COGNITION AND EMOTION
2008, 00 (00), 1–14

Psychology Press
Taylor & Francis Group

Audio-visual integration of emotional cues in song

William Forde Thompson
Macquarie University, Sydney, New South Wales, Australia

Frank A. Russo
Ryerson University, Toronto, Ontario, Canada

Lena Quinto
Macquarie University, Sydney, New South Wales, Australia

We examined whether facial expressions of performers influence the emotional connotations of sung materials, and whether attention is implicated in audio-visual integration of affective cues. In Experiment 1, participants judged the emotional valence of audio-visual presentations of sung intervals. Performances were edited such that auditory and visual information conveyed congruent or incongruent affective connotations. In the single-task condition, participants judged the emotional connotation of sung intervals. In the dual-task condition, participants judged the emotional connotation of intervals while performing a *secondary* task. Judgements were influenced by melodic cues and facial expressions and the effects were undiminished by the secondary task. Experiment 2 involved identical conditions but participants were instructed to base judgements on auditory information alone. Again, facial expressions influenced judgements and the effect was undiminished by the secondary task. The results suggest that visual aspects of music performance are automatically and preattentively registered and integrated with auditory cues.

Although music is an auditory phenomenon, music experiences often include a visual dimension arising from the facial expressions and gestures of performers. Indeed, before the invention of the gramophone and phonograph, all musical performances were heard and seen by the audience. In many contemporary musical idioms such as opera, musical theatre, and rock concerts, visual information is integral to the experience. When visual aspects of performance are perceptually available, music becomes a

Correspondence should be addressed to: Bill Thompson, Department of Psychology, Macquarie University, Sydney, New South Wales, Australia, 2109.
E-mail: Bill.Thompson@psy.mq.edu.au

www.psypress.com/cogemotion DOI: 10.1080/02699930701813974

multimodal experience in which auditory and visual sources of information are perceptually integrated.

A number of researchers have observed visual influences on judgements of music. Saldaña and Rosenblum (1993) found that visual information influences whether a musical tone is heard as plucked or bowed. Thompson, Graham, and Russo (2005) reported that the facial expressions and gestures used by legendary performers like Judy Garland and B. B. King influenced judgements of the emotional connotation of the music. Facial expressions also convey information about relative pitch (Thompson & Russo, 2007). More generally, performance gestures can augment or attenuate the perceived emotional tension of performances (Vines, Krumhansl, Wanderley, & Levitin, 2006), and are reliable indicators of musical expressiveness in the absence of auditory information (Davidson, 1993).

It is also well established that visual information arising from speakers greatly affects speech perception (McGurk & McDonald, 1976). When observers are exposed to a video with the audio syllable "ba" dubbed onto a visual "ga" the perceptual illusion of hearing "da" occurs, implicating a perceptual process that registers and integrates auditory and visual signals. *Integration* refers to the process by which cues arising from different modalities are combined to generate a unified experience that reflects a balance of available signals.

In a review of the issue, de Gelder, Vroomen, and Pourtois (1999) suggest that the McGurk effect occurs regardless of whether observers are asked to selectively attend to auditory information only or whether they are distracted (e.g., by a secondary task). These findings suggest that visual and auditory speech signals are integrated automatically, without the need for attentional processes. The robustness of the McGurk effect has been attributed to the redundancy of auditory and visual information that is encountered in face-to-face verbal interactions. As such, perceptual integration may function to optimise verbal decoding in the face of changes in signal-to-noise ratio (SNR). In particular, when there are sudden drops in auditory SNR, audio-visual integration reduces the probability of perceptual gaps in the signal.

Facial expressions of speakers also influence judgements of the emotional connotations of speech (de Gelder & Vroomen, 2000; Hietanen, Leppänen, Illi, & Surakka, 2004). As with the perception of speech sounds, facial expressions influence the perception of emotional prosody even when observers are engaged in a secondary task, suggesting that auditory and visual signals of emotional prosody are integrated at a perceptual level. Vroomen, Driver, and de Gelder (2001) asked participants to provide emotion judgements of audio-visual presentations of speech either as the sole task (single-task condition) or while they performed a demanding secondary task (dual-task condition). The researchers reasoned that if attention is needed to integrate auditory and visual information, then

reducing the amount of attention available by introducing a secondary task should reduce the extent of integration and, hence, the effect of visual information on emotion judgements. The results showed that the effect of visual information on judgements of emotional prosody was the same for single-task and dual-task conditions, suggesting that auditory and visual signals of emotion during speech are registered and balanced automatically and preattentively.

The aim of the current study was to examine whether emotional connotations of sung materials are influenced not only by melodic cues, but also by facial expressions of singers, and whether attentional resources mediate audio-visual integration. Based on the results of Thompson et al. (2005), we predicted that participants would be influenced by both auditory and visual information when judging the emotional valence of sung stimuli.

We also predicted that audio-visual integration would occur automatically and preattentively. That is, audio-visual integration should be observed even when participants are required to perform an attentionally demanding secondary task (dual-task condition) or when they are instructed to ignore visual cues and focus on the sounded music. However, if integration of audio-visual information is based on a conscious assessment of available cues (requiring attentional resources), then the effects of visual information on judgements of emotion should be attenuated in the dual-task condition (where attentional resources are occupied) or when participants are instructed to selectively attend to auditory input.

In Experiment 1, participants rated the emotional valence of audio-visual recordings of sung intervals. Ratings were made under both single-task and dual-task conditions. In the single-task condition, participants could direct their full attention to the audio-visual presentation. In the dual-task condition, participants had to perform a demanding secondary task at each of two levels of difficulty during the audio-visual presentation. Experiment 2 involved the same conditions and procedures as Experiment 1, but participants were instructed to focus only on auditory information when judging emotional valence.

EXPERIMENT 1

Experiment 1 was designed to assess whether audio-visual recordings of sung intervals provide auditory and visual affective cues that are recognised and balanced by listeners to achieve a unified impression of emotional meaning. We adapted the dual-task paradigm used by Vroomen et al. (2001, Experiment 2) to evaluate whether or not audio-visual integration of emotion cues occurs preattentively. Stimuli were audio-visual recordings of

a vocalist as she sang either of two ascending melodic intervals: a major third or a minor third.

Major and minor third intervals were selected because they communicate contrasting emotional valence. Indeed, there is widespread agreement that the ascending major third sounds "joyful" or "happy" and the ascending minor third sounds "mournful" or "sad" (Backus, 1969; Cooke, 1959; Dalla Bella, Peretz, Rousseau, & Gosselin, 2001; Gagnon & Peretz, 2003; Huron, 2006; Kastner & Crowder, 1990). The emotional distinction may find its origin in differences in implied consonance (Kameoka & Kuriyagawa, 1969), tonal stability (Bharucha & Stoeckig, 1987), or statistical occurrence in music (Huron, 2006).

Method

A vocalist was recorded singing ascending major and minor thirds to the syllable *la*. Audio and visual channels of the original recordings were then recombined to yield audio-visual pairings that were either congruent or incongruent with respect to emotional valence. For congruent conditions, a sounded major-third interval was combined with the visual clip taken from a sung major-third interval, and a sounded minor-third interval was combined with the visual clip taken from a sung minor-third interval. For incongruent conditions, a sounded major-third interval was combined with the visual clip taken from a sung minor-third interval, and a sounded minor-third interval was combined with the visual clip taken from a sung major-third interval.

Participants judged the valence of the four different audio-visual pairings under different attentional load conditions. Attentional load was manipulated by introducing a secondary task that was irrelevant to the task of judging emotional valence. The secondary task involved paying close attention to a sequence of digits that were presented at slow and fast rates.[1]

Participants. Participants were 50 undergraduate students (15 men and 35 women, mean age = 19.9, range = 15–33) enrolled in introductory psychology at the University of Toronto at Mississauga. They had an average of 2.4 years of formal training in music (range = 0–18 years) and were given course credit for their participation. All reported normal hearing.

[1] Our manipulation of attention was adapted from the approach used by Vroomen et al. (2001). The latter study involved only three conditions (two single-task conditions and one dual-task condition) and numbers only appeared at one rate (five times per second). In one single-task condition, numbers appeared but participants ignored them. In the other single-task condition, no numbers appeared. Thus, the two studies manipulated attention in similar ways for the same purpose, but numbers always occurred in the current study and were presented at slow and fast rates.

Stimuli and conditions. Stimuli consisted of audio-visual recordings of a vocalist singing ascending major thirds (i.e., happy) and ascending minor thirds (i.e., sad), which displayed the singer's head and top of shoulders. Prior to the recording, the vocalist was instructed to sing expressively but was not asked to exaggerate the emotional message. Each recording was 7 seconds in length and the sung interval began after 3 seconds had elapsed. The two sung notes were roughly 1000 ms in length and were separated by a 1000 ms pause. Facial expressions used while singing the major and minor intervals were consistent with those used to communicate happy and sad emotions, respectively (Ekman & Friesen, 1978). Examination of the recordings revealed a fluid use of facial expressions between the onset of the first note and the offset of the second note. The sung interval was presented to participants through high-quality headphones at a comfortable listening level (approximately 68 dB SPL) and the video was displayed on an eMac computer screen.

Audio and visual recordings were synchronised in iMovie to create four conditions: (1) major-third audio with major-third visual; (2) major-third audio with minor-third visual; (3) minor-third audio with major-third visual; and (4) minor-third audio with minor-third visual. Thus, audio and visual channels were congruent in two of the conditions and incongruent in two of the conditions.

In all four audio-visual conditions, a sequence of translucent "ones" and "zeros" was superimposed over the singer's face. The numbers were presented at a rate of either 700 ms or 300 ms per number (slow and fast presentation rates). They appeared as soon as the video began (i.e., 3 seconds before the first note was sung) and continued until the end of the video (1 second after the second note was sung). Participants were tested under single-task and dual-task conditions. In the single-task condition, participants provided an emotion rating and ignored the presentation of digits. In the dual-task condition, participants were required to count the number of zeros as well as provide an emotion rating. The four combinations of single-task and dual-task conditions with slow and fast presentation rates were defined as a single factor labelled *task condition*. Each of the 16 combinations of audio-visual condition (four levels) and task condition (four levels) was presented six times: twice each with one, two, and three zeros presented at random serial positions in the sequence of digits.

To summarise, there were four factors, as follows:

1. Audio-visual condition (four levels).
2. Task condition (four levels).
3. Number of zeros (three levels).
4. Exemplar (two levels).

The four factors yielded $4 \times 4 \times 3 \times 2 = 96$ trials. Factors 1 and 2 were of experimental interest, whereas factors 3 and 4 were introduced to vary the number of zeros (to prevent guessing) and to increase power (two exemplars of each condition). The four task conditions were presented in blocks consisting of 24 trials each. The order of trials within blocks and the order of blocks were counterbalanced. The experiment took approximately 40 minutes.

Procedure. Participants were presented with a video recording of a sung interval and were instructed to rate the emotion conveyed by the singer using any cues that they perceived to be relevant to the task. In the dual-task condition, they were also required to count the number of zeros that flashed on the performer's face during the trial. Ratings were made on a five-point bipolar scale with poles labelled "*very sad*" (1) and "*very happy*" (5). Although the use of unipolar scales is often prudent, the assumption of bipolarity of valence is typically valid (Ilie & Thompson, 2006; Schimmack & Grob, 2000; Schimmack & Reisenzein, 2002). In the secondary task, participants indicated whether they detected one, two, or three zeros. Participants received four practice trials.

Results and discussion

Preliminary analysis confirmed that participants attended to the secondary task. Pilot work indicated that sustained attention was required to monitor the number of zeros that appeared on the screen, and participants performed the task with a high degree of accuracy ($M = 85.1$, $SD = 13.3$). Accuracy was significantly lower for fast presentation rates ($M = 83.0$, $SD = 14.3$) than for slow presentation rates ($M = 87.2$, $SD = 12.0$), $t(49) = 2.7$, $p < .05$, confirming that greater attentional resources were required to monitor faster presentation rates. As the purpose of the secondary task was to occupy attentional resources as much as possible, we also set a criterion of 80% correct for inclusion in the analysis. Twelve participants fell below the inclusion criterion, so our analyses were based on the remaining 38 participants.

As displayed in Figure 1, major-thirds paired with the matching facial expressions were assigned high ratings ("very happy"), whereas minor thirds paired with matching facial expressions were assigned low ratings ("not happy"). When audio-visual information was incongruent, emotion ratings were intermediate, reflecting a balance between the two sources of information.

Emotion ratings were entered into a two-way analysis of variance (ANOVA) with repeated-measures on *Audio-Visual Condition* (four combinations of happy and sad audio and video clips) and *Task Condition*

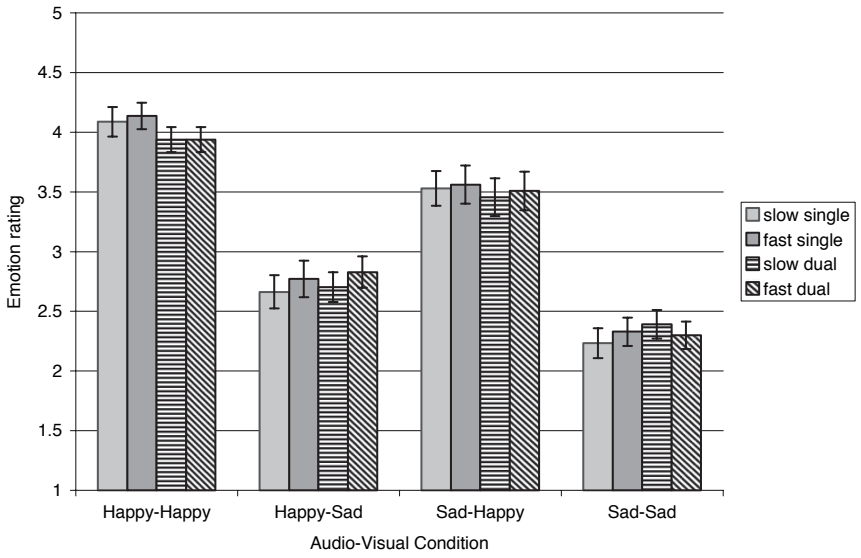

Figure 1. Mean rating of emotion (error bars are standard errors) across presentation conditions in each audio-visual condition.

(single-task slow; single-task fast; dual-task slow; dual-task fast). The score for each of the 16 combinations of the two independent variables was the average of six ratings (presentations involving one, two, or three zeros, each presented twice). The mean and median inter-subject correlation for these 16 conditions was 0.60 and 0.66, respectively (range -0.26 to 0.98, $p < .05$ in 73% of cases), indicating reasonably high agreement among subjects.

As predicted, there was a main effect of Audio-Visual Condition, $F(3, 111) = 76.49$, $p < .0001$, $\eta_p^2 = 0.67$. Planned comparisons confirmed the expected direction of influence for audio, visual, and audio-visual congruency manipulations. First, trials involving sounded major thirds were assigned higher ratings than trials involving sounded minor thirds, $F(1, 37) = 21.41$, $p < .0001$, $\eta_p^2 = 0.37$, confirming that the auditory signal influenced ratings of emotion. Second, trials involving facial expressions accompanying sung major thirds were assigned higher ratings than trials involving facial expressions accompanying sung minor thirds, $F(1, 37) = 132.40$, $p < .0001$, $\eta_p^2 = 0.78$, indicating that visual information influenced emotion ratings. Third, mean ratings for incongruent trials were not significantly different from mean ratings across happy and sad congruent trials, $F(1, 37) < 1$. That is, rating for incongruent audio-visual trials were intermediate between ratings for happy congruent trials and ratings for sad congruent trials, reflecting a balance between the two sources of information.

Inspection of Figure 1 suggests that among incongruent conditions, sad-audio/happy-video trials ($M = 3.51$, $SD = 0.71$) were assigned higher mean ratings than happy-audio/sad-video trials ($M = 2.74$, $SD = 0.55$). However, a post hoc analysis (Tukey) revealed that the difference did not reach statistically significance, $q = 1.88$, *ns*.

We found no evidence that audio-visual integration of emotion in music is influenced by attentional load. Specifically, the effect of *Task Condition* was not significant, $F(3, 111) = 1.29$, *ns*, and the interaction between *Task Condition* and *Audio-Visual Condition* was not significant, $F(9, 333) = 1.63$, *ns*. As seen in Figure 1, the pattern of judged emotion (happy audio, happy video > sad audio, happy video > happy audio, sad video > sad audio, sad video) was identical across the four task conditions. The results suggest that, as with emotion judgements for speech stimuli, audio-visual integration of emotional information in music occurs preattentively.[2]

EXPERIMENT 2

The results of Experiment 1 suggest that emotional interpretations of music are influenced by facial expressions of performers. This influence was undiminished by the introduction of a secondary task, suggesting that audio-visual integration occurs preattentively. Experiment 2 replicated the conditions of Experiment 1 but participants were specifically asked to assign judgements based on auditory information alone. If facial expressions still influence ratings of emotion in spite of these instructions, and the influence is still undiminished by the introduction of a secondary task, it would provide compelling evidence that audio-visual integration occurs automatically and preattentively.

[2] Two independent groups of participants with a similar distribution of music training provided emotion ratings for audio-only ($n = 21$) and visual-only ($n = 20$) presentations of the same stimuli. Sung major thirds presented as audio ($M = 3.96$, $SD = 0.50$) or video ($M = 4.21$, $SD = 0.74$) were assigned significantly higher ratings than minor thirds presented as audio ($M = 2.48$, $SD = 0.84$) or video ($M = 2.05$, $SD = 0.86$), $F(1, 39) = 179.28$, $p < .0001$. Mean ratings for incongruent audio-visual stimuli were intermediate between mean ratings for happy (major third) and sad (minor third) audio-alone and video-alone stimuli, confirming that visual and auditory cues both influenced emotion judgements. Ratings of happy audio-alone presentations were higher than ratings of happy-audio/sad-video presentations, $t(57) = 8.35$, $p < .0001$, and ratings of sad audio-alone presentations were lower than ratings for sad-audio/happy-video presentations, $t(57) = -4.99$, $p < .0001$. Similarly, ratings of happy visual-alone presentations were higher than ratings of happy-video/sad-audio presentations, $t(56) = 3.51$, $p < .001$, and ratings of sad video-alone presentations were lower than ratings of sad-video/happy-audio presentations, $t(57) = -3.70$, $p < .0001$. Thus, when presented with incongruent cues from two sensory modalities, participants did not rely on one channel as the basis of their rating, but integrated audio and visual signals to generate a rating that reflected a balance between the two cues.

Method

Participants. Participants were 29 undergraduate students (16 men and 13 women; mean age = 18.7; range = 18–22) enrolled in introductory psychology at the University of Toronto at Mississauga. Participants had an average of 2.5 years of formal training in music (range = 0–12 years). All were given course credit for their participation and all reported normal hearing. None participated in Experiment 1.

Stimuli and conditions. The stimuli and conditions were identical to those in Experiment 1 except that participants were always exposed to slow blocks before fast blocks (i.e., the rate at which numbers appeared).

Procedure. The procedure was identical to that used in Experiment 1 except that participants were instructed to make their judgements based on auditory information alone. Specifically, they were instructed to: "Ignore the facial expressions and make your rating based solely on what you hear".

Results and discussion

Accuracy in the secondary task was similar to that observed in Experiment 1 ($M = 85.9$, $SD = 10.4$) with nine participants falling below 80% correct. As the purpose of the secondary task was to occupy attentional resources maximally, these participants were removed from further consideration, and the analyses were based on the remaining 20 participants. The mean and median inter-subject correlation for these 16 conditions was 0.62 and 0.65, respectively (range -0.07 to 0.96, $p < .05$ in 74% of cases), which is comparable to the level of agreement observed in Experiment 1. Mean ratings of emotion across presentation and stimulus conditions were consistent with those observed in Experiment 1, and are displayed in Figure 2.

Emotion ratings were entered into a two-way ANOVA with repeated-measures on *Audio-Visual Condition* and *Task Condition*. As in Experiment 1, there was a main effect of audio-visual condition, $F(3, 57) = 44.59$, $p < .0001$, $\eta_p^2 = 0.70$. Planned comparisons confirmed the expected direction of influence for audio, visual, and audio-visual congruency manipulations. First, trials involving sounded major thirds were assigned higher ratings than trials involving sounded minor thirds, $F(1, 19) = 20.38$, $p < .001$, $\eta_p^2 = 0.52$. Second, even though participants were instructed to base judgements on auditory information only, trials involving facial expressions accompanying sung major thirds were assigned higher ratings than trials involving facial expressions accompanying sung minor thirds, $F(1, 19) = 64.03$, $p < .0001$, $\eta_p^2 = 0.77$. This remarkable finding provides strong

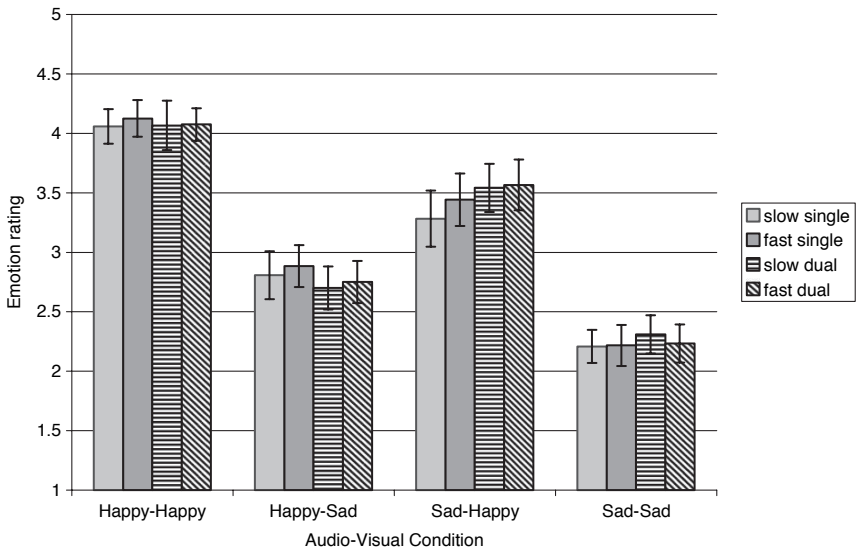

Figure 2. Mean rating of emotion (error bars are standard errors) across presentation conditions in each audio-visual condition with instruction to base judgement on audio information only.

evidence that auditory and visual cues to emotion are automatically registered and integrated with each other. Third, mean ratings for incongruent trials were not significantly different from mean ratings across happy and sad congruent trials, $F(1, 19) < 1$. That is, ratings for incongruent conditions were intermediate between ratings for happy and sad congruent trials. Post hoc (Tukey) analysis of incongruent conditions revealed that mean ratings for sad-audio/happy-video trials ($M = 3.45$, $SD = 0.71$) were not significantly higher than mean ratings for happy-audio/sad-video trials ($M = 2.78$, $SD = 0.56$), $q = 1.12$, *ns*.

As in Experiment 1, there was no significant main effect of *Task Condition* on ratings of emotion, $F(3, 57) < 1$, and no significant interactions involving *Task Condition*, Fs < 1.0. As may be seen in Figure 2, the pattern of judged emotion (happy audio, happy video > sad audio, happy video > happy audio, sad video > sad audio, sad video) did not vary across single-task and dual-task conditions. There were no other significant main effects or interactions. Finally, a three-way mixed-design ANOVA was conducted to compare the results of the two experiments. *Audio-Visual Condition* and *Task Condition* were within-subjects factors and *Experiment* was a between-subjects factor. There were no significant effects involving *Experiment* (all F values < 1.0).

GENERAL DISCUSSION

In two experiments, we found that facial expressions in performance influenced judgements of emotion in music. The finding is striking because facial expressions are extraneous to music and should be irrelevant to the task. The findings contribute to a growing body of research on the relation between music and emotion (see Juslin & Sloboda, 2001, for reviews) and the processes by which people decode emotional cues arising from music (Balkwill, Thompson, & Matsunaga, 2004; Ilie & Thompson, 2006). In particular, they illustrate that emotion perception for music performances is a multimodal phenomenon; one in which emotion cues from different perceptual systems are automatically and preattentively combined to form an integrated emotional interpretation.

An alternative explanation is that auditory and visual cues to emotion were processed separately and were consciously balanced with each other at the time of response. Although we observed no evidence for the involvement of conscious processes, two concerns about the experimental design could be raised. First, it is possible that our rating scale lacked the degree of sensitivity needed to detect subtle changes in audio-visual integration as a function of attentional resources, and a more sensitive scale might reveal effects of attentional load. This possibility warrants further investigation, but is an unlikely explanation of the results. The effects of audio-visual condition on ratings of emotion were extremely reliable, with highly significant differences occurring as a function of both audio and visual presentation conditions. If the rating scale lacked sensitivity, one would expect weak effects not only for manipulations of attentional load, but also for manipulations of audio-visual condition.

A second concern is that the secondary task might have been ineffective at depleting attentional resources. If attentional resources were not significantly reduced by the secondary task, then participants might have had the opportunity to consciously draw upon and integrate visual cues. We believe this interpretation is also unlikely. In both experiments, approximately one third of the sample failed to perform the counting task at the criterion level of 80% correct, suggesting the task was challenging. Only those participants who performed the counting task well were considered in the analysis. Moreover, when participants were asked to base their judgements on the auditory information alone, they were still influenced by visual information. If participants had used a conscious strategy to balance auditory and visual cues, then they should have been similarly capable of assigning less (or no) weight to visual information when instructed to do so. Instead, judgements of emotional meaning reflected visual cues, suggesting that audio-visual integration was not under conscious control.

Although research suggests that manipulations of either attentional or perceptual load can affect the processing of distractors (e.g., Lavie, 2005), emotional evaluations remained stable regardless of whether participants were engaged in a secondary task, or whether the task stimuli were presented slowly or quickly. A number of mechanisms could support automatic audio-visual integration of emotional cues. One possibility is that the integration depends on similarities between time-varying features in the audio and visual channel, as has been suggested for phonemic perception in speech (Calvert & Campbell, 2003; Summerfield, 1987). Cross-modal binding of time-varying features may be initiated because the sung melodic intervals involved temporally predictable auditory and visual signals (two notes of equal duration). A second possibility (which is not mutually exclusive from the first) is that seeing a vocal performance may initiate motor programmes that simulate the neural pattern of activity that would occur if the observer were performing. That is, the perception of visual (or audio) information about melody may invoke motor programmes that encompass auditory and visual aspects of the music. A number of proposals consistent with a motor theory of music perception have been suggested (Godoy, 2003; Lahav, Boulanger, Schlaug, & Saltzman, 2005; McAngus, Todd, O'Boyle, & Lee, 1999; Russo & Cuddy, 1999). Moreover, speech researchers have investigated and discussed motor theories for more than 50 years (Galantucci, Fowler, & Turvey, 2006; Lane, 1965; Liberman & Mattingly, 1985; Scott & Johnsrude, 2003).

The effect of facial expressions on judgements of affect was particularly remarkable in Experiment 2, where participants were instructed to base their judgements on the auditory information (i.e., to ignore visual information). One explanation is that observers are naturally more responsive to emotional connotations conveyed by visual cues than they are to emotional connotations conveyed by auditory cues. For example, sensitivity to happy and sad facial expressions may occur through innately determined processes, whereas emotional interpretations of major and minor intervals is thought to emerge through learning (e.g., Huron, 2006).

Recent speech perception research has revealed modulation of audio-visual integration by attention (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Tiippana, Andersen, & Sams, 2004). In contrast, the current data suggest that audio-visual integration of emotional cues in song is impervious to attentional demands. A number of theorists have suggested that music is evolutionarily older than speech (Darwin, 1871; Mithen, 2005) and there is little doubt that emotional processing is both adaptive and evolutionarily old. It is therefore possible that processing of emotion in music is more heavily reliant on primitive multisensory networks than is phonemic perception.

The findings add to an emerging literature demonstrating that visual aspects of music performance can powerfully affect our interpretations and

experience of the music. Visual effects on perception of music are observed not only for vocal music but also for instrumental music (Davidson, 1993; Thompson et al., 2005; Vines et al., 2006). In the course of a musical performance, an instrumentalist will often use body movements and facial expressions to anticipate and elaborate affective properties conveyed by the music. It remains to be clarified whether such ancillary aspects of performance are integrated in the same manner that has been demonstrated in the current study. Preattentive and automatic integration of facial expressions with sounded musical events may be unique to singing, where facial expressions are inextricably tied to sound production.

Manuscript received 13 February 2007
Revised manuscript received 4 September 2007
Manuscript accepted 14 November 2007
First published online day/month/year

REFERENCES

Alsius, A., Navarra, J., Campbell, R., & Soto-Faraco, S. (2005). Audiovisual integration of speech falters under high attention demands. *Current Biology*, *15*, 839–843.

Backus, J. (1969). *The acoustical foundations of music*. New York: W. W. Norton & Co.

Balkwill, L.-L., Thompson, W. F., & Matsunaga, R. (2004). Recognition of emotion in Japanese, Western, and Hindustani music by Japanese listeners. *Japanese Psychological Research*, *46*, 337–349.

Bharucha, J., & Stoeckig, K. (1987). Priming of chords: Spreading activation or overlapping frequency spectra? *Perception & Psychophysics*, *41*, 519–524.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57–70.

Cooke, D. (1959). *The language of music*. New York: Oxford University Press.

Dalla Bella, S., Peretz, I., Rousseau, L., & Gosselin, N. (2001). A developmental study of the affective value of tempo and mode in music. *Cognition*, *80*, B1–B10.

Darwin, C. (1871). *The descent of man*. New York: D. Appleton & Co.

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

de Gelder, B., & Vroomen, J. (2000). Perceiving emotions by ear and by eye. *Cognition and Emotion*, *14*, 289–311.

de Gelder, B., Vroomen, J., & Pourtois, G. (1999). Seeing cries and hearing smiles: Crossmodal perception of emotional expressions. In G. Aschersleben, T. Bachmann, & J. Musseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 425–437). Amsterdam: Elsevier Science.

Ekman, P., & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Palo Alto, CA: Consulting Psychologists Press.

Gagnon, L., & Peretz, I. (2003). Mode and tempo relative contributions to "happy–sad" judgements in equitone melodies. *Cognition and Emotion*, *17*, 25–40.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*, 361–377.

Godoy, R. I. (2003). Motor-mimetic music cognition. *Leonardo*, *36*, 317–319.

Hietanen, J. K., Leppänen, J. M., Illi, M., & Surakka, V. (2004). Evidence for the integration of audio-visual emotional information at the perceptual level of processing. *European Journal of Cognitive Psychology*, *16*, 769–790.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*, 319–329.

Juslin, P. N., & Sloboda, J. A. (Eds.). (2001). *Music and emotion: Theory and research*. Oxford, UK: Oxford University Press.

Kameoka, A., & Kuriyagawa, M. (1969). Consonance theory I. Consonance of dyads. *Journal of the Acoustical Society of America*, *45*, 1451–1459.

Kastner, M. P., & Crowder, R. G. (1990). Perception of the major/minor distinction: IV. Emotional connotations in young children. *Music Perception*, *8*, 189–202.

Lahav, A., Boulanger, A., Schlaug, G., & Saltzman, E. (2005). The power of listening: Auditory-motor interactions in musical training. *Annals of the New York Academy of Science*, *1060*, 189–194.

Lane, H. (1965). The motor theory of speech perception. A critical review. *Psychological Review*, *72*, 275–309.

Lavie, N. (2005). Distracted and confused? Selective attention under load. *Trends in Cognitive Sciences*, *9*, 75–82.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

McAngus Todd, N. P., O'Boyle, D. J., & Lee, C. S. (1999). A sensory-motor theory of rhythm, time perception and beat induction. *Journal of New Music Research*, *28*, 5–28.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Mithen, S. (2005). *The singing Neanderthals: The origins of music, language, mind and body*. New York: Weidenfeld & Nicholson.

Russo, F. A., & Cuddy, L. L. (1999). A common origin for vocal accuracy and melodic expectancy: Vocal constraints. *The Journal of the Acoustical Society of America*, *105*, 1217.

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

Schimmack, U., & Grob, A. (2000). Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, *14*, 325–345.

Schimmack, U., & Reisenzein, R. (2002). Experiencing activation: Energetic arousal and tense arousal are not mixtures of valence and activation. *Emotion*, *2*, 412–417.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organisation of speech perception. *Trends in Neurosciences*, *26*, 100–107.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audiovisual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lipreading* (pp. 3–51). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Tiippana, K., Andersen, T. S., & Sams, M. (2004). Visual attention modulates audiovisual speech perception. *European Journal of Cognitive Psychology*, *16*, 457–472.

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

Thompson, W. F., & Russo, F. A. (2007). Facing the music. *Psychological Science*, *18*, 756–757.

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.

Vroomen, J., Driver, J., & de Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective, & Behavioral Neuroscience*, *4*, 382–387.