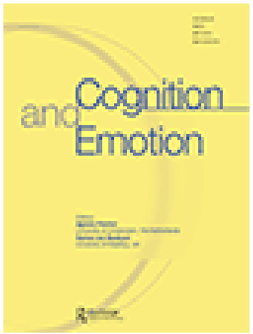

Routledge
Taylor & Francis Group

Cognition and Emotion

ISSN: 0269-9931 (Print) 1464-0600 (Online) Journal homepage: http://www.tandfonline.com/loi/pcem20

More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing

Piera Filippi, Sebastian Ocklenburg, Daniel L. Bowling, Larissa Heege, Onur Güntürkün, Albert Newen & Bart de Boer

To cite this article: Piera Filippi, Sebastian Ocklenburg, Daniel L. Bowling, Larissa Heege, Onur Güntürkün, Albert Newen & Bart de Boer (2016): More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing, Cognition and Emotion, DOI: 10.1080/02699931.2016.1177489

To link to this article: http://dx.doi.org/10.1080/02699931.2016.1177489

Published online: 03 May 2016.

COGNITION AND EMOTION, 2016
http://dx.doi.org/10.1080/02699931.2016.1177489

Routledge
Taylor & Francis Group

More than words (and faces): evidence for a Stroop effect of prosody in emotion word processing

Piera Filippi[a,b], Sebastian Ocklenburg[c], Daniel L. Bowling[d], Larissa Heege[e], Onur Güntürkün[b,c], Albert Newen[b,f] and Bart de Boer[a]

[a]Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Brussels, Belgium; [b]Center for Mind, Brain and Cognitive Evolution, Ruhr-University Bochum, Bochum, Germany; [c]Department of Biopsychology, Ruhr-University Bochum, Bochum, Germany; [d]Department of Cognitive Biology, University of Vienna, Vienna, Austria; [e]Department of General and Biological Psychology, University of Wuppertal, Wuppertal, Germany; [f]Institute of Philosophy II, Ruhr-University Bochum, Bochum, Germany

ABSTRACT

Humans typically combine linguistic and nonlinguistic information to comprehend emotions. We adopted an emotion identification Stroop task to investigate how different channels interact in emotion communication. In experiment 1, synonyms of "happy" and "sad" were spoken with happy and sad prosody. Participants had more difficulty ignoring prosody than ignoring verbal content. In experiment 2, synonyms of "happy" and "sad" were spoken with happy and sad prosody, while happy or sad faces were displayed. Accuracy was lower when two channels expressed an emotion that was incongruent with the channel participants had to focus on, compared with the cross-channel congruence condition. When participants were required to focus on verbal content, accuracy was significantly lower also when prosody was incongruent with verbal content and face. This suggests that prosody biases emotional verbal content processing, even when conflicting with verbal content and face simultaneously. Implications for multimodal communication and language evolution studies are discussed.

ARTICLE HISTORY
Received 8 October 2015
Revised 6 April 2016
Accepted 8 April 2016

KEYWORDS
Stroop task; emotion; multimodal communication; prosody; language evolution

Introduction

During emotive spoken communication, listeners use multiple sources of information spanning from verbal content to prosodic modulation, pragmatic context, facial expression, and gestures. In order to improve our understanding of emotion processing in spoken interaction, it is essential to bridge the study of each of these informative dimensions with empirical research on how they integrate and interact with each other during multimodal communication.

Research has shown that when emotional stimuli are conveyed only in one channel, emotion recognition is more accurate in the visual modality than in the auditory modality (Paulmann & Pell, 2011). However, in emotion communication, multiple channels can also strongly reinforce each other (Grandjean, Baenziger, & Scherer, 2006; Paulmann & Pell, 2011; Wilson & Wharton, 2006). Studies have shown that the integration of facial expression and prosody guides the perception of the speaker's emotional state (Belin, Fecteau, & Bédard, 2004; Campanella & Belin, 2007; Massaro & Egan, 1996). The integration of different sources of information in emotion comprehension has a relevant social function, as multiple emotional cues can be employed as appeals to appropriate behaviours, and ultimately, to regulate interpersonal interactions (Fischer & Roseman, 2007; van Kleef, De Dreu, & Manstead, 2004). Furthermore, the ability to use multiple channels to convey socially relevant information such as basic emotions might have provided adaptive advantages for the first hominins, paving the emergence of verbal language on a phylogenetic level (Mithen, 2005; Morton, 1977).

CONTACT Piera Filippi pie.filippi@gmail.com

Generally, audio and visual channels of emotion communication are integrated following two different dynamics: priming or simultaneous interaction. Numerous experiments have established that the verbal content and/or the prosodic modulation of segmental units prime the interpretation of a following target word (Nygaard & Lunders, 2002) or facial expression (Pell, 2002, 2005; Pell, Jaywant, Monetta, & Kotz, 2011; Schwartz & Pell, 2012) in an emotion-congruent manner. Moreover, it has been shown that emotional prosody also biases memory of affective words, again in an emotion-congruent manner (Schirmer, 2010; Schirmer, Kotz, & Friederici, 2002).

In emotion communication, linguistic and nonlinguistic channels can *simultaneously* express the same content or different content. Think, for instance about when someone says "I'm sad!", but does so in a happy prosody. Here, both verbal content and prosody express specific but conflicting emotions. The question, then, is: Which communicative channel are we most biased towards in identifying emotional content?

With the aim of exploring how communicative channels determine the degree and the direction of interference in emotion processing, much research has applied a Stroop-like task (MacLeod & MacDonald, 2000; Stroop, 1935). For instance, Stenberg, Wiking, and Dahl (1998) reported spontaneous attentional biases for emotional face processing over verbal content with emotional valence. In an emotional prosody–face Stroop task, De Gelder and Vroomen (2000) showed a bidirectional bias. The Stroop task paradigm has been applied to examine linguistic, cultural, and age biases in the interpretation of emotions conveyed by incongruent prosody and verbal content. For instance, Kitayama and Ishii (2002) have shown that English speakers are more attuned to the emotional valence of verbal contents, while Japanese speakers are more strongly influenced by prosody. In a follow-up study, Ishii, Reyes, and Kitayama (2003) found that Tagalog–English bilinguals in the Philippines showed an attentional bias for prosody regardless of the language used, a result that points in the direction of a cultural rather than a linguistic effect on emotion processing. Cultural effects on brain response to emotional expressions have been described in Liu, Rigoulot, and Pell (2015), who found that English native speakers are more attuned to emotional cues in the face rather than in the voice, when compared to Chinese native speakers. Wurm, Labouvie-Vief, Aycock, Rebucal, and Koch (2004), who adopted an emotional Stroop paradigm to investigate ageing effects in emotion recognition, found that older individuals are slower in processing emotion words that have high arousal intensity when prosody and verbal content mismatch. Finally, recent work in brain research showed sex differences (Schirmer & Kotz, 2003) and the activation of distinct brain areas (Grimshaw, 1998; Schirmer & Kotz, 2006) in processing emotions through interacting prosody and verbal contents. However, to our knowledge, no studies have analysed the interaction between more than two communication channels in conveying emotions simultaneously.

Our goal was to examine the relative saliency of multiple communication channels within an emotion identification task. To this end, we developed two Stroop experiments that address three issues in our understanding of emotion communication. First, although previous research using the Stroop task paradigm has examined the interaction of words and prosody in the expression of emotional valences, no behavioural Stroop experiment has ever combined emotional prosody with verbal content denoting emotions (for instance, "happy" and "sad"). For instance, Ishii et al. (2003) adopted verbal contents judged as pleasant or unpleasant (e.g. "refreshment" or "warm"), spoken with pleasant or unpleasant prosody. Similarly, Schirmer and Kotz (2003) adopted positive, neutral, or negative verbal contents, which were spoken with positive, neutral, or negative prosody. Contrasting emotional prosody with verbal contents (both referring to discrete emotional categories such as happy and sad) is advantageous in that it can help improve the understanding of how the verbal channel interacts with other expressive channels in recruiting cognitive resources for processing discrete emotions (see Bower, 1987), adding to the literature on the multi-channel interaction in processing dimensional emotions such as positive and negative valence. Furthermore, this design represents a more accurate variant of the original Stroop task, in which words denoting colours were contrasted with the actual colour of the text. Hence, to address this gap in the literature, we designed a Stroop experiment directly contrasting verbal content, prosody and, in experiment 2, also facial expression – as dimensions that interact to communicate the emotions happy and sad (Figure 1). We reasoned that using emotions as denotations that can be "transposed" across communication channels would enable a fine-tuned measurement of the interdimensional saliency effects.

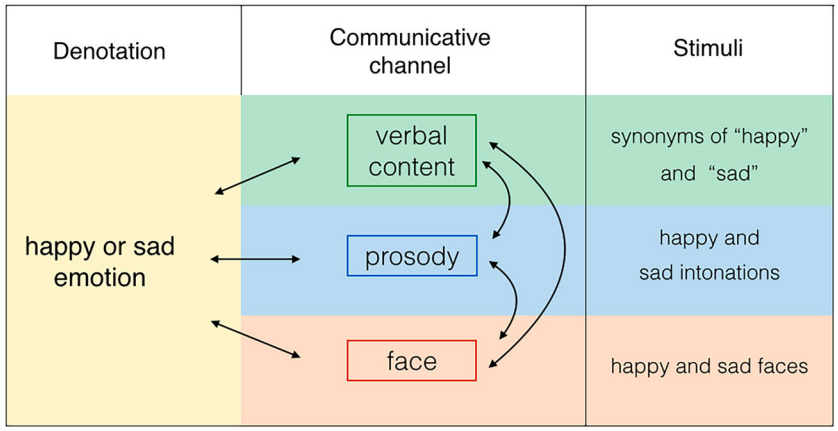

Figure 1. Representation of the study concept. The emotions happy and sad are used as denotations, which are communicated via multiple channels simultaneously. The aim of this study is to address the relative saliency of multiple communicative channels within an emotion identification Stroop task.

A second issue addressed by this work is that, although numerous Stroop experiments have focused on the relative role of verbal content and prosody in emotional speech processing, the outcomes of these studies do not always converge. Some of the Stroop experiments on emotion identification showed a bias towards the verbal content (Ishii et al., 2003), others claimed a bias towards prosody (Schirmer & Kotz, 2003), and still others failed to find any interference effects between prosody and verbal content (Wurm & Douglas, 1996). The resulting confusion warrants further exploration. Thus, in experiment 1, we investigated the relative saliency of prosody and verbal content in an emotion identification task, directly contrasting emotional congruence within these two channels.

Finally, as described above, much attention has been dedicated to the interference between two communication channels. With the aim of capturing more of the complexity associated with real-life interactions, we added facial expression as a third communicative channel in experiment 2. Specifically, experiment 2 was designed to assess how prosody, verbal content, and facial expression interact in emotional communication. We predicted that prosody would be more salient than verbal content and facial expression in an emotion identification task. Our prediction was built on previous studies showing that emotional prosody guides visual attention to emotional faces (Brosch, Grandjean, Sander, & Scherer, 2009; Rigoulot & Pell, 2012, but see Paulmann, Titone & Pell, 2012) and the perception of emotional verbal content (Nygaard, Herold, & Namy, 2009).

Experiment 1

Method and material

Participants. Twenty German native speakers (10 females, mean age = 22.65, SD = 3.89) were recruited at the Ruhr-University of Bochum. The experimental design adopted for this study was approved by the Ruhr-University of Bochum ethical review panel in accordance with the Helsinki Declaration. All participants gave written informed consents.

Selected stimuli description. In order to ensure that the emotional content of each stimulus was unequivocal across channels, for experiment 1 we adopted stimuli that were recognised as expressing the emotion "happy" or "sad" by at least 90% of a group of 24 participants in a previous stimulus validation experiment (Appendix 1). Here, respondents were asked to identify the emotion conveyed by each stimulus in a two-choice task (happy or sad), and to rate the emotional intensity of each stimulus on a 7-point Likert scale (0 = not intense, 6 = very intense). In order to keep the level of emotional intensity balanced across channels within each experimental trial, thus avoiding processing asymmetries across channels within each trial, only stimuli with a comparable level of emotional intensity were included in the same trial ($r = .803$, $p < .001$). The duration of the spoken stimuli is reported in Table 1.

Table 1. Duration (milliseconds) of the spoken stimuli used in experiments 1 and 2.

Spoken stimuli	Mean	Standard deviation
Synonyms of "happy"	757.733	117.565
Synonyms of "sad"	787.701	194.741

Verbal content. The stimuli selected for inclusion in the experiment consisted of two semantic sets: 16 German synonyms of "happy" and 16 German synonyms of "sad" (Appendix 2).

Prosody. For each of the two semantic sets selected for inclusion in the experiment, 8 verbal contents were spoken with happy prosody and 8 were spoken with sad prosody, for a total of 32 spoken items. Words spoken by four speakers (two males and two females) were selected based on the ratings made during the stimulus validation experiment (Appendix 1). The number of happy and sad verbal contents spoken by males and females in each emotional prosody was equal (Appendix 2).

Procedure. The experimental interface was created in *PsycoPy* (version 1.80.03; Peirce, 2007). Participants were individually tested in a quite laboratory, sitting at around 60 cm from an 18.5″ monitor. The entire procedure was computerised. Stimuli were played binaurally over Bayerdynamic DT990 pro headphones. Prior to starting the experiment, instructions were displayed on the screen and, in order to familiarise with the experimental procedure, participants ran four practice trials. Here, participants learned to identify the emotion conveyed by prosody while ignoring verbal content when the instruction "Prosody" preceded the trial (prosody task), and identify the emotion conveyed by verbal content while ignoring prosody when the instruction was "Word" (verbal content task).

The experiment consisted of 64 trials: 32 trials were played two times, each time with the instruction to focus on one emotion communication channel (32 with focus on prosody and 32 with focus on the verbal content). The order of trials was fully randomised across participants. Since the emotion expressed by each channel varied across trials, participants could not build any expectation on the exact combination of emotion conveyed by verbal content and prosody. Furthermore, by randomising the order of the channel that participants were instructed to focus on, they were prevented from being able to anticipate which channel they would have to ignore, and thus, from creating strategies to block unnecessary information. Responses were given by pressing one of two response keys that corresponded to two response options: the down arrow for "sad" and the up arrow for "happy". Participants were asked to respond as quickly as possible without sacrificing accuracy in judgment. Response time was measured in milliseconds from the onset of each stimulus. The inter-trial interval was 1500 ms. All statistical analyses were performed using SPSS for Mac OS X version 20.

Results and discussion

We first report the analysis of reaction times, followed by the analysis of accuracy data. The descriptive statistics for reaction times and accuracy are reported in Table 2.

Reaction time. Only correct responses were included in the analyses of reaction times. Overall, responses were very accurate (mean percentage correct = 94.5%, SD = 6.7%). Mean reaction times are displayed in Figure 2(a). A linear regression model was used within a repeated measures general linear model framework, to compare overall response times across and within experimental conditions. Task (prosody/verbal content) and congruence condition (congruent/incongruent) were entered as within-subject predictors. Reaction times were entered as the dependent variable.

The model revealed a significant main effect of task (Wald χ^2 (1) = 24.655, $p < .001$, $d = -0.57$), a significant main effect of congruence condition (Wald χ^2 (1) = 10.865, $p = .001$, $d = -0.36$), and no significant interaction between task and congruence condition (Wald χ^2 (1) = 0.274, $p = .601$).

These results provide evidence that in the prosody task, emotions were identified faster than in the verbal

Table 2. Experiment 1: Reaction time (milliseconds) and accuracy (% correct) for the different tasks and congruence conditions.

		Reaction time				Accuracy			
Task	Congruence condition	Min	Max	Mean	Standard deviation	Min	Max	Mean	Standard deviation
Prosody	Congruent	932.596	1522.089	1274.635	154.192	87.5	100	96.562	4.744
	Incongruent	1006.937	1622.102	1325.659	181.348	75	100	94.375	7.560
Verbal content	Congruent	1019.118	1599.955	1358.390	165.260	93.75	100	98.125	2.938
	Incongruent	1176.043	1722.434	1431.094	147.882	75	100	89.062	7.275

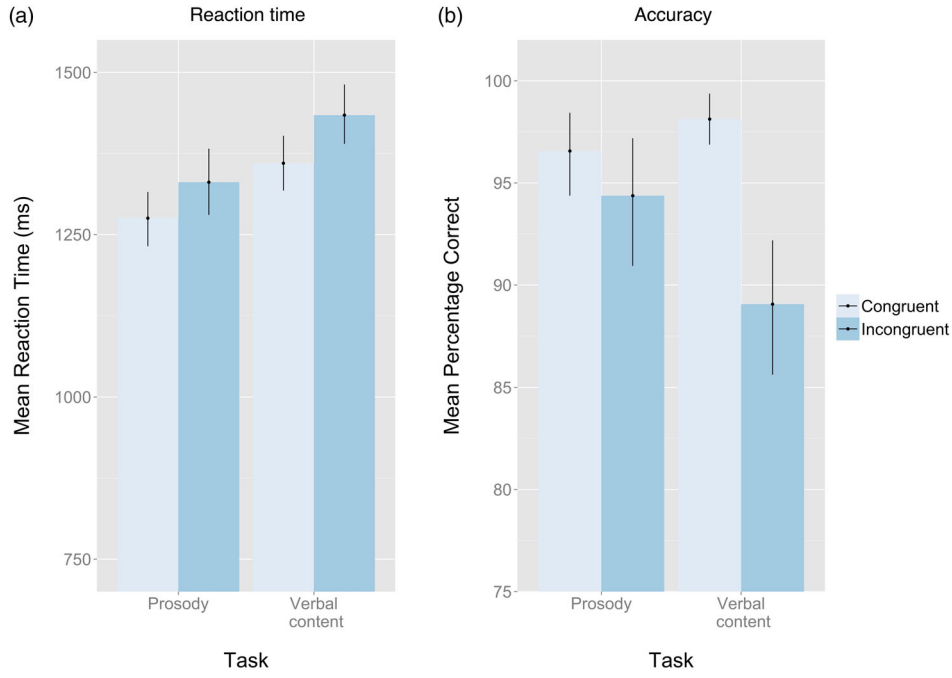

Figure 2. Experiment 1 – (a) Reaction times for the prosody task and the verbal content task averaged across participants. (b) Percentage of correct answers for the prosody task and the verbal content task averaged across participants. Results are shown separately for each congruence condition. Error bars represent 95% confidence intervals.

content task, in both congruence conditions. This suggests that prosody facilitates more rapid emotional understanding than verbal content. If collapsed across tasks, incongruent channels elicited slower responses than congruent ones overall. This is likely due in part to the fact that prosody and verbal content typically match in listeners' experience. In addition, faster reaction times on congruent trials might be caused by the use of prosody as a cue to word meaning (Kotz, Meyer, & Paulmann, 2006; Nygaard et al., 2009; Wildgruber, Ackermann, Kreifelts, & Ethofer, 2006).

Accuracy. Mean percentages of correct responses are displayed in Figure 2(b). The same model we used to analyse reaction times as a function of task and congruence condition was applied, with the only difference being that the percentages of correct responses were entered as the dependent variable.

The model revealed a significant main effect of task [Wald χ^2 (1) = 11.250, $p = .001$, $d = 0.28$], a significant main effect of congruence condition (Wald χ^2 (1) = 18.202, $p < .001$, $d = 0.90$), and a significant interaction between task and congruence condition (Wald χ^2 (1) = 17.410, $p < .001$). Pairwise comparisons were conducted within each task separately, using the sequential Bonferroni correction procedure (Holm, 1979). This analysis revealed a significant difference between congruence conditions within the verbal content task (Wald χ^2 (1) = 44.380, $p < .001$, $d = 1.65$), and no significant difference between congruence conditions in the prosody task (Wald χ^2 (1) = 1.604, $p = .205$).

These results provide evidence that in the prosody task, emotions were identified more accurately than in the verbal content task. If collapsed across tasks, incongruent channels elicited overall more incorrect responses than congruent ones. Pairwise comparisons computed within each task separately revealed that channels' incongruence within the verbal content task (but not within the prosody task) elicits significantly more incorrect responses than when the two channels are congruent. These results suggest that prosodic content is more salient than verbal content in emotion processing.

Experiment 2

Findings from experiment 1 suggest that in an emotional Stroop task, prosody is more salient than verbal content. In order to explore the interaction of prosody and verbal content with face in emotion communication, we designed a three-dimensional Stroop task.

Methods and materials

Participants. Twenty German native speakers (10 females, mean age = 22.2, SD = 3.87) were recruited at the Ruhr-University of Bochum.

Selected stimuli description

Verbal content and prosody stimuli were identical to those used in Experiment 1. The face stimuli consisted of colour photographs of a facial expression posed by two male and two female actors in a previous study (Pell, 2002) and were obtained through personal communication with the first author of that study. The faces used in experiment 2 were accurately recognised as conveying the emotion "happy" or "sad" by at least 90% of the participants in the stimulus validation experiment (Appendix 1). Here, as for the verbal content and prosody stimuli used in experiment 1, the respondents were asked to identify the emotion conveyed by each stimulus in a two-choice task (happy or sad), and to rate the emotional intensity of each stimulus on a 7-point Likert scale (0 = not intense, 6 = very intense). The selected happy and sad faces were equal in number ($n = 16$ per emotion, 4 for each actor). Spoken words were matched with individual faces posed by a member of the same sex. For each voice, four faces (two happy and two sad) posed by two actors of the same sex were selected. Care was taken that the level of emotional intensity across communication channels was of comparable strength. In order to keep the level of emotional intensity balanced across channels within each experimental trial, only stimuli with a comparable level of intensity were included in the same trial, as indicated by a Pearson's correlation coefficient (verbal content/face: $r = .715$, $p < .001$; verbal content/prosody: $r = .803$, $p < .001$; face/prosody: $r = .760$, $p < .001$).

Procedure. The setting of the experiment, interface, response procedure, and timings were identical to experiment 1. Within each experimental trial, a face was displayed while the spoken word was played. Specifically, the onset and offset of spoken words and faces were identical in each trial. In order to familiarise with the experimental procedure, participants ran eight practice trials. In this familiarisation phase, participants learned to identify the emotion conveyed by prosody, while ignoring verbal content and face when the instruction "Prosody" preceded the trial (prosody task), to identify the emotion conveyed by the verbal content, while ignoring prosody and face when the instruction "Word" preceded the trial (verbal content task), and to identify the emotion conveyed by the face, while ignoring prosody and verbal content when the instruction "Face" preceded the trial (face task). Congruence patterns were such that in each trial, two channels were congruent and the remaining one was incongruent, resulting in the following congruence conditions: congruent prosody–verbal content, congruent face–verbal content, and congruent prosody–face. In a fourth condition, all three channels were congruent (the cross-channel congruence control). Across trials, the channel that participants were instructed to attend to could be either one of the congruent channels or the incongruent channel. By comparing each condition against the cross-channel congruence control within each task, this design enabled the examination of how interference from two emotionally congruent channels affects a third incongruent channel in emotion processing.

Experiment 2 consisted of 96 trials: 32 trials were presented three times, each time with the instruction to focus on one of the three communication channels (32 with focus on prosody, 32 with focus on the verbal content, and 32 focus on the face). As in experiment 1, in order to prevent participants from building any expectation on the exact combination of emotion conveyed by each channel, we adopted an event-based design, in which the order of trials was fully randomised across tasks and congruence conditions.

Results and discussion

As for experiment 1, we first report the analysis of reaction times, followed by the analysis of accuracy data. The descriptive statistics for reaction times and accuracy are reported in Table 3.

Reaction time. Overall, responses were very accurate ($M = 95.3\%$, SD = 8.6%). Mean reaction times are displayed in Figure 3(a). We included only correct responses in the analyses of reaction times. A linear regression model was built within a repeated measure general linear model framework, to compare overall response times across and within experimental conditions. Task (prosody/face/verbal content) and congruence condition (congruent prosody–face/congruent prosody–verbal content/congruent face–verbal content/cross-channel congruence control) were entered as within-subject

Table 3. Experiment 2: Reaction time (milliseconds) and accuracy (% correct) for the different tasks and congruence conditions.

		Reaction time				Accuracy			
Task	Congruence condition	Min	Max	Mean	Standard deviation	Min	Max	Mean	Standard deviation
Face	Cross-channel	527.850	1691.615	983.545	315.416	75	100	98.75	5.590
	Prosody–verbal content	510.275	1695.783	964.018	337.485	62.5	100	91.25	10.806
	Prosody–face	534.179	1877.267	1010.862	380.598	87.5	100	98.125	4.579
	Face–verbal content	592.487	1754.187	1003.541	375.069	75	100	97.5	6.539
Verbal content	Cross-channel	955.973	1885.545	1391.243	254.049	87.5	100	98.75	3.847
	Prosody–verbal content	972.056	2008.672	1368.101	251.056	62.5	100	96.25	9.158
	Prosody–face	1003.972	2044.109	1481.164	284.279	75	100	91.25	9.158
	Face–verbal content	980.379	2023.217	1420.841	264.264	62.5	100	91.875	10.938
Prosody	Cross-channel	869.899	1948.171	1320.338	321.236	87.5	100	98.75	3.847
	Prosody–verbal content	955.384	1944.015	1385.575	277.292	62.5	100	95.625	9.314
	Prosody–face	1003.367	2897.121	1429.764	427.348	87.5	100	97.5	5.129
	Face–verbal content	784.405	1866.477	1413.809	283.258	62.5	100	88.75	12.098

predictors. Reaction times were entered as the dependent variable.

The model revealed a significant main effect of task (Wald χ^2 (2) = 127.285, $p < .001$), a significant main effect of congruence condition (Wald χ^2 (3) = 10.2485, $p = .017$), and no significant interaction between task and congruence condition (Wald χ^2 (6) = 4.393, $p = .624$). All pairwise comparisons were computed using the sequential Bonferroni correction procedure (Holm, 1979). Pairwise comparisons between

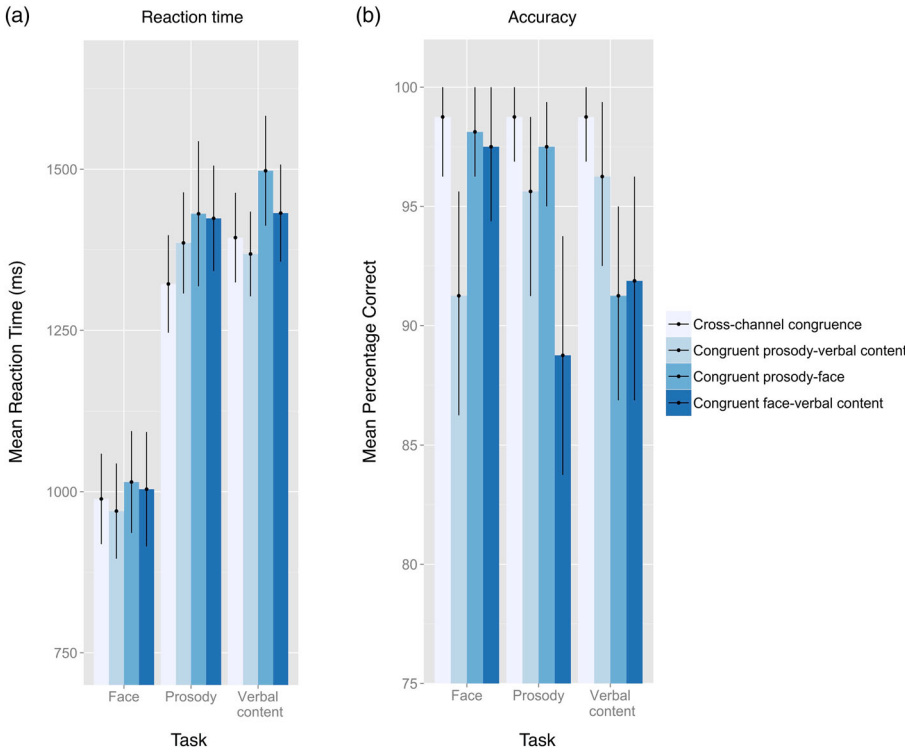

Figure 3. Experiment 2 – (a) Reaction times for the face task, the prosody task, and the verbal content task, averaged across participants. (b) Percentage of correct answers for the face task, the prosody task, and the verbal content task, averaged across participants. Results are shown separately for each congruence condition. Error bars represent 95% confidence intervals.

task conditions revealed a significant difference between face and prosody tasks (Mean Difference I–J (1) = 0.396, SE = 0.045, $p < .001$, $d = -1.175$) and between face and verbal content tasks (Mean Difference I–J (1) = 0.424, SE = 0.037, $p < .001$, $d = -1.384$). No significant difference was found between prosody and verbal content tasks (Mean Difference I–J (1) = 0.027, SE = 0.020, $p = .337$). Pairwise comparisons between congruence conditions and the cross-channel congruence control across tasks revealed a significant difference between the congruent prosody–face condition and the cross-channel congruence control (Wald χ^2 (1) = 9.308, $p = .007$, $d = 0.198$) and between the congruent face–verbal content condition and the cross-channel congruence control (Wald χ^2 (1) = 5.866, $p = .031$, $d = 0.023$).

These results provide evidence that in the face task, emotions were identified significantly faster than in the prosody and in the verbal content tasks. The inclusion of faces as a third emotional channel in this task levelled out the reaction time differences between the prosody and verbal content tasks detected in experiment 1. We observed slower reaction times in the congruent prosody–face and in the congruent face–verbal content conditions, but not in the congruent prosody–verbal content as compared to the cross-channel congruence control across tasks.

Accuracy. The same model used to analyse reaction times as a function of task and congruence was applied, with the only difference that the percentage of correct responses was entered as the dependent variable. Mean percentages of correct responses are displayed in Figure 3(b). The model revealed no significant effect of task (Wald χ^2 (2) = 4.505, $p = .105$), a significant main effect of congruence condition (Wald χ^2 (3) = 39.725, $p < .001$), and a significant interaction between task and congruence condition (Wald χ^2 (6) = 22.973, $p = .001$). Pairwise comparisons between the cross-channel congruence control and all the other congruence conditions were conducted within each task separately, using the sequential Bonferroni correction procedure (Holm, 1979). Within the face task, pairwise comparisons revealed a significant difference between congruent prosody–verbal content and cross-channel congruence control (Wald χ^2 (1) = 8.571, $p = .010$, $d = -0.875$). Within the prosody task, pairwise comparisons between congruence conditions and the cross-channel congruence control revealed a significant difference between congruent face–verbal content and cross-channel congruence control (Wald χ^2 (1) = 12.075, $p = .002$, $d = -1.123$). Within the verbal content task, pairwise comparisons revealed a significant difference between congruent prosody–face and cross-channel congruence control (Wald χ^2 (1) = 16.364, $p < .001$, $d = -1.075$), and between congruent face–verbal content and cross-channel congruence control (Wald χ^2 (1) = 8.094, $p = .009$, $d = -0.845$).

These results provide evidence that accuracy rate is comparable across all the attended channels we adopted in this experiment. Within each task condition, participants responded less accurately when the combination of any two channels was expressing an emotion that was incongruent with the attended channel, in comparison to the cross-channel congruence control.

Crucially, inspection of the congruence condition results shows that in the verbal content task, accuracy was significantly lower in the congruent prosody–face condition (where two channels were incongruent to the attended one) but also in the congruent face–verbal content condition, as compared to the cross-channel congruence control. The congruent face–verbal content effect is particularly striking as it suggests that in the verbal content task, accuracy was lower than in the cross-channel congruence control not only when verbal content (the attended channel) was incongruent to the other two channels, but also when prosody was incongruent to both verbal content and face. This result suggests that prosody biases responses in the process of identifying the emotion conveyed by verbal content, even when verbal content and face are both simultaneously expressing an emotion that is incongruent with the one expressed by prosody.

These data converge with, and extend the findings from experiment 1. The integration of reaction time and accuracy data suggests that, although processed significantly faster than prosody and verbal content, faces alone are not sufficient to interfere in emotion identification within a three-dimensional Stroop task. Faces and verbal content biased the process of emotion identification across tasks only when congruent with one of the other two channels. In contrast, prosody alone was sufficient to interfere against the combination of congruent face and verbal content, affecting accuracy in the verbal content task. This suggests that, in a task where emotions are simultaneously communicated by prosody, verbal content, and face – prosody exploits selective attentional resources within the process of identifying the emotion conveyed by verbal content.

General discussion

Humans typically combine two sources of information to comprehend emotions expressed in spoken interactions: linguistic, that is, verbal content, and nonlinguistic (e.g. body posture, facial expression, prosody, and pragmatic context). Prior studies on emotion processing using Stroop paradigms have shown that prosody, face, and verbal content influence each other (De Gelder & Vroomen, 2000; Ishii et al., 2003; Kitayama & Ishii, 2002; Nygaard & Queen, 2008; Schirmer & Kotz, 2003; Stenberg et al., 1998; Wurm & Douglas, 1996; Wurm et al., 2004), but have mainly focused on the interaction between two of these three dimensions (for instance, prosody and verbal content, or verbal content and faces). Furthermore, these studies used words judged for emotional valence rather than words directly denoting emotions as the verbal content channel. In our study, we adopted emotions as denotations that were signalled across multiple channels simultaneously.

In experiment 1, we tested a Stroop effect of happy and sad voice prosody over synonyms of "happy" and "sad". Experiment 1 indicated an interference effect for prosody in terms of accuracy rate. This result confirms earlier findings (Nygaard & Queen, 2008; Schirmer & Kotz, 2003) indicating that prosody recruits selective attention in verbal content processing. Moreover, our findings align with data from behavioural and brain studies suggesting that prosody elicits automatic responses in emotion processing, affecting verbal content processing (Mitchell, 2006; Schirmer & Kotz, 2003; Wildgruber et al., 2006).

In experiment 2, we extended the Stroop paradigm to an interference task where three emotion channels, specifically verbal content, prosody, and face, are simultaneously included. We thus simulated the complexity of ambiguous multimodal emotive communication using, for the first time in this research paradigm, three communicative channels within an interference task. Adopting stimuli that denote specific emotions enabled the examination of how the three different channels interact in conveying emotion-specific information, that is, the denotations "happy" and "sad", to the listeners at test. Importantly, findings from experiment 2 suggest that, while faces and verbal content interfered significantly with other channels only when they were in combination with a congruent channel (i.e. when the attended channel was incongruent to the other two channels), prosody alone was sufficient to interfere with both other channels when it was incongruent in the verbal content task. Notably, in experiment 2, reaction time and accuracy data did not mirror. The differences in these two analyses suggest that in the face task, the presence of two channels that are incongruent to face is strong enough to inhibit accuracy, but not to affect reaction time, and that in the verbal task, prosody alone is sufficient to inhibit accuracy, although it does not affect reaction times. This result warrants investigation at a neurobiological level. It might be that multimodal emotional processing results from the integration of multiple, independent sub-processes working in parallel (see Ho, Schröger, & Kotz, 2015) and that, if incongruent with both face and verbal content, prosody recruits selective attentional resources for identifying the emotion expressed by the incongruent verbal content.

Previous work has investigated emotion recognition in unimodal communication, providing evidence for the dominance of visual channels (specifically, of face expression and written words) over prosody in facilitating the recognition of disgust, happiness, surprise, and neutral expression (Paulmann & Pell, 2011). In the same study, prosody was found to be dominant over the visual channels for the recognition of sadness. Our work provides quantitative data on the perceptual saliency of prosody within a task where multiple congruent or incongruent emotions were expressed simultaneously through face, verbal content, and prosody. Our findings confirm and extend those of previous work in the field, indicating that emotional prosody provides a specific conceptual space that biases verbal content processing (Nygaard et al., 2009; Nygaard & Queen, 2008; Paulmann & Kotz, 2008), also in those cases where prosody is incongruous to both face and verbal content at the same time. Crucially, the pattern of results in both experiments is independent of any perceptual expectation, as the combination of the emotions expressed by the two (for experiment 1) or three (for experiment 2) channels, as well as the attended channel was randomly varied across trials.

Furthermore, our findings on the effect of congruence conditions in both experiment 1 and 2 complements findings from emotional priming research, which suggest that cross-modal congruence facilitates emotion processing (Nygaard & Lunders, 2002; Pell, 2002, 2005; Pell et al., 2011; Schwartz & Pell, 2012), but partially contrasts with data from Ishii et al. (2003), which indicate that Americans are more attuned to verbal content rather than to prosody in emotion valence judgments. As Ishii et al. (2003)

themselves argued, this inconsistency might be due to cultural differences in emotional information processing – in this case between Germans and Americans. Another possible explanation is that their study consisted of words with emotional valence rather than words denoting specific emotions. Future studies could examine differences in processing emotion words and emotional valences of non-emotion words, and cultural differences among further populations. Indeed, population sample, task design, and specific stimuli material used seem to be crucial in determining the extent of cross-channel interference (see Paulmann & Kotz, 2008).

Traditional models of lexical access and recognition have not typically recognised this crucial influential role of prosody in language processing (Ganong, 1980; Marslen-Wilson, 1987; McClelland & Elman, 1986; Morton, 1969). Conversely, one fundamental implication of our study is that a nonlinguistic dimension of communication, prosody, is remarkably important for online emotion word processing, since it biases the activation of specific emotional meaning representations. Intriguingly, the congruent prosodic modulation of verbal content or sentences conveying emotions can be considered a special case where linguistic and nonlinguistic channels refer to the same denotation (Nygaard et al., 2009).

Although the present study focused on two emotions (sad and happy) as denotations, it might be the case that other emotions are processed differently across communication channels and modalities. Within this research frame, future work could apply our experimental design to more emotions. Furthermore, based on our findings, we cannot exclude the existence of specific biases given by timing asymmetries between channels in the process of emotion recognition. For instance, it is plausible that prosody is detected earlier than the other channels, biasing decisions on the emotion conveyed by verbal content or face. Future research on behavioural and electrophysiological responses should address differences in processing timings between channels (see Eimer, Holmes, & McGlone, 2003; Pell et al., 2011; Scott, O'Donnell, Leuthold, & Sereno, 2009). This would further clarify the effect of each channel in the process of emotion recognition, pinpointing channel-specific biases. In addition, it is advisable for future studies to adopt highly sensitive response devices as suggested in Plant and Turner (2009).

This work could be further improved by also taking semantic and pragmatic concerns into account (Johnson-Laird & Oatley, 1989; Wilson & Wharton, 2006), and by adopting dynamic videos of faces time-synchronised with their voice, a feature that would permit examination of whether multimodal integration in emotional processing induces effects comparable to the McGurk illusion (Fagel, 2006; McGurk & MacDonald, 1976). In addition, due to the nature of our stimuli (which had to be recognised as expressing the same discrete emotion, at the same emotional intensity across channels in the validation experiment), the number of trials in our experiments was constrained to a limited number. Future studies adopting our experimental frame should include a higher number of stimuli to investigate interference effects across and within different communication channels.

Our findings on the salient role of prosody within contexts of cross-channel discrepancies is essential to understand what sensory cues favour the correct interpretation of ambiguous communications, including sarcastic, humorous or ironic messages (see Cutler, 1974). Furthermore, the investigation of the relative saliency and processing speed of communication channels within multimodal communicative situations might be applied to improve man–machine communication paradigms.

From an evolutionary perspective, our data fit with the hypothesis that the ability to communicate emotions through prosodic modulation of the voice – which appears to be dominant over verbal content – is evolutionary older than the emergence of segmental articulation (Fitch, 2010; Mithen, 2005). In line with this hypothesis, quantitative data suggest that prosody has a vital role in the perception of well-formed words (Johnson & Jusczyk, 2001), in the ability to map sounds to referential meanings (Filippi, Gingras, & Fitch, 2014), and in syntactic disambiguation (Soderstrom, Seidl, Nelson, & Jusczyk, 2003). This research could complement studies on iconic communication within visual and auditory domains.

Further work aimed at how emotional cues from different communication channels are integrated simultaneously will favour a better understanding of how humans interpret emotional contents in real-life interactions. Importantly, this research paradigm could provide crucial insights on what makes humans' ability to communicate unique.

Acknowledgements

We thank Nils Kasties, Charlotte Koenen, Sara Letzen, Sandra Linn, and Roland Pusch, for their help in recording the spoken

stimuli we adopted in this study. We are grateful to Annett Schirmer for her critical suggestions during the initial stages of experimental design and to Marc D. Pell for sharing the face stimuli. Piera Filippi developed the study concept. Piera Filippi, Dan Bowling, and Sebastian Ocklenburg contributed to the study design. Larissa Heege and Sebastian Ocklenburg performed testing and data collection. Piera Filippi performed data analysis. Piera Filippi drafted the manuscript, and all the other authors provided critical revisions. All authors approved the final version of the manuscript for submission.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by a research fellowship awarded to Piera Filippi by the Center for Mind, Brain and Cognitive Evolution (Ruhr-University Bochum, Germany) and by the European Research Council Starting Grant "ABACUS" [No. 293435] awarded to Bart de Boer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*, 129–135.

Bower, G. H. (1987). Commentary on mood and memory. *Behaviour Research and Therapy*, *25*, 443–455.

Brosch, T., Grandjean, D., Sander, D., & Scherer, K. R. (2009). Cross-modal emotional attention: Emotional voices modulate early stages of visual processing. *Journal of Cognitive Neuroscience*, *21*, 1670–1679.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*, 535–543.

Cutler, A. (1974). On saying what you mean without meaning what you say. In M. W. LaGaly, R. A. Fox, & A. Bruck (Eds.), *Papers from the tenth regional meeting, Chicago Linguistic Society* (pp. 117–127). Chicago: Chicago Linguistic Society.

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

Eimer, M., Holmes, A., & McGlone, F. P. (2003). The role of spatial attention in the processing of facial expression: An ERP study of rapid brain responses to six basic emotions. *Cognitive, Affective, & Behavioral Neuroscience*, *3*(2), 97–110.

Fagel, S. (2006). *Emotional McGurk effect*. Proceedings of the international conference on speech prosody, Dresden, 1.

Filippi, P., Gingras, B., & Fitch, W. T. (2014). Pitch enhancement facilitates word learning across visual contexts. *Frontiers in Psychology*, *5*, 1–8.

Fischer, A. H., & Roseman, I. J. (2007). Beat them or ban them: The characteristics and social functions of anger and contempt. *Journal of Personality and Social Psychology*, *93*, 103–115.

Fitch, W. T. (2010). *The evolution of language*. Cambridge: Cambridge University Press.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 100–125.

Grandjean, D., Baenziger, T., & Scherer, K. R. (2006). Intonation as an interference between language and affect. *Progress in Brain Research*, *156*, 1–13.

Grimshaw, G. M. (1998). Integration and interference in the cerebral hemispheres: Relations with hemispheric specialization. *Brain and Cognition*, *36*, 108–127.

Ho, H. T., Schröger, E., & Kotz, S. A. (2015). Selective attention modulates early human evoked potentials during emotional face–voice processing. *Journal of Cognitive Neuroscience*, *27*, 798–818.

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

Ishii, K., Reyes, J. A., & Kitayama, S. (2003). Spontaneous attention to word content versus emotional tone differences among three cultures. *Psychological Science*, *14*, 39–46.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 548–567.

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion*, *3*(2), 81–123.

Kitayama, S., & Ishii, K. (2002). Word and voice: Spontaneous attention to emotional utterances in two languages. *Cognition & Emotion*, *16*, 29–59.

van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2004). The interpersonal effects of anger and happiness in negotiations. *Journal of Personality and Social Psychology*, *86*, 57–76.

Kotz, S. A., Meyer, M., & Paulmann, S. (2006). Lateralization of emotional prosody in the brain: An overview and synopsis on the impact of study design. *Progress in Brain Research*, *156*, 285–294.

Liu, P., Rigoulot, S., & Pell, M. D. (2015). Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence. *Neuropsychologia*, *67*, 1–13.

MacLeod, C. M., & MacDonald, P. A. (2000). Interdimensional interference in the Stroop effect: Uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences*, *4*, 383–391.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition*, *25*, 71–102.

Massaro, D. W., & Egan, P. B. (1996). Perceiving affect from the voice and the face. *Psychonomic Bulletin & Review*, *3*, 215–221.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

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

Mitchell, R. L. C. (2006). How does the brain mediate interpretation of incongruent auditory emotions? The neural response to prosody in the presence of conflicting lexico-semantic cues. *European Journal of Neuroscience*, *24*, 3611–3618.

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

Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist*, *111*(981), 855–869.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*, 165–178.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: Acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science*, *33*, 127–146.
Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition*, *30*, 583–593.
Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: Linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 1017–1030.
Paulmann, S., & Kotz, S. A. (2008). An ERP investigation on the temporal dynamics of emotional prosody and emotional semantics in pseudo- and lexical-sentence context. *Brain and Language*, *105*, 59–69.
Paulmann, S., & Pell, M. D. (2011). Is there an advantage for recognizing multi-modal emotional stimuli? *Motivation and Emotion*, *35*(2), 192–201.
Paulmann, S., Titone, D., & Pell, M. D. (2012). How emotional prosody guides your way: evidence from eye movements. *Speech Communication*, *54*, 92–107.
Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python, *162*, 8–13.
Pell, M. D. (2002). Evaluation of nonverbal emotion in face and voice: Some preliminary findings on a new battery of tests. *Brain and Cognition*, *48*, 499–514.
Pell, M. D. (2005). Nonverbal emotion priming: Evidence from the facial affect decision task. *Journal of Nonverbal Behavior*, *29*, 45–73.
Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (2011). Emotional speech processing: Disentangling the effects of prosody and semantic cues. *Cognition & Emotion*, *25*, 834–853.
Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behavior Research Methods*, *41*(3), 598–614.
Rigoulot, S., & Pell, M. D. (2012). Seeing emotion with your ears: Emotional prosody implicitly guides visual attention to faces. *PLoS One*, *7*, e30740.
Schirmer, A. (2010). Mark my words: Tone of voice changes affective word representations in memory. *PloS One*, *5*, e9080.
Schirmer, A., & Kotz, S. A. (2003). ERP evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience*, *15*, 1135–1148.
Schirmer, A., & Kotz, S. A. (2006). Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, *10*, 24–30.
Schirmer, A., Kotz, S. A., & Friederici, A. D. (2002). Sex differentiates the role of emotional prosody during word processing. *Cognitive Brain Research*, *14*, 228–233.
Schwartz, R., & Pell, M. D. (2012). Emotional speech processing at the intersection of prosody and semantics. *PloS One*, *7*, e47279.
Scott, G. G., O'Donnell, P. J., Leuthold, H., & Sereno, S. C. (2009). Early emotion word processing: Evidence from event-related potentials. *Biological psychology*, *80*(1), 95–104.
Soderstrom, M., Seidl, A., Nelson, D., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*, 249–267.
Stenberg, G., Wiking, S., & Dahl, M. (1998). Judging words at face value: Interference in a word processing task reveals automatic processing of affective facial expressions. *Cognition & Emotion*, *12*, 755–782.
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.
Wildgruber, D., Ackermann, H., Kreifelts, B., & Ethofer, T. (2006). Cerebral processing of linguistic and emotional prosody: fMRI studies. *Progress in Brain Research*, *156*, 249–268.
Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38*, 1559–1579.
Wurm, L. H., & Douglas, A. V. (1996). Dimensions of speech perception: Semantic associations in the affective lexicon. *Cognition & Emotion*, *10*, 409–424.
Wurm, L. H., Labouvie-Vief, G., Aycock, J., Rebucal, K. A., & Koch, H. E. (2004). Performance in auditory and visual emotional stroop tasks: A comparison of older and younger adults. *Psychology and Aging*, *19*, 523–535.

Appendix 1: Stimuli validation procedure for experiments 1 and 2

Stimuli material was developed in two steps. First, we prepared sets of stimuli for each communication channel: verbal content, prosody and, for experiment 2, faces. The verbal content set consisted of 42 items (21 synonyms of "happy" and 21 synonyms of "sad"). The prosody stimuli set was built as follows: three male and three female lay actors recorded the 42 stimuli in happy and sad prosody. The actors were German native speakers. All stimuli were recorded in a quiet laboratory setting with a Stagg MD-500 Dynamic microphone at a 44.100 kHz sampling rate. Faces consisted of colour photographs of 8 females and 10 males' unobstructed facial expression (Pell, 2002).

Second, we developed a validation experiment with the aim to adopt only stimuli that convey the emotion of interest, that is, happy or sad, and to be able to associate stimuli with the same emotional intensity across communication channels, in each trial. For this validation experiment, a group of 24 native speakers of German (12 females, mean age: 26, 29) rated the emotional content and arousal intensity of a large set of stimuli within three communication channels: verbal content, prosody, face. Participants were instructed to identify the emotion expressed by (i) the verbal content displayed in written text, (ii) the intonation of the verbal contents played as spoken stimuli, and (iii) each face as "happy" or "sad" within a two-choice task. Participants also rated each item's emotional intensity on a 7-point Likert scale (0 = not intense, 6 = very intense). The order of the experimental blocks (verbal content, prosody and face) was counterbalanced across participants. The collected rating values for the selected stimuli are reported in Table A1.

Table A1. Emotional intensity ratings collected in a separate validation experiment for the stimuli used in Experiments 1 and 2.

Stimulus type	Emotion	Mean	Standard deviation
Face	Happy	3.595	0.703
	Sad	3.096	0.600
Prosody	Happy	3.543	0.770
	Sad	2.479	0.520
Verbal content	Happy	3.496	0.652
	Sad	3.314	0.643

Note: Participants rated each item's emotional intensity on a 7-point Likert scale (0 = not intense, 6 = very intense).

Appendix 2: Spoken stimuli used in experiments 1 and 2.

		Prosody			
		Happy		Sad	
		Male speaker	Female speaker	Male speaker	Female speaker
Verbal content	Happy	Beseelt (enlivened)	Begeistert (excited)	Fröhlich (cheerful)	Heiter (carefree)
		Hoffnungsfroh (hopeful)	Glücklich (happy)	Lustig (funny)	Vergnügt (jolly)
		Ausgelassen (frolic)	Freudig (joyful)	Erfreut (delighted)	Munter (chipper)
		Beschwingt (elated)	Erheitert (amused)	Motiviert (motivated)	Unbeschwert (happy-go-lucky)
	Sad	Bekümmert (distressed)	Verbittert (embittered)	Schmerzlich (grievous)	Jammernd (wailing)
		Trübselig (cheerless)	Traurig (sad)	Kummervoll (sorrowful)	Klagend (moanful)
		Trostlos (desolate)	Bedrückt (aggrieved)	Betrübt (unhappy)	Betroffen (sorrow-stricken)
		Deprimiert (depressed)	Freudlos (joyless)	Geknickt (broken)	Verzagt (disheartened)