BRILL *Multisensory Research (2020)* DOI:10.1163/22134808-bja10034 MULTISENSORY RESEARCH brill.com/msr

Crossmodal Visuospatial Effects on Auditory Perception of Musical Contour

Simon Lacey [1,2], **James Nguyen** [1], **Peter Schneider** [3,4] **and K. Sathian** [1,2,5,*]

[1] Department of Neurology, Milton S. Hershey Medical Center, Penn State College of Medicine, Hershey, PA 17033-0859, USA

[2] Department of Neural and Behavioral Sciences, Milton S. Hershey Medical Center, Penn State College of Medicine, Hershey, PA 17033-0859, USA

[3] Department of Neuroradiology, Heidelberg Medical School, Heidelberg, Germany

[4] Department of Neurology, Heidelberg Medical School, Heidelberg, Germany

[5] Department of Psychology, Milton S. Hershey Medical Center, Penn State College of Medicine, Hershey, PA 17033-0859, USA

Received 5 April 2019; accepted 8 July 2020

Abstract

The crossmodal correspondence between auditory pitch and visuospatial elevation (in which high- and low-pitched tones are associated with high and low spatial elevation respectively) has been proposed as the basis for Western musical notation. One implication of this is that music perception engages visuospatial processes and may not be exclusively auditory. Here, we investigated how music perception is influenced by concurrent visual stimuli. Participants listened to unfamiliar five-note musical phrases with four kinds of pitch contour (rising, falling, rising–falling, or falling–rising), accompanied by incidental visual contours that were either congruent (e.g., auditory rising/visual rising) or incongruent (e.g., auditory rising/visual falling) and judged whether the final note of the musical phrase was higher or lower in pitch than the first. Response times for the auditory judgment were significantly slower for incongruent compared to congruent trials, i.e., there was a congruency effect, even though the visual contours were incidental to the auditory task. These results suggest that music perception, although generally regarded as an auditory experience, may actually be multisensory in nature.

Keywords

Pitch contour, congruency effect, audiovisual, multisensory

[*] To whom correspondence should be addressed. E-mail: ksathian@pennstatehealth.psu.edu

DOI:10.1163/22134808-bja10034

1. Introduction

Crossmodal correspondences are near-universally experienced associations between apparently arbitrary stimulus features in different senses (Spence, 2011). A well-known example is one in which high and low auditory pitch are associated with high and low visuospatial elevation, respectively (e.g., Ben-Artzi and Marks, 1995; Bernstein and Edelstein, 1971; Evans and Treisman, 2010; Jamal *et al.*, 2017; Lacey *et al.*, 2016; McCormick *et al.*, 2018). In this study, we examine the implications that a dynamic variant of the pitch–elevation correspondence has for music perception.

Crossmodal correspondences may be important to music generally: for example, multiple different correspondences might be exploited in musical composition in order to evoke different effects (Walker, 2016). More specifically, however, the crossmodal correspondence between auditory pitch and visuospatial elevation has been suggested as the basis for Western musical notation in which low-pitched notes are written at the bottom of the stave and high-pitched notes at the top (Eitan, 2017). Such visuospatial connections to, and influences on, music processing have been widely reported. For example, in reading musical notation, the slope of the beam (the solid line connecting a rhythmic unit of notes) enables musicians and listeners to visually anticipate information about the auditory contour before actually playing or hearing the notes (Brodsky and Kessler, 2017). In addition, mental representation of auditory pitch occurs primarily along a vertical spatial axis, consistent with the pitch–elevation correspondence, although musicians are also able to employ a horizontal axis perhaps related to specific instruments, such as the piano keyboard (Lidji *et al.*, 2007).

However, the pitch–elevation crossmodal correspondence is typically studied using static, single-pitch stimuli. The problem with this, in terms of a connection with music perception, is that music does not arise from static single tones (Lidji *et al.*, 2007) but rather from dynamic changes in musical parameters such as pitch, loudness, tempo and so on (Eitan, 2013). While auditory pitch contour processing in the music domain (see Note 1) has a long research history, these studies typically involve unisensory auditory presentation with no accompanying visual stimuli (e.g., Dowling, 1978; Dowling and Fujitani, 1971; Eitan and Granot, 2006; Eitan and Tubul, 2010; Eitan *et al.*, 2012; Jeong and Ryu, 2016; Kohn and Eitan, 2016; Küssner *et al.*, 2014) thus they cannot directly address the importance of visuospatial processing to music perception. Some of these unisensory studies provide indirect evidence for visuospatial connections to music processing. For example, when asked to imagine a cartoon character's movement in response to short musical stimuli, participants reported imaging motion in all three spatial dimensions (Eitan and Granot, 2006; Eitan and Tubul, 2010). Similarly, participants produced body movements (Kohn and Eitan, 2016), hand gestures (Küssner *et al.*, 2014), or forced-choice verbal responses (Eitan *et al.*, 2012; Kohn and Eitan, 2016) that spatially matched auditory pitch contours.

Such studies establish an association between auditory pitch contour and visuospatial processing but not necessarily its functional importance — participants might imagine visual motion when asked to do so in an experimental setting but not otherwise — nor the strength of this association. Establishing functional relevance could be achieved using bisensory audiovisual stimuli and either interference or congruency effects (i.e., functional relevance can be demonstrated by showing that auditory pitch contour processing is impaired by either irrelevant visual stimuli or incongruence between auditory and visual contours) or training effects (i.e., that auditory pitch contour processing is improved in the presence of visual contours). In a Garner paradigm using auditory pitch glides and concurrent visual motion stimuli, Eitan and Marks (2012) showed that participants could discriminate between audiovisual ascending/descending stimuli faster when these were congruent compared to incongruent. But this study did not find Garner interference for these dynamic stimuli, suggesting that auditory and visual contour are not integrated (Eitan and Marks, 2012). In a bisensory study, Maeda *et al.* (2004) showed that auditory pitch glides could disambiguate visual grating motion but this study required a visual rather than an auditory judgment, thus its relevance to music perception is unclear. Foxton *et al.* (2004) report that pitch contour perception is improved by training with unisensory auditory tasks requiring same/different comparisons of pitch contour or actual pitches, but not by an audiovisual contour task. Although this study suggests that processing visuospatial and auditory pitch contours are independent, the visual contour was presented before the auditory pitch contour rather than concurrently. Non-concurrent auditory–visual presentation was also used in other studies, for example Wagner *et al.* (1981).

More recently, Lu *et al.* (2017a) presented auditory pitch contours with concurrent visual contours that could be either congruent or incongruent and showed that amusic individuals were less accurate than control participants at judging audiovisual congruency. However, it may be difficult to disentangle visuospatial effects from pitch perception effects in this study since amusic participants are not only impaired at pitch contour processing (Peretz *et al.*, 2003) but may also have spatial deficits (Douglas and Bilkey, 2007; Peretz and Vuvan, 2017; see also Stewart and Walsh, 2007). Additionally, the task in the study of Lu *et al.* (2017a) required an explicit congruency judgment and was made easier by the fact that presentation of the audiovisual contours was self-paced, advancing only when participants pressed a button; the visual contour remained visible for the duration of a trial; and incongruency was confined to a single note in a seven-note sequence. Here, we employed concurrent audiovisual presentation of ecologically valid musical phrases using visual contours that were congruent or incongruent with the auditory pitch contour at each point in a five-note sequence, without showing the whole contour at once (see Materials and Methods). Participants were asked to decide whether the final note in the auditory sequence was higher or lower in pitch than the first, such that the visual contours were merely incidental to the auditory task, rather than being explicitly processed as in the work of Lu *et al.* (2017a).

2. Materials and Methods

2.1. Participants

Twenty-four people (12 male, 12 female; mean age 25 years) took part after giving informed consent and were compensated for their time. All procedures were approved by the Institutional Review Board of Penn State College of Medicine. Eight participants reported that they had no musical experience of playing an instrument or singing (i.e., had received no musical training and were not self-taught musicians), while the remaining 16 had between 2 and 8 years of experience (mean ± SD 3.7 ± 1.8).

2.2. Materials

We created 12 examples for each of four types of auditory contours (rising, falling, rising–falling, and falling–rising: Fig. 1), using the piano setting in Sibelius Ultimate (Avid Technology Inc., Cambridge UK), for a total of 48 auditory stimuli. Each contour consisted of five notes, the total duration being 3 s. In order to avoid expectancies based on general musical knowledge, the pitch interval between successive notes was not linear (see, for example, the falling contour in Fig. 1). In order to introduce variability and further avoid expectancies about the contour, for the rising–falling and falling–rising contours, the contour changed direction after the second, third, or fourth note, with four examples of each in each group. For the rising–falling contours, the final note was always lower in pitch than the first and for falling–rising, the final note was always higher than the first. These auditory contours were then included in short animated movie clips created in Adobe After Effects (Adobe Systems, San Jose, CA, USA) with file conversion for software compatibility using the VLC media player (VideoLAN, Paris, France), Handbrake (https://handbrake.fr/), and Virtual Dub (http://www.virtualdub.org/). In the movie clips, each note was accompanied by a horizontal black bar, 30 mm wide, 9 mm high; subtending approximately 3° of visual angle at a viewing distance of approximately 60 cm [note that the bars provide a visual representation of pitch height and contour that is independent of expertise in reading musical notation (Brodsky and Kessler, 2017)]. In order to match the auditory

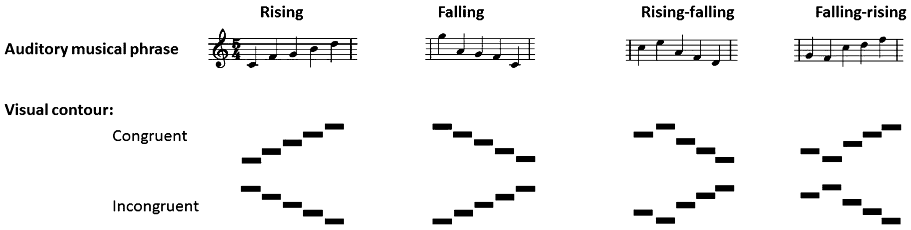

Figure 1. Examples of musical phrases and concurrent visual contours.

contour presentation, each bar was only present for the duration of the associated note before being replaced, i.e., unlike Lu *et al.* (2017a), the complete visual contour was not shown, and bars progressed from left to right across the screen. Additionally, there was a 30-mm gap between the positions of each bar on the screen so that the visual contour was over the full extent of the screen. The third visual bar, accompanying the middle note of each five-note contour, was always presented at the center of the screen. The vertical interval between successive bars (from the top edge of one to the bottom edge of the next) was fixed at 20 mm but the bar height always rose or fell corresponding to the auditory contour. The visual vertical interval was therefore only relative to the auditory pitch interval rather than matching it in absolute magnitude. This is in line with the weight of the evidence being that the pitch/elevation correspondence is indeed relative rather than absolute (Spence, 2019). The visual contour accompanying each auditory stimulus was either congruent (e.g., auditory rising/visual rising) or incongruent (e.g., auditory rising/visual falling), for a total of 48 audiovisual stimuli in each condition.

The stimuli were divided into two runs, each containing 24 congruent and 24 incongruent trials, six trials for each of the four contours; for the rising–falling and falling–rising contours, there were two trials for each of the three points at which the direction of the contour changed, i.e., after the second, third, or fourth note. A trial consisted of the 3-s stimulus presentation with 4 s for a response, totaling 7 s. Each run began with a blank 2-s interval for a total of 338 s. Within a run, stimuli were presented in a fixed but pseudorandom order; the order of the two runs was fully counterbalanced across participants and genders. Runs were presented using Presentation software (Neurobehavioral Systems Inc., Albany, CA, USA) which also recorded responses and response times (RTs).

2.3. Procedure

Participants were tested individually in a quiet, well-lit room. Prior to the main experiment, participants were familiarized with the task by watching two examples of each contour but without knowledge of the subsequent task; these trials were always congruent in order to avoid participants becoming aware that incongruity might be important. Participants listened to stimuli *via* noise-cancelling headphones at a volume comfortable for them and set individually when listening to the example stimuli. On each trial, participants decided whether the final note was higher or lower in pitch than the first and pressed the left or right buttons of a wireless computer mouse to indicate their answer. Nine participants chose response button pairings of left/low, right/high and 15 chose left/high, right/low; while this meant that response button pairings were not counterbalanced, it allowed participants to choose a pairing that felt natural, bearing in mind that pitch can be represented on a horizontal left/right axis (Lidji *et al.*, 2007). This avoided the potential for responses to be confounded by a response button pairing that was counter-intuitive to the participant.

After completing the main experiment, participants also completed the Clarity of Auditory Imagery Scale (CAIS: Willander and Baraldi, 2010), the scale test from the Montreal Battery for the Evaluation of Amusia (MBEA: Peretz *et al.*, 2003), the Object-Spatial Imagery and Verbal Questionnaire (OSIVQ: Blazhenkova and Kozhevnikov, 2009), and the Holistic-Spectral Pitch Perception test (HSPP: Schneider *et al.*, 2005). The HSPP identifies five classes of listeners by reference to a pitch perception preference index: strong fundamental pitch listeners (index -1 to -0.76), fundamental pitch listeners (-0.75 to -0.26), mixed listeners (-0.25 to 0.25), spectral pitch listeners (0.26 to 0.75) and strong spectral pitch listeners (0.76 to 1: Schneider *et al.*, 2005).

2.4. Data Analysis

Trials for which there was no response constituted 1.4% of all trials. Analysis of RTs was based on correct responses only (93.9% of all responses) and excluding outliers, defined as RTs greater than two standard deviations away from the mean and calculated separately for the two runs (5.8% of all correct responses: 3.2% of congruent trials and 2.6% of incongruent trials). RTs were calculated from the onset of the final note in each sequence since this was the earliest point at which a participant could make the first *versus* last pitch discrimination. For correlational analyses, we used the non-parametric Spearman's test because some variables (e.g., self-reported years of musical experience) were not normally distributed. Effect sizes (Cohen's *d*) were calculated using the online tool provided by Lenhard and Lenhard (2016).

3. Results

Response times were significantly slower for incongruent (mean ± SEM: 832 ± 52 ms) compared to congruent (774 ± 49 ms) trials (paired $t_{23} = -4.9$, $p < 0.001$, $d = 0.2$: Fig. 2a, top panel) and accuracy was significantly lower for incongruent (90.6 ± 2.9%) than congruent (96.9 ± 0.6%) trials (paired

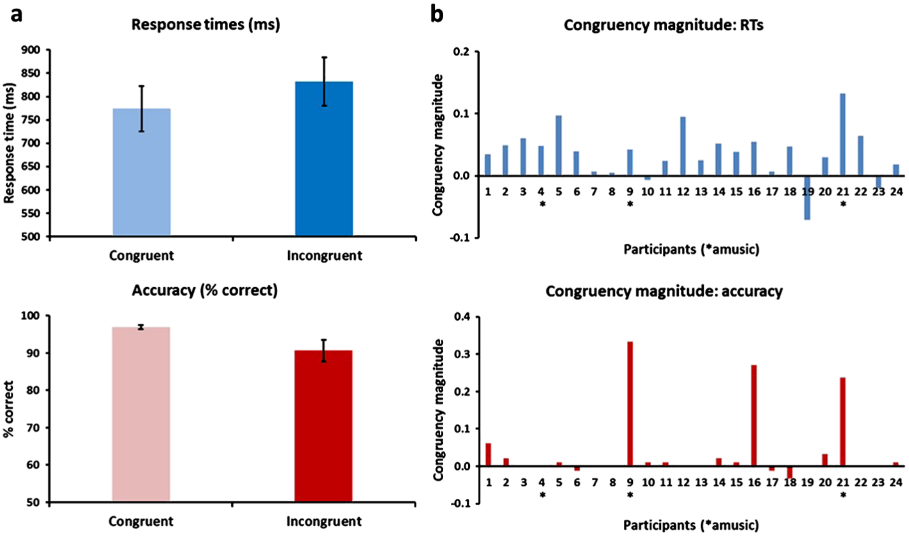

Figure 2. Responses were significantly slower (a: top) and less accurate (a: bottom) for incongruent, compared to congruent, trials. Magnitude of individual congruency effects for RTs (b: top) and accuracy (b: bottom).

$t_{23} = 2.2$, $p = 0.04$, $d = 0.5$: Fig. 2a, bottom panel). RTs and accuracy were uncorrelated for both congruent ($r = 0.03$, $p = 0.9$) and incongruent ($r = 0.1$, $p = 0.5$) trials, indicating that there was no speed–accuracy trade-off. The MBEA scale test scores identified three potentially amusic participants (MBEA score <23/30, Peretz *et al.*, 2003); the RT congruency effect remained significant if these participants were excluded from analysis ($t_{20} = -4.2$, $p < 0.001$, $d = 0.2$) but the accuracy congruency effect fell short of significance ($t_{20} = 1.9$, $p = 0.07$, $d = 0.5$).

The magnitude of the congruency effect was calculated for each subject by taking the difference of RT or accuracy (A) values in the congruent and incongruent conditions and dividing by their sum; i.e., $(\mathrm{RT_i} - \mathrm{RT_c})/(\mathrm{RT_i} + \mathrm{RT_c})$; $(\mathrm{A_c} - \mathrm{A_i})/(\mathrm{A_c} + \mathrm{A_i})$, where the subscripts refer to the congruent/incongruent conditions. Figure 2b shows individual congruency magnitudes for RTs (top panel) and accuracy (bottom panel). RT congruency magnitudes were not significantly correlated with scores on the MBEA, CAIS, or OSIVQ, nor with years of musical experience (Spearman's rho −0.3–0.1, all p values > 0.2). Congruency magnitudes for accuracy were significantly negatively correlated with OSIVQ spatial scores (Spearman's rho = −0.5, $p = 0.02$), indicating that the magnitude of the congruency effect for accuracy decreased as individual preference for spatial imagery increased. However, this correlation may have been affected by outliers, including two of the three potentially amusic individuals; excluding the outliers reduced this relationship to a trend that was short of significance (Spearman's rho $= -0.4$, $p = 0.07$). The HSPP identified 11 strong fundamental listeners, 11 fundamental pitch listeners and two mixed listeners, but no spectral listeners. Pitch perception preferences were not significantly correlated with congruency magnitudes for either RTs ($r = 0.25$, $p = 0.2$) or accuracy ($r = 0.35$, $p = 0.09$) but were significantly negatively correlated with performance on the MBEA scale test ($r = -0.43$, $p = 0.03$). This correlation is negative because MBEA scale scores increase as the preference index decreases (becomes less negative), but the actual relationship is positive in that MBEA scale test performance increases with increasingly stronger fundamental pitch perception.

One-way ANOVAs showed no effect of gender (congruency magnitude RT: $F_{1,22} = 0.06$, $p = 0.8$, $d = 0.1$; accuracy: $F_{1,22} = 0.9$, $p = 0.3$, $d = 0.4$). Additionally, the number of years of musical experience was uncorrelated with the congruency magnitudes for RTs (rho $= 0.02$, $p = 0.9$) and accuracy (rho $= -0.1$, $p = 0.5$); this remained so even when participants with zero years of experience were excluded.

4. Discussion

In this study, we showed that auditory pitch contour processing is impaired by a simultaneously presented incongruent visual contour even though the visual contour was incidental and irrelevant to the auditory task. The effect is likely not due to divided attention because accuracy was high in both congruent and incongruent conditions, although we acknowledge that in the absence of an auditory-only condition showing even higher performance, we cannot rule out a visual effect even on the congruent condition. Our results complement studies in which participants had to judge the visual element of an audiovisual stimulus (Romero-Rivas *et al.*, 2018), i.e., the effect is likely bidirectional. These findings add to earlier evidence for non-auditory influences on pitch perception; in these earlier studies, the auditory contour was presented in isolation (Eitan and Granot, 2006; Eitan and Tubul, 2010; Eitan *et al.*, 2012; Kohn and Eitan, 2016; Küssner *et al.*, 2014) or both auditory and visual contours were attended in order to make an explicit judgment of congruence (Lu *et al.*, 2017a), in contrast to the present study where concurrent auditory and visual stimuli were presented but only an auditory judgment was required.

The present results could not be explained by individual differences in pitch perception or variation in preferences for holistic *versus* spectral listening (although a caveat is that our sample did not contain any pure spectral listeners). Nor was performance related to individual differences in auditory imagery, although clarity of imagery, like vividness, may be a poor indicator of individual imagery ability and specific component processes (Lacey and Lawson, 2013).

However, consistent with other reports of spatial influences on different aspects of music processing (e.g., Brodsky and Kessler, 2017; Eitan, 2017; Lidji *et al.*, 2007), the visual effect on the pitch judgment task tended to decrease (i.e., congruency magnitudes became smaller) with increasing preference for visual spatial, as opposed to visual object imagery.

In addition, although the present study only required a judgment on the auditory stimuli, it is possible that attention was divided between the simultaneously presented auditory and visual contours, thus potentially impairing performance analogously to a dual-task design. This effect can be mitigated by relevant expertise or training which can reduce attentional demands (Cocchini *et al.*, 2017); thus, in the present study it was possible that more musical experience might have produced better pitch perception performance, despite the presence of a concurrent visual stimulus. In fact, consistent with earlier findings that musicians and non-musicians are equally impaired in dual-task performance (Cocchini *et al.*, 2017), we found no effect of musical experience on congruency effects for either RTs or accuracy. Although some participants reported as little as two years' playing or singing experience and thus would be unlikely to fit formal definitions as a 'musician', this and higher levels of experience undoubtedly reflected adequate ability to perform the simple pitch discrimination task required of them and yet such experience did not seem to help overcome the audiovisual incongruity. Note too, that even those participants who reported no musical experience, in the sense that they had not learned to play or sing *via* formal tuition or by being self-taught, were unlikely to lack *any* musical experience since they likely were exposed to music recordings and various media, e.g., films, television, etc. (apart from potentially amusic individuals, see below).

Our sample included three individuals (12.5%) who scored below the cutoff for amusia on the scale test of the MBEA (Peretz *et al.*, 2003). This is considerably higher than the current estimate for the prevalence of congenital amusia at just 1.5% (Peretz and Vuvan, 2017). However, it is important to note that we used the MBEA merely to screen for potential confounds rather than as a diagnostic procedure which would require more extensive testing (Vuvan *et al.*, 2018). In relation to these potentially amusic individuals, it has been suggested that the pitch perception deficit in amusia is accompanied by deficits in spatial processing (Douglas and Bilkey, 2007; see also Stewart and Walsh, 2007). Consistent with this suggestion, two of the three potential amusics in our sample had two of the strongest preferences for object, rather than spatial, imagery and also two of the largest congruency magnitudes for accuracy (Fig. 2b, bottom panel; the effect was less pronounced for RT congruency magnitudes), consistent with the poor performance for amusics compared to non-amusics shown by Lu *et al.* (2017a). However, the relationship between amusia and spatial deficits is likely complex. The mental rotation deficit reported by Douglas and Bilkey (2007) was not replicated by Tillmann *et al.* (2010) and may be confined to the more severe cases of amusia (Williamson *et al.*, 2011). Other studies suggest that the deficit is restricted to implicit, and does not affect explicit spatial processing (Lu *et al.*, 2016) and/or explicit pitch processing (Lu *et al.*, 2017b). Nonetheless, difficulty in spatial orientation is the only type of deficit consistently reported as comorbid with congenital amusia (Peretz and Vuvan, 2017). An interesting incidental finding was that MBEA scale test performance increased with increasingly stronger fundamental pitch perception; this is hard to interpret since our sample did not include any spectral listeners but may be worth following up in future work.

Neuroimaging studies suggest the involvement of spatial processes in music processing, particularly in the intraparietal sulcus (IPS), a region of which is also sensitive to the crossmodal pitch/elevation correspondence (McCormick *et al.*, 2018). For example, the IPS is involved in processing relative pitch during a task requiring recognition of transposed melodies (Foster and Zatorre, 2010) at a site close to that identified by McCormick *et al.* (2018), and is also engaged in spatial transformation tasks such as mental rotation (e.g., Jordan *et al.*, 2001; Zacks, 2008) as well as more general spatial imagery tasks (e.g., Mellet *et al.*, 1996; Sack *et al.*, 2008). It has also been shown, using multivariate pattern analysis, that different categories of melodic contour, either ascending or descending, can be decoded from activity in the left inferior parietal lobule, particularly at an IPS focus (Lee *et al.*, 2011). However, that study did not present a concurrent visual contour and it remains unknown whether decoding performance would track behavioral performance in being better on congruent compared to incongruent trials. Relatedly, recent work has shown that auditory contour (simple ascending/descending Shepherd tones) can be decoded from activity in early and extrastriate visual cortex, but only when the classifier is trained and tested within-modally (Ha *et al.*, 2019); when the classifier was trained on the auditory data and tested on the visual, or *vice versa*, decoding accuracy fell to chance levels. However, this study's focus on early visual processing areas does not preclude the possibility of crossmodal decoding in later visual cortical areas.

Finally, we acknowledge potential limitations of the current study. Firstly, it might be objected that the stimuli, in avoiding pitch expectancies based on tonal music, were not ecologically valid as most music is tonal in nature. Nonetheless, there is a large corpus of atonal music, for example, by Berg, Schoenberg, Webern, Messiaen, Bartok and others (see Lansky *et al.*, 2001) and some forms of jazz are atonal (see Robinson, 2002). We consider that our stimuli steer a middle course between being not obviously music [e.g., pitch glides (Eitan and Marks, 2012; Maeda *et al.*, 2004) and Shepherd tones (Ha *et al.*, 2019)] and actual tonal music which might be confounded with familiarity and/or the expectancies about note successions that come with long exposure to conventional tonal music. Thus, we consider our stimuli to be ecologically valid within the confines of experimental control over potential confounds. However, further work, using excerpts from familiar tonal pieces, should determine whether expectancies deriving from exposure to conventional tonal music can overcome the audiovisual incongruity tested here. Similarly, although we attend concerts and watch films with a musical soundtrack, music is not always accompanied by visual input, though it was a necessary part of the experimental design here. Secondly, two aspects of experimental procedure might have influenced the results: familiarization trials and task instructions. Prior to the main experiment, participants watched/listened to two examples of each contour which were always congruent but without being aware of the subsequent task. While we acknowledge that this might have led to a priming effect, this would also have been the case if we had presented examples of incongruent trials as well; not least because this might have alerted participants to the fact that incongruity was important. When performing the main experiment, participants were asked to say whether the final note of the sequence was higher or lower than the first, which might have invoked the metaphorical pitch–height relation (Lakoff and Johnson, 1980). While such explicit task instructions have been used in making pitch judgments (for example, Fernandez-Prieto *et al.*, 2017), a less explicit 'same/different' decision should also be tested given the relative weakness of the pitch–height correspondence in speakers of languages that use other spatial terms for pitch (e.g., thin–thick, see Dolscheid *et al.*, 2013).

5. Conclusion

In this study, the visual contour accompanying the auditory pitch contour can be seen as a dynamic variant of the pitch/elevation crossmodal correspondence. Responses were slower and less accurate when the visual contour was incongruent compared to congruent, as is also seen when the correspondence is between single auditory tones and visuospatial locations (e.g., Ben-Artzi and Marks, 1995; Bernstein and Edelstein, 1971; Evans and Treisman, 2010; Jamal *et al.*, 2017; Lacey *et al.*, 2016; McCormick *et al.*, 2018). While we do not claim to examine auditory or visual imagery exhaustively, our results show a closer connection to visuospatial, rather than auditory, imagery, and were unrelated to either listening preferences or musical training. Since music cannot arise from single notes (Eitan, 2013; Lidji *et al.*, 2007), these results extend the effect of the crossmodal correspondence into the musical domain and, since the visual contour was merely incidental to the auditory task, also suggest that our experience of music is inherently multisensory.

Acknowledgements

This work was supported by NIH grant R01EY025978 (PIs: KS and Lynne Nygaard); support for KS from the Veterans Administration is also acknowledged. We thank Eliza Donne, Multimedia Specialist, Harrell Health Sciences Library, Penn State College of Medicine, for creating the audiovisual stimuli.

Note

1. Note that in the speech domain, correspondence between auditory and visual contours may be useful in speech therapy or learning tonal languages (Hermes, 1998).

References

Ben-Artzi, E. and Marks, L. E. (1995). Visual–auditory interaction in speeded classification: role of stimulus difference, *Percept. Psychophys.* **57**, 1151–1162. DOI:10.3758/BF03208371.

Bernstein, I. H. and Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time, *J. Exp. Psychol.* **87**, 241–247. DOI:10.1037/h0030524.

Blazhenkova, O. and Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: theory and measurement, *Appl. Cognit. Psychol.* **23**, 638–663. DOI:10.1002/acp.1473.

Brodsky, W. and Kessler, Y. (2017). The effect of beam slope on the perception of melodic contour, *Acta. Psychol.* **180**, 190–199. DOI:10.1016/j.actpsy.2017.09.013.

Cocchini, G., Filardi, M. S., Crhonkova, M. and Halpern, A. R. (2017). Musical expertise has minimal impact on dual task performance, *Memory* **25**, 677–685. DOI:10.1080/09658211.2016.1205628.

Dolscheid, S., Shayan, S., Majid, A. and Casasanto, D. (2013). The thickness of musical pitch: psychophysical evidence for linguistic relativity, *Psychol. Sci.* **24**, 613–621. DOI:10.1177/0956797612457374.

Douglas, K. M. and Bilkey, D. K. (2007). Amusia is associated with deficits in spatial processing, *Nat. Neurosci.* **10**, 915–921. DOI:10.1038/nn1925.

Dowling, W. J. (1978). Scale and contour: two components of a theory of memory for melodies, *Psychol. Rev.* **85**, 341–354. DOI:10.1037/0033-295X.85.4.341.

Dowling, W. J. and Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies, *J. Acoust. Soc. Am.* **41**, 524–531. DOI:10.1121/1.1912382.

Eitan, Z. (2013). How pitch and loudness shape musical space and motion, in: *The Psychology of Music in Multimedia*, S.-L. Tan, A. J. Cohen, S. D. Lipscombe and R. A. Kendall (Eds), pp. 165–191. Oxford University Press, Oxford, UK.

Eitan, Z. (2017). Musical connections: crossmodal correspondences, in: *The Routledge Companion to Music Cognition*, R. Ashley and R. Timmers (Eds), pp. 213–224. Routledge, New York, NY, USA.

Eitan, Z. and Granot, R. Y. (2006). How music moves: musical parameters and listeners' images of motion, *Music Percept.* **23**, 221–247. DOI:10.1525/mp.2006.23.3.221.

Eitan, Z. and Marks, L. E. (2012). Garner's paradigm and audiovisual correspondence in dynamic stimuli: pitch and vertical direction, *See. Perceiv.* **25**, 70. DOI:10.1163/187847612X646910.

Eitan, Z. and Tubul, N. (2010). Musical parameters and children's images of motion, *Music Sci.* **14**(Suppl. 2), 89–111. DOI:10.1177/10298649100140S207.

Eitan, Z., Ornoy, E. and Granot, R. Y. (2012). Listening in the dark: congenital and early blindness and cross-domain mappings in music, *Psychomusicol. Music Mind Brain* **22**, 33–45. DOI:10.1037/a0028939.

Evans, K. K. and Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features, *J. Vis.* **10**, 6. DOI:10.1167/10.1.6.

Fernandez-Prieto, I., Spence, C., Pons, F. and Navarra, J. (2017). Does language influence the vertical representation of auditory pitch and loudness?, *i-Perception* **8**. DOI:10.1177/2041669517716183.

Foster, N. E. V. and Zatorre, R. J. (2010). A role for the intraparietal sulcus in transforming musical pitch information, *Cereb. Cortex* **20**, 1350–1359. DOI:10.1093/cercor/bhp199.

Foxton, J. M., Brown, A. C. B., Chambers, S. and Griffiths, T. D. (2004). Training improves acoustic pattern perception, *Curr. Biol.* **14**, 322–325. DOI:10.1016/j.cub.2004.02.001.

Ha, J., Kim, I. and Shim, W. (2019). Decoding melodic contours in early visual areas, in: Neuroscience 2019, Abstract 665.09, *Society for Neuroscience*, October 2019, Chicago, IL. USA.

Hermes, D. J. (1998). Auditory and visual similarity of pitch contours, *J. Speech Lang. Hear. Res.* **41**, 63–72. DOI:10.1044/jslhr.4101.63.

Jamal, Y., Lacey, S., Nygaard, L. and Sathian, K. (2017). Interactions between auditory elevation, auditory pitch and visual elevation during multisensory perception, *Multisens. Res.* **30**, 287–306. DOI:10.1163/22134808-00002553.

Jeong, E. and Ryu, H. (2016). Melodic contour identification reflects the cognitive threshold of aging, *Front. Aging Neurosci.* **8**, 134. DOI:10.3389/fnagi.2016.00134.

Jordan, K., Heinze, H.-J., Lutz, K., Kanowski, M. and Jäncke, L. (2001). Cortical activations during the mental rotation of different visual objects, *NeuroImage* **13**, 143–152. DOI:10.1006/nimg.2000.0677.

Kohn, D. and Eitan, Z. (2016). Moving music: correspondences of musical parameters and movement dimensions in children's motion and verbal responses, *Music Percept.* **34**, 40–55. DOI:10.1525/mp.2016.34.1.40.

Küssner, M. B., Tidhar, D., Prior, H. M. and Leech-Wilkinson, D. (2014). Musicians are more consistent: gestural cross-modal mappings of pitch, loudness and tempo in real time, *Front. Psychol.* **5**, 789. DOI:10.3389/fpsyg.2014.00789.

Lacey, S. and Lawson, R. (2013). Imagery questionnaires: vividness and beyond, in: *Multisensory Imagery*, S. Lacey and R. Lawson (Eds), pp. 271–282. Springer, New York, NY, USA. DOI:10.1007/978-1-4614-5879-1_14.

Lacey, S., Martinez, M., McCormick, K. and Sathian, K. (2016). Synesthesia strengthens sound-symbolic cross-modal correspondences, *Eur. J. Neurosci.* **44**, 2716–2721. DOI:10.1111/ejn.13381.

Lakoff, G. and Johnson, M. (1980). The metaphorical structure of the human conceptual system, *Cogn. Sci.* **4**, 195–208.

Lansky, P., Perle, G. and Headlam, D. (2001). Atonality, in: *The New Grove Dictionary of Music and Musicians*, 2nd edn., S. Sadie and J. Tyrrell (Eds), pp. 138–145. Macmillan Publishers, London, UK.

Lee, Y.-S., Janata, P., Frost, C., Hanke, M. and Granger, R. (2011). Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI, *NeuroImage* **57**, 293–300. DOI:10.1016/j.neuroimage.2011.02.006.

Lenhard, W. and Lenhard, A. (2016). *Computation of Effect Sizes*. *Psychometrica*. Accessed at https://www.psychometrica.de/effect_size.html. DOI:10.13140/RG.2.2.17823.92329.

Lidji, P., Kolinsky, R., Lochy, A. and Morais, J. (2007). Spatial associations for musical stimuli: a piano in the head?, *J. Exp. Psychol. Hum. Percept. Perform.* **33**, 1189–1207. DOI:10.1037/0096-1523.33.5.1189.

Lu, X., Ho, H. T., Sun, Y., Johnson, B. W. and Thompson, W. F. (2016). The influence of visual information on auditory processing in individuals with congenital amusia: an ERP study, *NeuroImage* **135**, 142–151. DOI:10.1016/j.neuroimage.2016.04.043.

Lu, X., Sun, Y., Ho, H. T. and Thompson, W. F. (2017a). Pitch contour impairment in congenital amusia: new insights from the self-paced audio-visual contour task (SACT), *PLoS ONE* **12**, e0179252. DOI:10.1371/journal.pone.0179252.

Lu, X., Sun, Y. and Thompson, W. F. (2017b). An investigation of spatial representation of pitch in individuals with congenital amusia, *Q. J. Exp. Psychol.* **70**, 1867–1877. DOI:10.1080/17470218.2016.1213870.

Maeda, F., Kanai, R. and Shimojo, S. (2004). Changing pitch induced visual motion illusion, *Curr. Biol.* **14**, R990–R991. DOI:10.1016/j.cub.2004.11.018.

McCormick, K., Lacey, S., Stilla, R., Nygaard, L. and Sathian, K. (2018). Neural basis of the crossmodal correspondence between auditory pitch and visuospatial elevation, *Neuropsychologia* **112**, 19–30. DOI:10.1016/j.neuropsychologia.2018.02.029.

Mellet, E., Tzourio, N., Crivello, F., Joliot, M., Denis, M. and Mazoyer, B. (1996). Functional anatomy of spatial imagery generated from verbal instructions, *J. Neurosci.* **16**, 6504–6512. DOI:10.1523/JNEUROSCI.16-20-06504.1996.

Peretz, I. and Vuvan, D. T. (2017). Prevalence of congenital amusia, *Eur. J. Hum. Genet.* **25**, 625–630. DOI:10.1038/ejhg.2017.15.

Peretz, I., Champod, A. S. and Hyde, K. (2003). Varieties of musical disorders: the Montreal battery of evaluation of amusia, *Ann. N.Y. Acad. Sci.* **999**, 58–75. DOI:10.1196/annals.1284.006.

Robinson, J. B. (2002). Free jazz, in: *The New Grove Dictionary of Jazz, Vol. 1*, 2nd edn., B. Kernfeld (Ed.), pp. 848–849. Grove's Dictionaries, New York.

Romero-Rivas, C., Vera-Constán, F., Rodríguez-Cuadrado, S., Puigcerver, L., Fernández-Prieto, I. and Navarra, J. (2018). Seeing music: the perception of melodic 'ups and downs' modulates the spatial processing of visual stimuli, *Neuropsychologia* **117**, 67–74. DOI:10.1016/j.neuropsychologia.2018.05.009.

Sack, A. T., Jacobs, C., De Martino, F., Staeren, N., Goebel, R. and Formisano, E. (2008). Dynamic premotor-to-parietal interactions during spatial imagery, *J. Neurosci.* **28**, 8417–8429. DOI:10.1523/JNEUROSCI.2656-08.2008.

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., Dosch, H. G., Bleeck, S., Stippich, C. and Rupp, A. (2005). Structural and functional asymmetry of lateral

Heschl's gyrus reflects pitch perception performance, *Nat Neurosci* **8**, 1241–1247. DOI:10.1038/nn1530.

Spence, C. (2011). Crossmodal correspondences: a tutorial review, *Atten. Percept. Psychol.* **73**, 971–995. DOI:10.3758/s13414-010-0073-7.

Spence, C. (2019). On the relative nature of (pitch-based) crossmodal correspondences, *Multisens. Res.* **32**, 235–265. DOI:10.1163/22134808-20191407.

Stewart, L. and Walsh, V. (2007). Music perception: sounds lost in space, *Curr. Biol.* **17**, R892–R893. DOI:10.1016/j.cub.2007.08.012.

Tillmann, B., Jolicoeur, P., Ishihara, M., Gosselin, N., Bertrand, O., Rossetti, Y. and Peretz, I. (2010). The amusic brain: lost in music, but not in space, *PLoS ONE* **5**, e10173. DOI:10.1371/journal.pone.0010173.

Vuvan, D. T., Paquette, S., Mignault Goulet, G., Royal, I., Felezeu, M. and Peretz, I. (2018). The Montreal protocol for identification of amusia, *Behav. Res. Methods* **50**, 662–672. DOI:10.3758/s13428-017-0892-8.

Wagner, S., Winner, E., Cicchetti, D. and Gardner, H. (1981). 'Metaphorical' mapping in human infants, *Child Dev.* **52**, 728–731. DOI:10.2307/1129200.

Walker, P. (2016). Cross-sensory correspondences: a theoretical framework and their relevance to music, *Pyschomusicol. Music Mind Brain* **26**, 103–116. DOI:10.1037/pmu0000130.

Willander, J. and Baraldi, S. (2010). Development of a new Clarity of Auditory Imagery Scale, *Behav. Res. Methods* **42**, 785–790. DOI:10.3758/BRM.42.3.785.

Williamson, V. J., Cocchini, G. and Stewart, L. (2011). The relationship between pitch and space in congenital amusia, *Brain Cogn.* **76**, 70–76. DOI:10.1016/j.bandc.2011.02.016.

Zacks, J. M. (2008). Neuroimaging studies of mental rotation: a meta-analysis and review, *J. Cogn. Neurosci.* **20**, 1–19. DOI:10.1162/jocn.2008.20013.