CROSS-MODAL PERCEPTION OF EMOTIONAL SPEECH

Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts

University of Tilburg
p.n.barkhuysen,e.j.krahmer,m.g.j.swerts@uvt.nl

ABSTRACT

We report on a perception experiment in which Czech participants rate the perceived emotional state of Dutch speakers. These speakers could either display a positive or a negative emotion, which was either real or acted. The Czech participants were confronted with these utterances, which they could not understand, in a bimodal (audio-visual) or a unimodal (audio- or vision-only) condition. It was found that acted emotional speech leads to significantly more extreme perceived emotion scores than non-acted emotional speech, where the difference between acted and real emotional speech is larger for the negative than for the positive conditions. Interestingly, the largest overall differences between acted and non-acted emotions were found for the audio-only condition, which suggests that acting has a particularly strong effect on the spoken realization of emotions.

Keywords: crossmodal perception, emotional speech, facial expressions, Velten technique.

1. INTRODUCTION

Previous research has brought to light that listeners can succesfully infer the emotional state of a speaker using information from different modalities. In the auditory domain, it has been shown that listeners can use various vocal cues to assess the emotion of a speaker (e.g. Bachorowski, 1999; Scherer, 2003; Banse & Scherer, 1996). Research on visual cues revealed that a speaker's emotional state can also be detected from facial expressions or gestures (e.g. Adolphs, 2002; Carroll & Russell, 1996; Schmidt & Cohn, 2002).

Still, a number of open questions remain. First, many such studies rely on "acted" data. The work of Ekman (1972), for instance, is based on posed photographs of actors, and also in speech research actors are frequently used. This raises the question as to whether such data are *ecologically valid*, in particular whether acted emotions are representative of real emotions (Gazzaniga & Smylie, 1990). Second, while we have gained much insight into how unimodal stimuli (either auditory or visual) are processed, far less is known about the extent to which these modalities interact with each other (but see, e.g., Aubergé & Cathiard, 2003; Hietanen et al., 2001). 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. There is some neurological evidence that integration of emotional information from the face and from the voice occurs at an early stage of processing, and involves low-level perceptual features (de Gelder et al., 1999), but much remains to be done.

The aim of this paper is to look into more detail at the perception of audiovisual expressions of acted (incongruent) and real (congruent) emotions in spoken language (both of which can be positive and negative). 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, Krahmer & Swerts (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, and then selected film clips (without sound) which were shown to observers who had 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 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.

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 the three conditions described by Velten

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

for the induction of real 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 (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 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 a (professional) 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 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. The study by Wilting et al. (2006) was conducted with vision-only stimuli presented to Dutch participants. It was not 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 audio-visual stimuli. Therefore the Dutch sentences were presented to Czech participants in a perception test, which is described next.

3. PERCEPTION TEST

3.1. Design

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

3.2. 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 7 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 acting. 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

Figure 2: The mean perceived emotional state per condition and modality.

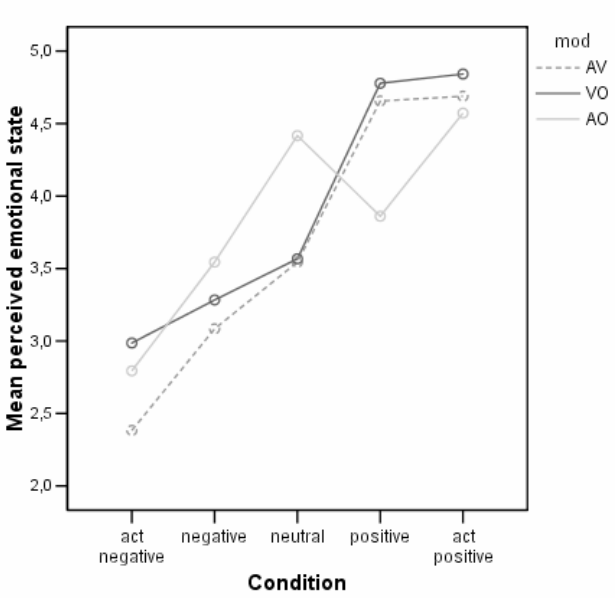
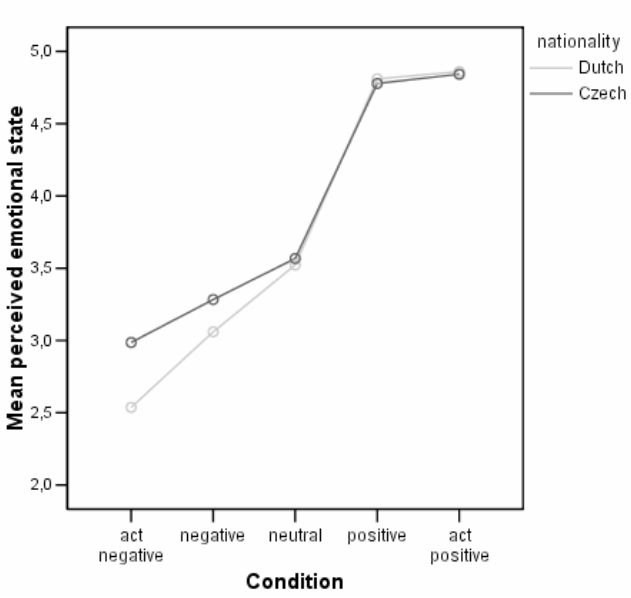


Figure 3: The mean perceived emotional state per condition and nationality.



shown only once. The experiment was preceded by a short training session consisting of 5 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 in front of the class room. The entire experiment lasted approximately 10 minutes.

3.3. Participants

Fifty-four people (18 per condition) participated in the experiment, 9 female and 45 male, with an average age of 23 (range 21-30). All were students and PhD-students from the Czech Technical University (Faculty of Electrical Engineering) and the Charles University (Faculty of Philosophy and Arts) in Prague, Czech Republic. None of them could understand Dutch.

3.4. Results

Figure 2 summarizes the results. A univariate analysis of variance (ANOVA) shows that *condition* has a significant effect on perceived emotional state ($F(4, 205) = 110.215$, $p < .001$).[1] Post hoc analyses using the bonferroni method reveal that all conditions lead to a significantly different perceived emotion ($p < .001$). It is interesting to observe that the acted emotions are perceived as more intense than the real ones. Speakers in the ACT POSITIVE condition are overall perceived as the most positive ($M = 4.70$, $SD = 0.53$), and speakers in the ACT NEGATIVE condition are perceived as the most negative ($M = 2.72$, $SD = 0.63$). Note that the perceptual difference between acted and non-acted emotional speech is larger for the negative emotions. In general, it seems that the acted emotions are "better" classified or interpreted as more intense than the real emotion.

Modality does not have a significant main effect on *perceived* emotional state ($F(2, 51) = 1.881$, $p = .163$), but interestingly there was an interaction between *condition* and *modality* ($F(8, 204) = 10.981$, $p < .001$). In all three modalities the acted moods are perceived as more intense than the real ones; speakers in the ACT POSITIVE condition are perceived as the most positive, and speakers in the ACT NEGATIVE condition are perceived as the most negative. However, contrasts showed that all levels of condition and modality interacted significantly with each other ($p < .01$). For both the AV and the VO modality the difference between POSITIVE and ACT POSITIVE is very small, while this difference is larger in the AO modality: 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 ACT NEGATIVE condition is substantial for the AO and the AV modality, but

[1] Because the assumption of sphericity was violated (Mauchly's $W = 0{,}126$, $p < .001$), the degrees of freedom were adapted. For the sake of transparency, however, we report on normal degrees of freedom.

here the VO modality stands out in the sense that the difference is relatively smaller. 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. Further, we compared the classification of the Czech participants for the fragments presented in the VO modality with the results of the Dutch perception test (Wilting et al., 2006). It turns out that the main effect of *nationality* was not significant ($F(1, 56) = 1.905$, $p = .173$). There was a significant interaction between *nationality* and *condition* ($F(4, 224) = 5.088$, $p < .01$); however, contrasts showed that this difference was only caused by the difference between the NEGATIVE and the ACT NEGATIVE stimuli ($F(1, 56) = 4.505$, $p = .038$).

4. CONCLUSION

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

It was found that acted emotional speech leads to significantly more extreme perceived emotion scores than non-acted emotional speech, where the difference between acted and real emotional speech is stronger for the negative than for the positive conditions. This is in line with past research (Wilting et al., 2006), suggesting that acted emotions are easier to recognize. Interestingly, the largest overall differences between acted and non-acted emotions were perceived in the audio-only condition, which suggests that acting has a particularly strong effect on the spoken realization of 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. 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 the emotions was not influenced by cultural differences (or by the fact that the Czech language may have different intonational prosody patterns). This is not implausible, considering that differences in facial expressions can be viewed as subtle differences in style, which become smaller by more frequent intercultural contact (Elfenbein & Ambady, 2003). An unsolved question remains whether acted emotions are easier to process than real emotions. This topic will be investigated in further research.

5. ACKNOWLEDGMENTS

This research was conducted within the framework of the FOAP project (http://foap.uvt.nl). The authors wish to acknowledge Karel Fliegel, Marie Nilsenova, and Carel van Wijk for various kinds of assistance.

6. REFERENCES

[1] Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. Beh and Cogn Neurosci Rev, 1(1), 21– 61.

[2] Aubergé, V., & Cathiard, M.-A. (2003). Can we hear the prosody of smile? Sp Comm, 40(1-2), 87-97.

[3] Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. J of Pers and Soc Psy, 70(3), 614-636.

[4] Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. Cur Dir in Psy Sci, 8(2), 53-57.

[5] Carroll, J., & Russell, J. (1996). Do facial expressions signal specific emotions? Judging emotion from the face in context. J of Pers and Soc Psy, 70(2), 205-218.

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

[7] Elfenbein, H. , & Ambady, N. (2003). When familiarity breeds accuracy: cultural exposure and facial emotion recognition. J of Pers and Soc Psy, 85(2), 276-290

[8] Gazzaniga, M. S., & Smylie, C. S. (1990). Hemispheric mechanisms controlling voluntary and spontaneous facial expressions. J of Cog Neurosci, 2(3), 239-245.

[9] 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. Neurosci Lett, 260, 133-136.

[10] Hietanen, J. K., Manninen, P., Sams, M., & Rusakka, V. (2001). Does audiovisual speech perception use information about facial configuration? Eur J of Cogn Psy, 13(3), 395-407.

[11] Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Sp Comm, 40, 227-256.

[12] Schmidt, K & Cohn, J (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. Yearb of Phys Anthr, 44, 3-24.

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

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