Effects of In-Car Noise-Conditions on the Recognition of Emotion within Speech

Björn Schuller[1]*, Gerhard Rigoll[1], Michael Grimm[2], Kristian Kroschel[2], Tobias Moosmayr[3], Günther Ruske[1]

[1] *Technische Universität München, Institute for Human-Machine Communication, 80290 München, Germany*

[2] *Universität Karlsruhe (TH), Institut für Nachrichtentechnik, 76128 Karlsruhe, Germany*

[3] *BMW Group, Forschungs- und Innovationszentrum, Akustik, Komfort und Werterhaltung, 80788 München, Germany*

Introduction

Integration of affective cues is considered a major factor towards future human-machine communication with respect to more naturalness. Today's in-car interfaces are already enhanced by speech recognition. In this contribution we therefore investigate the effects of acoustic in-car conditions on the recognition of emotion within speech, as it may be highly susceptible to noise [3]. Diverse driving situations and car-types and a public emotion database are used for tests. Results are presented for the general impact of noise and applying noise and speaker adaptation.

Interior Noise Masking

In contrast to a quiet test stand, emotion recognition in a real vehicle has to fight against hard conditions [2]: 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.

Compound of Interior Noise. Interior noise can be split up in four rough groups: First, wind noise generated by air turbulences at corners and edges of the vehicle, which arises equivalent to the velocity. Second, engine noise depending on load and number of revolutions. I.e. while accelerating (big load) the engine noise is more dominant than during the thrust. Third, noise caused by

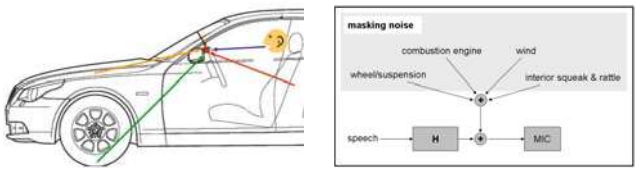

Figure 1: Speech and masking sound (left) and information flow (right).

wheels, driving, and suspension, basically influenced by road surface and tyre type. Thus a rough surface generates more wheel and suspension noise than a smooth road. Noise originated by wet roads or the windscreen banging of raindrops should not be discussed in the following. Finally, buzz, squeeks and rattles (BSR) basically caused by pounding or relative movement of interior components of a vehicle. Dependent on material matching, fixing and road stimulation BSR appears randomly and occurs with different peak levels (fig. 1).

Measuring Point. Herein, recognition is not accomplished in a real car but in an acoustic laboratory, allowing for reproducibility and variation of single parameters. However, the vehicle conditions must be emulated. According to existing speech recognition systems the microphone would be mounted in the middle of the instrument panel. Therefore, all masking noises for laboratory testing have been recorded exactly at the same point. The distance between mouth and microphone should be - as said - neglected. Masking noise is generated at very different places. In opposition to the mouth-to-microphone transfer function - due to large distances and room effects - the transfer function between noise source and microphone cannot be unvalued. Therefore, the microphone for noise recording and the standard microphone for speech recognition are exactly fixed at the same position. Thus, all transfer effects are included in the recording (fig. 1). So reality can be emulated adequately enough by superposing speech and the sum of recorded noise with the correct level.

Choice of Vehicles and Road Surfaces. Interior noise masking varies depending on vehicle class and derivatives. In order to cover a wide spectrum of car versions speech should be superposed by the interior noise of four very different vehicles as seen in tab. 1. In this

Table 1: Considered vehicles.

Vehicle	Derivative	Class
BMW 5 series	Touring	Executive car
BMW 6 series	Convertible	Executive car
BMW M5	Sedan	Exec. sports car
MINI Cooper	Convertible	Super-mini

vehicle choice, the influence and configuration of single noise sources differs (fig. 1). Although the soft top of both convertibles was closed during recording, interior noise is noticeably higher than in comparable sedans. There are big noise level differences between executive car and super-mini as well. While the engine noise dominates during the acceleration of the sportive M5, the similar constructed 5 series Touring is more gentle and comfortable. The worst case is represented by the MINI.

*Email:schuller@tum.de

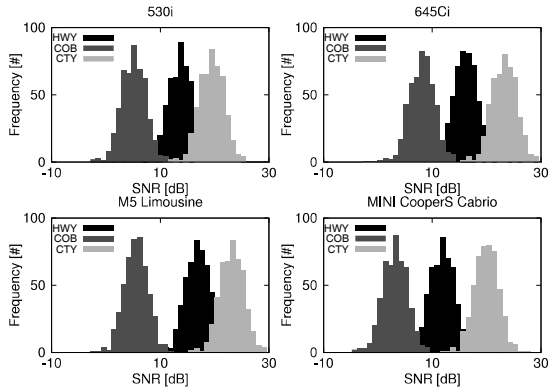

Figure 2: SNR distribution, database EMO-DB.

This super-mini unifies convertible, hard suspension and sportive engine. Just as the vehicle type, the road surface affects different interior noise. In order to get a good customer proximity, the sound while driving three different surfaces has been recorded as depicted in tab. 2.

Table 2: Considered road surfaces and velocities.

Surface	Velocity	Abbreviation
Big cobbles	30 km/h	COB
Smooth city road	50 km/h	CTY
Highway	120 km/h	HWY
Highway	accelerated	ACC

The lowest excitation provides a constant driving over a smooth city road at 50 km/h and medium revolution. Thus at this profile noise caused by wind, engine, wheel/suspension and BSR has its minimum. The subsequent higher excitation is measured at a highway drive (120 km/h). Wind noise is a multiple higher than for a drive at 50 km/h. The worst and loudest sound in the interior of a car provokes a road with big cobbles. Indeed, wind noise resides in lower levels, but the rough surface involves dominant wheel/suspension and BSR noise.

Database

To exclude other noise influences we decided for the popular studio recorded Berlin Emotional Speech Database ($EMO-DB$) [1], which covers the 'big six' emotion set (MPEG-4) besides boredom instead of surprise, and added neutrality. 10 (5f) professional actors speak 10 German emotionally undefined sentences. 494 phrases are marked as $\geq 60\%$ natural and $\geq 80\%$ assignable by 20 probands. 84.3% accuracy are reported for a human perception test. Noise is normalized to 125dB prior to addition to non-normalized speech. In the ACC 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. Fig. 2 shows the distribution of obtained SNRs.

Emotion Recognition and Adaptation

We use a set of 1.4k acoustic features basing on pitch, energy, duration, HNR, jitter, shimmer, formants 1-5 (amplitude, bandwidth, position), MFCC 1-16, and their derivatives. 19 functionals are applied to each contour covering extremes, ranges, positions, first four moments and quartiles. For classification we use Support Vector Machines (SVM) with linear Kernel and 1-vs.-1 multi-class discrimination. Two adaptation strategies are considered: First, noise adaptation (NA) by training within the noise and assuming matched conditions (e.g. based on velocity) throughout classification. Second, speaker adaptation (SA) by feature normalization with the whole individual speaker context. Results of a leave-one-speaker-out ($LOSO$) evaluation are presented in tab. 3. Note, that by feature optimization lower error rates can be obtained [3].

Discussion

As can be seen in tab. 3 noise adaptation clearly helps. Yet, an even stronger gain is obtained by speaker adaptation. Overall, other in-car noises as music or speaking persons can be expected as the main problem for speech based emotion recognition. Future works will investigate noise impact on feature groups.

Table 3: Speaker-independent error-rates for clean speech and diverse vehicles and road surfaces considering no ($-$), noise (NA), speaker (SA), and combined (NSA) adaptation. Classification by SVM, LOSO evaluation, database EMO-DB.

ERROR [%]		–	**NA**	**SA**	**NSA**
clean speech		**25.1**	–	**20.4**	–
530i	HWY	28.3	23.7	20.8	21.5
	COB	33.2	21.3	25.5	20.9
	CTY	27.1	24.9	21.1	24.9
645Ci	HWY	28.3	25.7	21.9	20.4
	COB	30.0	21.9	20.4	18.6
	CTY	26.3	23.9	20.4	18.4
M5	HWY	28.0	25.3	22.5	22.5
	COB	33.4	24.7	21.6	20.9
	CTY	26.7	25.5	21.6	21.9
	ACC	38.5	25.3	25.7	23.1
MINI	HWY	31.6	21.1	21.9	20.9
	COB	33.0	23.1	22.1	21.3
	CTY	25.5	24.5	22.7	22.1
mean		**30.0**	**23.9**	**22.1**	**21.3**

References

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

[2] M. Grimm, K. Kroschel, B. Schuller, G. Rigoll, T. Moosmayr: Acoustic Emotion Recognition in Car Environment Using a 3D Emotion Space Approach. *Proc. DAGA 2007*, Stuttgart, Germany, 2007.

[3] B. Schuller, D. Seppi, A. Batliner, A. Maier, S. Steidl: Towards More Reality in the Recognition of Emotional Speech. *Proc. ICASSP 2007*, Honolulu, 2007.