8	
90	9			10

Table 1. Combinations of visual and auditory noise used in the experiments. Percentages are the proportion of noise in the visual/auditory input in terms of contrast/volume (e.g., for an auditory noise level of 25% the volume of the voice was reduced from 90 to 75% and the volume of the noise increased from 10 to 25%). There were 150 trials for each noise condition. Conditions used in the experiment are indicated by a number that references the corresponding levels in the panels a-i of Fig. 5.

concludes he has seen a mule"[30]. Instead, Smeets and Brenner propose that overall consistency is not necessarily the criterion employed in perceptual processing, despite our beliefs that the laws of physics should be consistent. Rather, they suggest that we perceive a series of answers to sequential and specific questions related to the task at hand, and these answers need not be consistent with one another. In the present study, the inconsistency arises as a lack of trade-off between the effects of bottom-up noise and top-down attention on precision in the judgements of multi-modal combinations of gender cues, as required by optimal cue combination.

Methods

Procedure

Eighteen individuals (9 male) with normal or corrected-to-normal vision volunteered for the experiment (mean age +/- SD = 25.7 +/- 3.4 years). Each participant was randomly assigned to one of the three experimental conditions so that there were 3 male and 3 female observers per condition. All observers were naive, right-handed and native French speakers. All observers had normal (or corrected to normal) vision as assessed by the Freiburg Visual Acuity and Contrast Test (FrACT)[31], and normal color vision as assessed by the Farnsworth F2 plate observed under daylight fluorescent illumination (Naval Submarine Medical Research Laboratory, Groton, CT, USA). Normality of face perception was assessed by the Cambridge Face Memory Test (CFMT)[32]. All observers gave informed consent and were compensated for their participation. All studies were approved by the Comité de Protection des Personnes Sud-Est III Groupement Hospitalier Est, Hôpital Civil de Lyon Bron, France and conducted in agreement with the Declaration of Helsinki for the protection of human subjects.

Experiments were performed in a dark room and stimuli displayed on an Eizo FlexScan T562-T color monitor (42 cm) driven by a Power Mac G5 (3gHz) with screen resolution 832×624 pixels and run at a field rate of 120 Hz, noninterlaced. Calibration of the screen was performed with a Minolta CS-100 Chromameter. Observers were placed at a distance of 57.3 cm from the screen, and head stabilization was obtained with a chin and forehead rest. Auditory stimuli were presented through headphones (Sennheiser HD 449), which also served to mask any ambient noise. Sound calibration was performed with a Quest QE4170 microphone and a SoundPro SE/DL sound level meter.

The stimulus set, obtained from Watson et al.[13] and used in our previous study[15], consisted of video clips of the face of a person saying the phoneme "had", whose face and voice gender varied by morphing from feminine to masculine (18 levels of morphing for the face from 0.1 to 0.95 and 19 levels for the voice from 0.05 to 0.95). The endpoint faces/voices were obtained from averages of 10 male and female faces. From these we selected 4 levels of face gender morphing (0.1, 0.35, 0.65 and 0.95), and 4 levels of voice gender morphing (0.05, 0.35, 0.65 and 0.95) to obtain a 4×4 stimulus set (yielding 120 unique face-voice pairs). These levels are equally spaced on the gender morphing axis and generate a sparser sample of the curves obtained in Abbatecola et al. (2021)[15] (see Fig. 1b). The levels conform to the requirement for MLCM that the levels be easily ordered by a participant[17]. This corresponds to 150 trials for each noise condition, which should yield results with an equivalent level of accuracy from our previous paradigm (in which we used a 18×19 set, also with 1500 trials) according to a power simulation based on preliminary data for this project (see Abbatecola et al. (2021)[15] Supplementary section 1, Figure S2). The clips were converted to greyscale and matched for average luminance. An oval mask fitted around each face hid non-facial gender cues, such as the hair and the hairline.

The software PsychoPy3[33] (https://www.psychopy.org) was used to control stimulus presentation. Stimuli were displayed in the center of a grey background (31.2 cd/m^2, CIE xy = (0.306, 0.33)). Face luminance varied between 29.7 cd/m^2 (CIE xy = (0.306, 0.324)) for the eyes and 51.6 cd/m^2(CIE xy = (0.303, 0.326)) for the nose. Face diameter was fixed at 10 degrees of visual angle and voice volume between 85.2 and 86.7 dB SPL (A) - Peak.

Each observer performed 1500 trials allocated over 5 sessions of 300 trials each, randomly distributed among 10 noise conditions that combined different levels of visual and auditory noise, as indicated in Table 1. Visual noise was composed of white noise superimposed on top of the video images, obtained by randomizing the luminance of the image pixels. For the auditory noise, a random pink auditory noise was played during the audio. We defined 4 noise levels in the visual and auditory dimensions roughly equally spaced subjectively from being barely noticeable (10%) to rendering the stimuli unrecognizable (90%). We used pink rather than white noise for the auditory channel because high levels of white noise were perceived to be uncomfortably loud, and pink noise is more efficient at masking the human voice[34]. The 10% noise condition differed trivially from a

no-noise condition reported previously[15] (Supplementary section 1). We observed no evidence in our results to suggest that the effects on performance differed for the visual compared to the auditory noise.

Within the gender equality constraints specified above, participants were randomly assigned to one of three groups that differed only with respect to instructions to judge the video clips according to either the gender of the face, the gender of the voice or the gender of the stimulus (i.e., participants in this group were encouraged to use information from both modalities). We know from previous MLCM literature[19], including with the same stimulus set[15] that when instructed in this way participants are able to bias their attention to a particular stimulus feature.

On each trial two stimuli were randomly selected with the gender morphing scale values of the voice and face independently and randomly assigned and successively presented. The duration of each stimulus was fixed at 500ms with a minimum 500ms inter-stimulus interval between each pair. After the presentation, observers judged which stimulus (face, voice or stimulus, according to their assigned group) appeared more masculine. The next pair was presented following the observer's response via a button press using a Logitech gamepad f310.

Quantification and statistical analysis

Fitting MLCM

Curve fitting, simulation and statistical analyses were performed with R[35] (https://www.r-project.org) using the **MLCM**[36] (https://cran.r-project.org/web/packages/MLCM/index.html) and **lme4**[37] (https://cran.r-project.org/web/packages/lme4/index.html) packages. We summarize the MLCM signal detection model here that has been described in detail elsewhere[16,17].

For each trial, two non-identical items were randomly sampled from the set of visuo-auditory stimuli ordered along the face and voice gender physical continua (ordered by relative morphing between extreme gender examplars). Given a trial with the pair of physical gender levels $S_1 = (\phi_1^V, \phi_1^A)$ and $S_2 = (\phi_2^V, \phi_2^A)$, using V and A superscripts to signify visual and auditory components, respectively, we suppose that the two stimuli generate internal gender representations determined by a psychophysical function, ψ. For a single trial, the noisy gender comparison process can be modeled as:

$$\Delta_i(S_1, S_2) = \psi_1 - \psi_2 + \epsilon = \psi(\phi_1^V, \phi_1^A) - \psi(\phi_2^V, \phi_2^A) + \epsilon = \delta_i + \epsilon, \tag{1}$$

where ψ_1 and ψ_2 are internal representations for the gender of the first and second stimuli, respectively, determined by the psychophysical function, ϵ is a Gaussian random variable with mean $\mu = 0$ and variance σ^2 corresponding to judgement noise that accounts for random variation of observer responses when presented with the same stimulus pair , and Δ_i is the decision variable on the i^{th} trial.

We assume that the observer chooses the first stimulus when $\Delta_i > 0$, and otherwise the second. We code the observer's responses, R, by 1 or 0 depending on whether the choice is stimulus 1 or 2. This is a Bernoulli distributed random variable, and the log-likelihood, ℓ, of the model over all trials given the observer's responses is:

$$\ell(\Delta_i, R_i) = \sum_i R_i \cdot \log\left(\Phi\left(\frac{\delta_i}{\sigma}\right)\right) + (1 - R_i) \cdot \log\left(1 - \Phi\left(\frac{\delta_i}{\sigma}\right)\right), \tag{2}$$

where R_i is the response on the i^{th} trial and Φ is the cumulative distribution function for a standard normal variable. In all cases, the psychophysical scale values were estimated by maximizing the likelihood of the observers' choices across all trials. The estimated scale is unique only up to addition of a constant and/or multiplication by a scalor , thus requiring constraints to be imposed on the estimated values to render the model identifiable[15,17]. Here, we fixed the values of ψ to 0 at the most feminine values of the face and voice scales and parameterized σ so it corresponds to the unit along the perceptual scales. For this reason, we designated the ordinates in Figs. 1b and c and Fig. 3a and b as d'[17].

MLCM studies usually consider three nested models (independent, additive and interaction), which can be compared using likelihood ratio tests. In our case, following previous results with the same stimulus set and a comparable paradigm[15], we know that the independent model is not adequate, and we have previously determined that relevant interaction effects are small. We, therefore, focused on the additive model, under which we define the decision variable as a sum of the differences between the visual and auditory gender signals. Here the internal response to stimulus 1 is modeled as:

$$\psi(\phi_1^V, \phi_1^A) = \psi_1^V f(\phi_1^V) + \psi_1^A g(\phi_1^A), \tag{3}$$

where ψ_1^V and ψ_1^A are parameters representing internal gender responses evoked by the visual cue ϕ_1^V and the auditory cue ϕ_1^A from stimulus 1[17]. The decision variable of the first versus the second stimulus for each trial can then be derived using Eq. (1).

Our previous results with this stimulus set[15] allowed us to further parameterize the form of the scale of responses for each modality (Fig. 1c). Specifically, the voice contribution with respect to gender level was well described by a linear function of gender morphing (with slope varying across attentional task). Similarly, the face contribution was well described by a quadratic function (with coefficient varying across attentional task). In the earlier study, we refitted the MLCM models to the observers' choices with these fixed curves using a Generalized Linear Mixed-Effects Model[36] with participants as a random effect and the internal responses to an individual stimulus modeled as a linear predictor.

$$\psi(\phi^V, \phi^A) = \psi^V \cdot (\phi^V)^2 + \psi^A \cdot \phi^A, \tag{4}$$

where ψ^V and ψ^A are parameters estimated for face and voice gender levels ϕ^V and ϕ^A.

This parametrization allowed us to reduce the number of parameters necessary to estimate each model, as well as to model the effect of attention as a weighting factor[15] for each modality. The latter was done by fitting the model to the "Stimulus" condition (i.e.,where attention was not specifically directed to either modality), then allowing the two parameters to vary to account for data from the modality-specific attentional conditions (see Fig. 1d for an illustration of the resulting weights). The resulting internal response model for each bimodal stimulus is described as:

$$\psi(\phi^V, \phi^A) = \mathrm{w}^V_{\mathrm{att}} \cdot \psi^V \cdot (\phi^V)^2 + \mathrm{w}^A_{\mathrm{att}} \cdot \psi^A \cdot \phi^A, \tag{5}$$

where $\mathrm{w}^V_{\mathrm{att}}$ and $\mathrm{w}^A_{\mathrm{att}}$, constrained to fall between 0 and 1, indicate the weights attributed to attention in the visual and auditory modalities, respectively, in a given condition with $\mathrm{w}^V_{\mathrm{att}} = 1 - \mathrm{w}^A_{\mathrm{att}}$ (Fig. 1f). Analyzing the effects of visual and auditory noise is performed in the same way by fitting an initial model to the condition with the least amount of noise bimodally, then estimating the relative change of weights in all of the other noise conditions.

Combining both attention and noise weighting, we obtain the internal response model used in our analysis (Fig. 5a, b, c, d, e, f, g, h, i):

$$\psi(\phi^V, \phi^A) = \mathrm{w}^V_{\mathrm{att/noi}} \cdot \psi^V \cdot (\phi^V)^2 + \mathrm{w}^A_{\mathrm{att/noi}} \cdot \psi^A \cdot \phi^A, \tag{6}$$

where $\mathrm{w}^V_{\mathrm{att/noi}}$ and $\mathrm{w}^A_{\mathrm{att/noi}}$ are the weights corresponding to a particular combination of attention and noise level in the two modalities.

Simulated observer

We also applied the same analysis on simulated data to compare to the results of our actual participants (Fig. 3a, b, c, d). In the simulations, we chose attentional weights that roughly matched our previous empirical values[15]: 0.6 for the attended modality, 0.4 for the unattended, 0.5 for both modalities in the stimulus task.

Concerning the effect of noise, for optimal cue combination across modalities, under reasonable constraints we define weights for each modality as[5,6]:

$$\mathrm{w}^V_{\mathrm{noi}} = \frac{\sigma_V^{-2}}{\sigma_V^{-2} + \sigma_A^{-2}}; \mathrm{w}^A_{\mathrm{noi}} = \frac{\sigma_A^{-2}}{\sigma_V^{-2} + \sigma_A^{-2}}, \tag{7}$$

where σ_V^{-2} and σ_A^{-2} represent signal precision in the visual and auditory modalities in a given trial.

We created a dataset composed of all empirical trials across participants, for which we determined a simulated response using the formula:

$$\begin{aligned} \Delta_{\mathrm{sim}} &= \delta^V_{\mathrm{sim}} + \delta^A_{\mathrm{sim}} + \epsilon \\ \delta^V_{\mathrm{sim}}(\phi^V_1, \phi^V_2) &= \mathrm{w}^V_{\mathrm{att}} \cdot \mathrm{w}^V_{\mathrm{noi}} \cdot \left[\phi^V_1 - \phi^V_2 + \epsilon^V\right] \\ \delta^A_{\mathrm{sim}}(\phi^A_1, \phi^A_2) &= \mathrm{w}^A_{\mathrm{att}} \cdot \mathrm{w}^A_{\mathrm{noi}} \cdot \left[\phi^A_1 - \phi^A_2 + \epsilon^A\right], \end{aligned} \tag{8}$$

where Δ_{sim} is the decision variable, δ^V_{sim} and δ^A_{sim} are simulated perceived differences in the visual and auditory modalities, attentional and noise weighting are determined depending on the trial, ϵ^V and ϵ^A are Gaussian random variables with mean $\mu = 0$ and variance depending on the level of noise in the corresponding modality. The noisy decision rule is the same as used in fitting the empirical data.

Data availability

Source data for the figures in this paper are archived at https://github.com/ClementAbb/FV_noise_attention, Supplementary figures in a pdf file and the raw data are available as a supplementary text file.

Received: 4 December 2024; Accepted: 26 June 2025
Published online: 16 July 2025

References

1. Ghazanfar, A. A. & Schroeder, C. E. Is neocortex essentially multisensory?. *Trends Cogn. Sci.* **10**, 278–285 (2006).
2. Friston, K. Prediction, perception and agency. *Int. J. Psychophysiol.* **83**, 248–252 (2012).
3. Macaluso, E. et al. The curious incident of attention in multisensory integration: Bottom-up vs. top-down. *Multisens. Res.* **29**, 557–583 (2016).
4. Hartcher-O'Brien, J., Soto-Faraco, S. & Adam, R. A matter of bottom-up or top-down processes: The role of attention in multisensory integration. *Front. Integr. Neurosci.* **11**, 5 (2017).
5. Ernst, M. O. & Banks, M. S. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433 (2002).
6. Oruç, I., Maloney, L. T. & Landy, M. S. Weighted linear cue combination with possibly correlated error. *Vis. Res.* **43**, 2451–2468 (2003).
7. Young, M. J., Landy, M. S. & Maloney, L. T. A perturbation analysis of depth perception from combinations of texture and motion cues. *Vis. Res.* **33**, 2685–2696 (1993).

8. Yon, D. & Frith, C. D. Precision and the Bayesian brain. *Curr. Biol.* **31**, R1026–R1032 (2021).
9. Campanella, S. & Belin, P. Integrating face and voice in person perception. *Trends Cogn. Sci.* **11**, 535–543 (2007).
10. Rohe, T. & Noppeney, U. Cortical hierarchies perform Bayesian causal inference in multisensory perception. *PLoS Biol.* **13**, e1002073 (2015).
11. Nahorna, O., Berthommier, F. & Schwartz, J.-L. Binding and unbinding the auditory and visual streams in the McGurk effect. *J. Acoust. Soc. Am.* **132**, 1061–1077 (2012).
12. Macke, J. H. & Wichmann, F. A. Estimating predictive stimulus features from psychophysical data: The decision image technique applied to human faces. *J. Vis.* **10**, 22–22. https://doi.org/10.1167/10.5.22 (2010).
13. Watson, R. *et al.* Audiovisual integration of face–voice gender studied using "morphed videos". *Integrating Face and Voice in Person Perception* 135–148 (2013).
14. Latinus, M., VanRullen, R. & Taylor, M. J. Top-down and bottom-up modulation in processing bimodal face/voice stimuli. *BMC Neurosci.* **11**, 1–13 (2010).
15. Abbatecola, C., Gerardin, P., Beneyton, K., Kennedy, H. & Knoblauch, K. The role of unimodal feedback pathways in gender perception during activation of voice and face areas. *Front. Syst. Neurosci.* **15**, 669256 (2021).
16. Ho, Y.-X., Landy, M. S. & Maloney, L. T. Conjoint measurement of gloss and surface texture. *Psychol. Sci.* **19**, 196–204 (2008).
17. Knoblauch, K. & Maloney, L. T. *Modeling Psychophysical Data in R* (Springer Science & Business Media, London, 2012).
18. Aguilar, G. & Maertens, M. Toward reliable measurements of perceptual scales in multiple contexts. *J. Vis.* **20**, 19–19 (2020).
19. Maloney, L. T. & Knoblauch, K. Measuring and modeling visual appearance. *Ann. Rev. Vis. Sci.* **6**, 519–537 (2020).
20. Vincent, J., Maertens, M. & Aguilar, G. What Fechner could not do: Separating perceptual encoding and decoding with difference scaling. *J. Vis.* **24**, 5–5 (2024).
21. Feldman, H. & Friston, K. J. Attention, uncertainty, and free-energy. *Front. Hum. Neurosci.* **4**, 215 (2010).
22. Maguinness, C. & von Kriegstein, K. Cross-modal processing of voices and faces in developmental prosopagnosia and developmental phonagnosia. *Vis. Cogn.* **25**, 644–657 (2017).
23. Gerardin, P., Devinck, F., Dojat, M. & Knoblauch, K. Contributions of contour frequency, amplitude, and luminance to the watercolor effect estimated by conjoint measurement. *J. Vis.* **14**, 9–9 (2014).
24. Sun, H.-C., St-Amand, D., Baker, C. L. Jr. & Kingdom, F. A. Visual perception of texture regularity: Conjoint measurements and a wavelet response-distribution model. *PLoS Comput. Biol.* **17**, e1008802 (2021).
25. Bastos, A. M. et al. Canonical microcircuits for predictive coding. *Neuron* **76**, 695–711 (2012).
26. Markov, N. T. & Kennedy, H. The importance of being hierarchical. *Curr. Opin. Neurobiol.* **23**, 187–194 (2013).
27. Juni, M. Z., Gureckis, T. M. & Maloney, L. T. Effective integration of serially presented stochastic cues. *J. Vis.* **12**, 12–12 (2012).
28. Shayman, C. S. et al. Integration of auditory and visual cues in spatial navigation under normal and impaired viewing conditions. *J. Vis.* **24**, 7–7 (2024).
29. Smeets, J. B. & Brenner, E. The cost of aiming for the best answers: Inconsistent perception. *Front. Integr. Neurosci.* **17**, 1118240 (2023).
30. Senn, S. S. *Statistical issues in drug development* Vol. 69 (John Wiley & Sons, Hoboken, 2008).
31. Bach, M. The Freiburg visual acuity test-variability unchanged by post-hoc re-analysis. *Graefes Arch. Clin. Exp. Ophthalmol.* **245**, 965–971 (2006).
32. Duchaine, B. & Nakayama, K. The Cambridge face memory test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia* **44**, 576–585 (2006).
33. Peirce, J. W. Generating stimuli for neuroscience using PsychoPy. *Front. Neuroinform.* **2**, 343 (2009).
34. Saeki, T., Tamesue, T., Yamaguchi, S. & Sunada, K. Selection of meaningless steady noise for masking of speech. *Appl. Acoust.* **65**, 203–210 (2004).
35. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2025).
36. Knoblauch, K., Maloney, L. T. & Aguilar, G. *MLCM: Maximum Likelihood Conjoint Measurement* (2022). R package version 0.4.3.
37. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48. https://doi.org/10.18637/jss.v067.i01 (2015).

Acknowledgements

The authors thank Laurence T. Maloney and Steven K. Shevell for critical comments on an earlier draft of the paper. This study was supported by grants DUAL_STREAMS ANR-19-CE37-0025 (K.K.); LABEX CORTEX ANR-11-LABX-0042, Université de Lyon ANR-11-IDEX-0007 (H.K.); CORTICITY ANR-17-HBPR-0003, Connectome ANR-24-CE37-5022-01 (H.K.); PREDICTION ERC-2023-Adv 101142153 (H.K.) and PEP 69 (C.A.).

Author contributions

CA, KK and HK designed the experiments; CA programmed and performed the experiments; CA and KK analysed the data; CA, KK and HK wrote the manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-09542-6.

Correspondence and requests for materials should be addressed to C.A. or K.K.