Acoustic Emotion Recognition in Car Environment Using a 3D Emotion Space Approach

Michael Grimm[1]*, Kristian Kroschel[1], Björn Schuller[2], Gerhard Rigoll[2], Tobias Moosmayr[3]
[1] *Universität Karlsruhe (TH), Institut für Nachrichtentechnik, 76128 Karlsruhe, Germany*
[2] *Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany*
[3] *BMW Group, Forschungs- und Innovationszentrum, Akustik, Komfort und Werterhaltung, 80788 München, Germany*

Introduction

The automatic assessment of emotions conveyed in the speech signal has become a rapidly growing research interest in recent years. This paper focuses on a generalized framework to estimate emotions from the speech using an emotion space concept. The performance of such a system is studied in the acoustically demanding environment of vehicular noise while driving.

Due to the increasing number of speech-driven applications in the automotive environment, automatic emotion estimation from the speech signal of the driver has gained importance [1, 2]. Such assessment could provide information about the driver's satisfaction with the car's infotainment system, and, in particular, its human-machine-interface. Moreover it reflects the driver's perception of the traffic situation and thus reveals his/her stress level. The driver's emotion and his/her consciousness interact. Therefore the emotional state also affects the driving capability which is of utmost importance for the safety of all occupants [3, 4].

We describe emotions as points in a 3D emotion space consisting of three basic primitives (attributes): *valence* (negative–positive), *activation* (calm–excited), and *dominance* (weak–strong). All primitives are continuous and normalized to a range of [-1,+1]. Such a real-valued concept is helpful to distinguish spontaneous emotions, which we concentrate on.

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 [5]. These utterances were recorded from 47 speakers in a talk-show on TV. 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. They determined an appropriate value for each emotion primitive by means of assessment manikins [6]. The average standard deviation in their ratings was 0.31, and the average inter-evaluator correlation was 0.6.

Noise Scenario

To study the feasibility of emotion recognition in the car several noise scenarios of approx. 30 seconds were recorded. The microphone was mounted in the middle of the instrument panel, which is the standard for automatic speech recognition in the car. The recorded noise is a superposition of several influences: noise from the wheels/suspension, the combustion engine, interior squeak and rattle noise, and wind noise.

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

Table 1: Choice of Vehicles.

Choice of Vehicles. For this study we used four different cars as itemized in Tab. 1. Although the soft top of both convertibles was closed during the recordings the interior noise was noticeable higher than in comparable sedans. The supermini unifies convertible, hard suspension and sportive engine, and it thus provides the most demanding noise scenario.

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 are found with a constant driving over a smooth city road at 50 km/h and medium relovution. Higher noise levels are 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 rough surface involves dominant wheel/suspension noise as well as buzzes, squeaks and rattles.

Signal-to-Noise Ratio. The car noise 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. Due to the varying signal power in the speech recordings the result is a Gaussian-like distribution, which is shown in Fig. 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. The M5 resulted in 2-3 dB better results, the Mini in 2-3 dB worse results than the 530i or the 645Ci.

*Email: grimm@int.uni-karlsruhe.de
This work was supported by grants of the SFB 588 of the DFG.

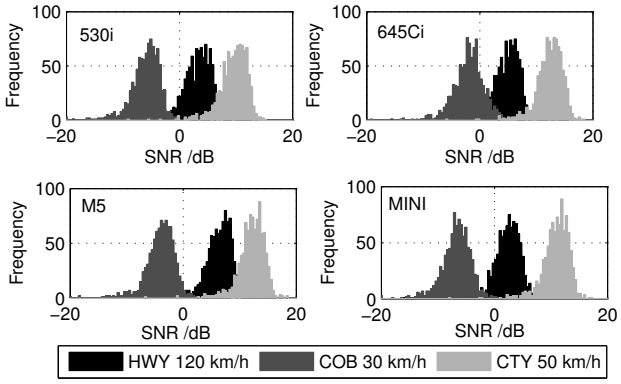

Figure 1: Signal-to-Noise Ratio distribution in the experiment.

Classification

The details of the emotion estimation system are described in [7]. We use a set of 20 prosodic features selected by Sequential Forward Selection. As a classifier we used Support Vector Regression (SVR) with a radial basis function as kernel. The output consists of one real-valued estimate for each emotion primitive. All results were calculated from 10-fold cross-validation experiments.

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.

Results

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 through the correlation between the estimates and the average ratings of the human evaluators. The results are reported exemplarily in Tab. 2 and 3 for our test b). The results for clean speech for both training and testing were added for comparison as a baseline.

The results show that the performance 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 show a similar degradation. 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 working fine. However, for the COB scenarios, the mean error increased notably by 16% indicating that emotion recognition is hardly possible in this case.

Conclusion

In this paper 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, and the results were calculated on a continuous-valued, three-dimensional description basis for emotions consisting of the three emotion primitives valence, activation and dominance.

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 we observed that the automatic emotion recognition results correlated with the SNR, which was found to be 10 to 12 dB for CTY, 2 to 6 dB for HWY, and -7 to -2 dB for COB. The emotion recognition still worked fine for CTY and HWY (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.

Our future work will investigate filtering the noisy speech before feature extraction, as well as an improved feature selection technique based on the noise scenario.

	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 2: Results: mean linear error. 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)[1]			0.82			0.81		

Table 3: Results: correlation coefficient between estimates and manual emotion labels. Baseline clean speech (CS) added for comparison.

References

[1] C.M. Jones and I.-M. Jonsson, "Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses," in *Proc. OZCHI*, 2005.

[2] B. Schuller et al., "Effects of In-Car Noise-Conditions on the Recognition of Emotion within Speech," in *Proc. DAGA*, 2007.

[3] J. Healey and R. Picard, "Smartcar: detecting driver stress," in *Proc. ICPR*, 2000, vol. 4, pp. 218–221.

[4] C. Nass et al., "Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion," in *Proc. CHI*, 2005.

[5] M. Grimm and K. Kroschel, "Rule-based emotion classification using acoustic features," in *Proc. ICTMC*, 2005.

[6] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proc. ASRU*, 2005, pp. 381–385.

[7] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Proc. ICASSP*, 2007, accepted.

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