COGNITION AND EMOTION, 2000, *14* (3), 289–311

The perception of emotions by ear and by eye

Beatrice de Gelder
Tilburg University, The Netherlands and Université Libre de Bruxelles, Belgium
Jean Vroomen
Tilburg University, The Netherlands

Emotions are expressed in the voice as well as on the face. As a first step to explore the question of their integration, we used a bimodal perception situation modelled after the McGurk paradigm, in which varying degrees of discordance can be created between the affects expressed in a face and in a tone of voice. Experiment 1 showed that subjects can effectively combine information from the two sources, in that identification of the emotion in the face is biased in the direction of the simultaneously presented tone of voice. Experiment 2 showed that this effect occurs also under instructions to base the judgement exclusively on the face. Experiment 3 showed the reverse effect, a bias from the emotion in the face on judgement of the emotion in the voice. These results strongly suggest the existence of mandatory bidirectional links between affect detection structures in vision and audition.

A now classical volume entitled *Language by Ear and by Eye* (Kavanagh & Mattingly, 1972) presents a number of studies drawing attention to the fact that language is presented to two different modalities, to the eyes in reading

Requests for reprints should be sent to Beatrice de Gelder, Department of Psychology, University of Tilburg, Warandelaan 2, PO Box 90153, 5000 LE Tilburg, The Netherlands; email: b.degelder@kub.nl.

Parts of this paper were presented at the Haskins Laboratories, January 1995, at the *4th Meeting of the European Society for Philosophy and Psychology*, Oxford, 30 August–2 September 1995 and at the 36th Annual Meeting of the Psychonomic Society, 10–11 November 1995. Preparation of this paper was partly supported by an ARC grant (Action de Recherche Concertee, Belgian Ministry of Education) to the first author. Research of the second author was funded by the Royal Academy of Sciences. Preparation of the manuscript was undertaken while the first author was a guest of the Department of Philosophy and Linguistics at MIT. We particularly thank P. Bertelson for comments on a previous draft, and the IPO, Eindhoven for their help in preparing the auditory stimuli. We gratefully acknowledge very useful comments by three anonymous referees.

and to the ears in speech perception. Another role for the visual modality in speech communication has been highlighted in later work which showed the importance of visual information sampled from the talker's face (McGurk & MacDonald, 1976). The importance of facial movements for speech perception (also called "lipreading" or "speechreading") has gained wide recognition with the work of McGurk and MacDonald showing that speech information from the voice and concurrent presentation of speech information from the face that is incompatible leads to illusory percepts. In some cases, subjects reported hearing sounds that were provided neither by the voice alone nor by the movements of the face alone, but involved some combination of the two. In the most spectacular example, an auditory "baba" combined with a visual "gaga" was often perceived as "dada". The McGurk effect, as the phenomenon is now generally called, offers a particularly striking example of audiovisual integration and adds to a body of knowledge which also includes cross-modal interactions observed in the localisation of sounds and lights (Bertelson, 1998).

Language is not unique in being perceived both by ear and by eye. In everyday life, the perception of emotions appears to be similarly bimodal. The ability to recognise emotions manifested in a variety of behaviours like face expressions, voice expressions, gestures, and gait is undoubtedly a very important basis for initiating action. Darwin was among the first theoreticians of emotion to consider emotions as closely related to action (Frijda, 1989). From such a "perception for action" vantage point, it matters little whether the information about somebody's affective state is obtained from seeing his face, hearing his voice, or from both. Perception of either an angry face or an angry voice generally leads to the conclusion that the person is angry, not that his face looks angry, nor for that matter that his voice sounds angry. This intuitive reasoning is in line with the assumption that emotions in the face as well as emotional expressions in the voice are processed by the same perceptual or cognitive mechanism. But the assumption of a common and thus amodal or abstract processing route is far from being supported by presently available research. As a matter of fact, most studies have been concerned with facial expressions of emotions, fewer studies are about affective prosody, and only very few have looked at the combination of both which is the topic of the present study. The issue of common processing resources is even more remote from presently available data.

The overwhelming majority of studies concerned with the perception of emotion have concentrated on face perception. The studies by Ekman and collaborators are widely known (see Ekman, 1992) to have established that at least some facial expressions are readily recognised in different age groups and in different cultures. In the present study we have chosen a small number or relatively uncontroversial emotions and our approach is neutral with respect to discussions on social or cultural factors determining

the scope of the emotion repertoire or affecting the ease of recognition of certain emotions in specific populations. Much uncertainty remains, though, about the perceptual processes underlying the recognition of expression as contrasted, for example, with the recognition of personal identity. One central question concerns the bearers of the facial expression. Do face parts each play a role independently, are the eyes more important than the mouth, or is it the whole face as a Gestalt that conveys the emotional information? For the case of personal identity it has generally been assumed that configural (Rhodes, Brake, & Atkinson, 1993) or holistic (Tanaka & Farah, 1993) processes starting from the whole face play a critical role. Configural/holistic processing must presumably be understood in the sense of Pomerantz (1981) as involving relations between component parts of the total input which are not available when those parts are presented in isolation. Configural rather than component-based recognition processes may be similarly important for expressions. In the case of identity, an argument for configural processing has traditionally been derived from the effect of face inversion (Rhodes et al., 1993; Yin, 1969). In experiments that compared the recognition of facial expressions from upright versus inverted faces, McKelvie (1995) and de Gelder, Teunisse, and Benson (1997d) obtained evidence for loss of recognition in the inverted condition for some, but not for all expressions. Thus, in some cases at least, affect-relevant information would be carried by the whole facial configuration and consequently lost with inverted presentation. The issue of whole versus parts in the perception of facial expressions gains added importance in studies of the combination of facial information with input from the voice. The eyes might be the most important source of affect information, but combining a voice with a face expression could direct attention to the mouth instead and thereby make the lower part of the face more important. Before turning to that issue we focus on some relevant aspects of studies of voice affect.

The way emotions are expressed in the voice has received considerable attention in recent years. Traditionally, researchers have either analysed the way speakers express emotions in prosodic parameters such as pitch and duration (encoding), or they investigated how well listeners were able to recognise an emotion as intended by the speaker (decoding). With the advent of new speech technology, a more integrative approach became possible in which the relevant parameters in the speech signal were manipulated or synthesised, and then presented to a listening subject for recognition. Most attention has been paid to the contribution of variations in pitch, duration, loudness, and voice quality as measured in natural or in simulated affective speech (e.g., Cummings & Clemments, 1995; Liberman & Michaels, 1962; Williams & Stevens, 1972). Many experiments have started from a set of emotional utterances in which one or more prosodic

features were eliminated, so that it could be determined how well these degraded utterances could still be labelled in terms of candidate emotions. If a prosodic feature is not used in communicating the emotion, then eliminating it from an utterance should have no effect on recognition, and if it is the only feature that is left, recognition should be at chance level. In a prototypical study, Lieberman and Michaels (1962) determined via a fixed-vowel POVO-type synthesiser the contribution of pitch and amplitude to the expression of several emotions/attitudes. The conclusion they reached has been arrived at in many similar studies (e.g., Carlson, Granström, & Nord, 1992; Cosmides, 1983; Fairbanks & Pronovost, 1939; Ladd, Silverman, Tolkmitt, Bergman, & Scherer, 1985; Protopapas & Lieberman, 1997; Williams & Stevens, 1972; for review see Frick, 1985; Scherer, 1986; and more recently Murray & Arnot, 1993): Many prosodic features contribute to the expression of emotion, but it is evident that the acoustic correlates are subject to large individual differences. The associations that have been found between prosodic features and affect vary from study to study (cf. Scherer, 1989; Williams & Stevens, 1972; for an overview, see Frick, 1985), and different speakers seem to favour different acoustic parameters for the same emotion (e.g., Lieberman & Michaels, 1962). However, despite the large interspeaker variability, there is some general consensus that if prosodic features are ranked in terms of their contribution, gross changes in pitch do contribute most to the transmission of emotions, duration is intermediate, whereas loudness seems to be least important (cf. Frick, 1985; Murray & Arnott, 1993).

An important question is what the perception of emotion in the face and the voice have in common. It appears that there are some differences in the effectiveness with which the face and the voice convey different emotions. Happiness is often the easiest facial expression to recognise and the only one that remains accessible when the face is presented upside down (de Gelder et al., 1998). Data form studies of patients with focal brain damage point in the same direction. In cases in which the damage impairs recognition of several facial expressions, recognition of happiness can still be partly preserved (Etcoff, Freeman, & Cave, 1991; de Gelder et al., 1997c). But when it comes to expression in the voice, happiness can sometimes be hard to tell apart from other emotions (e.g., Vroomen, Collier, & Mozziconacci, 1993, Experiment 1).

In studies of brain damaged patients the issue of an association of deficits of emotion recognition in the face as well as in the voice has been raised and the question of a common neuroanatomical basis was addressed. The main issue has been whether a deficit in the perception of face expression has a parallel in impaired recognition of voice expression and whether impaired voice expression recognition leaves the recognition of face expression intact (see van Lancker, 1997, for an overview). Research

by van Lancker and associates suggests that there is a relation between face and voice expression impairments. The issue of common processing structures for voice and face affect has been pursued in studies of amygdalectomy patients but at present there is an inconsistency between available evidence. One study reported that an amygdalectomy patient with a deficit in recognition of facial expressions was equally impaired in processing affective prosody (Scott et al., 1997). But another study reported a case where this similarity of deficits did not obtain (Anderson & Phelps, 1998).

The bimodal perception of emotions (i.e., the situation in which the face and voice are presented together), presents a relatively less explored topic. In a developmental study of intermodal perception of emotions, infants were presented with faces combined with voices. Five- to seven-month-old infants looked longer at a face that carried the same expression as the voice than at a face carrying a different expression (Walker & Grolnick, 1983). There are also a number of studies investigating the relative importance of the information from the auditory and visual channels. For example, Mehrabian and Ferris (1967) presented a slide of the face of a woman portraying "like", "dislike", or "neutral" with respect to another imagined person, and they orthogonally combined the slide of the face with the word "maybe" spoken in a "like", "dislike", or "neutral" tone of voice. They estimated that the facial information was 3/2 times more important than the vocal information. Their result is in line with the more general conclusion that the face is more important than the voice information for judging a portrayed emotion (e.g., Bugenthal, Kaswan, Love, & Fox, 1970; Hess, Kappas, & Scherer, 1988).

These studies have provided important information but none of them was focused on the integration mechanism underlying the combination of voice and face information in the course of perception. This was the goal of a study by Massaro and Egan (1996; see also Massaro, 1998). These authors presented their subjects with a single word recorded by a speaker in one of three affective tones (happy, angry, neutral) and showed them a computer-generated face displaying one of the same three moods. The instructions were to classify the emotion as happy or angry. The frequency of either response depended on the emotions expressed in both the face and the voice. The authors discussed these results mainly in terms of the better fit provided by Massaro's multiplicative model of feature integration (the FLMP) in comparison with an alternative additive model. An argument for the multiplicative model was also derived from the existence of a strong correlation between reaction time (RT) and a measure of the ambiguity of each input configuration regarding the target decision.

The study to be described in the present paper was similar to Massaro and Egan's (1996), but carried out independently and with a number of differences. Like the latter study, it started with the question whether

subjects who are presented simultaneously with affect-relevant information from a seen face and a heard voice combine the two sources to decide what emotion was presented. Our experimental situations differed, however, on a number of potentially important points. Our visual material was still photographs of faces posing particular emotions, and the auditory material consisted of a single sentence which had been recorded by a professional actor in several emotional tones. In addition, each of our experiments involved, in one modality, a continuum of expressions that was combined, in the other modality, with one of two corresponding extreme expressions. Each trial of Experiments 1 and 2 consisted of the presentation of a still face from a morphed continuum extending between happiness and sadness together with the delivery of a sentence pronounced in either a happy or a sad tone. In Experiment 3, a happy-fear voice continuum was used and combined with either one of the two facial expressions. These designs were adapted from earlier studies on bimodal speech perception, in which for example the video presentation of a face saying either /ba/ or /da/ was dubbed on to the delivery of a synthetic syllable from an auditory /ba/–/da/ continuum (Massaro, 1987). The main finding was that the identification curve of the syllable was shifted in the direction of either the /ba/ or the /da/ end of the continuum, depending on the visual stimulus. Similarly, we were encouraged to run the study with still photographs (for which morphed continua were available from a former study by de Gelder et al., 1997d) by the fact that static photographs of faces articulating particular utterances could bias the perception of heard speech sounds (Campbell, 1996).

Our first experiment was run to establish a cross-modal bias of a face and voice on affect identification when instructions specified to combine the two sources. It should confirm that subjects effectively take account of the two information sources. In the two following experiments, we tried to obtain more information regarding the mechanism of the combination by resorting to a focused attention paradigm, in which the subjects were instructed to base their response on one of the two sources only and to ignore the other one.

EXPERIMENT 1

This first experiment was run to determine whether in a bimodal situation in which information about emotional state is provided at the same time by a face and a voice, both information sources can influence recognition. On each bimodal trial, a still photograph of a face was presented on a screen while a voice was heard pronouncing a sentence in one of two tones. The faces were taken from a morphed continuum extending between extreme tokens expressing sadness and happiness, and the tone of the voice was either sad or happy. Participants were asked to indicate, by pressing one of

two keys, whether the person was happy or sad. Any reference to either the visual or the auditory modality was avoided.

Although the question addressed in this experiment is close to the one asked by Massaro and Egan (1996), the experimental situation differed from theirs on several potentially important aspects. As auditory inputs, we used a whole sentence instead of a single monosyllabic word; as visual inputs, we used still photographs of natural faces instead of a moving synthetic face simulating the required expression; we used faces from a continuum of expressions instead of three particular expressions; and as extreme expressions, we used sadness and happiness instead of anger and happiness. Because our experiment and Massaro and Egan's were designed (and run) independently of each other, no test of the effects of any of these differences was planned. Nevertheless, having two sets of results obtained under such contrasted conditions provides useful information concerning the robustness of the bimodal bias phenomenon.

Method

Participants. Sixteen right-handed undergraduates from Tilburg University, eight of each sex, were paid a small amount to participate in one experimental session.

Visual materials. Eleven black-and-white photographs making up a continuum between a sad and a happy expression were used. They were taken from the material used in an earlier study of the categorical perception of facial expression (de Gelder et al., 1997d). The two end-photographs were of a male model from the Ekman and Friesen (1976) series, one posing a sad and the other one a happy expression. Nine intermediate faces were obtained by a morphing procedure developed by Benson and Perrett (1991). Each photograph occupied a 9.5 $\times$ 6.5 cm rectangle on the computer screen, which at the mean viewing distance of 60 cm corresponds to a visual angle of 10.0 $\times$ 6.8 degrees.

Auditory materials. A sentence with an emotionally neutral content (*Zijn vriendin kwam met het vliegtuig* meaning, "His girlfriend came by plane") was spoken by a Dutch professional male actor who had been instructed to pronounce it once "as if he was happy" and the other time "as if he was sad". The sentences were recorded on digital audiotape and acoustic measurements showed that the duration of the happy utterance was 1.78 s, with a mean F_0 of 205 Hz (SD = 39.3); the sad utterance had a duration of 2.12 s, with a mean F_0 of 170 Hz (SD = 19.2). Further acoustic details and analysis of the intonation of the sentences are given in Vroomen et al., (1993).

Design and procedure

Three types of trials were run. On a *visual trial*, one of the 11 photographs was shown for 500 ms, 200 ms after a warning signal, without any auditory accompaniment. On an *auditory trial*, one of the two utterances (the sentence in the sad or in the happy tone) was delivered alone, 200 ms after a warning signal. On a *bimodal trial*, one of the utterances was delivered and 1 of the 11 photographs appeared on the screen at the onset of the last word (*vliegtuig*), and remained until the end of the word (i.e., for about 350 ms). The testing was organised into three blocks of 35 trials, each preceded by a short practice session. Each block consisted of the 22 possible bimodal trials (11 faces $\times$ 2 utterances), 11 visual trials, and 2 auditory trials, presented in a random order.

Participants were tested individually in a quiet experimental room. Each participant was seated at a 60 cm distance from the computer screen. Participants were informed of the different types of trails, which were demonstrated during the practice phase. They were instructed to listen to the voice and watch the screen, and to press as fast as possible one of two response keys, according to whether they felt the person was sad or happy. Similar instructions applied on the few trials where only an auditory stimulus was given. Responses were recorded by the computer and reaction times were measured from the onset of the presentation of the picture (and thus reaction times were not collected for the auditory trials, in which no picture was presented).

Results

On auditory trials, identification of the emotional tone of the voice was 100% correct. Figure 1 shows the percentage of "sad" responses as a function of the location of the face on the visual continuum for the visual trials and the bimodal trails with the "sad" and with the "happy" voice. All three curves rise along the continuum. In comparison with the visual condition, presentation of the "sad" voice increases the percentage of "sad" responses for all faces for which it is not yet at 100% in the visual condition, and presentation of a happy voice decreases it for all faces for which it is not at 0% in that condition.

The data in Figure 1 were submitted to ANOVA with Voice (none, sad, happy) and Face (steps 1–11 on the continuum) as within-subjects factors.[1] The main effect of the Face was significant, $F(10, 150) = 93.38$, $p < .001$, as

[1] We also ran the analyses with Greenhouse–Geisser correction and MANOVA. For Experiments 1 and 2 there was no difference in the sense that all that were significant by ANOVA were also significant by MANOVA. For Experiment 3, ANOVA and MANOVA are identical because the visual condition has only 1 *df*. We therefore report only ANOVA.

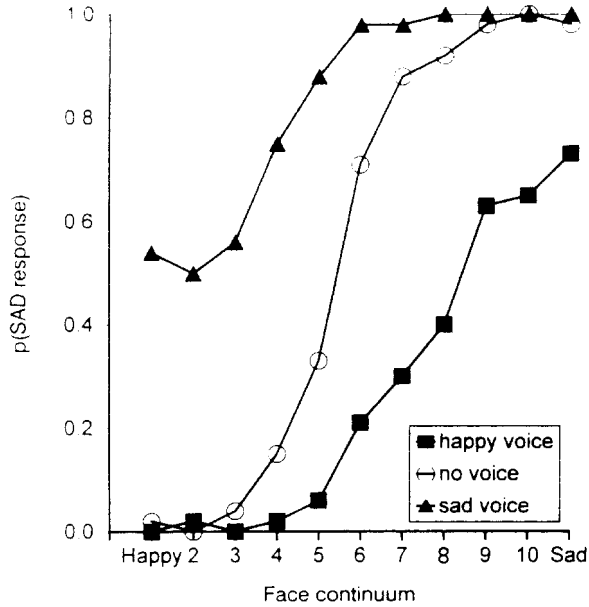

Figure 1. Proportion of "sad" responses as a function of the face continuum when combined with the happy, sad, and no voice.

well as that of the Voice, $F(2, 30) = 94.58$, $p < .001$. The Face × Voice interaction was also significant, $F(20, 300) = 9.24$, $p < .001$.

Reaction times are presented in Figure 2. In the visual condition, the function has an inverted U-shape: as one could expect, the ambiguous faces in the middle of the continuum caused the longest RTs. The bimodal conditions the longest RTs were obtained in the region of the continuum where the expression of the face was different from that of the voice. As argued by Massaro (1987, pp. 73–74) concerning similar results with the McGurk phenomenon, this effect of intermodal difference is probably due not to conflict, *per se*, but rather to the resulting ambiguity of the total input. An ANOVA on the reaction times with Face and Voice as within-subjects variables showed that there were main effects of Face, $F(10, 150) = 6.00$, $p < .001$, Voice, $F(2, 30) = 7.69$, $p < .002$, and a highly significant interaction between these two variables, $F(20, 300) = 15.70$, $p < .001$.

Discussion

When presented with a face and a voice expression, subjects appeared to combine the two sources of information. The combination was manifest in both the identification responses and in the RTs. The same type of

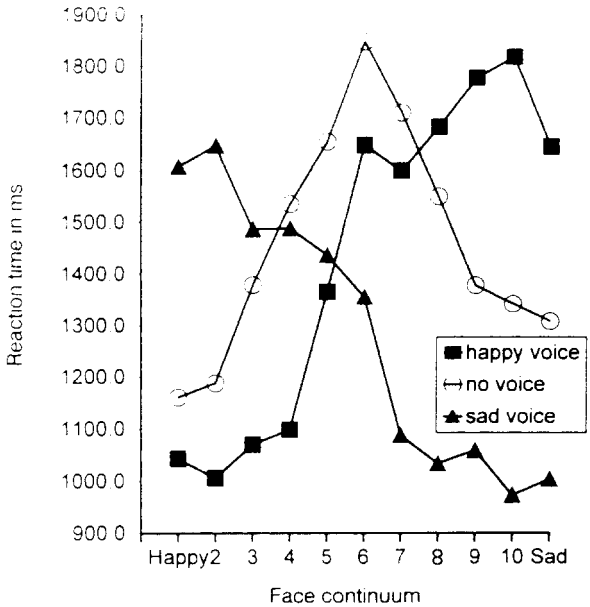

Figure 2. Mean reaction times of the identification responses as a function of the face continuum when combined with the happy, sad, and no voice.

combination of affect relevant information from the voice and the face has been obtained by Massaro and Egan (1996), but with materials quite different from ours. The existence of these differences in the experimental conditions makes the convergence of the results particularly instructive.

The fact that affect integration can be obtained with cross-model pairs as impoverished and distant from the normal ecological situation as a static face and a spoken sentence might suggest that the affect combination process is a particularly powerful one. However, one could also argue that a combination response was requested by the instructions, and that, as a consequence, the observed effect might not be a pure case of perceptual integration, but rather reflect a voluntary effort to obey the instructions. This issue can be addressed with a different version of the experimental manipulation, in which subjects are told explicitly to base their response on the inputs in one of the modalities and ignore those in the other modality. This focused attention to a specific modality is standard in research which considers conflicting inputs, like studies on ventriloquism (Bertelson, 1998). If there is an effect of the input channel the subject has been instructed to ignore, it suggests the existence of an automatic or mandatory process.

EXPERIMENT 2

Participants were presented the same stimuli as in Experiment 1, but the instructions were now to judge the face and ignore the voice.

Method

A new group of 16 right-handed undergraduates (8 of either sex), none of whom had participated in the previous experiment, was tested. The same unimodal and bimodal stimuli as in Experiment 1 served as materials. Subjects were told to listen to the voice and to watch the screen. The instructions emphasised that their task was to judge whether the face was angry or sad, and to ignore the expression in the voice.

Results

As can be seen in Figure 3, the "happy" sentence shifted the identification function towards the "happy" end of the continuum, and the "sad" utterance shifted it to the "sad" end. The data in this figure were submitted to an ANOVA with Voice (none, sad, happy) and Face (steps 1–11 on the continuum) as within-subjects factors. The main effect of Face was

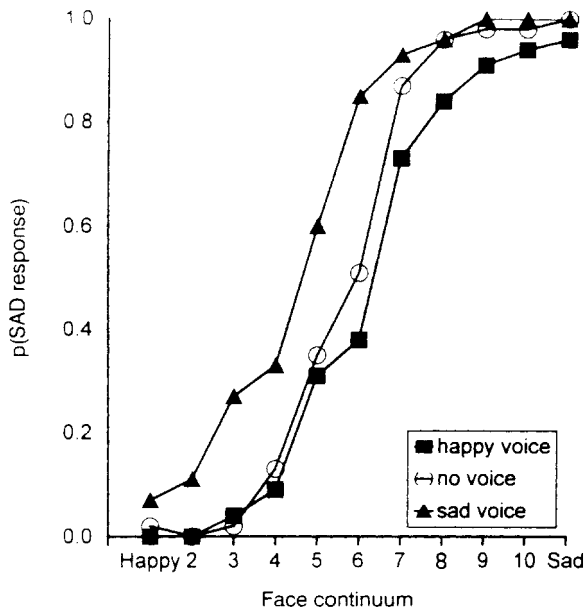

Figure 3. Proportion of "sad" responses as a function of the face continuum when combined with the happy, sad, and no voice.

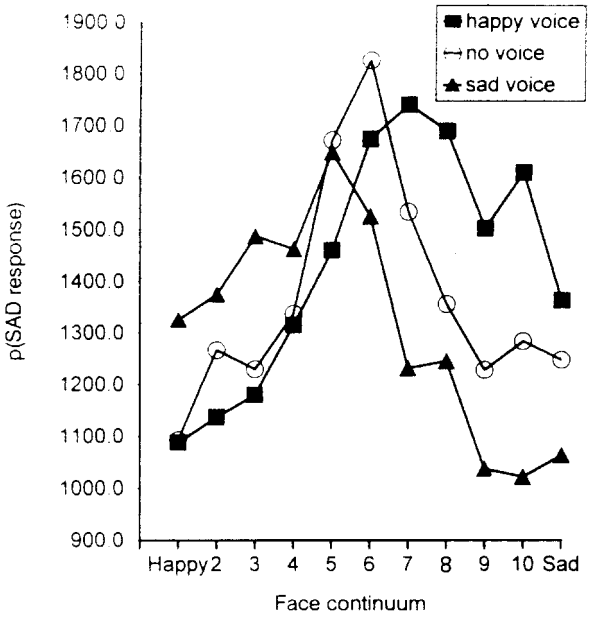

Figure 4. Mean reaction times of the identification responses as a function of the face continuum when combined with the happy, sad, and no voice.

significant, $F(10, 140) = 144.57$, $p < .001$. The effect of the Voice was also significant, $F(2, 28) = 7.96$, $p < .002$, as was the Face × Voice interaction, $F(20, 280) = 3.13$, $p < .001$. The interaction reflects the fact that the effect of the voice was larger at the ambiguous levels of the face continuum. When the two auditory sentences were presented in isolation they were recognised 99% correct.

The RTs are shown in Figure 4. In the no-voice condition, the same inverted U-shape function was obtained as in Experiment 1. But in the bimodal conditions, the functions were also U-shaped with peaks each displaced toward end corresponding to the emotion conveyed by the voice. Reaction times on the ends of the continuum were again slower with incongruent voice tones than with congruent ones. In the ANOVA, the main effect of Face was significant, $F(10, 140) = 10.95$, $p < .001$, that of the Voice was nonsignificant, $F(2, 28) = 2.35$, $p = .11$, and the interaction between the two variables was significant, $F(20, 280) = 5.64$, $p < .001$.

To determine whether the different instructions of Experiments 1 and 2 had an effect, we ran an overall ANOVA on the identification responses. Of interest was a significant second-order interaction between Experiment, Face, and Voice, $F(20, 580) = 2.98$, $p < .001$. The interaction signalled, as is also apparent in Figures 1 and 3, that the effect of the voice was

smaller in Experiment 2 than in Experiment 1. As a more direct test of the cross-modal effect, we computed, for each participant, the impact of the voice by subtracting the proportion of "sad" responses when a face was combined with a "happy" voice from a "sad" voice. The average mean difference between the proportion of "sad" responses with the "sad" and the "happy" voice was .56 in Experiment 1 and .17 in Experiment 2, which was significant in t-test, $t(29) = 5.20$, $p < .001$.

Discussion

Experiment 2 showed that, if compared with Experiment 1, the size of the cross-modal bias was smaller in the case that participants focused their attention on the face. This difference suggests that at least some part of the cross-modal effects observed in Experiment 1 were due to some voluntary effort to obey the integration instructions. Nevertheless, participants were still affected by the voice despite being instructed to ignore it and to focus on the face. With bimodal presentations, the identification function shifted in the direction of the emotion in the voice, and RTs were systematically slower for incongruent than for congruent trials. These findings therefore suggest that there is also a mandatory cross-modal interaction.

Given that there is some evidence for a mandatory influence of voice tone on judgements of facial expression, it remains to be determined if the reverse influence, from face to voice, can be demonstrated. The following experiment was run to answer that question.

EXPERIMENT 3

In this experiment, participants had to judge a voice taken from a continuum of voice tones. They were instructed to ignore a simultaneously presented face expressing one of the two end emotions of the voice continuum. Creating a continuum of voice tones posed some technical problems to be described in the Method section, which were more difficult to solve for some emotions than for other ones. In particular, we could not develop a happy-sad continuum that would have been the natural counterpart of the face-continuum used in Experiments 1 and 2, and had to use a more easily obtained continuum extending from happiness to fear.

Method

Participants. A new group of 12 right-handed undergraduates (6 of each sex) was tested. They received course credits for their participation.

Auditory materials. Preparation of the auditory stimuli started with the recording of two natural tokens of an actor pronouncing the same emotionally neutral sentence (*Zijn vriendin komt met het vliegtuig*; "His girlfriend is coming by plane") as in Experiments 1 and 2. The actor was instructed to pronounce the sentence once in a happy tone and another time in a fearful tone. His task was clarified by indicating prototypical circumstances for each emotion. The two emotions were chosen because their particular intonation patterns made it possible to create a continuum by changing simultaneously the duration, pitch range, and pitch register of one of the utterances (see Figure 5). This was achieved as follows. The utterance that expressed happiness served as the "source", and its duration, pitch range, and pitch register was shifted towards that of fear in a 7-step continuum. In order to change the pitch in equal steps, the original pitch contour was replaced by a minimal sequence of straight line approximations while the perceptual identity remained close to the original one.

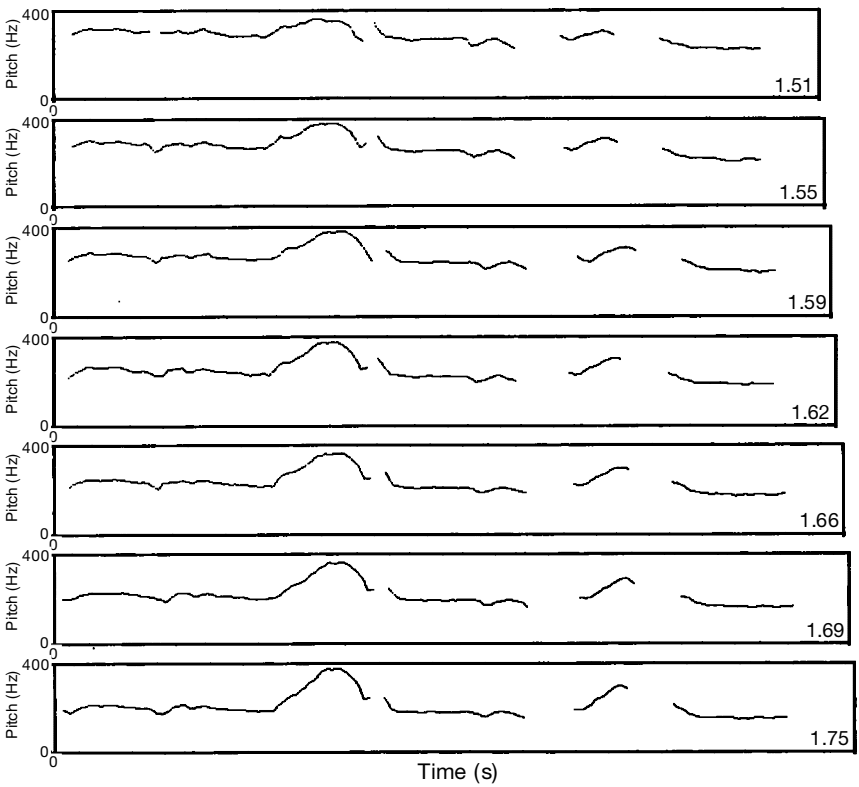

Figure 5. Pitch contours of the happy-fear voice continuum. The upper panel is the "fear" end of the continuum, the lowest panel is the "happy" end.

This artificial contour was generated by software (Zelle, de Pijper, & 't Hart, 1984) which takes into account the grammar of Dutch intonation. After marking the onset of the to-be-accented vowels, the program computes the various pitch movements by superimposing them on a declination line. Then, only two free parameters need to be set: The excursion size of the pitch movements in semitones and the end frequency of the utterance in hertz (Hz). The latter parameter determines the place in the pitch register. For the "happy" endpoint of the continuum, the excursion size was set at 10 semitones and the end-frequency at 150 Hz. For each successive stimulus in the continuum, the excursion size was decreased with 1 semitone and the end frequency was increased with 12 Hz. Thus, the seventh stimulus at the fear endpoint had an excursion size of 4 semitones and an end-frequency of 222 Hz. Finally, the duration of the obtained utterances was linearly compressed. The duration of the utterance at the happy endpoint was left at 100% (i.e., 1.58 s) and the duration of each next stimulus in the continuum was decreased with 2% so that duration at the fear endpoint was 88% (i.e., 1.39 s). All pitch and time manipulations were based on direct waveform manipulations (PSOLA, Charpentier & Moulines, 1989) so that the tokens sounded natural.

Visual materials. The visual stimuli consisted of two photographs, a happy one and a fearful one of the same male actor who pronounced the utterances. The faces (6 × 11 cm) were positioned in a frame (23 × 16 cm) and were shown on a black-and-white PC screen from a distance of approximately 60 cm. The face of a female actor with a neutral expression was used for catch trials.

Design and procedure. The experiment consisted of 70 experimental trials (5 repetitions of the 14 combinations: 2 Faces × 7 Voices). The instructions were to judge the emotion in the voice by pressing one of two keys labelled "happy" or "fear" while ignoring the emotion in the face. In addition, 25 catch trials were interpolated to ensure that subjects were looking at the screen while the auditory sentence was played. On catch trials, a female face was shown instead of the male one. When a female face appeared, participants were not to respond. All stimuli were presented in two pseudorandomly ordered blocks.

The auditory stimuli were played directly from the hard disk and presented at a comfortable listening level over headphones. Presentation of the face started 300 ms after the onset of the utterance and lasted until the end. Given the presence of catch trials, subjects were asked not to respond before the face was seen. Reaction time was measured from voice onset. The ITI was 2 s, and before testing proper started, there was a short practice session. Subjects were instructed to decide whether the voice

expressed "fear" (left button) or "happiness" (right button). They were told to base their judgement on the voice only and to ignore the face.

Results

The identification responses are presented in Figure 6. A 2 (Face) × 7 (Voice) ANOVA was performed on the proportion of "fear" responses with Voice and Face as within-subjects variables. The number of "fear" responses increased as the voice changed from the "happy" toward the "fear" end of the continuum, $F(6, 60) = 42.81$, $p < .001$. The effect of the Face was significant, $F(1, 10) = 5.57$, $p < .04$. The Face × Voice interaction was not significant, $F < 1$.

The RTs are presented in Figure 7, where they are plotted as a function of voice tone, separately for the two faces. The two curves were practically superposed, indicating that, unlike Experiment 2, RTs were not affected by the expression of the faces. In the two conditions, the voice tones close to the "happy" end of the continuum were identified faster than those close to the "fear" endpoint. In the ANOVA, the effect of the Voice was significant, $F(6, 60) = 8.96$, $p < .001$, whereas the main effect of the Face, $F < 1$, as well as the Face × Voice interaction, $F < 1$, were not significant.

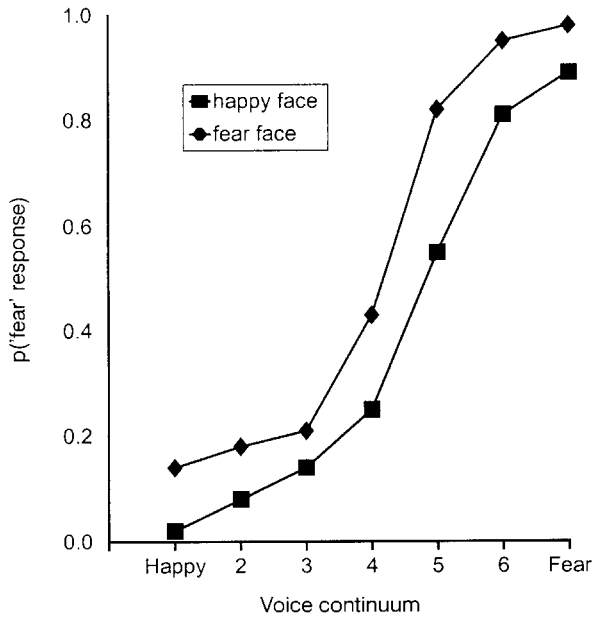

Figure 6. Proportion of "afraid" responses as a function of the voice continuum when combined with the happy or afraid face.

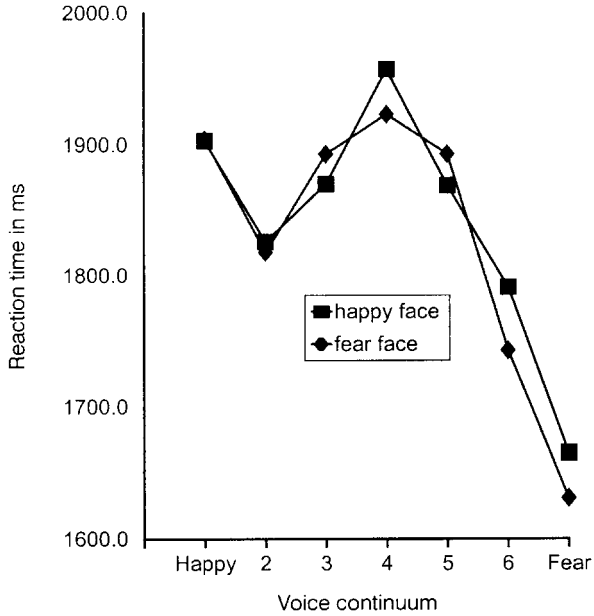

Figure 7. Mean reaction times of the identification responses as a function of the voice continuum combined with the happy or afraid face.

Discussion

Following the finding that perception of facial expression is greatly affected by concurrent information from the voice, we asked whether the converse would also hold. We examined the issue by using a situation similar to the previous one, in which the task was to judge the expression in the voice while ignoring the information concurrently conveyed by the face. A clear effect of the to be ignored information from the face was obtained at the level of identification responses, but not at the level of RTs.

The absence of an effect on RTs would seem to make the results of the present experiment different from those of Experiments 1 and 2, where the voice affected both the identification responses and their latencies. It must be emphasised, though, that in the present case, RTs were measured from the start of the utterance which lasted more than 1 s. This is very different from the case where RTs are measured to the visual presentation of a face in which all the information becomes available at once. It therefore seems that there is no necessity that measures made in such different situations show the same sensitivity to a particular effect, in this case the cross-modal bias.

Rather, the important result of Experiment 3 is that at the level of the dependent variable that is comparable across experiments (i.e., the identi-

fication responses, a significant bias by a facial expression has been demonstrated). That means that cross-modal biases between voice and face expressions are to a large extent bidirectional.

GENERAL DISCUSSION

The main objective of the present study was to explore the combination of information from facial expression and voice tone in the recognition of emotion. Experiment 1 showed that subjects can combine data from the two sources to arrive at a unique judgement. That result, however, did not mean that the combination is mandatory. It could simply be a response to the instructions which implied that the face and the voice be considered as belonging to the same person. The question of the mandatory character was examined in the two following experiments, using a focused attention paradigm. Experiment 2 showed that when subjects were asked to identify the expression of a face while ignoring a simultaneously heard voice, their choices were nevertheless influenced by the tone of the voice. Experiment 3 demonstrated the converse phenomenon. When asked to identify the tone of a voice while ignoring a simultaneously presented face, subjects were influenced by the expression in the face.

Our interpretation of the present cross-modal effect is that it presents a perceptual phenomenon reflecting the mandatory integration of inputs and not a post-perceptual decision under attentional control. Bimodal emotion perception concerns a single event presented in two different modalities. The finding that integration still occurs when subjects are instructed to ignore input in the other modality seems to speak against the idea of a post-perceptual conflict. In the latter case integration of heard and seen emotions would follow directly from a decision taken by the perceiver after both inputs were processed, whether or not to put the two kinds of information together and what final judgement to come up with. Moreover, the perceptual integration we have shown here is particularly striking because it was observed in such an impoverished situation as listening to a spoken sentence with watching a still photograph of a face. That situation is of course very different from the familiar social experience of seeing the moving face of a person while hearing what she/he says and having available a rich interpretative context.

A behavioural approach to perception like the one pursued here does shed light on the actual processing involved in bimodal perception of emotion and not just on the subjects' ability to perform a forced-choice identification of the stimuli (see also Massaro, 1998). However, one cannot address all aspects of the question of bimodal integration. As argued, for example by Stein and Meredith (1993), an answer to the question on the mechanism of audiovisual interactions and on the time course of integra-

tion requires converging evidence from other methods. It might be the case that audiovisual pairings like those observed here are a direct consequence of cells tuned to receive bimodal input. Also, specialised areas in the brain might be dedicated to cross-modal integration of auditory and visual input when there is a component of valence to the integration as argued for the amygdala (Nahm, Tranel, Damasio, & Damasio, 1993). Other researchers have pointed to the role of the basal ganglia in providing a context for processing stimuli with affective content (see LeDoux, 1996 for overviews), or to cortico-cortical connectivity (de Gelder et al., 1997b). Electrophysiological methods like recordings of event-related potentials (ERPs) are specifically appropriate to answer one aspect of the theoretical questions, as they allow us to trace the time course of cross-modal bias. In a recent study we investigated this issue by looking at scalp potentials while presenting audiovisual pairs (i.e., a sentence fragment combined with a still face expressing the same or a different emotion; de Gelder, Böcker, Tuomainen, Hensen, & Vroomen, 1999). The fact that the concurrent presentation of an incongruent facial expression had an impact on this auditory potential strongly suggests that the perceptual system integrated the two modalities early on.

Further support for the notion that the integration of the two input channels is mandatory is contained in evidence of the limited role of awareness in bimodal perception. This evidence comes from normal subjects as well as from brain-damaged patients. It must be noted that introspective reports obtained from participants testify to some degree awareness of the inconsistency between voice and facial expression. In some studies of the McGurk effect subjects were similarly aware of the cross-modal syllable discrepancy (Summerfield & McGrath, 1984; Rosenblum & Saldana, 1996). In the study by Green, Kuhl, and Meltzoff (1991) subjects were aware of the discrepancy of the gender of the face and the voice. Such awareness of inconsistency did not, however, overrule that the processing system combined the two sources and that integration took place.

Another valuable source of information about the limited role of awareness for cross-modal integration is provided by the performance of patients with focal brain lesions. We tested a prosopagnosic patient who could no longer recognise facial expressions. When tested with the materials and instructions of Experiment 2, her judgements of the face were entirely under the influence of the concurrently presented voice. But when the unrecognised facial expressions were paired with voice expressions and the task was to judge the voice (like in our Experiment 3), the face had an effect on the judgement of the voice, just as is the case in the normals reported here (de Gelder, Pourtois, Vroomen, & Bachoud-Levi, in press). A speculative explanation for this effect would be that the facial expression

is processed by the dorsal route and can take place in the absence of recognition routes relying on the ventral system and sustaining explicit recognition processes. On this view, combination of voice and face expressions would be early and automatic, a suggestion supported by the ERP data mentioned earlier. Similar evidence of nonconscious processing of facial expressions was also provided in our study of a patient, GY, who lost vision in his right visual field as a consequence of brain damage to his entire left striate cortex and showing 'blindsight' or correct visual discrimination in the absence of stimulus awareness (see Weiskrantz, 1997). When presented with facial expressions in his bad field and prompted to guess what expression is shown, this patient presents a highly reliable correct performance (de Gelder, Vroomen, & Weiskrantz, 1999).

The limited role of awareness in perceiving and integrating affective information suggests that we are dealing with a level of emotional processes that is more primitive than the one addressed in many studies of emotion, whether face or voice recognition, in which the emphasis is on the meaning of the stimuli and the way this meaning is accessed and constructed in a rich social context. The scope of this study was limited to only a few fairly noncontroversial emotions and no effort was made to address the many questions that surround the distinction between basic versus complex or blended emotions (see Frijda, 1989, for overview and critical discussion). Likewise, our study did not touch on the relation between processing emotional signals as tapped in the present study of audiovisual integration and the full emotional experience.

Whether or not voice and face expression recognition are organised around primitives or basic emotions, and whether or not these are the same in the two cases are questions for future research. We noted in the introduction that the idea of common amodal structures mediating perceptual processes in the different sensory systems in which emotions are expressed is often taken for granted. Some studies have argued for the existence at a functional level of a specialised emotion processor, or emotion module to be conceived along the lines of cognitive modules (Fodor, 1983) dealing with affective cognition. However, those suggestions were only based on results from facial emotion recognition (Etcoff & McGee, 1992). In the present state of our knowledge, various options must remain open. There are more events that signal potentially relevant affect information in our environment than just the movements of the human face. For example, would the perceptual system put together an environmental sound (e.g., an alarm bell) with a face expression? Does it combine written messages with voice expression? Or is audiovisual emotion perception really more like another case of the McGurk illusion in the sense that it only operates over inputs provided by the sights and the sounds of the face?

The question of the content sensitivity of intermodal pairing is critical for understanding the nature of audiovisual emotion perception and it is difficult for this to be settled by models that are intended to fit any situation where two sources of information are present (Massaro, 1987, 1998; Massaro & Egan, 1996). Such models deal with integration as only a quantitative issue, not considering the possibility of constraints from content-specificity on bimodal pairing. Is there a content-based component to audiovisual pairings above and beyond the mechanics of mandatory pairing? Research on bimodal integration presents a good tool for mapping the domain and the competence of what a possibly specialised emotion processor is.

Manuscript received 30 March 1998
Revised manuscript received 10 September 1999

REFERENCES

Anderson, A.K., & Phelps, E.A. (1998). Intact recognition of vocal expressions of fear following bilateral lesions of the human amygdala. *Neuroreport*, *9*, 3607–3613.

Benson, P.J., & Perrett, D. (1991). Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, *3*, 105–135.

Bertelson, P. (1998). Starting from the ventriloquism: The perception of multimodal events. In M. Sabourin, F.I.M. Craik, & M. Roberts (Eds.), *Advances in pyschological science: Vol. 1. Biological and cognitive aspects*. Hove, UK: Psychology Press.

Bugenthal, D.E., Kaswan, J.W., Love, L.R., & Fox, M.N. (1970). Child versus adult perception of evaluative messages in verbal, vocal, and visual channels. *Developmental Psychology*, *2*, 367–375.

Campbell, R. (1996). Seeing speech in space and time. *Proceedings of the International Conference on Spoken Language Processing*, pp. 1493–1498.

Carlson, R., Granström, B., & Nord, L. (1992). Experiments with emotive speech: acted utterances and synthesized replicas. *Proceedings of the International Congress of Speech and Language Processing*, Banff, 671–674.

Charpentier, F., & Moulines, E. (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis. *Proceedings of Eurospeech*, *'89*, 13–19.

Cosmides, L. (1983). Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, *9*, 864–881.

Cummings, K.E., & Clements, M.A. (1995). Analysis of the glottal excitation of emotionally styled and stressed speech. *Journal of the Acoustical Society of America*, *98*, 88–98.

de Gelder, B., Bachoud-Levi, A., & Vroomen, J. (1997a). Emotion by ear and by eye: Implicit processing of emotion using a cross-modal approach. *Proceedings of the Fourth Annual Meeting of the Cognitive Neuroscience Soceity*, No. 49, 73.

de Gelder, 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.

de Gelder, B., Bovet, P., Parnas, J., Vroomen, J., Popelier, T., & Innocenti, G. (1997b). Impaired integration of audition and vision in schizophrenics. *Experimental Brain Research*, *117*, 22.

de Gelder, B., Milders, M., van Deursen, M., Hansen, M., Vroomen, J., Haaxma, R., & Rouw, R. (1997, January). *Impaired perception of expression in voice and face contrasts with evidence of normal cross-modlal bias*. Poster session presented at the XVth European Workshop on Cognitive Neuropsychology, Bressanone, Italy.

de Gelder, B., Teunisse, J-P., & Benson, P.J. (1997d). Categorical perception of facial expressions: Categories and their internal structure. *Cognition and Emotion*, *11*, 1–23.

de Gelder, B., Vroomen, J., & Bertelson, P. (1998). Upright but not inverted faces modify the perception of emotion in the voice. *Cahiers de Psychologie Cognitive*, *17*, 1021–1032.

de Gelder, B., Vroomen, J., & Weiskrantz, L. (1999). Covert processing of facial expressions in a blindsight patient. *Neuroreport*, *10*.

de Gelder, B., Pourtois, G., Vroomen, J., & Bachoud-Levi, A.C. (in press). Covert processing of facial expressions in prosopagnosia. *Brain and Cognition*.

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

Ekman, P., & Friesen, W.V. (1976). Measuring facial movement. *Journal of Environmental Psychology and Nonverbal Behavior*, *1*, 56–75.

Etcoff, N., Freeman, R., & Cave, K.R. (1991). Can we lose memories of faces? Content specificity and awareness in prosopagnosia. *Journal of Cognitive Neuroscience*, *3*, 25–41.

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

Fairbanks, G., & Pronovost, W. (1939). An experimental study of the pitch characteristics of the voice during the expression of emotion. *Speech Monographs*, *6*, 87–104.

Fodor, J.A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Frick, R.W. (1985). Communicating emotion: the role of prosodic features. *Psychological Bulletin*, *97*, 412–429.

Frijda, N. (1989). *The emotions*. Cambridge: CUP.

Green, K., Kuhl, P., & Meltzoff, A. (1991). Integrating speech information across talkers, gender, and sensory modality: female faces and male voices in the McGurk effect. *Perception and Psychophysics*, *50*, 524–536.

Hess, U., Kappas, A., & Scherer, K. (1988). Multichannel communication of emotion: Synthetic signal production. In Scherer, K. (Ed.), *Facets of emotion: Recent research* (pp. 161–182). Hillsdale, NJ: Erlbaum.

Kavanagh, J.F., & Mattingly, I.G. (1972). *Language by ear and by eye: The relationship between speech and reading*. Cambridge, MA: MIT Press.

Ladd, D.R., Silverman, K.E.A., Tolkmitt, F., Bergman, G., & Scherer, K.R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signalling speaker affect. *Journal of the Acoustical Society of America*, *78*, 435–444.

LeDoux, J.E. (1996). *The emotional brain*. New York: Simon & Shuster.

Lieberman, P., & Michaels, S.B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to emotional content of speech. *Journal of the Acoustical Society of America*, *34*, 922–927.

McKelvie, S.J. (1995). Emotional expression in upside-down faces: Evidence for configurational and componential processing. *British Journal of Social Psychology*, *34*, 325–334.

Massaro, D.W. (1987). *Speech perception by ear and eye*. Hillsdale, NJ: Erlbaum.

Massaro, D.W. (1998). *Talking heads*. Cambridge, MA: MIT Press.

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

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

Mehrabian, A., & Ferris, S.R. (1967). Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, *31*, 248–252.

Murray, I.R., & Arnott, J.L. (1993). Toward the simulation of emotion in synthetic speech: A

review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, *93*, 1097–1108.

Nahm, F.D.D., Tranel, D., Damasio, H., & Damasio, A.R. (1993). Cross-modal associations and the human amygdala. *Neuropsychologia*, *31*, 727–744.

Pomerantz, J.R. (1981). Perceptual organization in information processing. In M. Kubovy & J.R. Pomerantz (Eds.), *Perceptual Organization*, Hillsdale, NJ: Erlbaum.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived stress. *Journal of the Acoustical Society of America*, *101*, 2267–2277.

Rhodes, G., Brake, S., & Atkinson, A.P. (1993). What's lost in inverted faces? *Cognition*, *47*, 25–57.

Rosenblum, L.D., & Saldana, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318–331.

Scherer, K.R. (1986). Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, *99*, 143–165.

Scherer, K.R. (1989). Vocal measurement of emotion. In R. Plutchik and H. Kellerman (Eds.), *Emotion: theory, research, and experience* (Vol. 4, pp. 233–259). San Diego, CA: Academic Press.

Scott, S., Young, A., Calder, A., Hellawell, D., Aggleton, J., & Johnson, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, *385*, 254–255.

Stein, B.E., & Meredith, M.A. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Summerfield, A.Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, *36A*, 51–74.

Tanaka, J.W., & Farah, M.J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology* (*A*), *46*, 225–245.

Tartter, V., & Braun, D. (1994). Hearing smiles and frowns in normal and whisper registers. *Journal of the Acoustical Society of America*, *96*, 2101–2107.

Van Lancker, D. (1997). Rags to riches: Our increasing appreciation of cognitive and communicative abilities of the human right hemisphere. *Brain and Language*, *57*, 1–11.

Vroomen, J., Collier, R., & Mozziconacci, S. (1993). Duration and intonation in emotional speech. *Proceedings of the Third European Conference on Speech Communication and Technology*, *Berlin*, 577–580. Berlin: ESCA.

Walker, A., & Grolnick, W. (1983). Discrimination of vocal expressions by young infants. *Infant Behavior and Development*, *6*, 491–498.

Weiskrantz, L. (1997). *Consciousness lost and found: a neuropsychological exploration*. New York: Oxford University Press.

Williams, C.E., & Stevens, K.N. (1972). Emotions and speech: Some acoustical factors. *Journal of the Acoustical Society of America*, *52*, 1238–1250.

Yin, R.K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*, 141–145.

Zelle, H.W., de Pijper, J.R., & 't Hart, J. (1984). Semi-automatic synthesis of intonation for Dutch and British English. *Proceedings of the 10th International Congress of Phonetic Sciences*, *Utrecht*, *IIB*, 247–251.