What you see is what you hear

What you see is what you hear: the importance of visual priming in music performer identification

Helen F Mitchell

Raymond AR MacDonald

Sydney Conservatorium of Music, University of Sydney, Australia

Edinburgh University, UK

Address for correspondence: Helen Mitchell, Conservatorium of Music C41, University of Sydney NSW 2006 Australia, e-mail: helen.mitchell@sydney.edu.au

Abstract

Visual information plays a critical role in the assessment of music performance. Audiovisual integration is well recognised in person perception, and people readily match talking faces to speaking voices. This effect exists in identifying music performers, but its strength is untested. This study investigated the importance of visual or audio priming in identifying a music performer from a line-up. Half the participants saw a target saxophonist (no sound) and then heard a line-up (no visuals) of saxophonists playing (2 to 5 saxophonists). In contrast, half the participants heard a target saxophonist (no visuals) and then saw the line-up (no sound). Participants identified the target saxophonist in visual and audio line-ups at a rate above chance, although identification accuracy decreased as the line-up number increased. Those who saw the targets identified significantly greater number of performers from the audio line-up than those who heard the targets and identified them from a visual line-up. As the task complexity and number of distractors increased, responses remained consistent and visual priming was robust and reliable in performer identification.

Keywords: Music performance; music perception, auditory identification, audiovisual integration, non-verbal communication.

Audiovisual integration is critical to the reception of music performance. While listeners maintain they are focused on sound when evaluating performers, recent music research suggests that audiences use visual information to complement their audio experience and that visual information alone can transmit musical intentions and quality. By isolating audio and visual representations of performance, we can ascertain where musical and extra-musical cues are available to audiences. Davidson (1993) established that pianists' communicate expressive manner or intentions through body movements, and audiences absorb and rely on this visual

information to interpret performers'. Viewing the performer's body movement played a critical role in communicating expressive intent. The visual-only conditions provided enhanced information about the performer's manner than audio-only and were essential to audiences' impressions of the performance style.

Non-verbal cues of expressivity would appear to be universal (Thompson, Graham, & Russo, 2005), and have been demonstrated in pianists (Thompson, & Luck, 2012), singers (Thompson, Russo, & Quinto, 2008), woodwind (Dahl & Friberg, 2007) and marimba players (Broughton & Stevens, 2009). The visual signal plays the predominant role in interpreting a variety of musical signals, like structural phrasing (Vines, Krumhansl, Wanderley, & Levitin, 2006), expressive intentions (e.g., Broughton & Stevens, 2009; Dahl & Friberg, 2007; Davidson, 1993; Vuoskoski, Thompson, Clarke, & Spence, 2013) emotional cues (e.g., Thompson, et al., 2008) and performance quality (Tsay, 2013). Indeed, visual information plays a more important role that audio in judges' ratings of expressivity and audiences obtain more information about expressivity by sight alone (Dahl & Friberg, 2007; Vuoskoski et al., 2013). 'Listeners' are so reliant on visual cues, they evaluate performers from the moment they step on stage (Platz & Kopiez, 2013) and favour performers who smile or make eye contact with the audience (Wapnick, Darrow, Kovacs, & Dalrymple, 1997; Wapnick, Mazza, & Darrow, 1998). While music research highlights the importance of visual information and its interaction with audio, there is still much to be learned about the extent to which listeners integrate the more subtle audiovisual information to identify individual performer characteristics.

The recent musical evidence that suggests listeners pay more attention to visual cues than audio is both controversial and challenging to the music profession (Tsay, 2013). Music is regarded

as an aural art, and listeners are conditioned to think that aesthetic judgments and critiques of music are based purely on sound. Yet, in a real world setting, person identification or recognition through multiple sensory channels plays a central role in our social interactions. Humans observe others and derive information about individuals beyond emotional cues and behaviours in order to identify them. Both aural and visual presentations are rich in information, not only about the performance but also about the performer.

In person recognition, the fusion of audio and visual information is critical to our ability to identify individual speakers as faces and voices present the same information about a speaker, but in a different way. In speaker identification, listeners integrate cross-modal perceptions (through sight and sound) to recognise the person speaking (Chartrand & Belin, 2006). Listener/viewers can match an unfamiliar voice to a face, or indeed, a face to a voice (Kamachi et al., 2003; Lachs & Pisoni, 2004). They match a muted clip of an unfamiliar face articulating a word to one of two audio clips (of the speaker, and a distractor), or voice to face, at a rate significantly above chance. For successful face to voice matching, the audio and visual clips need to be first *in motion* and then *temporally linked*, but audio or visual can be manipulated within these variables (Rosenblum, Yakel, Baseer, Panchal, Nodarse, & Niehus, 2002). We can match talking faces to speaking voices, but not static faces to speaking voices.

The cross-modal matching tasks extrapolate the way in which audio and visual modalities interact. Using pairs indicates that the cross-modal integration for speakers is a recognisable phenomenon, and that distractor in the pair does not diminish listener/viewers ability to match face to voice. The order of audiovisual priming (face or voice first) is critical, and it would appear that seeing a face before hearing a voice mimics naturalistic timing of speech and

viewer/listeners are more successful than listener/viewers in processing a speaker's identity (Maier, Di Luca, & Noppeney, 2011). Seeing the facial movement begin enables listeners to better process the speaker's identity and match it to sound. Crucially, face presentation first primes subsequent voice identification (Stevenage, Hugill, & Lewis, 2012). We are better at identifying faces than identifying voices. It seems we are more reliant on first seeing the face in order to match it to a voice, and are more susceptible to selecting face distractors and misidentifying the speaker when we try to match voice to face.

In musical contexts, the fusion of audio and visual information is also fundamental to identifying performers (Mitchell & MacDonald, 2014). Using the same design as the speaker identification studies, listener/viewers were asked to match the sight of a jazz saxophonist to his sound from a forced-choice pair (or vice versa from sound to sight). As in speaker studies, listeners were better at identifying a performer by sound after seeing him play.

Audiovisual matching is consistent across a number of studies when using forced-choice pairs. Visual cues prime audio identification, and seeing a speaker (or performer) better enables identification of speaker's voice (or performer's sound). The challenge for music researchers is to understand the strength of the audiovisual integration to identifying unique characteristics of performers. One way of achieving this is to make the experimental task more complex. Increasing the number of distractors challenges listener/viewers to perform the audiovisual matching task and confirms the importance of each sensory channel in performer identification. Discrete sound and sight information affects speaker identification and sound is more vulnerable to interference when the number of distractors increases (Stevenage, Neil, Barlow, Dyson, Eaton-Brown, & Parsons, 2013). Sound perception of an individual is more fragile and

susceptible to interference from distractors in comparison to visual perception. The question now is whether crossmodal performer recognition is affected by an increase in distractors, and whether sight or sound pathways are stronger in conveying unique information about an individual performer. The aim of this paper is to test the extent of visual priming in music performer identification.

Method

Ethics

The project was approved by the institutional Human Research Ethics Committee.

Participants

Thirty participants, known to the researchers via affiliations with the music faculty volunteered to take part. They were required to attend a single listening session lasting around 30 minutes.

Participants were aged between 20 and 67 (M = 29.5 years, SD = 10.1). All currently played one or more musical instruments/voice and had on average 14.5 years of training on that instrument. Twenty-three were currently enrolled in an undergraduate or postgraduate music degree and 19 currently taught a musical instrument.

Stimuli

Five professional male saxophonists who perform and teach in Sydney and who had studied at conservatoria in Australia volunteered to provide stimuli material for the project. They were informed that the purpose of the study was to investigate how listeners identify individual performers. Saxophonists were required to attend a single recording session.

Musical Tasks

The saxophonists gave an unaccompanied performance of jazz standard *Blue Bossa*. Saxophonists tuned to A = 440Hz and were set a consistent metronome pulse before playing.

Recording

The saxophonists performed in a recording studio, lined with sound absorbing curtains, and suitable for perceptual and acoustic recording and testing. The audio signal was captured using a matched pair of stereo microphones (Neumann KU100, ORTF configuration) 2.75m from the saxophonist. Levels were kept constant via a Millenia pre-amplifier with stepped controls (Millennia Media HV 3D-8 Microphone Pre-amplifier) so that the recordings of each saxophonist were comparable. The audio channels were digitised (Apogee AD-16X analogue to digital converter), and transferred to computer and saved as WAV files (24 bit, 48 kHz).

Saxophonists were video recorded against a black backdrop, and lit with two soft and focusable omni-lights (Lowel Rifa eX 44 kit) to ensure a lifelike depth of field and no shadows on the face, instrument or fingers. The video signal was captured via high-definition video camera (Sony PMW EX-3) onto removable memory card at broadcast quality, and white balanced to give depth of field contrast, colours and clarity for each musician. The camera operator used a preview monitor attached to the camera as a reference frame to position the participants at a constant height and distance in each shot. Video footage was ingested (SD Driver Macbook Pro) for storage and editing. Audio clips were prepared (as above) for a complete phrase of each song (c.8 seconds). Silent video clips of the same phrases were extracted in Final Cut Pro.

Experimental Design

Figure 1 shows the cross-modal matching task. The top row shows the visual-audio (V-A) order. The test clip contains the visual form of "Stan" playing *Blue Bossa*. Next, there are

Commented [p1]: When I looked at Figure 1 and saw that all of your saxophonists had the same first names as famous saxophonists, I was a bit confused and had to backtrack. (I.e., I thought maybe you had used found clips of the famous saxophonists instead of recruiting.) Don't want to be humourless about this - feel free to leave the names as is. Just wanted to point out. Maybe the added scare quotes can remove any confusion.

between two and five audio clips, one of which will be by "Stan" (target) and the up to four others randomised from the pool of saxophonists (distracters). Participants had to choose which clip matched the original performer in the test clip. The bottom row shows an example of the task carried out in the audio-visual (A-V) order. Presentation order of target and distracters was randomised across the 12 samples, and each participant saw three of each line-up (2, 3, 4 and 5). Figure 1 demonstrates the task for the 4-person line-up.

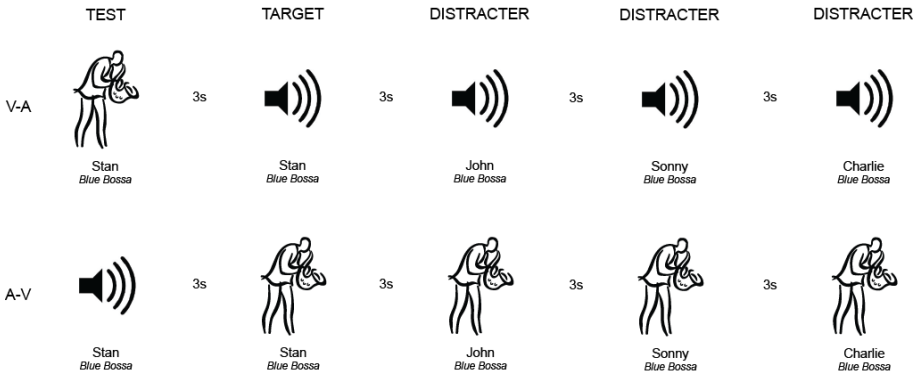

Figure 1: Schematic of the cross-modal matching task for the 4-person lineup. The top row shows the visual-audio (V-A) order. The test clip contains the visual form of Stan playing *Blue Bossa*. Next, there are between four audio clips, the first by Stan (target) and the second, third and fourth by other saxophonists (distracters). The bottom row shows the audio-visual order. Saxophonist pictures represent the silent video clips and the audio symbol represents the audio clip of the same musical phrase. There was a three-second pause between clips. Target/distracters clip presentation orders were randomised.

Video Preparation

Six videos were prepared in Final Cut Pro, three for V-A order and then V and A channels were swapped for the three in A-V order. Each included 12 cross-modal matching tasks. The test clips were introduced by sample number (1-12) and the target and distracter introduced by number. Participants knew from the feedback booklet how many clips to expect in each sample. There was a three-second fade to black between the test, target and distracter clips. Target/distracter clips and task orders were randomised in each video. In pilot testing, the three-second pause between clips was sufficient to prevent tapping being an aid to identification.

Procedure

The perceptual test was conducted in a quiet environment. Participants were seated in front of a MacBook Pro 15″ with circum-aural closed-back stereo monitoring headphones (Sennheiser HD 280).

Participants attended a single cross-modal perceptual test and were assigned to a V-A or A-V group. There was a short training period (three examples) before each participant's test session (12 samples). Finally, listeners were asked to rate their confidence in their responses and to explain how they made their choices. Results were coded numerically to either 0 or 1 for incorrect and correct identification.

Results

Participants identified the target saxophonist at an above chance level in line-ups of two (VA: $t(14) = 11.9$, $p < .001$, AV: $t(14) = 8.9$, $p < .001$), three (VA: $t(14) = 8.0$, $p < .001$, AV: $t(14) = 6.9$, $p < .001$), four (VA: $t(14) = 6.5$, $p < .001$, AV: $t(14) = 4.8$, $p < .001$) and five saxophonists (VA: $t(14) = 4.1$, $p < .005$, AV: $t(14) = 6.9$, $p < .005$). Table 1 shows mean identification rates for each line-up in each presentation order (VA and AV).

Table 1: Correct identification responses (Mean, SD and %) in each presentation order (VA or AV) and for each line-up length (2-, 3-, 4-, and 5-person line-up).

	VA			AV		
Line-up	Mean	(SD)	%	Mean	(SD)	%
Two	2.5	(0.6)	82%	2	(0.6)	67%
Three	1.9	(0.7)	62%	1.5	(0.6)	49%
Four	1.7	(0.9)	58%	1.2	(0.8)	40%
Five	1.8	(1.1)	60%	1.0	(0.8)	36%

The study design was a 4 (2, 3, 4 or 5 person line-up) x 2 (presentation order VA or AV) ANOVA with line-up length as the within-subject variable, and the presentation order (VA or AV) as the between-subjects factor. Figure 2 shows the estimated marginal means and main effects for different line-up lengths (2, 3, 4 and 5) in each presentation order (VA or AV). Mean identification rates decreased as line-up length increased. There was a significant main effect of line-up length, $F(3,26) = 8.28$, $MSE = 4.11$, $p < .001$, that is accuracy decreased with each addition to the line-up. Post-hoc Bonferroni tests showed that identification accuracy was

significantly higher in 2-person line-ups than 3-, 4- and 5-person line-ups ($p < .05$). All other comparisons were not significant. There was also a significant main effect of presentation order (VA or AV), $F(1,28) = 8.43$, $MSE = .2.13$, $p < .05$, that is, VA mean identification rates were significantly higher than AV rates. Independent samples *t* tests were performed to compare the mean accuracy of the VA and AV rates for each line-up and indicated that the difference between VA and AV accuracy rates was significant at the 5-person line-up ($p < .05$) and approached significance at the 2-person line-up ($p = .06$). The differences between VA and AV at the 3- and 4- person line-ups were not significant. There were no interaction effects for line-up and presentation order.

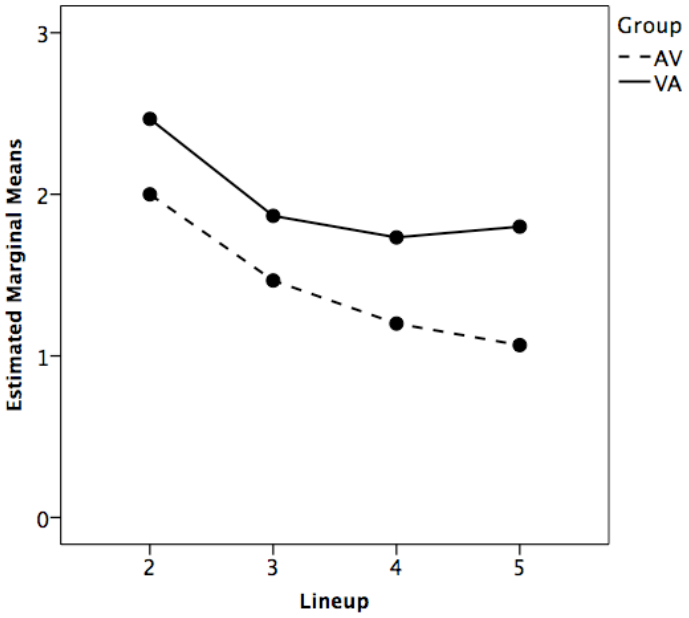

Figure 2: Main effect of line-up size in each presentation order (VA and AV), that is, between 2, 3, 4 and 5 person line-ups.

Neither group was very confident in their choices ($M = 4.7$, $SD = 1.8$). Participants in the V-A group were more confident ($M = 5.0$, $SD = 2.0$) than the A-V group ($M = 4.5$, $SD = 1.6$). Participants who saw the performer and then heard him were marginally more confident in their responses than those who heard and then saw the saxophonist but this difference was not statistically significant.

Perceptual processing

Listeners and viewers reported a variety of strategies to identify the targets from the line-ups. For both viewer (7 VA) and listener groups (7 AV), breathing, or breaths, seemed like an obvious place to differentiate between performers as it punctuated the playing they heard.

> *'Firstly I didn't know how to connect the video and the sound, after some examples I started to look at timing and breathing.' (VA5)*

> *'Placement of the breathing, finger movement indicating tempo.' (VA9)*

The search for breaths was not always successful, as most players, as expected, took breaths at the same time:

> *'Looking at how the players phrased (where/if they took breaths) it got confusing as a lot of people looked the same.' (AV24)*

For viewers in the VA order (12 VA), body movement was a critical concern in matching performer to sound. Some simply noted 'body movement' or 'sway', and some described the way in which they looked for tempo or timing through body movement to match with the audio. It quickly became apparent that the matching strategy was not effective for tempo or timing of the excerpt alone.

'Initially I was noting the fluidity of the performer's movement and how I perceived it to correlate with phrasing.' (VA13)

'I based my judgements on their hand movements, air blown from their mouths, and their body movements.' (AV28)

'I chose to focus on the movement and timing of the musicians, [but] matching this to the sound was more difficult than I thought it would be because they are quite similar.' (VA11)

Some viewers specifically linked body movement with the style of the performance, and the style of the audio in the samples in the audio line-up.

'Body parts - associated jauntier playing with more movement.' (VA2)

'Visual cues, not fingers, breathing after first phrase, panning from movement, or "static sound", feeling articulated body motion.' (VA8)

'Body language was also a factor, e.g., moving, if the sound example had more energy or still, if more calm.' (VA14)

Listeners (7 AV) shared similar ideas about the potential of watching body movements to monitor the visual tempo/timing of the excerpts. They tried to link their interpretation of the audio to their imagined vision of the performer:

'Looking at the body movement - if they were bopping to the rhythm in a way which sounded like the "groove" of the excerpt (55% of my decision making).' (AV30)

However, one listener was adamant that movement did not affect the way in which they picked the target from the line-up.

Body movement did not factor in trying to make a decision. (AV22)

Commented [p2]: From here on, there are no surrounding single quotes for participant comments. Can they be added here and below or removed above for consistency?

Both groups sought to match performer and performance by extra-musical means (7 VA, 4 AV). Viewers remembered the look of the performer and tried to match it to the sounds they heard.

Facial expression, keenness or engagement. Did same person play in same style? I thought so. (VA7)

I also then accounted for the tones in the samples and tried to match them with the saxophonist's disposition. (VA11)

Also, by making a judgement based on their appearance, even how old the saxophone appeared and what I gathered their personality to be, I matched those factors with the sound and style they played. (VA14)

Age of performers. Older players = more mellow, younger = brighter, punchier sound. Age of saxophones: older saxes = warm, mellow sound. (VA2)

Listeners worked in a similar fashion, trying to map the target sound to each of the visuals.

Whether the sound tended to suit the look of the player (AV19)

Mannerisms, the differences in their instruments, movements, slight variations in individual pieces felt 'different' in their colouration (synesthesia) which led me to focus on specific moments in the visuals that didn't sync up . . . (AV29)

Discussion

This task was designed to test the strength of visual priming in music performer identification. Listener/viewers were remarkably adept at matching both visual to audio and audio to visual performer presentations with increasing numbers of distractors, at rates above chance. Results confirmed visual cues provide more robust pathway and enhanced accuracy in performer identification regardless of the number of distractors. Viewers were more sensitive to the idiosyncrasies of a performer's sound when they had seen him play (V-A), and were more successful at identifying an audio clip from the line-up after seeing the saxophonist perform. This is an important finding as it highlights how sensitive listeners are to performers' visual cues.

Listener/viewers have already shown they are proficient at identifying target speakers, and indeed saxophonists, they see or hear from forced-choice pairs in the alternate sensory domain (Kamachi et al., 2003; Lachs & Pisoni, 2004; Mitchell & MacDonald, 2014). This novel line-up design indicated that while increasing the number of distractors affected listener/viewers' ability to identify individual performers, they could still consistently and reliably match audio and visual presentations of individual saxophonists. Separating audio and visual information seemed a curious and difficult task to these participants, but the results of this experiment follow results of not only speaker identification (Kamachi et al., 2003; Lachs & Pisoni, 2004), but also musical expressivity ratings (Vuoskoski et al., 2013) and musical judgments of performance quality (Tsay, 2013). In each case, *dynamic* and *temporally linked* sensory information achieves audiovisual integration.

Sound vulnerability contradicts conventional expectations in music, where sound is the primary focus for performer and listener. However, in this study visual priming was more important than audio priming to correctly identify the performer. While both audio and visual modalities provided sufficient information to achieve the task, visual provided more robust cues for successful identification. The relative timings of the action is crucial. In speech perception, speakers' faces move before the initiation of speech sound. As we see the sound preparation before we hear speech, it increases a speaker's intelligibility and likelihood of person identification (Maier et al., 2011). This may also explain why music performers' expressive body movements are more effective than audio in indicating expressivity and are sometimes all that is needed to recognise particular emotions (Dahl & Friberg, 2007; Vuoskoski et al., 2013). Here, the instigation, or 'sight of sound', also appeared to play a critical role in participants' ability to attribute sound to a particular saxophonist. It seems that visual cues are not restricted to global musical intentions, but more importantly reveal unique and identifying characteristics of individual performers' sound.

These participants gave some key insights into their perceptual processing. Listener/viewers were not accustomed to considering the relative impact of audio and visual channels on their evaluation of music, and despite their task success, listener/viewers were only moderately confident in their ability to match audio and visual information about saxophonists, which follows previous results in musical audiovisual matching (Mitchell & MacDonald, 2014). Audiovisual integration may be an instinctive phenomenon, but we are more familiar with considering one sense at a time, rather than the ways in which they might interact.

Many linked body movement to timing, or phrasing and were accustomed to tracking the microstructure of performers' metrical pulse. While all performers had established the same metronome pulse, the visual presentations, participants sought more global information as part of performers' style or body language. These results support universal non-verbal cues through visual means (Thompson, et al., 2005) and suggest that performers not only convey expressivity through their control of phrasing (Vines et al., 2006), but also transmit distinctive information that distinguishes them from other performers.

Both viewers and listeners were aware of their visual judgments, but viewers were more cognisant in conceptualising the cues they sought from the audio line-ups. This affirms the importance of the naturalistic of the action, and the importance of seeing the initiation of sound (Maier et al., 2011). The music domain is already aware that vision, more that sound, is required to transmit a performer's intent, and is in fact, more informative than the sound alone (Dahl & Friberg, 2007; Vuoskoski et al., 2013). The anticipation of performance, in the way a performer takes to the stage, already primes the audience to judge the performer (Platz & Kopiez, 2013), so it is a natural extension to suggest the preparation to play prepares the audience for a performer's sound. Participants gave shrewd analyses of their extra-musical judgments, and their gestalt impressions of the 'look' of a performer affected their interpretation of his sound. Visual evaluations included the saxophonists' commitment to the performance, his appearance, and his age. Music evaluators are influenced by the way in which a performer engages with them visually, regardless of the sound the produce (Wapnick et al., 1998) and make judgments by sight alone (Tsay, 2013) and these findings suggest that the 'look' of a performer prepares listeners for the sound to expect.

There are both intuitive and counterintuitive aspects to these results. In music, we believe sound forms the basis of our aesthetic and evaluative judgments, but we are already aware the vision plays a critical role in our reception of music performance. As such, audiovisual fusion is compelling, and this study confirmed that listeners could match performer identities by eye and ear. The effect of visual priming was robust, regardless of the number of distractors, and participants were better equipped to access unique details about individual performers' sound following sight. Results reflect the primacy of visual processing in the real world where it is more natural to see sound initiation, and sight effectively primes listeners for the sound they hear.

In music, sound is foremost in the minds of performer and listener but there is an untapped interplay between sound and sight when listeners perceive music performers. The role of visual information is critical to music performers, but not well understood by professionals and pedagogues who seek to create and hear a unique a distinctive sound. The idea that music listeners need multisensory cues to listen cannot simply be a scientific curiosity, and has profound implications for future music professionals. Music pedagogy must harness awareness of audiovisual processing, and visual priming, to effectively equip future generations of music professionals with knowledge and skills to capitalise on basic perceptual capacities, and also develop a more nuanced understanding of how multisensory processes shape music production to conceptualise, create and appraise a unique sound. Future studies will investigate how musicians can develop appropriate skillsets to be aware of and be able to access these multisensory capabilities in listening to music performers.

Acknowledgments

Sincere thanks the performers, listeners and listener/viewers who gave their time and insights in this study.

References

Broughton, M., & Stevens, C. (2009). Music, movement and marimba: an investigation of the role of movement and gesture in communicating musical expression to an audience. *Psychology of Music, 37*(2), 137–153. doi: 10.1177/0305735608094511

Chartrand, J.-P., & Belin, P. (2006). Superior voice timbre processing in musicians. *Neuroscience Letters, 405*(3), 164–167.

Dahl, S., & Friberg, A. (2007). Visual perception of expressiveness in musicians' body movements. *Music Perception, 24*(5), 433–454.

Davidson, J. W. (1993). Visual perception of performance manner in the movements of solo musicians. *Psychology of Music, 21*(2), 103–113. doi: 10.1177/030573569302100201

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). "Putting the face to the voice": Matching identity across modality. *Current Biology, 13*(19), 1709–1714.

Lachs, L., & Pisoni, D. B. (2004). Crossmodal source identification in speech perception. *Ecological Psychology, 16*(3), 159–187.

Maier, J., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance, 37*(1), 245–256.

Mitchell, H. F., & MacDonald, R. A. R. (2014). Listeners as spectators? Audio-visual integration improves music performer identification. *Psychology of Music, 42*(1), 112–127.

Platz, F., & Kopiez, R. (2013). When the first impression counts: Music performers, audience and the evaluation of stage entrance behaviour. *Musicae Scientiae, 17*(2), 167–197. doi: 10.1177/1029864913486369

Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition. *Perception & Psychophysics, 64*, 220–229.

Schweinberger, S. R., Robertson, D., & Kaufmann, J. M. (2007). Hearing facial identities. *The Quarterly Journal of Experimental Psychology, 60*(10), 1446–1456. doi: 10.1080/17470210601063589

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology, 24*(4), 409–419. doi: 10.1080/20445911.2011.642859

Stevenage, S. V., Neil, G. J., Barlow, J., Dyson, A., Eaton-Brown, C., & Parsons, B. (2013). The effect of distraction on face and voice recognition. *Psychological Research, 77*(2), 167–175. doi: 10.1007/s00426-012-0450-z

Thompson, M. R., & Luck, G. (2012). Exploring relationships between pianists' body movements, their expressive intentions, and structural elements of the music. *Musicae Scientiae, 16*(1), 19–40. doi: 10.1177/1029864911423457

Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica, 156*(1/4), 177–201.

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

Tsay, C.-J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences, 110*(36), 14580–14585. doi: 10.1073/pnas.1221454110

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition, 101*(1), 80–113.

Vuoskoski, J. K., Thompson, M. R., Clarke, E. F., & Spence, C. (2013). Crossmodal interactions in the perception of expressivity in musical performance. *Attention, Perception, & Psychophysics*, 1–14. doi: 10.3758/s13414-013-0582-2

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performance. *Journal of Research in Music Education, 45*(3), 470–479.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior and dress on violin performance evaluation. *Journal of Research in Music Education, 46*(4), 510–521.

Figure Legends

Figure 1: Schematic of the cross-modal matching task for the 4-person lineup. The top row shows the visual-audio (V-A) order. The test clip contains the visual form of Stan playing *Blue Bossa*. Next, there are between four audio clips, the first by Stan (target) and the second, third and fourth by other saxophonists (distracters). The bottom row shows the audio-visual order. Saxophonist pictures represent the silent video clips and the audio symbol represents the audio clip of the same musical phrase. There was a three-second pause between clips. Target/distracters clip presentation orders were randomised.

Figure 2: Main effect of line-up size in each presentation order (VA and AV), that is, between 2, 3, 4 and 5 person line-ups.