Incremental perception of acted and real emotional speech

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

Department of Communication and Cognition, University of Tilburg, The Netherlands
{p.n.barkhuysen, e.j.krahmer, m.g.j.swerts}@uvt.nl

Abstract

This paper[1] reports on an experiment using the gating paradigm to test the recognition speed for various emotional expressions from a speaker's face. In a perception experiment, subjects were presented with video clips of speakers who displayed negative or positive emotions, which were either acted or real. The clips were shown in successive segments (gates) of increasing duration. Results show that subjects 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, the recognition speed is faster for positive than negative emotions, in line with comparable valency effects reported by Leppänen and Hietanen (2003). Finally, the gating results confirm earlier findings that acted emotions are perceived as more intense than true emotions (Wilting et al., 2006), as the former get more extreme recognition scores than the latter, already after a short period of exposure.

Index Terms: incremental perception, gating paradigm, emotional speech, facial expressions, Velten technique

1. Introduction

Facial expressions are often considered to be *windows to the soul* as 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. Much research into the recognition of emotional expressions has been based on analyses of static images, such as photographs or drawings (e.g. Ekman, 1972). As a result, little is known about the perception of emotions through "fleeting changes in the countenance of a face" (Russell et al., 2003). The question we want to explore in this paper 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, both for positive versus negative emotions and for real versus acted ones.

First, consider the potential difference in recognition speed for true versus acted emotions. In particular, there is some work into timing-related differences between posed and spontaneous smiles (also known as non-Duchenne and Duchenne smiles). From a corpus study, Valstar et al. (2006) conclude that these two can be distinguished on the basis of the velocity, duration and order of occurrence of brow actions. Similarly, Cohn and Schmidt (2004) report that spontaneous as opposed to posed smiles are slow in onset, can have multiple rises of the mouth corners, and are accompanied by other facial actions, either simultaneously or immediately following. In addition, the valency, i.e., whether an emotion is positive or negative, may matter as well. 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, whereas Leppänen and Hietanen (2003) report that positive facial expressions are recognized faster than negative ones. Potentially, the valency effect on recognition speed, in whichever direction, may partly be due to timing-related differences in facial expressions.

The aim of this paper is to look in more detail at the recognition speed of dynamic expressions of positive and negative emotions, both acted and true. It describes a perception experiment 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 them produce sentences with an increasing emotional strength (Velten, 1968). The next section first describes previous work by Wilting et al. (2006), whose general approach was adopted for the current paper. We present a brief summary of their method and results of an experiment in which they first elicit real and acted emotional data from speakers using an adaptation of the Velten technique, and then selected film clips (without sound) which they showed to observers who have to judge the perceived emotional state of the recorded speakers. The later sections describe how the current study extends Wilting et al.'s research by testing the same experimental stimuli with a *gating paradigm* (Grosjean, 1986).

2. Wilting et al. (2006)

Wilting et al. (2006) used an adapted Dutch version of the original Velten (1968) technique, using 120 sentences evenly distributed over three conditions (POSITIVE, NEUTRAL and NEGATIVE). Besides these three conditions described by Velten for the induction of real emotions, 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 (ACT POSITIVE); in the other, positive sentences were shown and participants were instructed to utter these in a negative way (ACT 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.

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

[1] We thank Jean Vroomen for allowing us to use the PAMAR experimentation software. Thanks to Lennard van de Laar for technical support. This research was conducted within the framework of the FOAP project (http://foap.uvt.nl/), sponsored by NWO, the Netherlands Organisation of Scientific Research.

Figure 1: *Representative stills of acted (bottom) and real emotional (top) expressions, with on the left hand side the positive and on the right hand side the negative versions.*

the speakers with a digital camera, and a microphone connected to the camera. 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. Fifty Dutch speakers (10 per condition) were recorded in the data collection, 31 female and 19 male, none of them being an actor. Some representative stills are shown in Figure 1.

Wilting et al. (2006) reported 2 main findings. First, 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 non-acted conditions; the speakers in the acted conditions on average 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 acted versions led to more extreme scores than the non-acted ones, which suggests that the acted emotions were displayed more strongly than the non-acted ones. This raises the question in what sense the acted emotions differ from their non-acted counterparts. In this paper, we investigate the hypothesis that one difference is durational, especially in the onset, assuming that acted emotions appear quicker on the face than non-acted ones, though this may be different for positive versus negative emotions.

3. Perception test

3.1. Gating paradigm

The perception test is based on the *gating paradigm*, which is a well-known design in spoken word recognition research (Grosjean, 1986). In this paradigm, a spoken language stimulus is presented in segments of increasing duration and subjects 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*), the *confidence ratings* at various points in time and the *word candidates* proposed after each segment.

The current perception test resembles this gating design in that we present only parts of the original sentences used in Wilting et al. (2006), with an increasing duration. The first segment is very short, only consisting of 4 frames (160 ms). The later segments increase in steps of 160 ms until the last, sixth segment which is 960 ms. Each segment *S+1* includes the preceding segment *S,* and extends it with 4 extra frames (or 160 extra ms). We only used 6 frames, 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.

3.2. Design

The experiment uses a repeated measurements design with *condition* (with levels: ACT NEGATIVE, NEGATIVE, POSITIVE and ACT POSITIVE) and *gate* (with levels: ONE (i.e., 160 ms.) to SIX (i.e., 960 ms.)) as within-subject factors, and *certainty* (with levels: non-answers "don't know" versus answers "positive or negative") and *perceived emotional state* (with levels: "positive" and "negative") as the dependent variables.

3.3. Procedure

Participants took part one at a time. 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 6 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 3 answering possibilities: "*negative*", "*don't know*", and "*positive*". Three buttons on the keyboard were labeled with these answer possibilities, and after viewing a film clip, participants could press one of these buttons, after

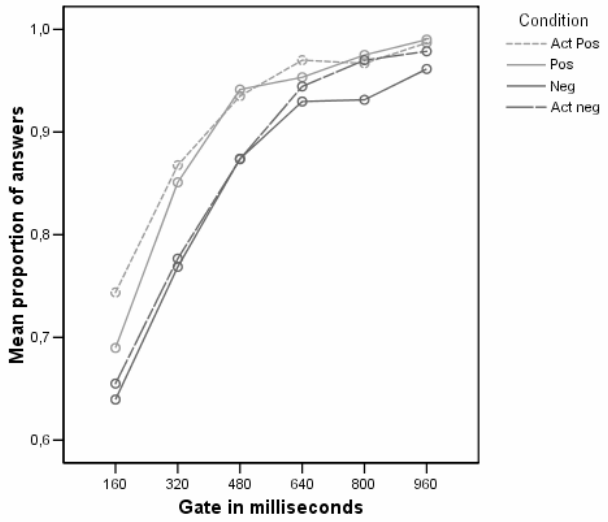

Figure 2: *The mean proportion of answers (vs non-answers) as a function of gate for different emotions.*

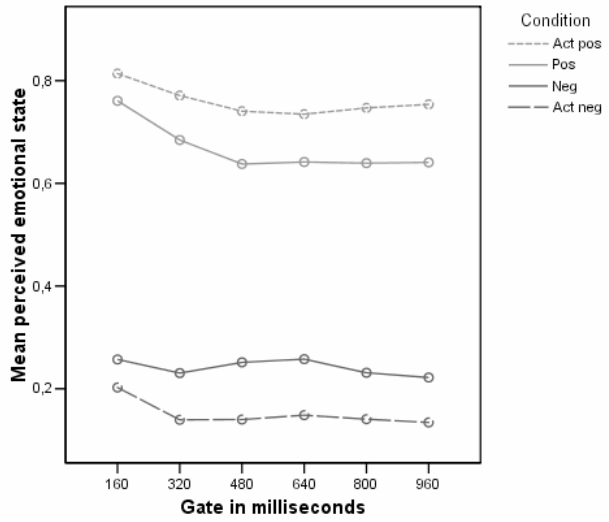

Figure 3: *The mean perceived emotional state as a function of gate for different emotions.*

which the next stimulus appeared. Participants were not informed about the fact that some of the speakers were acting.

The gates were presented in a *successive* format: that is, subjects viewed all the segments of a sentence, starting with the shortest and finishing with the longest. The gates were presented *forwards*, i.e. the first was cut from the beginning of the sentence while later segments were approaching the end ("left-to-right"). Stimuli (groups of 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 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 subjects' 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 1 stimulus group containing 6 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.

3.4. Participants

Forty people (10 per presentation order) participated in the experiment, with a roughly equal number of female and male participants. All were students from Tilburg University in The Netherlands, none had participated as a speaker in the study by Wilting et al. (2006), and all were ignorant of the experimental question.

3.5. 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 in section 3.5.1, next we look at the number of positive and negative answers as a function of gate in section 3.5.2.

3.5.1. 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 with the mean value over the 10 speakers per segment/gate. Figure 2 shows the percentage of answers as a function of gate. What this figure shows is that we find the most non-answers for the first gate, and the non-acted emotions get more non-answers than their acted 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 differs for positive versus negative emotions. Taking an 80% threshold, it can be seen that the recognition of *positive* emotions reaches this level already at gate 2 (real: $M = 0{,}83$, $SE = 0{,}028$; acted: $M = 0{,}87$, $SE = 0{,}025$), while the *negative* emotions reach this level only at gate 3 (real: $M = 0{,}87$, $SE = 0{,}031$; acted: $M = 0{,}87$, $SE = 0{,}026$).

A univariate analysis of variance shows that *condition* has a significant effect on the relative proportion of answers ($F(3, 117) = 8.051$, $p < .001$).[2] Post hoc analyses reveal that the negative conditions differ from the positive ones ($p < .05$) but the acted conditions do not differ significantly from the real ones. The relative proportion of answers also differs across the *gates* ($F(5, 195) = 47.138$, $p < .001$). 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* ($F(15, 585) = 2.914$, $p < .01$).

We also performed univariate analyses within a condition, in order to see how the relative proportion of answers across the gates differs between positive and negative emotions, both real and acted. For the ACT NEGATIVE condition, post hoc analyses show that gates 1 to 4 differ significantly from each other ($p < .05$). For the NEGATIVE condition, gates 1 to 3 differ significantly from each other ($p < .001$). For the POSITIVE condition, gates 1 to 3 differ significantly from each other ($p < .01$), as well as gates 4 and 6 ($p < .05$). For the ACT POSITIVE condition, 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, in order to see whether the differences between conditions are present from the beginning. For gate 1, post hoc analyses revealed that all

[2] Because the assumption of sphericity was violated (Mauchly's $W = 0.453$, $p < .001$), the degrees of freedom were adapted. For the sake of transparency, however, we report on normal degrees of freedom. All post-hoc analyses make use of the Bonferroni method.

conditions differ significantly from each other ($p < .05$) except the POSITIVE condition, which does not differ from any condition.

3.5.2. 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. For this analysis, there was a total of 1303 non-answers, which is 13,6% of all responses. Data are shown in Figure 3.

A univariate analysis of variance shows that *condition* has a significant effect on the perceived emotional state ($F(3, 117) = 219.238$, $p < .001$). Post hoc analyses reveal that all conditions differ significantly from each other ($p < .001$). It is interesting to observe that the acted moods are perceived as more intense than the real ones. Speakers in the ACT POSITIVE condition are overall perceived as the most positive ($M = 0.76$, $SE = 0.018$), and speakers in the ACT NEGATIVE condition are perceived as the most negative ($M = 0.15$, $SE = 0.021$). The perceived emotional state also differs across *gates* ($F(5, 195) = 9.689$, $p < .001$). Post hoc analyses show that only gate 1 differs significantly from all other gates ($p < .05$). Finally, there is *no* interaction between *condition* and *gate* ($F(15, 585) = 2.036$, $p = .06$).

As with the previous tests on relative proportion of answers, we also performed univariate analyses within a condition. For the ACT NEGATIVE condition, post hoc analyses revealed no significant differences. For the NEGATIVE condition, only gates 4 and 6 differ significantly from each other ($p < .05$), however the overall effect of gate is not significant ($F(5, 195) = 0.867$, $p = .442$). For the POSITIVE condition, only gate 1 differs significantly from all other gates ($p < .05$), except for gate 2, which does not differ significantly from any other gate. For the ACT POSITIVE condition, only gates 1 and gate 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. Because the certainty levels do not change substantially either in gates 4 to 6 (see section 3.5.1), it is interesting to look at the classification patterns within the first 3 gates. To test this, we performed a univariate analysis for the first 3 gates. Here, the effect of *condition* is again significant ($F(3, 117) = 212.042$, $p < .001$), as well as the effect of *gate* ($F(2, 78) = 10.551$, $p < .001$). Post-hoc analyses showed that only gate 1 differs significantly from gate 2 and 3 ($p < .01$). So, it seems that there is a transition point at gate 2. There was again no interaction between *condition* and *gate* ($F(6, 234) = 2.261$, $p = .06$).

Finally, it is interesting to see what the effect of condition is within gate 1, to explore how subjects recognize emotions within the shortest time interval. Within the first gate, the effect of *condition* is significant ($F(3, 117) = 127.729$, $p < .001$). Post hoc analyses show that the *negative* conditions differ from the *positive* ones ($p < .05$) but the *acted* conditions do not differ from the *real* conditions. The *positive* conditions are correctly classified as positive (real: $M = 0.76$, $SE = 0.027$; acted: $M = 0.81$, $SE = 0.027$) and the *negative* conditions are correctly classified as negative (real: $M = 0.26$, $SE = 0.036$; acted: $M = 0.20$, $SE = 0.03$).

4. Conclusions

This paper reported a *gating paradigm* to test the recognition speed for various emotional expressions from a speaker's face. In a perception experiment, subjects were presented with video clips of speakers who displayed negative or positive emotions, which were either acted or real. Using a *gating paradigm*, the clips were shown in successive segments of increasing duration. Results show that subjects 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, the recognition speed is faster for positive than negative emotions, in line with comparable valency effects reported by Leppänen and Hietanen (2003). It's interesting to consider that in their experiment people need 635 ms to correctly classify a picture of a happy face (95.5%), while in the current experiment 160-480 ms 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 (2003), 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 a person has in the correctness of this classification, which reaches the top level only after 640 ms. Finally, the gating results confirm earlier findings that acted emotions are perceived as more intense than true emotions (Wilting et al., 2006), as the former get more extreme recognition scores than the latter, already after a short period of exposure.

5. References

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

[2] 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.

[3] Ekman, P. (1972). *Emotion in the human face*. Pergamon Press.

[4] 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*, 61–92.

[5] Grosjean, F. (1996). Gating. *Language and Cognitive Processes, 11(6),* 597–604.

[6] 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.

[7] Russell, J.A., Bachorowski, J. & Fernández-Dols, J. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54*, 329–49.

[8] Valstar, M., Pantic, M., Ambadar, Z., & Cohn, J. (2006). Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. *ICMI 2006*, Banff, Canada.

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

[10] Wilting, J., Krahmer, E. & Swerts, M. (2006). Real vs. acted emotional speech. *Interspeech 2006*, Pittsburgh PA, USA.