Motiv Emot (2011) 35:192–201
DOI 10.1007/s11031-011-9206-0

ORIGINAL PAPER

Is there an advantage for recognizing multi-modal emotional stimuli?

Silke Paulmann · Marc D. Pell

Published online: 9 April 2011

Abstract Emotions can be recognized whether conveyed by facial expressions, linguistic cues (semantics), or prosody (voice tone). However, few studies have empirically documented the extent to which multi-modal emotion perception differs from uni-modal emotion perception. Here, we tested whether emotion recognition is more accurate for multi-modal stimuli by presenting stimuli with different combinations of facial, semantic, and prosodic cues. Participants judged the emotion conveyed by short utterances in six channel conditions. Results indicated that emotion recognition is significantly better in response to multi-modal versus uni-modal stimuli. When stimuli contained only one emotional channel, recognition tended to be higher in the visual modality (i.e., facial expressions, semantic information conveyed by text) than in the auditory modality (prosody), although this pattern was not uniform across emotion categories. The advantage for multi-modal recognition may reflect the automatic integration of congruent emotional information across channels which enhances the accessibility of emotion-related knowledge in memory.

Keywords Emotional prosody · Emotional semantics · Emotional facial expressions

S. Paulmann (✉)
Department of Psychology, University of Essex,
Wivenhoe Park, Colchester, Essex CO4 3SQ, UK
e-mail: paulmann@essex.ac.uk

M. D. Pell
School of Communication Sciences and Disorders,
McGill University, Montreal, QC, Canada

Introduction

Every day, telephone conversations end with one party wondering whether the person on the other end *really* meant what they said, for example, whether they were really looking forward to meeting their parents the following weekend. It is very likely that afterwards, the uncertain party wished "if only I had seen his face", somehow implying that either a different channel of expression (e.g. face) would be better to decode the emotional meaning of the utterance, or alternatively that an additional channel of information would lead to more accurate results when judging the emotional communicative intention of the other person. This situation does not only hold true for telephone conversations: just think how many times you have wondered whether the email or text you received was a bad joke or was actually meant to be taken seriously. In this case, weren't you hoping for a button that would read out the email in the appropriate emotional tone of voice? Again, this implicitly suggests that a combination of different sources of verbal and non-verbal cues makes it easier to perceive emotions from others than relying on a single cue alone.

But is it necessarily the case that more information leads to better recognition during emotional communication? Is it not also true that one can readily interpret that the cashier at the grocery store is angry by simply looking at her face? Or by listening to the tone of the voice alone, that someone is actually unhappy when they say "no, really, I'm fine"? In fact, there is an established literature which shows that during uni-modal emotion processing—principally, when adults were asked to categorize the emotional meaning of facial or vocal expressions in the absence of other emotionally-biasing cues—that basic emotions can be recognized quite accurately in a forced-choice response format

(e.g., Banse and Scherer 1996; Juslin and Laukka 2003; Borod et al. 2000; Ekman 1992; Elfenbein and Ambady 2002; Paulmann et al. 2008; Pell et al. 2009). For instance, Banse and Scherer (1996) explored how accurately listeners can infer the emotionality of a sentence based on the prosody only in sentences that did not contain any meaningful lexical-semantic information (pseudo-sentences). They presented stimuli belonging to one of fourteen emotional categories and report overall recognition rates of 48% (ranging from 78% accuracy to infer *hot anger* from the tone of voice to 22% correct in identifying *shame*), that is approximately seven times higher than predicted by chance (~7%). Similar good recognition rates have been reported for emotional recognition from facial expressions (e.g. Ekman and Friesen 1976) and lexical-semantic stimuli (e.g. Borod et al. 2000). Unfortunately, these data do not clearly exemplify whether additional channels of information about emotions (i.e., multi-modal stimuli) would promote significantly better recognition than when only one channel is present, and the influence of specific communication channels during multi-modal emotion processing remains poorly understood. The goal of the current study was to test the commonly held assumption: is the recognition of multi-modal emotional stimuli more accurate than for uni-modal stimuli? Given that much of our social interactions depend on the successful decoding of emotional information, it is critical to understand how we make use of different sources of emotional information and to identify whether we base emotional inferences on a particular hierarchy of information channels.

Surprisingly, this topic has received relatively little empirical attention. Some test batteries (in different languages) were developed that allow insight into how multi-modal stimulus processing may differ from uni-modal stimulus processing. For instance, the Diagnostic Analysis of Nonverbal Accuracy (DANVA; Nowicki and Duke 1994) test battery contains test-stimuli from the visual (static face) and auditory (voice) modality for four different emotions (*anger*, *sadness*, *fear*, *happiness*). In addition, the Profile of Nonverbal Sensitivity (PONS; Rosenthal et al. 1979) test battery contains *dynamic* stimuli that convey emotional cues (*anger*, *love*, *jealousy*, *gratitude*, *seduction*) as displayed in the face, voice, or body. These are conveyed either uni-modally or multi-modally. Finally, Baenziger et al. (2009) developed the multimodal emotion recognition test (MERT) which contains dynamic emotional expressions from the auditory and visual modality alone or in combination. Their test results show that emotion recognition is better when emotional cues are present in the face and voice at the same time as opposed to cues from the voice only. No such advantage was found between dynamic stimuli that contained face and voice information versus face information only, implying that facial cues are more easily interpreted than voice cues. Also, their data suggests that it is not necessarily the case that multi-modal information leads to better recognition rates for emotional expressions.

However, there is some evidence that emotional expressions encoded by more than one information channel are recognized with greater accuracy, consistent with the idea of a multi-modal "advantage". For instance, some studies interested in the integration and/or combination of emotional information have presented multi-modal emotional stimuli in paradigms that include mismatching or incongruent information between channels (also referred to as 'conflict situations'; see DeGelder and Bertelson 2003; Kreifelts et al. 2007, 2010; DeGelder and Vroomen 2000). During an fMRI experiment that looked at the audio-visual integration of non-verbal emotion stimuli, Kreifelts et al. (2007) required participants to classify the emotion of expressions presented in the auditory modality (single words spoken in emotional prosody), in the visual modality (emotional faces), or in combined multi-modal stimuli. Their behavioural results showed an advantage for recognizing audio-visual stimuli when compared to auditory or visual stimuli alone. These findings are in line with a report by DeGelder and Vroomen (2000) who demonstrated better recognition of *happy* and *sad* facial expressions when presented with a matching emotional tone of voice than no tone of voice. Similarly, Collignon et al. (2008) have shown that *fear* and *disgust* expressions are recognized more accurately when presented in a bi-modal condition (dynamic facial and vocal expressions) than when corresponding uni-modal stimuli are presented. Collectively, these findings imply that there is a discernable advantage to processing emotional displays when they are encountered in more than one communication channel.

In line with Baenziger et al. (2009), some of this research also promotes the idea that particular emotional channels dominate other channels when encountered in a uni-modal or multi-modal context. For instance, Collignon et al.'s (2008) results imply that facial stimuli dominate auditory stimuli when presented in an incongruency paradigm, as classification judgements were more accurate in the visual modality (at least for displays of *fear* and *disgust*). Similarly, other reports suggest higher recognition rates for facial emotional expressions when compared to vocal emotional expressions (e.g., Johnstone and Scherer 2000; Pell 2002; Hawk et al. 2009), or for emotional words when compared to emotional faces or voices (Borod et al. 2000). When prosody and semantics are compared, electrophysiological data imply that the literal meaning (semantics) of an utterance can predominate the processing of its voice tone during emotion processing (e.g., Kotz and Paulmann 2007), although it is not always clear whether this effect is related to the presentation of uni-modal (prosody) versus multi-modal (prosody + semantics)

stimuli, or because semantic content is systematically easier to recognize than emotional prosody. Finally, when the effects of uni-modal prosody, uni-modal semantic cues, and combined prosody and semantic cues about emotion were compared in a priming paradigm, there was no advantage to encountering emotions in one (uni-modal) versus two (bi-modal) speech channels as inferred from emotional congruency effects on decisions about a related facial expression (Pell et al., in press). The notion that some information channels dominate others, or at least under certain conditions, could be explained by differences in attention induced by task goals (Welch and Warren 1980) and/or by underlying differences in the 'information reliability' of each channel when encountered in a particular context (Schwartz et al. 1998). It is not immediately clear whether specific channel effects should emerge during the processing of multi-modal emotional displays when all available channels are relevant and when the task does not focus attention on a particular information channel, as is the case in the current study.

Thus, while our knowledge of how emotional information is integrated and recognized across channels is advancing steadily, the present literature is limited in a number of ways. Most of these studies have evaluated a very small number of emotions (sometimes as few as two) and/or did not include a neutral baseline; moreover, due to the nature of the tasks employed, the emotional exemplars presented in many of these studies are often highly atypical of natural emotional expressions (i.e., still picture frames or single words, rather than dynamic faces and ongoing speech). Frequently, these stimuli are presented in "conflict situations" which tend to have relatively low ecological validity (but see DeGelder and Bertelson 2003). Finally, many studies do not adequately differentiate the two distinct information sources available in the verbal channel, namely prosody and linguistic-semantic content (Pell et al., in press). To address some of these issues, here we presented dynamic emotion stimuli—short sentences—which always contained a congruent set of cues to express one of five basic emotions or no recognizable emotion (hereafter referred to as *neutral*). We then manipulated the availability of cues in three major communication channels—prosody (how something is said, or the tone of voice), semantics (what is said, or the literal meaning of the sentence), and facial expressions—to evaluate the recognition of each emotion under different uni-modal and multi-modal processing conditions. To effectively isolate prosody and semantics, in some of our conditions we presented grammatically well-formed sentences with an emotional semantic context ("lexical sentences"), and in others we presented pseudo-sentences or "nonsense speech" which were emotionally intoned but contained no semantic content (e.g., *Someone nestered the flugs* spoken in an angry prosody). Through the stepwise manipulation of the three emotion information channels, our approach allowed us to investigate recognition accuracy for uni-modal, bi-modal, and multi-modal displays of emotion, as well as to briefly explore the relative weight of each channel for recognizing particular emotion categories.

Based on the literature that implies that additional channel information promotes increased recognition of emotional displays, we expected to find a clear trend which shows that recognition accuracy is significantly better for multi- versus uni-modal stimuli (and possibly that accuracy increases when three versus two emotional channels are present). In light of data which show that emotions are often explicitly categorized more easily from the semantic content of an utterance or from facial expressions (Borod et al. 2000; Johnstone and Scherer 2000; Pell 2002; Paulmann et al. 2008), in the uni-modal condition we expected to find the lowest recognition rates when only prosodic cues were present. Importantly, our data allowed us to explore whether these general patterns are true for all emotions or possibly limited to certain emotions due to channel "dominance" effects.

Methods

The study consisted of two phases: first, a stimulus construction and validation experiment was undertaken to define perceptual attributes of videotaped emotional expressions, allowing us to select a subset of "valid" exemplars that could be manipulated to present expressions with uni-modal, bi-modal, and multi-modal cues; then, in the main experiment a forced-choice recognition experiment was performed by a new group of 72 participants who judged the emotional meaning of items prepared in the initial phase, based on different types and combinations of emotional cues.

Stimulus construction and validation

Participants

The participants in the validation phase of the study were six native English speakers (three female) who posed the emotional expressions, described previously by Pell (2002). As well, 20 young English-speaking listeners (ten female, mean age: 21.0 years) provided perceptual ratings of the recordings for validation purposes. All participants responded to electronic advertisements at McGill University and received compensation for their involvement.

Stimulus recording procedure

The materials were video recordings of short English sentences spoken by six amateur actors (three female, three

male) to convey one of five target emotions (anger, disgust, sad, happy, pleasant surprise) and neutral affect. As described by Pell (2002), all emotional meanings were portrayed (simulated) by each speaker/actor using common procedures for eliciting discrete emotions (e.g., actors were first shown pictures and/or were described situations associated with the target emotion). Each actor produced a series of sentences to express each target emotion; the six emotion types were blocked for the recording sequence, which was individually randomized for the six speakers. Video recordings framed the actor's head and shoulders as they produced each sentence and were captured directly onto digital media. To manipulate which speech cues would be available for listeners to judge the emotion being expressed, each actor produced two distinct types of sentences while being video recorded: for each emotion, they produced five "lexical" utterances which were grammatically well-formed English sentences with a semantic context that biased the target emotion (e.g., surprise: *He won the lottery*); and, they produced five semantically-anomalous "pseudo" utterances of comparable length, which resemble English in their grammatical and phonotactic properties, but contained no semantic information for recognizing emotion (e.g., *He nestered the flugs* spoken to express surprise). The same five pseudo-utterances were produced to express each of the six emotion types, whereas five distinct lexical items were constructed with semantic context to bias each emotion. To facilitate the ability of actors to produce pseudo-utterances in a manner that was as natural as possible, actors always produced the list of lexical items to convey the target emotion first, followed by the list of pseudo-utterances to convey the same emotion, during the recording session. This procedure has been employed successfully by other researchers who have simultaneously recorded "lexical" and "pseudo" utterances conveying emotion from the same actors (Castro and Lima 2010; Pell et al. 2009). A total of 360 video stimuli were recorded for editing and validation (6 speakers × 6 emotion types × 2 sentence types × 5 items).

Perceptual validation procedure

It was expected that some of our recorded items would not adequately portray the intended emotional meaning due to difficulties at the *encoding* stage (i.e., in the ability to simulate particular emotions in a laboratory setting, e.g., Pell et al. 2009). As our main focus was how emotions are recognized or decoded according to channel availability, a perceptual validation experiment was run to restrict items in our main experiment to the best exemplars of each target emotion based on ratings of the source videoclips prior to any editing (i.e., the recordings of actors producing lexical and pseudo-utterances, from which all other conditions were eventually constructed). Twenty young participants viewed all of the unmodified videos which were presented in random order in a single experiment, and they judged the emotion of the actor in a six forced-choice response format. Based on the group consensus about each item (% correct target identification), the best three exemplars of each emotion expression were selected for each of the six actors, when producing lexical utterances (i.e., stimuli containing face + prosody + semantics cues) and pseudo-utterances (i.e., stimuli containing face + prosody cues). Due to low recognition of one emotion expressed by three of the speakers, three "good" exemplars could not be chosen in three exceptional cases (one female speaker contributed only one surprise stimulus; another female speaker contributed only one anger stimulus; and one male speaker contributed only two disgust stimuli). This resulted in the selection of 16 angry, 17 disgust, 18 happy, 18 neutral, 18 sad, and 16 surprise stimuli for editing and presentation in the main experiment, in conditions involving both lexical and pseudo-utterances. The emotion identification rate for the selected stimuli, prior to editing, was high overall for source recordings containing both lexical and pseudo-utterances ($M = 80.6$ and 82.1%, respectively, where chance recognition was 16.7%). The validation data also revealed emotion-specific differences when listeners judged lexical versus pseudo-utterances, although these patterns were not analyzed since they were expected to re-emerge in the main experiment when all six conditions involving uni-modal, bi-modal, and multi-modal cues were constructed and judged by a new group of participants.

Stimulus editing/task construction

All selected videos were edited using Adobe Premiere software to construct six distinct "cue" conditions: three conditions that provided emotional cues in only one communication channel (*uni-modal*: prosody, semantics, or face); two conditions in which two of these channels were simultaneously available and emotionally congruent (*bi-modal*: face + prosody, prosody + semantics); and one condition in which all three channels were simultaneously present and emotionally congruent (*multi-modal*: face + prosody + semantics). By necessity, the method for editing stimuli to isolate specific cue combinations was achieved in different ways. For the three conditions that did *not* contain emotional semantics (uni-modal prosody, uni-modal face, and bi-modal face + prosody), the appropriate stimuli were constructed from the pseudo-utterance recordings produced by each actor to ensure that prosodic information was present in speech in the absence of semantic information. Specifically, the uni-modal prosody stimuli were constructed by extracting the audio track of videos in which actors expressed emotions via pseudo-

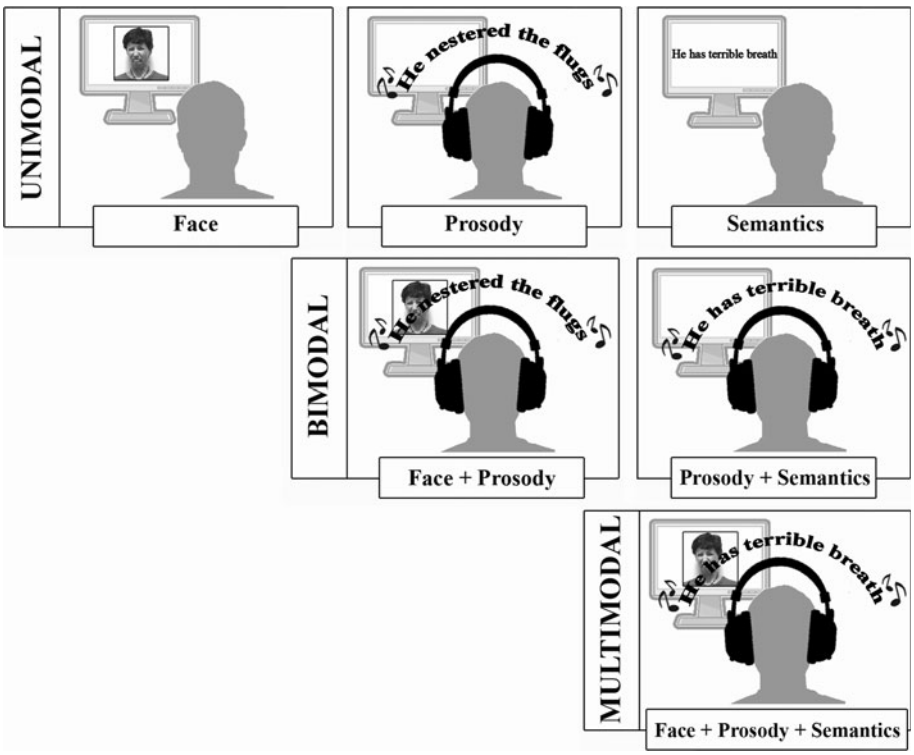

Fig. 1 Illustration of the six tasks presented in the experiment, according to whether emotional stimuli contained facial, prosodic, and/or semantic cues

sentences (saved as .wav audiofiles), and the uni-modal face stimuli were constructed by extracting only the video track of the same stimuli (saved as silent .avi videofiles).[1] The bi-modal face + prosody condition was achieved by presenting the corresponding, unaltered videos used to construct the uni-modal prosody and uni-modal face conditions.

In contrast, cue conditions that included the semantic channel (bi-modal prosody + semantics, multi-modal face + prosody + semantics) were constructed from the recordings of *lexical* sentences produced by the six actors: the bi-modal prosody + semantics stimuli were constructed by extracting the audio track of these videos (saved as.wav audiofiles), and the multi-modal face + prosody + semantics stimuli were the unaltered video recordings of actors producing lexical sentences. The only stimulus condition that could not be constructed from the video recordings was the uni-modal semantics condition; since prosodic information is always present in auditory language, stimuli for this condition were constructed by presenting the lexical sentences in written (text) format to eliminate the existence of both prosodic and facial cues. Figure 1 illustrates the six cue conditions in the experiment and how they were constructed.

Main experiment

Participants

Seventy-two native English speakers participated in the main experiment. To minimize potential carry-over and stimulus repetition effects in the experiment, participants were randomly assigned to one of three test groups who were presented only the uni-modal, bi-modal, or multi-modal stimuli (24 participants/group, half female). Participants in the three test groups were matched on a one-to-one basis for sex, age and education; there were no differences in mean group age (uni-modal = 21.9 years ± 2.4, bi-modal = 21.8 years ± 2.4, multi-modal = 21.8 years ± 2.7) or mean group education (uni-modal = 16.4 years ± 2.0, bi-modal = 16.3 years ± 1.6, multi-modal = 15.6 years ± 2.3). All participants were compensated for their involvement.

Task and procedure

Participants were tested individually in a quiet laboratory, seated at a comfortable viewing distance from a computer monitor. In each test group, each participant judged all of the respective stimuli (i.e. all of the uni-modal, bi-modal, or multi-modal stimuli). In total, the uni-modal test group

[1] For the uni-modal face condition, stimuli were initially extracted from both the lexical and pseudo-utterances, saved as silent .avi videoclips, which were presented to a group of raters. There was no statistically significant effect of identifying emotions from uni-modal face stimuli extracted from videoclips containing lexical versus pseudo-utterances; since including all of these items would yield twice as many items in this one condition, only uni-modal face stimuli from pseudo-utterances were used.

judged 236 stimuli (103 stimuli each in the face-only and prosody-only conditions, and 30 text stimuli in the semantics-only condition), the bi-modal group judged 206 stimuli (103 stimuli each in the face + prosody and prosody + semantic conditions), and the multi-modal group judged 103 stimuli (face + prosody + semantics condition). Stimuli were fully randomized for presentation within each cue condition, and the order of tasks was fully counterbalanced within the uni-modal and bi-modal groups. Visual stimuli (videos or text) were presented in the centre of the computer screen, and auditory stimuli were presented over high-quality headphones at a comfortable hearing level. After the participant listened to and/or viewed each stimulus, they were instructed to categorize the emotion that was being expressed from among six alternatives: anger, disgust, sadness, happiness, pleasant surprise, neutral (labels were varied in their position on the screen across participants to prevent response bias). Participants indicated their decision by clicking on the appropriate response label displayed on the screen, and there was no time limit to respond. Each testing session lasted between 30 min (multi-modal group) and 60 min (uni-modal and bi-modal groups). The experiment always began with instructions from the examiner and a set of practice trials before each task.

Results

To test whether multi-modal stimuli are better recognized than uni-modal stimuli, the emotional target unbiased hit rates (Hu; see Hawk et al. 2009 for similar approach) were analyzed with a repeated-measures analysis of variance (ANOVA) using the PROC MIXED function in Statistical Analysis System (SAS). Tukey–Kramer adjustments were applied when multiple comparisons were explored with *t*-tests. In a first step, responses were analyzed in a 3 × 6 design with the between-subject factor of *condition* (uni-, bi-, or multi-modal emotional information) and the within-subject factor of *emotional category* (anger, disgust, sadness, happiness, surprise, neutral). To investigate the influence of individual communication channels on emotion recognition, the data were analyzed in a second ANOVA with repeated factors of *emotional category* (see above) and *channel* (face, prosody, semantics). Only significant effects are reported in the text. The unbiased hit rates (Hu) for each of the six emotional categories are provided in Table 1 for each of the three test groups, according to what cues were available to aid target recognition.

Condition (number of channels) analysis

The main 3 × 6 (condition × emotional category) ANOVA yielded a significant main effect of *condition,* $F(2, 67) = 59.39$, $p < .0001$, indicating overall differences for recognizing emotional displays in the uni-modal, bi-modal, and multi-modal conditions. Tukey–Kramer adjusted *t*-tests revealed significantly lower Hu scores for uni-modal (.48) than for both bi-modal (.64), $t(67) = 5.97$, $p < .0001$, and multi-modal stimuli (.77), $t(67) = 10.87$, $p < .0001$. Also, multi-modal stimuli were recognized better overall than bi-modal stimuli, $t(67) = 4.85$, $p < .0001$. The general impact of condition on emotion recognition is illustrated in Fig. 2.

The analysis also yielded a significant main effect for *emotional category*, $F(5, 67) = 75.20$, $p < .0001$, and an interaction of *emotional category* and *condition,* $F(10, 67) = 4.16$, $p < .0001$. Post-hoc elaboration of the interaction revealed a significant recognition advantage for each emotional category when presented with multi-modal cues as opposed to uni-modal (and usually bi-modal) cues; statistical values for the *condition* effect for each emotional category were: anger, $F(2, 67) = 68.00$, $p < .0001$; disgust, $F(2, 67) = 27.53$, $p < .0001$; sadness, $F(2, 67) = 34.59$, $p < .0001$; happiness, $F(2, 67) = 44.74$, $p < .0001$; surprise: $F(2, 38) = 35.86$, $p < .0001$; and neutral, $F(2, 67) = 21.06$, $p < .0001$. Generally speaking, multi-modal stimuli were almost always recognized more accurately than bi- or uni-modal stimuli, and bi-modal stimuli were almost always recognized better than uni-modal stimuli (all t's > 2.76, and adjusted $p < .05$). The only exceptions to this pattern occurred when comparing accuracy rates for recognizing neutral affect from bi-modal versus uni-modal stimuli, and when comparing accuracy rates for recognizing pleasant surprise from multi-modal versus bi-stimuli stimuli (no significant difference in either case).

Channel dominance analysis

For the 6 × 3 (emotional category × channel) ANOVA which included stimuli from the different *uni-modal* channels, a main effect for *channel* was found, $F(2, 23) = 28.19$, $p < .0001$. Tukey–Kramer adjusted *t*-tests revealed more accurate emotion recognition for both face-only stimuli (.53), $t(23) = 6.38$, $p < .0001$, and semantics-only stimuli (.52), $t(23) = 5.19$, $p < .0001$, when compared to prosody-only stimuli (.40), implying an advantage for visually presented stimuli in the uni-modal condition. In addition, we again found a significant main effect of *emotional category*, $F(5, 23) = 44.09$, $p < .0001$, and interaction of *emotional category* and *channel*, $F(10, 23) = 168.79$, $p < .0001$. In the following, comparisons between the different channels and their adjusted *t*-values are listed by emotional category.

For anger recognition, there was no significant accuracy difference between the different uni-modal channels (a score of .50 for prosody-only, .55 for semantics-only, and .53 for face-only stimuli). For disgust recognition, accuracy

Table 1 Unbiased hit rates (Hu scores) for each emotional category according to group (uni-modal, bi-modal, multi-modal) and the communication channels available (standard errors in parentheses)

Group	Channel	Emotion						All emotions
		Anger	Disgust	Sadness	Happiness	Surprise	Neutral	
Uni-modal group	Prosody	0.50 (0.02)	0.26 (0.03)	0.63 (0.03)	0.28 (0.02)	0.26 (0.02)	0.45 (0.04)	0.40 (0.15)
	Semantics	0.55 (0.04)	0.73 (0.05)	0.59 (0.04)	0.43 (0.03)	0.18 (0.03)	0.61 (0.04)	0.52 (0.19)
	Face	0.53 (0.03)	0.59 (0.04)	0.54 (0.03)	0.65 (0.02)	0.43 (0.03)	0.42 (0.03)	0.53 (0.09)
Bi-modal group	Prosody + semantics	0.80 (0.03)	0.69 (0.03)	0.69 (0.03)	0.45 (0.02)	0.43 (0.03)	0.55 (0.04)	0.60 (0.15)
	Face + prosody	0.72 (0.03)	0.72 (0.03)	0.77 (0.02)	0.66 (0.02)	0.60 (0.03)	0.60 (0.04)	0.68 (0.07)
Multi-modal group	Face + prosody + semantics	0.89 (0.02)	0.82 (0.03)	0.85 (0.02)	0.73 (0.02)	0.60 (0.03)	0.75 (0.03)	0.77 (0.10)
All groups	All channels	0.67 (0.16)	0.64 (0.20)	0.68 (0.12)	0.53 (0.17)	0.42 (0.17)	0.56 (0.12)	0.58 (0.17)

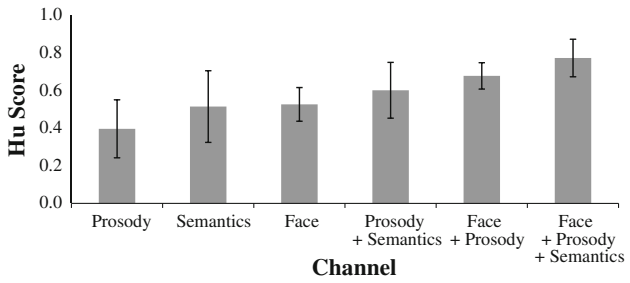

Fig. 2 General impact of channel availability on emotion recognition (averaged across the six emotion types)

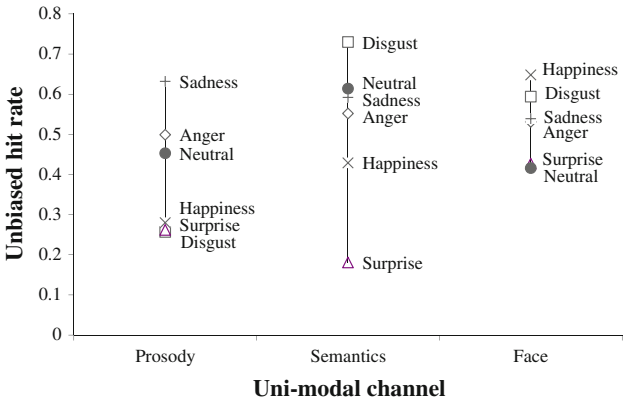

Fig. 3 Differences in the recognition of the six emotions in the uni-modal condition, when only facial, prosodic, or semantic cues are available

was significantly greater based on semantics (.73), $t(23) = 10.57$, $p < .0001$, and facial cues (.53), $t(23) = 7.44$, $p < .0001$, when compared to prosody (.26); the difference between semantic and face cues for recognizing disgust was marginally significant, $t(23) = 2.38$, $p = .06$. For sad recognition, prosody (.63) led to more accurate responses than facial expressions (.54), $t(23) = 2.69$, $p < .05$, but using semantics (.59) did not differ from either prosody or face for sadness. For happy recognition, facial cues (.64) promoted much better accuracy than semantic cues (.43), $t(23) = 5.97$, $p < .0001$, and both facial, $t(23) = 17.45$, $p < .0001$, and semantic cues, $t(23) = 4.49$, $p < .001$, promoted greater accuracy for happiness than prosody (.28). For surprise recognition (which tended to be lowest overall), accuracy for facial cues (.43) was greater than for both prosody (.26), $t(23) = 5.03$, $p = .0001$, and semantic cues (.18), $t(23) = 7.68$, $p < .0001$, and accuracy for prosody also exceeded semantics, $t(23) = -2.49$, $p < .0001$. Finally, for neutral stimuli, semantic cues (.61) promoted superior recognition than facial cues (.42), $t(23) = 3.23$, $p = .01$, and prosodic cues (.45), $t(23) = 2.81$, $p < .05$ (which did not differ for neutral). Taken together, these results suggest the dominance of both the semantic and face channel over the prosodic channel to recognize disgust, happiness, surprise, and neutral expressions; for neutral and disgust expressions, the semantic channel dominated the face channel, whereas the opposite was true for the two positive emotional categories where the face dominated (surprise and happy). Prosodic cues promoted better recognition of only sadness in our dataset, and interestingly, there were no significant differences in the recognition of anger as a function of uni-modal channels. These relationships are illustrated in Fig. 3.

Discussion

This study set out to investigate whether processes for recognizing displays of emotion conveyed by speech and/or face are facilitated by multi-modal when compared to uni-modal stimuli, through the step-wise addition of emotional channels. Secondly, we explored if particular channels for recognizing emotional information involving prosody, semantics, and/or facial expressions are associated with systematically different recognition rates, and whether this is similar for all emotions. In contrast to most previous studies, we explored the processing of several emotions at once (and a neutral category) which were encoded in dynamic stimuli which never presented a "conflict situation" during emotion processing. In general,

our results exemplify that as emotional channel availability increases, there is a corresponding increase in how accurately emotional displays are explicitly recognized, i.e., multi-modal stimuli were recognized significantly better than bi-modal stimuli, and bi-modal stimuli were recognized significantly better than uni-modal stimuli. Thus, while there is evidence that emotions can be recognized fairly well from only one channel in many instances (e.g., Borod et al. 2000; Ekman 1992; Paulmann et al. 2008; Pell et al. 2009)—confirmed here, where we found that unbiased hit rates in the uni-modal conditions ranged from .18 to .65 correct overall—our data establish that emotion recognition is facilitated by an enriched stimulus characterized by redundancy among the major communication channels.

Our results align with other data which show that emotional judgements tend to improve when more than one source of *congruent* information about the intended emotion is available (e.g., Collignon et al. 2008; DeGelder and Vroomen 2000; Massaro and Egan 1996; Pell 2005). It can be argued that more accurate recognition of multi-modal versus uni-modal stimuli provides indirect evidence for the integration of different information channels during emotional processing; it is reasonable to assume that emotional information channels need to be compared and/or integrated at some point to allow a holistic impression of the emotion being communicated. For example, Borod et al. (2000) have suggested that each emotional channel is first treated independently by separate sensory modality systems and then processed by a "general affective processor". In fact, the audio-visual integration of emotional information may be a mandatory, automatic process (see Massaro and Egan 1996; DeGelder et al. 1999; Kreifelts et al. 2007). While our findings do not directly inform the nature of emotion integration, they are nonetheless consistent with the idea that emotion recognition processes incorporate all available emotion cues, possibly in an involuntary manner, leading to systematically higher accuracy rates as observed here. Interestingly, this process did not appear limited to the processing of overtly emotional stimuli since we witnessed a similar advantage for neutral (i.e., non-emotional) displays when presented in multi-modal versus uni-modal stimuli.

One explanation for the advantage in recognizing stimuli with multiple, redundant channels could lie in the processes underlying (emotional) information processing. For instance, models of information processing assume that after encountering a stimulus, populations of neurons in modality-specific (e.g., auditory, visual, affective) input systems get activated and the systems can each act individually, but they are also highly interconnected allowing for a fusion of information (e.g., Niedenthal 2007). In emotion processing specifically, the notion of emotion concepts or nodes, or "mental processes that transform raw data of experience into manageable units", have also been discussed by several researchers; presumably, these concepts would have associative links with a variety of stimuli, including different types of expressive cues that convey a particular emotion (Bower 1981; Niedenthal and Halberstadt 1995; Russell and Lemay 2000). Assuming that information in each emotional channel activates shared conceptual knowledge about an emotion, one can speculate that as more congruent information gets activated, corresponding knowledge about emotions becomes increasingly more accessible for use during emotion recognition tasks. Certainly, more work is needed to replicate our findings and to test these claims.

In addition to showing that multi-modal stimuli facilitate emotion recognition, we explored whether particular channels are more effective for recognizing specific emotions. Overall, we noted that emotions presented in visual information channels (facial expressions, semantic content as conveyed via text) were recognized more accurately than in the auditory channel, at least when auditory information is restricted to the prosody. It is possible that these patterns reflect broad differences in how visual versus auditory information activate related emotion concepts during emotional communication; specifically, one of the unique characteristics of emotional expressions conveyed through prosody is that they are inherently dynamic and their meaning unfolds over a protracted time period. Researchers have argued that vocal emotion expressions are perceived categorically but in a probabilistic manner over time (Juslin and Laukka 2003). In contrast, the physical features which signify emotions in the facial channel can be processed instantaneously and are known to demonstrate strong category boundaries in perception (e.g., Etcoff and Magee 1992). Even when facial expressions are presented dynamically as was accomplished here, it is therefore likely that emotion-related knowledge triggered by faces can be activated differently over time than for corresponding vocal expressions (e.g., physical feature extraction can occur at any given point in time for facial expressions while the necessary interplay of acoustic cues over time may not engage a similar process for vocal expressions). This could have strengthened underlying knowledge about the emotion over time leading to improved performance in the face condition. Similarly, the emotional semantic information was highly prototypical of the target emotions and may have activated the underlying emotional conceptual knowledge strongly. Future studies may find that these differences in the physical properties and/or time-course for recognizing emotions in auditory versus visual information may be partly responsible for the apparent dominance of visual channels during emotion processing.

At the same time, recent data imply that the visual dominance effect should be viewed as flexible rather than absolute and that it often depends on the context (e.g., Collignon et al. 2008). Another important caveat is that our methods did not allow us to properly isolate semantic information presented in the *auditory* modality because this condition would have always included prosodic information which biased or conflicted with semantics (see Pell 2006 for a methodological discussion). Thus, we did not achieve a definitive test of how emotional information is naturally used in the auditory modality. Finally, it bears emphasizing that not all emotions were recognized less accurately from prosody; in fact, sadness was recognized most accurately in this channel when the uni-modal conditions are compared, as commonly reported (e.g., Banse and Scherer 1996; Pell et al. 2009). Also, it has been argued that expressions of fear, which were not investigated in our study, are often recognized advantageously in the vocal channel (Levitt 1964; Pell et al. 2009). This suggests that the salient features for recognizing discrete emotions are not always of equal value in the auditory and visual modality when these cues are presented simultaneously.

Differences in how particular emotions were recognized according to the channel may have been influenced by several factors. First, despite our efforts to perceptually validate our source recordings, we cannot exclude the possibility that the actors who portrayed the emotional stimuli in our study were not equally capable of *encoding* salient features in the vocal and facial channel, as there are well known individual differences in the ability to pose emotional expressions (e.g., Banse and Scherer 1996; Paulmann et al. 2008; Pell et al. 2009). However, given the strong precedence for the observation that some emotions are recognized systematically better from auditory or visual signals during emotional communication (e.g., De Silva et al. 1997; Pell 2002), it is unlikely that encoding difficulties explain our channel effects. In an interesting study, Busso et al. (2004) used a computer emotion recognition system to investigate how well facial expressions, vocal expressions, and fused expressions could be categorized according to major physical classifiers of each modality. For the auditory modality, several acoustical features were identified (e.g., pitch, energy, and durational cues) and for the visual modality several facial expression classifiers were selected (e.g., forehead, eyebrow, cheek position). As we have also argued here, their results demonstrated that uni-modal emotion recognition is highly successful in these computer based systems ($\sim$71% correct in the auditory modality and $\sim$85% correct in the visual modality), and in addition, that combining the classifiers for the auditory and visual modalities led to a further, significant increase in emotion recognition (to almost 90%). Interestingly, emotions that were misclassified in one modality were often more easily classified in the other modality. Along similar lines, Borod et al. (2000) reported channel-related differences during an emotional discrimination task which suggest that the physical properties underlying certain emotions in the facial or vocal channel vary in their perceptual complexity (e.g., those with upturned vs. downturned mouth; high vs. low pitch, see Borod et al. 2000). From an evolutionary perspective, these data fit with the notion that emotional features which are unique to the visual or auditory modality may have greater signal value for humans when communicating certain emotions.

In conclusion, our behavioural data establish that emotion recognition tends to be more successful when several information channels are simultaneously present. Assuming that emotional information in each channel is somehow integrated to form a unified impression about a speaker's emotion, the fact that multiple, congruent channels enhance recognition processes may be explained by increased activation of emotion-related knowledge or "emotion concepts" which are used during emotional communication, and in the formation of social impressions which revolve around emotional cues. Nonetheless, while it seems clear that multi-modal stimuli benefit emotion recognition processes over uni-modal stimuli, our data caution that the recognition of discrete emotions is not always equally successful in many "impoverished" contexts when communication channels are missing. As such, there will continue to be doubt about the intended significance of many telephone messages and emails that lack critical markers and redundancy about emotion, owing to differences in the channels typically used to communicate discrete emotions.

Acknowledgments The authors would like to thank Meg Webb and Catherine Knowles for help with the stimuli and data acquisition. This work was supported by a Postdoctoral fellowship from the German Academic Exchange Service (DAAD) awarded to the first author, and by a Discovery grant awarded to the second author by the Natural Sciences and Engineering Research Council of Canada.

References

Baenziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body. The Multimodal Emotion Recognition Test (MERT). *Emotion, 9*(5), 691–704.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 3*, 614–636.

Borod, J. C., Cicero, B., Obler, L. K., Welkowitz, J., Erhan, H. M., Santschi, C., et al. (1998). Right hemisphere emotional perception. Evidence across multiple channels. *Neuropsychology, 12*, 446–458.

Borod, J. C., Pick, L. H., Hall, S., Sliwinski, M., Madigan, N., Obler, L. K., et al. (2000). Relationships among facial, prosodic, and

lexical channels of emotional perceptual processing. *Cognition and Emotion, 14*, 193–211.

Bower, G. H. (1981). Mood and memory. *American Psychologist, 36*, 129–148.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of ACM 6th International Conference on Multimodal Interfaces (ICMI 2004)*, State College, PA, 2004.

Castro, S. L., & Lima, C. F. (2010). Recognizing emotions in spoken language: A validated set of Portuguese sentences and pseudo-sentences for research on emotional prosody. *Behavior Research Methods, 42*(1), 74–81.

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

De Silva, L. C., Miyasato, T., & Natatsu, R. (1997). Facial emotion recognition using multimodal information. In *Proceedings of IEEE International Conference on Information, Communications and Signal Processing (ICICS'97)*, pp. 397–401.

DeGelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences, 7*, 460–467.

DeGelder, B., Böcker, K. B. E., Tuomainen, J., Hensen, M., & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses. *Neuroscience Letters, 260*, 133–136.

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

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6*, 169–200.

Ekman, P., & Friesen, W. (1976). *Pictures of facial affect*. Palo Alto, CA: Consulting Psychologist's Press.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin, 128*, 203–235.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44*, 227–240.

Hawk, S. T., van Kleef, G. A., Fischer, A. H., & van der Schalk, J. (2009). Worth a thousand words: Absolute and relative decodability of nonlinguistic affect vocalizations. *Emotion, 9*(3), 293–305.

Johnstone, T., & Scherer, K. R. (2000). Vocal communication of emotion. In M. Lewis & J. Haviland (Eds.), *Handbook of emotions* (2nd ed., pp. 220–235). New York: Guilford Press.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin, 129*, 770–814.

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.

Kotz, S. A., & Paulmann, S. (2007). When emotional prosody and semantics dance cheek to cheek: ERP evidence. *Brain Research, 1151*, 107–118.

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *Neuroimage, 37*, 1445–1456.

Kreifelts, B., Ethofer, T., Huberle, E., Grodd, W., & Wildgruber, D. (2010). Association of trait emotional intelligence and individual fMRI-activation patterns during the perception of social signals from voice and face. *Human Brain Mapping, 31*(7), 979–991.

Levitt, E. A. (1964). The relationship between abilities to express emotional meanings vocally and facially. In J. R. Davitz (Ed.), *The communication of emotional meaning* (pp. 87–100). New York: McGraw-Hill.

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

Niedenthal, P. M. (2007). Embodying emotion. *Science, 316*, 1002–1005.

Niedenthal, P. M., & Halberstadt, J. B. (1995). The acquisition and structure of emotional response categories. *The Psychology of Learning and Motivation, 33*, 23–63.

Nowicki, S., & Duke, M. (1994). Individual differences in the nonverbal communication of affect. *Journal of Nonverbal Behavior, 18*, 9–36.

Paulmann, S., Pell, M. D., & Kotz, S. A. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104*, 262–269.

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*, 499–504.

Pell, M. D. (2005). Nonverbal emotion priming: evidence from the 'facial affect decision task'. *Journal of Nonverbal Behavior, 29*(1), 45–73.

Pell, M. D. (2006). Cerebral mechanisms for understanding emotional prosody in speech. *Brain and Language, 96*(2), 221–234.

Pell, M. D., Jaywant, A., Monetta, L., & Kotz, S. A. (in press). Emotional speech processing: disentangling the effects of prosody and semantic cues. *Cognition & Emotion*. doi: 10.1080/02699931.2010.516915.

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics, 37*, 417–435.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test*. Baltimore: John Hopkins University Press.

Russell, J., & Lemay, G. (2000). Emotion concepts. In M. Lewis & M. J. Haviland-Jones (Eds.), *Handbook of emotion* (2nd ed., pp. 491–503). New York: Guilford Press.

Schwartz, J.-L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell (Ed.), *Hearing by eye: The psychology of lipreading* (pp. 3–51). Hove, UK: Lawrence Erlbaum Associcates.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88*, 638–667.