sempre:
Society for Education, Music and Psychology Research

Article

Psychology of Music
2017, Vol. 45(2) 204–215

DOI: 10.1177/0305735616654216
journals.sagepub.com/home/pom
SAGE

Attention to affective audio-visual information: Comparison between musicians and non-musicians

Janne Weijkamp[1] and Makiko Sadakata[1,2,3,4]

Abstract

Individuals with more musical training repeatedly demonstrate enhanced auditory perception abilities. The current study examined how these enhanced auditory skills interact with attention to affective audio-visual stimuli. A total of 16 participants with more than 5 years of musical training (musician group) and 16 participants with less than 2 years of musical training (non-musician group) took part in a version of the audio-visual emotional Stroop test, using happy, neutral, and sad emotions. Participants were presented with congruent and incongruent combinations of face and voice stimuli while judging the emotion of either the face or the voice. As predicted, musicians were less susceptible to interference from visual information on auditory emotion judgments than non-musicians, as evidenced by musicians being more accurate when judging auditory emotions when presented with congruent and incongruent visual information. Musicians were also more accurate than non-musicians at identifying visual emotions when presented with concurrent auditory information. Thus, musicians were less influenced by congruent/incongruent information in a non-target modality compared to non-musicians. The results suggest that musical training influences audio-visual information processing.

Keywords

audio-visual, emotion, perception, speech, training

A large body of evidence suggests that musical training affects auditory information processing. Individuals who received musical training are more accurate at discriminating fundamental acoustic features, such as pitch in music and language (Besson, Schön, Moreno, Santos, & Magne, 2007; Fujioka, Ross, Kakigi, Pantev, & Trainor, 2006; Musacchia, Strait, & Kraus,

[1]Artificial Intelligence Department, Radboud University, the Netherlands
[2]Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University, the Netherlands
[3]Musicology department, University of Amsterdam, the Netherlands
[4]Institute for Logic, Language and Communication, University of Amsterdam, the Netherlands

Corresponding author:
Makiko Sadakata, Donders Institute, Radboud University, Montessorilaan 3, 6525HR, Nijmegen, the Netherlands.
Email: m.sadakata@donders.ru.nl

2008; Wong, Skoe, Russo, Dees, & Kraus, 2007), temporal information (Milovanov, Huotilainen, Esquef, Alku, Välimäki, & Tervaniami, 2010; Rammsayer & Altenmüller, 2006; Sadakata & Sekiyama, 2011), as well as higher level features, such as emotions in speech prosody (Lima & Castro, 2011). To put it simply, musical training contributes to the development of "good ears." In our everyday communication, however, auditory information is only one of multiple input streams, including visual and tactile channels, for instance. The current study addresses the influence of having such "good ears" in a multi-modal environment, and, more specifically, in perception of affective audio-visual information.

Previous studies have convincingly demonstrated that multi-modal sensory integration processes are not hard-wired but can vary across individuals. The combination of auditory and visual information can, for instance, be biased by a perceiver's cultural background, as is the case with the McGurk effect (McGurk & Macdonald, 1976), where the combination of two phonemes (separately projected as audio /ba/ and visual /ga/) results in the perception of yet another phoneme (/da/). Notably, the strength of this effect varies depending on the prior experience of individuals. For example, native speakers of Japanese demonstrate a weaker McGurk effect than native English-speakers: native Japanese-speakers appear to rely less on visual information than native English-speakers when perceiving phonemes (Burnham & Sekiyama, 2012; Sekiyama & Tohkura, 1991).

The current study addresses if such an effect of prior experience of individuals could be observed in higher level audio-visual processing. We thus address the possible influence of musical training on multi-modal perception. If one has an enhanced ability to perceive auditory information in the audio-visual context, it is possible that this influences how our perceptual system weighs and combines information from multiple sensory inputs. Previous research revealed that musicians are more accurate at judging audio-visual synchrony of piano performances than non-musicians (Lee & Noppeney, 2011). However, this study used musical materials as stimulus material and this may have contributed to such findings: it is therefore not surprising that participants with more musical training outperformed those with less musical training. The present study investigates whether such an advantage reported in individuals with more musical training also extends to non-musical audio-visual material. More specifically, we address how having "good ears" influences these individuals' processing of audio-visual information that is not related to a musical context, in the multisensory perception of emotion.

To address this question, we used a version of the Stroop task. The Stroop task presents participants with stimuli containing two concurrent features that either conflict or match with each other while having the participant focus on one source of information (Stroop, 1935). The task measures the degree of interference; that is, if attention to one feature can be affected by information from another feature. Although the original Stroop task was used to investigate two concurrent features in the visual modality (Stroop, 1935), it has been successfully extended to a multi-modal context (e.g., Tanaka et al., 2010). Previous studies revealed that, while trying to recognizing the emotional state of the target modality, it is not possible to ignore an irrelevant emotional state expressed in a non-target modality (e.g., Collignon et al., 2008; de Gelder & Vroomen, 2000; Donohue, Appelbaum, Park, Roberts, & Woldorff, 2013, Donohue, Todisco, & Woldorff, 2013; Jeong et al., 2011; Thompson, Russo, & Quinto, 2008).

Additionally, previous studies (e.g., Collignon et al., 2008) suggest that the interaction of auditory and visual information perception varies depending on, e.g., stimulus type used (e.g., music and speech) and the nature of the emotion presentation (clear and ambiguous). Generally, visual information seems to serve as the primary modality used, while the auditory has a secondary importance modality in the recognition of unambiguous emotional speech (de Gelder & Vroomen, 2000; Donohue, Appelbaum, et al., 2013).

Happy, sad, and neutral facial expressions were presented simultaneously with voice tones (humming), creating congruent and incongruent combinations. Participants were instructed to focus on the emotional state in one of the two modalities, and ignore the other, while their perceptual accuracy of the stimuli was measured. With this setup, we asked whether the degree of interference from concurrent visual information on auditory judgment varies between musicians and non-musicians. According to previous findings showing that musicians have better skills for processing auditory information, we expected musicians to be more robust and reliable in perceiving auditory information than non-musicians. Furthermore, we expected to see less of a visual interference effect on auditory tasks in the audio-visual condition for musicians than non-musicians. It is possible to see the opposite interaction effect: that the degree of interference from auditory information on visual judgment varies between groups. On the one hand, enhanced auditory skills may increase the significance of auditory information inputs as compared to that of visual information inputs, and, therefore, musicians may show a larger auditory interference effect than non-musicians on visual tasks in the audio-visual condition. On the other hand, musicians may be better at audio-visual integration and, if so, they may show a smaller auditory interference effect on these tasks.

Methods

Participants

Two groups, each with 16 individuals, took part in the current study. Participants in the musician group had more than 5 years of formal musical lessons and were actively practicing playing their instrument(s) for more than 2.5 hours per week at the time of the experiment (eight males, mean age = 29.31, *SD* = 11.02). Singers were not included. Participants in the non-musician group had less than 2 years of formal musical lessons and had not been practicing playing any instrument(s) for at least the last 2 years (eight males, mean age = 21.81 *SD* = 1.47).[1]

Separate advertisements for musicians and non-musicians were created so that the aim of the study (comparing perceptual skills of musicians and non-musicians) was not disclosed to participants prior to the experiment. Participants were recruited via acquaintances as well as via the participant recruitment system at Radboud University Nijmegen. Among 16 musicians, seven belonged to an amateur orchestra in the Netherlands, three were students of Hogeschool Arnhem Nijmegen. They participated in the experiment without any compensation. The rest of the participants (both musicians and non-musicians) were students of Radboud University Nijmegen and some of them received course credits for their participation.

Visual stimuli

Three male and female volunteers were told to pose with sad, neutral, and happy faces for the visual stimuli creation. We first created six sets of three facial expressions in total (18 images). Using a 9-point Likert scale (with 9 being happy, 5 being neutral, 1 being sad), 6 pilot participants rated perceived emotions of the faces. Among the six volunteers, we selected images of four volunteers that best reflected the target emotion. Figure 1 illustrates 12 images of visual stimuli (black-and-white) by the four volunteers. The experimental stimuli consisted of two male and two female faces with three different emotional states each (happy/neutral/sad). The practice stimuli consisted of another set of one male and one female face with the same three states.

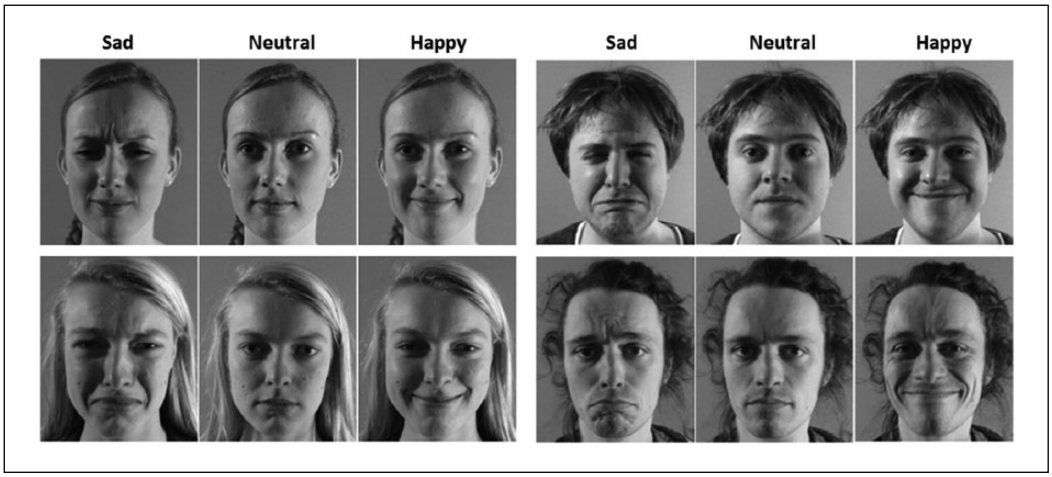

Figure 1. Sad (left), neutral (middle), and happy (right) images of four faces used for the experiment.

Auditory stimuli

Humming voices with sad, happy, and neutral emotional states were recorded as stimuli. Three male and female volunteers were instructed to express the target emotions by humming without opening their mouth. Humming with closed mouth instead of, for example, speaking or singing, was used here to increase the matching impression of voice and face because the visual stimuli included faces with closed mouth. As with the visual stimuli, using the same 9-point Likert scale (9 = happy, 5 = neutral, 1 = sad), six pilot participants rated perceived emotions of the voices. Among the six volunteers, we selected voices of four volunteers that best reflected the target emotion. The final stimuli set consisted of two male and female voices each, resulting in 12 stimuli (mean duration = 578 ms, *SD* = 182 ms). The average intensity level of stimuli was normalized to about 70dB. Figure 2 illustrates the pitch and intensity contour of these stimuli. The practice stimuli consisted of another set of one male and one female voice. The pitch (F0) and intensity contours for the neutral emotion are flat compared to the happy and sad emotions, reflecting that there is less emotional information in these voices. Pitch contour and intensity patterns were extracted using Praat (version 5.3.61).

Apparatus

A linear PCM recorder (Sony PCM–D1) was used to record all audio stimuli at a sampling rate of 96 kHz. A Compaq notebook computer with an Intel Pentium Dual Core processer (1 GB RAM) with 15 inch LCD screen and an Apple MacBook pro with IntelCoreDuo processor (4GB RAM) with a 15-inch TFT screen were used to perform the experiment. For the presentation of auditory stimuli, Sony MDR-7506 headphones with an average sound pressure level of approximately 70 dB SPL were employed. The experimental task was programmed with PsychoPy (version 14.3), used for both presenting instructions and stimuli as well as collecting behavioral responses. The keyboards of the computers were used to register responses of participants.

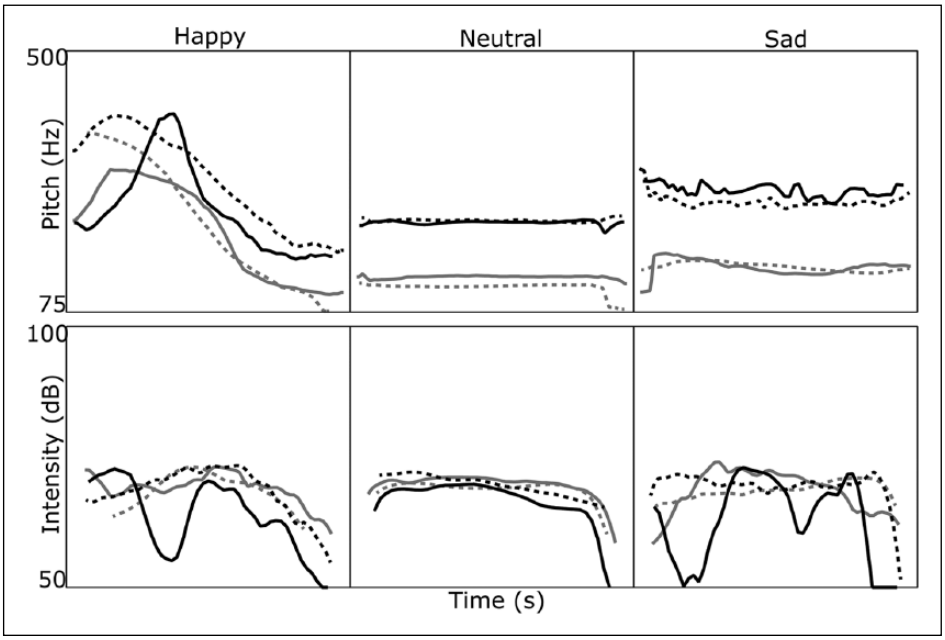

Figure 2. The black lines (solid and dotted) are from the female voices, and the grey lines (solid and dotted) are from the male voices.

Task and procedure

The experiment consisted of four tasks: a visual-only task (V task), an audio-only task (A task), and two audio-visual tasks with their focus on either visual information (AV-V task) or auditory information (AV-A task). For all participants, the AV-V and AV-A tasks occurred first (the order of these two counterbalanced) followed by the V task and the A task (the order of these two counterbalanced: for the half of the participants, the V task took place first, and for the other half, A task took place first). Presentation order of stimuli within each task was randomized. Correct response rates and reaction times from the onset of each stimulus were measured. All three emotions (happy, neutral, and sad) were used as response options and they were counted as correct when matched with the emotion of the target modality. Participants were instructed to respond as accurately and rapidly as possible. There was a practice session of AV-V and AV-A tasks using a practice stimulus set.

AV-V and AV-A task. Both audio-visual tasks were initiated by a fixation cross appearing for 1,500 ms followed by one of the 12 faces for 600 ms simultaneously with one of the 12 voices. All possible combinations of visual and auditory stimuli were presented with respect to gender: 3 auditory emotions (happy, sad, neutral) × 3 visual emotions (happy, sad, neutral) × 2 faces × 2 voices leading to 36 female and 36 male trials, resulting in 72 trials. Participants were instructed to report the emotional state of either face (AV-V) or voice (AV-A) using three buttons: happy, neutral, and sad. Each of the two tasks lasted approximately 8 minutes. In order to make sure that participants paid attention to the visual stimuli,

Figure 3. Two positions of face presentation.

Table 1. Overall frequency of responses for the four tasks (%).

		Target response				Target response		
Given response	***A-task***	Happy	Neutral	Sad	***V-task***	Happy	Neutral	Sad
	Happy	91.4	1.6	1.2	Happy	90.2	2.0	2.7
	Neutral	7.8	92.6	7.4	Neutral	6.3	95.3	10.2
	Sad	0.8	5.9	91.4	Sad	3.5	2.7	87.1
Given response	***AV-A task***	Happy	Neutral	Sad	***AV-V task***	Happy	Neutral	Sad
	Happy	88.3	2.0	2.5	Happy	81.1	5.5	5.0
	Neutral	8.9	93.6	10.2	Neutral	15.8	84.8	14.0
	Sad	2.9	4.5	87.3	Sad	3.1	9.7	81.0

especially during the AV-A task, the vertical position of the fixation cross and faces were varied across trials (Figure 3). Participants were explicitly instructed to watch the faces and to keep their eyes open.

V-task. A fixation cross appeared for 1,500 ms followed by one of the 12 faces presented at the same location for 600 ms. Participants were instructed to report the emotional state of the presented face by pressing one of three buttons on a keyboard: happy, neutral, and sad. All faces were presented two times (24 trials) resulting in approximately 2 minutes.

A-task. One of the 12 voices was presented per trial, while a fixation cross was presented on the screen. Participants were instructed to report the emotional state of a voice using buttons. All voices were presented two times (24 trials), resulting in approximately 2 minutes.

Results

Overall response patterns

Table 1 presents a summary of response rates. In general, participants tended to judge target emotions correctly in all tasks, with most mistakes resulting from the confusion between sad and neutral or happy and neutral responses. In other words, the confusion between sad and happy did not happen often.

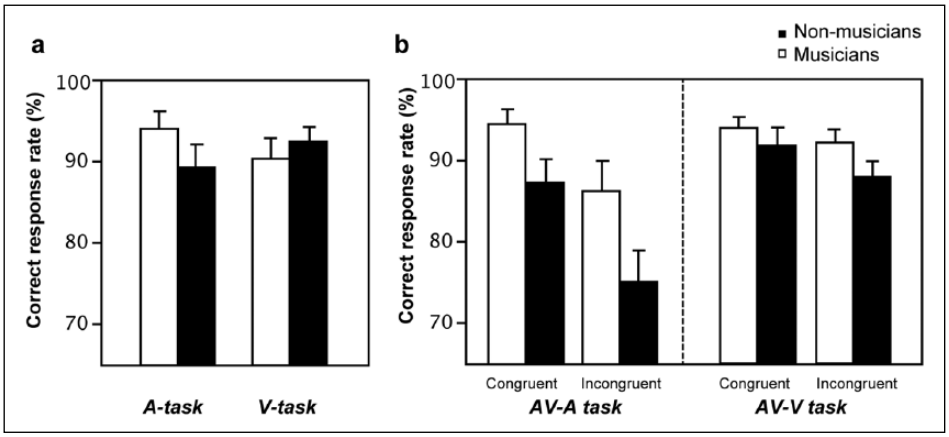

Figure 4. Mean correct response rates and cost measures.
a) Auditory and Visual only tasks (A-/V-tasks).
b) Audio-visual tasks with focus on Auditory modality (AV-A) and Visual modality (AV-V). The error bars indicate the standard error.

Correct response rates

Figure 4(a) presents mean correct response rates for the auditory-only (A task) and the visual-only task (V task). A two-way mixed model ANOVA was performed with Musical training (musician, non-musician) as between-subject factor and Modality (audio, visual) as within-subject factor on correct response rates. Neither Musical training nor Modality resulted in significant main effects, Musical training, $F(1, 30) = 0.283$, ns; Modality, $F(1, 30) = 0.029$, ns. This suggests that musicians and non-musicians performed equally well on both the visual-only and the auditory-only task with no difference between the two tasks.

Figure 4(b) presents the mean correct response rates of the AV-A task and AV-V task. A three-way mixed model ANOVA was performed on the correct response rate with Musical training (musician/non-musician) as between-subjects factor and Condition (congruent/incongruent) and Modality (auditory/visual) as within-subjects factors. All factors indicated significant main effects, Musical training: $F(1, 30) = 7.11$, $p < .05$, $\eta_p^2 = .192$; Modality: $F(1, 30) = 14.09$, $p < .001$, $\eta_p^2 = .320$; Condition: $F(1, 30) = 34.59$, $p < .001$, $\eta_p^2 = .536$. No interaction effect with regard to musical training was found indicating that the correct response rates of musicians were overall higher (91.6% $SE = 1.6$) than that of non-musicians (85.5%, $SE = 1.6$).

A significant interaction effect between Condition and Modality was found. A simple-effect test further confirmed that the performance accuracy on the visual task was significantly higher than that of audio task in the incongruent condition ($p < .001$), but not in the congruent condition. This suggests that attending to auditory information while being presented with concurrent incongruent visual information was more difficult than the other AV conditions.

The effect of additional information (congruent/incongruent) on each modality (auditory/visual) was analysed by subtracting the correct response rate of the AV-A task from that of the A task, and that of the AV-V task from the V task (hereafter referred to as cost measure). A positive cost measure indicates interference between target and non-target modality present in the task resulting in a decreased accuracy for identifying emotions. A negative cost measure can be interpreted as a facilitation effect, that is, the accuracy in identifying emotions increased

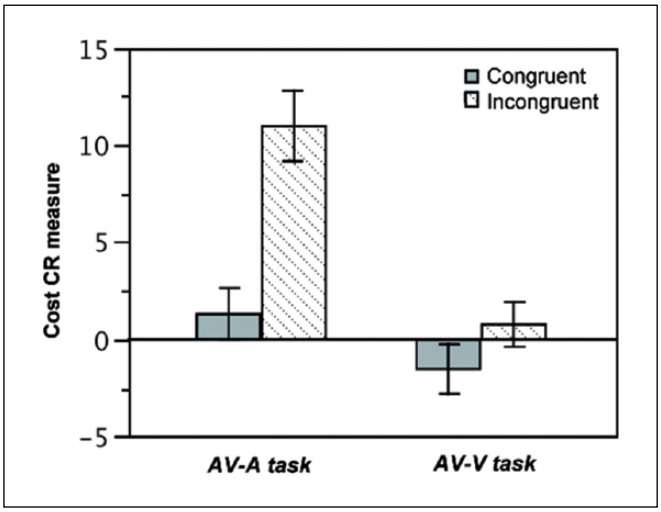

Figure 5. Cost measure for the audio-visual tasks with focus on Auditory modality (AV-A) and Visual modality (AV-V). The error bars indicate standard errors.

because of the additional information presented in the non-target modality. A three-way mixed model ANOVA was performed on the cost measure with Musical training (musician/non-musician) as between-subjects factor and Condition (congruent/incongruent) and Modality (auditory/visual) as within-subjects factors. All factors indicated significant main effects: Musical training: $F(1,30) = 8.42$, $p < .01$, $\eta_p^2 = .219$; Modality: $F(1, 30) = 15.14$, $p < .001$, $\eta_p^2 = .335$; Condition: $F(1, 30) = 34.59$, $p < .001$, $\eta_p^2 = .536$). Only one significant interaction effect was found between Modality and Condition, $F(1, 30) = 13.49$, $p < .001$, $\eta_p^2 = .310$. No interaction effect with regard to musical training was significant. The overall mean cost measure for musicians was −2.7 (standard error = 1.5) and that for non-musicians was 2.5 ($SE = 1.7$), indicating that musicians demonstrated a robust tendency towards a lower cost measure for both congruent and incongruent conditions. Notably, this was the case not only for the AV-A task but also for the AV-V task. This suggests that emotional judgment was less influenced by concurrent information in a non-target modality in musicians as compared to on-musicians.

The interaction effect between Modality and Condition is illustrated in Figure 5. Simple effect analyses revealed that the cost measure was significantly greater for the incongruent than the congruent condition, indicating a stronger interference effect for the incongruent information. Furthermore, the difference between the incongruent and congruent condition was more pronounced for the AV-A task than AV-V task: the performance accuracy of auditory tasks was more susceptible for concurrent incongruent visual information.

Reaction time

Figure 6(a) presents mean reaction times for the A- and V-task. A two-way mixed model ANOVA was performed with Musical training (musician/non-musician) as between-subject factor and Modality (auditory/visual) as within-subject factor. Reaction time served as dependent variable. There was no significant main effect of Musical training, $F(1, 30) = 0.271$, ns, whereas a significant main effect of Modality revealed shorter reaction times for the V-task than the A-task, $F(1, 30) = 146.35$, $p < .001$, $\eta_p^2 = .830$. However, this could be due to the different

nature of stimulus presentation in V- and A-tasks. The RT was measured as the stimuli onset to the button press time. While the presentation of visual stimuli was instantaneous, that of auditory stimuli took 578 ms on average. Nevertheless, an important point to note is that musicians and non-musicians did not differ significantly with respect to their response time for these tasks.

The effect of additional information (congruent/incongruent) in the second modality (auditory/visual) on reaction time was analysed by subtracting the reaction time of the AV-A task from that of the A task, as well as that of the AV-V task from that of the V task (cost RT measure). Figure 6(b) presents the mean reaction time of the auditory task (AV-A task) and that of the visual task (AV-V task), and Figure 6(c) presents the cost RT measure for the congruent and incongruent conditions. A three-way mixed model ANOVA was performed with Musical training (musician/non-musician) as between-subject factor and Condition (congruent/incongruent) and Modality (auditory/visual) as within-subject factors on cost RT measure. There were trends of main effects of Modality and Condition: modality, $F(1, 30) = 4.09$, $p = .052$, $\eta_p^2 = .120$; condition, $F(1, 30) = 32.19$, $p < .0001$, $\eta_p^2 = .518$, with significant interactions between them, $F(1, 30) = 12.31$, $p < .001$, $\eta_p^2 = .291$. No main effect or significant interaction related to Musical training was present, $F(1, 30) = .86$, ns. Further simple effect tests revealed that incongruent information led to a larger effect both when judging visual and auditory emotion, with the interference effect being stronger for auditory than visual emotion.

Discussion

Using an audio-visual Stroop task with emotion judgment, the current study tested whether the degree of interference of concurrent visual information on auditory judgment varies between musicians (defined here as individuals with more than 5 years of training) and non-musicians (defined here as individuals with fewer than 2 years of training).

In general, in accordance with previous studies indicating that visual information serves as the primary modality and auditory as the second modality (de Gelder & Vroomen, 2000; Donohue, Appelbaum, et al., 2013), our results indicate a greater interference effect of visual information on auditory processing than the other way around. Furthermore, as expected, musicians were more accurate than non-musicians in identifying auditory emotions when presented with concurrent (both congruent and incongruent) visual information. This enhancing effect is in proportion to a large body of research showing the enhanced ability of musicians to deal with auditory inputs (e.g., Besson et al., 2007; Fujioka et al., 2006; Lima & Castro, 2011; Sadakata & Sekiyama, 2011; Thompson et al., 2008; Musacchia et al., 2008). The current study demonstrates that the earlier proven benefit that musicians have in processing audio-visual information when presented with musical stimuli (Lee & Noppeney, 2011) extends to a more general context in the multisensory perception of emotion. Additionally, the present results showing differences between musicians and non-musicians again support that the process of multi-modal sensory integration is not hard-wired but flexibly adapted to one's environment.

Furthermore, musicians were also more accurate than non-musicians in identifying visual emotions when presented with concurrent auditory information. Differently than predicted, musicians did not show a larger auditory interference effect than non-musicians. It seems that musicians' enhanced ability to deal with auditory information did not distract them from focusing on the visual information. There are at least three possible different explanations to account for such a fact. Firstly, musical training may indeed enhance sub-processes of audio-visual processing, such as visual attention (Rodrigues, Loureiro, & Caramelli, 2013) and/or integration (Lee & Noppeney, 2011). In particular, Rodrigues' report on musicians' enhanced ability to

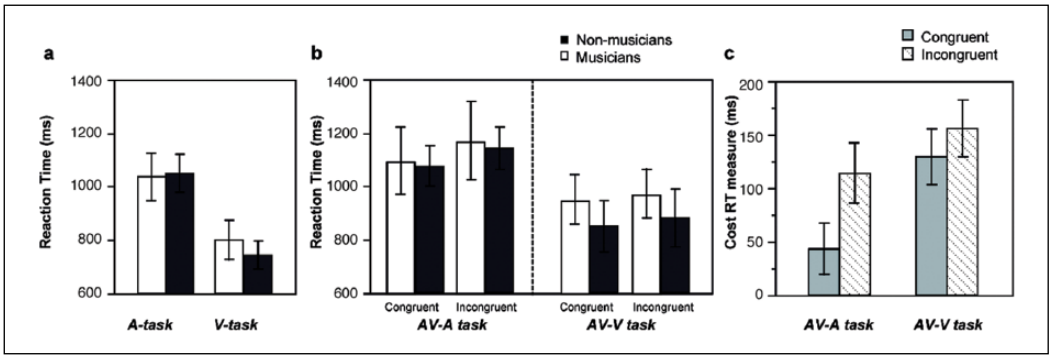

Figure 6. Mean reaction times.
a) Auditory and Visual only tasks (A-/V-tasks).
b) Audio-visual tasks with focus on Auditory modality (AV-A) and Visual modality (AV-V). The error bars indicate standard errors.
c) Cost RT measures for Audio-visual tasks with focus on Auditory modality (AV-A) and Visual modality (AV-V). The error bars indicate standard errors.

selectively attend to a visual object may be beneficial to the present task. Secondly, musical training may not enhance skills related to visual perception per se (as suggested by Strait et al., 2010), but rather support efficient encoding of auditory information, which results in better processing of concurrent information inputs from another sensory modality. For example, musical training may help with processing auditory information with smaller cognitive resources (Chandler & Sweller, 1991), leaving more cognitive resources available for the processing the concurrent visual information. Finally, musical training may help in not only attending to but also ignoring auditory inputs, which would predict similar results as shown in the present experiment. Future studies will need to address which of the possible explanations may account for such findings.

Surprisingly, the two groups did not differ with regard to their performances for unimodal tasks (audio and visual only conditions). In line with previous studies (Lima & Castro, 2011; Rodrigues et al., 2013), we expected musicians to outperform non-musicians at least in the audio-only task. However, previous studies used more challenging tasks than ours, such as judging an intended emotion of full speech sentences with more options of emotional states (Lima & Castro, 2011) and visual attention tasks that are more subject to individual differences (Rodrigues et al., 2013). One explanation for not finding unimodal group differences here is a possible ceiling effect: the task was too easy for both groups. The true facilitation effect of musical training might therefore only become evident when the task difficulty increases with audio-visual conditions.

Focusing on emotional states of one modality and ignoring the other is not a common task in our everyday communication. However, an incongruent combination of multisensory information does occur outside an experimental context. For example, while one of the audio-visual markers for "irony/sarcasm" is a combination of a blank face and exaggerated pitch patterns, these two are not inevitably congruent in their emotional characters (Attardo, Eisterhold, Hay, & Poggi, 2003). In other words, although there are incongruent combinations of multisensory inputs around us, we do not perceive them as such: all multisensory inputs tend to be integrated to induce a uniform percept. With this in mind, it is questionable to what extent the presented effect of musicians' enhanced ability to selectively attend to auditory (or visual) input is

supportive in their daily communication. However, having access to a more accurate input stream should serve as a good basis for inferring the most probable emotional state and musicians may show a benefit in such conditions. Further study investigating identification of more complex emotion expression and the interaction with individual skills in auditory/visual perception would be a promising next step.

Acknowledgements

The authors are grateful to Ida Sprinkhuizen-Kuyper, Christian Hoffmann, Jana Krutwig and M. Paula Roncaglia-Denissen for their comments.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Note

1. The difficulty in comparing musicians and non-musicians is to match the group profile completely. In the current study, although we tried our best to match the groups and testing environments as much as possible, the average age of musicians was higher than that of non-musicians. Although we think that the age difference is not very likely to explain the better performance (e.g., Lima & Castro, 2011), we suggest that future studies control for such group differences or explicitly incorporate it into the experimental design.

References

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, *16*(2), 243–260.

Besson, M., Schön, D., Moreno, S., Santos, A., & Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorative Neurology and Neuroscience*, *25*, 399–410.

Burnham, D., & Sekiyama, K. (2012). Investigating auditory-visual speech perception development. In G. Bailly, P. Perrier, & E. Vatikiotis-Bateson (Eds.), *Audiovisual speech processing* (pp. 62–75). Cambridge, UK: Cambridge University Press.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*, 293–332.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126–135.

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

Donohue, S. E., Appelbaum, L. G., Park, C. J., Roberts, K. C., & Woldorff, M. G. (2013). Cross-modal stimulus conflict: The behavioral effects of stimulus input timing in a visual-auditory Stroop task. *PLOS ONE*, *8*(4), 1–13.

Donohue, S. E., Todisco, A. E., & Woldorff, M. G. (2013). The rapid distraction of attentional resources toward the source of incongruent stimulus input during multisensory conflict. *Journal of Cognitive Neuroscience*, *25*(4), 623–635.

Fujioka, T., Ross, B., Kakigi, R., Pantev, C., & Trainor, L. J. (2006) One year of musical training affects development of auditory cortical-evoked fields in young children. *Brain*, *129*, 2593–2608.

Jeong, J., Diwadkar, V. A., Chugani, C.D., Sinsoongsud, P., Muzik, O., Behen, M. E., . . .Chigani, D. C. (2011). Congruence of happy and sad emotion in music and faces modifies cortical audiovisual activation. *NeuroImage*, *54*, 2973–2982.

Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences*, *108*(51), E1441–E1450.

Lima, C. F., & Castro, S. L. (2011). Speaking to the trained ear: Musical expertise enhances the recognition of emotions in speech prosody. *Emotion*, *11*, 1021–1031.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Milovanov, R., Huotilainen, M., Esquef, P.A., Alku, P., Välimäki, V., & Tervaniemi, M. (2010). The role of musical aptitude and language skills in preattentive duration processing in school-aged children. *Neuroscience Letters*, *460*(2), 161–165.

Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hearing Research*, *241*(1), 34–42.

Rammsayer, T., & Altenmüller, E. (2006). Temporal information processing in musicians and nonmusicians. *Music Perception*, *24*, 37–47.

Rodrigues, A. C., Loureiro, M. A., & Caramelli, P. (2013). Long-term musical training may improve different forms of visual attention ability. *Brain and Cognition*, *82*(3), 229–235.

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica*, *138*, 1–10.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, *90*, 1797–1805.

Strait, D. L., Kraus, N., Parbery-Clark, A., & Ashley, R. (2010). Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance. *Hearing Research*, *261*, 22–29.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Tanaka, A., Koizumi, A., Imai, H., Hiramatsu, S., Hiramoto, E., & de Gelder, B. (2010). I feel your voice: Cultural differences in the multisensory perception of emotion. *Psychological Science*, *21*(9), 1259–1262.

Thompson, W. F., Russo, F. A., & Quinto, L. (2008). Audio-visual integration of emotional cues in song. *Cognition and Emotion*, *22*(8), 1457–1470.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*, 420–422.