On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving

Michael Grimm, Kristian Kroschel – *Universität Karlsruhe (TH), Germany*
Helen Harris, Clifford Nass - *Stanford University, USA*
Björn Schuller, Gerhard Rigoll – *TU München, Germany*
Tobias Moosmayr – *BMW Group, München, Germany*

On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving

Michael Grimm[1⋆], Kristian Kroschel[1], Helen Harris[2], Clifford Nass[2], Björn Schuller[3], Gerhard Rigoll[3], and Tobias Moosmayr[4]

[1] Universität Karlsruhe (TH), Institut für Nachrichtentechnik, 76128 Karlsruhe, Germany
[2] Stanford University, Department of Communication, Stanford, CA 94305-2050, USA
[3] Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany
[4] BMW Group, Forschungs- und Innovationszentrum, Akustik, Komfort und Werterhaltung, 80788 München, Germany

Abstract. This paper brings together two important aspects of the human-machine interaction in cars: the psychological aspect and the engineering aspect. The psychologically motivated part of this study addresses questions such as *why* it is important to automatically assess the driver's affective state, which states are important and how a machine's response should look like. The engineering part studies *how* the emotional state of a driver can be estimated by extracting acoustic features from the speech signal and mapping them to an emotion state in a multidimensional, continuous-valued emotion space. Such a feasibility study is performed in an experiment in which spontaneous, authentic emotional utterances are superimposed by car noise of several car types and various road surfaces.

1 Introduction

In recent years there has been a growing number of speech-driven applications in the car [1]. Therefore, current research on improvements of both comfort and safety in the car needs to pay attention to the speech interface between the driver and the infotainment system of the car. This paper focuses on one major aspect of the human factors: the driver's emotion.

In a previous study it was found that matching the emotional state of the driver and the expressiveness of a synthetic voice has a major impact on the driving performance [2, 3]. Thus, it is necessary to automatically recognize the emotional state of a person while driving. This led us to the following experiment: Spontaneous emotional utterances were superimposed with car noise of several scenarios. The emotion conveyed in these utterances was automatically estimated using a set of 20 selected acoustic features. For the representation of the emotion, a three-dimensional emotion space concept was applied. Thus, an emotion is described in terms of three continuous-valued primitives (attributes), namely

⋆ Contact: grimm@int.uni-karlsruhe.de

valence (positive vs. negative), activation (calm vs. excited), and dominance weak vs. strong).

There are only a few other works on emotion recognition in the car. Jones and Jonsson [4] presented a method to detect five emotional states of drivers in a driving simulator. They use neural network classifiers but did not investigate the impact of the car noise. Schuller *et al.* [5] also based their experiment on driving simulator data, recognizing four different emotions using Support Vector Machines. However, preliminary studies on emotional speech superimposed by white noise showed that the recognition performance depends very much on the signal-to-noise ratio [6]. Thus, we study the impact of car noise on the emotion recognition in this paper.

The rest of the paper is organized as follows. Section 2 discusses the impact of the driver's emotion on the communication with the car and its importance regarding the safety. Section 3 presents the data used for the automatic emotion recognition experiment. Section 4 introduces the car noise conditions. Section 5 details the classification of emotions from the speech signal. Section 6 presents the recognition results. Section 7 summarizes the study and outlines future work.

2 The Role of Emotion in Driving Experiments

No human thought or action takes place in a vacuum. Temperaments, moods, and emotions shape how people view the world and how they react to it. Although temperaments and personality traits display more stability over time, and predict nuances of behavior, moods and emotions are easier to detect and classify in a real-time scenario. Furthermore, emotional states are more readily mapped to behavioral consequences.

2.1 Emotional States and Driving Behavior

In the context of driving, three distinct groups of emotional states have emerged as states of interest. The first state is defined by a slightly positive valence and a moderate level of arousal, closely associated with the emotional state of *happy*. The optimal state, thought of as a *flow* state [7], involves a moderate level of arousal, allowing for attention, focus, and productivity. A state of high arousal or extreme positive valence can potentially lead to distraction. States of a positive affect have also been shown to improve performance in non-driving contexts [8–10].

The second state of interest is characterized by an extreme level negative valence and high arousal, usually classified as anger. Frustration is distinguished from anger by the degree of negativity and arousal. Often, frustration is referred to as a gateway emotion that leads to anger, and ultimately to aggression and road rage [11]. With an increasing number of vehicles on the roadways, drivers encounter more frustration-inducing scenarios. As a result, road rage is now an escalating problem, and is the primary cause of many accidents and driving fatalities.

The third state is characterized by very low arousal and sometimes accompanied by a slight negative valance; when broadly defined, this state encompasses

both sadness and drowsiness. A sad or negative state degrades task performance, and within the car this state is manifested as inattention.

Given these implications of driving under the influence of particular emotional states, it is unquestionably important to be able to identify such states at drivers. However, once the state of the driver is known, what is the best strategy to improve driver emotion and optimize driving behavior?

2.2 Appropriate Responses to Driver Emotion

Previous research has experimentally tested social responses to driver emotion. Nass and colleagues used a 2 (inducement emotion) x 2 (voice interface emotion) between-subject factorial design [3]. Without the benefit of a naturalistic setting and real-time assessment of driver emotion, researchers relied on the method of using emotionally-charged clips to induce emotion [12, 13]. Two five minute videos, one inducing the state of *happy*, and the other for *sad/subdued* were created from a pre-tested image database. Half of the participants in the study were induced to be *happy*, and the other half were *sad/subdued*; the effectiveness of the inducement method was verified by self-report data from the Differential Emotional Scale (DES) [14].

In keeping with the factorial design, half of the participants from each inducement group drove through a simulated driving course while engaged in conversation with a *happy* voice interface; the other half interacted with the *sad/subdued* voice. The voice interface was actually a series of brief questions and comments, played at exactly the same time in the simulator for every participant. The same female voice talent recorded both the *happy* and *sad/subdued* versions of the script; there was no difference in content between the two versions, only a distinction in intonation and expression.

Contrary to common beliefs that *happy* is always best, results from the study showed that matching the voice emotion to the driver emotion proved more beneficial to emotion than simply presenting all drivers with a *happy* voice. This surprising result reveals the importance of designing socially appropriate in-car voice interfaces. Drivers who were induced to be *sad/subdued* expected their conversation partner to be aware of their state and respond accordingly, in a more subdued manner. Thus, the ability to detect driver emotion is not only helpful in predicting driver behavior, but necessary for designing smart, adaptive, and beneficial driver interfaces.

In order to implement socially appropriate interfaces [15, 16], several hurdles must be overcome before the technology and knowledge can be integrated in vehicles. Future research must continue to explore the implications of addressing the driver with an in-car voice interface. Some first steps have been made, but the studies that have been completed only begin to touch upon the range of human emotional states, not to mention the effects of individual differences. However, suitable interactions cannot occur without first an adequate classification of emotions along the valence, activation, and dominance dimensions, as well as a robust capability to detect such emotions. The field of psychology, among others, provides a significant body of work to aid in distinguishing emotions. The work presented here further contributes to the future of intelligent in-car

interface design by demonstrating the feasibility of detecting emotions in variety of contexts and under challenging noise conditions.

3 Data and Evaluation

For this study we used the VAM corpus, a database consisting of 947 spontaneous emotional utterances in German, which was first used in [17]. These utterances were recorded from 47 speakers in a talk-show on TV. The emotions arose from spontaneous, unscripted discussions, mainly on family issues or friendship questions.

The mean utterance duration was 3.0 s. The mean Signal-to-Noise Ratio was 19.2 ± 3.0 dB, reflecting the relatively good recording conditions of the close talk microphones. All signals were sampled at 16 kHz and 16 bit resolution.

The emotional content was manually labeled by a group of 18 human evaluators. An appropriate value for each emotion primitive was assigned to each utterance by means of the Self Assessment Manikins [18–20]. One out of five icons per primitive could be selected. The choice was then mapped to a scale of [-1,+1].

The average standard deviation in the human evaluator's ratings was 0.31, and the average inter-evaluator correlation was 0.6. This shows moderate to high inter-evaluator agreement for a rather difficult task of labeling spontaneous, non-acted emotions.

4 Noise Scenario

Robust automatic speech recognition (ASR) under the influence of car noise is still being researched [21]. Also, emotion recognition in the car is much more demanding than in clean speech. To study the feasibility, several noise scenarios of approx. 30 seconds were recorded in the car while driving. The microphone was mounted in the middle of the instrument panel, which is the standard for ASR applications in the car. The recorded noise was a superposition of several influences: noise from the wheels/suspension, the combustion engine, interior squeak and rattle noise, and wind noise. The influence of the signal path between the speaker and the microphone was neglected.

4.1 Choice of Vehicles

For this study we used four different cars as itemized in Table 1. Although the soft top of both convertibles was closed during the recordings, the interior noise was noticeable higher than in comparable sedans. While the engine noise dominates during the acceleration of the sportive M5, the similarly constructed 5 series Touring is more gentle and comfortable. The supermini unifies convertible, hard suspension and sportive engine, and it thus provides the most demanding noise scenario.

Table 1. Choice of Vehicles.

Notation	Vehicle	Derivative	Class
530i	BMW 5 series	Touring	Executive Car
645Ci	BMW 6 series	Convertible	Executive Car
M5	BMW M5	Sedan	Executive Sports Car
Mini	MINI Cooper S	Convertible	Supermini

4.2 Road Surfaces

Just as the vehicle type, the road surface affects the interior noise. We recorded the interior noise in all cars on the following surfaces:

- Smooth city road, 50 km/h (CTY)
- Highway, 120 km/h (HWY)
- Big cobbles, 30 km/h (COB)

The lowest noise levels were found with a constant driving over a smooth city road at 50 km/h and medium relovution. Higher noise levels were measured at a highway drive due to the increased wind noise. The worst scenario was found in the recordings on a road with big cobbles. The wind noise resided in lower levels but the rough surface involved dominant wheel/suspension noise as well as buzzes, squeaks and rattles.

4.3 Signal-to-Noise Ratio

The car noise signal of the different scenarios was chopped to fit the length of each utterance and then overlaid additively. To determine the noise conditions quantitatively, the Signal-to-Noise Ratio (SNR) was calculated for each utterance in the speech database and each noise scenario,

$$SNR = 10 \log_{10} \frac{P_{\text{sig}}}{P_{\text{N,car}} + P_{\text{N,mic}}}. \tag{1}$$

In addition to the car noise $P_{\text{N ,car}}$, the recording noise $P_{\text{N,mic}}$ in the speech signal was taken into account and compared to the signal power P_{sig}. The signal power was measured in voiced segments only, whereas the recording noise power was measured in speech pauses only. Due to the varying signal power in the speech recordings, and due to the varying durations of the noise segments, the result was a Gaussian-like distribution, which is shown in Figure 1.
It can be summarized that the road surface has a major impact on the scenario. The SNR for the CTY scenarios was best with a mean value of 11 dB. It was followed by the HWY scenarios (4 dB) and the COB scenarios (-5 dB).

The vehicle has a minor influence. Still, the M5 and the 645Ci resulted in 2 dB better result than the 530i or the Mini for the CTY scenarios. On the highway (HWY), the M5 outperformed the 645Ci by 2 dB and the 530i or Mini by 3 and 4 dB, respectively. Interestingly, on the cobbled road (COB), the 645Ci outperformed the M5 by 2 dB and the 530i or Mini by 4 and 5 dB, respectively.

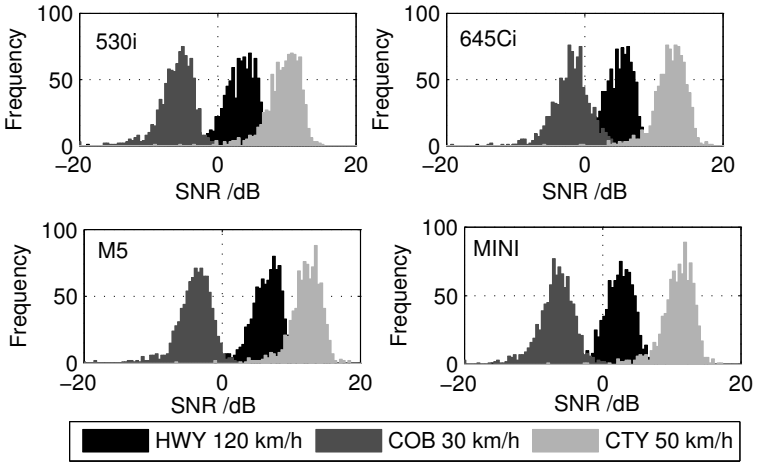

Fig. 1. Signal-to-Noise Ratio distribution in the experiment.

5 Classification

For the automatic estimation of the emotion conveyed in the speech signal, a two-step procedure was applied. First, a set of characteristic prosodic features was extracted from the speech signal. Second, these features were mapped to the values of the three emotion primitives using multidimensional regression techniques and a training set to determine the parameters of the regression curve. The details of the emotion estimation system are described in [22].

5.1 Acoustic features

We used a set of 20 prosodic features selected by Sequential Forward Selection from a total of 137 features. These features include

- pitch related features
- energy related features
- features related to duration and timing
- spectral features using Mel Frequency Cepstral Coefficients (MFCCs)

The pitch and energy related features are the statistics, such as mean value, standard deviation, range, and quartiles of the fundamental frequency F0 and the energy as well as their first and second derivatives, respectively. Together with the temporal features, which are, e.g., speaking rate and mean speaking pause duration, such characteristics describe the intonation of the utterance. The spectral characteristics describe the sentence-dependent voice characteristics in several subbands selected to match the perception characteristics of the human ear. Similar feature sets were used in a number of other studies on emotion recognition in speech [23]. All features were normalized to the range of [0,1].

5.2 Estimation technique

As a classifier, we used kernel-based Support Vector Regression (SVR) [22]. Such method was shown to give superior results compared to other classifiers [5, 22]. The algorithm minimizes the structural risk, in contrast to many other classification techniques minimizing the empirical risk only [24, 25]. A radial basis function ($\sigma = 3.5$) was used as kernel. Parameter optimization was achieved by a grid search on a logarithmic scale and a subsequent search in the region of minimum error. The output of the SVR consists of one real-valued estimate for each emotion primitive [22].
We performed two different experiments for the automatic emotion estimation:

(a) train the algorithms with undisturbed speech and test them with the noisy speech, and
(b) use noisy speech for both training and testing.

While the first method allows for a more convenient training procedure and less effort to create training data, the latter one might provide better training conditions to the algorithms due to the same nature of training and test data.

6 Results

All results were calculated from 10-fold cross-validation experiments. The automatic emotion estimation under noise was compared to the reference given by the human evaluators. For each scenario the mean linear error was calculated. The accuracy of the tendency in the estimates was measured by the correlation between the estimates and the average ratings of the human evaluators (Pearson's empirical correlation coefficient).

The results for clean speech for both, training and testing, were added for comparison as a baseline. This baseline shows that, provided acoustically good conditions, emotion primitives estimation is possible with a mean error of 0.14 and a correlation of 0.42, 0.81 and 0.82 for valence, activation, and dominance, respectively. While the correlation coefficient for valence is only moderate, which is due to a very flat distribution, the emotion primitives activation and dominance were estimated with high reliability in clean speech.

6.1 Experiment (a)—training with clean speech and testing with noisy speech

The results of experiment (a) are reported in Tables 2 and 3, respectively. The performance of the emotion estimation mainly depends on the road surface and therefore on the SNR. In experiment (a), the mean error increased by 2% for the CTY scenario, which is almost neglectable. For the HWY and the COB scenarios, the mean error increased notably by 18% and dramatically by 44%, respectively.

The correlation coefficients have to be read skeptically for valence, since the statistical confidence was only moderate for this primitive ($p \geq 10^{-3}$). For the statistically significant correlation coefficients, however, a moderate (CTY: -4%) to remarkable (HWY: -14%, COB: -40%) decrease was observed.

Table 2. Results of experiment (a)—training with clean speech and testing with noisy speech: mean linear error. Baseline clean speech (CS) added for comparison.

	Valence			Activation			Dominance		
	HWY	COB	CTY	HWY	COB	CTY	HWY	COB	CTY
530i	0.15	0.17	0.13	0.18	0.24	0.16	0.18	0.20	0.15
645Ci	0.14	0.15	0.13	0.18	0.22	0.15	0.16	0.20	0.14
M5	0.13	0.16	0.13	0.18	0.25	0.16	0.18	0.21	0.13
Mini	0.14	0.17	0.13	0.20	0.25	0.17	0.18	0.22	0.14
CS	0.13			0.15			0.14		

Table 3. Results of experiment (a)—training with clean speech and testing with noisy speech: correlation coefficient between estimates and manual emotion labels. Baseline clean speech (CS) added for comparison.

	Valence			Activation			Dominance		
	HWY	COB	CTY	HWY	COB	CTY	HWY	COB	CTY
530i	(0.34)[1]	(0.14)	(0.43)	0.74	0.52	0.79	0.67	0.50	0.76
645Ci	(0.41)	(0.30)	(0.45)	0.75	0.57	0.81	0.70	(0.48)	0.78
M5	(0.39)	(0.19)	(0.46)	0.74	(0.46)	0.79	0.68	(0.40)	0.80
Mini	(0.35)	(0.10)	(0.45)	0.69	(0.40)	0.77	0.61	(0.38)	0.76
CS	(0.42)			0.82			0.81		

Thus, in the CTY scenario, the emotion recognition still works fine, almost independent of the vehicle type. On the highway (HWY) there is a notable decay in performance, but the recognition is still feasible. There is a clearly better result for the executive cars over the supermini in this case. On the cobbled road, the automatic recognition is not feasible any more. In this scenario, the 6 series convertible outperformed the other vehicles, but, still the results imply that the recognition is hardly possible.

6.2 Experiment (b)—training and testing with noisy speech

The results of experiment (b) are reported in Tables 4 and 5, respectively. In experiment (b), the results were better, which was probably the case because the calculated regression hyperplane could adapt to the noise scenarios. Still, the mean error increased by 2% and 7% for the CTY and the HWY scenarios, respectively, which implies that in this case the emotion recognition is sill possible with a mean error of 0.13 to 0.16. However, for the COB scenarios, the mean error increased by 16% indicating that emotion recognition is possible, but with a notable decay in performance. The correlation coefficients emphasize the fact that providing noisy data at the training state is very helpful.

[1] All correlation coefficients in brackets are only moderately statistically significant at $p \geq 10^{-3}$.

Table 4. Results of experiment (b)—both training and testing with noisy speech: mean linear error. Baseline clean speech (CS) added for comparison.

	Valence			Activation			Dominance		
	HWY	COB	CTY	HWY	COB	CTY	HWY	COB	CTY
530i	0.14	0.14	0.13	0.16	0.18	0.16	0.15	0.17	0.14
645Ci	0.14	0.14	0.13	0.16	0.17	0.16	0.15	0.16	0.14
M5	0.13	0.14	0.13	0.16	0.18	0.16	0.15	0.17	0.14
Mini	0.14	0.15	0.13	0.16	0.19	0.16	0.15	0.17	0.14
CS	0.13			0.15			0.14		

Table 5. Results of experiment (b)—both training and testing with noisy speech: correlation coefficient between estimates and manual emotion labels. Baseline clean speech (CS) added for comparison.

	Valence			Activation			Dominance		
	HWY	COB	CTY	HWY	COB	CTY	HWY	COB	CTY
530i	(0.38)	(0.32)	(0.43)	0.79	0.73	0.81	0.75	0.69	0.79
645Ci	(0.40)	(0.37)	0.44	0.79	0.78	0.81	0.76	0.71	0.79
M5	0.44	(0.34)	(0.43)	0.79	0.75	0.80	0.76	0.69	0.79
Mini	(0.39)	(0.35)	0.45	0.79	0.72	0.80	0.77	0.67	0.79
CS	(0.42)			0.82			0.81		

6.3 Discussion

It was found that experiment (b) gave clearly better results. In this case the noisy speech was already provided at the training step and thus the feature representation used for the determination of the regression hyperplane was more significant with respect to the test data. However, for the practical application, it is difficult to gather emotionally labeled training samples of the driver under different noise conditions. It is much easier to provide a large set of emotional training data if these can be gathered from clean speech. Therefore the question is, whether the good results achieved with noisy training data could also be achieved by a combination of providing clean speech training data and introducing filter techniques before extracting the acoustic features.

6.4 Noise reduction

A preliminary analysis of the noise signals revealed that these signals were highly concentrated in the low-frequency bands. They almost vanished in frequency bands above 130 Hz. However, cutting the noisy speech at 130 Hz is not reasonable since the fundamental frequency of several male speakers in our corpus were (locally) as low as 60 Hz. Therefore, we decided for a compromise and used a highpass filter that combined a very narrow stop band of 48 Hz and a rather wide transition band of 272 Hz. Thus, a great part of the noise was suppressed. In the critical frequency range of [48 Hz, 130 Hz], the noise was at least damped while still providing the crucial frequency information on the fundamental frequency of the speaker. For the implementation of the filter we used a FIR filter

Table 6. Results of emotion estimation in noisy, highpass filtered speech: mean linear error. Baseline clean speech (CS) added for comparison.

	Valence			Activation			Dominance		
	HWY	COB	CTY	HWY	COB	CTY	HWY	COB	CTY
530i	0.13	0.13	0.13	0.16	0.17	0.17	0.15	0.15	0.15
645Ci	0.13	0.13	0.13	0.16	0.17	0.16	0.15	0.15	0.15
M5	0.13	0.13	0.13	0.17	0.16	0.16	0.15	0.15	0.15
Mini	0.13	0.13	0.13	0.16	0.16	0.16	0.15	0.15	0.15
CS	0.13			0.15			0.14		

Table 7. Results of emotion estimation in noisy, highpass filtered speech: correlation coefficient between estimates and manual emotion labels. Baseline clean speech (CS) added for comparison.

	Valence			Activation			Dominance		
	HWY	COB	CTY	HWY	COB	CTY	HWY	COB	CTY
530i	(0.43)	0.44	(0.43)	0.79	0.79	0.78	0.77	0.77	0.77
645Ci	(0.43)	0.44	0.46	0.80	0.78	0.79	0.77	0.77	0.77
M5	0.45	0.44	(0.43)	0.79	0.79	0.79	0.77	0.77	0.77
Mini	0.45	0.44	(0.45)	0.79	0.79	0.79	0.77	0.77	0.77
CS	(0.42)			0.82			0.81		

of order 155 using the Parks-McClellan algorithm [26]. Such high filter order was necessary because of the very low cut-off frequency, which was only 0.006 times the sampling frequency.

The results of such highpass pre-processing were very promising. Tables 6 and 7 show the individual errors and correlation coefficients. The error was almost the same as the baseline: 0.13, 0.16, and 0.15 for valence, activation and dominance, respectively. It increased on average only by 6%, which indicates that the automatic emotion recognition is still possible. Furthermore, it was observed that the results now were almost the same for all vehicles. This can be explained by the fact that the more demanding noise scenario in the Mini, for example, was caused by more noise energy, but in the same frequency bands than with comparable executive cars.

7 Conclusion and Outlook

This paper reports current research on the emotion in human-computer interaction in the car. The first part of this study stressed the fact that detecting the driver's emotional state is indeed important. Such knowledge reveals information on the communication between the driver and the car instruments, and, in addition, can be used to design the car's answer in a way to provide best conditions for safe driving.

We presented results of emotion recognition in the speech when the signal is superimposed by car noise. Several vehicle types and road surfaces were

tested. The results were calculated on a continuous-valued, three-dimensional description basis for emotions consisting of the three emotion primitives valence, activation and dominance, each normalized to [-1,+1].

The results show that although sedan and executive type cars provide 2-3 dB better SNR than superminis, the road surface has more impact on the results than the car type. With our speech corpus consisting of spontaneous, unscripted emotional utterances, we observed that the automatic emotion recognition results correlated with the SNR, which was found to be 10 to 12 dB for city scenarios, 2 to 6 dB for the highway, and -7 to -2 dB for cobbled roads. The emotion recognition still worked fine for city and highway (only when noisy data was provided for training already) with a degradation of 2 and 7%, respectively. On rough cobbled roads the emotion recognition did not give acceptable results any more.

As an improvement pre-filtering was proposed for the highly relevant case of only clean speech training data being available. The application of a highpass filter with cut-off frequency as low as 48 Hz led to remarkable improvements. In this case the degradation from clean speech experiments was only 6% and emotion recognition was feasible with error rates of 0.13, 0.16, and 0.15 for valence, activation, and dominance, respectively.

While these results are based on a manual superposition of clean speech utterances and recorded noise signals of the cars, our future work will investigate the application of such an emotion recognition within the car in real time. Additionally, the emotion recognition results might be used to formulate behavior rules for the car's infotainment system once they are provided as a human-in-the-loop feedback signal.

Acknowledgment

This work was supported by grants of the Collaborative Research Center (SFB) 588 "Humanoid Robots" of the Deutsche Forschungsgemeinschaft (DFG).

References

1. Hitzenberger, L.: Man Machine Interaction in Car Information Systems. In: Proceedings of the First International Conference on Language Resources and Evaluation, Granada (1998) 179–182
2. Jonsson, I.M., Nass, C., Harris, H., Takayama, L.: Matching In-Car Voice with Driver State : Impact on Attitude and Driving Performance. In: Proceedings of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. (2005) 173–181
3. Nass, C., Jonsson, I.M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L.: Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In: Proc. CHI. (2005)
4. Jones, C., Jonsson, I.M.: Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In: Proc. OZCHI. (2005)
5. Schuller, B., Lang, M., Rigoll, G.: Recognition of Spontaneous Emotions by Speech within Automotive Environment. In: Proc. 32. Deutsche Jahrestagung für Akustik (DAGA), Braunschweig, Germany (2006) 57–58

6. Schuller, B., Arsic, D., Wallhoff, F., Rigoll, G.: Emotion Recognition in the Noise Applying Large Acoustic Feature Sets. In: Proc. Speech Prosody, Dresden, Germany (2006)
7. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper & Row, New York (1991)
8. Hirt, E., Melton, R., McDonald, H., Harackiewicz, J.: Processing goals, task interest, and the mood-performance relationship: A mediational analysis. Journal of Personality and Social Psychology **71** (1996) 245–261
9. Groeger, J.: Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task. Psychology Press, Philadelphia, PA (2000)
10. Isen, A., Rosenzweig, A., Young, M.: The influence of positive affect on clinical problem solving. Medical Decision Making **11**(3) (1991) 221–227
11. Galovski, T., Blanchard, E.: Road rage: A domain for psychological intervention? Aggression and Violent Behavior **9**(2) (2004) 105–127
12. Gross, J., Levenson, R.: Emotion elicitation using films. Cognition & Emotion **9** (1995) 87–108
13. Detenber, B., Reeves, B.: A bio-informational theory of emotion: Motion and image size effects on viewers. Journal of Communication **46**(3) (1996) 66–84
14. Izard, C.: Patterns of Emotions. Academic Press, New York (1972)
15. Nass, C., Gong, L.: Social aspects of speech interfaces from an evolutionary perspective: Experimental research and design implications. Communications of the ACM **43**(9) (2000) 36–43
16. Nass, C., Brave, S.: Wired for Speech: How Voice Activates and Enhances the Human-Computer Relationship. MIT Press, Cambridge, MA (2005)
17. Grimm, M., Kroschel, K.: Rule-based emotion classification using acoustic features. In: Proc. 3rd International Conference on Telemedicine and Multimedia Communication (ICTMC). (2005)
18. Lang, P.: The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders. Lawrence Erlbaum, Hillsdale, NJ (1985)
19. Grimm, M., Kroschel, K.: Evaluation of natural emotions using self assessment manikins. In: Proc. ASRU. (2005) 381–385
20. Grimm, M., Mower, E., Kroschel, K., Narayanan, S.: Primitives-based evaluation and estimation of emotions in speech. Speech Communication Journal (2007) accepted for publication.
21. Setiawan, P., Suhadi, S., Fingscheidt, T., Stan, S.: Robust Speech Recognition for Mobil Devices in Car Noise. In: Proc. Interspeech, Lisbon, Portugal (2005)
22. Grimm, M., Kroschel, K., Narayanan, S.: Support vector regression for automatic recognition of spontaneous emotions in speech. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Proccessing (ICASSP). (2007) Accepted for Publication.
23. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine **18**(1) (2001) 32–80
24. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
25. Smola, A., Schölkopf, B.: A tutorial on support vector regression. Technical report, NeuroCOLT2 (1998)
26. Kammeyer, K.D., Kroschel, K.: Digitale Signalverarbeitung. 4. edn. Teubner Stuttgart (1998)