Jyväskylän yliopiston julkaisuarkisto
Jyväskylä University Digital Archive

UNIVERSITY OF JYVÄSKYLÄ

This is an electronic reprint of the original article.
This reprint *may differ* from the original in pagination and typographic detail.

Author(s): Vuoskoski, Jonna K.; Thompson, Marc; Spence, Charles; Clarke, Eric F.

Title: Interaction of sight and sound in the perception and experience of musical performance

Year: 2016

Version:

Please cite the original version:

Vuoskoski, J. K., Thompson, M., Spence, C., & Clarke, E. F. (2016). Interaction of sight and sound in the perception and experience of musical performance. Music Perception, 33(4), 457-471. https://doi.org/10.1525/mp.2016.33.4.457

All material supplied via JYX is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

Interaction of Sight and Sound in the Perception and Experience of Musical Performance

Jonna K. Vuoskoski
University of Oxford, Oxford, United Kingdom

Marc R. Thompson
University of Jyväskylä, Jyväskylä, Finland

Charles Spence, & Eric F. Clarke
University of Oxford, Oxford, United Kingdom

Recently, Vuoskoski, Thompson, Clarke, and Spence (2014) demonstrated that visual kinematic performance cues may be more important than auditory performance cues in terms of observers' ratings of expressivity perceived in audiovisual excerpts of piano playing, and that visual kinematic performance cues had crossmodal effects on the perception of auditory expressivity. The present study was designed to extend these findings, and to provide additional information about the roles of sight and sound in the perception and experience of musical performance. Experiment 1 investigated the relative contributions of auditory and visual kinematic performance features to participants' subjective emotional reactions evoked by piano performances, while Experiment 2 was designed to explore the effect of visual kinematic cues on the perception of loudness and tempo variability. Experiment 1 revealed that visual performance cues seem to be just as important as auditory performance cues in terms of the subjective emotional reaction of the observer, thus highlighting the importance of non-auditory cues for music-induced emotions. The results of Experiment 2 revealed that visual kinematic cues only affected ratings of loudness variability, but not ratings of tempo variability.

Received: November 26, 2013, accepted February 20, 2015.

Key words: Multisensory integration, piano performance, expressivity, music-induced emotion, audiovisual perception

Music is an inherently multisensory phenomenon, comprising auditory, visual, and somatosensory components. In musical performance, a performer's body movements and gestures can convey a range of meaningful information to audiences and co-performers alike, including emotional expression (Castellano, Mortillaro, Camurri, Volpe, & Scherer, 2008; Dahl & Friberg, 2007; Davidson, 1993, 1994) and phrasing (Juchniewicz, 2008; Vines, Krumhansl, Wanderley, & Levitin, 2006), as well as musical ideas and timing (Glowinski et al., 2013; Goebl & Palmer, 2009; Williamon & Davidson, 2002). The salience of visual kinematic information (i.e., visual information about performers' body movements and gestures) for an observer's perception and experience of a musical performance has been widely documented (e.g., Chapados & Levitin, 2008; Davidson, 1993; Tsay, 2013; Vines, Krumhansl, Wanderley, Dalca, & Levitin, 2011; Vines et al., 2006), and a recent meta-analysis by Platz and Kopiez (2012) revealed that, compared to audio-only presentations, audiovisual information has a moderate effect on participants' evaluations of a musical performance.

Although it has been established that visual information about the performer's movements consistently enhances the appreciation of a musical performance (Platz & Kopiez, 2012), previous studies have not reliably estimated the relative contributions of visual and auditory performance cues to observers' experience. Although previous investigations have shown that the effect size of the visual component on observers' evaluations could on average be characterized as "medium" (Platz & Kopiez, 2012), it is not known how that relates to the effect size of auditory performance cues – especially across different levels of expressivity. Variations in performance features – often collectively referred to as "expressivity" – are what differentiate performances of the same notated work, and serve to articulate musical structure (Clarke, 1988), communicate emotional meaning (see Juslin, 2001, for a review), and convey a sense of biological motion (Juslin, 2003). In order to

Music Perception, volume 33, issue 4, pp. 457–471, issn 0730-7829, electronic issn 1533-8312. © 2016 by the regents of the university of california all rights reserved. please direct all requests for permission to photocopy or reproduce article content through the university of california press's Reprints and Permissions web page, http://www.ucpress.edu/journals.php?p=reprints. DOI: 10.1525/mp.2016.33.4.457

investigate the relative contributions of auditory and visual performance cues to observers' evaluations, an experimental method is needed in which the expressivity conveyed by the visual and auditory components of a performance can be manipulated independently, so as to result in matched and mismatched audiovisual pairings. Such experimental designs have been successfully used to investigate the interaction of auditory and visual components in the perception of note duration (Schutz & Lipscomb, 2007), loudness (Rosenblum & Fowler, 1991), timbre (Saldaña & Rosenblum, 1993), pitch (Thompson, Graham, & Russo, 2005), and interval affect (Thompson, Russo, & Quinto, 2008), demonstrating that visual information can significantly alter the perception of various auditory features. However, the difficulty with applying such a design to a complex action such as musical performance is that musicians find it very difficult to control expressivity in one modality independent of the other (e.g., Thompson & Luck, 2012), and the temporal properties of a musical performance also vary greatly from one performance to the next.

Previous studies have attempted to tackle this issue by combining a constant auditory stimulus with visual information of actors portraying different expressive intentions (e.g., Juchniewicz, 2008; Morrison, Price, Geiger, & Cornacchio, 2009), or have settled for combining structurally incongruent auditory and visual components, thus resulting in functionally incongruent and temporally asynchronous stimuli (e.g., Krahé, Hahn, & Whitney, 2013; Petrini, McAleer, & Pollick, 2010). The former approach is problematic because of the limited validity of "faked" expressive movements, and the latter because the movements and gestures in musical performance have been found to arise from a representation of the musical structure, and thus convey meaning in association with specific musical passages (e.g., MacRitchie, Buck, & Bailey, 2013).

To address these limitations, a recent study by Vuoskoski et al. (2014) presented a novel method for creating matched and mismatched audiovisual combinations of different expressive intentions. By utilizing motion-capture animations of piano performances and time-warping algorithms, Vuoskoski et al. were able to investigate the relative contributions of auditory and visual kinematic performance cues to the perception of expressivity in a systematic and balanced way. In contrast to previous studies, the mismatched stimuli utilized by Vuoskoski et al. were temporally synchronized and structurally congruent (i.e., the visual kinematic information always represented the same composed structure as the auditory information). Vuoskoski et al. also explored potential crossmodal effects in the perception of auditory and visual expressivity, addressing the question of whether simultaneously presented visual kinematic information might alter the way in which auditory expressivity is perceived, or vice versa. They found that relative to auditory cues, visual kinematic cues actually contributed slightly more to a participant's overall evaluation of expressivity, and that there appeared to be crossmodal interactions at play in the evaluation of both auditory and visual expressivity.

Although Vuoskoski et al.'s (2014) study provides preliminary evidence for the existence of crossmodal effects in the evaluation of expressivity – as well as shedding light on the relative salience of auditory and visual kinematic performance cues – there are some limitations and questions that require further investigation. First, when considering the relative importance of auditory and visual cues from the observer's point of view, the evaluation of perceived expressivity may not capture the most salient or essential aspects of an observer's experience of a musical performance. Instead of the objective appraisal of the expressive components of a musical performance, it is arguably an observer's subjective emotional experience of the performance that ultimately determines their evaluation (cf. Hargreaves & North, 2010). Although there is some evidence to indicate that visual information might enhance emotional reactivity to musical performances (Chapados & Levitin, 2008), it is not yet known how the effect of visual performance cues relates to that of auditory performance cues with regard to the subjective emotional reaction of the observer. Furthermore, the explicit instructions used by Vuoskoski et al. to take both auditory and visual aspects of the performance into account in the evaluations of overall expressivity might have affected which aspects of the material the participants attended to (for details, see Vuoskoski et al., 2014; Experiment 1). In other words, it may be that as a result of the instructions, the participants paid more attention to visual kinematic performance features than they otherwise would.

Second, although the study by Vuoskoski et al. (2014, Experiment 2) demonstrated that visual kinematic cues can have an impact on evaluations of auditory expressivity, the exact nature of these crossmodal effects remains unclear. It is not yet established whether there are crossmodal effects at play in the perception of lower-level auditory features such as, for example, loudness. Furthermore, it is possible that the outcome reflects response bias, the participants' *ratings* of auditory expressivity being affected by the expressive qualities of the simultaneously presented visual kinematic information without their *perception* of the auditory features

actually having been affected (see, e.g., Schutz & Kubovy, 2009).

The aim of the present study was therefore to extend the findings of Vuoskoski et al. (2014), and to provide new information regarding the roles of visual kinematic and auditory cues in the subjective emotional reactions evoked by musical performance, as well as investigating the possible effect of visual kinematic cues on the evaluation of specific auditory performance features. Experiment 1 was designed to investigate the relative contributions of auditory and visual kinematic performance features on participants' subjective emotional reactions, and thus to provide a more ecologically relevant account of the roles of sight and sound in an observer's experience of a musical performance. The difference between the previous Experiment 1 reported by Vuoskoski et al. (2014) and the current Experiment 1 mirrors the well-established distinction between *perceived* and *felt* emotion (see, e.g., Sloboda & Juslin, 2010). The former experiment investigated evaluations of a perceived characteristic of the performances (i.e., perceived expressivity), while the current experiment investigates the subjective emotional reactions experienced by participants (i.e., felt emotion). Previous research has suggested that visual information may have a significant impact on an observer's emotional reactions to a musical performance (Chapados & Levitin, 2008; Krahé et al., 2013; Vines et al., 2006), but the effect size of visual kinematic performance cues relative to that of auditory performance cues has yet to be investigated.

The aim of Experiment 2 was to explore the effect of visual kinematic cues on the evaluations of auditory expressivity in more detail. The two main auditory characteristics contributing to expressivity in piano performance are variations in timing and dynamics (i.e., tempo and loudness variation; e.g., Gabrielsson, 1999; Palmer, 1997), with the amount of variation being positively associated with perceived expressivity (e.g., Bhatara, Tirovolas, Marie Duan, Levy, & Levitin, 2011). Perceived expressivity is also positively associated with how much a performer moves (e.g., Davidson, 1994; Thompson & Luck, 2012). Since the size of a performer's movements reflects the physical energy used to play the notes, the kinematic visual information specifying a performer's movements might be expected to affect the perception of loudness, which is directly related to physical energy. Visual kinematic cues have previously been shown to affect loudness perception in the context of simple clapping sounds, with the size of clapping motions positively associated with perceived loudness (Rosenblum & Fowler, 1991).

By comparison, it is less obvious how temporally aligned visual kinematic information could affect the auditory perception of tempo variability. Previous research has shown the temporal resolution of the auditory modality to be superior to that of the visual modality (e.g., Burr, Banks, & Morrone, 2009; Freides, 1974; Repp & Penel, 2002), resulting in superior auditory rhythm and beat perception (e.g., Grahn, 2012). However, previous research has also shown that visual kinematic information can influence the perceived duration of notes played on a marimba (Schutz & Kubovy, 2009; Schutz & Lipscomb, 2008), and that the sensitivity to rhythmic deviations can be modulated by point-light animations of a bouncing person (Su, 2014). Nevertheless, since the movements of the pianists were temporally synchronized with the music in all of our stimuli, we hypothesize that the visual kinematic information will have an effect on the perception of loudness variability but not on the perception of tempo variability.

Experiment 1

METHOD

Participants. Nineteen participants (7 male, 12 female) aged 18-31 years ($M = 23.1$, $SD = 4.1$) were recruited from the University of Oxford community. Fourteen participants (73.7%) reported having received at least some music training on an instrument (ranging from 1 to 18 years; $M = 10.6$, $SD = 5.0$). The participants received a monetary incentive (£5) for taking part in the study. All of the experimental procedures followed the University of Oxford Policy on the Ethical Conduct of Research Involving Human Participants and received approval from the Research Ethics Committee.

Stimuli. The stimuli were obtained from a recent study by Vuoskoski et al. (2014), where the stimulus generation process is reported in some detail. However, the process is briefly outlined here, as the method of stimulus generation is crucial for the questions addressed in the study. Two pianists – one male and one female – performed Chopin's Prelude in E minor (Op. 28, No. 4) with three different levels of expression: *Deadpan* (reduced level of expressive intensity); *Normal* (normal level of expressive intensity); and *Exaggerated* (maximum level of expressive intensity); while their movements were captured at 120 frames per second using an 8-camera optical motion capture system (Qualisys Pro-Reflex). In addition, the MIDI output of the digital piano keyboard used in the performances was recorded, providing a complete record of the performances. To create the audio stimuli, the MIDI data were imported

TABLE 1. *Mean Tempo, Tempo Variation, Mean Root-mean-square Energy, and Total Amount of Movement in the Six Performance Excerpts*

	Performance type	Mean tempo (bpm)	Tempo variability (%)	Mean RMS (SD)*	Amount of movement (m)
Pianist 1	Deadpan	65.48	5.84	2.05 (0.79)	15.69
	Normal	62.22	15.08	2.45 (1.22)	36.70
	Exaggerated	58.87	17.31	3.23 (1.50)	44.03
Pianist 2	Deadpan	59.29	8.24	1.72 (0.73)	18.27
	Normal	58.72	16.18	2.23 (1.30)	29.61
	Exaggerated	57.39	24.26	2.32 (1.54)	33.36

Tempo variability reflects the standard deviation of the divergence from mean tempo, calculated for each eighth note. Root-mean-square energy reflects the mean loudness (and loudness variability) of the audio excerpts. RMS values were calculated for 500 millisecond segments. Amount of movement indicates the total distance travelled by the motion capture markers. *RMS values and standard deviations have been multiplied by 1000.

into GarageBand '11 (version 6.0.5), running on Mac OS X. The "Grand Piano" software instrument with 50% reverb was used to generate high-quality renditions of the performances. The segment from the beginning of measure 13 to the end of measure 20 was used to create the experimental stimuli, as this section includes the expressive climax of the piece (Sloboda & Lehmann, 2001), and should therefore allow for the greatest amount of variation in terms of expressive intensity. The duration of the resulting six performance excerpts (2 performers x 3 expressive intentions) ranged from 29 to 33 s ($M = 31.3$, $SD = 1.5$). The descriptive details of the performance excerpts (mean tempo, mean loudness, tempo and loudness variability, and the total amount of movement) are displayed in Table 1.

In order to generate audiovisual stimuli that would be incongruent in terms of their expressive intention (e.g., *deadpan* audio + *normal* movement) yet temporally synchronized, the motion capture data from each performer were temporally aligned to each of the three audio tracks of that performer using a time-warping algorithm (Verron, 2005; see also Wanderley, Vines, Middleton, McKay, & Hatch, 2005). This procedure involved the generation of timing profiles for each performance by annotating the timing of each eighth-note chord played by the left hand, producing an average resolution of 2.04 time points per s. The motion capture data was then functionalized using cubic splines. Using the annotated timing profiles for each performance, curve-stretching algorithms (see Verron, 2005, for details) were used to stretch and compress the motion capture data of a given performance so that it matched the timing profile of another performance. More specifically, the splines between each note onset were made to match the time separation of the corresponding note onsets in the other performance. Two time-warped versions of each performance were generated to match the timing profiles of the other two performances by the same performer. Note that only the movement data were time-warped while the audio data remained unaltered. Finally, the resulting splines were sampled to create time-warped motion capture data that could be used to generate point-light animations. This method has previously been used for analysis purposes (see Wanderley et al., 2005), as it enables the comparison of different performances independent of original tempo or timing variations. However, the present study (and the previous one by Vuoskoski et al., 2014) are – to the best of our knowledge – the first to use the method to generate time-warped point-light animations.

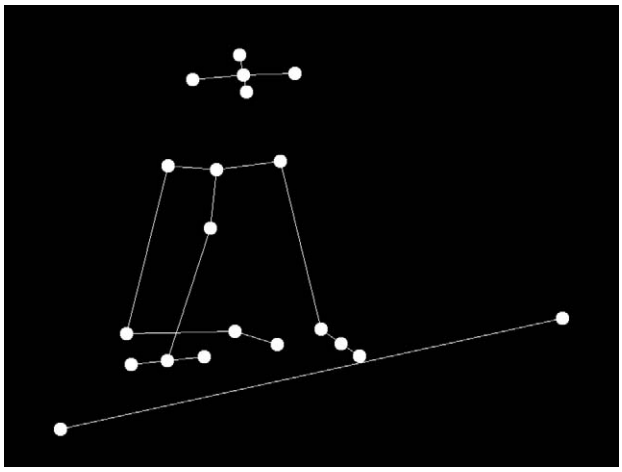

FIGURE 1. A sample frame of the point-light animations used in Experiments 1 and 2.

Point-light animations of the original and time-warped motion capture data were generated using MATLAB and the Motion Capture Toolbox (Burger & Toiviainen, 2013). Light points – connected by lines to form a stick-figure shape – represented each pianist's hands, wrists, elbows, shoulders, head (midpoint and four markers around the head), torso (mid-shoulder and mid-torso), and hips. The keyboard was represented by two markers connected by a line (see Figure 1 for a sample frame).

The 18 animations were combined with the appropriate audio to create 6 matching (e.g., *normal* audio +

normal video) and 12 mismatching (e.g., *exaggerated* audio + *deadpan* video) audiovisual stimuli (example stimuli can be downloaded from https://dl.dropboxusercontent.com/u/311821/Video_examples.zip). Note that the audio and video from different performers were never combined. In addition, unimodal versions of the stimuli (6 audio-only and 18 video-only stimuli) were also generated.

Procedure. The Max/MSP (version 5.1.9; Cycling '74) graphical programming environment (running on Mac OS X) was used to present the stimuli and to collect the data. The animations were presented with a resolution of 800 x 600 pixels and a frame rate of 30 fps. The audio was presented in WAV format through high quality headphones (Sennheiser HD 219). The participants were instructed to evaluate the intensity of their subjective emotional reactions to the performances, and were informed that a given performance might leave them cold, while another performance might move them in a profound way. The evaluations were made using a horizontal analog scale (width 278 pixels) ranging from "did not move me at all" to "moved me very strongly." The participants were instructed to base their ratings on their own emotional reactions rather than any specific aspect of the performances (such as the auditory or visual components of the stimuli), so as not to direct the participants towards perceived rather than felt emotion. The output of the scale, as a default property of the Max/MSP-object, provided data in the range 0-127. The participants were instructed to make their evaluations immediately after each excerpt had ended.

The experiment started with two practice trials using audiovisual excerpts that were similar to – but not part of – the actual stimulus set, to which the participants were instructed to respond. These responses were not included in the data. The practice trials were followed by the 18 audiovisual excerpts, which were presented in a different random order for each participant. The audiovisual block was followed by two unimodal blocks (audio-only, consisting of 6 audio excerpts; and video-only, consisting of 18 video excerpts), in which evaluations of felt emotional impact were based only on what was perceived in the presented modality. The audiovisual block was always presented first, as the audiovisual condition was the main focus of interest in the current study. Furthermore, the initial exposure to the audiovisual excerpts provided participants with a relevant framework in which to view the silent point-light animations, which might have seemed strange or arbitrary if presented first. The video-only condition included both the six original animations as well as the twelve time-warped animations that had been altered to match the different audio tracks. The order in which the two unimodal blocks were presented was counterbalanced across participants. Again, the excerpts within the blocks were presented in a different random order for each participant. After the experiment, the participants completed a short questionnaire about their music training and music listening habits, and were fully debriefed.

RESULTS

Emotional impact in unimodal rating conditions. In order to investigate whether the unimodal (audio-only and video-only) representations of different expressive intentions resulted in differing evaluations of felt emotional impact, repeated-measures ANOVAs were conducted to investigate the ratings obtained in the two unimodal conditions. There were two within-participant factors; Performance Condition (*Deadpan, Normal*, or *Exaggerated*) and Pianist (Pianist 1 or 2), and one between-participants factor; Block Order. The latter factor was added in order to investigate whether the presentation order of the unimodal blocks (audio-only first or video-only first) had any effect on participants' ratings. Note that the audiovisual block always preceded the two unimodal blocks. In the audio-only condition, there was a significant main effect of Performance Condition; $F(2, 34) = 7.07$, $p < .01$, η^2_G (generalized eta squared; Bakeman, 2005) $= .17$, as well as a significant main effect of Pianist; $F(1, 17) = 5.84$, $p < .05$, $\eta^2_G = .04$. There was no main effect of Block Order, and no interaction effects. Multiple comparisons of means (paired t-tests, $p < .05$ significance level adjusted using the Holm-Bonferroni method; Holm, 1979) revealed that ratings of emotional impact for the *Deadpan* performances were significantly lower than those for the *Normal* and *Exaggerated* performances, but that the latter two did not differ significantly from each other. A comparison of means also revealed that the performances of Pianist 2 were rated as having a stronger emotional impact on average than those of Pianist 1. The mean ratings for the three different types of performances by the two pianists are displayed in Figure 2.

A similar repeated-measures ANOVA was conducted to analyze the ratings of felt emotional impact obtained in the video-only rating condition, with the difference that two factors regarding performance condition were included: Type of Video, and Type of Time-warp. As the video component of the mismatched stimuli had been slightly altered to fit the accompanying audio track, Type of Time-warp was included to determine whether there were any differences between the

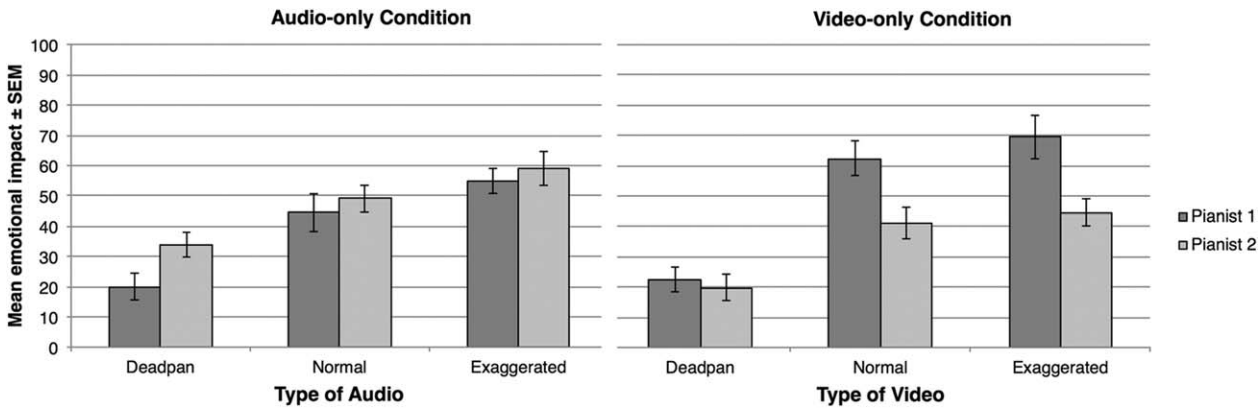

FIGURE 2. The mean ratings of emotional impact (± standard error of the mean) obtained in the unimodal audio-only and video-only conditions of Experiment 1. The ratings have been scaled to a range of 0-100.

different time-warped and original animations. Type of Video and Type of Time-warp both had three levels: *Deadpan*, *Normal*, and *Exaggerated*. Due to a technical failure, one participant's video-only ratings were not recorded, and thus $n = 18$ for this analysis. The analysis revealed significant main effects of Type of Video; $F(2, 32) = 29.17$, $p < .001$, $\eta_G^2 = .26$, Type of Time-warp; $F(2, 32) = 4.32$, $p < .05$, $\eta_G^2 = .01$, and Pianist; $F(1, 16) = 18.04$, $p < .001$, $\eta_G^2 = .07$. There were no main or interaction effects related to Block Order. Multiple comparisons of means revealed that the emotional impact of the *Deadpan* video type was rated as significantly weaker than the impact of the *Normal* or *Exaggerated* video types, as expected; but that the difference between the latter two – although in the expected direction – was not statistically significant. Multiple comparisons regarding the effect of Type of Time-warp did not reveal any significant differences between the different time-warped and original animations after the Holm-Bonferroni correction had been applied. A comparison of means confirmed that the emotional impact of the performances by Pianist 1 were evaluated as significantly stronger than for those by Pianist 2. There was also a significant interaction between Type of Video and Pianist; $F(2, 32) = 10.02$, $p < .001$, $\eta_G^2 = .04$. Multiple comparisons of means revealed that the emotional impact of the performances by Pianist 1 was rated as significantly stronger (than those of Pianist 2) only in the *Normal* and *Exaggerated* video conditions. The mean ratings given for the three different types of performance by the two pianists are shown in Figure 2.

Ratings of emotional impact in the audiovisual condition. In order to investigate the salience of the auditory and visual modalities with regard to the emotional impact induced by the audiovisual performance excerpts, a repeated-measures ANOVA was conducted. There were three within-participant factors in the ANOVA: Type of Audio, Type of Video, and Pianist. The analysis yielded significant main effects of Type of Audio: $F(2, 36) = 11.22$, $p < .001$, $\eta_G^2 = .10$; and Type of Video: $F(2, 36) = 9.12$, $p < .001$, $\eta_G^2 = .09$. The mean ratings (grouped by Type of Audio and Type of Video) are displayed in Figure 3. Multiple comparisons of means (paired t-tests, $p < .05$ significance level adjusted using the Holm-Bonferroni method) revealed that all three types of audio were significantly different from each other, with the *Deadpan* condition receiving the lowest and the *Exaggerated* condition receiving the highest ratings of felt emotional impact. Regarding the different

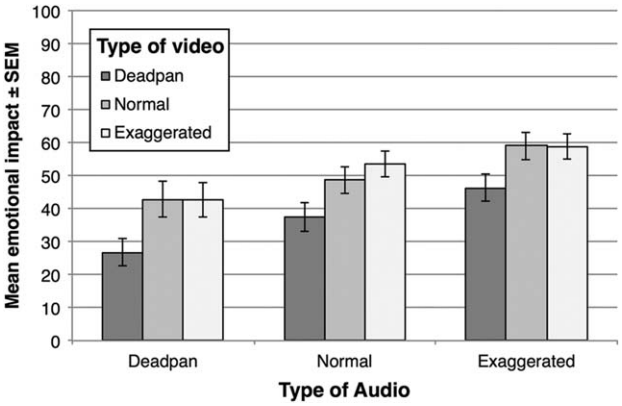

FIGURE 3. The mean ratings of felt emotional impact (± standard error of the mean) obtained in the audiovisual condition of Experiment 1, grouped by Type of Audio and Type of Video. The ratings have been scaled to a range of 0-100.

video types, multiple comparisons revealed that the *Deadpan* videos received significantly lower ratings of emotional impact than the *Normal* and *Exaggerated* videos, but that the latter two did not differ significantly from one other. There was no main effect of Pianist, and no interaction effects.

To further investigate the relative contribution of auditory and visual cues to the emotional impact evoked by the performance excerpts, a linear regression analysis was conducted. The dependent variable was the mean ratings of felt emotional impact for audiovisual stimuli, while the mean ratings of emotional impact for audio-only and video-only stimuli were the independent variables. The two predictor variables were not significantly intercorrelated, $r(16) = -.06$, *ns*, but both were significantly correlated with the dependent variable: $r(16) = .67$, $p < .01$, for audio-only ratings, and $r(16) = .62$, $p < .01$, for video-only ratings. Audio-only and video-only ratings of emotional impact both significantly predicted felt emotional impact in the audiovisual condition, $\beta = .71$, $t(17) = 7.94$, $p < .001$, and $\beta = .66$, $t(17) = 7.42$, $p < .001$, respectively. Together they explained a significant proportion of the variance in the emotional impact felt in the audiovisual condition; $R^2 = .88$, $F(2, 17) = 55.93$, $p < .001$.

DISCUSSION

The results of Experiment 1 demonstrate that each audio type – *Deadpan, Normal,* and *Exaggerated* – was rated as eliciting a different level of emotional impact in the audio-only condition. The effect size of audio type ($\eta_G^2 = .17$) was notably smaller than that in a previous experiment measuring perceived expressivity (using the same stimuli; $\eta_G^2 = .59$; Vuoskoski et al., 2014). This difference in effect size may be attributable to the more subjective and internal character of participants' own emotional reactions as compared to the more manifest and external character of the expressive intentions on which participants were asked to focus in the previous study. Indeed, previous research on music-induced emotions has found that there tends to be more inter-individual variability in evaluations of felt emotion compared to evaluations of perceived emotion (e.g., Juslin, 2009).

Interestingly, the effect of video type in the video-only rating condition ($\eta_G^2 = .26$) was somewhat larger than the effect of audio type in the audio-only condition, though there was no statistically significant difference between the *Normal* and *Exaggerated* video types in terms of their emotional impact. Although this effect size is smaller than that observed in a previous experiment investigating the perception of expressivity ($\eta_G^2 = .61$; Vuoskoski et al., 2014), it is nevertheless striking that point-light animations of pianists performing were nonetheless able to evoke significantly differentiated emotional responses in participants. However, it may also be that participants' evaluations were affected by demand characteristics (e.g., Orne, 1962). When asked to evaluate the emotional impact of stimuli that clearly represent an emotional expression of some kind, it might be that even in the absence of genuine emotional reactions the participants nonetheless move the slider on the basis of perceived expressivity rather than felt emotion (cf. Konečni, 2008). This possibility is supported by the fact that three of the participants reported extremely low (or nonexistent) levels of emotional impact in response to the video-only stimuli (but not in response to the audiovisual or audio-only stimuli), perhaps reflecting a more rigorous rating strategy on their part than for the other participants. Furthermore, having already responded to an audiovisual block (which was always presented first) it is possible that the participants' unimodal ratings were influenced by previous audio-visual associations. Since the participants were exposed to both matched and mismatched combinations in the audiovisual block, it is unlikely that they would have associated a specific audio-only stimulus with a specific video component or vice versa; but it may be that a more generic association between the two modalities may nonetheless have been induced.

The results of the audiovisual rating condition revealed that both Type of Audio and Type of Video had a significant effect ($\eta_G^2 = .10$ and .09, respectively) on the emotional impact of the piano performances. The effect sizes of audio type and video type were comparable, in contrast to the differences observed in the unimodal rating conditions. This pattern of results is somewhat different from that found for the perception of expressivity (Vuoskoski et al., 2014), where Type of Video ($\eta_G^2 = .29$) revealed a stronger effect compared to Type of Audio ($\eta_G^2 = .23$). Again, the overall difference in effect size may be related to the subjective and elusive character of emotional reactions as compared with perceived expressive intentions, but the difference in the relative contribution of auditory and visual modalities suggests that while visual kinematic cues may be more salient than auditory cues in communicating expressive intentions, their contribution to the emotional impact of performances is comparable to that of auditory performance cues. The results of the linear regression analysis support this conclusion, by showing that audio-only ratings and video-only ratings explain comparable proportions of the variance in the audiovisual ratings of emotional impact. As in the case of the unimodal rating

blocks, it is possible that some of the participants based their ratings of emotional impact on perceived expressivity rather than their actual emotional reactions. Note, though, that this is an issue that affects all studies aiming to investigate music-induced emotions using self-report measures, and can be minimized by giving clear instructions to participants (see e.g., Konečni, 2008). We gave our participants explicit instructions to focus on the "emotional effect that the performance has on you," and used unambiguous labels ("did not move me at all" and "moved me very strongly") to signify the extremes of the rating scale.

Finally, the contribution of either modality to the emotional impact of a performance may depend on the performer and her or his efficacy in conveying expressive intentions via body movements and auditory cues. In the present study, the audio-only excerpts of Pianist 2 were evaluated as having a stronger emotional impact than those of Pianist 1, while the video-only ratings revealed the opposite pattern. These results are in line with the objective measures of auditory and kinematic features (see Table 1), with Pianist 2 displaying more tempo variability, and Pianist 1 displaying more movement overall. However, there was no effect of Pianist in the ratings obtained in the audiovisual condition (and no interaction effects), thus suggesting that the relative contribution of the auditory and visual modalities to the emotional impact of audiovisual performances may not be significantly affected by differences in expressive efficacy. Furthermore, it should be noted that the facial expressions of performers – which would sometimes be visible to the audience in real-life performance situations and which are eliminated in this study by the use of stick figures – may add significantly to the overall emotional impact of a musical performance.

Experiment 2

The results of Experiment 1 revealed that auditory and visual kinematic performance cues seem to account for comparable proportions of participants' subjective emotional reactions to piano performance excerpts. However, the potential crossmodal effects involved in the process remain unclear. A previous study by Vuoskoski et al. (2014) revealed that visual kinematic cues can affect the ratings of perceived auditory expressivity, but it is not yet known whether this effect reflects actual crossmodal interactions between the auditory and visual modalities, or whether instead it could be attributed, for example, to some kind of response bias. Furthermore, if the observed effects were due to crossmodal interactions, it is unclear which aspects of perceived auditory expressivity are affected by visual kinematic cues. Thus, the aim of Experiment 2 was to investigate whether visual kinematic cues might affect the perception of the key auditory features contributing to perceived expressivity, namely loudness and tempo variability. Since the aim was to obtain as reliable and consistent an evaluation of loudness and tempo variation as possible, only those participants with musical instrument training were recruited to take part in this experiment.

METHOD

Participants. Seventeen participants (7 male, 10 female) aged 18-61 years ($M = 26.3$, $SD = 11.7$) were recruited from the University of Oxford community. All of the participants had received a minimum of two years of music training on an instrument (ranging from 2 to 17 years; $M = 10.2$, $SD = 4.9$). The participants received a monetary incentive (£5) for taking part in the study.

Stimuli. The stimuli were the same as those in Experiment 1.

Procedure. The procedure was almost identical to that of Experiment 1, with the difference that instead of emotional impact, the participants were asked to evaluate the amount of loudness (dynamic) variation, and the amount of tempo variation, in the performances. They were instructed that "A performance with no variation in dynamics or timing would sound flat and mechanical, while a performance with an extreme amount of variation would have continuous changes in tempo and loudness." Both evaluations were made using horizontal visual analog scales (width 278 pixels) ranging from "No variation at all" to "An extreme amount of variation." The order in which the scales were presented was balanced across participants. The same rating scales were also used in two unimodal rating conditions. In the video-only condition, the participants were instructed to "try to imagine how the music produced by the pianists' movements would sound, and evaluate the amount of variation in the timing and dynamics of the imagined performances." The audiovisual block was always presented first, followed by the audio-only and video-only blocks. As the presentation order of the unimodal blocks had no significant effect on participants' ratings in Experiment 1, all participants in this experiment completed the unimodal blocks in the same order.

RESULTS

Unimodal perception of loudness and tempo variability. In order to determine whether the perceived amount of loudness and tempo variation differed significantly between the different performance conditions, the ratings

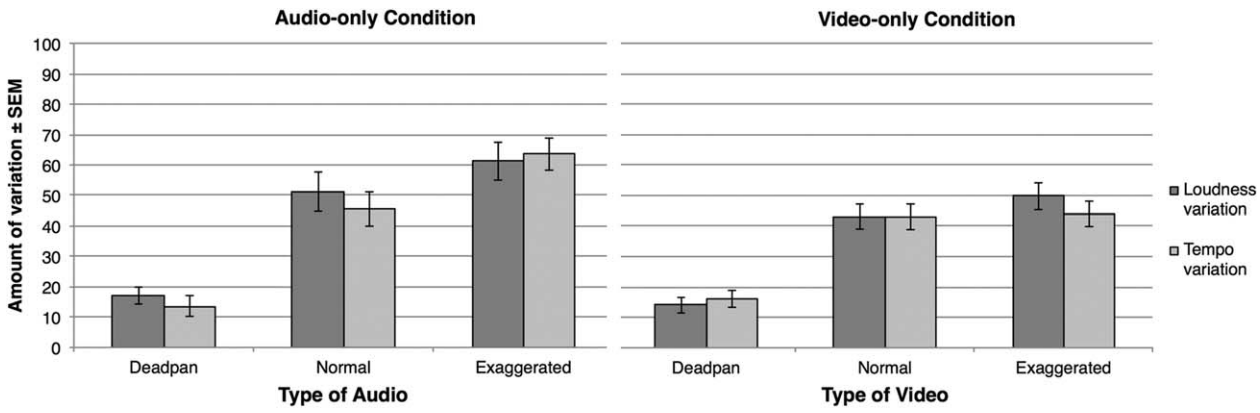

FIGURE 4. The mean ratings of loudness and tempo variability (± standard error of the mean) obtained in the unimodal audio-only and video-only conditions of Experiment 2. The ratings have been scaled to a range of 0-100.

obtained in the unimodal audio-only rating condition were analysed using repeated-measures ANOVAs. The mean ratings are displayed in Figure 4. There were two within-participant factors: Type of Audio (*Deadpan, Normal,* or *Exaggerated*), and Pianist (1 or 2). One participant's audio-only ratings were not recorded due to a technical failure, and thus $n = 16$ for this analysis. In the ratings of the perceived amount of loudness variation, there was a significant main effect of Type of Audio; $F(2, 30) = 41.01$, $p < .001$, $\eta_G^2 = .40$, but no effect of Pianist nor any interaction. Multiple comparisons of means (paired *t*-tests, level of statistical significance adjusted using the Holm-Bonferroni method) revealed that all three audio types differed significantly from each other in terms of the perceived amount of loudness variation, with the *Deadpan* audio type receiving the lowest and the *Exaggerated* audio type the highest ratings. A similar analysis was conducted on the ratings of the amount of tempo variation. This analysis yielded a significant main effect of Type of Audio; $F(2, 30) = 46.61$, $p < .001$, $\eta_G^2 = .48$, but once again no effect of Pianist and no interaction effect was observed. Multiple comparisons of means revealed that all three audio types differed significantly from each other in terms of the perceived amount of tempo variation, with the *Deadpan* audio type receiving the lowest and the *Exaggerated* audio type the highest ratings.

The next step was to investigate the ratings of loudness and tempo variation obtained in the video-only condition, where the participants were instructed to base their ratings on how they imagined the music produced by the observed movements would sound. Repeated-measures ANOVAs with three within-participants factors – Type of Video, Type of Time-warp, and Pianist – were conducted to investigate whether the participants were able to consistently estimate the amount of loudness and tempo variation based on the pianists' movements alone. Type of Time-warp was included as a factor in order to see whether there were any differences between the time-warped and the original animations, since time-warping changes the timing of the movements. In the ratings of loudness variation, there were significant main effects of Type of Video; $F(2, 32) = 49.35$, $p < .001$, $\eta_G^2 = .38$, and Pianist; $F(1, 16) = 26.29$, $p < .001$, $\eta_G^2 = .09$, but no main effect of Type of Time-warp. Multiple comparisons of means revealed that all three video types differed significantly from one another in terms of loudness variability, with the *Deadpan* video type receiving the lowest and the *Exaggerated* video type the highest ratings. Furthermore, a comparison of means revealed that Pianist 1 was rated as exhibiting more loudness variation. There were also interaction effects between Type of Video and Pianist; $F(2, 32) = 20.77$, $p < .001$, $\eta_G^2 = .07$, and between Type of Time-warp and Pianist; $F(2, 32) = 7.30$, $p < .01$, $\eta_G^2 = .02$. Multiple comparisons of means revealed that Pianist 1 was rated as exhibiting more loudness variation than Pianist 2 only in the *Normal* and *Exaggerated* video types. Multiple comparisons investigating the interaction effect between Type of Time-warp and Pianist failed to reach statistical significance after the Holm-Bonferroni correction had been applied.

A similar analysis was conducted on the ratings of tempo variation obtained in the video-only condition, yielding significant main effects of Type of Video; $F(2, 32) = 38.73$, $p < .001$, $\eta_G^2 = .26$, Type of Time-warp; $F(2, 32) = 4.94$, $p < .05$, $\eta_G^2 = .02$, and Pianist; $F(1, 16) = 6.18$, $p < .05$, $\eta_G^2 = .02$. Multiple comparisons of means revealed that the *Deadpan* video type was rated as significantly lower in tempo variation than the *Normal* and *Exaggerated* video types, but that there was no

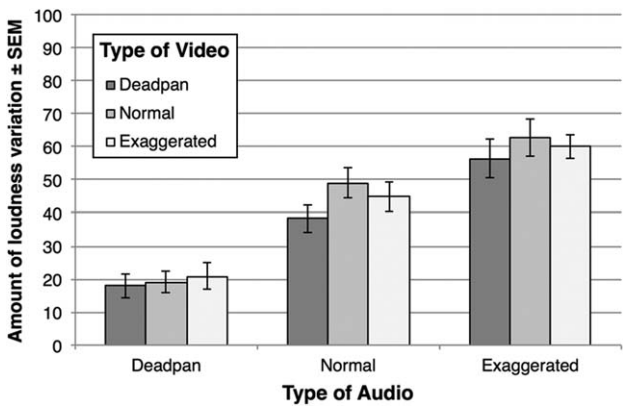

FIGURE 5. The mean ratings of loudness variability (± standard error of the mean) obtained in the audiovisual rating condition of Experiment 2, grouped by Type of Audio and Type of Video. The ratings have been scaled to a range of 0-100.

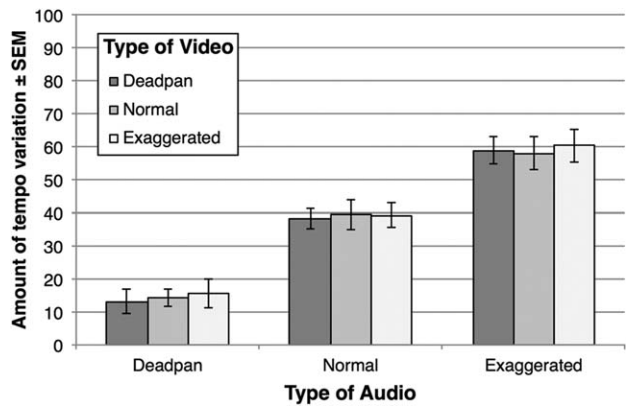

FIGURE 6. The mean ratings of tempo variability (± standard error of the mean) obtained in the audiovisual rating condition in Experiment 2, grouped by Type of Audio and Type of Video. The ratings have been scaled to a range of 0-100.

statistically significant difference between the latter two. Multiple comparisons for the main effect of Type of Time-warp failed to reach statistical significance after the Holm-Bonferroni correction had been applied. A comparison of means also revealed that Pianist 1 was rated as exhibiting more tempo variation than Pianist 2, with interaction effects between Type of Video and Pianist; $F(2, 32) = 3.36$, $p < .05$, $\eta_G^2 = .02$, and between Type of Time-warp and Pianist; $F(2, 32) = 5.78$, $p < .01$, $\eta_G^2 = .01$. Multiple comparisons revealed that Pianist 1 was rated as exhibiting more tempo variation than Pianist 2 only in the case of the *Exaggerated* video type. Furthermore, multiple comparisons revealed that Type of Time-warp only had a significant effect on the ratings of tempo variation in the case of Pianist 2, with the videos warped to *Exaggerated* audio receiving higher ratings than those warped to *Normal* or *Deadpan* audio.

Bimodal perception of loudness and tempo variability. In order to investigate the potential effect of visual cues on the perception of loudness variation, the ratings of loudness variation – obtained in the audiovisual rating condition – were analysed using a repeated-measures ANOVA. The mean values are displayed in Figure 5. There were three within-participants factors: Type of Audio, Type of Video, and Pianist. The analysis yielded significant main effects of Type of Audio; $F(2, 32) = 72.69$, $p < .001$, $\eta_G^2 = .38$, Type of Video; $F(2, 32) = 3.71$, $p < .05$, $\eta_G^2 = .01$, and Pianist; $F(1, 16) = 6.70$, $p < .05$, $\eta_G^2 = .03$. There were no interaction effects. Multiple comparisons of means revealed that all three audio types were rated as significantly different in terms of the amount of loudness variation, with the *Deadpan* audio type receiving the lowest and the *Exaggerated* audio type the highest ratings. For the effect of Type of Video, multiple comparisons of means revealed that there was a statistically significant difference only between the *Deadpan* and *Normal* video types, with the *Deadpan* video type receiving significantly lower ratings of loudness variation. A comparison of the means also revealed that Pianist 2 was rated as exhibiting more loudness variation than Pianist 1.

Finally, the potential effect of visual cues on the perception of tempo variation was investigated by conducting a similar repeated-measures ANOVA on the ratings of tempo variation (see Figure 6 for mean ratings). Once again, there were three within-participants factors: Type of Audio, Type of Video, and Pianist. The analysis yielded significant main effects of Type of Audio; $F(2, 32) = 61.47$, $p < .001$, $\eta_G^2 = .45$, and Pianist; $F(1, 16) = 7.70$, $p < .05$, $\eta_G^2 = .03$, but no effect of Type of Video, nor any interaction effects. Multiple comparisons of means revealed that all three audio types were rated as significantly different in terms of the amount of tempo variation, with the *Deadpan* audio type receiving the lowest and the *Exaggerated* audio type the highest ratings. A comparison of means also revealed that Pianist 2 was rated as exhibiting more tempo variation than Pianist 1.

DISCUSSION

The ratings of loudness and tempo variability obtained in the audio-only condition demonstrated – in line with the objective measures of loudness and tempo variability (see Table 1) – that the performances produced under all three expressive conditions were evaluated as significantly different in terms of the perceived loudness and

timing variation. Furthermore, there were no significant differences between the two pianists in terms of perceived loudness and tempo variability. In the silent video-only rating condition, where the participants were instructed to imagine how the music produced by the pianists' movements would sound, the participants rated all three video types as significantly different in terms of their loudness variability. Since the total amount of movement increased significantly from *Deadpan* to *Exaggerated* performances (see Table 1), this suggests that participants used the size of movements as the cue in their evaluations. This conclusion is further supported by the finding that Pianist 1 – who showed more movement variation across the different performance types (see Table 1, right hand column) – was evaluated as exhibiting more loudness variation than Pianist 2 in the video-only condition. In the video-only ratings of tempo variability, the notably larger effect size for Type of Video relative to Type of Time-warp (which represented the timing model to which the animation was time-warped and matched) suggests that participants used the simple *amount* of movement – rather than the *pattern of timing* of those movements – as a cue. This finding may be explained by the limited temporal resolution of the visual modality (e.g., Freides, 1974; Welch, DuttonHurt, & Warren, 1986), as well as the strong real-world association between the size of performers' movements and the amount of tempo and loudness variation.

Although Pianist 1 was evaluated as exhibiting more loudness and tempo variation than Pianist 2 in the video-only condition, this pattern of results was reversed in the audiovisual rating condition. The audiovisual ratings revealed that Pianist 2 was evaluated as exhibiting more loudness and tempo variation than Pianist 1 – a result that is in line with the objective measures of audio features (see Table 1). Interestingly, however, there was no effect of Pianist in the audio-only condition. As in the audio-only condition, all three audio types were evaluated as significantly different in terms of their loudness and tempo variability in the audiovisual rating condition. The effect of Type of Audio on the evaluations of loudness variability was comparable to that observed in the audio-only condition, but Type of Video also had a statistically significant effect. More specifically, when the different audio types were presented in combination with the *Deadpan* video type, they received lower ratings of loudness variability than when presented together with the *Normal* video type; while for the ratings of tempo variability, the effect of Type of Audio was comparable to the audio-only ratings, and showed no effect of Type of Video.

These results are consistent with the hypothesis that visual kinematic information exerts a crossmodal influence on the perception of loudness variability, but not on the perception of tempo variability. However, the pattern of crossmodal effects observed in the two experiments reported here was not entirely straightforward. The theory of optimal sensory integration (e.g., Alais & Burr, 2004; Ernst & Banks, 2002), which proposes that more weight is given to the modality that provides the more reliable sensory information, does not fully explain why the loudness variability of the *Normal* video type was evaluated as significantly higher than that of the *Deadpan* video type while the *Exaggerated* video type was not. An alternative account is offered by those studies that have demonstrated that when sounds and sights are perceived as originating from a common event (i.e., the *unity assumption*), the process of sensory integration is altered in a way that differs from the traditional understanding of optimal integration (e.g., Schutz & Kubovy, 2009). However, studies that have investigated the unity assumption using musical instrument stimuli have reported conflicting findings, either succeeding (Schutz & Kubovy, 2009) or failing (Vatakis & Spence, 2008) to find an effect of the unity assumption. Mitterer and Jesse (2010) propose that multisensory integration may actually be driven by learned co-occurrences of visual and auditory stimuli rather than their perceived common causation: using piano stimuli showing either a key stroke or the actual sound-producing hammer stroke, they demonstrated that multisensory integration was stronger in the case of key strokes. As there is a strong real-world correlation between auditory and visual cues of musical expressivity – with performers finding it difficult to retain their normal level of expression while restricting their movements (Thompson & Luck, 2012) – this account may also reflect the process underlying the effects observed in the present study.

In line with this proposal, it may be that the degree of crossmodal effect observed in the perception of loudness variability varied depending on the ecological plausibility of the audio-video combinations, suggesting that only those cues that could be meaningfully paired with cues in the other modality resulted in crossmodal effects (cf. Warren, Welch, & McCarthy, 1981). This interpretation is in line with the findings of Vuoskoski et al. (2014), who observed that the more contrasting audio-visual combinations resulted in weaker crossmodal effects.

Finally, there is a need to consider the potential effect of response bias on the observed effects. It may be that only participants' evaluations of loudness variability

were affected by visual cues, while their perceptions of loudness variability remained unaltered. We did not explicitly instruct the participants to base their evaluations *only* on the auditory modality, as we expected musically trained participants to have an established understanding of loudness and tempo variability as musical features; and asking participants to base their ratings on one modality while still attending to the other, risks drawing participants' attention to the phenomenon under investigation, thus increasing the likelihood of demand characteristics. The fact that the ratings of tempo variability were not affected by the simultaneously presented visual kinematic information, and that visual information affected ratings of loudness variability only in the case of certain audio-visual pairings (across both pianists), suggests that the observed crossmodal effects cannot be explained solely in terms of response bias. However, further investigation is undoubtedly required to clarify whether visual information about a piano performance could affect the perception of loudness at a sensory level.

General Discussion

This study provides further evidence for the significance of visual kinematic cues in the perception and experience of musical performance. Although previous studies have shown that visual information can influence the emotions induced by a musical performance (e.g., Chapados & Levitin, 2008; Krahé et al., 2013; Timmers, Marolt, Camurri, & Volpe, 2006, Vines et al., 2011), they haven't been able to reliably estimate the effects size of visual performance cues relative to that of auditory performance cues. The present study revealed that – in terms of the emotional impact of musical performances – the contribution of visual kinematic performance cues appears to be comparable to that of auditory performance cues. This is not to say that the effect of visual cues would be equal to that of musical cues as a whole, since there is the significant impact of the music's composed structure to consider in addition to auditory performance features. The emotions conveyed and induced by music emerge from the combination of structural and performance features, and are also affected by individual and situational factors (e.g., Scherer & Zentner, 2001). In relation to this complex range of factors, the present study was only designed to investigate the relative contributions of auditory and visual kinematic performance cues by comparing different performances (and combinations of different performances) of the same musical piece. Thus, the results of this study suggest that in terms of the effect of performance cues on observers' subjective emotional reactions to a musical performance, the visual modality appears to be just as important as the auditory modality.

The significant contribution of visual cues to our participants' emotional experiences is striking, since the effects of performance features on the perception and induction of *emotion* have often been considered only from an auditory perspective (see e.g., Juslin & Timmers, 2010) – despite more widespread recognition of the role of visual factors in judgements of performance *expressivity* (e.g., Davidson, 1993,1994; Tsay, 2013). There is some evidence to suggest that the type of emotional expression communicated via visual kinematic cues can have an effect on the type (and intensity) of emotions perceived and experienced by the observer of a musical performance (Chapados & Levitin, 2008; Krahé et al., 2013; Timmers et al., 2006; Vines et al., 2011), but more controlled and systematic investigations (e.g., within-participants rather than between-participants designs, and more systematically generated stimuli) are needed to explore this issue further. Moreover, recent findings suggest that the emotions felt by a performer also alter the way in which he or she moves, since observers seem to perceive visually and audiovisually presented (but not solely auditorily presented) violin performances as sadder when the performer was actually *feeling* sad, compared to when they were only expressing sadness (Van Zijl & Luck, 2013). These findings – as well as those of the present study – support the view that observers of a musical performance are able to detect very subtle yet informative cues from visual kinematic information – without necessarily attending to them in a conscious manner (Tsay, 2013).

The results of the present study also provide evidence to support the view that visual kinematic information can have an effect on the judgment of certain auditory performance cues. The results of Experiment 2 revealed that visual kinematic information had an impact on ratings of loudness variability – but not on ratings of tempo variability – suggesting that the crossmodal effects in the perception of auditory expressivity observed in a previous study (Vuoskoski et al., 2014) may be attributed to the effect of visual cues on perceived loudness (rather than tempo) variability. In order to tease out the relative contributions of timing and loudness variability – as well as the effects of visual kinematic cues – on perceived auditory expressivity in more detail, future studies could apply time-warping algorithms to MIDI data as well.

In the case of both experiments, there seemed to be a clearer difference between the *Deadpan* and *Normal* performance types than between the *Normal* and *Exaggerated* performance types. This is in line with the

findings of Vuoskoski et al. (2014), as well as those of Davidson (1993) suggesting that performers may find it easier to "withhold expression from the piece than exaggerate the expressivity of a piece beyond its normal level" (Davidson, 1993, p. 109). It should also be noted that performers can differ greatly in terms of how much they move while performing, as well as how much loudness and timing variation they use when communicating their expressive intentions. Indeed, this was the case in the present study, where Pianist 1 displayed more movement variability, whereas Pianist 2 exhibited more tempo variability. Although the effects observed in this study were consistent across pianists (as evidenced by the lack of interaction effects related to Pianist), it may be that the relative contributions of auditory and visual kinematic performance cues may vary across different pianists, especially in the case of more extreme performance styles. Indeed, differences in expressive efficacy – between different performers and between different instruments – may explain the contrasting findings observed in the present study and a previous study by Vines et al. (2011), where different expressive intentions led to differing emotional reactions only in the audio-visual and video-only conditions, but not in the audio-only condition. However, it might also be argued that the pianists included in this study – music students rather than professional concert pianists – utilize more conventional (i.e., less idiosyncratic) expressive devices in their performances, and thus represent the majority of musicians better than do professional concert pianists.

In conclusion, the results of the two experiments reported here demonstrate that visual information about a performer's movements not only has an impact on the intensity of emotional reactions evoked by the performance, but can also change how that performance sounds to an observer. The study has shown that visual performance cues may be just as important as auditory performance cues in terms of the subjective emotional experience of the observer, suggesting that non-auditory cues may contribute more to music-induced emotions than has previously been established. These results confirm the significant role of visual kinematic cues for audience members, and encourage further investigations into the ways in which visual information may interact with auditory information in our perception and experience of a musical performance.

Author Note

We are grateful to three anonymous reviewers and the Action Editor for their helpful comments on an earlier version of the paper. This research was supported by the Andrew W. Mellon Foundation.

Correspondence concerning this article should be addressed to Jonna K. Vuoskoski, Faculty of Music, University of Oxford, St Aldate's, OX1 1DB, Oxford, United Kingdom. E-mail: jonna.vuoskoski@music.ox.ac.uk

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257-262.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*, 379-384.

Bhatara, A., Tirovolas, A. K., Marie Duan, L., Levy, B., & Levitin, D. J. (2011). Perception of emotional expression in musical performance. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 921-934.

Burger, B., & Toiviainen, P. (2013). MoCap Toolbox – A Matlab toolbox for computational analysis of movement data. In R. Bresin (Ed.), *Proceedings of the 10th Sound and Music Computing Conference*. Stockholm, Sweden: KTH Royal Institute of Technology.

Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, *198*, 49-57.

Castellano, G., Mortillaro, M., Camurri, A., Volpe, G., & Scherer, K. (2008). Automated analysis of body movement in emotionally expressive piano performances. *Music Perception, 26*, 103-119.

Chapados, C., & Levitin, D. J. (2008). Cross-modal interactions in the experience of musical performances: Physiological correlates. *Cognition*, *108*, 639-651.

Clarke, E. F. (1988). Generative principles in music performance. In J. A. Sloboda (Ed.), *Generative processes in music: The psychology of performance, improvisation, and composition* (pp. 1-26). Oxford: Oxford University Press.

Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception*, *24*, 433-454.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*, 103-113.

Davidson, J. W. (1994). What type of information is conveyed in the body movements of solo musician performers? *Journal of Human Movement Science, 6*, 279-301.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433.

Freides, D. (1974). Human information processing and sensory modality: Cross-modal functions, information complexity, memory, and deficit. *Psychological Bulletin*, *81*, 284-310.

Gabrielsson, A. (1999). The performance of music. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 501-602). San Diego, CA: Academic Press.

Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C., & Doherty, C. (2013). The movements made by performers in a skilled quartet: A distinctive pattern, and the function that it serves. *Frontiers in Psychology*, *4,* 841.

Goebl, W., & Palmer, C. (2009). Synchronization of timing and motion among performing musicians. *Music Perception*, *26*, 427-438.

Grahn, J. A. (2012). See what I hear? Beat perception in auditory and visual rhythms. *Experimental Brain Research*, *220*, 51-61.

Hargreaves, D. J., & North, A. C. (2010). Experimental aesthetics and liking for music. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 515-546). Oxford: Oxford University Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65-70.

Juchniewicz, J. (2008). The influence of physical movement on the perception of musical performance. *Psychology of Music*, *36*, 417-427.

Juslin, P. N. (2001). Communicating emotion in music performance: A review and a theoretical framework. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 309-337). Oxford: Oxford University Press.

Juslin, P. N. (2003). Five facets of musical expression: A psychologist's perspective on music performance. *Psychology of Music*, *31*, 273-302.

Juslin, P. N. (2009). Emotional responses to music. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 131-140). Oxford: Oxford University Press.

Juslin, P. N., & Timmers, R. (2010). Expression and communication of emotion in music performance. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 453-489). Oxford: Oxford University Press.

Koneni, V. J. (2008). Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 115-129.

Krahé, C., Hahn, U., & Whitney, K. (2013). Is seeing (musical) believing? The eye versus the ear in emotional responses to music. *Psychology of Music* [online before print]. DOI: 0305735613498920.

MacRitchie, J., Buck, B., & Bailey, N. J. (2013). Inferring musical structure through bodily gestures. *Musicae Scientiae*, *17*, 86-108.

Mitterer, H., & Jesse, A. (2010). Correlation versus causation in multisensory perception. *Psychonomic Bulletin and Review*, *17*, 329-334.

Morrison, S. J., Price, H. E., Geiger, C. G., & Cornacchio, R. A. (2009). The effect of conductor expressivity on ensemble performance evaluation. *Journal of Research in Music Education*, *57*, 37-49.

Orne, M. T. (1962). On the social psychology of the psychological experiment with particular reference to demand characteristics and their implications. *American Psychologist, 17*, 776-783.

Palmer, C. (1997). Music performance. *Annual Review of Psychology*, *48*(1), 115-138.

Petrini, K., McAleer, P., & Pollick, F. (2010). Audiovisual integration of emotional signals from music improvisation does not depend on temporal correspondence. *Brain Research, 1323*, 139-148.

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance. *Music Perception*, *30*, 71-83.

Repp, B. H., & Penel, A. (2002). Auditory dominance in temporal processing: New evidence from synchronization with simultaneous visual and auditory sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *28*, 1085-1099.

Rosenblum, L. D., & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 976-985.

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception and Psychophysics*, *54*, 406-416.

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1791-1810.

Schutz, M., & Lipscomb, S. (2007). Hearing gestures, seeing music: Vision influences perceived tone duration. *Perception, 36*, 888-897.

Sloboda, J. A., & Juslin, P. N. (2010). At the interface between the inner and outer world: Psychological perspectives. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 73- 98). Oxford: Oxford University Press.

Sloboda, J. A., & Lehmann, A. C. (2001). Tracking performance correlates of changes in perceived intensity of emotion during different interpretations of a Chopin piano prelude. *Music Perception*, *19*, 87-120.

Su, Y. H. (2014). Audiovisual beat induction in complex auditory rhythms: Point-light figure movement as an effective visual beat. *Acta Psychologica*, *151*, 40-50.

Thompson, M. R., & Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae*, *16*, 19-40.

Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica, 156*, 203-227.

Thompson, W. F., Russo, F. A., & Quinto, L. (2008). Audio–visual integration of emotional cues in song. *Cognition and Emotion, 22*, 1457-1470.

Timmers, R., Marolt, M., Camurri, A., & Volpe, G. (2006). Listeners' emotional engagement with performances of a Scriabin étude: An explorative case study. *Psychology of Music*, *34*, 481-510.

Tsay, C. J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences of the USA, 110*, 14580-14585.

Van Zijl, A. G., & Luck, G. (2013). The sound of sadness: The effect of performers' emotions on audience ratings. In G. Luck, & O. Brabant (Eds.), *Proceedings of the 3rd International Conference on Music & Emotion (ICME3).* Jyväskylä, Finland: ICME3.

Vatakis, A., & Spence, C. (2008). Evaluating the influence of the "unity assumption" on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*, 12-23.

Verron, C. (2005). *Traitement et visualisation de données gesturalles captées par Optotrak* [Processing and visualizing gesture data captured by Optotrak]. Unpublished Report. Input Devices and Music Interaction Laboratory, McGill University. Retrieved from http://www.idmil.org/publications

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., Dalca, I. M., & Levitin, D. J. (2011). Music to my eyes: Cross-modal interactions in the perception of emotions in musical performance. *Cognition, 118,* 157-170.

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.

Vuoskoski, J. K., Thompson, M. R., Clarke, E. F., & Spence, C. (2014). Crossmodal interactions in the perception of expressivity in musical performance. *Attention, Perception, and Psychophysics*, *76*, 591-604.

Wanderley, M., Vines, B. W., Middleton, N., McKay, C., & Hatch, W. (2005). The musical significance of clarinetists' ancillary gestures: An exploration of the field. *Journal of New Music Research*, *34*, 97-113.

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception and Psychophysics*, *30*, 557-564.

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception and Psychophysics*, *39*, 294-300.

Williamon, A., & Davidson, J. W. (2002). Exploring co-performer communication. *Musicae Scientiae, 6*, 53-72.