Cahiers de Psychologie Cognitive/
Current Psychology of Cognition
1998, 17 (4-5), 1021-1031 (special issue)

Upright but not inverted faces modify the perception of emotion in the voice

Beatrice de Gelder, Jean Vroomen, and Paul Bertelson

Tilburg University (The Netherlands) and
Université Libre de Bruxelles (Belgium)

Abstract. Previous work has demonstrated the existence of automatic mutual crossmodal biases between emotions conveyed by seen faces and heard sentences. The goal of the present study was to determine whether, in the particular case of the influence of facial expression on the identification of expressive voice tone, the effect survives face inversion, a manipulation known to affect the recognition of facial expression as well as that of personal identity. Young adults were given the task of categorizing as either happy or fearful the tone of voice of sentences from a continuum extending in seven steps from happiness to fear. The sentences were presented alone or together with a still photograph of a face expressing happiness or fear, and either upright or upside-down. For the bimodal trials, the subjects were instructed to base their judgement on the voice only and to ignore the expression conveyed by the face. Judgements of voice tone were found to be biased in the direction of the expression of the face when the latter was shown upright, not when it was shown upside-down. These results are discussed in relation to the mechanisms subtending the multimodal recognition of emotion.

Key words: Crossmodal interaction, emotion perception, face expression, voice tone.

Correspondence should be sent to Beatrice de Gelder, Tilburg University, PO Box 90153, 5000 Le Tilburg, The Netherlands (e-mail: degelder@kub.nl).

INTRODUCTION

Emotions are manifested in a variety of behaviors like face expressions, voice expressions, and gait. But the overwhelming majority of the studies on the perception of emotion have concentrated on facial expression, whether they have addressed its normal processing (Ekman & Friesen, 1975), its development (e.g., Diamond & Carey, 1986), or deficits following neurological impairment (Damasio, Tranel, & Damasio, 1990; Scott, Young, Calder, Hellawell, Aggleton, & Johnson, 1997). A recent addition to this line of work, made possible by advances in image synthesis, is the use of stimulus continua, which permits addressing the question of the categorical perception of emotions (Calder, Young, Perrett, Etcoff, et al., 1996; Etcoff & McGee, 1992; de Gelder, Teunisse, & Benson, 1997). But conclusions from these studies must in fact be restricted to how emotions are perceived in the face, and do not extend to the question of general emotion categories per se.

The understanding of emotions in the voice has progressed more slowly, and very few studies are available (de Gelder & Vroomen, 1995; see Vroomen, Collier, & Mozziconacci, 1993, for an overview). In fact, a major theoretical stumble-block for such studies concerns the phonetic realization of affective prosody.

Researchers have often tacitly assumed the existence of an underlying functional-neurological system common to the perception of emotions in the voice and in the face. But the fact that emotion is perceived in the visual as well as in the auditory modality raises a host of empirical questions. In the few studies that have looked at voice expressions the goal has generally been to obtain converging evidence for a deficit that would run parallel to disorders in the perception of facial expressions (van Lancker, 1997). Thus, conclusions about a common processing stucture have been based on the association of deficits in the visual and auditory modalities. The problem with those studies is that the evidence is only indirect and correlational. It does not allow for any conclusions about the combining of information from the voice and the face by the processing system, and even less, about whether there exists a shared underlying representation system for emotions.

Only a handful of studies have looked directly at relations between input modalities. In one of these, Tartter and Braun (1994) showed that subjects listening to spoken syllables could infer the facial expression of the speaker pronouncing them. This finding is intriguing, among other reasons because it suggests a natural link in production between what appears from the outside as two separate channels, the tone of the voice and the expression on the face. Massaro and Egan (1996) came the closest to addressing the issue of bimodal perception in experiments that used a synthetic face presenting two different expressions, combined with a word spoken in either of two emotional tones. Their results clearly suggest that the processing system samples information from both the auditory and the visual sources. However, one can wonder whether synthetic facial expressions involve exactly the same underlying neurological basis as do natural faces. Likewise, it is a matter of debate what the optimal acoustic and phonetic carrier of an emotional message is in spoken language. The emotional validity of a single word is unlikely to be as strong as that of a full sentence.

Pursuing this issue, de Gelder, Vroomen, and Teunisse (1995) and de Gelder and Vroomen (1998) reported a study in which static photographs of natural faces with varying expressions were presented in combination with a full sentence pronounced in one of several tones. In two of the experiments, faces with expressions from a continuum extending from "angry" at one end to "sad" at the other end were presented either alone or together with a sentence pronounced in either an angry or a sad tone. In comparison with unimodal visual presentations, the identification function of facial expressions in the bimodal presentations was shifted in the direction of the expression conveyed by the voice. This crossmodal influence of the voice on judgements of facial expression was still observed when the subjects were told to ignore the voice. Basically the same pattern of results was obtained in a further experiment in which the task was to judge the emotion conveyed by voices from a continuum presented together with a face expressing one of the two extreme emotions. From the analysis of response frequencies and latencies it was concluded that the processing system combines affect-relevant information from simultaneously presented faces and voices, and that this combination takes place in the course of perceptual processing rather than post-perceptually. That conclusion was supported in a further study which did not use any stimulus continuum, but only a set of different facial expressions combined with different voice tones (de Gelder, Böcker, Tuomainen, Hensen, Vroomen, & Bertelson, 1998, Exp. 1).

The present paper reports a first step in exploring what properties of the face contribute to the obtained bias in the perception of voice tone.

For this purpose, we contrasted upright and inverted face presentation. The effects of up-down face inversion has attracted much attention in the literature on face identification (see Valentine, 1988, for a review). Starting with Yin's (1969) findings, the fact that inversion results in a stronger reduction of identification performance for faces than for other objects has been taken as evidence for a special status of faces. The general notion has been that the recognition of upright faces takes advantage of configural attributes which are not available in inverted faces.

Is the configuration also important in extracting emotion from the face? This question has received less attention, no doubt because very little is known about how the different regions of the face convey expression: do they operate in concert, or is it the facial configuration that is the bearer of the expression? The well-known phenomenon called *the Thatcher illusion* (Thompson, 1980) suggests the latter. Inverting the eyes and the mouth in an upright face causes a perception of gruesome expression. But when such a face is presented upside-down, the unpleasant impression tends to disappear. The inverted stimulus does not prompt recognition of a familiar person, but there is nothing particularly striking about the expression. More directly relevant evidence was obtained by McKelvie (1995) who had subjects judge the emotion expressed by faces presented upright or upside-down. Whereas upright faces were identified with a high degree of accuracy, matching classical data by Ekman and Friesen (1975), performance with inverted faces varied considerably across emotions, from substantial for happiness and surprise, to very low for emotions like sadness, fear, and anger. Related evidence was provided by de Gelder et al. (1997) in a study in which stimuli from several facial expression continua were presented either upright or upside-down. Inversion made discrimination along an angry-sad continuum impossible. For two other continua, happy-sad and angry-fearful, non-random discrimination was observed in spite of the inversion. However, all evidence of categorical processing obtained for upright faces disappeared with inversion.

The present study compares the effects on judgements of the emotional tone in which a sentence is pronounced of the expression conveyed by a face presented simultaneously in either the upright or the upside-down orientation.

METHOD

Participants

Twelve right-handed first-year students from Tilburg University, 6 female and 6 male, aged 19 to 26 years, were tested. They received course credit for their participation.

Auditory materials

Preparation of the auditory stimuli started with the recording of two natural tokens of an actor pronouncing a short sentence with hopefully (but the point is by no means essential) semantically neutral content ("Zijn vriendin komt met het vliegtuig", meaning "His girlfriend is coming by plane"). The actor was instructed to pronounce the utterance once in a happy tone and another time in a fearful tone, and prototypical circumstances for such expressions were indicated. These two emotions were chosen because their intonation patterns made it possible to create a continuum by simultaneously changing the duration, the pitch range, and the pitch register of one of the utterances.

The continuum was obtained using the following procedure. The utterance that expressed happiness served as the 'source', and its duration, pitch range, and pitch register were shifted towards that of fear in seven steps. 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. This artificial contour was generated by software (Zelle, de Pijper, & 't Hart, 1984) which takes into account the grammar of Dutch intonation. After the onset of the to-be-accented vowels was marked, the program computed the various pitch movements by superimposing them on a declination line. Then, only two free parameters needed to be set: the excursion size of the pitch movements (in semitones) and the end frequency of the utterance (in 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 of the following stimuli in the continuum, the excursion size was decreased by 1 semitone and the end frequency was increased by 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 utterances was linearly compressed. The duration of the utterance at the happy endpoint was left at 1.58 sec, and the duration of each successive stimulus on the continuum was decreased by 2%, so that the duration at the fear endpoint matched the natural value of 1.39 sec. All pitch and time manipulations were based on direct waveform manipulations (PSOLA, Charpentier & Moulines, 1989) in such a way that the tokens always sounded natural.

Visual materials

The visual stimuli consisted of two facial expressions of the same actor, a happy one and a fearful one. The faces were presented upright or upside-down. 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.

Design and procedure

Subjects were tested individually in a soundproof booth. The experiment consisted of 140 bimodal trials, i.e., 5 repetitions of the 28 combinations of 2 faces × 2 orientations × 7 voice tones, presented randomly in two blocks. The auditory stimuli were played directly from hard disk and presented at a comfortable listening level over headphones. The voice and the face were presented simultaneously for the duration of the utterance. The offset-to-onset ITI was 2 sec, and before testing started, subjects were given a short practice session.

Subjects were instructed to decide whether the voice was fearful or happy. They were asked to base their judgement on the voice only and to ignore the face.

RESULTS

The percentages of "fearful" responses to the different voices in the continuum, combined with the happy and the fearful faces, respectively, are shown in Figure 1 for upright faces and in Figure 2 for inverted faces. Under each of the four conditions, the proportion of fearful responses increased as voice tone changed from the happy toward the fearful end of the continuum. With upright faces, there was a strong effect of facial expression, manifested by a systematic separation between the curve for the fearful face and that for the happy face. With inverted faces, the separation disappeared.

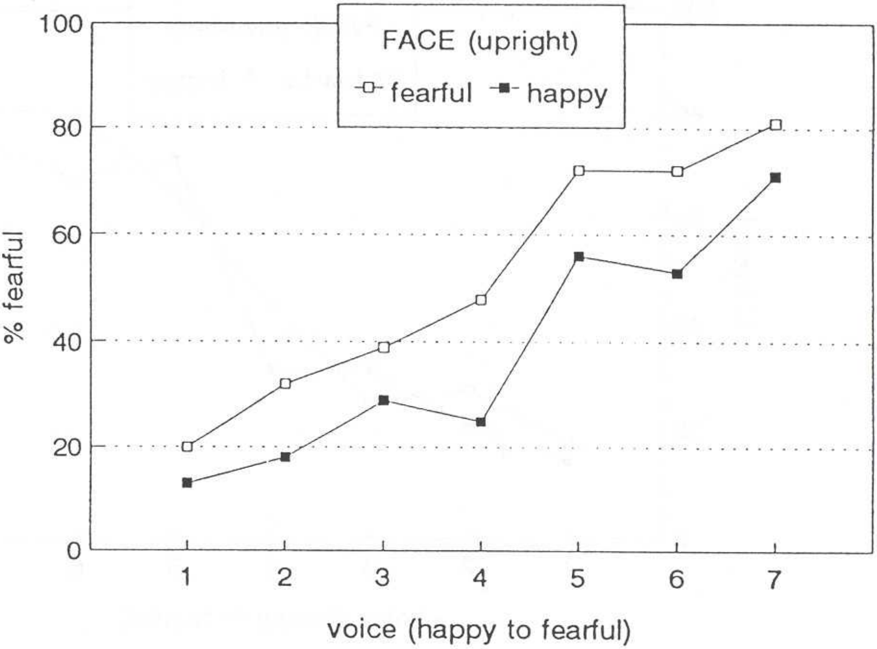

Figure 1. Percent of "fearful" responses as a function of voice expressive tone (abcissa), with emotion expressed by the face as parameter: Condition with face upright.

A repeated measure analysis of variance (ANOVA) including Facial expression (*happy* versus *fearful*), Orientation (*upright* versus *inverted*) and Voice tone (7 levels), as within-subject factors, was performed on these data. The main effects of Voice tone, Facial expression, and Orientation were all significant, $F(6, 66) = 45.41, p < .001$; $F(1, 11) = 6.91, p < .03$, and $F(1, 11) = 5.07, p < .05$, respectively. Most importantly, the Facial expression by Orientation interaction, which measures the reduction of bias resulting from face inversion, was significant, $F(1, 11) = 6.30, p < .03$.

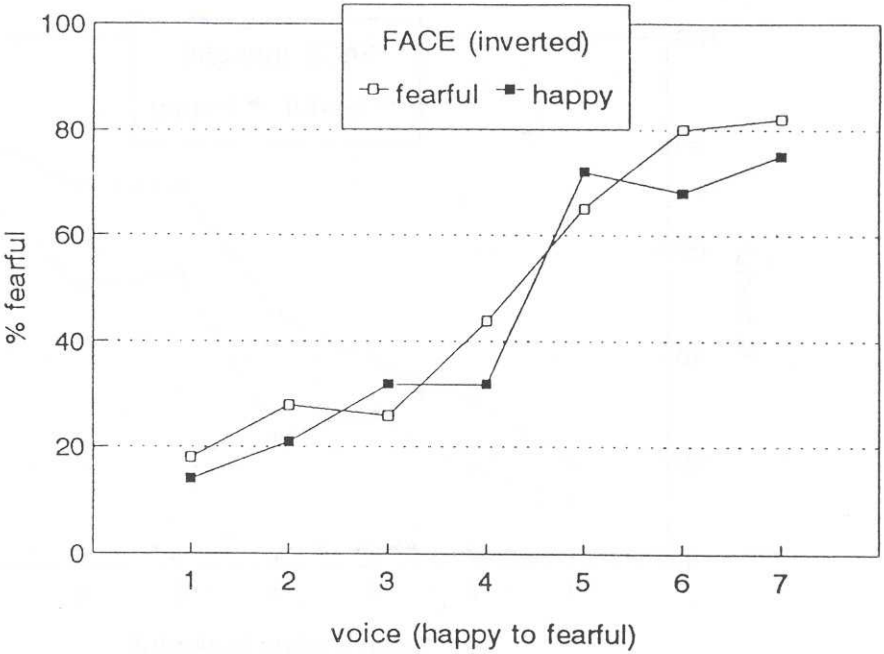

Figure 2. Percent of "fearful" responses as a function of voice expressive tone (abcissa) with emotion expressed by the face as parameter: Condition with face inverted.

DISCUSSION

The goal of this study was to determine whether the effect of facial expression on judgments of voice tone depends on the orientation of the face. It was found that the previously reported bias indeed occurs with an upright face, but not with an inverted one.

The absence of bias from inverted faces is interesting when compared with previous findings regarding the difficulty of explicit recognition of emotion from inverted faces (de Gelder et al., 1997; McKelvie, 1995). It would seem that the two manifestations of facial emotion processing, explicit identification and crossmodal bias, are influenced by the same variables.

The bias observed with upright faces confirms earlier results. Just as in Experiment 3 of de Gelder and Vroomen (1998), it was obtained with instructions to focus attention on the voice and ignore the expression in the face, supporting the conclusion that crossmodal affective bias is an automatic perceptual phenomenon which cannot be reduced to some post-perceptual voluntary adjustments.

The present study, which is a first exploration into the mechanisms of crossmodal bias in the recognition of emotion, involved just one particular pair of emotions. Obviously, one could ask how far the results would be replicated with other combinations of emotions. An important factor to be taken into account in further studies concerns the actual emotion presented in the voice and the face, respectively. There are important variations in the effectiveness with which different emotions are conveyed in the face (Ekman & Friesen, 1975; McKelvie, 1995). Similar variations exist for the voice. Moreover, these differences are not necessarily correlated across modalities. For example, happiness has generally been found a clearer facial expression than, for example, sadness, but in the voice, the prosodic pattern for sadness is clearer than the one for happiness. Such basic facts must be taken into account if general principles concerning the multimodal perception of emotion are to be derived.

As discussed in other papers from our research program (de Gelder et al., 1997; de Gelder & Vroomen, 1998), a framework fitting the present approach is that of a specialized functional processing unit, or module (Fodor, 1983), whose domain is the perception of emotional expressions and whose operations are triggered by expressive information in the different sense modalities that convey such information. The best documented example of this kind of multimodal module is presently offered by language and the different modalities like hearing, lipreading, or sign language that can convey it. The notion of a module specialized for the recognition of emotional expression contrasts with traditional views on two major counts. One concerns the semantics of the representations of emotion. The other concerns the question of whether those representations are modality-specific. However, the evidence provided by the present paradigm cannot tell us whether emotional expressions conveyed by the voice and by the face, respectively, are each represented in a separate system, whether there exist in the mind/brain abstract representations that can be matched to manifestations in either of the two input modalities.

ACKNOWLEDGEMENTS

Preparation of this paper was partly supported by a grant from the Ministry of Education and Scientific Research of the French-speaking Community of Belgium (Concerted Research Action 96/01-2037).). The research by the second author was funded by the Dutch Academy of Sciences (KNAW).

RÉSUMÉ

Des travaux antérieurs ont démontré l'existence d'influences mutuelles automatiques entre les émotions exprimées par des visages présentés au regard et par des phrases entendues. L'objectif de la présente étude était de déterminer si, dans le cas particulier de l'influence de l'expression faciale sur l'identification du ton affectif de la voix, l'effet résiste à l'inversion du visage, une manipulation dont on sait qu'elle détériore la reconnaissance de l'expression tout comme celle de l'identité personnelle. De jeunes adultes ont reçu pour tâche de classer comme heureux ou effrayé le ton de voix des phrases d'un continuum allant en 7 échelons du bonheur à la peur. Les phrases étaient présentées seules ou en même temps qu'une photo statique d'un visage exprimant soit le bonheur, soit la peur, en orientation normale ou inversée. La consigne pour les essais bimodaux était de baser le jugement sur la seule voix et d'ignorer le visage. Un biais de ce jugement dans la direction de l'expression du visage est apparu pour l'orientation normale, mais non pour l'orientation inversée. On discute la relation de ces résultats avec les mécanismes qui sous-tendent la reconnaissance multimodale des émotions.

REFERENCES

Calder, A., Young, A., Perrett, D., Etcoff, N., et al. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3*, 81-117.

Charpentier, F., & Moulines, E. (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Eurospeech 89, 2*, 13-19.

Damasio, A. R., Tranel, D., & Damasio, H. (1990). Face agnosia and the neural substrate of memory. *Annual Review of Neuroscience, 13*, 89-109.

de Gelder, B., Böcker, K., Tuomainen, J., Hensen, M., & Vroomen, J. (1998). The combined perception of emotion from voice and face: early interaction revealed by electric brain responses (submitted).

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

de Gelder, B., & Vroomen, J. (1995). Categorical perception of emotional speech. *Journal of the Acoustical Society of America, 100,* 4 Pt. 2, 2818-2819.

de Gelder, B., & Vroomen, J. (1998). Emotion by ear and by eye (submitted).

de Gelder, B., Vroomen, J., & Teunisse, J.-P. (1995). Hearing smiles and seeing cries: The bimodal perception of emotion. 36th Annual Meeting of the Psychonomic Society, Los Angeles, 309, 30.

Diamond, R., & Carey, S. (1986). Why faces are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115,* 105-117.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face.* Englewood Cliffs: Prentice-Hall.

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

Fodor, J. A. (1983). *The modularity of mind.* 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.

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

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.

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

Thompson, P. (1980). Margaret Thatcher: A new illusion. *Perception, 9,* 483-484.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology, 79,* 161-204.

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. In *Proceedings of the Third European Conference on Speech Communication and Technology* (pp. 577-580). Berlin.

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.

Received 23 December, 1997
Accepted 15 May, 1998