Speaker, Noise, and Acoustic Space Adaptation for Emotion Recognition in the Automotive Environment

Björn W. Schuller

Institute for Human-Machine Communication (MMK), Technische Universität München (TUM), D-80333 München
E-Mail: schuller@tum.de
Web: http://www.mmk.ei.tum.de/

Abstract

Emotional surveillance of drivers possesses significant potential for increased security within passenger transport. In an automotive setting the interaction can further be improved by social awareness of an MMI. Within this scope the detection of security relevant behavior patterns as aggressiveness or sadness is discussed. The focus lies on real-life usage respecting online processing, subject independency, and noise robustness. The approach introduced employs multivariate time-series analysis by brute-force feature generation. Extensive results are reported on two public standard corpora. The influence of noise is discussed by representative car-noise overlay. Thereby impact per low-level-descriptor is considered.

1 Introduction

Recognition of the emotional state of a car driver posseses great potential with respect to safety and enhanced comfort [1-5].

However, the noise present in an automotive environment downgrades performance of engines [6-9]. In contrast to a quiet test stand, emotion recognition in a real vehicle has to fight against hard conditions: first, speech is not recorded close to the speaker's mouth, but in an immediate adjacency of the instrument panel. Thus the signal can be modified by the room acoustics dependent mouth-to-microphone transfer function. Second, while driving speech is superposed by several background noises. Their acoustic masking effect may hide important information and is large compared to that of the mouth-to-microphone transfer function, which is therefore neglected in the ongoing.

Likewise we consider the impact of car-noise on the recognition performance by a variety of cars and driving situations. Further, we investigate how to cope with noise by three simple adaptation strategies: first matched conditions learning with noisy speech samples. This cannot easily be guaranteed in an actual engine, as matching will not always be possible without error. However, by speed and GPS data, an estimate of the best noise samples can be made. Second, we consider feature selection in the noise. This too requires condition matching. Finally, the models can also be adapted to the driver's voice. The knowledge of the emotion is thereby not necessary. However, it seems desirable to have a broad variation of emotions in the adaptation material.

Apart from selection of mixed features spanning all types of low-level-contours as pitch or MFCC, we further analyse the impact of each adaptation strategy on contour types.

The paper is structured as follows: we first introduce our acoustic features in sec. 2. Next, the speech and noise data is detailed in sec. 3. Sec. 4 shows our experimental results, which are concluded in sec. 5.

2 Acoustic Features

In order to represent a state-of-the-art emotion recognition engine, we cover prosodic, articulatory and voice quality features known to carry information about emotion by use of 1,406 acoustic systematically generated acoustic features as used in [10].

Table 1: Audio Low-Level-Descriptors and functionals.

LLD (2x37)	Functionals (19)
(Δ) Pitch	Mean, Centr., Std. Dev.
(Δ) Energy	Skewness, Kurtosis
(Δ) Envelope	Quartile 1,2,3
(Δ) Formant 1-5 Amplitude	Quartile 1 - Minimum
(Δ) Formant 1-5 Bandwidth	Quartile 2 - Quartile 1
(Δ) Formant 1-5 Frequency	Quartile 3 - Quartile 2
(Δ) MFCC Coefficient 1-16	Maximum - Quartile 3
(Δ) HNR	Max., Min., Rel. Pos.
(Δ) Shimmer	Range, ZCR
(Δ) Jitter	Pos. 95% Roll-Off

These base on 37 typical Low-Level-Descriptors (LLD) as seen in table 1 and their first order delta coefficients [2]. These 37x2 LLD are next smoothed by low-pass filtering with an SMA-filter. Such systems next derive statistics per speaker turn by a projection of each univariate time series, respectively LLD, onto a scalar feature independent of the length of the turn [6]. This is realised by use of a functional, such as statistical moments or extremes. 19 functionals are applied, herein, to each LLD on speaker-turn-level covering extremes, ranges, positions, first four moments and quartiles as shown in table 1. Note that three functionals are related to position, known as duration in traditional phonetic terminology, as their physical unit is msec.

3 Emotion and Noise Data

3.1 EMO-DB

First, the Berlin Emotional Speech Database (EMO-DB) [11] is an audio only German emotion database of 10

professional actors (5 female). The recordings took place in an anechoic chamber with 16 kHz, 16 bit, thus allowing for systematic noise overlay. For each of 7 emotions (anger, boredom, disgust, fear, happiness, sadness, and neutrality), 10 sentences of emotionally neutral content were spoken by each speaker. The final data-set consists of 494 samples.

3.2 eNTERFACE

The eNTERFACE corpus is a further public, yet audio-visual emotion database [12]. It consists of the 'big six' emotion set (MPEG-4: surprise instead of boredom and no neutrality in comparison to EMO-DB), and contains 44 subjects from 14 nations. As EMO-DB, it consists of studio recordings of pre-defined spoken content, but in English. Each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion. They then had to react to each of the situations and two experts judged whether the reaction expressed the emotion in an unambiguous way. Only if this was the case, the sample was added to database. The audio sample rate is 48 kHz, 16-bit. Overall, the database consists of 1,170 samples.

3.3 Car-Noise

To study the impact on recognition in the car noise environment, several noise scenarios were recorded [1,5]. A condenser microphone was therefore mounted in the middle of the instrument panel of diverse cars. In order to cover a wide spectrum of car versions, speech from EMO-DB and eNTERFACE is superposed by the interior noise of four very different vehicles, namely a BMW 5 series Touring and 6 series Convertible as executive cars, an M5 Sedan as sports car, and a MINI Cooper Convertible as Super-mini. In this vehicle choice, the influence and configuration of single noise sources differs. The worst case is represented by the MINI. 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), and accelerated highway drive (ACC, only for M5). Eventually, a total of 13 car-noise scenarios is simulated. Every noise scenario takes approximately 30 seconds. Additionally, ambient babble noise was recorded to simulate voice over-talk deriving e.g. from a car-stereo, a communication device, or passengers. The recording was carried out during business hour in a pedestrian street in downtown Munich, Germany, with the same microphone, and takes approximately one hour.

Noise is normalised to 125dB prior to addition to non-normalised speech. In the ACC and babble scenario we connected samples of each emotion separately prior to noise-stream-overlay. In any other scenario a clip of the according length of the spoken utterance was cut from the beginning of the noise recordings.

4 Experimental Results

We provide results in a Leave-One-Subject-Out (LOSO) manner. This ensures speaker independency at any time, as required in a reasonable in-car system [9]. As classifier Support Vector Machines (SVM) with polynomial kernel and a one-vs.-one multiclass discrimination strategy are used. Learning is carried out by Sequential Minimal Optimisation (SMO) [13].

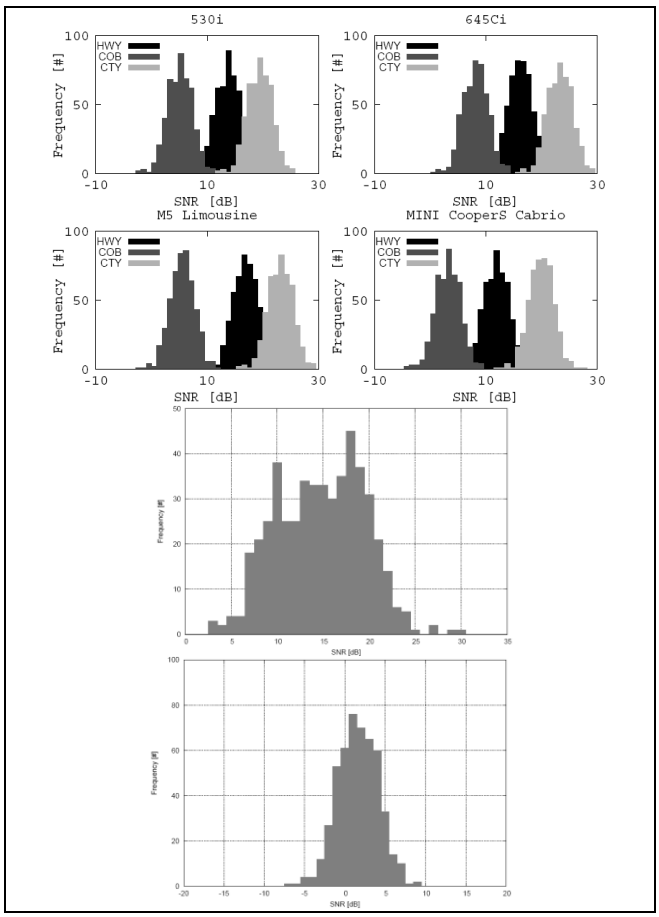
First, figure 1 depicts the effect of additive noise overlay for the two databases EMO-DB and eNTERFACE with respect to SNR level distribution.

Second, table 2 shows observed accuracies for emotion recognition on the databases EMO-DB and eNTERFACE. Note that car noise is summarised by the mean over all car types and driving situations. As a worst case scenario we also provide results for the MINI on big cobbles (COB), which proved the hardest challenge, overlaid with the babble noise.

Since noise clearly degrades performance, we herein introduce four compensation strategies: first, noise adaptation (NA) by training in the noise and recognition assuming matched conditions (as can be realised by speed indicator or GPS data); second, speaker adaptation (SA) by mean and standard deviation normalisation for each speaker, individually. Thereby the whole speaker context is used. Note that in a real adaptation scenario no emotion information is needed for SA; third, combined speaker and noise adaptation (NSA). Finally, as a novel strategy, we combine noise and speaker adaptation with noise specific feature selection (NSAFS) by Correlation-based Feature Subset Selection (CFSS) [13] with Sequential Floating Forward Search to avoid NP-hard exhaustive search. As a result, accuracies can be step-wisely “repaired” by combination of methods, even in the worst case noise scenario.

Tables 3 and 4 detail these results by car type and surface, respectively driving situation. Apparently, the car type has less influence on the accuracy degradation than the driving situation. The worst cases thereby are the slow drive over big city cobbles, and the accelerated highway drive, as one would also assume from experience.

We finally provide details on the impact of noise with respect to feature type as in [14]. Features are thereby grouped by their according LLD type: prosodic, voice quality, and spectral/cepstral. Table 5 shows the different effect of the diverse named strategies to cope with noise. As can be seen by the shown absolute gain, adaptation generally clearly helps to improve. However, the combination of noise and speaker adaptation is not necessarily the best choice - depending on the LLD type. Generally, eNTERFACE shows a more evenly distributed downgrade over all types. This may derive from the on average lower SNR ratio in comparison to EMO-DB. On EMO-DB the energy related features suffer most from noise-overlay. Independent of the database, spectral features seem a good choice.



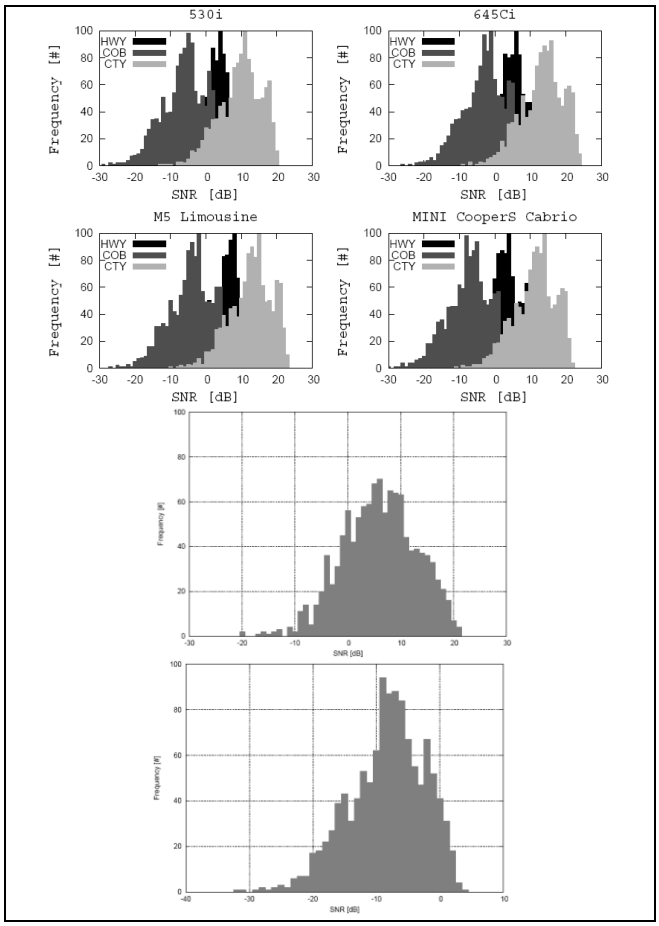


Figure 1 (left): SNR distribution EMO-DB (top) and eNTERFACE (bottom). Each: upper four: diverse cars; middle: babble noise; bottom: MINI + babble.

Table 2: Emotion recognition, diverse noises.

Acc. [%]	-	NA	SA	NSA	NSAFS
EMO-DB					
Clean	74.9	-	79.6	-	**80.4**
Car Noise	70.0	76.1	77.9	78.7	**80.5**
Bab. Noise	60.5	72.1	75.1	76.3	**77.3**
Bab.+MINI	46.6	70.4	75.7	76.1	**79.5**
eNTERFACE					
Clean	54.2	-	61.4	-	**62.8**
Car Noise	42.1	53.2	54.2	61.0	**61.6**
Bab. Noise	38.5	48.3	51.8	56.7	**59.7**
Bab.+MINI	30.6	49.8	46.2	55.8	**58.6**

Table 3: Emotion recognition, cars/surfaces, EMO-DB.

Acc. [%]	-	NA	SA	NSA	NSAFS
530i					
HWY	71.7	76.3	79.2	78.5	**80.6**
COB	66.8	78.7	74.5	79.1	**81.2**
CTY	72.9	72.9	78.9	75.1	**79.7**
645Ci					
HWY	71.7	74.3	78.1	79.6	**80.8**
COB	70.0	78.1	79.6	81.4	**83.0**
CTY	73.7	76.1	79.6	81.6	**82.2**
M5					
HWY	72.0	74.7	77.5	77.5	**81.6**
COB	66.6	75.3	78.4	79.1	**78.7**
CTY	73.3	74.5	**78.4**	78.1	76.7
ACC	61.5	74.7	74.3	76.9	**81.0**
MINI					
HWY	68.4	78.9	78.1	79.1	**82.0**
COB	67.0	76.9	77.9	78.7	**81.2**
CTY	74.5	77.3	77.3	77.9	**79.4**

Table 4: Emotion recognition, cars/surfaces, eNTERFACE.

Acc. [%]	-	NA	SA	NSA	NSAFS
530i					
HWY	43.2	53.5	55.4	61.0	**60.2**
COB	35.7	53.5	50.9	**63.2**	59.6
CTY	49.9	52.8	56.6	61.5	**62.5**
645Ci					
HWY	41.8	53.1	54.2	61.4	**61.6**
COB	37.4	56.2	53.4	**61.6**	60.4
CTY	49.9	53.2	58.2	60.9	**63.2**
M5					
HWY	42.6	54.0	55.6	62.7	**63.9**
COB	35.6	53.8	48.9	58.5	**60.9**
CTY	47.4	54.2	56.3	**63.0**	61.2
ACC	43.1	49.7	52.4	56.6	**60.3**
MINI					
HWY	40.3	52.6	55.7	**62.1**	60.7
COB	32.9	51.6	48.8	59.6	**60.1**
CTY	47.2	53.3	57.4	51.3	**65.0**

Table 5: Influence on low-level-descriptor types (number per type provided, c.f. table 1) for speaker-independent emotion recognition, diverse noises. Prosodic (pitch (F0) and energy plus envelope (EN)), voice quality (VQ, jitter, shimmer, HNR), and spectral (formants 1-5 amplitude, bandwidth, frequency (FO), MFCC 1-16 (MF)) plus delta, each. Per type: accuracy clean speech, absolute loss noisy speech, and absolute gain (+) by adaptation.

[%] #	Type	F0 38	EN 76	VQ 114	FO 570	MF 608
EMO-DB						
Clean	**acc.**	**48.2**	**57.5**	**47.8**	**65.8**	**48.2**
	SA	8.9	2.0	0.6	4.6	8.9
MINI	**loss**	**-0.9**	**-17.4**	**-5.3**	**-5.5**	**-0.9**
	+NA	0.7	12.1	11.1	4.3	0.7
	+SA	7.2	3.6	3.7	6.7	7.2
	+NSA	6.1	8.7	6.1	9.3	6.1
MINI+ Babble	**loss**	**-4.3**	**-26.5**	**-5.9**	**-22.1**	**-4.3**
	+NA	0.4	16.1	7.3	23.5	0.4
	+SA	7.9	9.7	-3.8	16.2	7.9
	+NSA	6.9	19.4	2.2	25.0	6.9
eNTERFACE						
Clean	**acc.**	**39.0**	**33.4**	**31.9**	**43.6**	**48.8**
	SA	1.0	2.5	-0.4	6.1	6.1
MINI	**loss**	**-13.7**	**-12.7**	**-12.2**	**-11.5**	**-15.0**
	+NA	0.3	8.2	2.8	15.2	5.4
	+SA	7.9	11.6	12.4	8.0	15.3
	+NSA	13.7	16.1	14.2	24.7	14.0
MINI+ Babble	**loss**	**-13.7**	**-12.4**	**-11.8**	**-12.5**	**-17.6**
	+NA	7.1	11.4	11.4	8.5	15.9
	+SA	0.6	6.6	-0.8	8.6	11.1
	+NSA	9.1	14.6	11.9	12.4	20.5

5 Conclusion

In this paper we introduced recognition of emotion in the car. The requirement of real-time capability could be fulfilled. Extensive results were presented facing subject independency, and considering diverse noise scenarios as to be expected in real-life application. While noise heavily downgraded recognition accuracies, this loss could be overcome by the introduced novel combined noise and speaker adaptation with matched feature selection.

Further efforts will have to investigate realistic data to allow for more accurate insights and model constructions.

Acknowledgements

The author would like to thank Tobias Moosmayr at the BMW Group in Munich, Germany for provision of the audio recordings in the car. Further this work highly benefits from the cooperation with the student researcher Naijiang Lu.

Literature

[1] Grimm, M.; Kroschel, K.; Harris, H.; Nass, C.; Schuller, B.; Rigoll, G.; Moosmayr, T.: On the Necessity and Feasibility of Detecting a Driver's Emotional State While Driving. *Proc. ACII 2007*, Lisbon, 2007, ACM, Springer, pp. 126-138.

[2] Jonsson, I.M., Nass, C., Harris, H., Takayama, L.: Matching In-Car Voice with Driver State : Impact on Attitude and Driving Performance. *Proc. of the Third International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, 2005, pp. 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. *Proc. CHI*, 2005.

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

[5] Schuller, B., Lang, M., Rigoll, G.: Recognition of Spontaneous Emotions by Speech within Automotive Environment. *Proc. DAGA*, Braunschweig, Germany, 2006, pp. 57–58.

[6] Schuller, B., Arsic, D., Wallhoff, F., Rigoll, G.: Emotion Recognition in the Noise Applying Large Acoustic Feature Sets. *Proc. Speech Prosody*, Dresden, Germany, 2006.

[7] You, M.; Chen C.; Bu, J.; Liu, J.; Tao, J.: Emotion Recognition from Noisy Speech. *Proc. ICME*, 2006, pp. 1653 - 1656.

[8] Lee, K.-K.; Cho, Y.-H.; Park, K.-S.: Robust Feature Extraction for Mobile-Based Speech Emotion Recognition System In: *Intelligent Computing in Signal Processing and Pattern Recognition*, LNCIS, Springer Berlin/Heidelberg, Volume 345/2006, 2006, pp. 470-477.

[9] Schuller, B.; Seppi, D.; Batliner, A.; Maier, A.; Steidl, S.: Towards More Reality in the Recognition of Emotional Speech. *Proc. ICASSP*, IEEE, Vol. IV, Honolulu, Hawaii, 2007, pp. 941-944.

[10] Batliner, A.; Schuller, B.; Schaeffler, S.; Steidl, S.: Mothers, Adults, Children, Pets - Towards the Acoustics of Intimacy. *Proc. ICASSP*, IEEE, Las Vegas, Nevada, 2008, pp. 4497-4500.

[11] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B.: A Database of German Emo-tional Speech, *Proc. INTERSPEECH*, ISCA, Lisbon, Portugal, pp.1517-1520, 2005.

[12] Martin, O.; Kotsia, I.; Macq, B.; Pitas, I.: The enterface'05 Audio-Visual Emotion Database. *Proc. IEEE Workshop on Multimedia Database Management*, Atlanta, 2006.

[13] Witten, I.H.; Frank, E.: *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000, pp. 133.

[14] Schuller, B.; Batliner, A.; Seppi, D.; Steidl, S.; Vogt, T.; Wagner, J.; Devillers, L.; Vidrascu, L.; Amir, N.; Kessous, L.; Aharonson, V.: The Relevance of Feature Type for the Automatic Classification of Emotional User States: Low Level Descriptors and Functionals, *Proc. INTERSPEECH 2007*, Antwerp, Belgium, 2007, ISCA, pp. 2253-2256.