PROCEEDINGS OF THE ROYAL SOCIETY B

Proc. R. Soc. B (2011) **278**, 1365–1372
doi:10.1098/rspb.2010.1888
Published online 20 October 2010

A general rule for sensory cue summation: evidence from photographic, musical, phonetic and cross-modal stimuli

M. P. S. To[1,*], R. J. Baddeley[2], T. Troscianko[2] and D. J. Tolhurst[1]

[1]*Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, UK*

[2]*Department of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, UK*

The Euclidean and MAX metrics have been widely used to model cue summation psychophysically and computationally. Both rules happen to be special cases of a more general Minkowski summation rule $(\mathrm{Cue}_1^m + \mathrm{Cue}_2^m)^{1/m}$, where $m = 2$ and ∞, respectively. In vision research, Minkowski summation with power $m = 3$–4 has been shown to be a superior model of how subthreshold components sum to give an overall detection threshold. Recently, we have previously reported that Minkowski summation with power $m = 2.84$ accurately models summation of *supra*threshold visual cues in photographs. In four suprathreshold discrimination experiments, we confirm the previous findings with new visual stimuli and extend the applicability of this rule to cue combination in auditory stimuli (musical sequences and phonetic utterances, where $m = 2.95$ and 2.54, respectively) and cross-modal stimuli ($m = 2.56$). In all cases, Minkowski summation with power $m = 2.5$–3 outperforms the Euclidean and MAX operator models. We propose that this reflects the summation of neuronal responses that are not entirely independent but which show some correlation in their magnitudes. Our findings are consistent with electrophysiological research that demonstrates signal correlations ($r = 0.1$–0.2) between sensory neurons when these are presented with natural stimuli.

Keywords: natural images; music; phonetics; cross-modal; feature integration; neuronal correlation

1. INTRODUCTION

Our interactions with the environment require the ability to accurately interpret and integrate the surrounding perceptual cues [1–4]. In vision, straightforward combination rules for neural channels have been proposed in detection (e.g. [5]) and visual search experiments [6–10]. These include classical models such as linear (city-block) addition, energy (Euclidean) summation and the maximum (MAX) rule. Of these, the Euclidean metric has been widely used to define psychological space (e.g. [6,7]) and has often been the benchmark against which psychological models of cue summation are tested. However, the MAX metric has also been studied computationally and biologically (e.g. [11–13]). These two rules are therefore particularly important in the modelling of perception.

However, there is evidence to suggest that cue summation may diverge systematically from these classical rules, which are special cases of a more general Minkowski summation rule:

$$\mathrm{Minkowski\_sum} = (\mathrm{Cue}_1^m + \mathrm{Cue}_2^m)^{1/m} \qquad (1.1)$$

where $m = 1$, 2 and ∞ result in linear addition, Euclidian summation and the MAX rule, respectively.

Visual detection experiments using compound stimuli such as pairs of sinusoidal gratings or even complex natural images have shown that a Minkowski rule with power $m = 3$–4 (rather than Euclidean or MAX) is a good model of how subthreshold components sum to give an overall detection threshold (e.g. [14–18]). Furthermore, we have recently shown in *supra*threshold discrimination experiments that perceptual combination of features such as colour changes and shape changes in photographs of *natural* visual scenes are best modelled using a Minkowski summation with power $m \approx 3$ [19,20].

We are interested in whether such perceptual summation of suprathreshold cues also occurs when an observer listens to composite natural sounds, such as music or speech. In hearing, there has been much debate on how auditory cues combine in multi-tone stimuli (e.g. [21]) and more complex sounds like music (e.g. [22]) or speech (e.g. [23]). Although there exist parallels between visual and auditory processing, research in auditory feature integration arises from a different tradition with a different point of view, so that it is difficult to draw direct comparisons between summation models used in vision and hearing. However, Green [21] did report that two subthreshold tones summed to reach threshold with less summation than expected for power (i.e. Euclidean) summation, implying that simple auditory cue summation might also follow a Minkowski rule with $m > 2$.

The present paper compares cue or feature combination in different auditory and visual natural stimuli by asking observers to give magnitude estimation ratings for pairs of stimuli that differ along a number of different dimensions. Cue combination is studied by considering how ratings to changes along single dimensions compare with

* Author for correspondence (to@cantab.net).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspb.2010.1888 or via http://rspb.royalsocietypublishing.org.

Received 2 September 2010
Accepted 1 October 2010

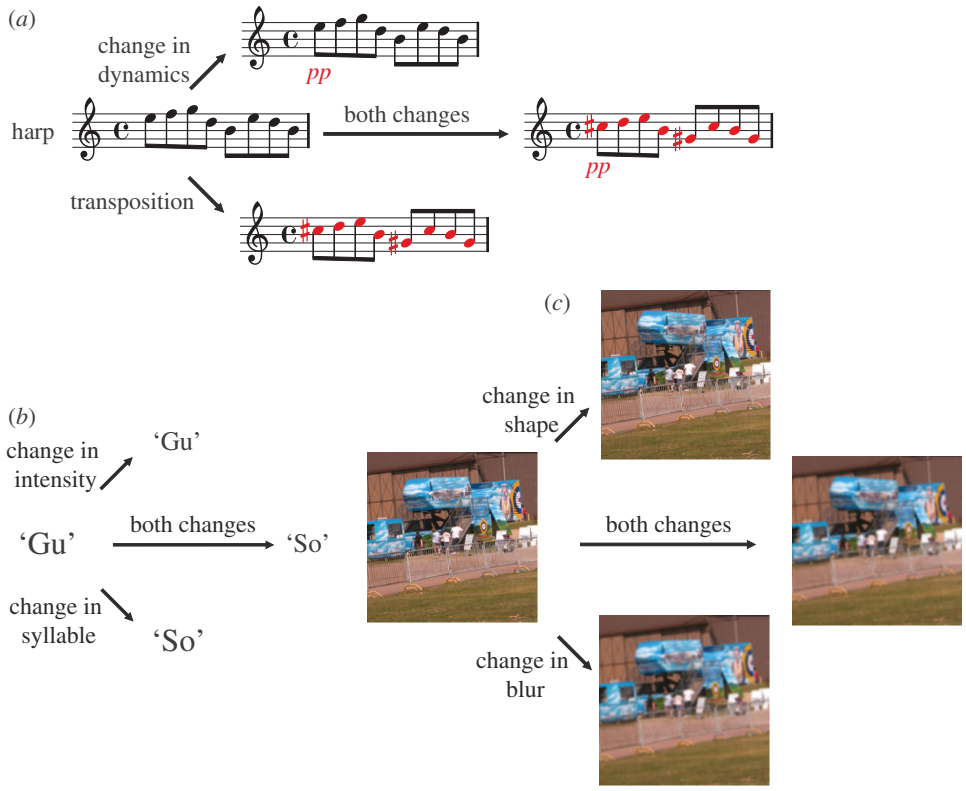

Figure 1. Examples of *combination sets* used in the (*a*) music, (*b*) phonetic and (*c*) visual scene experiments. Starting from one reference stimulus, the first pair (component pair) differed in one dimension, the second pair (a second component pair) differed in a second dimension and the final pair (the composite) differed in both dimensions.

ratings to changes along a combination of these dimensions. We will consider auditory cue summation in musical sequences and in phonetic utterances, and we will show the same kinds of systematic failure of the Euclidean and MAX models found with cue combination in photographs of natural scenes. We will also investigate how visual cues are summed with auditory cues in a cross-modal situation, where observers are simultaneously presented with photographs and natural sounds and where the stimulus changes can be visual, auditory or a combination of both. Again, our results will demonstrate that Minkowski summation with power $m = 2.5–3$ is most successful in describing feature integration in cross-modal natural stimuli and a significantly better description than Euclidean summation and MAX rule. We suggest that Minkowski summation with power $m = 2.5–3$ reflects the summation of neuronal responses that are not entirely independent but that show some correlation in their magnitudes. This would make sense since information from the world is correlated [24]. Our conclusions are supported by electrophysiological research that demonstrates consistent signal correlations between sensory neurons when they are presented with natural stimuli (see §4).

2. METHODS

(a) *Experimental stimuli*

(i) *Construction of stimuli*

In the music experiment (experiment 1), observers were presented with 160 musical sequence pairs (each lasting 2 s) that differed in one or two of the following dimensions: intensity (by changing the dynamics to *pp* or *ff*), timbre (by changing the instrument), pitch (by transposing the sequence upward or downward by various chromatic or diatonic intervals) and/or content (by changing, adding or removing one or more notes). The magnitudes of these changes varied between the different parent sequences. There were 16 reference sequences, each providing four single dimension changes and six composite changes (see §2*a*(ii)). All sequences were generated using a free evaluation copy of *Notion Demo* (Notion Music Software, v. 1.5.4.0), a piece of music composition and performance software. (Examples of sequences and differences are shown in figure 1*a* and in the electronic supplementary material, figure S1*a*,*b*.)

The phonetic experiment (experiment 2) used 320 stimulus pairs (each lasting 1–2 s) made from recordings of single spoken syllables. All phonetic utterances were recorded and modified using *Audacity* (v. 1.3.4—beta software that is freely downloaded online). The stimuli in the pair could differ in intensity (increased by 5 dB or decreased by 15 dB), low-pass (roll-off = 18 dB/octave) and high-pass (roll-off = 24 dB/octave) filtering, tempo (faster or slower) and pitch (very high, high, low or very low). The magnitudes of these changes were different for different syllables. A reference syllable could be paired with one of these variants or with a different syllable (e.g. 'ta' versus 'na'), and might change in one (component) or two (composite) ways (see §2*a*(ii)). There were 20 reference syllables, each with 15 variants, and a total of 320 phonetic utterances (see electronic supplementary material, table S1).

In the vision experiment (experiment 3), observers were presented with a total of 300 pairs of natural image scenes based on 20 parent images, each matched with 14 variants.

Some variants were a second photograph of the same scene taken when, say, an object had moved. In other variants, using MATLAB, the scene could be blurred or sharpened to varying degrees, the contrast could be reduced, or the hue and saturation of the whole scene could be changed while leaving the brightness relatively unaffected (see examples in figure 1*c* here and in the electronic supplementary material, figure S1*d*). In total, 20 pairs were identical, 100 pairs differed only along one of the dimensions described above and 180 pairs differed in a combination of two dimensions (see §2*a*(ii)).

The stimuli in the cross-modal experiment (experiment 4) were natural images coupled with natural sounds, i.e. a visual image was presented simultaneously with an auditory sequence. Observers were presented with a total of 648 pairs of these image-sound combinations and were asked how different the overall audio-visual 'experience' appeared to them. The stimuli were based on 36 reference photographs (nine each of animals, musical instruments, objects and people) and 36 appropriate reference sound effects. Seventy-two original photographs were purchased from the website *iStockphoto* (from various artist members) and then modified using MATLAB. Each reference image could be changed in two ways (e.g. blur, contrast or hue) and a third variant consisted of a second photograph of the same object or scene taken when a target object had moved or changed shape. The natural sound sequences were generated from 36 reference sound effects chosen from a database called 'INSTANTSOUNDFX'. Each reference sequence was subsequently modified using the software program *Audacity* and cropped to have a duration of 1 s. Three variants from each reference sound sequence were made by increasing or decreasing the intensity by values between 3 and 10 dB, low-pass (roll-off between 6 and 36 dB/octave) and high-pass (roll-off between 6 and 48 dB/octave) filtering, and lowering or heightening the pitch by 15–40 dB. In 324 image-sound pairs, the images and sounds were chosen so that their content was congruous (e.g. bird images–chirping sounds), but in the other 324 pairs, the images and sounds were drawn from different categories so that their content was incongruous (e.g. bird images–telephone ringing sounds). See the electronic supplementary material online for examples, and details on how congruous and incongruous sounds were matched with visual images.

The electronic supplementary material online also contains further details on the construction and presentation of all the stimuli described above.

(ii) *Combination sets*

The experiments were based around *combination sets* of three stimulus pairs. Starting from one reference stimulus, the first pair (component pair) differed in one dimension such as intensity, the second pair (a second component pair) differed in a second dimension such as pitch, and the final pair (the composite) differed in both the dimensions. For example, in the music experiment, a first pair (a component pair) might differ in one dimension such as dynamics, the second pair (a second component) might differ in a second dimension such as scale (transposition), and the final pair (the composite) would exhibit differences in *both* dynamics *and* scale dimensions (figure 1*a*). The magnitude of the two changes in the composite was the same as in the component pairs, and each component pair contributed to more than one combination set. There were in total 96 musical combination sets in the experiment 1, 200 phonetic combination sets in experiment 2 and 180 visual scene combination sets in experiment 3. See examples of each in figure 1.

In the cross-modal experiment (experiment 4), among the 648 image-sound pairs, 216 contained only image changes (while the sound remained unchanged), 216 contained only sound changes (while the images remained unchanged) and 216 contained both image and sound changes; i.e. there were 216 image-sound combination sets. Overall, the 216 combination sets consisted of 108 congruous and 108 incongruous sets. An example of congruous combination set is presented in the electronic supplementary material, figure S2*b–d*).

(b) *Experimental procedure*

The procedure was the same for all experiments, and the details of training and experimental design are given in the electronic supplementary material. Human observers (naive to the purpose of the experiments) were presented with pairs of stimuli and asked to make subjective numerical ratings of the perceived difference between the items in each pair [19]). During the experiments, observers were frequently presented with the same standard stimulus pair (specific for each experiment, see electronic supplementary material, figures S1*a*,*c*, S2*a* and table S1), whose magnitude difference was defined as '20'. They were instructed that their ratings of the subjective difference between any other test pair should be based on this standard pair: if they perceived the difference between the test pairs to be lesser, equal or greater than the standard pair, their ratings should be less, equal or greater than 20, respectively. For the experiments, the presentation sequence of stimulus pairs was randomized differently for each observer.

(c) *Data collation*

In each experiment, the magnitude estimation ratings of the 10–15 observers were averaged together for further analysis. The results for each observer were first divided by their median value (typically about 20). The scaled rating for each stimulus was then averaged across observers, and the average was multiplied by the grand average of all the observers' original ratings [25].

3. RESULTS

(a) *Cue combination for natural sounds and natural visual scenes*

Figure 2*a–c* examines whether Euclidean summation can predict the measured rating ($R3$) to the composite stimulus in each combination set from the separate ratings ($R1$ and $R2$) given for its two component pairs in the first three experiments. The figure panels plot the predicted value of R3 against the measured value of R3. By analogy with equation (1.1), the predicted rating for the compound is:

$$R3_{\text{predicted}} = (R1^m + R2^m)^{1/m}, \tag{3.1}$$

where $m = 1$ for linear summation, $m = 2$ for Euclidean summation and $m = \infty$ for the MAX rule. The predicted ratings for the composite stimuli (ordinate) are well correlated with the actual ratings (abscissa); Pearson's r is 0.87, 0.94 and 0.90 for figure 1*a* (musical sequences), figure 1*b* (phonetics) and figure 1*c* (visual scenes), respectively. However, there is a systematic failure of the prediction:

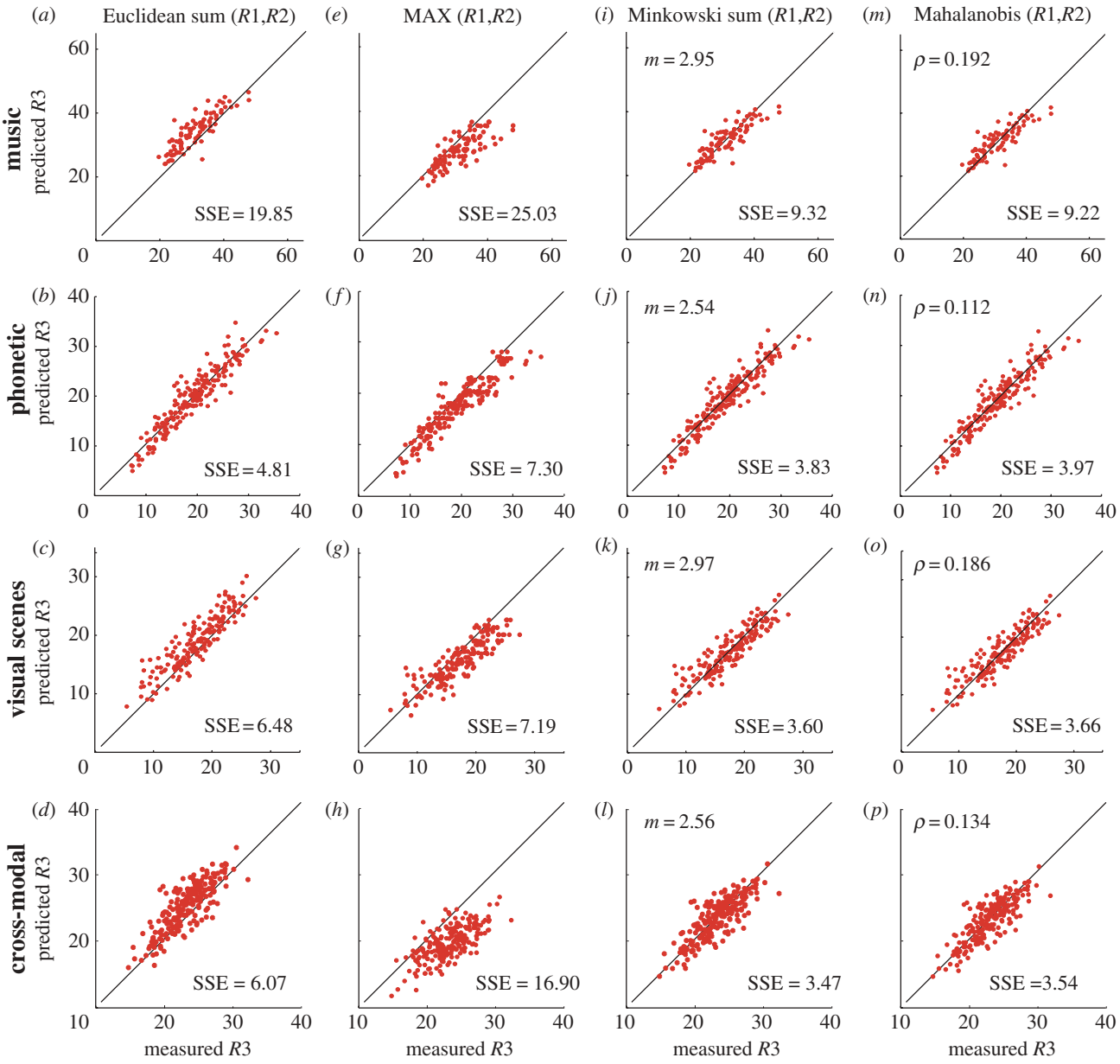

Figure 2. For the separate music, phonetics, visual and cross-modal experiments, predictions of the rating (*R*3) given to the composite stimulus pair in a combination set calculated from the individual ratings (*R*1 and *R*2) for the two separate component stimuli. Lines of equality are shown. SSE is the summed squared error between predicted and actual *R*3 divided by the number of combination pairs in the experiment, and expressed as 'squared rating units'. The results from experiments 1–4 are presented in the first, second, third and fourth rows, respectively. In (*a–d*), the Euclidean sum (when $m = 2$ in equation (3.1)) of *R*1 and *R*2 is plotted against the measured *R*3. In (*e–h*), the MAX (when $m = \infty$ in equation (3.1)) of *R*1 and *R*2 is plotted against the measured *R*3. In (*i–l*), the Minkowski sum (equation (3.1)) of *R*1 and *R*2 is plotted against *R*3 (the best-fitting Minkowski exponents are, *m*: 2.95, 2.54, 2.97, 2.56 in (*i–l*), respectively). In (*m–p*), the Mahalanobis sum (equation (4.1)) of *R*1 and *R*2 is plotted against *R*3 (covariance parameter, ρ: 0.192, 0.112, 0.186, 0.134 in (*m–p*), respectively).

the data points tend to lie above the line of equality (especially evident in figure 2*a*,*c*, where the summed squared-error per point (SSE) is greater than in figure 2*b*), showing that Euclidean summation of the component ratings has slightly overestimated the rating given by the observers to the compound stimuli.

The MAX rule (equation (3.1) with $m = \infty$) also performs poorly; figure 2*e–g* shows that data points were mostly below the line of equality, meaning that taking the maximum of the component ratings mostly underestimated the rating of the combination stimuli. Indeed, we have previously shown for cue integration in visual stimuli that the MAX operator model underestimated observers' ratings while linear addition dramatically overestimated them. The same trend was observed in the present experiments; table 1 shows that the SSE per point for the linear addition rule was very high and the model is less acceptable even than the Euclidean model. The SSE values for the MAX rule were close to but were consistently higher when compared with those for Euclidean summation.

Euclidean summation and the MAX operator are special cases of a general Minkowski summation rule (equation (3.1)), a summation rule frequently used to model threshold and suprathreshold visual performance [14–18,25] as well as elsewhere [26]. An iterative

Table 1. The summed squared error per point (SSE) for linear summation, MAX rule, Euclidean summation, Minkowski summation and Mahalanobis distance are shown. The predictions from the Minkowski and Mahalanobis models were consistently better than those for the more usually considered linear addition, MAX and Euclidean summation models.

		SSE/point				
experiment	number of combination sets	linear sum	maximum rule	Euclidean sum	best Minkowski sum	best Mahalanobis distance
1 (music)	96	290.64	25.03	19.85	9.32	9.22
2 (phonetic)	200	64.09	7.39	4.82	3.83	3.97
3 (visual scenes)	180	83.00	7.19	6.48	3.60	3.66
4 (cross-modal)	216	129.27	16.90	6.07	3.47	3.54

search was used to determine the value of the exponent m that minimized the sum of squared deviations SSE between the predicted value of $R3$ (ordinate, equation (3.1)) and the measured value on the abscissa; 95 per cent confidence intervals were obtained by Monte Carlo bootstrapping. Figure 2*i–k* show how well Minkowski summation of the ratings to component stimuli predicts the rating given to the composite stimulus in the first three experiments. This serves to compare our present experiments (particularly with natural sounds) with our previous visual experiments [19] but, more importantly, by allowing the summation exponent m to be a free parameter, it illustrates that there *is* a systematic failure of the Euclidean and MAX predictions.

In the two experiments with natural sounds (figure 2*i*,*j*), the generalized Minkowski summation rule outperformed the specific Euclidean and MAX rules, in that the SSE in both cases was reduced significantly (table 1). For instance, for the phonetics experiment where the improved fit is least obvious, adding the one parameter has caused SSE to fall from 4.81 to 3.83 on average for each data point: $F_{1,199} = 50.91$, $p \approx 0$. The Minkowski summation exponents are 2.95 (95% confidence interval, 2.62–3.36) and 2.54 (95% confidence interval, 2.31–2.78) in the music and phonetics experiment, respectively (figure 2*i*,*j*). The correlation coefficients between predicted and measured ratings are 0.86 (music) and 0.95 (phonetics).

To examine whether the Minkowski model was equally successful in modelling cue combination across the different feature dimensions (e.g. pitch and timbre versus intensity and tempo), we first calculated the squared errors between the averaged observers' ratings for each combination stimulus and the corresponding Minkowski predictions, and then analysed the mean-squared errors in a one-way repeated measures analysis of variance (ANOVA) for the different types of combination (listed in electronic supplementary material, table S3). In both auditory experiments, the Minkowski summation model was uniformly efficient in predicting the ratings for all the different combination types: $F_{5,95} = 1.69$, $p = 0.14$ for music and $F_{9,190} = 0.82$, $p = 0.60$ for phonetics. Post hoc Bonferroni analyses found no differences in the SSE between predicted and measured ratings among the different types of combinations.

Results from the visual experiment (figure 2*k*) also showed that Minkowski summation with a very similar exponent ($m = 2.97$; 95% confidence interval, 2.78–3.36) was a better model than Euclidean or MAX summation (table 1), confirming our previous reports on a different set of image stimuli [19]. The correlation between measured and predicted ratings was 0.91, and a one-way repeated measures ANOVA showed that the Minkowski summation model was equally accurate in modelling all nine types of combinations (electronic supplementary material, table S3): $F_{8,171} = 1.21$, $p = 0.30$. In addition, a post hoc Bonferroni analysis revealed no differences in the SSE between predicted and measured ratings among the nine types of visual combinations.

(b) *Combination of audio-visual cues in bimodal stimuli*

In the previous three experiments, observers were presented either with pairs of sounds or with pairs of visual scenes so that they could rate perceived differences either in the sound or in the visual stimulus, respectively. Here, we show the results of an experiment where each stimulus in a pair was a natural image visual stimulus *coupled with* a natural sound. The stimuli encompassed 216 combination sets, composed of three stimulus pairs: in the first pair only the images changed, in the second pair only the sounds changed, and in the third pair both images and sounds changed.

Figure 2*d*,*h* shows that Euclidean and MAX summations of the auditory cue (or rating) with the visual cue failed to predict the rating given by observers to the composite stimuli, where both visual and auditory components change. The discrepancy was similar to those of figure 2*a–c*,*e–g* for the separate auditory and visual cases. Minkowski summation (with power $m = 2.56$; 95% confidence interval, 2.42–2.67) of the ratings for the separate visual and auditory changes was a superior fit to Euclidean or MAX rules (figure 2*l*; table 1 summarizes all the SSE values). The correlation coefficient between predicted and measured ratings was 0.82. When combination sets from the congruous and incongruous conditions were analysed separately (see electronic supplementary material, figure S3), the best predictions were obtained with the general Minkowski summation rule with power $m = 2.62$ and 2.50, respectively. The absence of an effect of congruency suggests a general rule rather than a stimulus-specific phenomenon.

The remaining parts of figure 2*m–p* will be discussed below.

4. DISCUSSION

There has a long-been debate about the arithmetic rules governing feature integration: that is, the way in which a person combines multiple sensory or cognitive cues. It

has been asked whether successful demonstration of a rule would represent some universal Law of Mentation [6,7]. The present experiments have focussed on the summation of perceptual cues in *natural* auditory and visual stimuli. We have been able to look more subtly than many previous studies at the precise applicability of different summation rules, and our experiments have revealed that Minkowski summation with power m in the range of 2.5–3.0 provides a significantly better fit to the data than linear addition, Euclidean summation and MAX rules, where m would exactly be 1, 2 and ∞, respectively. We find systematic deviations from all three classical models for feature integration in natural auditory, visual and audio-visual stimuli.

(a) *Minkowski summation in feature integration*

Minkowski summation with $m = 3$–4 has long been used in the study of vision (e.g. [14,15]) to describe how multiple subthreshold visual stimuli sum towards an overall detection process. This may be consistent with Green [21], who found that the summation of pure tones in auditory experiments was less than expected on a power summation rule. The same Minkowski rule (typically with $m = 3$–4) has been used to model the detection of changes in natural visual images, when multiple tiny cues are contributed across very many visual channels or model neurons [16–18]. Such modelling has been extended to the perceived magnitudes of suprathreshold differences in natural images similar to those that we have described here [18,25]. We have shown here that the same sort of Minkowski exponent describes the perception of suprathreshold changes in naturalistic auditory stimuli, as well as visual changes. This suggests that the combination of cues, whether subthreshold for detection or suprathreshold for perceived differences or similarity judgements, follows one general rule (cf. [7]).

We have also shown that the same Minkowski summation rule describes the summation of auditory cues with visual ones in cross-modal experiments, for incongruous as well as congruous pairings. There is an interesting evaluation of multi-sensory integration by Laurienti *et al.* [27], who summarize the degree of summation of auditory and visual responses in the cat superior colliculus. Here, some neurons respond only to auditory stimuli, some only to visual stimuli and some to both (even when they are not necessarily congruous visual and auditory events). Laurienti *et al.* [27] estimate the overall population response of the superior colliculus (for comparison with functional magnetic resonance imaging studies (fMRI)) and report that the response to an auditory/visual stimulus combination is greater than the response to either alone, but is less than the arithmetic sum. The summing exponent seems to be consistent with the overall appearance of Euclidean or Minkowski summation with exponent like those we have fit in figure 2.

Minkowski summation with exponent m of 2.5–3 provides a convenient numerical description of the results of our present experiments, but it does not provide a physical or neural explanation for the cognitive processes involved. Vision scientists who model detection processes have called the Minkowski summation rule the 'probability summation model', presuming that the Minkowski exponent is a parameter associated with the steepness of the psychometric function for detection [15]. On the other hand, for the suprathreshold integration of binocular and binaural signals, the Minkowski exponent has been suggested to reflect the strength of reciprocal inhibition between two neurons prior to summation [28]. These interpretations do not seem to be immediately applicable to an overall suprathreshold sensation, in the context described above. In the following section, we speculate about what neural mechanisms might lead to the failure of Euclidean summation.

(b) *The Mahalanobis distance*

Euclidean summation has been widely discussed as a general model for cue summation (e.g. [6,7]). However, this straightforward rule, as well as the MAX operator, is contradicted by our empirical results where Minkowski summation with power m of 2.5–3 is clearly a better fit to the experimental data. It is possible that this value of the Minkowski exponent m is related to the amount of correlation between different neuronal signals responding to natural stimuli. Euclidean summation might be appropriate if activity is independent, as each neuron would convey a uniquely important signal. However, if responses were highly correlated, the information given by only one neuron would be sufficient and the MAX rule (where m is ∞) would apply. So, if the neuronal signals to natural stimuli are slightly correlated, then the most appropriate summating exponent should be only a little greater than 2. A Minkowski exponent between 2.5 and 3 therefore suggests some small degree of signal correlation between actual neuronal responses.

In this case, an appropriate measure of cue combination should therefore readily account for potential correlation in signals, and an adjusted Euclidean measure of distance between stimuli is needed: the Mahalanobis distance [29] with covariance parameter ρ. The Mahalanobis distance has one free parameter, like the Minkowski distance. For the case where we are summing just two cues, the following formulation for a measure of feature integration is based on the Mahalanobis distance [29] and its relation to the Euclidean sum (equation (3.1) with $m = 2$) is clear:

$$\begin{aligned}\text{Mahalanobis\_sum} &= R3_{\text{predicted}} \\ &= (R1^2 + R2^2 - 2 \times \rho \times R1 \times R2)^{1/2}\end{aligned} \tag{4.1}$$

where ρ is the correlation between the dimensions represented by $R1$ and $R2$; it is the covariance of the sensory messages, if the sensory dimensions each have the same overall variance. The true Mahalanobis distance would be given by equation (4.1) after division by $(1 - \rho^2)$ but we omit this division from our measure since we would have to apply the same scaling to the *measured* values of $R3$ as well. The term $2\rho \times R1 \times R2$ is the amount by which Euclidean summation overestimates (figure 2*a–d*) the rating to $R3$. For our four experiments, figure 2*m–p* plots the value of $R3$ (ordinate) predicted (equation (4.1)) by the Mahalanobis sum of $R1$ and $R2$ against the actual values of $R3$; the value of ρ shown in each panel was found for each of the four experiments separately by iteratively searching for the value that gave the least SSE and the 95 per cent

confidence intervals for the ρ values for experiments 1 (music), 2 (phonetic), 3 (visual scenes) and 4 (cross-modal) are 0.16–0.23, 0.07–0.15, 0.15–0.21 and 0.13–0.17, respectively. (Electronic supplementary material, figure S3 shows the separate Mahalanobis fits for the congruous and incongruous conditions in experiment 4.) The Mahalanobis fits are significantly superior to the Euclidean and MAX operator fits (all F-tests highly significant), and are about the same as for the Minkowski fits.

The Mahalanobis distance is closely related to the 'law of cosines' where 'ρ' is replaced by 'cos τ' (e.g. [28]). In this case, τ is a measure of the interaction between two non-orthogonal (i.e. correlated) sensory dimensions. Therefore, the present data might also represent an interaction between two sensory dimensions within a non-orthogonal coordinate system. Although the present data fit both the Mahalanobis and Minkowski models nicely, we should bear in mind that there might exist conditions where the behaviour of the Minkowski sum and the Mahalanobis sum do not overlap; indeed, it is only for small values of ρ that a Minkowski exponent can give a satisfactory alternative fit.

(c) *Correlations in neural signals*

While a Minkowski exponent m of 2.5–3 has little neural meaning in cross-dimensional feature integration, the values (0.11–0.19) of the covariance parameter ρ for the Mahalanobis sum *do* have a potential neurophysiological significance. It has long been known that stimulus-evoked responses and spontaneous activity are correlated between neurons in the cerebral cortex [30], though Ecker *et al.* [31] argue that some of this correlation could be removed in well-controlled experiments. Indeed, the idea of widespread correlation is implicit in our understanding of the origin of the electroencephalogram, and such correlation probably underlies the spontaneous fluctuations in the BOLD signal seen in the connected cortical areas during fMRI [32]. Such widespread correlations might be related to changes in overall alertness or in attention to a task [33], and changes in neuronal responsiveness might lead to trial-by-trial apparent correlation of neural messages about truly-independent sensory dimensions. If attentional or other factors operate across large areas of cerebral cortex, we might even expect such correlations to be cross-modal and, of course, we have found the same putative correlations in our experiments between natural visual changes and natural auditory changes. Moreover, such widespread correlations of neural activity might explain why the summation of cross-modal stimuli was the same for incongruous pairings as for congruous ones.

The responses of sensory neurons are also likely to be correlated for several other, less global reasons. First, sensory input signals are likely to be correlated because information from the world is correlated. For instance, when we see a small elongated element in one part of our visual field, it is very likely that, beyond it and along its long axis, we may see elongated elements of very similar orientation [24] since the small elements are all part of one collinear or slightly curved object. A sharp luminance boundary will activate multiple visual cortex neurons whose receptive fields are of different spatial scales [34]. Secondly, visual receptive-field construction is not orthogonal and the visual cortex code is redundant [35,36] so that neurons' stimulus response-spaces overlap to some extent with the spaces of other neurons. Neurons close together within cerebral cortex are likely to respond to similar stimuli because of the columnar layout of the cortex [37] and also because nearby neurons share their (noisy) inputs and modulatory controls.

Thus, there are many reasons why we should expect the neural signals to paired stimuli to be correlated. Correlations in nearby stimulus features or in the receptive-field structure of nearby neurons with shared connectivity might explain why a Mahalanobis rule governs summation *within* audition or *within* vision. It is harder to see how such correlations explain why the same rule should govern cross-modal summation between audition and vision, especially in the incongruous case.

We still have an incomplete understanding of the magnitudes of typical correlation coefficients between neural responses, especially when the systems are studied with natural stimuli [38]. There have been many studies, particularly in various cortical areas of the visual system (e.g. [38–42]) and other cortical systems (e.g. [43]) that have measured the correlation in activity of simultaneously recorded neurons. The correlation coefficients actually vary widely between pairs of neurons and are generally highest for neurons recorded very close together. However, noticeable positive correlations have been reported even for neurons recorded more than 10 mm apart in cortex and for neurons that do not respond to the same visual features [41]. Some studies have tried to distinguish correlations in the trial-by-trial noise ('perceptual independence') from correlations in the underlying coded signals ('perceptual separability') by asking whether neuronal responses show overall correlations or whether the correlations are only at a trial-by-trial level of modulation. Despite the wide variety of behaviours, all these studies have generally reported typical or average correlations in the noise and the coded signals of about 0.2 (but see [31]). A typical inter-neural correlation of about 0.2 compares remarkably well with our estimates (figure 2 and electronic supplementary material, figure S3) of the signal correlations implicit in human sensory integration.

This research was supported by EPSRC/Dstl research grant (E037097/1 and EP/E037372/1) to D.J.T. and T.T., under the Joint Grants Scheme. M.P.S.T. was employed by that grant. We are very grateful for the helpful comments and suggestions of Prof. Nick Chater and his laboratory members, Dr Ian Cross and members of the Centre for Music and Science, Prof. Sarah Hawkins, Dr Jeroen Van Boxtel and Dr Patrick Rebuschat. We also thank Prof. Nigel Bennett and the three anonymous reviewers. The results from experiment 1 have been reported briefly in To *et al.* [20].

REFERENCES

1 von der Malsburg, C. 1995 Binding in models of perception and brain function. *Curr. Opin. Neurobiol.* **5**, 520–526. (doi:10.1016/0959-4388(95)80014-X)
2 Treisman, A. 1996 The binding problem. *Curr. Opin. Neurobiol.* **6**, 171–178. (doi:10.1016/S0959-4388(96)80070-5)
3 Ghose, G. M. & Maunsell, J. 1999 Specialized representations in visual cortex: a role for binding? *Neuron* **24**, 79–85. (doi:10.1016/S0896-6273(00)80823-5)
4 Wolfe, J. M. & Cave, K. R. 1999 The psychophysical evidence for a binding problem in human vision. *Neuron* **24**, 11–17. (doi:10.1016/S0896-6273(00)80818-1)

5 Ennis, D. M., Palen, J. J. & Mullen, K. 1988 A multidimensional stochastic theory of similarity. *J. Math. Psychol.* **32**, 449–465. (doi:10.1016/0022-2496(88)90023-5)
6 Shepard, R. N. 1964 Attention and the metric structure of stimulus space. *J. Math. Psychol.* **1**, 54–87. (doi:10.1016/0022-2496(64)90017-3)
7 Shepard, R. N. 1987 Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323. (doi:10.1126/science.3629243)
8 Li, Z. 2002 A saliency map in primary visual cortex. *Trends Cogn. Sci.* **6**, 9–16. (doi:10.1016/S1364-6613(00)01817-9)
9 Koene, A. R. & Zhaoping, L. 2007 Feature-specific interactions in salience from combined feature contrasts: evidence for a bottom-up saliency map in V1. *J. Vis.* **7**, 6: 1–14. (doi:10.1167/7.7.6)
10 Zhaoping, L. & May, K. A. 2007 Psychophysical tests of the hypothesis of a bottom-up saliency map in the primary visual cortex. *PLoS Comput. Biol.* **3**, e62. (doi:10.1371/journal.pcbi.0030062)
11 Riesenhuber, M. & Poggio, T. 1999 Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025. (doi:10.1038/14819)
12 Gawne, T. J. & Martin, J. M. 2002 Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophysiol.* **88**, 1128–1135.
13 Zhou, C. H. & Mel, B. W. 2008 Cue combination and color edge detection in natural scenes. *J. Vis.* **8**, 4: 1–25. (doi:10.1167/8.4.4)
14 Quick Jr, R. F. 1974 A vector-magnitude model of contrast detection. *Kybernetik* **16**, 65–67. (doi:10.1007/BF00271628)
15 Robson, J. G. & Graham, N. 1981 Probability summation and regional variation in contrast sensitivity across the visual field. *Vis. Res.* **21**, 409–418. (doi:10.1016/0042-6989(81)90169-3)
16 Rohaly, A. M., Ahumada Jr, A. J. & Watson, A. B. 1997 Object detection in natural backgrounds predicted by discrimination performance and models. *Vis. Res.* **37**, 3225–3235. (doi:10.1016/S0042-6989(97)00156-9)
17 Párraga, C. A., Troscianko, T. & Tolhurst, D. J. 2005 The effects of amplitude-spectrum statistics on foveal and peripheral discrimination of changes in natural images, and a multi-resolution model. *Vis. Res.* **45**, 3145–3168. (doi:10.1016/j.visres.2005.08.006)
18 Lovell, P. G., Parraga, C. A., Troscianko, T., Ripamonti, C. & Tolhurst, D. J. 2006 Evaluation of a multiscale color model for visual difference prediction. *ACM Trans. Appl. Percept.* **3**, 155–178. (doi:10.1145/1166087.1166089)
19 To, M., Lovell, P. G., Troscianko, T. & Tolhurst, D. J. 2008 Summation of perceptual cues in natural visual scenes. *Proc. R. Soc. B* **275**, 2299–2308. (doi:10.1098/rspb.2008.0692)
20 To, M. P. S., Troscianko, T. & Tolhurst, D. J. 2009 Music and natural image processing share a common feature-integration rule. In *Proc. 31st Annual Conf. of the Cognitive Science Society* (eds B. C. Love, K. McRae & V. M. Sloutsky), pp. 2481–2486. Austin, TX: Cognitive Science Society.
21 Green, D. M. 1958 Detection of multiple component signals in noise. *J. Acoust. Soc. Am.* **30**, 904–911. (doi:10.1121/1.1909400)
22 Meng, A., Ahrendt, P. & Larsen, J. 2005 Improving music genre classification by short-time feature integration. *IEEE Int. Conf. Acoust., Speech, and Signal Process.* **5**, 497–500.
23 Oden, G. C. & Massaro, D. W. 1978 Integration of featural information in speech perception. *Psychol. Rev.* **85**, 172–191. (doi:10.1037/0033-295X.85.3.172)
24 Geisler, W. S., Perry, J. S., Super, B. J. & Gallogly, D. P. 2001 Edge co-occurrence in natural images predicts contour grouping performance. *Vis. Res.* **41**, 711–724. (doi:10.1016/S0042-6989(00)00277-7)
25 To, M. P. S., Lovell, P. G., Troscianko, T. & Tolhurst, D. J. 2010 Perception of suprathreshold naturalistic changes in colored natural images. *J. Vis.* **10**, 12: 1–22. (doi:10.1167/10.4.12)
26 Kruskal, J. B. 1964 Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27. (doi:10.1007/BF02289565)
27 Laurienti, P. J., Perrault, T. J., Stanford, T. R., Wallace, M. T. & Stein, B. E. 2005 On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. *Exp. Brain Res.* **166**, 289–297. (doi:10.1007/s00221-005-2370-2)
28 Lehky, S. R. 1983 A model of binocular brightness and binaural loudness perception in humans with general applications to nonlinear summation of sensory inputs. *Biol. Cybernet.* **49**, 89–97. (doi:10.1007/BF00320389)
29 Ashby, F. G. & Perrin, N. A. 1988 Toward a unified theory of similarity and recognition. *Psychol. Rev.* **95**, 124–150. (doi:10.1037/0033-295X.95.1.124)
30 Li, C. L. 1959 Synchronization of unit activity in the cerebral cortex. *Science* **129**, 783–784. (doi:10.1126/science.129.3351.783)
31 Ecker, A. S., Berens, P., Keliris, G. A., Bethge, M., Logothetis, N. K. & Tolias, A. S. 2010 Decorrelated neuronal firing in cortical microcircuits. *Science* **327**, 584–587. (doi:10.1126/science.1179867)
32 Fox, M. D. & Raichle, M. E. 2007 Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8**, 700–711. (doi:10.1038/nrn2201)
33 Moran, J. & Desimone, R. 1985 Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784. (doi:10.1126/science.4023713)
34 Marr, D. 1982 *Vision*. San Francisco, CA: W. H. Freeman.
35 Olshausen, B. A. & Field, D. J. 1997 Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* **37**, 3311–3325. (doi:10.1016/S0042-6989(97)00169-7)
36 Willmore, B. & Tolhurst, D. J. 2001 Characterising the sparseness of neural codes. *Network, Comput. Neural Syst.* **12**, 255–270.
37 Hubel, D. H. & Wiesel, T. N. 1962 Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243.
38 Yen, S. C., Baker, J. & Gray, C. M. 2007 Heterogeneity in the responses of adjacent neurons to natural stimuli in cat striate cortex. *J. Neurophysiol.* **97**, 1326–1341. (doi:10.1152/jn.00747.2006)
39 Gawne, T. J. & Richmond, B. J. 1993 How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **13**, 2758–2771.
40 Reich, D. S., Mechler, F. & Victor, J. D. 2001 Independent and redundant information in nearby cortical neurons. *Science* **294**, 2566–2568. (doi:10.1126/science.1065839)
41 Smith, M. A. & Kohn, A. 2008 Spatial and temporal scales of neuronal correlation in primary visual cortex. *J. Neurosci.* **28**, 12591–12603. (doi:10.1523/JNEUROSCI.2929-08.2008)
42 Zohary, E., Shadlen, M. N. & Newsome, W. T. 1994 Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143. (doi:10.1038/370140a0)
43 Petersen, R. S., Panzeri, S. & Diamond, M. E. 2001 Population coding of stimulus location in rat somatosensory cortex. *Neuron* **32**, 503–514. (doi:10.1016/S0896-6273(01)00481-0)