

available at [www.sciencedirect.com](http://www.sciencedirect.com)[www.elsevier.com/locate/brainres](http://www.elsevier.com/locate/brainres)


---



---

**BRAIN  
RESEARCH**


---



---



---

**Research Report**

# Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence

Karin Petrini\*, Phil McAleer, Frank Pollick

Department of Psychology, University of Glasgow, 58 Hillhead Street, Glasgow G12 8QB, UK

---

**ARTICLE INFO**
**Article history:**

Accepted 3 February 2010

Available online 11 February 2010

**Keywords:**

Audiovisual integration

Emotion

Music

Psychophysics

**ABSTRACT**

In the present study we applied a paradigm often used in face–voice affect perception to solo music improvisation to examine how the emotional valence of sound and gesture are integrated when perceiving an emotion. Three brief excerpts expressing emotion produced by a drummer and three by a saxophonist were selected. From these bimodal congruent displays the audio-only, visual-only, and audiovisually incongruent conditions (obtained by combining the two signals both within and between instruments) were derived. In Experiment 1 twenty musical novices judged the perceived emotion and rated the strength of each emotion. The results indicate that sound dominated the visual signal in the perception of affective expression, though this was more evident for the saxophone. In Experiment 2 a further sixteen musical novices were asked to either pay attention to the musicians' movements or to the sound when judging the perceived emotions. The results showed no effect of visual information when judging the sound. On the contrary, when judging the emotional content of the visual information, a worsening in performance was obtained for the incongruent condition that combined different emotional auditory and visual information for the same instrument. The effect of emotionally discordant information thus became evident only when the auditory and visual signals belonged to the same categorical event despite their temporal mismatch. This suggests that the integration of emotional information may be reinforced by its semantic attributes but might be independent from temporal features.

© 2010 Elsevier B.V. All rights reserved.

---

**1. Introduction**

In everyday life we need to respond appropriately to disparate affective signals in order to successfully behave and interact in our social environment. To reach this goal we need to integrate different kinds of non-verbal affective information such as face and body expressions as well as voice and sound prosody. Thus both non-verbal visual and auditory information needs to be fused to create a unique emotional percept.

Several studies have examined the effect of audiovisual integration on perceived emotions by using face–voice stimuli (Massaro and Egan, 1996; De Gelder and Vroomen, 2000; Kreifelts et al., 2007; Collignon et al., 2008; Campanella and Belin, 2007), showing that under normal conditions the visual information dominates the auditory when judging an emotion (Collignon et al., 2008), and that the two affective signals are integrated even when emotionally discordant (Massaro and Egan, 1996; De Gelder and Vroomen, 2000).

---

\* Corresponding author. Fax: +44 141 330 4606.

E-mail address: [karin@psy.gla.ac.uk](mailto:karin@psy.gla.ac.uk) (K. Petrini).

On the contrary, very few studies have investigated the same process for body–sound by means of music. Indeed, although there has long been interest in how music communicates different emotions (Scherer, 1995; Juslin and Sloboda, 2001), and much more recently an interest in how the musician's body movement alone can affect the perceived emotion (Laukka and Gabrielsson, 2000; Dahl and Friberg, 2007; Castellano et al., 2008; Davidson, 1993, 1994), so far little interest has been shown in the multimodal affective aspects of music performance, although music, beside speech and dance, is a very important tool of non-verbal social communication and interaction.

Vines et al. (2006) investigated the dynamics of sensory integration for perceiving musical performance and indicated that the expressive movements of two clarinetists influenced the way participants experienced tension during musical pieces when auditory and visual information were presented together. Thus the authors suggested that the integration of the two signals can give rise to an emergent quality of musical performance augmenting knowledge about communicative processes. The displays of Vines et al. (2006) included the head–face as well as the torso–arms of the musicians, making it difficult to discriminate between the effects of facial expressiveness as opposed to body expressiveness. We know that facial expressions can influence the judgements of emotion in song, showing enhanced emotional ratings when presented along with congruent auditory information (Thompson et al., 2008). Davidson (1994) also examined the impact of different body parts on perceived emotion in pianists' displays, showing that the head movements were the most influential in affecting musicians' expressiveness. However, Davidson (1994) used only pianists in her study, and in piano playing the head movements are not instrumental but expressive, while the torso and arm movements can be both instrumental and expressive. In other words the movements of head and face are not strictly required in order to technically perform the piece, while the torso and arm movements are. Thus, the expressive contribution of the different parts of the body to the multimodal percept might vary from instrument to instrument depending on whether or not they are an essential part of the action finalised to create the musical piece.

In a previous study by Davidson (1993) participants were found to be better at distinguishing three levels of performance expressiveness (no expressiveness, standard expressiveness, and exaggerated expressiveness), when only viewing the performers. This result is in contrast with that of Vines et al. (2006) which showed that sound plays a dominant role in determining the contour and trajectory of emotional experience. The discrepancy in these results might be easily explained by differences in task instructions; indeed, asking for either a “tension” judgement or an “expressivity” rating might lead to a different impact of audio-only or visual-only information, respectively (Vines et al., 2006). Thus it is important to use a task that does not differentiate a priori between the relevance of one or the other signal if one wants to achieve a better understanding of how auditory and visual cues integrate when perceiving emotions. Nevertheless, in line with the findings of Davidson (1993), as well as Broughton and Stevens (2009) on perceived expressiveness, Vines et al.

(2006) showed that visual information affects perceived tension, and in some cases (over short durations of time) does so in a way similar to sound (as shown by similar contours of the audio-only and visual-only tension curve). Finally, Vines et al. (2006) reported, in line with the face–voice literature (Collignon et al., 2008), that the audiovisual condition enhanced the perceived tension with respect to the audio-only and visual-only condition; furthermore, that when the auditory and visual information from the performers were discordant (for example, when the performer was visually very expressive in playing a somewhat subdued and quiet sound), the tension perceived by the participants increased from that perceived when only hearing the sound. This result suggested that participants were integrating the two signals even when emotionally discordant, instead of basing their judgements on one signal.

In the current study we aim to examine the nature of the multimodal processes for non-verbal emotional signals by using three-second solo improvisation displays of an experienced drummer and a saxophonist. The decision to use solo improvisations instead of existing musical excerpts was taken to avoid any effect of familiarity on observers' emotional judgements as well as the presentation of incomplete musical ideas resulting from selecting short excerpts from longer compositions. Additionally, using improvisations allows the researcher to know the emotional intention of the composer–performer (Behrens and Green, 1993), allowing strict control over the stimulation. Moreover, the decision to use two different instruments was based firstly on evidence that different instruments are better at expressing different emotions (Behrens and Green, 1993), and secondly that the upper-body movements (Broughton and Stevens, 2009) permitted by drumming are much more evident and less expressively restricted than those permitted by the saxophone: in addition it is noted that torso and arm movements are instrumental in both cases.

Finally, given that one experimental question was what the perception of emotion from the fusion of body expressiveness and sound has in common with that of facial expressions and voice prosody, we used an experimental design often used in emotional face–voice studies (Massaro and Egan, 1996; De Gelder and Vroomen, 2000; Collignon et al., 2008; Hietanen et al., 2008). Participants were presented with audiovisually congruent, audio-only, visual-only and audiovisually incongruent conditions. In the incongruent conditions the visual and auditory signals were always mismatching temporally (i.e. the musicians' movements did not correlate to the sound), but only some of them were also mismatching semantically (i.e. the visual stimulation from the saxophone was combined with the auditory stimulation from the drum or vice versa). This differentiation between incongruent stimuli was important to understand the importance of both temporal and semantic correspondence in the audiovisual integration of emotional signals. We chose to use the bimodal incongruent conditions, beside the congruent and unimodal, because we wanted to disentangle the effect of incongruent non-emotional factors (e.g. musical instrument) from the effect of emotional incongruence. Using other kinds of paradigms to manipulate the reliability of the two signals, such as the noise paradigm (Landy et al., 1995; Ernst and Banks, 2002; Collignon

et al., 2008) would not have been optimal for this purpose because the two incongruent signals would have been affected in the same way (e.g. adding noise would have degraded the non-emotional incongruent information as much as the emotional incongruent information). Up to now it has been completely unclear whether emotional, semantic and temporal characteristics are integrated together or separately, and this is the first attempt to give an answer to this important aspect of multisensory integration. To this end, in a first experiment participants were asked to rate and categorise each display for six emotions universally recognised in facial expressions (happiness, sadness, anger, fear, disgust, surprise: Ekman and Friesen, 1975) and a neutral. Additionally, in a second experiment participants had to pay attention to only one modality and disregard the other in two separated and counterbalanced blocks, to allow us to further understand if the irrelevant information would affect the emotional content of the attended modality, and if so, whether this effect would change with the stimulus type (e.g. bimodal incongruent for instrument and emotion vs. bimodal incongruent only for emotion).

## 2. Results

### 2.1. Experiment 1

Binomial analysis for the audiovisual congruent condition of each stimulus category revealed that the intended emotion was chosen at a level above chance (14.28%) for all the three saxophone stimuli (happiness:  $p < .001$ ; sadness:  $p < .001$ ; surprise:  $p < .001$ ), for the drummer stimulus representing anger ( $p < .001$ ) and for a neutral emotion ( $p < .001$ ). The stimulus for the drummer representing the happy emotion did not reach significance ( $p = .054$ ), and so this display was not further investigated, though its data were used for the analysis of the other displays.

The analysis was carried out separately on the averaged responses and ratings<sup>1</sup>, after the data were transformed in such a way as to eliminate eventual biases in the emotional responses. To this end, the responses and ratings for a target emotion (e.g. happiness responses and ratings for the happy display: see Supplementary material online) were adjusted by subtracting from them the average responses and ratings for the nontarget emotions (e.g. happiness responses and ratings for the non-happy displays). This transformation was carried out on a subject-wise basis before averaging across the different emotions for each instrument category.

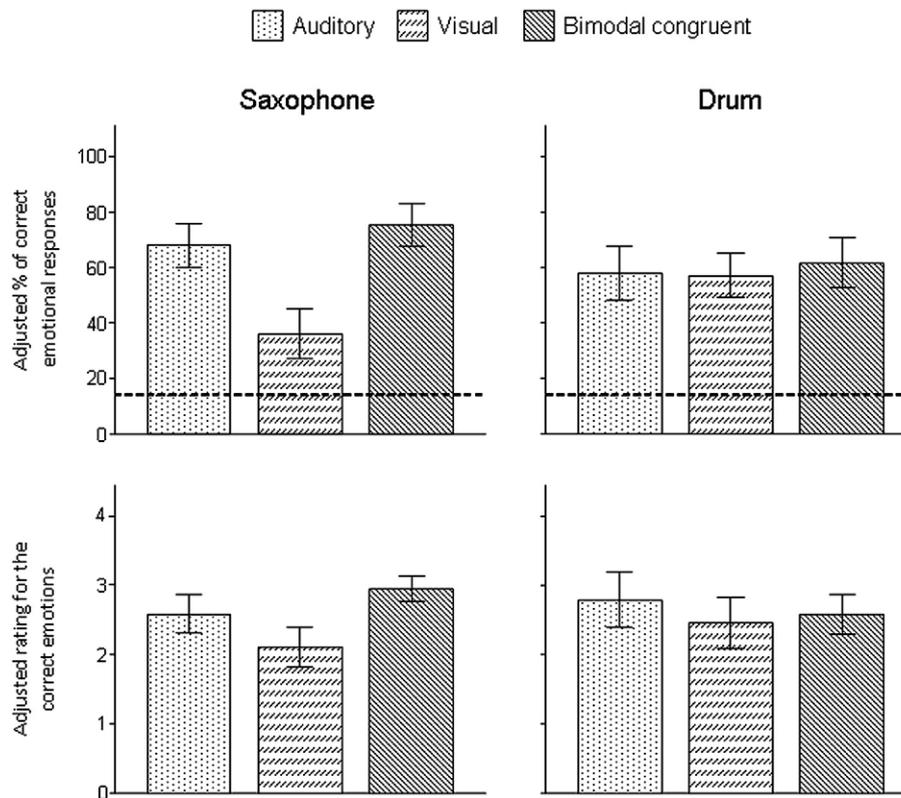
The adjusted correct discriminations and ratings were analysed by carrying out a 2 (Instrument: saxophone or drum)  $\times$  3 (Stimuli: visual, auditory or bimodal congruent) repeated measures ANOVA. Furthermore, to acquire knowledge about the differences between bimodal congruent and unimodal visual and auditory conditions a series of simple contrast measures was obtained and Bonferroni's corrected for two comparisons.

The ANOVA carried out on the correct emotional responses showed that the within factor "Instrument" did not reach significance ( $F(1, 19) = 0.54, p = .818$ ), while the factor "Stimuli" did ( $F(2, 18) = 14.923, p < .001$ ). Thus the two instruments received overall the same number of correct emotional responses, although this number varied with the signal modality (top diagrams in Fig. 1). A significant interaction between "Instrument" and "Stimuli" was also found ( $F(2, 18) = 4.597, p = .024$ ), indicating that some stimuli were more effective in eliciting the intended emotion for one of the two instruments. The simple contrast measures between the audiovisual congruent condition and the auditory-only condition showed no significant difference in correct emotional responses ( $F(1, 19) = 4.400, p = .100$ ), while the contrast between the audiovisual congruent condition and the visual-only condition showed a significant difference ( $F(1, 19) = 29.048, p < .001$ ) due to the smaller number of correct emotional responses received by the visual-only condition. Also the simple contrasts indicated that the significant interaction we found between the two within factors was due to the higher number of correct emotional responses received by the visual-only drum condition compared (top diagrams in Fig. 1) with the saxophone's visual-only condition ( $F(1, 19) = 9.335, p = .007$ ).

The same analysis of variance carried out on the ratings for the correct emotions showed that the within factor "Instrument" was once again not significant ( $F(1, 12) = .045, p = .835$ ), indicating that the two instruments received overall the same emotional rating for the correct emotions (bottom diagrams in Fig. 1). The within factor "Stimuli" was instead significant ( $F(2, 11) = 6.054, p = .017$ ), indicating that the emotional rating attributed to the displays depended on the modality of presentation. Finally, no significant interaction was found between the two within factors ( $F(2, 11) = 1.623, p = .241$ ), showing that the effect of audiovisual modality on emotional rating was similar for the two instruments. The simple contrast measures between the audiovisual congruent condition and the visual-only condition ( $F(1, 12) = 9.020, p = .011$ ) were significant, while the contrast between the audiovisual congruent condition and the auditory-only condition ( $F(1, 12) = .131, p = .723$ ) was not significant.

Because there are no "correct" responses or ratings per se for audiovisual incongruent stimuli, we calculated the tendency to respond correctly or rate the correct emotion when it was presented auditorily or visually. The tendency was estimated by subtracting the proportion of "visual correct" responses from the proportion of "auditory correct" responses for the two incongruent conditions: for instrument and emotion (e.g. auditory information from the sad saxophone together with the visual information from the anger drum and vice versa); for emotion (e.g. auditory information from the happy saxophone together with the visual information from the sad saxophone and vice versa) to investigate whether the semantic non-correspondence between the signals would further affect the emotional judgements. The estimated indices varied from 1 (subjects always responded correctly to the auditory information) to  $-1$  (subjects always responded correctly to the visual information) and were analysed by carrying out a *t* test to compare the two incongruent conditions. The results clearly showed a tendency toward choosing the emotion expressed

<sup>1</sup> Due to a technical problem the rating data of the first 7 subjects could not be included in the analysis.



**Fig. 1 – Adjusted percentage of correct emotional responses and adjusted ratings for the correct emotion received by the two instrument categories: (a) saxophone; (b) drum. The audiovisual conditions distinguish between bimodal congruent, auditory-only, visual-only. The error bars represent the standard error of the mean.**

by the auditory information (Fig. 2a). This tendency did not change for the two incongruent conditions ( $t(19)=-1.055$ ,  $p=.305$ ).

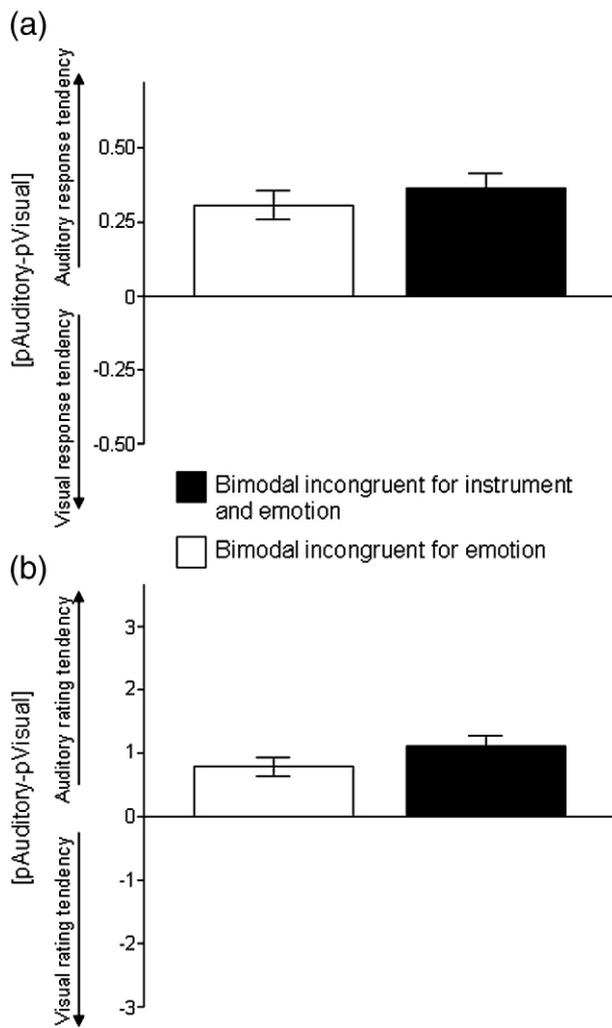
For the rating we estimated the tendency by subtracting the rating received by the correct visual emotion from the rating received by the correct auditory emotion for the two incongruent conditions. The estimated indices varied from 6 (subjects attributed the highest rating to the correct auditory information and the lowest to the visual) to  $-6$  (subjects attributed the highest rating to the correct visual information and the lowest to the auditory) and were analysed by carrying out a  $t$  test to compare the two incongruent conditions. The results once again showed a tendency toward rating as stronger the emotion expressed by the auditory information (Fig. 2b), which was similar for the two kinds of incongruency ( $t(12)=-1.599$ ,  $p=.136$ ).

Finally, we compared the response time (RT) for the audiovisual congruent, unimodal, and audiovisual incongruent conditions with a three level one-way ANOVA. This further analysis was carried out to examine whether participants took longer to judge the audiovisual incongruent displays than to judge the audiovisual congruent and unimodal displays. We found that participants took on average the same time ( $F(2, 18)=1.061$ ,  $p=.367$ ) to judge the emotional valence of congruent (RT mean:  $2.380s \pm .271$ ), unimodal (RT mean:  $2.242s \pm .190$ ), and incongruent displays (RT mean:  $2.558s \pm .210$ ).

## 2.2. Experiment 2

In this second experiment we took into account the response speed as well as the accuracy, in order to calculate the Inverse Efficiency (IE) scores, as in Collignon et al. (2008). These scores were derived by dividing the response times by correct response rates separately for each condition, carried out in such a way that the higher the score was, the worse was the performance. The IE scores averaged for each stimulus condition were submitted to a 2 (Attention: auditory signal attended, visual signal attended)  $\times$  4 (Stimuli: unimodal, bimodal congruent, bimodal incongruent for instrument and emotion, and bimodal incongruent for emotion) repeated measures analysis of variance. Furthermore, to acquire knowledge about the differences between the IE scores of the four stimuli categories, a series of repeated contrast measures was obtained and Bonferroni's corrected for three comparisons.

The ANOVA carried out on the IE scores showed that both the within factors "Attention" ( $F(1, 15)=6.305$ ,  $p=.024$ ), and "Stimuli" ( $F(3, 13)=4.209$ ,  $p=.028$ ) significantly affected the participants' performance. Fig. 3 shows that both significant results were caused by the much higher IE scores obtained, in the visual attended block, by the bimodal incongruent condition combining different emotional signals within the same instrument. To test this observation repeated contrast measures were carried out to compare the IEs received by the unimodal stimuli to those received by the bimodal congruent



**Fig. 2 – Tendency towards choosing the auditory and visual emotional content estimated by subtracting the proportion of “correct visual” responses from the proportion of “correct auditory” responses in Experiment 1. Participants tend to report the emotion expressed in the auditory modality for both incongruent conditions.**

( $F(1, 15) = .184, p = .674$ ), the bimodal congruent to the bimodal incongruent for instrument and emotion ( $F(1, 15) = .040, p = .844$ ), and finally the bimodal incongruent for instrument and emotion with the incongruent for emotion ( $F(1, 15) = 11.163, p = .012$ ). Further testing by means of repeated contrast measures also showed a significant interaction between “Attention” and “Stimuli” when comparing the two incongruent conditions ( $F(1, 15) = 6.216, p = .025$ ), confirming that this effect emerged only for the attended visual task.

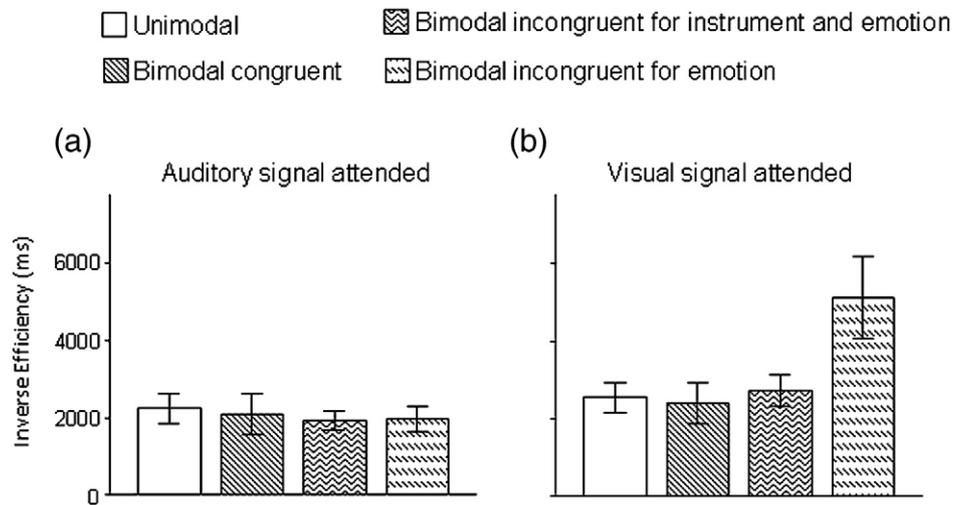
### 3. Discussion

In the present study, we asked participants to rate and discriminate the perceived emotion for brief solo improvisation musical excerpts displayed either auditorily, visually,

audiovisually congruently or audiovisually incongruently. Our results suggest that the way non-verbal affective body–sound signals from music performance are integrated is very similar to the way in which this occurs for affective face–voice signals (Collignon et al., 2008). Under normal conditions of sight and sound one of the two modalities (the visual or the auditory) dominates over the other, depending on the kind of stimulus being processed. This similarity, along with other demonstrated similarities in the way voice and music express emotion (Scherer, 1995; Juslin and Laukka, 2003), allow us to assume that if a specific brain network has developed to optimise our social behaviour by integrating emotional sound and sight from the environment (Kreifelts et al., 2007), then it might possibly subtend both body–sound and face–voice affective processes.

Participants were found to perceive correctly the intended emotion for all audiovisually congruent conditions but one (the happiness–drum condition). We think that this result for the happiness–drum condition was determined by the changes in the experimental context. Indeed, happiness was the only emotion represented by both instruments in the main experiment, while in the preliminary phase six different emotions (happiness, sadness, anger, fear, disgust, surprise), and neutrality were represented by both instruments. Thus it is probable that the smaller number of stimuli made it more difficult for participants to perceive the drumming display as being as happy as the saxophone display, thereby shifting the emotional content of the drumming display towards a neutral emotion (see Supplementary material online). This finding, together with the preliminary findings, adds evidence to the assumption that different instruments are better at expressing different emotions (Behrens and Green, 1993).

In line with Vines et al. (2006), and despite slight differences between the stimulus categories, we found that the auditory signal dominates in determining the perceived emotion, and that the visual information, when in agreement, appears to support and confirm the auditory information. In terms of cross-modal terminology this effect might be called “auditory capture” as opposed to the “visual capture” (Ernst and Bühlhoff, 2004) found by Collignon et al. (2008) in their face–voice emotion study. The dominance of the auditory signal over the visual in music performance might have different explanations. One possible explanation, as pointed out by Vines et al. (2006), is that the making of music requires a specific coupling between the performer and the instrument. The necessity to maintain this perfect coupling limits the movements that the musician can convey, and consequently the complexity of information delivered by the music sound (e.g. pitch, intensity, rhythm and so on, for review see Juslin and Sloboda, 2001) cannot often be expressed with body movements. However, some instruments limit body movements more than others. For example, the clarinetists in Vines et al. (2006), as well as the saxophonist in our study, had more limited arm movements when compared, for example, with drummers and pianists. Vines et al. (2006) suggested that it would be informative to compare instruments with different restrictions upon movements of the body to examine whether, when the movements of the musician are less restricted, they are able to deliver the same emotional content as that evoked by the sound.



**Fig. 3 – Mean IE scores and standard errors obtained in Experiment 2 for unimodal stimuli, bimodal congruent stimuli, bimodal incongruent stimuli semantically and emotionally, bimodal incongruent stimuli emotionally.**

**(a) Performance when participants were instructed to judge the emotion perceived auditorily. (b) Performance when participants were instructed to judge the emotion perceived visually.**

In our study, besides the saxophone, we examined the drum, a very different instrument that allows a much larger range of upper-body movements, and we showed that for this instrument the visual information alone was as good as the auditory information in eliciting the intended emotion. Thus it looks like that the privileged status of the auditory signal over the visual, when delivering emotion through music, has some exceptions. However, despite the higher reliability of the drum visual information than that of the saxophone, we found that the two instruments did not differ for the incongruent conditions. Indeed, for both instruments participants had the tendency to choose the emotion communicated by the auditory information when presented with visual information which was discordant for instrument and/or emotion. To further examine whether the difference in visual expressivity we found between instruments can completely be attributed to the instruments' characteristics rather than the difference in expressivity between the two musicians, a similar paradigm to ours could involve more musicians for each instrument as well as one musician playing different instruments. However, we already know from the study of Vines et al. (2006) that even though one clarinetist was found to be more expressive than the other, observers weighted the auditory and visual information from both musicians in a similar way when judging the perceived tension.

So far our findings are in line with the prediction of “inverse effectiveness”, which states that the result of multisensory integration is inversely proportional to the effectiveness of the relevant stimuli (Stein and Meredith, 1993; Collignon et al., 2008). That is, we failed to find a significant difference between the auditory and the audiovisual congruent condition because the auditory information was highly effective in delivering the intended emotions. Therefore, no evident integration between the signals was found when participants judged the affective content of the musical stimuli. Still, although the findings of both Experiment 1 and 2 support the dominance of the

auditory signal, the results of the second experiment indicate that, when forced to, participants are able to use the visual information and disregard the auditory information, with one exception. Indeed, when participants were free to use the most reliable information (the sound) to judge the perceived emotion, no effect of the visual information was found on both bimodal incongruent stimuli. However, in Experiment 2 when forced to use the less reliable information (the visual), the auditory information was found to affect the perceived emotion only when it belonged to the same event category, as dictated by the “Unity assumption” theory. The finding that the auditory information can affect the visual only when they are discordant emotionally within the same instrument category supports the claims of Vatakis and Spence (2007) that the “unity assumption” (the assumption made by an observer that two signals are part of the same multisensory event) can modulate the cross-modal binding of multisensory information at the perceptual level. That is, when observers assume that the sound and the sight belong to the same event because both represent a saxophone and a drum, then they find it more difficult to ignore the irrelevant signal, despite the temporal mismatch between the musicians' movements and the produced sound. On the contrary, when the two signals, in addition to being discordant emotionally, are also discordant semantically (e.g. drum's sound combined with saxophone's visual information), it is easier for participants to disregard the emotional content of the irrelevant information. This builds on Collignon et al.'s (2008) suggestion that the emotional information is integrated at a perceptual level by attributing less weight to the uncertain sensory information, by indicating that this attribution might rely on factors other than the uncertainty of the information. For example, the relevance of the auditory signal could reflect the influence of different kinds of top-down factors on multisensory processes (Ernst and Bühlhoff, 2004; Vatakis and Spence, 2007). In other words, when left free to use both modalities, participants would

probably rely on the auditory information due to prior knowledge, considering that music has been predominantly accessed via auditory means since the introduction of technologies like the phonograph and radio (Broughton and Stevens, 2009). However, when forced to use the visual information, they would be affected by the emotional content of the auditory signal only when assuming that the two signals belonged to the same event. Thus, the weighting of the auditory signal might be a consequence of bottom-up as much as top-down factors and our brains may rely on both to maximise the efficacy of the process of integration (Collignon et al., 2008; Ernst and Bühlhoff, 2004).

One could argue that the very effect of incongruent information on perceived emotion does not depend at all on perceptual processes, but rather on high-level cognitive processes (Bertelson and De Gelder, 2004). That is, it might be that participants who are aware of the mismatch between sight and sound consciously decide that the auditory information, for example, should influence their emotional judgements. However, we can exclude this explanation for different reasons. First of all, if participants consciously took into account the incongruency when judging the perceived emotion then it is not clear why the auditory information should be weighted more in their decision than the visual. That is, it is difficult to understand why participants would judge and rate as much less sad the incongruent display combining the sad video with the other emotional sounds than the sad sound with the other emotional videos (see Supplementary material). Secondly, in Experiment 1 if participants had made a conscious decision only when the bimodal incongruent displays were presented, while instead making perceptual judgements when presented with the unimodal and bimodal congruent displays, we would expect a longer response time in the incongruent situations. However, participants were found to take approximately the same amount of time to make their judgements on unimodal, bimodal congruent and bimodal incongruent emotional displays. This result suggests that the participants made the same kind of judgement on all of these different displays. Thirdly, in Experiment 2, when asked to pay attention to only one of the two modalities and give a quick judgement, participants had a corrected reaction time (IE) that did not change between stimulus categories, if we exclude the bimodal incongruent condition in the visual attended task. Also this effect of discordant information found in the visual attended task for the emotionally incongruent condition within instrument category can hardly be explained as a product of conscious decision-making. Indeed, if participants consciously decided to take into account the emotional content of the discordant information because they were aware of the mismatch, why they would not do so when the mismatch was made even clearer by the semantic discordance between the signals. For all these reasons we conclude that the found effect of the discordant information on the perceived emotion is not cognitive but perceptual, although this perceptual process is affected by top-down factors.

Finally, in Experiment 2 the irrelevant auditory information was found to influence the visual emotional content only when it was concordant semantically (Vatakis and Spence, 2007, 2008; Schutz and Kubovy, 2008) despite the temporal

mismatch between the signals (McGurk and MacDonald, 1976; Arrighi et al., 2006; Vatakis and Spence, 2006a,b; Petrini et al., 2009a,b) being present in both the semantic discordant and concordant stimuli. That is, if the temporal mismatch (i.e. the lack of temporal covariation between the musicians' movements and the sound) between the signals had any effect on the perceived emotion, we should have found an effect of the irrelevant auditory information on the attended visual information for both types of incongruent stimuli in Experiment 2, although it could have varied quantitatively when also semantically discordant. Instead participants' judgements were affected only for the emotionally incongruent condition within the same instrument category (e.g. when the happy sound from the saxophone was combined with the sad visual signal from the saxophone). Thus, whereas the semantic correspondence appears to reinforce the process of integration between the emotional signals at least when forced to attend to the less reliable visual information, the temporal mismatch between the auditory and the visual signal appears to be irrelevant. The possibility of separately processing the temporal and emotional aspects from the disparate sensory signals would be adaptively important, because depending on the situation, our brain could base its response only on some multisensory aspects while ignoring the others. Therefore keeping these integration processes segregated could save resources and make the process faster and more effective. A recent fMRI study by Kreifelts et al. (2007) on the integration of non-verbal affective signals from face and voice showed that there probably is a specific network (pSTG-Thalamus) subtending the integration of audiovisual emotional aspects, and our results would support the authors' findings.

## 4. Experimental procedures

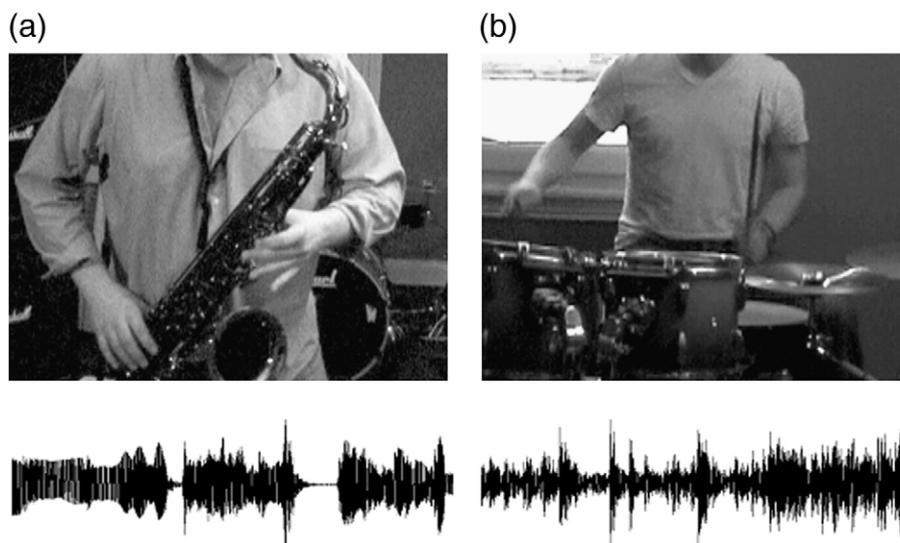
### 4.1. Pre-experiment

#### 4.1.1. Participants

Fifteen participants (8 females and 7 males with an average age of 23.5) of UK nationality (to exclude possible cultural effects) with no musical training participated in a preliminary experiment to select the best emotional displays. The study received IRB approval and all participants gave informed consent to participate. Participants received a monetary incentive for their participation.

#### 4.1.2. Movement recordings and stimuli production

Two hours of brief audiovisual musical recordings were obtained by asking two experienced musicians (a drummer with 13 years experience and a saxophonist with 23 years experience) to improvise with their instruments in order to communicate six different emotions (happiness, sadness, anger, fear, disgust, surprise) and one neutral. Recordings were made in a professional music studio by using two digital video-cameras (a Panasonic HDC-SD5 and a Sony DCR-TRV950E) to record the musicians' performances from both a front and side view. Using Adobe Premiere 1.5 we created 130 black and white displays (PAL — 720×526) of 3 s duration where only the torso and arms were presented (Fig. 4).



**Fig. 4 – (a) Frame and wav sample from the saxophonist surprise display; (b) frame and wav sample from the drummer anger display.**

#### 4.1.3. Apparatus and procedure

The visual stimuli were presented on a Sony Trinitron screen with resolution of 1280×1024 pixels and a refresh rate of 60 Hz, by a Dell laptop running Windows XP. Auditory stimuli were presented through headphones (Beyer Dynamic DT Headphones).

Participants were shown all of the 130 congruent displays in a randomised order by using Presentation 13.1, and asked to choose the perceived emotion from the seven possibilities (happiness, sadness, anger, fear, disgust, surprise and neutral). Immediately following this decision they were asked to rate the strength with which they perceived the chosen emotion by using a rating scale from 1 (“not at all”) to 6 (“very much”).

#### 4.1.4. Results

In order to obtain the appropriate displays for the main experiment, two criteria were set for the results: 1) 73% (corresponding to 11 out of 15 participants) or more of participants should have correctly discriminated the expressed emotion; 2) the displays should have received an averaged rating of  $\geq 4$  for emotional intensity. Six displays were obtained: three for the drummers (happiness: 73%, 4; anger: 80%, 5; neutral: 73%, 4) and three for the saxophonist (happiness: 93%, 5; sadness: 93%, 4; surprise: 73%, 5).

### 4.2. Experiment 1

#### 4.2.1. Participants

Twenty further participants (12 females and 8 males) of UK nationality with no musical experience and an average age of 23 were recruited to participate in this experiment. The study received IRB approval and all participants gave informed consent to participate. Participants received a monetary incentive for their participation.

#### 4.2.2. Apparatus and stimuli

The visual stimuli were presented on a Sony Trinitron screen with a resolution of 1280×1024 pixels and a refresh rate of 60 Hz, by a

Dell laptop running Windows XP. Auditory stimuli were presented through headphones (Beyer Dynamic DT Headphones).

The six displays selected from the preliminary experiment were manipulated using Adobe Premiere 1.5 to create the audio-only, visual-only and audiovisually incongruent conditions for a total of 48 stimuli. The audio-only condition was created by showing a black screen together with the auditory signal, while the visual condition was created by showing only the visual display without any sound. Five incongruent conditions were created for each emotional category by adding the sound from the other conditions to each of the six visual displays. Of these five, two incongruent displays maintained the instrument correspondence but not the emotional content (e.g. saxophone visual information for happiness with saxophone auditory information for sadness), another two mismatched both emotional and instrumental correspondence (e.g. saxophone visual information for happiness with drumming auditory information for anger) and finally a further condition maintained the emotional correspondence but not the instrumental (e.g. saxophone visual information for happiness with drum auditory information for happiness). The lack of temporal correspondence was maintained for all the five incongruent conditions.

#### 4.2.3. Procedure

The stimuli were presented to participants twice in two separate blocks by using Presentation 13.1. Before starting the experiment participants performed a brief set of practice trials (three displays presented in a randomised order) to familiarise themselves with the task. In the first block, the 48 stimuli were presented once in a randomised order and the participants were asked, for each stimulus, to give a rating judgement by rating the strength with which they perceived each one of the seven conditions (happiness, sadness, anger, fear, disgust, surprise and neutral) on a scale from 1 (“not at all”) to 6 (“very much”). The order in which participants were asked to rate the seven conditions was randomised every time after each display.

In the second block the stimuli were presented again once in a randomised order and participants were asked to give a categorical judgement by choosing which of the seven emotions they perceived using a 7 alternative forced choice task. We presented each display only once during each block because we wanted to avoid the possibility of participants having the time to learn that only five emotions out of seven were presented, and also to decrease the chance of participants having the time to acquire cognitive strategies with which to make their emotional judgements. The rating judgements were always asked for before the categorical judgements to avoid possible rating biases toward the chosen category.

### 4.3. Experiment 2

#### 4.3.1. Participants

Sixteen further participants (10 females and 6 males) of UK nationality with no or little musical experience and an average age of 23 were recruited to participate in this experiment. The study received IRB approval and all participants gave informed consent to participate. Participants received a monetary incentive for their participation.

#### 4.3.2. Apparatus and stimuli

The apparatus and stimuli were the same as for Experiment 1.

#### 4.3.3. Procedure

In Experiment 2 the new group of participants was asked to judge the perceived emotion from a list of five categories (happiness, sadness, anger, surprise, and neutral). However, this time participants were requested to focus their attention on the visual information and disregard the auditory, or vice versa, and to be as fast as possible in giving their answer. This change of task was aimed to test whether the irrelevant information would affect the emotion communicated by the attended modality, and if so, whether this effect would change for the different stimuli (e.g. bimodal incongruent for instrument and emotion vs. bimodal incongruent only for emotion). In the visual block of the experiment participants saw stimuli displayed only visually, audiovisually congruently, and audiovisually incongruently. In the auditory block participants were presented with auditory, audiovisually congruent, and audiovisually incongruent stimuli. The order of the auditory and visual blocks was counterbalanced across participants. In each block 42 stimuli were presented, 6 unimodal, 6 bimodal congruent, 12 bimodal incongruent for emotion, and 18 bimodal incongruent for instrument and emotion.

## Acknowledgements

This work was supported by grants from the ESRC (RES-060-25-0010). We would also like to thank the two musicians Simon Pauley and Szymon Ostasz for their valuable contribution.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.brainres.2010.02.012.

## REFERENCES

- Arrighi, R., Alais, D., Burr, D., 2006. Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *J. Vis.* 6, 260–268.
- Behrens, G.A., Green, S.B., 1993. The ability to identify emotional content of solo improvisations performed vocally and on three different instruments. *Psychol. Music* 21, 20–33.
- Bertelson, P., De Gelder, B., 2004. The psychology of crossmodal attention. In: Spence, C., Driver, J. (Eds.), *Crossmodal Space and Crossmodal Attention*. Oxford University Press Inc., pp. 141–177.
- Broughton, M., Stevens, C., 2009. Music, movement and marimba: an investigation of the role of movement and gesture in communicating musical expression to an audience. *Psychol. Music* 37, 137–153.
- Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. *Trends Cogn. Sci.* 11, 535–543.
- Castellano, G., Mortillaro, M., Camurri, A., Volpe, G., Scherer, K., 2008. Automated analysis of body movement in emotionally expressive piano performances. *Music Percept.* 26, 103–119.
- Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., Lepore, F., 2008. Audio-visual integration of emotion expression. *Brain Res.* 1242, 126–135.
- Dahl, S., Friberg, A., 2007. Visual perception of expressiveness in musicians' body movements. *Music Percept.* 24, 433–454.
- Davidson, J.W., 1993. Visual perception of performance manner in the movements of solo musicians. *Psychol. Music* 21, 103–113.
- Davidson, J.W., 1994. Which areas of a pianist's body convey information about expressive intention to an audience? *J. Hum. Mov. Stud.* 26, 279–301.
- De Gelder, B., Vroomen, J., 2000. The perception of emotions by ear and by eye. *Cogn. Emot.* 14, 289–311.
- Ekman, P., Friesen, W.V., 1975. *Unmasking the Face*. Englewood Cliffs, NJ, Prentice-Hall.
- Ernst, M.O., Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433.
- Ernst, M.O., Bühlhoff, H.H., 2004. Merging the senses into a robust percept. *Trends Cogn. Sci.* 8, 162–169.
- Hietanen, J.K., Leppänen, J.M., Illi, M., 2008. Evidence for the integration of audiovisual emotional information at the perceptual level of processing. *Eur. J. Cogn. Psychol.* 16, 769–790.
- Juslin, P.N., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* 129, 770–814.
- Juslin, P., Sloboda, J.A., 2001. *Music and Emotion: Theory and Research*. Oxford Univ. Press, Oxford.
- Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., Wildgruber, D., 2007. Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. *NeuroImage* 37, 1445–1456.
- Landy, M.S., Maloney, L.T., Johnston, E.B., Young, M., 1995. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vis. Res.* 35, 389–412.
- Laukka, P., Gabrielsson, A., 2000. Emotional expression in drumming performance. *Psychol. Music* 28, 181–189.
- Massaro, D.W., Egan, P.B., 1996. Perceiving affect from the voice and the face. *Psychon. Bull. Rev.* 3, 215–221.
- McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. *Nature* 264, 746–748.
- Petrini, K., Dahl, S., Rocchesso, D., Waadeland, C.H., Avanzini, F., Pollick, F., Puce, A., 2009a. Multisensory integration of drumming actions: musical expertise affects perceived audiovisual asynchrony. *Exp. Brain Res.* 198, 339–352.

- Petrini, K., Russell, M., Pollick, F., 2009b. When knowing can replace seeing in audiovisual integration of actions. *Cognition* 110, 432–439.
- Scherer, K.R., 1995. Expression of emotion in voice and music. *J. Voice* 9, 235–248.
- Schutz, M., Kubovy, M., 2008. The effect of tone envelope on sensory integration: support for the 'unity assumption'. *J. Acoust. Soc. Am.* 123 3412.
- Stein, B.E., Meredith, M.A., 1993. *The Merging of the Senses*. MIT, Cambridge (MA).
- Thompson, W.F., Russo, F.A., Quinto, L., 2008. Audio-visual integration of emotional cues in song. *Cogn. Emot.* 22, 1457–1470.
- Vatakis, A., Spence, C., 2006a. Audiovisual synchrony perception for speech and music using a temporal order judgment task. *Neurosci. Lett.* 393, 40–44.
- Vatakis, A., Spence, C., 2006b. Audiovisual synchrony perception for music, speech, and object actions. *Brain Res.* 1111, 134–142.
- Vatakis, A., Spence, C., 2007. Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. *Percept. Psychophys.* 69, 744–756.
- Vatakis, A., Spence, C., 2008. Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychol.* 127, 2–23.
- Vines, B.W., Krumhansl, C.L., Wanderley, M.M., Levitin, D.J., 2006. Cross-modal interactions in the perception of musical performance. *Cognition* 101, 80–113.