SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PSYCHOLOGY
PERCEPTION
ATTENTION
DECISION

Received
11 June 2013

Accepted
15 November 2013

Published
6 December 2013

Correspondence and requests for materials should be addressed to W.R. (wjroseboom@gmail.com)

The cross-modal double flash illusion depends on featural similarity between cross-modal inducers

Warrick Roseboom, Takahiro Kawabe & Shin'ya Nishida

NTT Communication Science Laboratories, Kanagawa, Japan.

Despite extensive evidence of the possible interactions between multisensory signals, it remains unclear at what level of sensory processing these interactions take place. When two identical auditory beeps (inducers) are presented in quick succession accompanied by a single visual flash, observers often report seeing two visual flashes, rather than the physical one — the double flash illusion. This compelling illusion has often been considered to reflect direct interactions between neural activations in different primary sensory cortices. Against this simple account, here we show that by simply changing the inducer signals between featurally distinct signals (e.g. high- and low-pitch beeps) the illusory double flash is abolished. This result suggests that a critical component underlying the illusion is perceptual grouping of the inducer signals, consistent with the notion that multisensory combination is preceded by determination of whether the relevant signals share a common source of origin.

The cross-modal double flash illusion (DFI; often referred to as the sound induced flash illusion) occurs when two brief auditory or tactile events (inducers) are presented in quick succession ($<\sim$120 ms) accompanied by a single visual flash (target). Under these conditions, observers are inclined to report that two, rather than one, visual flashes had occurred[1,2]. This compelling illusion had a huge impact on subsequent multisensory research as it appeared to provide clear behavioural evidence of the strong direct interactions possible between primary sensory cortices. A simple neural account of the illusion implies that activation of auditory cortex in near temporal synchrony with visual cortex activation produces non-veridical visual cortex activation for an illusory additional flash[3–6]. This co-activation may be facilitated by direct neural projections between different sensory cortices (see[7] for review). According to this view, the critical stimulus parameters of the illusion are temporal proximity and the number of events of target and inducer.

Proposals regarding the computational structure underlying the DFI suggest that the illusion results from statistically optimal combination of inconsistent cross-modal signals[8,9]. When estimating the numerosity of visual multi-sensory signals, the brain gives a larger weighting to the modality that is more precise (a general quality shared with psychological accounts such as the modality appropriateness hypothesis[10,11]). As auditory perception is typically more precise than visual perception in the temporal domain, this process will often result in auditory dominance over vision when two auditory pips and one flash are combined.

In this study, we were interested in whether these apparently simple characterisations of the DFI are true. To address this issue, we wanted to further examine what types of sensory information contribute to the illusory flash percept. It has previously been demonstrated that the temporal relationship between signals is critical[1]. Other studies provide mixed results regarding the contribution of spatial relation[12,13]. Despite the fact that apparently equivalent phenomena have been demonstrated using many different combinations of sensory events, in all cases the inducer signals consist of repeating, featurally identical, signals from the same sensory modality as one another (e.g. visual target/tactile inducers[9,14]; visual target and inducers[13,15,16]; audio/visual with the roles reversed[11]; or audio inducers and tactile targets[17,18]). To date, no investigation has examined the role of featural relation between the inducer signals.

The results of a recent study[19] indicate the importance of featural relation among cross-modal signals in determining visual perception. That study examined the effect of a sequence of cross-modal events (auditory or tactile) on perception of a directionally ambiguous visual apparent motion sequence. As with the DFI, for this phenomenon both temporal[20,21] and spatial relation[22–26] had previously been demonstrated to contribute. The results demonstrated that organisation of the cross-modal event sequence on the basis of featural similarity alone could determine visual apparent motion perception. Featural similarity was manipulated both between sensory

modalities (auditory and tactile) and within a single sensory modality (pure tone and broadband noise auditory signals). In both cases, visual perception was determined by featural similarity among the cross-modal events. On the basis of these results, it was suggested that the role of the cross-modal event sequence may be to segment the visual event sequence into pairs determined by the apparent segmentation of the non-visual event sequence by featural (dis)similarity.

While the DFI concerns apparent visual numerosity rather than visual motion direction, it is possible that the role of the non-visual cues is similar in both cases – to disambiguate ambiguous visual perception. Consequently, it may be reasonable to predict that manipulations of featural similarity such as those used in the above cited study may also contribute in situations that typically induce the DFI.

Results

Is the featural similarity of inducer signals critical to the double flash illusion? To examine whether featural similarity of inducer signals is critical to the DFI, we used a stimulus similar to that typically used. Given that both auditory and tactile signals have separately been demonstrated to be effective in eliciting a DFI, in Experiment 1 we used different combinations of tactile and auditory events. As shown in Figure 1, there could be one or two visual events. These events could be accompanied by either one or two cross-modal events (auditory or tactile). When two cross-modal events were present, they could consist of either the same signal repeating twice (the same as previous DFI demonstrations; condition Same), or could switch between the two signal types (i.e. the first cross-modal signal could be auditory while the second would be tactile or the reverse relationship; condition Different). When two visual events were presented, their onsets were separated by 100 ms (see Methods for extensive experimental details). If the featural similarity of inducer signals is critical to the DFI, when the inducer signals are Different we would predict that the DFI is reduced compared with when they are the Same.

Shown in Figure 2A is the average number of visual flashes reported by five naïve participants and one author (WR). The critical conditions are those containing only a single visual flash. The single flash can be accompanied by one or two cross-modal inducer events. The DFI is typically revealed when there are two cross-modal events, and a single visual flash. This condition is outlined by a broken red line in Figure 2 for emphasis. Comparisons of the different signal types showed no difference in reports for any condition between the auditory and tactile inducers, or the different directions of alternation (i.e. tactile first followed by auditory or the reverse; results not shown) and so the presented data is in each case collapsed across these conditions.

To test whether similarity between inducer events is critical to the DFI, we conducted Friedman's analysis of variance by rank comparing reports in the six different conditions (Shapiro-Wilk tests showed that data in some conditions was not normally distributed: 1 flash/1 cross-modal, $p = 0.57$; 1 flash/2 cross-modal Same, $p = 0.1$; 1 flash/2 cross-modal Different, $p = 0.27$; 2 flash/1 cross-modal, $p = 0.19$; 2 flash/2 cross-modal Same, $p = 0.03$; 2 flash/2 cross-modal Different, $p = 0.06$). This analysis revealed a significant difference among these conditions ($\chi^2_5 = 26.15$, $p < 0.01$). Directly addressing the DFI, and the role of similarity between the cross-modal inducer events, comparisons revealed that when a single visual flash is accompanied by a pair of identical cross-modal events participants reported more flashes (1 flash/2 cross-modal Same = 1.66 +/− 0.14) than when the flash is accompanied by a single cross-modal event (1 flash/1 cross-modal = 1.09 +/− 0.02; $t_5 = 4.34$, $p < 0.01$, Cohen's $d = 2.23$; paired samples). This result is consistent with previous reports[1,2]. However, the number of reported flashes in the Same cross-modal inducer condition was also significantly greater than when the two cross-modal inducers were different (i.e. one was tactile and one was auditory noise; 1 flash/2 cross-modal Different = 1.13 +/− 0.05; $t_5 = 4.36$, $p < 0.01$, Cohen's $d = 2.04$; paired samples). An

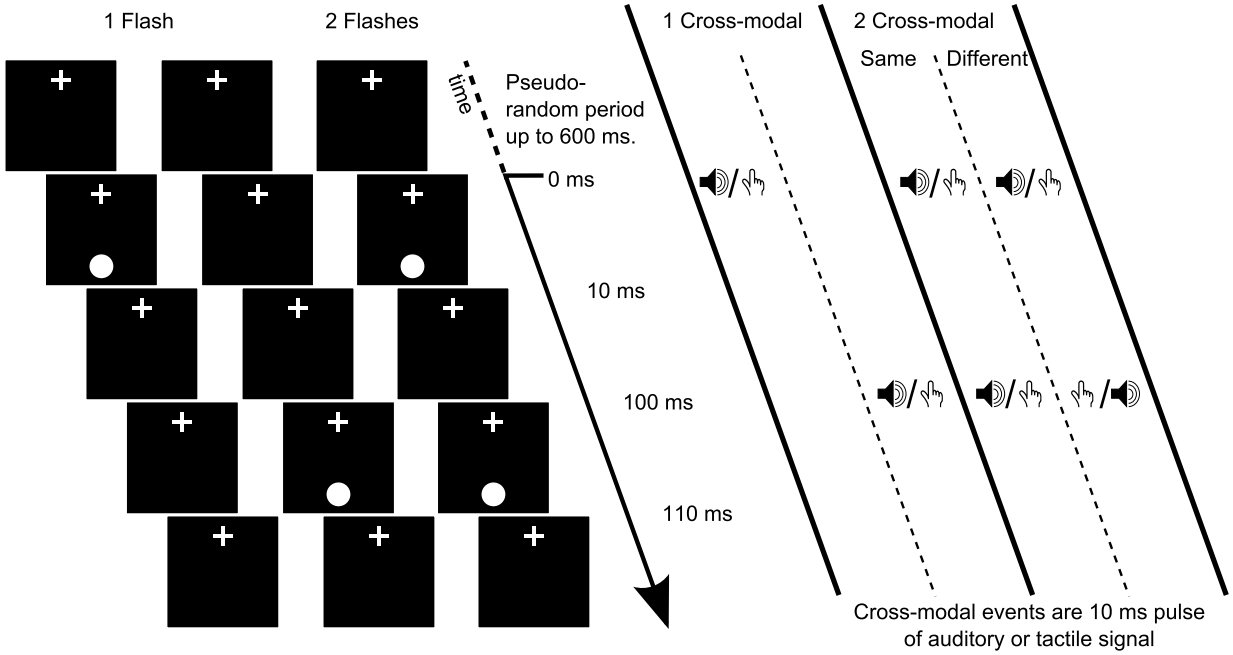

Figure 1 | Depiction of the stimulus used in Experiment 1. Each trial presentation began with a pseudo-random period of up to 600 ms where only the fixation cross was presented. The visual stimulus was a white disc presented for 10 ms. There could be either one or two visual presentations. There could also be either one or two cross-modal events (inducers). These could be auditory or tactile signals. When there were two cross-modal events, they could both be the same signal type (Same), or one could be auditory and the other tactile (Different). When there was only one visual and one cross-modal event, they could be synchronous, or the cross-modal event could lead or trail the visual event by 100 ms. When there was one visual event and two cross-modal events, the visual event could be synchronous with either the first or second cross-modal event. When there were two visual and two cross-modal presentations they were always presented as two successive synchronous visual/cross-modal pairs separated by 100 ms.

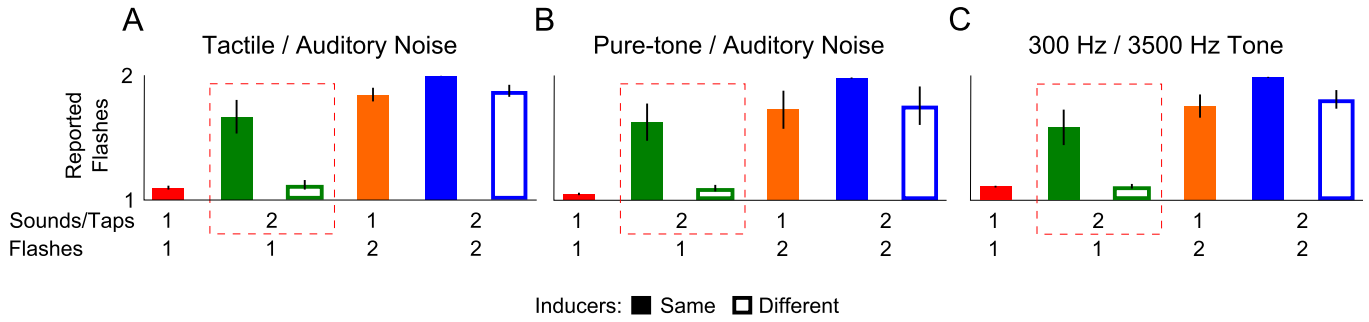

Figure 2 | (A–C) Bar plots depicting the mean number of reported flashes in Experiment 1 and 2 for six participants. (A) Data from Experiment 1 where a Tactile/Auditory Noise stimulus combination was used. (B–C) Data from Experiment 2 where Pure-tone/Auditory Noise and 300 Hz/3500 Hz Tone combinations were used. In all cases there could be either one or two visual flashes that could be accompanied by one or two cross-modal events. When two cross-modal events were presented, they could be either the same (e.g. both auditory or both tactile) or different signals (e.g. tactile synchronous with first visual flash and auditory noise synchronous with second or vice versa). For each stimulus combination the data outlined in the broken red line indicates the condition under which the DFI is typically obtained. Regardless of stimulus combination, a strong DFI was found when the two cross-modal events were the same, though was abolished when they were different. Error bars indicate +/− standard error of the mean.

additional comparison confirmed that the number of reported flashes when the cross-modal inducers were different was also not significantly different from that when only a single cross-modal event was presented ($t_5 = 0.91$, $p = 0.41$; paired samples). See Supplemental Materials for an additional experiment, Supplemental Experiment 1, using different timing conditions.

Feature or sensory modality based? The results of Experiment 1 suggest that similarity between sequential inducer events is critical to inducing the DFI. When the cross-modal inducer signals did not match (Different cross-modal signal conditions), the DFI was completely abolished. However, in the Different cross-modal signal conditions, the switch in signal types was between sensory modalities, from audition to tactile (or vice versa). This leaves the possibility of several alternative explanations rather than the simple effect of featural similarity of the inducer events. First, it may be that switching between the two sensory modalities contains some additional attentional cost that may prevent determination of the relationship between the different events. Alternatively, the obtained result is largely consistent with what is expected under a statistically optimal combination of the three events. When the inducer events differ, there is only a single event presented in each sensory modality. Combination across the three modalities would indicate the presence of only a single tri-modally presented event (i.e. $(1 + 1 + 1)/3 = 1$). This outcome has previously been reported for stimulus arrangements using slightly different temporal properties[9]. To investigate these alternative possibilities, we examined two scenarios in which the inducer signal changed though remained within the same modality (audition). If a similar pattern of results are found when both inducer signals are presented from the same sensory modality but differ in feature, it would suggest that featural similarity of inducer events, rather than any explanation related to a switch between sensory modalities, is critical to the DFI.

In Experiment 2 all auditory signals consisted of a 10 ms pulse. We examined two different auditory signal combinations: a pure-tone and auditory noise signal combination and a 300 Hz and 3500 Hz pure tone combination. Different stimulus combinations were used in different blocks of trials. As shown in Figure 2B–C, the different auditory inducer combinations provided results similar to those found for auditory and tactile signal combinations. We conducted analyses similar to those described in Experiment 1 for each of the Pure-tone/auditory noise (PN), and 300 Hz/3500 Hz (PP) auditory signal combinations. Shapiro-Wilk tests again showed that data from some of these conditions was not normally distributed (1 flash/1 cross-modal$_{PN}$, $p = 0.57$; 1 flash/2 cross-modal Same$_{PN}$, $p = 0.08$; 1 flash/2 cross-modal Different$_{PN}$, $p = 0.56$; 2 flash/1 cross-modal$_{PN}$, $p = 0.03$; 2 flash/2 cross-modal Same$_{PN}$, $p = 0.04$; 2 flash/2 cross-modal Different$_{PN}$, $p = 0.01$; 1 flash/1 cross-modal$_{PP}$, $p = 0.73$; 1 flash/2 cross-modal Same$_{PP}$, $p = 0.34$; 1 flash/2 cross-modal Different$_{PP}$, $p = 0.98$; 2 flash/1 cross-modal$_{PP}$, $p = 0.07$; 2 flash/2 cross-modal Same$_{PP}$, $p = 0.20$; 2 flash/2 cross-modal Different$_{PP}$, $p = 0.35$). Friedman's analysis of variance by rank revealed a significant difference among the Pure-tone/auditory noise conditions ($\chi^2_5 = 24.76$, $p < 0.01$). A repeated measures analysis of variance also revealed significant differences among the 300 Hz/3500 Hz conditions ($F_{(1.14,5.69)} = 19.43$, $p < 0.01$, *partial* $\eta^2 = 0.8$; Greenhouse-Geisser correction for violation of sphericity). Contrasts regarding our hypothesis that inducer event similarity is important to the DFI revealed an identical pattern of results as those reported in Experiment 1. A strong DFI was found when the cross-modal inducers were the same type, for both the pure-tone/auditory noise signals (1 flash/1 cross-modal$_{PN}$ = 1.05 +/− 0.02; 1 flash/2 cross-modal Same$_{PN}$ = 1.62 +/− 0.16; $t_5 = 3.72$, $p = 0.01$, Cohen's $d = 2.07$; paired samples) and 300 Hz/3500 Hz pure-tone combinations (1 flash/1 cross-modal$_{PP}$ = 1.06 +/− 0.01; 1 flash/2 cross-modal Same$_{PP}$ = 1.58 +/− 0.15; $t_5 = 3.72$, $p = 0.01$, Cohen's $d = 1.99$; paired samples). Furthermore, comparing the number of reported flashes when there was 1 visual flash and 2 cross-modal inducers we found that for both the pure-tone/auditory noise and the 300 Hz/3500 Hz pure-tone combinations the number of reported flashes was significantly reduced when the two cross-modal events were different, compared to when they were the same (pure-tone/auditory noise; $t_5 = 3.55$, $p = 0.02$, Cohen's $d = 1.86$; 300 Hz/3500 Hz pure-tones; $t_5 = 3.58$, $p = 0.02$, Cohen's $d = 1.78$; paired samples). Finally, when the two cross-modal inducers were different (i.e. one was pure-tone and one was auditory noise) the number of reported flashes did not differ from that in the single cross-modal presentation for either the pure-tone/auditory noise signals (1 flash/2 cross-modal Different$_{PN}$ = 1.1 +/− 0.03; $t_5 = 1.39$, $p = 0.22$; paired samples) or the 300 Hz/3500 Hz pure-tone signals (1 flash/2 cross-modal Different$_{PP}$ = 1.06 +/− 0.01; $t_5 = 2.39$, $p = 0.06$; paired samples).

Discussion

The purpose of this study was to determine whether featural similarity between inducer signals was critical to the DFI. In the first experiment, we established that when the signal type of the two cross-modal inducers alternated between different sensory modalities, the DFI was abolished. In the second experiment we confirmed that equivalent effects also occurred when the two cross-modal signals originated from the same sensory modality but differed in feature (pure-tone and auditory noise or high and low frequency pure-tones; see also Supplemental Experiment 1 for data obtained under different timing

conditions). These results support the idea that featural similarity among inducer signals contributes critically to the DFI.

An interesting aspect to note regarding the stimuli used in this study is that, especially for the stimulus in which the two auditory signals were both pure-tones but differed in pitch, the stimulus manipulations were similar to those used in studies of perceptual organisation in the context of auditory streaming (grouping). Many factors, including the influence of top-down processes such as attention[27], have been demonstrated to contribute to the likelihood that a sequence of auditory events is perceived as a single continuous sequence or segregated into multiple perceptual streams (see[28] for recent review). However, one of the strongest cues to stream segregation is the basic stimulus properties. When using pure-tone auditory stimuli, increasing differences in the temporal frequency (pitch) produce clear stream segregation effects[29]. There is also evidence to suggest this is may be true even in single presentation stimuli, similar to that used in the present study[30]. Consequently, we believe that the effect of switching between feaurally different inducer events may be to change the basic perceptual organisation within the non-visual event stream in a conceptually similar fashion to that often described by studies of perceptual grouping in the auditory domain (or indeed perceptual grouping phenomena in vision such as visual apparent motion; see[31–33]). This speculative interpretation is consistent with the results of a recent study mentioned in the Introduction[19] and suggests that perceptual grouping among the inducer signals affects how those signals are combined with the visual signal(s) and thus the generation of the DFI.

The above proposal is also broadly consistent with the hierarchy of multisensory processing previously suggested in different contexts. Several studies have demonstrated that determination of within-modal perceptual grouping is critical to determination of the overall multisensory percept (e.g.[34–39]). This kind of processing hierarchy seems appropriate given that accurate estimation of cross-modal (or cross-attribute within a single sensory modality) relationships is impossible at much larger temporal offsets than those that are resolvable by the uni-modal mechanisms[40–42]. It also provides an interesting problem for existing proposals regarding the possible process underlying the DFI.

As mentioned in the Introduction, the DFI is well described by a statistically optimal combination strategy[8,9]. The optimal combination process has previously been placed within a broader hierarchy of causal inference processing[43,44]. In this hierarchy, optimal combination occurs between signals that are determined to have a common source of origin. Previously it has been shown that spatial proximity between cross-modal signals is a useful indicator for such source determination[43]. Based on the results presented in this study, we suggest that featural (dis)similarity of signals within a within-modal stimulus sequence is also a critical part of the source determination process. As mentioned above, this proposal is consistent with the long established literature on source segregation within the auditory modality (see[28]). Regarding the DFI in particular, the results presented here suggest that the simple computational structure previously suggested for the DFI[8,9] is insufficient for an accurate depiction of the phenomenon. For the DFI to occur, a pair of auditory (or tactile) stimuli should be perceived as coming from a common source of origin, to which the visual stimulus also belongs. Under these conditions, the observer combines the two auditory signals with one visual signal, resulting in the DFI. However, if the two auditory stimuli are perceived as coming from different sources, the observer combines the visual stimulus with only one of the auditory stimuli and no DFI results. Therefore, computational accounts of the DFI require an additional level of complexity in that source determination has to be accomplished prior to the optimal combination process, as has been shown to be true for multisensory spatial localisation[43] and has been suggested to be true of perceptual combination processes in general (see[44] for review). Our results demonstrate that this source determination occurs within-modally and can be accomplished using basic featural cues such as auditory pitch.

A final point of interest is whether after combination of the multisensory signals, the multisensory representation can feed back into the lower levels of uni-sensory representation. Some neurophysiological data supports the idea that neural regions often associated with general multisensory processing[3,4] and mechanisms of selective attention[45] may be critical to the DFI and that low level visual representations may be modulated in the presence of the illusion[3–6]. The existence of some kind of feedback system may also be supported by behavioural results. For example, it has been demonstrated that while the DFI is partially attributable to simple changes in decisional criterion, there also appears to be some change in visual sensitivity associated with the presence of the illusory flash[46].These results provide some evidence to support the notion that the final combined multisensory representation may play a role in determination of the lower level representations through feedback, though this issue certainly remains a matter of debate.

In this study we manipulated the relationship among inducer signals by changing the apparent featural correspondence. The results of previous studies[2,16] indicate that manipulations of temporal proximity are also effective in decreasing the apparent correspondence between inducers, while spatial correspondence may also be an effective cue[13] (though see also[12]). The DFI has previously been supposed to represent a basic example of cross-modal processing. That featural, along with temporal and spatial, information is a key determinant of the DFI suggests this conception to be untrue. Rather, the DFI appears to be subject to the complex interactions between spatial, temporal and featural properties of sensory signals, along with top-down processes such as attention[47], common to other multisensory interactions.

Methods

Experiment 1. Participants included one of the authors (WR) and five participants who were naïve as to the experimental purpose. All reported normal or corrected to normal vision and hearing. Naïve participants received ¥1000 per hour for their participation. Ethical approval for this study was obtained from the ethical committee at Nippon Telegraph and Telephone Corporation (NTT Communication Science Laboratories Ethical Committee). The experiments were conducted according to the principles laid down in the Helsinki Declaration. Written informed consent was obtained from all participants except the authors.

Visual stimuli were generated using a VSG 2/3 from Cambridge Research Systems (CRS) and displayed on a 21″ Sony Trinitron GDM-F520 monitor (resolution of 800 × 600 pixels and refresh rate of 100 Hz). Participants viewed visual stimuli from a distance of ~105 cm. Audio signals were presented via a loudspeaker at a distance of ~60 cm, while tactile signals were presented via a vibration generator (EMIC Corp.) placed at a distance of ~50 cm from the participant. Participants placed their right arm on a cushioned arm-rest and rested their finger on the vibration generator. Audio and tactile stimulus presentations were controlled by a TDT RM1 Mobile Processor (Tucker-Davis Technologies). Auditory presentation timing was driven via a digital line from a VSG Break-out box (CRS), connected to the VSG, which triggered the RM1. Participants responded using a CRS CT3 response box.

Stimulus and procedures. The visual stimulus consisted of a white (CIE 1931 x = 0.297, y = 0.321, 123 cd/m^2) disc (0.4 degrees of visual angle in diameter) centered 4.75 dva below a white central fixation point (0.25 dva in width and height) against a black (~0 cd/m^2) background (see Figure 1, for depiction). Visual stimulus presentations were 10 ms in duration. Broadband auditory noise was presented continuously throughout the experiment at ~65 db SPL to mask any audible noise produced by the tactile stimulator. Auditory signals consisted of a 10 ms pulse, including 1 ms cosine onset and offset ramps of a transient amplitude increase in the broadband noise (~70 db SPL). Tactile signals consisted of a 10 ms, pulse containing 1 ms cosine onset and offset ramps, of the vibration generator driven at 100 kHz.

Each trial was preceded by a pseudo-random period of up to 600 ms where only the fixation cross-hair was presented. Regarding the visual stimulus, there were two types of presentations, one flash, or two flashes. When two flashes appeared, their onsets were separated by 100 ms. The visual flashes could be accompanied by either one or two cross-modal (audio or tactile) events. When there was only a single visual and single cross-modal event, on 50% of trials they occurred synchronously, while on 25% of trials the cross-modal event occurred prior to the visual event by 100 ms and on the final 25% of trials the cross-modal event occurred following the visual event by 100 ms. When there were two visual events and a single cross-modal event, the cross-modal event occurred synchronously with the first presented visual event on 50% of trials and synchronously with the second presented visual event on the other 50% of

trials. Similarly, when there were two cross-modal events and a single visual event, the single visual event occurred synchronously with the first presented cross-modal event on 50% of trials and with the second cross-modal event on the other 50% of trials. When there were two of each visual and cross-modal events, they always appeared as two synchronous cross-modal/visual pairs separated by 100 ms.

When only one cross-modal event was presented, on 50% of these trials the event was tactile and on 50% it was auditory noise. In presentations where two cross-modal events were presented, there were two conditions: Same or Different. In the Same condition, on 50% of trials both events were tactile and on 50% both were auditory noise. In the Different condition, on 50% of trials auditory noise was presented first and tactile second, while the other 50% were the reverse order. Each block of trials consisted of 256 individual trials, 64 of which contained 1 visual and 1 cross-modal event, 64 which contained 1 visual and 2 cross-modal events (32 of Same and 32 of Different conditions), 64 which contained 2 visual and 1 cross-modal events, and 64 which contained 2 visual and 2 cross-modal events (32 of Same and 32 of Different conditions). The order of completion of the trials was pseudo-random. Participants completed two blocks of trials.

Experiment 2. The methods of Experiment 2 were identical to Experiment 1 with the following exceptions. Only auditory signals were used. All signals consisted of a 10 ms pulse, containing 1 ms cosine onset and offset ramps. In the pure-tone/auditory noise experiment the signals were either a transient amplitude increase in the auditory noise (as in Experiment 1) or a 1500 Hz sine-wave carrier pure-tone. In the 300 Hz/3500 Hz experiment, the signals were either a 300 Hz or 3500 Hz sine-wave carrier pure-tone. Different stimulus combinations were used in different blocks of trials.

1. Shams, L., Kamitani, Y. & Shimojo, S. Illusions. What you see is what you hear. *Nature.* **408**, 788 (2000).
2. Shams, L., Kamitani, Y. & Shimojo, S. Visual illusion induced by sound. *Cogn. Brain Res.* **14**, 147–152 (2002).
3. Watkins, S., Shams, L., Tanaka, S., Haynes, J. D. & Rees, G. Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage.* **31**, 1247–1256 (2006).
4. Watkins, S., Shams, L., Josephs, O. & Rees, G. Activity in human V1 follows multisensory perception. *Neuroimage.* **37**, 572–578 (2007).
5. Mishra, J., Martinez, A., Sejnowski, T. J. & Hillyard, S. A. Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *J. Neurosci.* **27**, 4120–4131 (2007).
6. Mishra, J., Martinez, A. & Hillyard, S. A. Effect of attention on early cortical processes associated with the sound-induced extra flash Illusion. *J. Cogn. Neurosci.* **22**, 1714–1729 (2009).
7. Driver, J. & Noesselt, T. Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgements. *Neuron.* **57**, 11–23 (2008).
8. Shams, L., Ma, W. J. & Beierholm, U. Sound induced flash illusion as an optimal percept. *Neuroreport.* **16**, 1923–1927 (2005).
9. Wozny, D. R., Beierholm, U. R. & Shams, L. Human trimodal perception follows optimal statistical inference. *J. Vis.* **8**, 1–11 (2008).
10. Welch, R. B. & Warren, D. H. Immediate perceptual response to intersensory discrepancy. *Psychol. Bull.* **88**, 638–667 (1980).
11. Andersen, T. S., Tiippana, K. & Sams, M. Factors influencing audiovisual fission and fusion illusions. *Cogn. Brain Res.* **21**, 301–308 (2004).
12. Innes-Brown, H. & Crewther, D. The impact of spatial incongruence on an auditory-visual illusion. *PLoS One.* **4**, e6450 (2009).
13. Bizley, J. K., Shinn-Cunningham, B. G. & Lee, A. K. C. Nothing is irrelevant in a noisy world: Sensory illusions reveal obligatory within-and across-modality integration. *J. Neurosci.* **32**, 13402–13410 (2012).
14. Violentyev, A. C. A., Shimojo, S. & Shams, L. Touch-induced visual illusion. *Neuroreport.* **16**, 1107–1110 (2005).
15. Chatterjee, G., Wu, D. A. & Sheth, B. R. Phantom flashes caused by interactions across visual space. *J. Vis.* **11**(2): 14, 1–14 (2011).
16. Apthorp, D., Alais, D. & Boenke, L. T. Saccade adaptation goes for the goal. *J. Vis.* **13**(5): 3, 1–15 (2013).
17. Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V. & Kheddar, A. Feeling what you hear: Auditory signals can modulate tactile tap perception. *Exp. Brain Res.* **162**, 172–180 (2005).
18. Hötting, K. & Röder, B. Hearing cheats touch, but less in congenitally blind than in sighted individuals. *Psychol. Sci.* **15**, 60–64 (2004).
19. Roseboom, W., Kawabe, T. & Nishida, S. Direction of visual apparent motion driven by perceptual organization of cross-modal signals. *J. Vis.* **13**(1): 6, 1–13 (2013).
20. Freeman, E. & Driver, J. Direction of visual apparent motion driven solely by timing of a static sound. *Curr. Biol.* **18**, 1262–1266 (2008).
21. Kafaligonul, H. & Stoner, G. R. Auditory modulation of visual apparent motion with short spatial and temporal intervals. *J. Vis.* **10**(12): 31, 1–13 (2010).
22. Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C. & Kingstone, A. The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. Cogn. *Brain. Res.* **14**, 139–146 (2002).
23. Soto-Faraco, S., Spence, C. & Kingstone, A. Multisensory contributions to the perception of motion. *Neuropsychologia.* **41**, 1847–1862 (2003).
24. Sanabria, D., Soto-Faraco, S., Chan, J. & Spence, C. Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. *Neurosci. Lett.* **377**(1), 59–64 (2005).
25. Sanabria, D., Spence, C. & Soto-Faraco, S. Perceptual and decisional contributions to audiovisual interactions in the perception of apparent motion: A signal detection study. *Cognition.* **102**, 299–310 (2007).
26. Hidaka, S., Teramoto, W., Sugita, Y., Manaka, Y., Sakamoto, S. & Suzuki, Y. Auditory motion information drives visual motion perception. *PLoS One.* **6**, e17499 (2011).
27. Thompson, S. K., Carlyon, R. P. & Cusack, R. An objective measurement of the build-up of auditory streaming and of its modulation by attention. *J. Exp. Psychol.: Human Percept. Perf.* **37**, 1253–1262 (2011).
28. Moore, B. C. J. & Gockel, H. Properties of auditory stream formation. *Phil. Trans. R. Soc. B.* **367**, 919–931 (2012).
29. Rose, M. M. & Moore, B. C. J. Effects of frequency and level on auditory stream segregation. *J. Acoust. Soc. Am.* **108**, 1209–1214 (2000).
30. Strybel, T. Z. & Menges, M. L. Auditory apparent motion between sine waves differing in frequency. *Perception.* **27**, 483–495 (1998).
31. Kolers, P. A. *Aspects of motion perception.* Oxford, UK: Pergamon (1972).
32. Ullman, S. *The interpretation of visual motion.* MIT Press, Cambridge, MA (1979).
33. Gepshtein, S. & Kubovy, M. The lawful perception of apparent motion. *J. Vis.* **7**(8): 9, 1–15 (2007).
34. Watanabe, K. & Shimojo, S. When sound affects vision: effects of auditory grouping on visual motion perception. *Psychol. Sci.* **12**, 109–116 (2001).
35. Roseboom, W., Nishida, S. & Arnold, D. H. The sliding window of audio-visual simultaneity. *J. Vis.* **9**(12): 4, 1–8 (2009).
36. Cook, L. A. & Van Valkenburg, D. L. Audio-visual organisation and the temporal ventriloquism effect between grouped sequences: evidence that unimodal grouping precedes cross-modal integration. *Perception.* **38**(8), 1220–33 (2009).
37. Klink, P. C., Montijn, J. S. & van Wezel, R. J. A. Crossmodal duration perception involves perceptual grouping, temporal ventriloquism, and variable internal clock rates. *Atten Percept. Psychophys.* **73**, 219–236 (2011).
38. Roseboom, W., Nishida, S., Fujisaki, W. & Arnold, D. H. Audio-Visual Speech Timing Sensitivity Is Enhanced in Cluttered Conditions. *PLoS One.* **6**(4), e18309 (2011).
39. Kawabe, T., Roseboom, W. & Nishida, S. The sense of agency is action-effect causality perception based on cross-modal grouping. *Proc. R. Soc. Lond. B.* **280** (2013).
40. Fujisaki, W. & Nishida, S. Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Exp. Brain Res.* **166**, 455–464 (2005).
41. Fujisaki, W. & Nishida, S. Audio-tactile superiority over visuo-tactile and audio-visual combinations in the temporal resolution of synchrony perception. *Exp. Brain Res.* **198**, 245–259 (2009).
42. Fujisaki, W. & Nishida, S. A common perceptual temporal limit of binding synchronous inputs across different sensory attributes and modalities. *Proc R. Soc Lond. B.* **277**, 2281–2290 (2010).
43. Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B. & Shams, L. Causal inference in multisensory perception. *PLoS One.* **2**, e943 (2007).
44. Shams, L. & Beierholm, U. R. Causal inference in perception. *Trends Cogn. Sci.* **14**, 425–432 (2010).
45. Kamke, M. R., Veith, H. E., Cottrell, D. & Mattingley, J. B. Parietal disruption alters audiovisual binding in the sound-induced flash illusion. *Neuroimage.* **62**, 1334–1341 (2012).
46. McCormick, D. & Mamassian, P. What does the illusory-flash look like? *Vis. Res.* **48**, 63–69 (2008).
47. Werkhoven, P. J., Van Erp, J. B. F. & Philippi, T. G. Counting visual and tactile events: The effect of attention on multisensory integration. *Atten Percept. Psychophys.* **71**(8), 1854–1861 (2009).

Author contributions

W.R., T.K. and S.N. contributed to conception and design of the experiments and wrote the manuscript. W.R. and T.K. collected and analysed the data.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors are employees of Nippon Telegraph and Telephone Communication Science Laboratories, which is a basic-science research section of Nippon Telegraph and Telecommunication. There are no patents, products in development or marketed products to declare.

How to cite this article: Roseboom, W., Kawabe, T. & Nishida, S. The cross-modal double flash illusion depends on featural similarity between cross-modal inducers. *Sci. Rep.* **3**, 3437; DOI:10.1038/srep03437 (2013).