EMOTION RECOGNITION FROM DISTURBED SPEECH - TOWARDS AFFECTIVE COMPUTING IN REAL-WORLD IN-CAR ENVIRONMENTS

Alicia Flores Lotz[1], *Fabian Faller*[2], *Ingo Siegert*[1], *Andreas Wendemuth*[1]

[1] *Institute for Information and Communications Engineering, Cognitive Systems Group, Otto-von-Guericke University, 39016 Magdeburg, Germany*

[2] *Interior Systems & Technology Advanced Development, Continental Automotive GmbH, 64832 Babenhausen, Germany*

alicia.lotz@ovgu.de

Abstract: Certain emotions can have a negative effect on the driver's capability of safely operating the vehicle and can ultimately lead to accidents. Therefore, it would be beneficial if the vehicle was able to detect the emotional state of the driver and provide appropriate assistance to mitigate these effects. This study investigates the influence of in-car acoustic characteristics and driving noises on emotion recognition from speech. The quality of the noisy speech samples was analyzed by calculation of SNR and CER[%]. Afterwards, classification experiments on high quality, in-car and noisy speech samples were carried out and evaluated. Data was recorded inside a car cabin in a simulator environment, resulting in realistic conditions where perturbations are being convoluted with the speech samples. For comparability with the state of the art, standard emotional speech databases were used for the evaluations conducted in this study. By considering the evaluated quality and classification measures, we conclude that high quality emotional speech is most severely impaired in the car, and that highway noise reduces the performance of the emotion classifier strongly. This leads to further requirements for in-car emotion recognition.

1 Introduction

It is well known that emotions can affect the behavior of a driver in negative ways [1, 2]. These emotions can be caused by internal sensations from out body, external factors or prior experiences [3, 4]. To mitigate these negative safety aspects, the car should be able to detect the emotional state of the drive and enhance the driving safety and comfort, for example by offering automation in critical situations (from driver assistance to highly automated driving).

From speech emotion recognition it is known that recognition rates can reach high values ($> 90\%$) for acted emotions under clear recording conditions. Further, research has shown to date that these rates drop considerably ($< 60\%$) if naturalistic emotions are considered [5], and if data is disturbed with artificial noise or if recordings are performed in noisy environments [6]. A less investigated scenario are in-car settings. Here, particular naturalistic emotions will be observed, and the acoustic recordings will be naturally perturbed by convoluted in-car noise which is of a special nature. Previous investigations have only superimposed clean speech to different car noise types and noise levels [7]. Some studies concentrate on convoluted in-car noise but disregard the presence of the in-vehicle acoustics [8], others are simply not designed to evaluate the driver's emotional state [9].

The present paper will investigate the quality and recognition performance of emotional speech under replayed simulated highway noise in a driving simulator, recorded with high quality directional microphones.

2 Database

As database the data samples of the Berlin Database of Emotional Speech (EMODB) [10] and the Vera am Mittag (VAM) corpus [11] were replayed under controlled test conditions in a driving simulator of Continental Automotive GmbH (see section 3). By choosing well-known datasets we obtain comparability with published results under different conditions. The original samples were normalized to a similar volume level and edited into one sound file interrupted by pauses of one second silence per utterance. Afterwards, the recorded sound files were separated into the original snippet, containing the identical utterance of the original sample but of disturbed nature. By replaying the samples in the simulator, the simulated car noise was not superimposed but convoluted with the speech samples.

Berlin Database of Emotional Speech :
The EMODB Dataset contains 494 utterances recorded in an anechoic chamber by 10 professional actors (5 male/5 female). Each actor simulated 10 sentences in seven different emotions (neutral, anger, fear, joy, sadness, disgust and boredom). In total 800 sentences were recorded. By conducting a perception test regarding the recognizability of the emotions and their naturalness, all utterances with a recognizability of over 80% and naturalness of over 60% were chosen as final samples of the dataset [10].

Vera am Mittag :
The VAM corpus is an audio-visual emotional speech database containing recordings of the German talk show "Vera am Mittag". The dataset comprises 946 utterances of 47 non-professional speakers (11 male/36 female). All utterances were labeled, in terms of the emotional dimensions *valence*, *arousal* and *dominance* using the SAMs, by 17 human listeners [11]. The labels were then mapped onto the four quadrants of the valence-arousal level (q1, q2, q3, q4) [5]. As the samples were taken from natural-like conversations, the number of utterances is unequally distributed among the emotional categories.

3 Recording Setup

3.1 The Simulator

The study is conducted in a fixed-base driving simulator, located at the premises of Continental Automotive GmbH in Babenhausen, Germany. The simulator consists of a BMW 5-series chassis, connected to the simulation environment and placed in front of a wide screen (Figure 1). Environmental noise as well as engine sound is generated by the simulation environment and fed into the vehicle chassis by a set of strategically placed speakers/actuators. The engine sound is generated by an actuator placed underneath the vehicle's engine hood using it as a resonator in order to create a realistic engine sound. Environmental noise is generated by speakers, placed in each front door and the rear window shelf, providing the experience of surrounding sound.

3.2 Microphone Integration

The recordings of the replayed emotional samples were conducted using two directional shotgun microphones placed at the A-pillars of the simulator vehicle directed towards the driver. The samples were played back from loudspeakers mounted at head heights on the driver's seat. Two different recording scenarios were used to obtain samples only influenced by the in-car acoustics (simulator turned off, no environmental noise present) and disturbed with simulated highway noise (simulator turned on).

Figure 1 – The driving simulator used in the study.

3.3 The Route

During the part of the experiment when the simulator is turned on, it is operated in automated driving mode in order to allow the placement of a speaker in the area where the driver's head would usually be located (see section 3.2). The simulated route models a two-lane highway and has been designed with varying traffic density in order to diversify the environmental noise during the course of the experiment. In addition to that, the vehicle's velocity has been manually adapted frequently and a number of lane changes have been manually triggered from the simulator's control room.

4 Methodology

This section gives a detailed description of the used methodologies. By replaying the original speech samples inside of the simulator, we assume a decrease in the quality of the samples. To evaluate this assumption the Signal-to-Noise Ratio (SNR) and Compression Error Rate (CER) were calculated. To also evaluate the influence of noise in speech emotion recognition, state-of-the-art classification experiments were carried out.

Peak Normalization :
The calculation of the SNR and CER are both based on the signal power of the speech samples. To obtain correct SNR and CER values the signal power of the waveform of the clean and noisy speech samples need to be of the same loudness. Because of the absorbing characteristics of the in-vehicle acoustic characteristics and the distance between the microphones and loudspeaker, the signals were peak normalized to overcome this difference. This method normalizes the speech sample to a desired maximum amplitude of the Waveform (dB). In the presented work a maximum amplitude of -1dB was chosen for all speech samples (clean and noisy).

Signal-to-Noise Ratio :
The SNR is defined as the ratio between the power of the clean speech sample (P_s) and the power of the noise signal (P_n) superimposed to the clean speech and is denoted in dB:

$$SNR = 10 \cdot \log_{10}(\frac{P_s}{P_n}) \tag{1}$$

For the recording setup presented in section 3, the noise signal is convoluted with the replayed audio sample (P_{ns}). A separate recording of the in-car noise was not realized. Therefore eq. (1) needs to be adapted. It is assumed, that the noise power can be estimated by subtracting the power of the noisy speech from the power of the clean speech: $P_n = P_{ns} - P_s$. This assumption is only valid for ideal recordings, where the power of the clean speech is identical to the power of the speech part contained in the noisy speech. In the presented study this is not the case, as:

1. The acoustic characteristics inside the simulator vehicle suppresses the signal replayed by the microphones. This leads to a general reduction of the signals' power.

2. For the noisy speech signal the noise amplitude can exceed the amplitude of the speech part. By computing a peak normalization, the noisy speech signal is normalized to the maximum amplitude of the noise and not of the speech content.

These effects may lead to incorrect SNR values and a negative denominator in eq. (1). Therefore a positive constant α was introduced by the authors of [9]. Botinhau and Yamagishi introduce the constant parameter α to compress the clean speech power and, by doing so, overcome the problem of a negative denominator:

$$SNR = 10 \cdot \log_{10}\left(\frac{\alpha \cdot P_s}{P_{ns} - \alpha \cdot P_s}\right) \tag{2}$$

In [9] the α value was chosen so that $P_{ns} - \alpha \cdot P_s > 0$ holds true for every speech sample. We adapted this approach to our needs. By calculating α only using those speech samples recorded while the simulator was turned off, α takes into account the in-vehicle acoustic characteristics and the chosen recordings' setup. This value was then utilized to calculate the SNR of the speech samples recorded while the simulator was turned on. By using this approach, a negative denominator may occur in eq. (2), caused by the peak normalization. These SNRs, comprising only 4 samples, were excluded from further evaluations.

Compression Error Rate :
The CER as presented in [12] measures the absolute difference between the original and noisy signal's spectrogram in dB. To make results comparable over different datasets and experimental setups, the measure was adapted so that the difference between the two spectrograms is given as a percentage value. The spectrogram is computed and standardized as described in [12]. Negative values of the CER[%] characterize a decrease of power in the noisy signal compared to the original sample and positive values an increase of power, respectively.

Classification Experiments :
The classification experiments were carried out using the software tool WEKA [13]. As state-of-the-art classifier we opt for Support Vector Machines, which were trained and tested via the leave-one-speaker-out approach, using the "emobase" features obtained from the openSMILE feature extraction toolkit [14].

Unweighted Average Recall :
The Unweighted Average Recall (UAR) is used as measure to indicate the performance of a conducted classification experiment over its mean average recall of one speaker [15]. For the recognition experiments the UAR was calculated for each speaker over all present emotions. Afterwards, these UARs per speaker were averaged over all speakers.

5 Realization

The classification experiments were conducted using the non-normalized speech samples. By doing so we overcome power differences of samples within the recording setups (left/right microphone and simulator on/off), caused by effect 2 described in the section 4. This also means, that the power level of the original replayed speech samples can differ significantly from the power of the recorded speech samples. To test if this difference in the signals' power result in significant changes of the classification experiments, the classification results of the original recorded data and normalized recorded data were tested against each other using analysis of variance (ANOVA). As the results of these pre-tests showed that the difference between the original and normalized samples are not significant for all recordings (P-value > 0.6), we assume that the signals' power does not affect the classification results significantly, and we can therefore use the original non-normalized recordings for further analysis.

An overview on the conducted experiments is given in table 1. The experiments were all carried out for each corpus (EMODB/VAM), microphone (left/right) and simulator setting

(on/off) separately, resulting in 8 classification experiments used for evaluation (#experiment: 1-8) and 2 experiments as baseline using the original data sample (#experiment: 9,10).

Table 1 – Overview of the conducted classification experiments used for evaluation.

# experiment	corpus	microphone	simulator
1	EMODB	left	off
2	EMODB	right	off
3	EMODB	left	on
4	EMODB	right	on
5	VAM	left	off
6	VAM	right	off
7	VAM	left	on
8	VAM	right	on
9	EMODB	-	-
10	VAM	-	-

6 Evaluation of Results

6.1 Quality Assessment

Table 2 – Means and standard deviation of SNR and CER for EMODB and VAM.

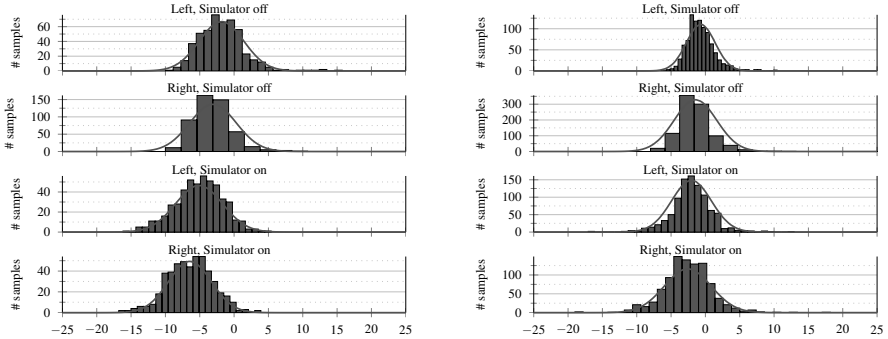
# setting	SNR	# setting	CER
1	-1.83 (3.24)	**1**	-5.77 (6.40)
2	-3.07 (3.38)	**2**	-5.09 (6.53)
3	-5.10 (3.43)	**3**	13.34 (8.26)
4	-6.43 (3.12)	**4**	14.36 (9.54)
5	-0.68 (2.09)	**5**	-7.23 (3.94)
6	-1.45 (3.09)	**6**	-7.64 (5.82)
7	-2.03 (2.89)	**7**	9.34 (7.04)
8	-2.57 (3.41)	**8**	11.22 (9.09)

For the evaluation of the quality assessment via SNR and CER only the normalized recorded data samples of the left and right microphone were used. It was assumed that for the "simulator off"-setting, only the in-vehicle acoustics and the recording setup influenced the quality of the recording. The quality measures of this setting will therefore give information on how much these influences affect the recording quality. It was also assumed that the left microphone recordings, located closer to the loudspeaker, receive better quality values than the right one.

The histograms (and fitted normal distributions) of the SNR distributions for the EMODB and VAM corpus are shown in figure 2, with all microphone and simulator conditions. Their means and standard deviations are given in table 2. It can be stated that the differences in the SNR are highly significant for both, simulator setting and microphone placement ($P-\text{value} < 0.001$). As the volumes of the replayed original speech samples for EMODB and VAM were adjusted to a similar level, the differences in the SNR-level can be explained by the different recording qualities of the original datasets: EMODB was recorded in an anechoic chamber with no surround sound/noise being present, while VAM was recorded in a television studio, with audience and other interference factors. This makes the EMODB data samples more prone to the in-vehicle acoustic characteristics and the highway noise of the simulator.

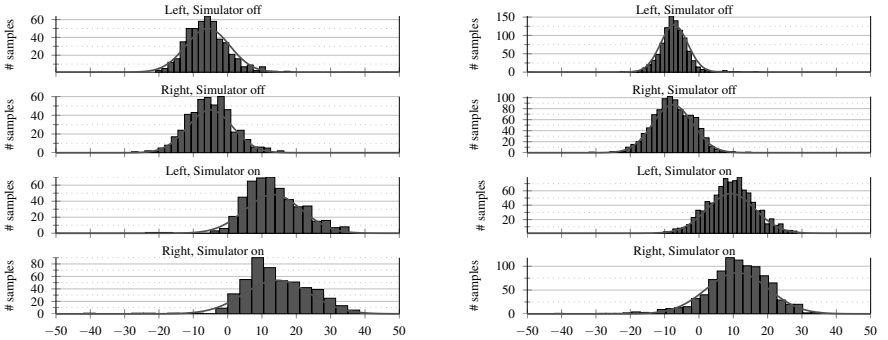
The results of the CER are given in figure 3 and table 2. By using the CER, information on the spectral difference between the original and recorded speech samples is obtained. For both, EMODB and VAM, the "simulator off"-setting results on average in negative CER-values which indicates a decrease of spectral power of the recorded signal compared to the original data samples. This can be explained by absorbing characteristic of the in-vehicle acoustics. For the

"simulator on"-setting the average CER-values are positive, indicating an increase of spectral power, caused by the convoluted highway noise of the simulator. The differences between the "simulator off" and "simulator on"-setting of the left and right microphone are highly significant for both datasets ($P-\text{value} < 0.001$).



(a) SNR of EMODB data samples. **(b)** SNR of VAM data samples.

Figure 2 – Histograms of the SNR-distributions and fitted normal distributions of the "simulator on" and "simulator off"-setting for the left and right microphone of the recorded EMODB and VAM samples.



(a) CER of EMODB data samples. **(b)** CER of VAM data samples.

Figure 3 – Histograms of the CER-distributions and fitted normal distributions of the "simulator on" and "simulator off"-setting for the left and right microphone of the recorded EMODB and VAM samples.

6.2 Classification

The results of the classification experiments described in section 5 can be taken from figure 4. The results of the reference classifiers, trained and tested on the original data samples of EMODB ($UAR = 79.61\%$) and VAM ($UAR = 55.18\%$), are displayed on the left side of the bar plots (crosshatched - blue). The classification results of the "simulator off" and "simulator on"-setting are marked in red and green for both microphone placements (left/right). The differences in the performance of the EMODB and VAM data samples are reasonable, as the datasets contain different levels of naturalness and recording quality. Because of the EMODB samples being more prone to the in-vehicle acoustics and highway noise of the simulator, the performance of the trained classifiers decreases, respectively.

For EMODB the performance of the classifier trained on the "simulator on" samples is

significantly lower ($P-\text{value} < 0.02$), compared to the baseline result. The positioning of the microphones does not influence the performance of the classifiers significantly. The UAR of the classifier obtained from the left and right microphone samples are similar considering both simulator settings.

For VAM the emotion recognition of the in-car settings do not differ significantly from the baseline recognition. This may be attributed to the fact that the naturalistic VAM data was recorded in a talk show environment (as compared to EMODB's studio environment) with perturbing acoustic characteristics, hence the in-car setting has little further disturbing effect. The performance of the evaluated classifiers does not differ significantly from each other, but it can be noticed that a classifier trained on the "simulator on"-setting of the left microphone samples outperforms the classifier of the right microphone samples. This is reasonable, as the left microphone was positioned closer to the loudspeaker and therefore absorbed less noise, compared to the right microphone.

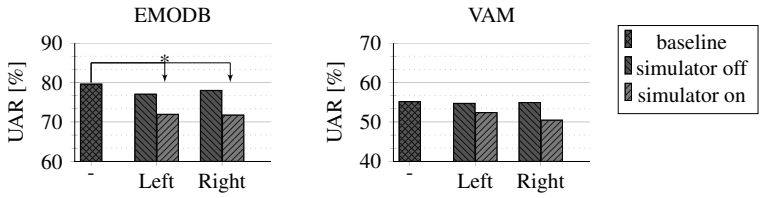


Figure 4 – Recognition results of the implemented classifiers for EMODB and VAM. The Stars denote the significance level: * ($P-\text{value} < 0.02$) using ANOVA

7 Conclusion and Outlook

By calculating the Signal-to-Noise Ratio and the Compression Error Rate [%] we could draw a conclusion on the recorded signals' quality. For the SNR a significant quality difference could be shown for all recordings' setups, independent of the simulator setting (on/off) and microphone positioning (left/right). It could be shown, that the "simulator on"-setting leads to a significant decrease of the SNR compared to the "simulator off"-setting. This drop in signal quality was also shown by an increase of the CER, which indicates an increase of the signals' spectral power. From the classification results we could identify a relation between the decrease of signal quality and the decrease of the classifiers performance. For EMODB the average signal quality was indicated lower as for VAM. This also resulted in a higher performance decrease for all classification experiments. It can also be stated that the simulated highway noise impairs the performance of the classifier strongly, for EMODB even significantly ($\alpha = 0.05$).

As outlook a multi-style training of the presented recordings is planned. Leave-one-speaker-out classification experiments will be carried out, trained on data samples of the original dataset and disturbed sample recordings. Additionally the performance of classifiers trained on both recorded samples of the left and right microphone for each simulator setting will be evaluated.

Acknowledgment

This paper has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 68890.

References

[1] GARASE, M. L.: *Road Rage*. LFB Scholarly Publishing LLC, 2006.

[2] GRIMM, M., K. KROSCHEL, H. HARRIS, C. NASS, B. SCHULLER, G. RIGOLL, and T. MOOSMAYR: *On the Necessity and Feasibility of Detecting a Drivers Emotional State While Driving*. In *Proc. of the ACII-2007*, vol. 4738 of *LNCS*, pp. 126–138. Springer, Berlin, Heidelberg, Lisbon, Protugal, 2007.

[3] AMERICAN AUTOMOTIVE ASSOCIATION: *Americans feel unsafe sharing the road with fully self-driving cars*. http://newsroom.aaa.com/2017/03/americans-feel-unsafe-sharing-road-fully-self-driving-cars/, 2017.

[4] KOO, J., J. KWAC, W. JU, M. STEINERT, L. LEIFER, and C. NASS: *Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance*. *International Journal on Interactive Design and Manufacturing*, 9(4), pp. 269–275, 2015.

[5] SCHULLER, B., B. VLASENKO, F. EYBEN, G. RIGOLL, and A. WENDEMUTH: *Acoustic Emotion Recognition: A Benchmark Comparison of Performances*. In *Proc. of the IEEE ASRU-2009*, pp. 552–557. 2009.

[6] CHENCHAH, F. and Z. LACHIRI: *Speech emotion recognition in noisy environment*. In *Proc. of the ATSIP-2016*, pp. 788–792. Monastir, Tunisia, 2016.

[7] GRIMM, M., K. KROSCHEL, B. SCHULLER, G. RIGOLL, and T. MOOSMAYR: *Acoustic Emotion Recognition in Car Environment Using a 3D Emotion Space Approach*. In *Proc. of the DAGA-2007*. Stuttgart, Germany, 2007.

[8] JONES, C. M. and I.-M. JONSSON: *Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces*. In C. STEPHANIDIS (ed.), *Proc. of the UAHCI-2007*, vol. 4555 of *LNCS*, pp. 411–420. Springer, Berlin, Heidelberg, Beijing, China, 2007.

[9] BOTINHAO, C. V. and J. YAMAGISHI: *Speech intelligibility in cars: the effect of speaking style, noise and listener age*. In *Proc. of the INTERSPEECH-2017*, pp. 2944–2948. Stockholm, Sweden, 2017.

[10] BURKHARDT, F., A. PAESCHKE, M. ROLFES, W. SENDLMEIER, and B. WEISS: *A database of german emotional speech*. In *Proc. of the INTERSPEECH-2005*, pp. 1517–1520. Lisbon, Portugal, 2005.

[11] GRIMM, M., K. KROSCHEL, and S. NARAYANAN: *The vera am mittag german audio-visual emotional speech database*. In *Proc. of the IEEE ICME-2008*, pp. 865–868. Hannover, Germany, 2008.

[12] LOTZ, A. F., I. SIEGERT, M. MARUSCHKE, and A. WENDEMUTH: *Audio compression and its impact on emotion recognition in affective computing*. In *Elektronische Sprachsignalverarbeitung 2017*, vol. 86 of *Studientexte zur Sprachkommunikation*, pp. 1–8. 2017.

[13] HALL, M., E. FRANK, G. HOLMES, B. PFAHRINGER, P. REUTEMANN, and I. H. WITTEN: *The weka data mining software: An update*. *SIGKDD Explor. Newsl.*, 11(1), pp. 10–18, 2009.

[14] EYBEN, F., M. WÖLLMER, and B. SCHULLER: *openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor*. In *Proc. of the ACM MM-2010*, p. s.p. Firenze, Italy, 2010.

[15] ROSENBERG, A.: *Classifying skewed data: Importance weighting to optimize average recall*. In *Proc. of the INTERSPEECH-2012*, pp. 1392–1395. Portland, USA, 2012.