interspeech 2008 incorporating SST 08

Detection of Security Related Affect and Behaviour in Passenger Transport

Björn Schuller[1], Matthias Wimmer[2], Dejan Arsic[1], Tobias Moosmayr[3], Gerhard Rigoll[1]

[1]Institute for Human-Machine Communication, Technische Universität München, München, Germany
[2]Perceptual Computing Lab, Waseda University, Tokyo, Japan
[3]BMW Group, Forschungs- und Innovationszentrum, München, Germany
schuller@tum.de

10.21437/Interspeech.2008-94

Abstract

Surveillance of drivers, pilots or passengers possesses significant potential for increased security within passenger transport. In an automotive setting the interaction can e.g. be improved by social awareness of an MMI. As further example security marshals can be efficiently positioned guided by according systems. Within this scope the detection of security relevant behavior patterns as aggressiveness or stress 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 for the synchronization and data reduction of audio and video by brute-force feature generation. By combined optimization of the large audiovisual space accuracy is boosted. Extensive results are reported on aviation behavior, as well as in particular for the audio channel on numerous standard corpora. The influence of noise will be discussed by representative car-noise overlay.

Index Terms: Audiovisual Emotion Recognition, Affective Computing, Security-critical Systems, Transport Surveillance, Automotive Environment

1. Introduction

In many security critical environments emotion or mental states such as tiredness, intoxication or high stress level of users of technical systems can play a life decisive role. Examples of such settings comprise e.g. operators of (nuclear) power plants, surgeons in intelligent operating rooms, or tank or fighter aircraft pilots in a tactical mission. But also in everyday's driving, affect may become security relevant. In [5], three main emotions of interest are named: happiness as positive influence as opposed to anger easily leading to road rage, and sadness leading to inattention. However, clearly also the aforementioned states may highly negatively influence attention and reaction times. Possible strategies if such states are recognized include socially aware accommodation in the system behavior proven to be beneficial [5], or activation of fail-safe strategies as system emergency take-over in extreme cases. This however demands for reliable recognition of such dangerous behavior and negative emotion in the first place. Following the terrorist attacks on 9/11 growing interest in such technology to help prevent follow-up attacks can also be observed: automatic camera-based surveillance of airplanes is one good example of such advances [1]. However, also surveillance of public ground transportation as in local trains may help fight vandalism or person assaults. Granting passenger's safety is thereby an expensive task, as additionally to closed circuit television systems a large staff is required to analyze video streams online. Therefore it seems desirable to automatically monitor passenger compartments in trains, aircrafts and buses to automatically detect passengers which might be a threat to others or themselves and effectively position security staff units.

This paper therefore introduces an early fusion approach to emotion and behavior recognition by audio and video sources. These are well suited modalities for the aimed at surveillance application: speech technology is already present in upper-class limousines, and can be easily installed in public transport systems. Cameras can accordingly easily be mounted in a car or in seat's back rests in front of passengers to capture facial expression. We will first focus on the audio-part, as it is known to provide more reliable results at the time. Thereby we provide an impression of obtainable results on popular and application-specific databases: SUSAS, EMO-DB, eNTERFACE, and ABC. The challenging requirements of our use-cases will be respected in particular: subject-independency, coping with typical noise conditions as in automotive environment, planes, and public transportation, and real-time capability. Furthermore we investigate the benefit derived from additional vision information exploitation.

2. Audiovisual Behavior Detection

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. 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 [2]. This is realized 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.

To extract video features, a mask is applied to a face on an image, consisting of a set of 133 points (fig. 1, left). This face mask model is fully described by position, rotation and deformation, as first introduced in [4], and called Point

Distribution Model. Its parameters hold the translation describing the position of the mask, the scaling factor, and the rotation being sufficient for placing the face mask on a 2-dimensional image. Furthermore, the model is described by a second component, the deformation vector, extracted from the 133 points of the face mask by applying a Principal Component Analysis (PCA) and taking only the 17 best Principal Axes (PA), which already cover 95% of the deformation information of the mask. Fig. 1, right, shows 3 examples of the 17 top-ranked PA.

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

For finding the optimal parameters of the face mask, a model fitting approach is chosen. Further an image dependent classifier was created for each image sequence. For more detailed explanation, we refer to [8]. Based on the skin color image, the edges of the observed face can be found and used for the fitting algorithm optimizing the objective function. The result is the fitting parameterization of the model. Apart from this algorithm, methods for e.g. eye detection are used for error reduction. The fitting algorithm works in real time, and is based on a greedy algorithm, applied on the search-space for finding the best matching model parameters.

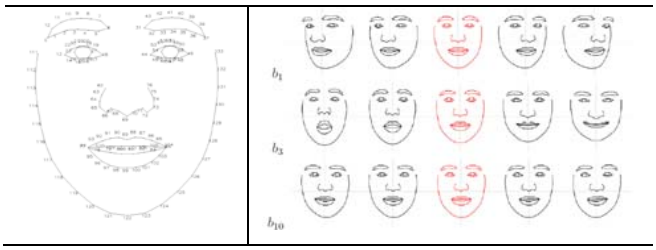

Figure 1. *Face mask model (left) and deformations of the face mask when changing one PA of the face mask b1-b17 (right). Parameter b1 changes the rotation of the head, b3 the opening of the mouth, b10 moves the pupils.*

As the goal is to classify emotions in a video sequence, the dynamic change of face parameters has to be considered. This leads to the usage of optical flow as a more reasonable approach for emotion classification. The optical flow is a measure of the apparent visual motion in two successive video frames. It can be visualized with a vector-field showing motion vectors for a grid of pixels. Fig. 2 shows an example for such a vector field overlaying its source image. It is based on the assumption that if a pixel (x,y) is available in the source image at time t, it is also available in the target image at time t+1 displaced by (δx,δy) - the optical flow of this point. The metric for the similarity of two points is the comparison of its brightness values, assuming that the brightness of the source pixel and the displaced pixel do not change. This metric is found by Taylor series expansion and ignorance of higher order terms of the Taylor series under the assumption that the motion is very small. This however leads to non predetermined variables, and as a consequence to the not exactly solvable aperture problem: there is not enough information to uniquely determine motion. However, it is possible to estimate the optical flow with iterative approaches, as the Lucas and Kanade algorithm, used herein. Its idea is that the optical flow of a pixel and pixels in its neighborhood is approximately the same. For this algorithm a small neighborhood of pixels is chosen, here a grid of 5x5 pixels. As a consequence, the optical flow can never be determined exactly, and does not always map to the real motion. To extract the optical flow within a face, the already fitted face mask can be used to define the boundary of the area of interest. In detail, the measuring points of the optical flow are equally distributed in the area bounded by the contour of the face mask. The resulting raster of measuring points is 10 points wide and 14 points high as shown in fig. 2 (left). Fig. 2 (right) shows two examples of the optical flow at two different points in time.

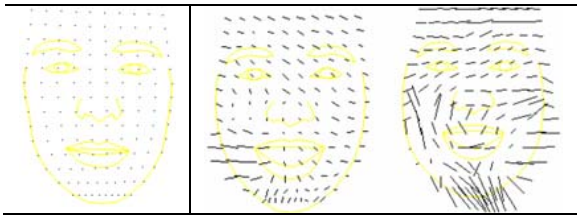

Figure 2. *Face mask with raster (left), applied optical flow (right).*

Next, the multivariate time-series analysis approach introduced is taken over for video feature space construc-tion. This allows for early integration and simple synchronization of audio and video features in one super-vector. As a bonus, a combined audiovisual feature space optimization can be carried out. Table 2 shows applied video feature LLD and functionals in detail. The total dimension for video results in 4 x 297 = 1188 features.

After systematic feature generation as introduced, selection of the most relevant ones is important to save computation time considering real-time processing, and boost performance as most classifiers are susceptible to high dimensionality. We choose Sequential Forward Floating Search (SFFS) employing the classifier's error as optimization criterion for audio and video feature selection, herein. The optimal number of features is determined in accordance to the highest observed accuracy throughout selection. For classification we use Support Vector Machines (SVM) with linear Kernel and pair-wise multi-class discrimination [2].

Table 2. *Video Low-Level-Descriptors and functionals.*

Low-Level-Descriptors (297)	**Functionals (4)**
Face Mask Principal Axis 1-17	Mean
Optical Flow Angle 1-140	Std. Dev.
Optical Flow Abs. Val. 1-140	Max., Min. Value

3. Behavior and Emotion Databases

Only two sets, namely SUSAS and ABC, are originally intended for our use-case. However, the further evaluated EMO-DB and eNTERFACE databases are popular sets, and well suited for the later described controlled noise analysis. First, we selected the Speech Under Simulated and Actual Stress (SUSAS) audio only database [6] as a reference for spontaneous recordings in field noise. We decided for the 3,663 actual stress speech samples recorded in subject motion fear and stress tasks in a helicopter environment. 7 speakers, 3 of them female, in free fall actual stress situations are contained. Two different stress conditions have been collected: medium and high stress. Further samples cover neutrality, fear, and screaming as classes. Likewise a total of five emotions, respectively speaking styles, are covered. SUSAS samples are constrained to a 35 words vocabulary of short aircraft communication commands. All files are sampled in 8 kHz, 16 bit. The recordings are partly overlaid with heavy noise and background over-talk. However, this resembles realistic acoustic recording conditions, as also given in our scenarios of interest such as automotive speech interfaces or public transport surveillance.

Next, the Berlin Emotional Speech Database (EMO-DB) [3] 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, as opposed to SUSAS. 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 database was annotated by 20 subjects with respect to naturalness and assignability. The final data-set with >60% naturalness and >80% assignability to an emotion consists of 494 samples.

The eNTERFACE corpus is a further public, yet audio-visual emotion database [7]. 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.

As public audiovisual emotion data apart from the named eNTERFACE corpus is sparse, we decided to use a database which is crafted for our special target application of public transport surveillance: the Airplane Behavior Corpus (ABC) [8]. 8 subjects in gender-balance from 25 to 48 years (mean 32 years) took part in the recording by mood induction. The language throughout recording is German. A total of 11.5h video was recorded and annotated independently after pre-segmentation by three experienced male labelers within a closed set: aggressive, cheerful, intoxicated, nervous, neutral, and tired behavior. The average length of the 396 inter-labeler-agreed samples in total is 8.4s.

To study the feasibility of recognition in the car as representative noise environment, several noise scenarios were recorded [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 (ACC, only for M5). Eventually, a total of 13 car-noise scenarios are simulated. Every noise scenario takes approx. 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 of Munich, Germany, with the same microphone, and takes approximately one hour. The SNR distributions for each noise scenario are shown in fig. 3.

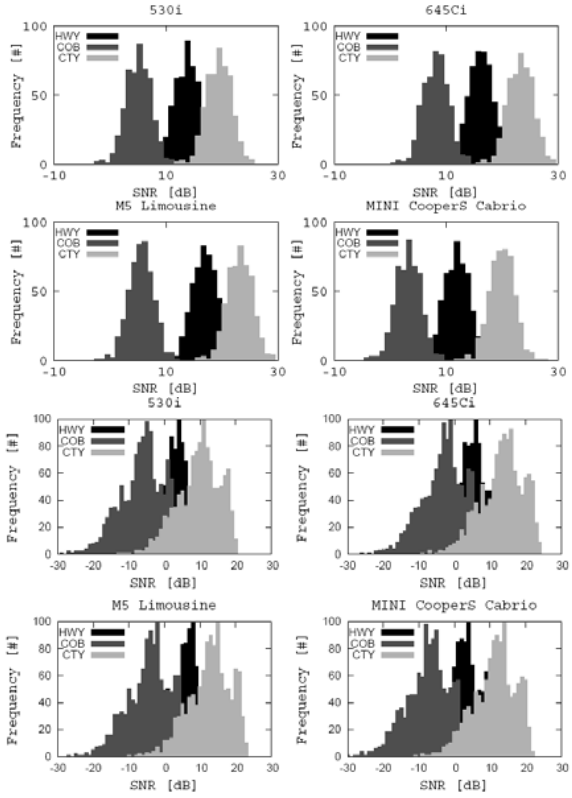

Figure 3. *SNR distribution for the databases EMO-DB (top) and eNTERFACE (bottom) for diverse car and driving situations.*

4. Experimental Results

As surveillance usually demands for subject independent recognition, we provide results in a Leave-One-Subject-Out (LOSO) manner. First, we consider audio exclusively. Thereby we show results for clean speech and speech in the diverse noise situations, as described in sec. 3.2. Table 3 shows observed accuracies for SUSAS, EMO-DB, and eNTERFACE. Note that car noise is summarized 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. As SUSAS is originally recorded in the noise (helicopter), no additional noise overlay is considered.

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 realized by speed indicator). Second, speaker adaptation (SA) by mean and standard deviation normalization 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 (NSA+FS). As a result, accuracies can be step-wisely "repaired" by combination of methods, even in the worst case noise scenario. SUSAS thereby clearly demonstrates the problem of speaker independency: if evaluated in a 10-fold stratified cross-validation for each speaker, 83.8% accuracy are observed.

Next, we show results on the ABC database for audio-visual processing in table 4 to focus on the advantages of combination of audio and video for robust behavior detection. Accuracies are shown for audio and video, individually, and combined. Thereby the gain of feature selection (FS) is also shown for each modality.

Table 3. *Accuracies for speaker-independent audio-based emotion recognition in diverse noise situations with diverse compensation strategies, as described.*

Accuracy [%]	**-**	**NA**	**SA**	**NSA**	**NSA+FS**
SUSAS					
Noise	50.4	50.4	49.1	49.1	**51.8**
EMO-DB					
Clean Speech	74.9	-	79.6	-	**80.4**
Car Noise	60.5	72.1	75.1	76.3	**77.3**
Babble Noise	70.0	76.1	77.9	78.7	**80.5**
Babble+MINI	46.6	70.4	75.7	76.1	**79.5**
eNTERFACE					
Clean Speech	54.2	-	61.4	-	**62.8**
Car Noise	38.5	48.3	51.8	56.7	**59.7**
Babble Noise	42.1	53.2	54.2	61.0	**61.6**
Babble+MINI	30.6	49.8	46.2	55.8	**58.6**

In the case of audiovisual processing two methods can be considered: indivdual (iFS), and combined (cFS) feature selection. This reveals the effectiveness of the introduced multivariate time-series-analysis by functionals: not only is the accuracy further boosted, but the audiovisual feature space size could be further compressed from 248 features in the case of iFS to 200 features in the case of cFS, thus also forwarding real-time capability.

Table 4. *Accuracies for speaker-independent behaviour recognition with diverse modalities and optimization strategies, as described. Database ABC, SVM.*

Acc. [%]	**Audio**		**Video**		**Audiovisual**		
Selection	-	FS	-	FS	-	iFS	cFS
ABC							
Noise	69.4	73.7	51.8	61.1	71.2	77.3	81.8

5. Conclusion

In this paper we introduced audiovisual recognition of emotion and behavior for surveillance in public transport systems. 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. Addition of video information and combined audiovisual feature space optimization could further be shown highly effective. As an aid to assist human security staff, such systems may already be of help in the near future. Further efforts will have to provide realistic data to allow for more accurate insights and model constructions. And finally, it will be left to decide where such oncoming automatic surveillance is appropriate, and whether it is more convenient than human observation.

6. References

[1] Arsic, D; Schuller, B.; Rigoll, G.: Suspicious Behavior Detection In Public Transport by Fusion of Low-Level Video Descriptors, *Proc. ICME 2007*, IEEE, Beijing, China, 2007.

[2] 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.

[3] 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.

[4] Cootes, T. F.; Taylor, C. J.: Active shape models – smart snakes. *Proc. of the 3rd British Machine Vision Conference 1992*, pp. 266 – 275. Springer, 1992.

[5] 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.

[6] Hansen, J.H.L.; Bou-Ghazale, S.: Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, *Proc. EUROSPEECH-97*, Rhodes, Greece, Vol. 4, pp. 1743-1746, 1997.

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

[8] Schuller, B.; Wimmer, M.; Arsic, D.; Rigoll, G.; Radig, B.: Audiovisual Behavior Modeling by Com-bined Feature Spaces, *Proc. ICASSP 2007*, Honolulu, Hawaii, USA, 15.-20.04.2007. IEEE., Vol. II, pp. 733-736.