

# Cross-modal and incremental perception of audiovisual cues to emotional speech

Barkhuysen, P.; Krahmer, E.J.; Swerts, M.G.J.

*Published in:*
Language and Speech

*Publication date:*
2010

*Document Version*
Publisher's PDF, also known as Version of record

# *Crossmodal and incremental perception of audiovisual cues to emotional speech*

**Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts**

*Tilburg University*

**Key words**

*audiovisual speech*

*emotional speech*

*multimodality*

*perception*

*timing*

**Abstract**

In this article we report on two experiments about the perception of audiovisual cues to emotional speech. The article addresses two questions: (1) how do visual cues from a speaker's face to emotion relate to auditory cues, and (2) what is the recognition speed for various facial cues to emotion? Both experiments reported below are based on tests with video clips of emotional utterances collected via a variant of the well-known Velten method. More specifically, we recorded speakers who displayed positive or negative emotions, which were congruent or incongruent with the (emotional) lexical content of the uttered sentence. In order to test this, we conducted two experiments.

The first experiment is a perception experiment in which Czech participants, who do not speak Dutch, rate the perceived emotional state of Dutch speakers in a bimodal (audiovisual) or a unimodal (audio- or vision-only) condition. It was found that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, where the difference between congruent and incongruent emotional speech is larger for the negative than for the positive conditions. Interestingly, the largest overall differences between congruent and incongruent emotions were found for the audio-only condition, which suggests that posing an incongruent emotion has a particularly strong effect on the spoken realization of emotions.

*Acknowledgments:* The research presented above was conducted within the framework of the FOAP project (http://foap.uvt.nl), sponsored by the Netherlands Organization of Scientific Research. We thank Marie Nilsenová, Janneke Wilting, Lennard van de Laar, Carel van Wijk, and Karel Fliegel for help with the collection of the audiovisual stimuli, the execution of the perception experiments, and statistical advice. We thank Jean Vroomen for allowing us to make use of the PAMAR software for our gating experiment. We also thank the participants in our experiments for agreeing to let us include their photographs in our article.

*Address for correspondence*. M. Swerts, P.O. Box 90153, NL-5000 LE Le Tilburg; e-mail: <M.G.J.Swerts@uvt.nl>





*Language and Speech* 0023-8309; Vol 53(1): 3–30; 348993; DOI:10.1177/0023830909348993
http://las.sagepub.com

The second experiment uses a *gating paradigm* to test the recognition speed for various emotional expressions from a speaker's face. In this experiment participants were presented with the same clips as experiment I, but this time presented vision-only. The clips were shown in successive segments (gates) of increasing duration. Results show that participants are surprisingly accurate in their recognition of the various emotions, as they already reach high recognition scores in the first gate (after only 160 ms). Interestingly, the recognition scores raise faster for positive than negative conditions. Finally, the gating results suggest that incongruent emotions are perceived as more intense than congruent emotions, as the former get more extreme recognition scores than the latter, already after a short period of exposure.

# 1 Introduction

Facial expressions are often considered to be windows to the soul, because they are thought to reveal the emotional state of a speaker. From a face, we may tell whether a person is feeling happy, sad, angry, anxious, etc. (e.g., Adolphs, 2002; Carroll & Russell, 1996; Schmidt & Cohn, 2001). However, previous research has brought to light that the emotional state of a speaker can also be derived from other modalities. In the auditory domain, it has been shown that listeners can infer the emotional state from the expression of a speaker's voice (Bachorowski, 1999; Banse & Scherer, 1996; Scherer, 2003). Scherer (2003) states that acoustic emotional expressions occur at various stages (and levels) within the communication process. There is a wealth of neurobiological evidence suggesting that the recognition of emotion is a complex process that involves the cooperation of processes across various brain structures (Adolphs, 2002; Vuilleumier & Pourtois, 2007).

However, while we have gained much insight into how unimodal stimuli (either auditory or visual) are processed, far less is known about the extent into which these modalities interact with each other. There is some preliminary evidence that one modality may have an effect on another one, as is, for example, clear from the fact that people are able to detect from a speaker's voice whether he or she is showing a smile (Aubergé & Cathiard, 2003). It is very likely that the brain tends to bind information received through different modalities (referred to as *intermodal* or *crossmodal binding*) (see e.g., Ghazanfar, Maier, Hoffman, & Logothetis, 2005), because often it receives information simultaneously through different sensory systems but from the same distal source (Pourtois, de Gelder, Vroomen, Rossion, & Crommelinck, 2000), especially because the "sender" tends to transmit information across different modalities (Graf, Cosatto, Ström, & Huang, 2002). Generally speaking, this multimodal integration is very useful, because input from one modality can substitute another one in deteriorated circumstances. For example, lip-reading can be useful for speech comprehension in noisy environments (Calvert, Brammer, & Iversen, 1998), or vice versa, in darkness, auditory signals can replace visual signals (Calvert et al., 1998). Neurological studies have already brought to light what the nature is of different networks activated in different brain areas during crossmodal binding, for example when involved in audio-visual speech processing (Calvert, 2001; Calvert et al., 1998; Sekiyama, Kanno, Miura, & Sugita, 2003). In the past, the binding and interaction of different modalities has been demonstrated very spectacularly in the so-called McGurk effect, which shows that the auditory perception of a sound can be altered by the display of incongruent visual information (McGurk & MacDonald, 1976). The McGurk paradigm has been

a source of inspiration for studies on the perception of emotion which also use stimuli with congruent and incongruent auditory and visual cues to emotions (Aubergé & Cathiard, 2003; de Gelder & Vroomen, 2000; Hietanen, Marinnen, Sams, & Rusakka, 2001). However, while much research has been done about crossmodal integration during *audiovisual speech* processing, much more needs to be done about crossmodal integration during the processing of *emotions* (e.g., de Gelder, Bocker, Tuomainen, Hensen, & Vroomen, 1999; Pourtois et al., 2000), when combined with audiovisual speech. It has been shown that the ability to integrate information from emotional faces with emotional prosody is already present in 7-month-old infants (Grossmann, Striano, & Friederici, 2006). Unfortunately, many of these studies investigating the recognition of emotional expressions have been based on analyses of static images, such as photographs or drawings (see e.g., Ekman, Friesen, & Ellsworth, 1972, pp.49–51), rather than dynamic images. As a result, little is known about the perception of emotions through "fleeting changes in the countenance of a face" (Russell, Bachorowski, & Fernández-Dols, 2003). Often, a realistically varying speech signal is combined with a static face, resulting in knowledge about online auditory speech but not about online visual speech. Consequently, we do not yet fully understand whether auditory and visual cues of emotional speech differ in perceptual strength, and how people deal with input coming from two modalities when they have to make judgments about a speaker's emotional state (in contrast to judging an emotional state *without* speech). This knowledge could be very useful for the development of computerized speech systems, for instance (Cohn & Katz, 1998). Therefore, the first question we want to explore in this article is whether the processing of emotional speech is integrated across modalities, that is, whether the perception of a combination of two modalities is more successful than the perception of a single modality alone.

A second question we want to explore is to what extent the recognition of emotion varies as a function of the time that people are exposed to the facial expressions of a speaker. There are reasons to believe that this temporal recognition process may vary for different kinds of emotions, such as positive versus negative emotions. That is, it has been argued that positive and negative emotions are not recognized equally fast, although there is some controversy about the direction of this effect. Fox et al. (2000) claim that angry facial expressions are detected more rapidly than happy expressions, whereas Leppänen and Hietanen (2004) report that positive facial expressions are recognized faster than negative ones. (Note, however, that closer inspection of the stimuli used in these studies reveals that the angry stimuli of the last two experiments reported in Fox et al., 2000, are similar to the sad stimuli in the experiments of Leppänen & Hietanen, 2004, basically using very similar stylized emoticons to reflect these emotions.) Potentially, the valency effect on recognition speed, in whichever direction, may partly be due to timing-related differences in facial expressions. In addition, there is work on the time-course of intermodal binding of emotions, where it appears that integration of emotional information from the face and from the voice occurs at an early stage of processing (before both modalities have been fully processed), and uses low-level perceptual features (de Gelder et al., 1999). According to Pourtois et al. (2000), intermodal binding of emotions occurs around 110 ms post-stimulus, which is earlier than the processing of intermodal speech, which lies around 200 ms post-stimulus (Pourtois et al., 2000; see also Sekiyama

et al., 2003). However, as mentioned above, these studies work with the presentation of static rather than dynamic faces. There is neurological evidence that moving faces are processed by a fundamentally different path than static faces (Humphreys, Donelly, & Riddoch, 1993).

As mentioned above, many emotion studies rely on "acted" data. The work of Ekman (e.g., 1993; Ekman et al., 1987), for instance, is based on posed photographs of actors, and also in speech research actors are frequently used. Additionally, many studies, in line with the McGurk paradigm, make use of stimuli that consist of incongruent cues to various emotions (e.g., conflicting visual and auditory cues). An important question is whether such stimuli are *ecologically valid*, in that acted or incongruent emotions may be more "controlled" than the spontaneous displays of emotions in natural interactions. Neurological studies have shown that voluntary expressions are fundamentally different in nature from spontaneous expressions (Gazzaniga & Smylie, 1990; Rinn, 1984, 1991). From a corpus study, Valstar, Pantic, Ambadar, and Cohn (2006) conclude that these two can be distinguished on the basis of the speed, duration, and sequence of brow actions. Similarly, there is some work into timing-related differences between spontaneous and posed smiles (also known as Duchenne and non-Duchenne smiles; see e.g., Ekman, 2004, p.204–209, for a description; Ekman, Davidson, & Friesen, 1990). Cohn and Schmidt (2004) report that spontaneous, as opposed to posed smiles, have a smaller amplitude, have an onset that is more related to the duration (i.e., longer smiles are slower in onset), can have multiple rises of the mouth corners, and are accompanied by other facial actions, either simultaneously or immediately following.

In sum, the aim of this article is to look in more detail at the perception of audio-visual expressions of positive and negative emotions (both congruent and incongruent) in spoken language, and to explore the recognition speed of these dynamic expressions of positive and negative emotions (both congruent and incongruent). It describes two perception experiments and an observational study for which we used Dutch data collected via a variant of the Velten technique. This is an experimental method to elicit emotional states in participants, by letting speakers produce sentences increasing in emotional strength (Velten, 1968). The next section first describes previous work by Wilting, Krahmer, and Swerts (2006), whose data were used in the current article. We present a brief summary of their method and the results of an experiment in which they first elicit congruent and incongruent emotional data from speakers using an adaptation of the Velten technique, and then selected film clips (without sound) that they showed to observers who had to judge the emotional state of the recorded speakers. The later sections describe how the current study uses the data collected by Wilting et al.'s research by testing these experimental stimuli in both bimodal and unimodal conditions. For reasons described below, the participants in the current study were native speakers of Czech, who were not able to understand the lexical content of the presented utterances. In the second experiment we test the original experimental stimuli (but presented without sound) on Dutch participants using a *gating paradigm* (Grosjean, 1996). Our final study consists of observational analyses of various facial expressions in the upper and lower areas of a speaker's face to see whether certain features correlate with reported or perceived emotions from speakers.

## 2 Audiovisual recordings

Wilting et al. (2006) used an adapted Dutch version of the original Velten (1968) induction procedure, using 120 sentences evenly distributed over three conditions (POSITIVE, NEUTRAL and NEGATIVE).[1] Besides the three conditions described by Velten for the induction of congruent emotions (POSITIVE, NEUTRAL, NEGATIVE), two "acting" conditions were added. In one of these, participants were shown negative sentences and were asked to utter these as if they were in a positive emotion (INCONGRUENT POSITIVE); in the other, positive sentences were shown and participants were instructed to utter these in a negative way (INCONGRUENT NEGATIVE). The sentences showed a progression, from neutral ("Today is neither better nor worse than any other day") to increasingly more emotional sentences ("God I feel great!" and "I want to go to sleep and never wake up" for the positive and negative sets, respectively), to allow for a gradual build-up of the intended emotional state.

Participants were told that the goal of the experiment was to study the effect of mood on memory recall (earlier work has revealed that mood induction procedures become more effective when the induction serves a clear purpose, e.g., Westermann, Spies, Stahl, & Hesse, 1996). The instructions, a slightly abridged version of the original instructions from Velten, were displayed on the computer screen, and participants were instructed to first silently read the texts, after which they had to read them aloud. For the congruent conditions, the participants were instructed to try to "feel" and "display" the emotion that the sentence was representing, while for the incongruent conditions, the participants were instructed to try to "feel" and "display" the opposite emotion.[2]

During the data collection, the sentences were displayed on a computer screen for 20 seconds, and participants were instructed to read each sentence, first silently and then out loud. Recordings were made from the face and upper body of the speakers with a digital camera, and a microphone connected to the camera. Fifty Dutch speakers (10 per condition) were recorded in the data collection, 31 female and 19 male, none of them being a (professional) actor. The advantage of using different speakers across conditions is that, in the perception tests, observers could not base their judgments upon the familiarity of the faces, therefore preventing learning effects. Some representative stills are shown in Figure 1.

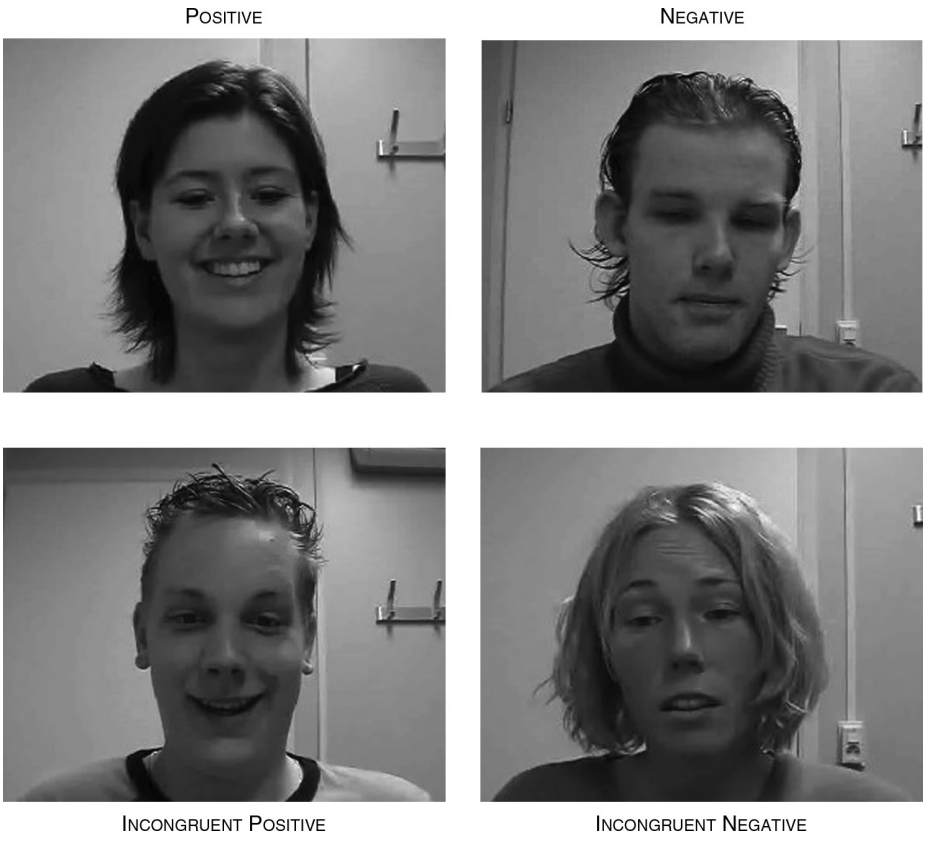
Immediately following this phase, participants had to fill in a short mood questionnaire ("At this moment, I feel ... ") derived from Mackie and Worth (1989) and Krahmer, van Dorst, and Ummelen (2004), consisting of six seven-point bipolar

1 We chose to classify the emotions under investigation according to their valence, i.e., positive and negative, instead of using a subjective term such as "happy" or "depressed," because we were only interested in the valence of an emotion and not in specific properties of an individual emotion.

2 Note that although the terminology in our instruction reflected only the valence of the emotion, the list designed by Velten should invoke the emotions "elation" and "depression" (Velten, 1968). However, these two emotions differ primarily along one dimension, i.e., positive to negative, according to the dimensional view upon emotions (e.g., Bachorowski, 1999). By instructing the participants to feel and display the opposite emotion as the one reflected in the sentences, we tried to direct the way they would "act" by the content of the list rather than by terminology.

**Figure 1**

Representative stills of congruent (top) and incongruent (bottom) emotional expressions, with on the left-hand side the positive and on the right-hand side the negative versions




semantic differential scales, using the following adjective pairs (English translations of Dutch originals): happy/sad, pleasant/unpleasant, satisfied/unsatisfied, content/discontent, cheerful/sullen, and in high spirits/low-spirited. The order of the adjectives was randomized; for ease of processing, negative adjectives were mapped to 1 and positive ones to 7.

Wilting et al. (2006) reported two main findings. First, from the survey presented to participants after the elicitation phase, it turned out that the Velten technique was very effective in that the positive and negative emotions could indeed be induced through this method, but only for speakers in the congruent conditions; the speakers in the incongruent conditions did not feel different from the speakers in the neutral condition. Second, observers turned out to be able to reliably distinguish between positive and negative emotions on the basis of visual cues; interestingly, the incongruent versions led to more extreme scores than the congruent ones, which suggests that the incongruent emotions were displayed more strongly than the congruent ones. This raises the question in what sense the *positive* emotions differ from their negative counterparts. In this article, we investigate the hypothesis that one difference is *durational*, especially

in the onset, assuming that positive emotions appear quicker on the face than negative ones, though this may be different for congruent versus incongruent emotions. Also, we are interested in the question whether the perception of positive versus negative emotions differs across *modalities*, and whether the perception of *congruent* versus *incongruent* emotions differs across modalities, and/or whether there is an interaction between these two.

In the next study we test these data in both bimodal and unimodal conditions, on Czech participants.

# 3 Experiment I: Crossmodal perception

## 3.1 Stimuli

From each of the speakers in the recordings, the last sentence was selected. These sentences captured the speakers at the maximum height of the induced emotion. We chose to use maximum height stimuli, because Horstmann (2002) reported that prototypical emotions resemble the most intense expression of an emotion. The previous study by Wilting et al. (2006) was conducted with vision-only stimuli presented to Dutch participants. It would not have been possible to present the auditory or audiovisual variants to Dutch participants, as the lexical information would be a give away clue for the speaker's emotional state. Still, we are interested in the perception of the audio-only and audiovisual stimuli. Therefore the Dutch sentences were presented to Czech participants in the perception test, as they did not understand Dutch.

## 3.2 Design

The experiment uses a repeated measurements design with modality as between-subjects factor (with levels: AUDIOVISUAL: AV, VISION-ONLY: VO and AUDIO-ONLY: AO), condition as within-subjects factor (with levels: INCONGRUENT NEGATIVE, NEGATIVE, NEUTRAL, POSITIVE, and INCONGRUENT POSITIVE), and perceived emotional state as the dependent variable.

## 3.3 Procedure

Participants were told that they would see or hear 50 speakers in different emotional states, and that their task was to rate the perceived state on a seven-point valency scale ranging from 1 (= *very negative*) to 7 (= *very positive*). Participants were not informed about the fact that some of the speakers were displaying an incongruent emotion. Within each modality, there were two subgroups of participants, who were presented with the same stimuli but in a different random order to compensate for potential learning effects. Stimuli were preceded by a number displayed on the screen indicating which stimulus would come up next, and followed by a 3 second interval during which participants could fill in their score on an answer form. Stimuli were shown only once. The experiment was preceded by a short training session consisting of five stimuli of different speakers uttering a non-experimental sentence to make participants acquainted with the stimuli and the task. If all was clear, the actual experiment started, after which there was no further interaction between the participants and the experimenter. The perception tests in the three conditions were conducted as a group experiment with the material presented on a large screen at the front of the classroom. The entire experiment lasted approximately 10 minutes.

### 3.4 Participants

Fifty-four people (18 per condition) participated in the experiment, nine female and 45 male, with an average age of 23 (range 21–30). All were students and Ph.D. students from the Czech Technical University (Faculty of Electrical Engineering) and the Charles University (Faculty of Philosophy and Arts) in Prague, Czech Republic. The choice of Czech participants was arbitrary; the only real constraint was that the participants should not understand Dutch.

### 3.5 Statistical analyses

All tests for significance were performed with a repeated measures analysis of variance (ANOVA). Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse-Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. Post hoc analyses were performed with the Bonferroni method.

### 3.6 Results

Figure 2 and Table 1 summarize the results. A repeated measures ANOVA, with modality as between-subjects factor, condition as within-subjects factor, and perceived emotional state as the dependent variable, shows that *condition* has a significant effect on *perceived emotional state*, $F(4, 204) = 145.042$, $p < .001$, $\eta^2_p = .740$. Repeated contrasts revealed that all conditions (level 1: INCONGRUENT NEGATIVE, level 2: NEGATIVE, level 3: NEUTRAL, level 4: POSITIVE, and level 5: INCONGRUENT POSITIVE) lead to a significantly different perceived emotion, $F_{12}(1, 51) = 89.558$, $p < .001$, $\eta^2_p = .637$; $F_{23}(1, 51) = 50.167$, $p < .001$, $\eta^2_p = .496$; $F_{34}(1, 51) = 43.855$, $p < .001$, $\eta^2_p = .462$; $F_{45}(1, 51) = 20.052$, $p < .001$, $\eta^2_p = .282$. It is interesting to observe that the *incongruent* emotions are perceived as more intense than the congruent ones. Speakers in the INCONGRUENT POSITIVE condition are overall perceived as the most positive, $M = 4.70$, $SD = 0.53$, and speakers in the INCONGRUENT NEGATIVE condition are perceived as the most negative, $M = 2.72$, $SD = 0.63$. Note that the perceptual difference between incongruent and congruent emotional speech is larger for the *negative* emotions. In general, it seems that the incongruent emotions are classified "better," or interpreted as more intense than the congruent emotion.

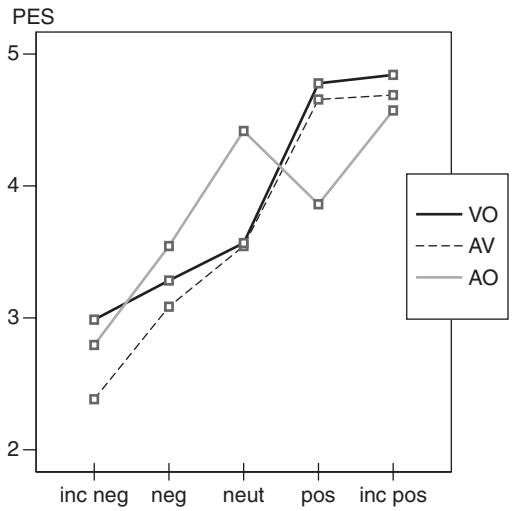
**Table 1**

Perceived emotional state on a seven-point scale (1 = very negative, 7 = very positive) as a function of condition (standard deviations between brackets) as well as condition split by modality

| *Condition* | *AV* | *VO* | *AO* | *Total* |
|---|---|---|---|---|
| INCONGR POSITIVE | 4.69 (.35) | 4.84 (.35) | 4.57 (.78) | 4.70 (.53) |
| POSITIVE | 4.66 (.46) | 4.78 (.46) | 3.86 (.95) | 4.43 (.77) |
| NEUTRAL | 3.54 (.31) | 3.57 (.46) | 4.42 (.49) | 3.84 (.59) |
| NEGATIVE | 3.08 (.49) | 3.28 (.47) | 3.54 (.77) | 3.30 (.61) |
| INCONGR NEGATIVE | 2.38 (.36) | 2.99 (.64) | 2.79 (.72) | 2.72 (.63) |
| Total | 3.67 (.98) | 3.89 (.91) | 3.84 (.98) | 3.80 (.96) |

**Figure 2**

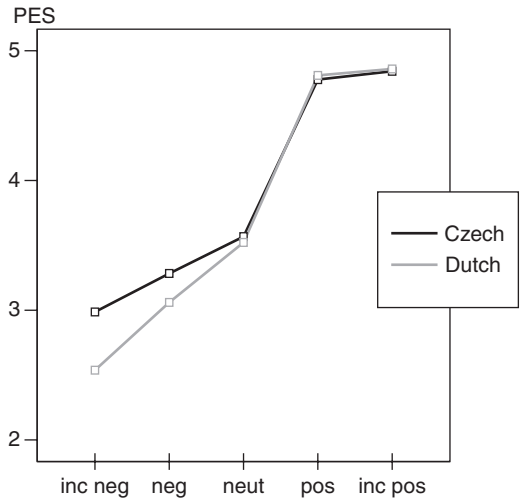
The mean perceived emotional state (1 = very negative, 7 = very positive) per condition and modality




*Modality* does not have a significant main effect on *perceived emotional state*, $F(2, 51) = 1.881$, $p = .163$, $\eta^2_p = .069$, but interestingly there was an interaction between *condition* and *modality*, $F(8, 204) = 10.981$, $p < .001$, $\eta^2_p = .301$. In all three modalities the *incongruent* emotions are perceived as more intense than the congruent ones; speakers in the INCONGRUENT POSITIVE condition are perceived as the most positive, and speakers in the INCONGRUENT NEGATIVE condition are perceived as the most negative. However, repeated contrasts showed that all levels of condition and modality interact significantly with each other, $F_{12}(1, 51) = 5.438$, $p < .01$, $\eta^2_p = .176$; $F_{23}(1, 51) = 5.254$, $p < .01$, $\eta^2_p = .171$; $F_{34}(1, 51) = 41.526$, $p < .001$, $\eta^2_p = .620$; $F_{45}(1, 51) = 13.475$, $p < .001$, $\eta^2_p = .346$. For both the AV and the VO modality the difference between POSITIVE and INCONGRUENT POSITIVE is very small, $D_{AV} = 0.03$, and $D_{VO} = 0.06$, while this difference is much larger in the AO modality, $D_{AV} = 0.71$: for this modality, the POSITIVE condition even scored lower on the valency scale than NEUTRAL. On the other side of the spectrum, the difference between the NEGATIVE and the INCONGRUENT NEGATIVE condition is substantial for the AO and the AV modality, $D_{AV} = 0.75$, and $D_{AV} = 0.70$, but here the VO modality stands out in the sense that the difference is relatively small, $D_{VO} = 0.29$. In other words, the classification pattern for the AV modality resembles the VO modality for the positive moods, while for the negative moods the pattern of the AV modality is similar to the AO modality. Note also that the difference between the two incongruent emotions is larger in the AV modality, $D_{AV} = 2.31$, somewhat smaller in the VO modality, $D_{VO} = 1.85$, and the smallest in the AO modality, $D_{AV} = 1.78$. Another interesting point is the difference between the facial expressions and vocal expressions in the POSITIVE condition, $D_{VO\text{-}AV} = 0.92$. This difference is very large in comparison to the other conditions, apart from the NEUTRAL condition, where, in contrast to the POSITIVE condition, the AO modality scores higher than the VO modality, $D_{VO\text{-}AV} = -0.85$.

**Figure 3**

The mean perceived emotional state (1 = very negative, 7 = very positive) per condition and nationality




Further, we compared the classification of the Czech participants for the fragments presented in the VO modality with the results of the earlier Dutch perception test (Wilting et al., 2006), by a repeated measures ANOVA, with nationality as between-subjects factor, condition as within-subjects factor, and perceived emotional state as the dependent variable. It turns out that the main effect of *nationality* was not significant, $F(1, 56) = 1.905$, $p = .173$, $\eta^2_p = .033$. There was a significant interaction between *nationality* and *condition*, $F(4, 224) = 5.088$, $p < .01$, $\eta^2_p = .083$; however, repeated contrasts showed that this difference was only caused by the difference between the NEGATIVE and the INCONGRUENT NEGATIVE stimuli, $F_{12}(1, 56) = 4.505$, $p = .038$, $\eta^2_p = .074$ (see Figure 3).

## 3.7 Summary

We have reported on a perception experiment in which Czech participants rated their perceived emotional state of Dutch speakers. These speakers could either display a positive or a negative emotion, which was either congruent or incongruent. The Czech participants were confronted with these utterances in a bimodal (audiovisual) or a unimodal (audio-only or vision-only) condition.

There was no overall effect of modality. Further, it was found that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, where the difference between incongruent and congruent emotional speech is larger for the negative than for the positive conditions. Interestingly, the largest overall differences between incongruent and congruent emotions were perceived in the audio-only condition, which suggests that displaying an incongruent emotion has a particularly strong effect on the spoken realization of emotions. This difference between the congruent and the incongruent conditions is in particular larger for the *positive* emotions. In addition, comparing the different

modalities suggests that positive emotions are clearer in the visual modality (since the highest scores were obtained in the AV and VO modalities), while the classification of negative emotions in the AV modality follows the pattern of the AO modality. Another interesting point is the difference between facial and vocal expression within the separate conditions. It seems that the Velten procedure did not elicit recognizable vocal expressions in the POSITIVE condition, whereas it elicited recognizable facial expressions. On the other hand, the senders in the INCONGRUENT POSITIVE condition were able to display recognizable facial *and* vocal expressions. We also compared the classification of the Czech participants for the VO fragments with the results of the Dutch perception test with the same stimuli (Wilting et al., 2006), which lead to essentially the same results.

Although we have shown that participants can correctly classifiy dynamical expressions of (congruent and incongruent) emotions, we did not investigate the speed with which these expressions were classified. This is interesting in the light of the above-discussed timing differences between spontaneous and voluntary expressions. We also do not know whether there are timing differences between positive and negative emotions. The second experiment will investigate whether positive and negative emotions (both congruent and incongruent) differ with respect to the speed with which they are recognized as such.

# 4 Experiment II: Incremental perception

## 4.1 Stimuli

The second perception test is based on the *gating paradigm*, which is a well-known design in spoken word recognition research (Grosjean, 1996). In this paradigm, a spoken language stimulus is presented in segments that increase in length and participants are asked to propose the word being presented and to give a confidence rating after each segment. The dependent variables are the *isolation point* of the word (i.e., the *gate*[3]), the *confidence ratings* at various points in time and the *word candidates* proposed after each segment.

The current perception test resembles this gating design, but only in that we present parts of the original sentences used in Wilting et al. (2006), increasing in length. To enable comparisons across experiments, the fragments were cut from the start of the original fragment as it was used in experiment I. The first segment is very short, only consisting of four frames (160 ms). The size of the later segments increases in steps of 160 ms until the last, sixth segment, which is 960 ms long. Each segment S+1 thus includes the preceding segment S, and extends it by four extra frames (or 160 extra ms). We only used six segments, because a pilot study indicated that adding longer segments did not lead to a substantial increase in recognition accuracy.

The current set-up differs from the "standard" gating approach, in that we do not ask participants to give confidence ratings. Rather, after each gate, participants have to indicate whether they believe that the speaker is in a positive or in negative mood, *or* whether they cannot make this distinction on the basis of the current gate.

3 In our perception test, the isolation point is rather the gate at which a fragment is correctly recognized and where responses for following gates are no longer changed.

### 4.2 Design

The experiment uses a repeated measurements design with condition (with levels: INCONGRUENT NEGATIVE, NEGATIVE, NEUTRAL, POSITIVE and INCONGRUENT POSITIVE) and gate (with levels: ONE (i.e., 160 ms), TWO (i.e., 320 ms), THREE (i.e., 480 ms), FOUR (i.e., 640 ms), FIVE (i.e., 800 ms), to SIX (i.e., 960 ms)) as within-subjects factors, and confidence (with levels: non-answers "don't know" versus answers "positive or negative") and perceived emotional state (with levels: "positive" and "negative") as the dependent factors.

### 4.3 Procedure

Participants were tested individually. They were invited into a quiet room, and asked to take place in front of the computer. Participants were told that they would see 40 speakers in different emotional states, and that for each speaker they would see six short, overlapping fragments (the gates). The task of the participants was to determine, for each gate, whether the speaker was in a positive or in a negative mood. They were given three answering possibilities: "negative," "don't know," and "positive." Three buttons on the keyboard were labeled with these answer possibilities, and *only after* viewing a film clip, could participants press one of these buttons, after which the next stimulus appeared. Therefore, they could take as much time as they needed for judging the film clip, while they were viewing a blank screen. However, the instruction encouraged the participants to respond quickly. If they were not sure yet about the emotion of the clip, they could use the "don't know" button, which was designed for this purpose. Participants were not informed about the fact that some of the speakers were acting an incongruent emotion.

The gates were presented in a *successive* format: that is, participants viewed all the segments of a sentence, starting with the shortest and finishing with the longest. The gates were presented *forwards*, that is, the first was cut from the beginning of the sentence and then increasingly longer stretches were added, thus later segments were approaching the end ("left-to-right"). Stimulus groups (containing six gates) were preceded by a number displayed on the screen indicating which stimulus group would come up next, and followed by the first segment only after which the participants could press the appropriate button to indicate their answers. Stimuli were shown only once. Stimulus groups were presented in one of four random orders, to compensate for potential learning effects. The fragments were only presented visually, without the corresponding sound; therefore the lexical or grammatical content could not influence the participants' decision. Also, no feedback was given to participants about the correctness of their scores.

The experiment was preceded by a short training session consisting of one stimulus group containing six gate-segments, uttered by a single speaker uttering a non-experimental, neutral sentence to make participants acquainted with the stimuli and the task. If all was clear, the actual experiment started, after which there was no further interaction between the participants and the experimenter. The entire experiment lasted approximately 25 minutes.

### 4.4 Participants

Forty people (10 per presentation order) participated in the experiment, 33 female and seven male, with an average age of 19 (range 18–27). All were students from Tilburg University in The Netherlands, none had participated as a speaker in the study by Wilting et al. (2006) or in experiment I, and all were unaware of the experimental question.

**Table 2**

Perceived emotional state as a function of condition (standard errors between brackets) as well as condition split by gate

| *Gate* | *1* | *2* | *3* | *4* | *5* | *6* | *Total* |
|---|---|---|---|---|---|---|---|
| INCONGR | 0.81 | 0.77 | 0.74 | 0.74 | 0.75 | 0.75 | 0.76 |
| POSITIVE | (.03) | (.03) | (.02) | (.02) | (.02) | (.02) | (.02) |
| POSITIVE | 0.76 | 0.67 | 0.64 | 0.64 | 0.64 | 0.64 | 0.67 |
| | (.03) | (.03) | (.02) | (.02) | (.02) | (.02) | (.02) |
| NEGATIVE | 0.26 | 0.23 | 0.25 | 0.26 | 0.23 | 0.22 | 0.24 |
| | (.04) | (.03) | (.02) | (.03) | (.02) | (.03) | (.02) |
| INCONGR | 0.20 | 0.14 | 0.14 | 0.15 | 0.14 | 0.13 | 0.15 |
| NEGATIVE | (.03) | (.02) | (.03) | (.03) | (.03) | (.02) | (.02) |
| Total | 0.51 | 0.46 | 0.44 | 0.45 | 0.44 | 0.44 | |
| | (.02) | (.01) | (.01) | (.01) | (.01) | (.01) | |

## 4.5 Statistical analyses

All tests for significance were performed with a repeated measures ANOVA. Mauchly's test for sphericity was used, and when it was significant or could not be determined, we applied the Greenhouse-Geisser correction on the degrees of freedom. For the sake of transparency, we report on the normal degrees of freedom in these cases. Post hoc analyses were performed with the Bonferroni method.

## 4.6 Results

We report on the results in two steps. First we look at the percentages of answers and non-answers as a function of gate, and next we look at the number of positive and negative answers as a function of gate.

First of all, we present the general distribution of responses across the conditions in Table 2.[4]

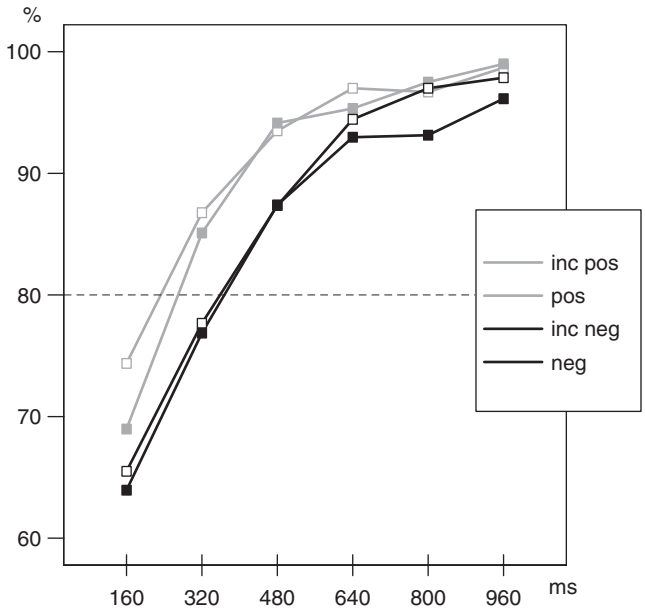
### 4.6.1 Non-answers versus answers

For this analysis, we recoded the responses such that non-answers ("don't know") were mapped to a value of 0 (= *no decision made*), and answers ("negative" or "positive") were mapped to 1. There were 1112 non-answers, which is 11.6% of all responses. There were a total of 191 missing values, which is 2% of all responses; these were replaced

4 There seems to be a response bias towards negative responses, i.e., the number of "positive responses" for the positive and the incongruent positive conditions is higher than the number of "don't know" responses. Therefore, within these conditions, the mean perceived emotional state "drops" in the later gates. This could be caused by the successive forward presentation format. According to Grosjean (1996), in this design potential artefacts may occur: "The successive presentation format may induce response perseveration and negative feedback. This in turn may yield a slightly conservative picture of recognition." However, the tendency for less extreme or more negative responses in the positive condition is in line with the results of Wilting et al. (2006) and with the results in the first perception experiment. Therefore, we do not consider this to be a problem.

**Figure 4**

The mean proportion of answers (vs. non-answers) as a function of gate (in ms) for different emotions




with the mean value over the 10 speakers per segment/gate. Figure 4 shows the proportion of answers as a function of gate. We assumed that the proportion of answers is a reflection of the level of confidence that the participants have in their ability to make a correct judgment at that particular gate. What this figure shows is that we find the most non-answers for the first gate, and the *congruent* emotions get more non-answers than their incongruent counterparts. In all conditions, the percentage of answers increases over the next gates, and seems to reach a plateau after the fourth gate (640 ms). Also, the speed of recognition (i.e., how much visual information, defined as the number of gates, is needed) differs for *positive* versus negative emotions. Taking an 80% threshold,[5] it can be seen that the recognition of *positive* emotions reaches this level already at gate 2 (congruent: $M = 0.83$, $SE = 0.028$; incongruent: $M = 0.87$, $SE = 0.025$), while the *negative* emotions reach this level only at gate 3 (congruent: $M = 0.87$, $SE = 0{,}031$; incongruent: $M = 0.87$, $SE = 0.026$).

A repeated measures ANOVA with condition and gate as within-subjects factors and proportion of answers (i.e., the confidence) as the dependent variable shows that *condition* has a significant effect on the relative proportion of answers, $F(3, 117) = 8.051$, $p < .001$, $\eta^2_p = .171$. Post hoc analyses reveal that the *positive* conditions differ from the negative ones ($p < .05$) but the *congruent* conditions do not differ significantly from the incongruent ones. The relative proportion of answers also differs across the

5 Grosjean (1996) reports on a study that used this threshold as a recognition point, although there is no consensus about which threshold reflects the "real" recognition point.

gates, $F(5, 195) = 47.138$, $p < .001$, $\eta^2_p = .547$. Post hoc analyses reveal that all gates differ significantly from each other, $p < .01$, except gate 4 and 5, $p = 1$. Finally, there is an interaction between *condition* and *gate*, $F(15, 585) = 2.914$, $p < .01$, $\eta^2_p = .070$.

We also performed univariate analyses within a condition, with gate as within-subjects factor and proportion of answers as the dependent variable, in order to see how the relative proportion of answers across the gates differs between *positive* and negative emotions, both *congruent* and incongruent. Within the INCONGRUENT NEGATIVE condition, $F(5, 195) = 33.529$, $p < .001$, $\eta^2_p = .462$, post hoc analyses show that gates 1 to 4 differ significantly from each other, $p < .05$. Within the NEGATIVE condition, $F(5, 195) = 34.622$, $p < .001$, $\eta^2_p = .470$, gates 1 to 3 differ significantly from each other, $p < .001$. Within the POSITIVE condition, $F(5, 195) = 40.511$, $p < .001$, $\eta^2_p = .510$, gates 1 to 3 differ significantly from each other, $p < .01$, as well as gates 4 and 6, $p < .05$. Within the INCONGRUENT POSITIVE condition, $F(5, 195) = 30.774$, $p < .001$, $\eta^2_p = .441$, gates 1 to 3 differ significantly from each other, $p < .01$, as well as gates 3 and 6, $p < .05$.

Finally, we performed univariate analyses within gate 1, with condition as within-subjects factor and proportion of answers as the dependent variable, in order to see whether the differences between conditions are present from the beginning. For gate 1, $F(3,117) = 5.949$, $p < .01$, $\eta^2_p = .132$, post hoc analyses revealed that all conditions differ significantly from each other, $p < .05$, except the POSITIVE condition, which does not differ from any condition.

#### 4.6.2 Perceived emotional state

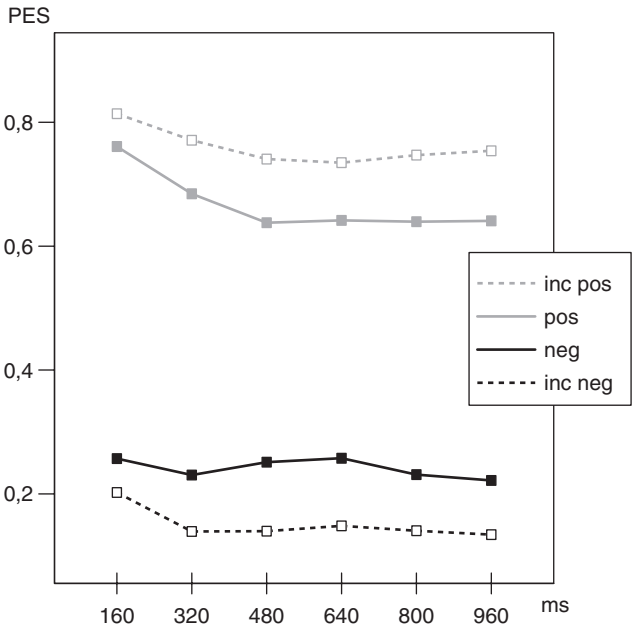
For this analysis, we recoded the original responses such that the "negative" responses obtained a value of 0, and the "positive" responses obtained a value of 1. The "don't know" responses were treated the same as the missing values. All these non-answers were subsequently replaced by the mean of the 10 presented speakers per segment/gate. We used this strategy because the "don't know" responses were already processed in the first step of the statistical analyses. In this *successive* step we want to know whether the distribution of *positive* versus negative answers differs across the conditions for all those cases where the participants were certain about their classification and therefore *did* choose an answer. So, while the first step reflects the level of uncertainty across all responses, this step reflects the 'correctness'[6] of the answers for all the 'certain' responses. For this analysis, there was a total of 1303 non-answers, which is 13.6% of all responses. Data are shown in Figure 5.

A repeated measures ANOVA, with condition and gate as within-subjects factors and perceived emotional state as the dependent variable, shows that *condition* has a significant effect on the perceived emotional state, $F(3, 117) = 219.238$, $p < .001$, $\eta^2_p = .849$. Post hoc analyses reveal that all conditions differ significantly from each other, $p < .001$. It is interesting to observe that the *incongruent* emotions received more extreme mean classification scores than the congruent ones. Speakers in the INCONGRUENT POSITIVE condition are overall classified as the most positive, $M = 0.76$, $SE = 0.018$, and speakers in the INCONGRUENT NEGATIVE condition are classified as the most negative, $M = 0.15$, $SE = 0.021$. The perceived emotional state also differs across gates, $F(5, 195) = 9.689$, $p < .001$, $\eta^2_p = .199$. Post hoc analyses show that only

6 Therefore, this level is comparable with the variable word candidates in the standard gating paradigm.

**Figure 5**

The mean perceived emotional state (0 = negative, 1 = positive) as a function of gate for different emotions




gate 1 differs significantly from all other gates, $p < .05$. Finally, there is *no* interaction between *condition* and *gate*, $F(15, 585) = 2.036$, $p = .06$, $\eta^2_p = .050$.

As with the previous tests on relative proportion of answers, we also performed univariate analyses within a condition, with gate as within-subjects factor and perceived emotional state as the dependent variable. Within the INCONGRUENT NEGATIVE condition, $F(5,195) = 3.298$, $p < .05$, $\eta^2_p = .078$, post hoc analyses revealed no significant differences. Within the NEGATIVE condition, only gates 4 and 6 differ significantly from each other, $p < .05$; however, the overall effect of gate is *not* significant, $F(5, 195) = 0.867$, $p = .442$, $\eta^2_p = .022$. Within the POSITIVE condition, $F(5, 195) = 9.586$, $p < .001$, $\eta^2_p = .197$, only gate 1 differs significantly from all other gates, $p < .05$, except for gate 2, which does *not* differ significantly from any other gate.[7] Within the INCONGRUENT POSITIVE condition, $F(5, 195) = 4.736$, $p < .01$, $\eta^2_p = .108$, only gates 1 and 4 differ significantly from each other, $p < .05$. Therefore, it seems that in general, after gate 1, there are *no* substantial differences anymore in the classification patterns.

7 It is important to realize that these scores reflect the patterns after participants were certain about their classification, because the "don't know" responses were treated as non-answers. In the first step it was found that the recognition speed was faster for the positive than for the negative emotions. Therefore, it is possible that the more positive classification in the first gate reflects the part of the population that is more certain about their answers, i.e., that an interaction is possible between the level of confidence and the extremity of the responses.

Because the *confidence levels* do not change substantially either in gates 4 to 6, it is interesting to look at the classification patterns within the first three gates. To test this, we performed a repeated measures ANOVA, with condition and gate as within-subjects factors and perceived emotional state as the dependent variable, within the first three gates. Here, the effect of *condition* is again significant, $F(3, 117) = 212.042$, $p < .001$, $\eta^2_p = .845$, as well as the effect of *gate*, $F(2, 78) = 10.551$, $p < .001$, $\eta^2_p = .213$. Post hoc analyses showed that only gate 1 differs significantly from gates 2 and 3, $p < .01$. So, it seems that there is a *transition point* at gate 2, which can be compared with the *isolation point* in the standard gating paradigm. There was again *no* interaction between *condition* and *gate*, $F(6, 234) = 2.261$, $p = .06$, $\eta^2_p = .055$.

Finally, because we were interested in the effect of condition within gate 1, we performed a univariate ANOVA with condition as within-subjects factor and perceived emotional state as the dependent variable, in order to explore how participants recognize emotions within the shortest time interval. Within the first gate, the effect of *condition* is significant, $F(3, 117) = 127.729$, $p < .001$, $\eta^2_p = .766$. Post hoc analyses show that the *positive* conditions (i.e., the positive and the incongruent positive) differ from both negative ones, $p < .05$, but the *congruent* conditions (i.e., the positive and the negative) do not differ from the incongruent conditions. The *positive* conditions are correctly classified as more positive (congruent: $M = 0.76$, $SE = 0.027$; incongruent: $M = 0.81$, $SE = 0.027$) and the *negative* conditions are correctly classified as more negative (congruent: $M = 0.26$, $SE = 0.036$; incongruent: $M = 0.20$, $SE = 0.03$).

### 4.7 Summary

In this study, we used a *gating paradigm* to test the recognition speed for various emotional expressions from a speaker's face. Participants were presented with video clips of speakers who displayed positive or negative emotions, which were either congruent or incongruent. Using a gating paradigm, the clips were shown in successive segments which increase in length.

We first calculated the *confidence* scores, which are the number of times that the subjects made a classification related to the number of times that they could not yet make a classification. We found the most non-answers for the first gate, and the *congruent* emotions got more non-answers than their incongruent counterparts. Further, in all conditions, the percentage of answers increased over the next gates, and reached a plateau after the fourth gate (640 ms). Also, the proportion of answers increased faster for the *positive* than for the negative emotions.

Next, we analyzed the *valence* of answers. Results show that participants are surprisingly accurate in their recognition of the various emotions, as they already reach high recognition scores in the first gate (after only 160 milliseconds). Interestingly, this recognition plateau is reached earlier for *positive* than negative emotions. Finally, *incongruent* emotions get more extreme recognition scores than congruent emotions, and already after a short period of exposure, perhaps because the incongruent recordings contain more expressive displays.

Given the previous two perception experiments, the next section discusses an observational analysis that aims to find possible visual correlates of emotional expressions, both in the upper and lower area of the face.

## 5 Observational analyses

To gain further insight into which facial cues could have influenced the subjects' categorization, we annotated all fragments in terms of a number of facial features. Although much is known about the prototypical expressions of emotions (Ekman, 1993), less is known about the difference in facial cues displayed in *congruent* and incongruent emotions (Wilting et al., 2006). Also, while past research has shown which facial cues are prototypical for pictures of emotions of joy and sadness, a second question is whether temporal dynamics such as the *duration* and the *intensity* of these cues can be successful in distinguishing between these positive and negative emotions, as these dynamics have already been shown to be successful in signaling the difference between congruent and incongruent displayals (Cohn & Schmidt, 2004; Valstar et al., 2006). Because temporal aspects of facial features are extremely difficult to assess manually and often require the use of advanced computer models (Cohn & Katz, 1998; Valstar et al., 2006), we chose to annotate solely whether or not a (number of chosen) feature(s) occurred, and the *subjective* intensity of these cues, rather than their exact duration and amplitude.

We concentrate on a small set of features. The chosen features are roughly comparable with Action Units described by Ekman and Friesen (1978), though there is not necessarily a one-to-one mapping to these Action Units. The choice of these features was based upon two restrictions: we wanted to score the *upper* as well as the *lower* face, and further we chose a set of features we assumed to reflect a *positive* as well as a *negative* emotion.

For the *upper* face we chose the following two features:

1. *Raising the brows*. This feature resembles the Action Unit combination 1 + 2.
2. *Frowning upwards*, that is, raising the brows and frowning. This feature resembles the Action Unit combination 1 + 4.

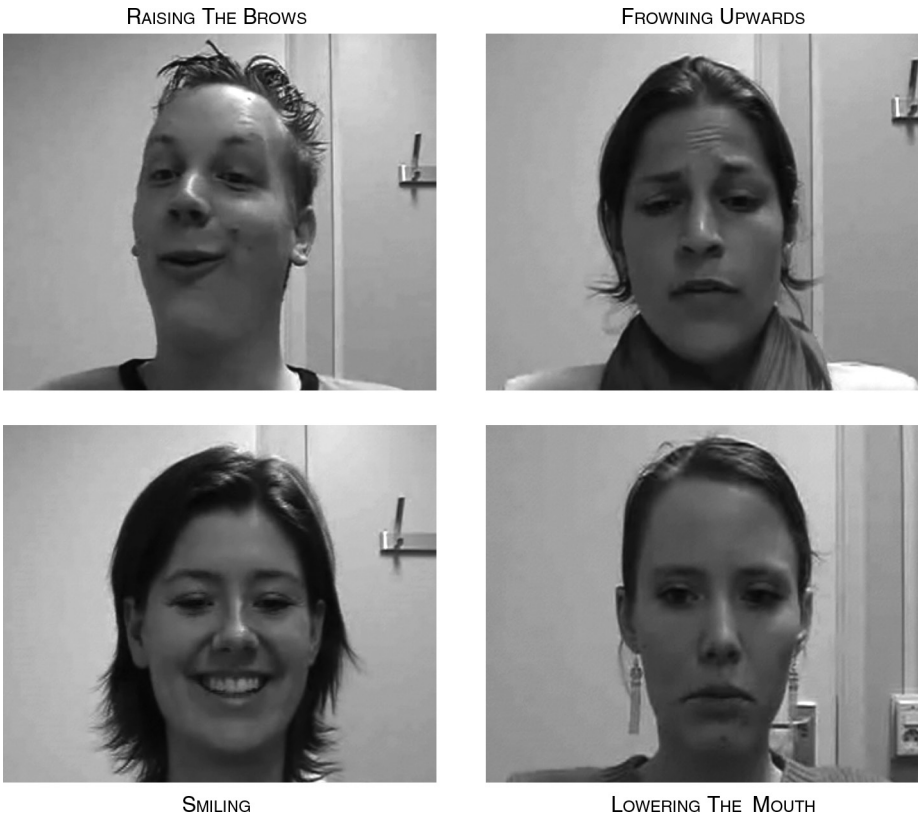
For the *lower* face we chose the features:

3. *Smiling*, that is, pulling the corners of the mouth aside and up. This feature resembles the Action Unit 12.
4. *Lowering the mouth*, that is, pulling the corners of the mouth down. This feature resembles the Action Unit 15.

The labeling was performed by three judges, the first author of this article and two independent Ph.D. students, who were unfamiliar with the purpose of the current study, but who were experienced with visual annotations. The procedure was as follows. The judges watched the film fragments and labeled them using the set of four features. Each judge labeled each feature individually. The labeling process took place blind for condition. We asked the labelers to score the maximum intensity that the feature reached in the entire film clip. The presence of the feature was largely determined on the labelers' subjective impression of whether the feature occurred or not. Each feature was given a number between 0 and 2 to reflect different strengths, where 0 stands for a complete absence and 2 represents a very clear presence of the facial feature. The scores for the features were subsequently summed across the three judges resulting in

**Figure 6**

Representative examples of the four annotated features: upper face (top) and lower face (bottom) expressions, with on the left-hand side the positive and on the right-hand side the negative versions




an overall score between 0 and 6 for the respective features. For instance, when coder 1 scored a 2, and the other two coders scored a 1, the overall score was a 4. This way of computing of the intensity by summing up the scores of the three labelers is consistent with the method of Hirschberg, Litman, and Swerts (2004) and Barkhuysen, Krahmer, and Swerts (2004) to label auditory and visual degrees of hyperarticulation.

For each labeled feature, we computed the Pearson correlation. The correlation was significant between all three coders for all the four features (*raising the brows*, $r_{12} = 0.61$, $p < .01$; $r_{13} = 0.67$, $p < .01$; $r_{23} = 0.74$, $p < .01$; *frowning upwards*, $r_{12} = 0.76$, $p < .01$; $r_{13} = 0.56$, $p < .01$; $r_{23} = 0.41$, $p < .01$; and *smiling*, $r_{12} = 0.68$, $p < .01$; $r_{13} = 0.74$, $p < .01$; $r_{23} = 0.77$, $p < .01$). The correlation was somewhat lower for *lowering the mouth*, $r_{12} = 0.38$, $p < .01$; $r_{13} = 0.35$, $p < .01$; $r_{23} = 0.44$, $p < .01$, but still significant.

### 5.1 Results

First of all, we present the general distribution of responses across the conditions in Table 3. According to Table 3, the two features within either the upper (brows) or lower

**Table 3**

Distribution of utterances from experiment I in terms of their mean scored intensity (standard errors in brackets) as a function of condition

| *Condition* | *Incongruent negative* | *Negative* | *Neutral* | *Positive* | *Incongruent positive* | *Total* |
|---|---|---|---|---|---|---|
| Raising the brows | 2.20 (.61) | 0.80 (.25) | 0.80 (.59) | 1.20 (.44) | 2.70 (.86) | 1.54 (.27) |
| Frowning upwards | 0.60 (.40) | 0.30 (.15) | 1.00 (.68) | 0.70 (.47) | 0.00 (.00) | 0.52 (.19) |
| Smiling | 0.40 (.40) | 0.20 (.13) | 0.40 (.16) | 2.90 (.71) | 3.70 (.54) | 1.52 (.28) |
| Lowering the mouth | 2.90 (.48) | 2.10 (.43) | 1.70 (.50) | 1.30 (.40) | 0.20 (.13) | 1.64 (.22) |

face (mouth) behave in an opposite way. Further, the intensity of the mouth is dependent upon condition, while the brows are independent from the valency of the condition.

### 5.1.1 Valency of the emotion of the speaker in the fragment

In this section, we explore to what extent there is a relation between the *valence* of the emotional state of the speaker in the fragment and the intensity of the annotated visual features described above. A univariate ANOVA was performed for each of the separate features, with condition as independent factor (INCONGRUENT NEGATIVE, NEGATIVE, NEUTRAL, POSITIVE, INCONGRUENT POSITIVE) and the feature as dependent factor (RAISING THE BROWS, FROWNING UPWARDS, SMILING and LOWERING THE MOUTH). There was a significant effect of *condition* on SMILING, $F(4, 45) = 13.727$, $p < .001$, $\eta^2_p = .55$, in the sense that the intensity of SMILING increases in the (congruent as well as incongruent) *positive* conditions (congruent: $M = 2.9$, $SE = 0.446$, and incongruent: $M = 3.7$, $SE = 0.446$). Post hoc analyses revealed that for SMILING, the *positive* conditions differ from the negative ones, $p < .01$, but the congruent conditions do not differ significantly from their incongruent counterparts (e.g., POSITIVE did not differ from INCONGRUENT POSITIVE). Further, the NEUTRAL condition differed from the positive ones, $p < .01$. There was also a significant effect of *condition* on LOWERING THE MOUTH, $F(4, 45) = 5.940$, $p < .01$, $\eta^2_p = .346$, in the sense that the intensity of LOWERING THE MOUTH increases in the (congruent as well as incongruent) *negative* conditions (congruent: $M = 2.1$, $SE = 0.410$ and incongruent: $M = 2.9$, $SE = 0.410$). Post hoc analyses revealed that for LOWERING THE MOUTH only the INCONGRUENT POSITIVE condition differed from the two negative conditions. So, SMILING occurs more in the *positive* conditions, while LOWERING THE MOUTH occurs more often in the *negative* conditions (the latter only across congruent *and* incongruent conditions). This validates the data along with the well-known literature on facial expressions. The upper face did not vary consistently across conditions: the other two features were non-significant.

### 5.1.2 Incongruent vs. congruent emotions of the speaker in the fragment

Also, we are interested in whether there was a relationship between the intensity of these features and whether the speaker was displaying an emotional expression which was

incongruent with the lexical content of the utterance. Although the univariate ANOVA did not show an overall effect for raising the brows, inspection of Figure 6 and Table 3 tells us that the intensity of raising the brows tends to increase in the incongruent conditions (negative: $M = 2.2$, $SE = 0.586$ and positive: $M = 2.7$, $SE = 0.586$), while the other three features do not seem to have a correlation. In order to test this further, we performed separate *t*-tests for each feature. In these tests, both incongruent conditions (negative and positive) as an 'incongruent' group were compared with a second group containing the two congruent conditions. It was shown that indeed only the feature RAISING THE BROWS was significant, $t = -2.529$, d.f. = 38, $p < .05$. Therefore, the brows are raised more intensely in the incongruent conditions.

#### 5.1.3 Emotional intensity of each feature in the fragment as perceived by the judges

Next, we are interested in whether there is a relationship between the *intensity* of the annotated features for each fragment (as it was scored by the three coders) and the *perceived* emotional state of the fragment such as it was classified in perception experiment I (by the Czech judges). Figure 7 shows the mean intensity of the scored feature as a function of mean perceived emotional state (1 = very negative, 7 = very positive) in experiment I (in the VO condition). Again, the intensity of the mouth movements increases as the perceived valency of the emotional state grows stronger (in either direction), while the brows seem uncorrelated.

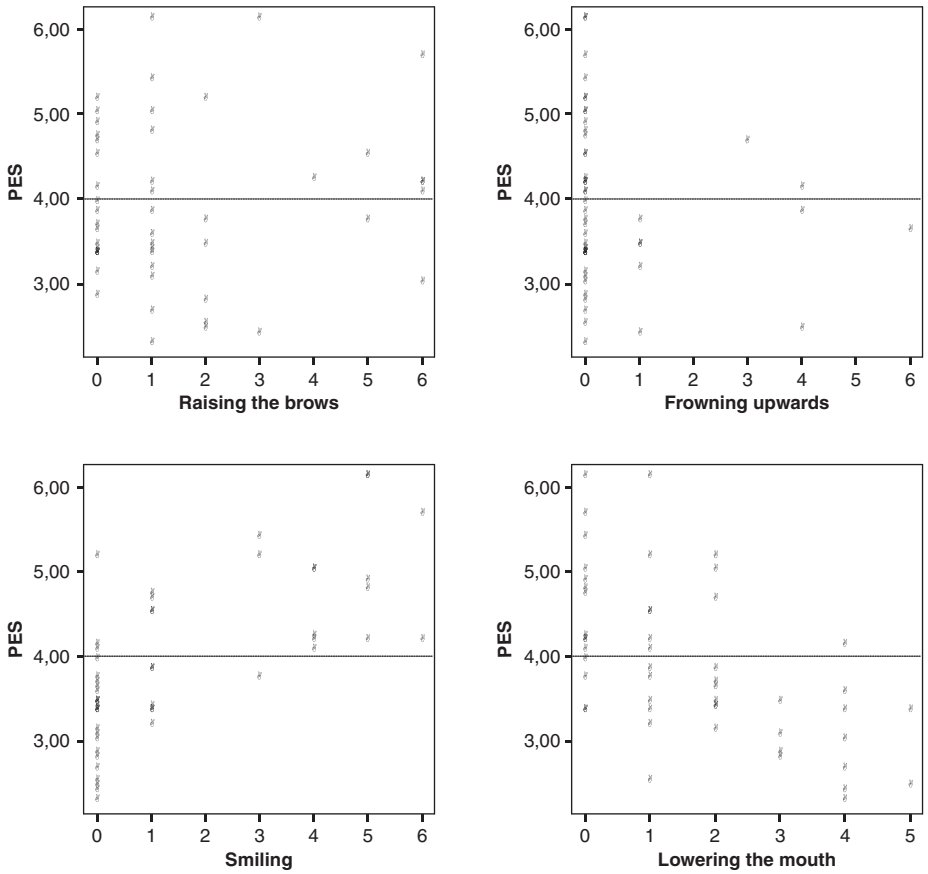
In order to test this, correlational analyses were performed between the four features and the mean perceived emotional state in experiment I (in the VO condition). The Pearson correlations for the features SMILING, $r = 70.4$, $p < .01$, and LOWERING THE MOUTH, $r = -58.4$, $p < .01$, were significant, though in opposite directions. The other two features were non-significant. Therefore, the more a fragment was perceived as positive, the more smiling occurred in the fragment. Vice versa, when the fragment was perceived as less positive, lowering the mouth was scored as more intense.

### 5.2 Summary

We were interested in the difference in occurrence of facial cues displayed in *positive* and negative conditions (both *congruent* and incongruent), and whether the intensity of facial cues can be useful for distinguishing between these conditions. The annotation analyses revealed that the occurrence of the features SMILING, LOWERING THE MOUTH and RAISING THE BROWS varies consistently across conditions. The data showed that SMILING and LOWERING THE MOUTH correlated with the perceived emotion: SMILING is scored as more intense in the positive conditions, while LOWERING THE MOUTH is scored as more intense in the negative conditions. Also, because RAISING THE BROWS is scored as more intense in the incongruent conditions, RAISING THE BROWS can be used to detect whether a speaker is displaying an emotion that is opposite to the lexical content of the sentence. Another question was whether there is a relationship between the emotional state of the fragment as it was perceived in experiment I, and the intensity of the annotated features as they were displayed in the fragments. The data showed that the more a fragment was perceived as positive, the higher the scored intensity of the SMILING was. Vice versa, when the fragment was perceived as less positive, LOWERING THE MOUTH was scored as more intense.

**Figure 7**

The mean perceived emotional state (1 = very negative, 4 = neutral, 7 = very positive) such as each fragment was classified by the Czech participants in experiment I, as a function of the mean intensity of each feature (0 = no intensity, 6 = very intense) for that fragment such as it was scored by the three coders




## 6 Discussion and conclusion

In this article, we investigated whether dynamic auditory and visual cues of emotional speech differ in perceptual strength, and how people deal with input coming from two modalities when they have to make judgments about a speaker's emotional state. In addition, we were interested in how fast people would recognize various emotions when presented with fragments of speech. Previous research has brought to light that listeners can successfully infer the emotional state of a speaker using information from different modalities (see e.g., Adolphs, 2002; Bachorowski, 1999; Banse & Scherer, 1996; Carroll & Russell, 1996; Scherer, 2003; Schmidt & Cohn, 2001). However, while there is much insight into how unimodal stimuli (either auditory or visual) are processed, less is known about the extent to which these modalities interact with each other.

Also, while much research has been done in the field of audiovisual speech processing, less work has been done about crossmodal integration in the context of *emotional speech*. Next, there is more knowledge available about online auditory speech than about online visual speech, because many studies combined a dynamic speech signal with *static* facial images. In order to answer such research questions, we collected utterances in a semi-spontaneous way using a experimental paradigm eliciting *positive* and *negative* emotions. In this paradigm, the participants, while being videotaped, had to reproduce sentences increasing in emotional strength. The display of the negative or positive emotions could be congruent or incongruent with the lexical content of the sentences. Using these utterances, two perception experiments were carried out.

The first experiment was a classification experiment with Czech participants to make sure that the participants could not rely on lexical cues. These participants were confronted with a selection of the recorded fragments, presented in three formats: AUDIOVISUAL (AV), VISION-ONLY (VO) and AUDIO-ONLY (AO). The task for participants was to indicate on a seven-point scale whether the speaker in the fragment was in a *positive* or a *negative* emotion. It was found that the highest scores were found in the AV and VO modalities, suggesting that the positive emotions are more clear in the visual modality, while the lowest scores were found in the AV and AO modality, suggesting that the negative emotions are more clear in the auditory modality. This is consistent with other findings (Scherer, 2003, pp.235–236). Further, the AV modality was always scored best, suggesting that the combination of two modalities contains more information than a single modality, although the difference between the AV modality and the two single modalities was not significant. We also compared the classification of the Czech participants for the VO fragments with the results of the Dutch perception test with the same stimuli (Wilting et al., 2006), which lead to essentially the same results. Therefore, it seems that the recognition of emotions was not influenced by cultural differences (or by the fact that the Czech language may use different intonational patterns). See Elfenbein and Ambady (2003) for more discussion on such issues.

A second question we explored in this article is to what extent the recognition of emotion varies as a function of the *time* that people are exposed to the facial expressions of a speaker. In order to answer this question, a second experiment was conducted. In a gating experiment participants were offered short parts of the original fragments increasing in length, from 160 ms (4 video frames) to 960 ms (24 video frames). After each gate participants had to indicate whether they believed that the speaker was in a positive or negative mood, and whether they could make the distinction on the basis of the current gate. The results showed that the participants already reached high recognition scores in the first gate. The confidence of the participants, determined as the moment where they chose either a positive or a negative emotion rather than the neutral option, reached a plateau in the fourth gate. Interestingly, this recognition plateau is reached earlier for positive than negative emotions, which is comparable to the valency effects reported by Leppänen and Hietanen (2004). It is interesting to consider that in the latter experiment people need 635 ms processing time to correctly classify a picture of a happy face (95.5%), while in the current experiment 160–480 ms of information seems to be sufficient for classifying a film clip of a speaker in a positive state. As our *confidence* scores reach a plateau after 640 ms, which is consistent with

the scores reported by Leppänen and Hietanen (2004), it might be useful to make a distinction between the capability of correctly classifying an emotion, which is already possible after only 160 ms, and the confidence people have in their ability to make a correct classification, which reaches the top level only after 640 ms.

To ensure the ecological validity of the emotions studied, one has to consider several problems. A problem with many emotion studies is that they often rely on "acted" data. The work of Ekman (e.g., 1993; Ekman et al., 1987), for instance, is based on posed photographs of actors; actors are also frequently used in speech research. Additionally, the comparison of the role of different modalities is often investigated by using congruent versus incongruent speech analogous to McGurk tasks. This raised the question whether the incongruent emotions are representative of acted, voluntary emotions or whether they are representative of real, spontaneous emotions. Wilting et al. (2006) addressed this problem by creating an "acting" condition: by asking the participants to display an emotion that was opposite to the lexical content of the sentences in the Velten task, such "incongruent" sentences become similar to "acted emotions" as speakers are displaying an emotion they are not feeling. The participants in the congruent task, on the other hand, were free to express the emotion invoked by the sentences. We can be sure that they were indeed feeling the congruent emotion because Wilting et al. (2006) tested which emotion they felt by presenting a survey afterwards. Although the survey indicated that the participant's emotions in the incongruent conditions was not different from the neutral condition, it would be interesting to further refine this test in the future, for example, to find out whether there is indeed an absence of emotion or whether they may have started to feel a mixture of emotions. It would be nice if future studies could supplement the current study with findings of brain research or arousal measures such as galvanic skin response.

The first perception test showed that incongruent emotional speech leads to significantly more extreme perceived emotion scores than congruent emotional speech, while the difference between incongruent and congruent speech is larger for the negative than for the positive emotions. This is in line with past research (Wilting et al., 2006), suggesting that incongruent emotions are perceived as more intense than congruent ones (possibly because they are displayed more intensely). It is interesting to note, though, that especially the negative incongruent expressions appear to be "ironic," which may have been caused by the mismatch between the form and the lexical content (see e.g., Attardo, Eisterhold, Hay, & Poggi, 2003, for a discussion about multimodal markers of irony). It would be interesting to replicate the experiment in the future, where the participants have to utter a sentence containing a neutral lexical content after the last sentence of the (positive or negative) list, which may be used in the perception studies instead. Further, de Gelder, and Vroomen (2000) report on the relative importance of the face above the voice for judging a (portrayed) emotion. Here, the difference between the two incongruent emotions was indeed somewhat larger in the VO modality than in the AO modality. Another interesting point is the difference between facial and vocal expression within the separate conditions. It seems that the Velten procedure did not elicit recognizable vocal expressions in the POSITIVE condition, whereas it elicited recognizable facial expressions. On the other hand, the senders in the INCONGRUENT POSITIVE condition were able to display recognizable facial *and* vocal expressions. According to de Gelder and Vroomen (2000), there are differences

in the effectiveness with which the face and the voice convey different emotions. The recognition of happiness, for example, remains accessible when the face is presented upside down, and also in focal brain damage patients where the recognition of several facial expressions is impaired, while, on the other hand, in the voice happiness is sometimes hard to tell apart from other emotions. Our results suggest that happiness in the voice can be detected when the senders are *acting* that they are happy, while, in fact, they do not necessarily feel that way.

The second perception test showed that the *incongruent* emotions received these more extreme recognition scores already after a short period of exposure. The gating results confirm earlier findings where incongruent emotions are perceived as more intense than congruent emotions (Wilting et al., 2006), as in the current experiment the former get more extreme recognition scores than the latter, and already after a short period of exposure, perhaps because the incongruent recordings contain more expressive displays. Horstmann (2002) reported that prototypical emotions resemble the most intense form of expressing an emotion. Perhaps when acting an incongruent emotion, the senders tend to display more prototypical expressions, in contrast to when they are free to express spontaneously whatever emotion they are feeling.

To gain further insight into which facial cues could have influenced the subjects' categorization, we annotated all fragments in terms of a number of facial features. According to some models (Cohn & Schmidt, 2004; Valstar et al., 2006), dynamic facial expressions consist of an initial onset phase, a peak, and an offset phase. In the *onset phase* of an expression, the facial muscles contract until the facial expression reaches its *apex*. In the next phase, the facial feature is at its peak and does not change any further until the start of the *offset* phase. Here, the facial muscles start to relax until the facial expression has returned to its neutral position (Valstar et al., 2006). The onset phase is usually very quick, ranging from 0.40 to 0.70 seconds in the case of smiles (Cohn & Schmidt, 2004). The subjects in our experiments needed only 160–480 ms for classifying a film clip of a speaker in a positive state, and their confidence scores reach a plateau after 640 ms, equalling the duration of an average onset phase. However, it is perfectly possible that displayed facial cues in the fragments were already at their apex, as we captured the speakers at the height of the induced emotion (by using only the last sentence of the list as a stimulus in the perception test).

We chose to annotate solely whether or not a (number of chosen) feature(s) occurred, and the *subjective intensity* of these cues, rather than their exact duration and amplitude. These features were RAISING THE BROWS, FROWNING UPWARDS, SMILING and LOWERING THE MOUTH. The occurrence of two other possible candidates, that is, gaze and head movements, was too low, but these features seem to be correlated with end-of-utterance marking (Barkhuysen, Krahmer, & Swerts, 2008). We felt that the intensity of the scored features is a reflection of the displayed apex in the offered fragments. We investigated to what extent there is a relation between the valence of the emotional state of the speaker in the fragment and the annotated visual features described above, that is, whether the *intensity* of facial cues can be successful in distinguishing between positive and negative emotions. It was shown that the intensity of the mouth was correlated with the intensity of the perceived emotion, in that when the mouth is lowered, the fragment is perceived as more negative, while the fragment is perceived as more positive when the mouth is smiling.

Further, we expected that the final intensity of the displayed cue can discriminate between congruent, "spontaneous," and incongruent, "acted" emotions, because posed smiles have a smaller amplitude (e.g., Cohn & Katz, 1998), and also the intensity of brow actions has been shown to be successful for distinguishing between spontaneous and posed expressions (Valstar et al., 2006), although it is not clear in what direction this relationship was. Our data showed that only raising the brows tends to increase in the incongruent conditions.

Next, we were interested in whether there is a relationship between the emotional state of the fragment as it was perceived in perception experiment I, and the intensity of the annotated features as they were displayed in the fragments. The data showed that the more a fragment was perceived as positive, the more smiling occurred in the fragment. Vice versa, when the mouth was lowered more intensely, the fragment was perceived as less positive.

Possibly, the *configuration* of features may be more important than simply distinguishing "which feature is responsible for what." Neurological research shows that faces are processed as a whole, apart from the processing path of individual features (Adolphs, 2002), and further there are even more specialized routes for the processing of moving faces, that is, dynamic, changeable configurations of facial features (Adolphs, 2002; Humphreys et al., 1993), although there are multiple interactions between the several pathways (Vuillemier & Pourtois, 2007). Also, the timing and coordination of the various regions of the face are usually off the mark in posed expressions (Ekman & Friesen, 1978). However, based upon the annotation results it is very likely that at least information from the *mouth* could have been very useful. Although the upper face in general, in particular the eyes, is reported as the most important source for emotion recognition, combining vocal expressions with facial expressions may draw attention to the mouth, unintentionally making the lower part of the face the most important source (de Gelder & Vroomen, 2000). It is therefore possible that in emotional *speech*, other facial features are more important than in emotional expressions without speech.

## References

ADOLPHS, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Review*, **1**(1), 21– 61.

ATTARDO, S., EISTERHOLD, J., HAY, J., & POGGI, I. (2003). Multimodal markers of irony and sarcasm. *Humor: International Journal of Humor Research*, **16**(2), 243–260.

AUBERGÉ, V., & CATHIARD, M.-A. (2003). Can we hear the prosody of smile? *Speech Communication*, **40**(1–2), 87–97.

BACHOROWSKI, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, **8**(2), 53–57.

BANSE, R., & SCHERER, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, **70**(3), 614–636.

BARKHUYSEN, P., KRAHMER, E., & SWERTS, M. (2004). Problem detection in human–machine interactions based on facial expressions of users. *Speech Communication*, **45**(3), 343–359.

BARKHUYSEN, P., KRAHMER, E., & SWERTS, M. (2008). The interplay between the auditory and visual modality for end-of-utterance detection. *The Journal of the Acoustical Society of America*, **123**(1), 354–365.

CALVERT, G. A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, **11**, 1110–1123.

CALVERT, G. A., BRAMMER, M. J., & IVERSEN, S. D. (1998). Crossmodal identification. *Trends in Cognitive Sciences*, **2**(7), 247–253.

CARROLL, J. M., & RUSSELL, J. A. (1996) Do facial expressions signal specific emotions? Judging emotions from the face in context. *Journal of Personality and Social Psychology*, **70**(2), 205–218.

COHN, J. F., & KATZ, G. S. (1998, September). Bimodal expression of emotion by face and voice. Paper presented at the Workshop on Face/Gesture Recognition and Their Applications, The Sixth ACM International Multimedia Conference, Bristol, UK.

COHN, J. F., & SCHMIDT K. L. (2004). The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, **2**, 1–12.

de GELDER, B., BOCKER, K., 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., & VROOMEN, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, **14**(3), 289–311.

EKMAN, P. (1993). Facial expression and emotion. *American Psychologist*, **48**, 384–392.

EKMAN, P. (2004). *Emotions revealed: Recognizing faces and feelings to improve communication and emotional life*. New York: Owl Books.

EKMAN, P., DAVIDSON, R. J., & FRIESEN, W. V. (1990). The Duchenne smile: Emotional expression and brain physiology II. *Journal of Personality and Social Psychology*, **58**(2), 342–353.

EKMAN, P., & FRIESEN, W. V. (1978). *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, CA: Psychologists Press.

EKMAN, P., FRIESEN, W. V., & ELLSWORTH, P. (1972). *Emotion in the human face: Guidelines for research and an integration of findings*. New York: Pergamon Press.

EKMAN, P., FRIESEN, W. V., O'SULLIVAN, M., CHAN, A., DIACOYANNI-TARLATIS, I., HEIDER, K., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, **53**(4), 712–717.

ELFENBEIN, H., & AMBADY, N. (2003). When familiarity breeds accuracy: Cultural exposure and facial emotion recognition. *Journal of Personality and Social Psychology*, **85**(2), 276–290.

FOX, E., LESTER, V., RUSSO, R., BOWLES, R., PICHLER, A., & DUTTON, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently? *Cognition and Emotion*, **14**(1), 61–92.

GAZZANIGA, M. S., & SMYLIE, C. S. (1990). Hemispheric mechanisms controlling voluntary and spontaneous facial expressions. *Journal of Cognitive Neuroscience*, **2**(3), 239–245.

GHAZANFAR, A. A., MAIER, J. X., HOFFMAN, K. L., & LOGOTHETIS, N. K. (2005). Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience*, **25**(20), 5004–5012.

GRAF, H. P., COSATTO, E., STRÖM, V., & HUANG, F. J. (2002, May). Visual prosody: Facial movements accompanying speech. Paper presented at the Fifth IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA.

GROSJEAN, F. (1996). Gating. *Language and Cognitive Processes*, **11**(6), 597–604.

GROSSMANN, T., STRIANO, T., & FRIEDERICI, A. (2006). Crossmodal integration of emotional information from face and voice in the infant brain. *Developmental Science*, **9**(3), 309–315.

HIETANEN, J. K., MANNINEN, P., SAMS, M., & RUSAKKA, V. (2001). Does audiovisual speech perception use information about facial configuration? *European Journal of Cognitive Psychology*, **13**(3), 395–407.

HIRSCHBERG, J., LITMAN, D., & SWERTS, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, **43**, 155–175.

HORSTMANN, G. (2002). Facial expressions of emotion: Does the prototype represent central tendency, frequency of instantiation, or an ideal? *Emotion*, **2**(3), 297–305.

HUMPHREYS, G. W., DONNELLY, N., & RIDDOCH, M. J. (1993). Expression is computed separately form facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia*, **31**(2), 173–181.

KRAHMER, E. J., VAN DORST, J., & UMMELEN, N. (2004). Mood, persuasion and information presentation: The influence of mood on the effectiveness of persuasive digital documents. *Information Design Journal + Document Design*, **12**(3), 40–52.

LEPPÄNEN, J., & HIETANEN, J. K. (2004). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research/Psychologische Forschung*, **69**, 22–29.

MACKIE, D., & WORTH, L. (1989). Processing deficits and the mediation of positive affect in persuasion. *Journal of Personality and Social Psychology*, **57**, 27–40.

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

POURTOIS, G, de GELDER, B., VROOMEN, J., ROSSION, B., & CROMMELINCK, M. (2000). The time-course of intermodal binding between seeing and hearing affective information. *Cognitive Neuroscience*, **11**(6), 1329–1333.

RINN, W. E. (1984). The neuropsychology of facial expressions: A review of the neurological and psychological mechanisms for producing facial expressions. *Psychological Bulletin*, **95**(1), 52–77.

RINN, W. E. (1991) Neuropsychology of facial expression. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 3–30). Cambridge: Cambridge University Press.

RUSSELL, J. A., BACHOROWSKI, J., & FERNÁNDEZ-DOLS, J. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, **54**, 329–49.

SCHERER, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, **40**, 227–256.

SCHMIDT, K., & COHN, J. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Yearbook of Physical Anthropology*, **44**, 3–24.

SEKIYAMA, K., KANNO, I., MIURA, S., & SUGITA, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience Research*, **47**, 277–287.

VALSTAR, M., PANTIC, M., AMBADAR, Z., & COHN, J. (2006, November). Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. Paper presented at ICMI 2006, Banff, Canada.

VELTEN, E. (1968). A laboratory task for induction of mood states. *Behavior Research Therapy*, **6**, 473–482.

VUILLEUMIER, P., & POURTOIS, G. (2007). Distributed and interactive brain mechanisms during emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia*, **45**, 174–194.

WESTERMANN, R., SPIES, K., STAHL, G., & HESSE, F. W. (1996). Relative effectiveness and validity of mood induction procedures: A meta analysis. *European Journal of Social Psychology*, **26**, 557–580.

WILTING, J., KRAHMER, E., & SWERTS, M. (2006). Congruent vs. incongruent emotional speech. *Interspeech 2006*, Pittsburgh PA, USA.