An Experimental Paradigm for Inducing Emotions in a Real World Driving Scenario: Evidence from Self-Report, Annotation of Speech Data and Peripheral Physiology

Alicia Flores Requardt*, Marc Wilbrink, Ingo Siegert*, Meike Jipp**, Andreas Wendemuth*, Klas Ihme****

**Otto-von-Guericke-University Magdeburg,Cognitive Systems Group, Universitätsplatz 2, 39106 Magdeburg, Germany (e-mail: {alicia.requardt, ingo.siegert, andreas.wendemuth}@ovgu.de).*
***German Aerospace Center (DLR), Institute of Transportation Systems, Lilienthalplatz 7, 38018 Braunschweig, Germany (e-mail:{marc.wilbrink, meike.jipp, klas.ihme}@dlr.de)*

Abstract: Empathic vehicles are a promising concept to increase the safety and acceptance of automated vehicles. However, on the way towards empathic vehicles a lot of research in the area of automated emotion recognition is necessary. Successful methods to detect emotions need to be trained on realistic data that contain the target emotion and come from a setting close to the final application. At the moment, data sets fulfilling these requirements are lacking. Therefore, the goal of this work is to present an experimental paradigm that induces four different emotional states (neutral, positive, frustration and mild anxiety) in a real-world driving setting using a combination of secondary tasks and conversation-based emotional recall. An evaluation of the paradigm using self-report data, annotation of speech data and peripheral physiology indicates that the methods to induce the target emotions were successful. Based on the insights of the experiment, finally a list of recommendations for the induction of emotions in real world driving settings is given.

Keywords: **Empathic vehicles, emotion recognition, annotation of speech data, frustration, physiology**

1. INTRODUCTION

Nowadays, the automotive industry is moving from manual driving over assisted driving towards highly automated driving. Most new vehicles are already equipped with advanced driver assistant systems (ADAS) that support the driver in critical situations caused by the vehicle environment, or the driver himself (e.g. collision avoidance, blind spot detection, lane-change assist) and increase the comfort of driving (e.g. adaptive cruise control, stop and go assist, parking assist). In the future, it is expected that vehicles will not only be able to take into account their environment, but also to monitor the drivers in order to adapt to their state and by this provide assistance and support tailored to their current needs. Generally, three types of vehicles are differentiated that describe the variant development stages:

The *cognitive car*, which perceives and analyzes the vehicle environment and traffic situation, monitors the interaction between driver, car and traffic and reacts in relevant situations (Heide and Henning, 2006; Gadsden and Habibi, 2009). With the availability of the abovementioned assistance functionalities, cognitive cars can already be seen as state-of-the-art implementation in the automotive industry. *Intelligent vehicles* are cognitive cars that are additionally able to monitor critical states of the driver, such as sleepiness and inattention, and react to it by warning the driver in dangerous circumstances or by partly/fully taking over control from the driver (Flemisch et al., 2013). Despite being a hot topic in research, the concept mostly neglects the presence of emotions and their importance in human-vehicle interaction. Thus, recently the concept of *empathic vehicles* has been coined. These are not only able to react appropriately to current critical driver states, but also detect the emotional state (also including stress) of the driver and respond empathically, for example by mirroring or balancing the emotions of the driver (Hernandez et al., 2014; Drewitz et al., 2017). Empathic vehicles can be seen as a future technology trend, which is obvious from the fact that their feasibility is currently investigated in the large-scale project ADAS&ME (http://www.adasandme.com/). Potential adaptation strategies currently discussed in research include the adaptation of the interior light to create a relaxing atmosphere, biofeedback or implementing spoken-dialog-system-based assistants that sympathize with the user or support through active listening (e.g. Plitnick et al., 2010; Löcken et al., 2017; Nass et al., 2005; Klein et al., 2002). In this field of research, mostly the focus is drawn towards automated driving. Then, the vehicle not only serves as a transportation means, but also as a companion technology, interacting with the driver in a human-like interaction (cf. Biundo and Wendemuth, 2017), for example using spoken dialog systems as already visible in present day premium segment vehicles. Still, at the moment fully autonomous vehicles are rather a future vision, so that the human in the car will, for safety reasons, likely need to be ready to take over the control of the car in different critical situations. Therefore, driver monitoring systems for safety will be relevant in the foreseeable future with speech potentially being one available indicator for safety critical driver states, such as certain emotions.

Thus empathic properties of vehicles are not merely a gadget for marketing, but serve an important purpose from a human factors engineering perspective (Drewitz et al., 2017). Drivers are similarly affected by emotionally challenging situations in manual driving as in automated driving. Emotions can influence cognitive processes of the driver which are relevant for the driving task, such as the built-up of a sufficient situation representation (Jeon, 2015) or decision making (Freese and Jipp, 2015), in both positive and negative ways. Recent research proposes that negative emotions, such as frustration and anxiety, but also positive emotions are relevant during driving because they have an effect on the driver. For example, anxiety, on one hand, may lead to an increase of situation awareness, such that the driver will adapt his driving behavior towards the given circumstances (Lu et al., 2013), while, on the other hand, it may cause a decrease of the driver's attention focus (Jeon et al., 2014). In contrast, frustration may lead to aggressive driving increasing the risk of causing an accident (Shinar, 1998). In general it can be stated that emotions often affect aspects of driving safety (Pêcher et al., 2011). For instance, Pêcher et al. (2009) show that not only negative, but also positive emotions can affect the driving performance in a negative way and may result in reckless driving behavior. Crucially, humans barely compensate for these effects, because, unlike when distracted by a smart phone or conversations, they are often not aware of the resulting impairments (Jeon, 2015). In consequence, the experience of strong emotions in the car can affect driving quality and thus reduces road safety. In addition, emotions, especially negative ones, impact the feeling of comfort and hence may also affect user experience as well as acceptance of technical systems (Picard and Klein, 2002). Therefore, empathic vehicles that detect these emotions and support the driver to balance their emotional states or reduce the potentially negative consequences of emotions, bear good prospects to increase safety and acceptance of highly automated driving.

Recognizing relevant emotions, such as positive emotional states (e.g. joy), frustration or mild anxiety, in automated driving is a challenging endeavor. A prerequisite for this is the availability of multimodal data of persons experiencing these emotions while driving in realistic scenarios, which needs adequate experimental designs inducing emotions in real-word settings. So far, most data sets come from simulator set-ups which offer easy to use, flexible and standardized ways to build experiments. However, the data coming from simulators are limited by the artificial environment, thus reducing the generalization of the acquired results. Therefore, the goal of the current study is to describe and evaluate an experimental paradigm for inducing four driving relevant emotional states (neutral, positive, frustrated and anxious) in a real vehicle set-up. Self-report, annotation of speech data as well as physiological signals will be used to evaluate the adequacy of the emotional scenarios.

2. METHODS

2.1 Design

The goal of the experiment was to evaluate the use of a combination of secondary tasks and conversation-driven emotional recall to induce the target emotional states neutral, positive, frustration and mild anxiety in a realistic driving setting. Therefore, participants drove four different driving scenarios that served the induction of the emotion in a within-design (see also Lotz et al., 2018 for details).

2.2 Participants

In sum, 30 volunteers (seven female) with a mean (*M*) age of 30.6 years (standard deviation [*SD*] = 5.0 years, range 22 to 41 years) participated in the experiment. All of them possessed a valid German driver's license and were standard German native speaker without speech-related or neurological disorders. For insurance reasons, all participants were employees of the DLR by the time of the experiment. The volunteers received a financial compensation of 30 € for their participation.

2.3 Ethics statement

The study procedure was in accordance with the guidelines of the German Aerospace Center (DLR) and approved by the ethics committee of Otto-von-Guericke-University Magdeburg (reference number 153/17). Before the start of the experiment, all volunteers provided written informed consent to participate.

2.4 Experimental set-up

The study was conducted on the DLR site in Braunschweig, Germany, which is a designated test ground for driving experiments. On the site, driving is allowed with a maximum speed of 30 km/h. As test vehicle, the research car FASCar II of the DLR was used (Fischer et al., 2014). The FASCar II is a vehicle for testing driver assistance systems and automated driving functions. For safety, an additional brake pedal is available at the passenger seat. This combination of test site and vehicle ensured natural driving experience and driving environment, comparable to quiet residential areas. A fixed driving course of roughly 900 meters on the available streets in the site was determined which took approx. 2.5 min to drive. To ensure comparability of all recordings, the data was collected during day light and under similar and constant weather conditions. Termination criteria were strong rain and/or thunderstorm. In addition to the participant, two further persons were in the car: One investigator sitting on the passenger seat, and one technician for the supervision of the sensor data recording sitting on the rear bench behind the passenger seat.

2.5 Driving scenarios

Each participant drove four experimental scenarios with the goal to induce the four different target emotional states (neutral, positive, frustration, and anxiety). In each scenario five rounds of the course had to be driven. In the first round, participants only had to drive the car without conversation or secondary task. In the remaining four rounds, a mixture of secondary task and emotional recall through conversations was used to induce the target emotional state. A cover story was designed to conceal the true nature of the experiment. This told participants that the study's purpose was to test recently developed assistance systems. The experimenter always initiated the conversations starting from the current situation (mostly the just accomplished secondary task). For this, a list of pre-defined questions and topics was available to keep the conversations alive; however, the goal was to individualize the conversations as much as possible in order to really stimulate the targeted emotional experience in a natural-like conversation. The experimenter always took care that the participants did not talk about experiences not-related to the specific target emotional state and directed participants back to the topic when necessary. The details of the four driving scenarios are presented in the following:

Neutral: To induce a neutral state, the experimenter initiated a conversation on neutral topics, such as the weather, the job or the daily commute to work (2nd and 3rd round). Afterwards, participants had to solely drive the car without conversation and secondary task (4th and 5th round). In the framework of the cover story, this was presented to the participants as training.

Positive: For the induction of the positive state, participants were told that a test of the audio set-up was necessary. For this, a funny radio show ("Wir sind die Freeses", Altenburg, 2017) was presented via the loudspeakers (2nd and 3rd round) and participants had to listen to it as secondary task. Then the experimenter initiated a conversation on positive topics starting from the funny phases of the radio show to stimulate recall of positive experiences. Topics here could for example include humor (e.g. "What kind of show do you find funny?") or other positive experience such as vacations (4th and 5th round).

Frustration: A Wizard-of-Oz (WOZ) navigation system was used to induce frustration. Participants were told that they should evaluate a recently developed speech-based, touch-free navigation system as secondary task. For this, they had to enter a specific address and start the routing. However, indeed the system was controlled by the technician on the rear seat, who purposely misunderstood commands of the participants to frustrate him/her (2nd and 3rd round). Afterwards the experimenter initiated a conversation on similarly frustrating experiences, for example with other technical systems to stimulate the recall of frustration (4th and 5th round).

Anxiety: For the induction of mild anxiety, again a WOZ setup was used. Participants were asked to evaluate the usability of a brake assistant using the speaking aloud technique as secondary task. The brake assistant was introduced as having the capability to detect traffic cones at the side of the street and to automatically brake at these while playing a brief warning sound to the driver. Indeed, the brakes were controlled by the experimenter with the additional brake pedal at the passenger seat, who applied the brakes sometimes at random locations. In addition, the experimenter could play the warning tone without braking (2nd and 3rd round). After this, the experimenter initiated a conversation on similar experiences that elicited anxiety, for instance near crashes or critical incidents during driving to stimulate the recall of anxiety (4th and 5th round).

2.6 Measures

Self-report: To assess participants' subjective emotional experience, we employed three different self-report measures:

- The Geneva Emotion Wheel (GEW, Scherer et al., 2013), which is composed of 20 discrete emotion terms that should be rated on a scale from 1 to 5.
- The Self-Assessment Manikin (SAM, Bradley and Lang, 1994), which assesses emotional experience on the dimensions valence (negative to positive) and arousal (low to high) on a Likert scale from 1 to 5.
- Free text input to describe their current emotional state in their own words.

Before the start of the experiment, self-report was provided using the GEW, the SAM and the free text input. After each driving scenario for emotion induction, the GEW was applied directly. After the entire drive, participants were asked to provide a detailed self-report using the SAM separately for the conversation and the task phases of the emotion inductions. Participants could provide free text input to rate their experience specifically when conducting the secondary tasks (radio show, navigation system, and brake assistant).

Audio recording: The audio speech stream was recorded using two Shure VP 82 shotgun microphones attached to the dashboard above the steering wheel and close to the right A-pillar using elastic mounting to dampen the car's movement. Additionally, to collect high quality reference recordings, a Sennheiser HSP-4 EW-3 headset microphone was worn by participants. The microphone tracks were synchronized using a Steinberg UR44 audio interface. The microphone data from the headset microphone was used for the annotation of the speech data with respect to the emotion.

Physiology: Peripheral physiological data was recorded using the wireless sensor system Heally (SpaceBit, Eberswalde, Germany) to measure electrocardiogram (ECG), finger temperature and skin resistance. ECG was measured with a standard 3-lead set-up with a sampling rate of 500 Hz. Finger temperature and skin resistance were measured using a finger sensor at the index finger of the non-dominant hand at a sampling rate of 20 Hz. For one participant, no physiological signal was recorded for the anxiety scenario due to technical problems with the data acquisition.

2.7 Procedure

Participants were welcomed and informed about the purpose of the study (partly concealed by the cover story). Then they provided written informed consent, read the instructions in a self-paced way and filled a standard demographic questionnaire as well as the abovementioned self-report measures. Afterwards the physiological sensors and the headset microphone were applied and participants boarded the experimental vehicle in the garage. The experimenter and the technician also took a seat and the participant was asked to drive to the initial position of the round course. At this position all driving scenarios started and before each drive, participants were instructed about the upcoming drive (in the framework of the cover story). Then participants drove the respective scenario for five rounds. After each scenario, the participant parked the car at the start position and filled in the GEW and another short questionnaire on fatigue (that was not analyzed, but only served to conceal the focus on the emotions). After the last scenario, participants drove back to the garage, were relieved from the sensors as well as the microphones and answered the final questionnaire that included the detailed SAM and text input. Finally, participants were debriefed about the true aim of the study and received their reimbursement.

3. DATA ANALYSIS

3.1 Self-report

In order to compare the self-report as measure of emotional experience, we employed the following steps:

GEW: From the GEW we selected items to represent the three non-neutral target emotions based on a semantic analysis of the GEW. A composite positive affect subscale was formed by building the average of the items *amusement*, *joy*, *pleasure* and *contentment* as these matched the positive target emotional state in the best way. In addition, the items anger and fear where chosen to best represent frustration and anxiety, respectively. These three scales were then compared between the four scenarios by a series of repeated-measures analyses of variance (ANOVAs) with the factor scenario. In addition, in case of a significant main effect of the scenario, we also report the results of post-hoc comparisons between the conditions (Bonferoni-corrected).

SAM: We extracted the values for valence and arousal of each assessment time point. As there were two values (one for the task and one for the conversation) for the positive, the frustration and the anxiety scenario, we averaged across these to acquire one value per scenario. Finally, repeated-measures ANOVAs were calculated with driving scenarios as factor. In addition, the results of the post-hoc comparisons are reported.

Free text input: The free text input was analyzed in three steps: First, the text was digitalized. However, many participants did not only write about their emotional experience, but left general comments on the situation or task. Therefore, in a second step, the text was reduced to include only the content related to their current experience. This included removal of all none experience-related words and transferring all remaining words to adjectives of experience (e.g. Frustration [German: Frustration] was transferred to frustrated [frustriert] or “it was amusing” [“es war lustig”] was transferred to amused [belustigt]). In this step, words mentioned more than once per scenario and participant were also removed. Finally, in a third step, we counted the words and present the word counts separately for each secondary task.

3.2 Annotation of speech data

In order to evaluate the emotionality of the speech data, all extracted speech samples were annotated by three independent female labelers of the same age group (20 to 35 years). The speech samples were extracted from the raw high quality recordings of the headset microphone. All voiced speech segments of the audio file were used and subdivided into samples of 2 s length, if possible. If divided, the remaining sample length should not be below 0.5 s, to ensure a reliable annotation of the labelers. To avoid this, these short segments were added to the previous sample coming from the same speech segment, such that the sample could reach a maximal length of 2.5 s. The generated speech samples were annotated by the labelers in the following three stages using the ikannotate labelling tool (Böck et al., 2011):

1. Annotation of the dimensions of valence and arousal level on the 5-point SAM scale (Bradley and Lang, 1994, as also used for the self-report).

2. Annotation of the four emotion categories: neutral, positive, frustrated (including angry) and anxious. Additionally, the possibility for free text input to enter a different emotional state was available to the labelers

3. Annotation of the labelers’ satisfaction level of the current labelling on a 5-point Likert scale from 1 (very dissatisfied) to 5 (very satisfied).

To ensure reliability of the annotation results, the interrater reliability (IRR) of the labelers was determined by calculating Krippendorff’s alpha (Siegert et al., 2014). The IRR for objective tasks provides information on how good the raters understood the annotation task. For subjective tasks as emotion annotation, the IRR helps to assess how accurately the problem was identified. By post-hoc tests, the IRR can be used to identify raters having problems in aligning to the problem description. Based on this, a labeler lowering the reliability of the annotation by 0.05 was excluded before assigning a certain label to the considered speech sample. The exclusion of a rater due to a low IRR results in a rather conservative labelling. Only cases where the two raters with a high IRR contradict each other and the rater with the low IRR agrees with one of the other's statement are affected by this decision. Thus, this procedure gives less but more reliable labels. For the dimensional assignment, an average of the results over all included labelers was calculated respectively for valence and arousal, resulting in labels for all annotated speech samples. To assign the emotion categories, a majority voting of all labelers was carried out. In case of labeler's exclusion due to low IRR, the remaining two labelers' hat to be fully conform in their annotation result. All samples with an ambiguous majority voting were not assigned with a label and therefore, not used for the further evaluation. The rating of the labelers' satisfaction level was used to verify the suitability of the annotation.

In order to evaluate the success of the emotion induction, we provide the absolute number of speech samples assigned to a certain emotion dimension and category. In addition, the share of those samples originating from the corresponding emotion scenario in relation to the total number of samples assigned to this emotion category is given. In this paper, we report the results of a subset of the participant sample that was readily annotated by the time of the deadline.

3.3 Physiology

Initially, the heart rate (HR) in beats per minute (bpm) was determined from the ECG signal by counting the number of R waves per minute. Finger temperature was extracted from the raw signal. Skin conductance level (SCL) was calculated by the inverse of the skin resistance. For each participant, we calculated a reference value, which was the mean of the time from one minute after start of driving to the end of the 1st round. In addition, the mean of the induction phase (round 2 to 5) was calculated. In order to account for inter-individual variability in the physiological activity, we subtracted the reference value from the mean for the emotion-induction. Finally, these reference-corrected values in the four scenarios were compared to each other by a series of repeated-measures ANOVAs with the factor scenario. Similarly, as for the self-report data, in case of a significant main effect of the scenario, we also report the results of post-hoc comparisons between the conditions (Bonferoni-corrected).

4. RESULTS

4.1 Self-report

In the following, the results regarding the self-report are presented with the goal to compare participants' emotional experience between the four scenarios.

GEW: There was a significant effect of the scenarios on the composite positive affect subscale (*F*(2.1, 56.7) = 9.5, *p* < .05, Greenhouse-Geisser-corrected). Post-hoc tests revealed that the value for the neutral scenario was higher than for frustration (*p* < .05, Bonferoni-corrected). In addition, the positive scenario was experienced significantly more positive than the frustration and the anxiety scenario (*p* < .05, Bonferoni-corrected). No other comparison was significant. Moreover, the second ANOVA revealed a significant effect of the scenarios on experienced anger (*F*(1.5, 42.2) = 4.7, *p* < .05, Greenhouse-Geisser-corrected). However, although the value was descriptively highest during the frustration scenario, none of the post-hoc test was significant after Bonferoni correction (all *p*s > .05). Finally, the ANOVA for the item fear revealed no significant main effect of the scenarios (*F*(3,84) = 1.2, *p* = .35) despite the descriptive value being highest during the induction of anxiety. The descriptive statistics for the GEW can be found in Table 1.

Table 1. Descriptives (mean [*M*] and standard deviation [*SD*]) of the Geneva Emotion Wheel in the four scenarios for the composite positive affect scale as well as for the items anger and fear.

	Positive affect		**Anger**		**Fear**	
	M	**SD**	**M**	**SD**	**M**	**SD**
Neutral	3.4	0.9	0.0	0.2	0.1	0.6
Positive	3.6	1.0	0.1	0.3	0.1	0.4
Frustration	2.9	1.3	0.6	1.2	0.1	0.4
Mild anxiety	3.1	1.1	0.2	0.8	0.3	0.9

SAM: The descriptive statistics (*M* and *SD*) for the SAM describing the experience in the secondary tasks are presented in Table 2. Valence was highest during the radio show and lowest during the usage of the navigation system. The highest arousal was experienced during the navigation system task, while the lowest was experienced during the radio show. The ANOVA regarding valence revealed a significant effect of the scenario (*F*(2.3, 63.2) = 61.9, *p* < .001, Greenhouse-Geisser-corrected). Post-hoc comparison show that participants experienced higher valence in both, the neutral and the positive scenario, compared to the frustration and the anxiety scenario (*p*s < .05, Bonferoni-corrected). There were no significant differences between neutral and positive as well as between frustration and anxiety. For arousal, a significant effect of the scenarios was revealed, too (*F*(2.2, 57.6) = 6.2, *p* < .01, Greenhouse-Geisser-corrected). Here, post-hoc comparisons showed a significantly higher experienced arousal during the anxiety scenario compared to the neutral and the positive scenario (*p*s < .05). No other difference was significant.

Free text input: At baseline, participants mostly chose positive emotional words describing their excitement and interest in the upcoming experiment (excited, interested, curious, expectant), but also "happy" and "neutral". After the induction of the positive state, generally participants expressed rather positive emotions (amused, entertained, relaxed), however, some also mentioned negative emotions (irritated) and distraction (distracted). After the induction of frustration, mostly negative emotions were used (irritated, frustrated, upset, uncertain). In addition, some participants felt misunderstood, but also amused. After the induction of uncertainty, the words having the highest frequency were related to negative, uncertain emotional states (insecure, puzzled, uncertain, and surprised). Still also the positive words interested and excited were mentioned. For an overview of the original German words mentioned more than twice and their English translations, see Table 3.

Table 2. Mean (*M*) and standard deviations (*SD*) of the valence and arousal rating in the self-assessment manikin (SAM) in the four driving scenarios.

	Valence		Arousal	
	Mean	**SD**	**Mean**	**SD**
Neutral	4.1	0.6	2.1	1.0
Positive	4.4	0.6	2.1	1.1
Frustration	2.9	0.8	2.5	0.9
Anxiety	3.0	0.6	2.6	1.0

Table 3. Results of the free text input regarding the experience at baseline and during the secondary tasks. German words, their count (if > 2) and their English translations are presented. Note that, during the radio show, “amused” is mentioned twice in the English translation, because both German words “belustigt” and “amüsiert” are translated with “amused”.

	Words (count > 2)	
	German (original)	**English (translation)**
Baseline	gespannt (9), interessiert (6), neugierig (4), neutral (4), glücklich (3), erwartungsvoll (3)	excited (9), interested (6), curious (4), neutral (4), happy (4), expectant (3)
Radio show	belustigt (8), entspannt (4), genervt (4), unterhalten (3), abgelenkt (3), amüsiert (3)	amused (8), relaxed (4), irritated (4), entertained (3), distracted (3), amused (3)
Navigation	genervt (11), frustriert (6), verärgert (4), missverstanden (3), belustigt (3), unsicher (3)	irritated (11), frustrated (6), upset (4), misunderstood (3), amused (3), uncertain (3)
Brake assistant	verunsichert (5), verwundert (4), interessiert (4), gespannt (4), überrascht (3), unsicher (3)	insecure (5), puzzled (4), interested (4), excited (4), surprised (3), uncertain (3)

4.2 Annotation of speech data

The audio material of 24 participants (six female), resulting in 5.68 hours of speech material, comprising 13802 speech samples (10267 male, 3535 female), was annotated and evaluated, which leads to an average number of 570 samples for male participants and 589 samples for female participants. Considering the emotion scenarios separately, this results in 3259 samples originating from the neutral scenario, 2897 from the positive scenario, 3375 from the frustration scenario and 4271 from the anxiety scenario. A repeated-measures ANOVA revealed that the number of samples recorded from the anxiety scenario is significantly higher than the number of samples recorded from other emotion scenarios (main effect scenario: $F(3,92) = 12.7$, $p < .001$, post-hoc: anxiety vs. other emotions, all ps < .005).

Annotation time: The annotation of all 13802 speech samples took on average 36 hours for each labeler. In this time, the labelers annotated the dimensions of valence and arousal, the four emotion categories neutral, positive, frustration/anger, anxiety/fearful, and rated the satisfaction level of their annotation.

Interrater reliability: Based on the IRR, a labeler who annotated contrarily to the other labelers was excluded for further evaluation. Table 4 shows the average IRR over all evaluated participants for all possible combinations of the three labelers. For the dimensional approach, a compromise between a good annotation of valence and arousal needed to be made. Because of the low IRR of the arousal level compared to the very high IRR of the valence level when leaving out labeler 2, the compromise of leaving out labeler 1, which results in a satisfactory IRR for both valence and arousal, was made. This decision was made individually for every considered participant, resulting in a leaving out of labeler 1 in 14 out of 24 cases. The annotation of labeler 2 was left unconsidered in the remaining ten cases. Leaving out labeler 3 never resulted in an increase of interrater reliability. By considering these cases we achieved an average IRR of .44 for valence and .34 for arousal.

Table 4. Average IRR of all possible combinations of labelers for dimensional and categorial annotation.

	Valence	Arousal	Categorial
All	.37	.17	.25
w/o Labeler 1	.37	.25	.21
w/o Labeler 2	.48	.18	.32
w/o Labeler 3	.22	-.04	.19

For the categorial annotation the best IRR was achieved when leaving out labeler 2. This was the case for the annotation of 16 out of 24 participants. For seven cases the annotation results of all labelers were used, because the difference between the IRR of the best two and all labelers, was considerably small. Labeler 1 was left unconsidered only in one case. Considering these cases an average IRR of 0.32 was achieved for the categorial annotation.

Table 5. Confusion matrix of the categorially and dimensionally labeled audio samples (in percent).

Dimensional/ categorial	n	q1	q2	q3	q4	low	positive	negative
Neutral	44.86%	0	48.75%	16.27%	0	84.92%	0	0
Positive	20.53%	99.60%	48.39%	0	0	.46%	100%	0
Frustrated	22.04%	.40	1.08%	17.38%	97.73%	3.48%	0	93.22%
Anxious	12.57%	0	1.79%	66.36%	2.27%	11.14%	0	6.78%

Label assignment: The labels of the *dimensional* approach were assigned to the desired audio samples by averaging the annotation results of the considered labelers. The averaged values of the valence/arousal level were then mapped onto the four quadrants (q1 to q4, see Figure 1) and the origin of the valence/arousal space (n [=neutral], see Figure 1). Samples lying directly on either the valence or the arousal axis were labeled as “high”, “low”, “positive”, and “negative”, respectively. This resulted in nine mappings: n (count = 11377), q1 (268), q2 (439), q3 (841), q4 (95), high (1), low (579), positive (131) and negative (70). The large amount of samples mapped onto the neutral region of the valence/arousal level is striking, but reasonable for this kind of highly natural and low expressive recorded audio data.

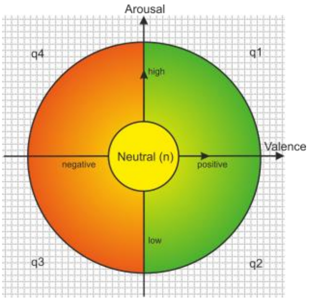

Figure 1. Mapping of the valence/arousal values on the four quadrants of the valence arousal space.

The majority voting of the remaining labelers regarding emotion category resulted in a total number of 8640 categorially labelled audio samples, which corresponds to 63% of the original samples. Considering all emotion categories separately, this resulted in 3676 neutral, 1910 positive, 1771 frustrated, and 1283 anxious samples. The high number of neutral samples is explicable by the experimental setup, as a neutral emotional state will naturally occur in all designed scenarios, without a need of being induced.

To confirm that both annotation approaches correspond to each other, the confusion matrix of both approaches was determined. The results are shown in Table 5. Because of the low number of “high” dimensionally annotated samples, this label was left unconsidered. In Table 5, green entries denote a correlated assignment between both annotation approaches, while red entries denote an uncorrelated assignment. A high correlation of the annotation approaches is indicated by high values in green and low values in red entries. For the stated confusion matrix, a high consistency between the annotation results can be concluded. The high number of labels lying in the neutral region of the valence/arousal annotation, but being assigned to a different emotion category is explicable by the low expressiveness of the data. Already slight changes in valence and arousal indicate a change of the emotional state. Therefore, the assumption can be drawn, that the true neutral region lies closer around the origin of the valence/arousal dimensions than assumed. The high number of labels assigned to the neutral emotion category but mapped to the region of low arousal and neutral valence indicates an elongation of the neutral region in the valence/arousal-diagram towards the low arousal axis, which is also assumed in the emotion models presented by Holzapfel et al., (2002) and Almeida et al. (2016).

Table 6. Confusion matrix of samples coming from the emotion scenarios and samples coming from the categorial annotation.

Scenario/ Emotion	Neutral	Positive	Frustration	Anxiety
Neutral	1256	806	723	891
Positive	510	826	290	284
Frustrated	159	141	946	525
Anxious	151	79	190	863

Evaluation of experimental setup: As the experimental setup was designed such that the induced emotions are similar to the considered emotion categories used for annotation, a clear statement on the performance of the experimental setup can be given by determining the share of samples originating from the scenario and the labelling. A confusion matrix of the results is stated in Table 6. It can be seen, that the largest entries lie on the main diagonal of the matrix. This indicates that for each driving scenario, except anxiety, a majority of the samples were also labeled as the corresponding emotion. The large number of samples labelled as neutral (first row) is reasonable, as neutral speech was uttered in all the designed scenarios. The same holds for the number of samples labeled as positive (second row) as most of the participants were very positive while conversing with the interviewer. Also the low number of samples labelled as frustrated and anxious in the neutral and positive scenario is reasonable as the participants also talked about frustrating situations they experienced beforehand. As the mild anxiety scenario was conducted after the frustration scenario and they were both based on the evaluation of a technical system which did not work properly, some of the participants also experienced strong frustration during the anxiety task. This explains the high number of samples labelled as frustrated in the mild anxiety scenario.

4.3 Physiology

Table 7. Descriptive statistics (Mean [*M*] und standard deviation [*SD*]) of heart rate (HR), finger temperature (FT) and skin conductance level (SC) in the four scenarios.

	HR		FT		SC	
	M	**SD**	**M**	**SD**	**M**	**SD**
Neutral	.70	2.72	-.12	.79	5.33	8.08
Positive	2.85	3.12	.23	.73	2.87	7.43
Frustration	1.59	3.74	-.22	.75	3.17	17.9
Anxiety	2.24	3.44	-.30	.73	.53	12.4

Scale: HR in beats per minute, FT in °C, SC in 10^{-4} micro Siemens.

Participants' heart rate descriptively increased compared to the reference value in all scenarios with the highest increase in the positive followed by the anxiety scenario. The ANOVA revealed a significant effect of the scenario ($F(2,84) = 3.5$, $p < .05$). Post-hoc comparisons showed that the heart rate was higher in the positive than in the neutral scenario ($p < .05$). All other comparisons were not significant (all *p*s > .05). Finger temperature was only higher than the reference value in the positive scenario. The ANOVA revealed a significant effect of the scenario ($F(3,84) = 5.5$, $p < .01$). Post-hoc comparisons revealed that the finger temperature was significantly higher in the positive compared to the anxiety scenario ($p < .05$, all other *p*s > .05). Skin conductance was descriptively lowest in the anxiety scenarios and highest during the neutral scenario. However, the ANOVA did not reveal a significant effect of the scenario on the skin conductance level ($F(1.7,49.6) = 1.1$, $p = .326$, Greenhouse-Geisser-corrected). For an overview on the descriptive statistics of the physiological values, see Table 7.

5. DISCUSSION

The goal of this work was to evaluate an experimental set-up combining secondary tasks and conversation-based emotional recall to induce emotions in highly realistic, real-world driving scenarios. We used self-report, annotation of speech data and peripheral physiology as measures to determine whether or not the experimental manipulations were successful. In the following, we will discuss the results for each of the emotion induction scenarios, consider the limitations of the study and finally provide recommendations for future real-world driving studies with the aim to induce emotions.

5.1 Evaluation of the scenarios

Neutral scenario: In the neutral scenario, participants experienced subjectively lower arousal and higher valence than in the two negative scenarios (frustration and anxiety); however, no differences were revealed as compared to the positive scenario. Interestingly, also the composite positive affect from the GEW did not differ between the neutral and positive scenario. With respect to the annotation of speech data, the largest share of the neutral scenario was also labeled as neutral. Though, it has to be mentioned that some samples of the neutral scenario were also labeled as positive, frustrated or anxious. This is likely due to the fact that although the experimenter attempted to keep the conversation as neutral as possible, some mentioned topics such as the job or the weather triggered also positive, frustrated and anxious experiences in the participants. Regarding physiology, it was shown that the heart rate was significantly lower than in the positive scenario. This supposes the low arousal in the neutral scenario indicating a successful induction of neutral state. In total, we can conclude that the neutral scenario seemed to actually induce neutral experiences, although self-report indicates that valence and arousal did not significantly differ from the positive scenario.

Positive scenario: The positive scenario was experienced as more positive than the two negative scenarios (frustration and anxiety) as indicated by the GEW composite positive subscale and the SAM valence scale. Arousal was experienced slightly lower than in the anxiety scenario. The free text input showed that participants frequently reported being amused, relaxed and entertained by the scenario, despite some mentioning being annoyed or distracted. These negative terms may be due to the fact that a radio show targeting a specific kind of humor was chosen to elicit positive emotions, which may strongly differ between participants. The annotation of the speech samples point into a similar direction indicating that mostly a positive emotional state was induced, but additionally rather neutral experiences as well as frustration and anxiety were triggered. Interestingly, heart rate was higher in the positive compared to the neutral scenario, and finger temperature was higher in the positive in comparison to the anxiety scenario. This is in line with a review of Kreibig (2010), who states that happiness comes along with increased heart rate and increased finger temperature. In addition, recent work suggests that skin temperature can also be seen as a measure of control over the situation, in the sense that higher control over the situation is associated with higher finger temperature (see Fontaine et al., 2016; Zhang et al., 2018), which characterizes one aspect of the difference between the positive and the anxious scenario. Altogether, it seems that this scenario was suitable to elicit a positive emotional state most of the time for most of the participants.

Frustration scenario: The frustration scenario was experienced as less positive than the positive scenario according to the GEW and the SAM, but did not differ from any other scenario regarding arousal. This makes sense as frustration is mostly seen as having rather negative valence, but only very moderate arousal (Russell, 1980; Ihme, Unni, Zhang, Rieger, Jipp, 2018; Ihme, Dömeland, Freese, Jipp, 2018). Interestingly, this is also backed up by the fact that the physiological signals did not show any significant effect in relation to the frustrated scenario. It has to be mentioned that the item anger of the GEW did not show a significant effect for frustration, which indicates that participants did not experience so much anger here, but rather milder negative feelings. The free text input provides some deeper insights: participants mostly mention words being very close to frustration, such as irritated, frustrated, upset and misunderstood, but also amused and uncertain. "Angry" was not mentioned. The amusement may be seen as a sign of a "grim sense of humor", because the navigation system just did not understand them. Grippingly, earlier studies on frustration in human-computer interaction have shown that participants even smile when experiencing frustration (Hoque et al., 2012). The annotation of the audio data also argues in favor of a successful induction of frustration, because the largest share of the frustrating scenario was labeled as frustrated. In total, it seems that the induction of frustration has worked very well in this study.

Anxiety scenario: The results regarding the anxiety scenario are a bit more complicated to interpret. Participants rated the scenario as having lower valence and positive affect as well as higher arousal than the neutral and the positive scenario (without differing from frustration), which is in line with the classification of anxiety in the valence and arousal space (e.g. Fontaine et al., 2016). This higher arousal may be related to the anticipation of negative events (e.g. unexpected brake reactions). Interestingly, as mentioned above, participants also show lower finger temperature (compared to the positive scenario), which has been associated with the low control over a situation in relation to fear or anxiety (cf. Fontaine et al., 2016; Zhang et al., 2018). This low control may have well been experienced by the participants when interacting with the unforeseeable brake assistant. In addition, a large share of the samples of the anxiety scenario has been labeled as anxious (~ 34 %). Still, other emotions have been triggered in this scenario as well. A very interesting insight can be drawn from the free text input here, which revealed that participants rather felt insecure, puzzled, surprised or uncertain instead of anxious (which was only mentioned once). This indicates that we did not accomplish to induce strong anxiety, but a "milder" state which is rather uncertainty or insecurity. To sum up, there are indicators that induced anxiety (from the speech annotation), while other indicators (free text input) rather suggest that this state was a bit milder (uncertainty).

5.2 Limitations

A few limitations have to be mentioned regarding this set-up. First, it seems as if the GEW was not the best measure to acquire self-reports as the items of the GEW do not perfectly match with the target emotional states. In the future, it may be worthwhile to use self-report questionnaires including emotional words for the target emotional states. As a second limitation, we did not randomize the scenarios for the sake of a trustworthy cover story. This could mean that effects of motivation or fatigue may add different variance to later scenarios (anxiety) than the earlier ones. The rather positive experience of participants during the neutral drive may in part be explained by the fact that participants here were still alert and motivated. Future studies using the paradigm should consider adjusting the cover story to acquire the possibility to randomize the order of the scenarios. In addition, we did not ask participants for free text input regarding the neutral scenario, which should be considered in future studies to further improve the comparability of the scenarios.

5.3 Recommendations for the induction of emotions

The emotion induction methods presented in this study appeared to have worked very well. The extensive evaluation of the scenarios allows us to give some recommendations for future studies that want to induce emotions in real vehicles: First of all, it seems that a cover story provides a way to conceal the true nature of the study and get people motivated to take part and engage in the tasks. Engagement seems to be a prerequisite for successfully inducing emotions. Second, participants were relatively positive in the neutral scenario as well, likely because they were motivated and the situation was novel. It may be worthwhile to add a very boring secondary task to the driving in order to even produce less positive, rather neutral experiences in participants. Third, the induction of positive experiences through humor worked fairly well. Still, some participants found the radio show annoying, which is likely due to the fact that humor differs between people. To reliably induce amusement it could help to have a selection of funny shows available and let the participants choose their preferred one. Fourth, the introduction of frustration appeared to work relatively well, if participants have a goal which is blocked from time to time. Thus malfunctioning technical systems are a good choice to induce frustration. Fifth, the induction of strong anxiety appears really hard without actually threatening the participants. The method here worked well to induce mild anxiety or uncertainty, which are relevant emotions in human-machine interaction. Still, for inducing anxiety maybe controlled critical incidents (near crashes) could be used, if possible within ethical and safety regulations. Finally, it has to be mentioned that it is almost not possible to induce one (and only one) emotional state constantly over several minutes. For instance, the use of secondary tasks, as used for the induction of frustration in our study, may, in addition to emotions, induce mental workload that is also accompanied by arousal effects. Therefore, even with good paradigms to induce emotions, a post-hoc annotation of the data based on the information from speech and physiology is recommended to extract the exact phases in which an emotion was experienced and thus to generate a ground truth.

6. CONCLUSION

Here, we presented an experimental paradigm that enables researchers and engineers to induce different emotional states in a real-world driving scenario. Experimental paradigms like this are an important tool to generate data needed for the development of methods to detect emotional states of drivers. Thus, the presented work is a crucial brick when finally building the future empathic vehicle.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research in innovation programme under grant agreement No. 68890 (ADAS&ME).

REFERENCES

Almeida, P. R., Ferreira-Santos, F., Chaves, P. L., Paiva, T. O., Barbosa, F., and Marques-Teixeira, J. (2016). Perceived arousal of facial expressions of emotion modulates the N170, regardless of emotional category: Time domain and time–frequency dynamics. *International Journal of Psychophysiology* 99, 48–56. doi: 10.1016/j.ijpsycho.2015.11.017

Altenburg, A. (2017). Wir sind die Freeses, www.ndr.de/ndr2/wir_sind_die_freeses/podcast4250.html.

Biundo, S. and Wendemuth, A. (2017). Companion Technology – A Paradigm Shift in Human-Technology Interaction. Springer International Publishing. doi: 10.1007/978-3-319-43665-4

Böck, R., Siegert, I., Haase, M., Lange, J., and Wendemuth, A. (2011). ikannotate - A Tool for Labelling, Transcription, and Annotation of Emotionally Coloured Speech. In Sidney D'Mello, et al., editors, Proc. of the ACII-2011, volume 6974 of LNCS, pages 25–34, Memphis, TN, USA. Springer Verlag Berlin, Germany.

Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 49–59.

Drewitz, U., Kaul, R., Jipp, M., and Ihme, K. (2017). "Verstehende und mitdenkende Assistenz für hochautomatisierte Fahrzeuge," VDI Tagung: Der Fahrer im 21. Jahrhundert, 21 Nov 2017, Braunschweig, Germany.

Fischer, M., Richter, A., Schindler, J., Plättner, J., Temme, G., Kelsch, J., et al. (2014). "Modular and Scalable Driving Simulator Hardware and Software for the Development of Future Driver Assistence and Automation Systems," in *Proceedings of the Driving Simulation Conference 2014*, 223–229, Paris, France.

Flemisch, F., Meier, S., Neuhöfer, J., Baltzer, M., Altendorf, E., and Özyurt, E. (2013). Kognitive und kooperative Systeme in der Fahrzeugführung: Selektiver Rückblick über die letzten Dekaden und Spekulation über die Zukunft. *Kognitive Systeme*. doi: 10.17185/duepublico/31356

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2016). The World of Emotions is not Two-Dimensional. *Psychol Sci* 18, 1050–1057. doi: 10.1111/j.1467-9280.2007.02024.x

Freese, M., and Jipp, M. (2015). Zwischen Rational und Emotional – Ein Überblick über Entscheidungen und deren Einflussgrößen in kooperierenden Teams. *Kognitive Systeme, 2, doi:* 10.17185/duepublico/40722.

Gadsden, S., and Habibi, S. (2009). "The Future of Automobiles: Cognitive Cars," in *22nd Canadian Congress of Applied Mechanics (CANCAM)*, 111-11, Halifax, Nova Scotia, Canada.

Heide, A., and Henning, K. (2006). The "Cognitive Car": A Roadmap for Research in the Automotive Sector. *IFAC Proceedings Volumes* 39, 44–50. doi: 10.3182/20060522-3-FR-2904.00008

Hernandez, J., McDuff, D., Benavides, X., Amores, J., Maes, P., and Picard, R. (2014). "AutoEmotive: bringing empathy to the driving experience to manage stress," in *Proceedings of the 2014 companion publication on Designing interactive systems*, 53–56, Vancouver, BC, Canada.

Holzapfel, H., Fuegen, C., Denecke, M., and Waibel, A. (2002). "Integrating emotional cues into a framework for dialogue management," in *Fourth IEEE International Conference on Multimodal Interfaces*, 141–146, Pittsburgh, PA, USA .

Hoque, M. E., McDuff, D. J., and Picard, R. W. (2012). Exploring Temporal Patterns in Classifying Frustrated and Delighted Smiles. *IEEE Trans. Affective Comput.* 3, 323–334. doi: 10.1109/T-AFFC.2012.11

Ihme, K., Dömeland, C., Freese, M., and Jipp, M. (2018). Frustration in the Face of the Driver: A Simulator Study on Facial Muscle Activity during Frustrated Driving. *Interaction Studies*, 19:3, 487-498.

Ihme, K., Unni, A., Zhang, M., Rieger, J.W., Jipp. M. (2018). Recognizing Frustration of Drivers from Video Recordings of the Face and Measurements of Functional Near Infrared Spectroscopy Brain Activation. *Frontiers in Human Neuroscience*, 12:327. doi: 10.3389/fnhum.2018.00327Jeon, M. (2015). Towards affect-integrated driving behaviour research. *Theoretical Issues in Ergonomics Science* 16, 553–585. doi: 10.1080/1463922X.2015.1067934

Jeon, M., Walker, B. N., and Yim, J.-B. (2014). Effects of specific emotions on subjective judgment, driving performance, and perceived workload. *Transportation Research Part F: Traffic Psychology and Behaviour* 24, 197–209. doi: 10.1016/j.trf.2014.04.003

Klein, J., Moon, Y., Picard, R.W. (2002). This computer responds to user frustration. *Interacting with Computers* 14:2, 119–140. dDoi: 10.1016/S0953-5438(01)00053-4

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological psychology* 84, 394–421. doi: 10.1016/j.biopsycho.2010.03.010

Löcken, A., Ihme, K., Unni, A. (2017): Towards Designing Affect-Aware Systems for Mitigating the Effects of In-

Vehicle Frustration. in *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications Adjunct*, Oldenburg, Germany, 24 - 27 Sep. New York, NY, USA: ACM, 88–93. doi: 10.1145/3131726.3131744

Lotz, A., Ihme, K., Charnoz, A., Maroudis, P., Dmitriev, I., and Wendemuth, A. (2018). "Recognizing Behavioral Factors while Driving: A Real-World Multimodal Corpus to Monitor the Driver's Affective State," in *LREC 2018: Eleventh International Conference on Language Resources and Evaluation : May 7-12, 2018, Miyazaki, Japan*, ed N. Calzolari (Paris, France: European Language Resources Association (ELRA)).

Lu, J., Xie, X., and Zhang, R. (2013). Focusing on appraisals: How and why anger and fear influence driving risk perception. *Journal of Safety Research* 45, 65–73. doi: 10.1016/j.jsr.2013.01.009

Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L. (2005): Improving automotive safety by pairing driver emotion and car voice emotion. in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. April 2-7, 2005, Portland, OR, USA. (ACM New York, NY, USA) doi: 10.1145/1056808.1057070

Pêcher, C., Lamercier, C., and Cellier, J.-M. (2011). "The influence of emotions on driving behavior," in *Traffic psychology: An international perspective*, ed D. Hennessy (New York, NY: Nova Science Publ), 145–158.

Pêcher, C., Lemercier, C., and Cellier, J.-M. (2009). Emotions drive attention: Effects on driver's behaviour. *Safety Science* 47, 1254–1259. doi: 10.1016/j.ssci.2009.03.011

Picard, R. W., and Klein, J. (2002). Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with Computers* 14, 141–169. doi: 10.1016/S0953-5438(01)00055-8

Plitnick, B., Figueiro, M., Wood, B., Rea, M. (2010): The effects of red and blue light on alertness and mood at night. *Lighting Research and Technology* 42:4, 449–458. doi: 10.1177/1477153509360887Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 1161–1178. doi: 10.1037/h0077714

Scherer, K. R., Shuman, V., Fontaine, J. R. J., and Soriano, C. (2013). "The GRID meets the Wheel: Assessing emotional feeling via self-report," in *Components of Emotional Meaning*, eds J. R. J. Fontaine, K. R. Scherer, and C. Soriano (Oxford University Press), 281–298.

Shinar, D. (1998). Aggressive driving: the contribution of the drivers and the situation. *Transportation Research Part F: Traffic Psychology and Behaviour* 1, 137–160. doi: 10.1016/S1369-8478(99)00002-9

Siegert, I., Böck, R., and Wendemuth, A. (2014). Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements. *J Multimodal User Interfaces* 8, 17–28. doi: 10.1007/s12193-013-0129-9

Zhang, M., Ihme, K., and Drewitz, U. (2018). "Discriminating Drivers' Fear and Frustration through the Dimension of Power," in *Humanist Conference 2018, The Hague, The Netherlands*.

An Experimental Paradigm for Inducing Emotions in a Real World Driving Scenario

Requardt, Alicia Flores; Wilbrink, Marc; Siegert, Ingo; Jipp, Meike; Wendemuth, Andreas; Ihme, Klas

In: Kognitive Systeme / 2018 - 1

DOI: https://doi.org/10.17185/duepublico/48595

URN: urn:nbn:de:hbz:464-20190417-115016-0

Link: https://duepublico.uni-duisburg-essen.de:443/servlets/DocumentServlet?id=48595