Driver Frustration Detection from Audio and Video in the Wild

Irman Abdić[1,2] Lex Fridman[1], Daniel McDuff[1], Erik Marchi[2], Bryan Reimer[1], and Björn Schuller[3]

[1] Massachusetts Institute of Technology (MIT), Cambridge, USA
abdic@mit.edu
[2] Technische Universität München (TUM), Munich, Germany
[3] Imperial College London (ICL), London, UK

1 Introduction

The question of how to design an interface in order to maximize driver safety has been extensively studied over the past two decades [13]. Numerous publications seek to aid designers in the creation of in-vehicle interfaces that limit demands placed upon the driver [10]. As such, these efforts aim to improve the likelihood of driver's to multi-task safely. Evaluation questions usually take the form of "Is HCI system A better than HCI system B, and why?". Rarely do applied evaluations of vehicle systems consider the emotional state of the driver as a component of demand that is quantified during system prove out, despite of numerous studies that show the importance of affect and emotions in hedonics and aestetics to improve user experience [8]. The work in this paper is motivated by a vision for

(a) Class 1: Satisfied with Voice-Based Interaction

(b) Class 2: Frustrated with Voice-Based Interaction

Fig. 1. Representative video snapshots from voice navigation interface interaction for two subjects. The subject (a) self-reported as not frustrated (satisfied) with the interaction and the (b) subject self-reported as frustrated (frustrated).

an adaptive system that is able to detect the emotional response of the driver and adapt, in order to aid driving performance. The critical component of this vision is the detection of emotion in the interaction of the human driver with the driver vehicle interface (DVI) system. We consider the binary classification problem of a "frustrated" driver versus a "satisfied" driver annotated based on a self-reported answer to the following question: "To what extent did you feel frustrated using the car voice navigation interface?" The answers were on a scale of 1 to 10 and naturally clustered into two partitions as discussed in Sect. 2. As presented in Fig. 1, the "satisfied" interaction is relatively emotionless, and the "frustrated" interaction is full of affective facial actions.

The task of detecting drivers' frustration has been researched in the past [1]. Boril *et al.* exploited the audio stream of the drivers' speech and discriminated "neutral" and "negative" emotions with 81.3 % accuracy (measured in Equal Accuracy Rate – EAR) across 68 subjects. This work used SVMs to discriminate between classes. The ground truth came from one annotation sequence. A "humored" state was presented as one of the 5 "neutral" (non-negative) emotions. This partitioning of emotion contradicts our findings that smiling and humor are often part of the response by frustrated subject.

Contributions. We extend this prior work by (a) leveraging audiovisual data collected under real driving conditions, (b) using self-reported rating of the frustration for data annotation, (c) fusing audio and video as complimentary data sources, and (d) fusing audio and video streams across time in order to characterize the trade-off between decision time and classification accuracy. We believe that this work is the first to address the task of detecting self-reported frustration under real driving conditions.

2 Dataset for Detecting Frustration

The dataset used for frustration detection was collected as part of a study for multi-modal assessment of on-road demand of voice and manual phone calling

and voice navigation entry across two embedded vehicle systems [9]. Participants drove one of two standard production vehicles, a 2013 Chevrolet Equinox (Chevy) equipped with the MyLink system and a 2013 Volvo XC60 (Volvo) equipped with the Sensus system.

For the frustration detection task we selected 20 subjects from the initial dataset of 80 such that our selection spanned both vehicles, gender (male, female) and four age groups (18–24, 25–39, 40–54, 55 and older). This pruning step was made for two reasons. First, a significant amount of videos had poor lighting conditions where extraction of facial expressions was not possible or was very difficult. To address this issue, we discarded subjects where less than 80 % of video frames contained a successfully detected face. We applied the face detector described in [4] that uses a Histogram of Oriented Gradients (HOG) combined with a linear SVM classifier, an image pyramid, and a sliding window detection scheme. Second, a substantially higher proportion of subjects self-reported low frustration level (class "satisfied"), thus we had to select our subjects viligantly to keep the dataset balanced and have both classes represented equally.

It is important to note that all subjects drove the same route and all tasks were performed while driving. For this paper, we focused in on the navigation task. After each task, subjects completed a short written survey in which they self-reported the workload and rated an accomplished task, including their frustration level on a scale from 1 to 10, with 1 being "not at all" and 10 "very". The question that the subjects were asked to answer is as follows: "To what extent did you feel frustrated using the car voice navigation system?". We found that the navigation system task had a clustering of responses for self-reported frustration that naturally fell into two obvious classes, after removing the minority of "neutral" responses with self-reported frustration level from 4 to 6. The "frustrated" class contained all subjects with self-reported frustration level between 7 and 9, and "satisfied" class contained all subjects with self-reported frustration level from 1 to 3. There are two different types of epochs: (1) audio epochs, where subjects are dictating commands to the machine, and (2) video epochs, where subjects are listening to a response from the machine and signaling frustration through various facial movements.

3 Methods

3.1 Audio Features

In contrast to large scale brute-force feature sets [11], a smaller, expert-knowledge based feature set has been applied. In fact, a minimalistic standard parameter set reduces the risk of over-fitting in the training phase as compared to brute-forced large features sets, which in our task is of great interest. Recently, a recommended minimalistic standard parameter set for the acoustic analysis of speaker states and traits has been proposed in [2]. The proposed feature set is the so-called Geneva Minimalistic Acoustic Parameter Set (GeMAPS). Features were mainly selected based on their potential to index affective physiological changes in voice production, for their proven value in former studies, and for

their theoretical definition. Acoustic low-level descriptors (LLD) were automatically extracted from the speech waveform on a per-chunk level by using the open-source openSMILE feature extractor in its 2.1 release [3].

3.2 Video Features

We used automated facial coding software to extract features from the videos. The software (Affdex - Affectiva, Inc.) has three main components. First, the face is detected using the Viola-Jones method [14] (OpenCV implementation). Thirty-four facial landmarks are then detected using a supervised descent based landmark detector and an image region of interest (ROI) is segmented. The ROI includes the eyes, eyebrows, nose and mouth. The region of interest is normalized using rotation and scaling to 96×96 pixels. Second, histogram of oriented gradient (HOG) features are extracted from the ROI within each frame. Third, support vector machine classifiers are used to detect the presence of each facial action. Details of how the classifiers were trained and validated can be found in [12]. The facial action classifiers return a confidence score from 0 to 100. The software provided scores for 14 facial actions. In addition to facial actions we used the three axes of head pose and position of the face (left and right eye corners and center of top lip) as observations from which to extract features. For each epoch the mean, standard deviation, minimum and maximum values for each action, head pose and position metric were calculated to give 60 video features ((14 actions + 3 head pose angles + 3 landmark positions)*4).

3.3 Classifier

We used a Weka 3 implementation of Support Vector Machines (SVMs) with the Sequential Minimal Optimization (SMO), and audio and video features described in Sect. 3 [5]. We describe a set of SMO complexity parameters as:

$$C \in \{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, ..., 1\}. \quad (1)$$

For each SMO complexity parameter C from (1) we upsampled the feature vectors (one per epoch) from the original datasets to balance the number of epochs per class by calculating the upsampling factors. An average upsampling factor across four folds is 1.03 for the "frustrated" class and 1.24 for the "satisfied" class. We kept the original datasets, and produced an additional upsampled dataset for further experiments. We then (a) normalized and (b) standardized both upsampled and original datasets for each SMO complexity parameter C, and obtained 36 different configurations per fold. We carried out 144 experiments across four folds, computed accuracy, and selected the configuration that gave us the best average result. The term "accuracy" stands for Unweighted Average Recall (UAR).

4 Results

We used features and a classifier as described in Sect. 3 and achieved an accuracy of 77.4 % for "audio" epochs and 81.2 % for "video" epochs as presented in Table 1. The *epoch type* column indicates whether the human or the machine are speaking and *data source* indicates the source of the signal which is being used for extracting features. The presented results are the average accuracy for the subject-independent cross-validation over four folds.

Table 1. Results for predicting frustration from a single epoch of audio and video.

Epoch type	Data source	C	Acc. (%)
Machine speaking	Video	$1e^{-3}$	81.2
Human speaking	Audio	$5e^{-3}$	77.4

In order to characterize the tradeoff between classification accuracy and the duration of the interaction, we fused the predictions from consecutive epochs for both video and audio using a majority vote fusion rule [7]. The interaction of the driver with the voice-based system is a sequence of mostly-alternating epochs of face video data and voice data. In presenting the results, we consider two measures of duration: (1) d_e is the duration in the number epochs and (2) d_s is the duration in the number of seconds. Both measures are important for the evaluation of systems performance, since classifier decisions are made once per epoch (as measured by d_e) but the driver experiences the interaction in real-time (as measured by d_s). The fused results for up to 17 epochs are presented in Fig. 2 where duration d_e is used. The average accuracy is shown with the red line and the accuracy for each of the four folds is shown with the gray line. The average accuracy does not monotonically increase with the number of predictions fused. Instead, it slightly fluctuates due to a broad variation in complexity of the underlying subtasks. An average accuracy of 88.5 % is achieved for an interaction that lasts approximately 1 min but a lower average accuracy of 82.8 % is achieved for an interaction that lasts approximately 2 minutes. Evaluation over one of the folds in Fig. 2 achieves 100 % accuracy after 9 epochs. This is possible due to the fact that the number of epochs for total interaction varies between subjects, and the reported accuracy for a specific duration d_e is averaged over only the interactions that last at least that long. It follows that with the longer durations d_e (x-axis), the number of subjects over which the accuracy is averaged decreases and the variance of the accuracy increases.

We used a Weka implementation of the Information Gain (IG) feature evaluation to rank video features [6]. Then, we grouped features into the feature categories by summing corresponding category IG ranking values for mean, maximum, minimum and standard deviation. Each feature category represents one action, *i. e.,* inner brow rise, nose wrinkle or lip depressor. The 5 best discriminating feature categories are: (1) horizontal location of the left eye corner,

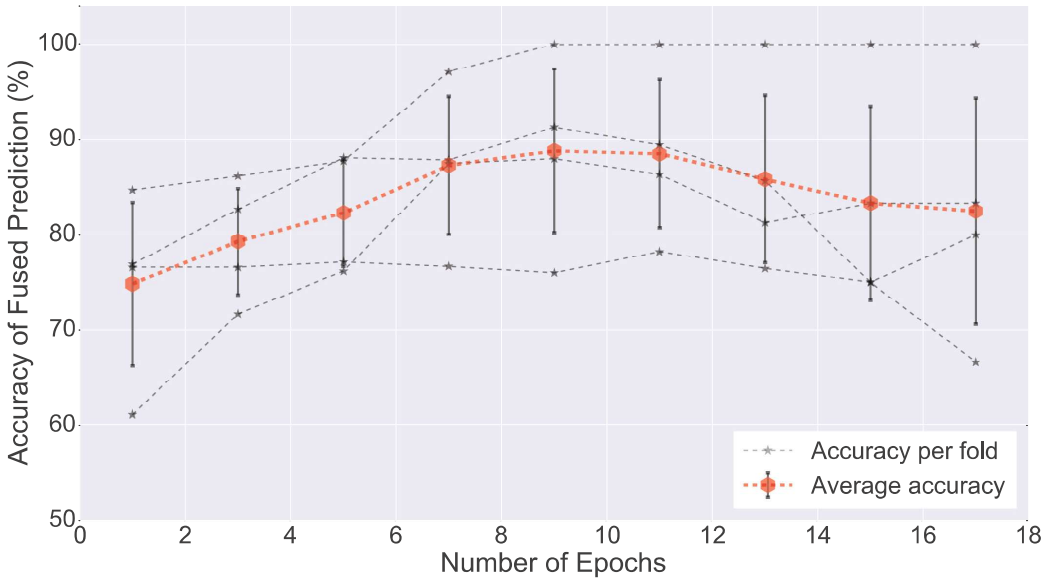

Fig. 2. Trade-off between fused prediction accuracy and the number of epochs per interaction (d_e).

(2) horizontal location of the top of the mouth, (3) horizontal location of the right eye corner, (4) the angle of head tilt (i.e. rotation of the head about an axis that passes from the back of the head to the front of the head), and (5) smile confidence (on a scale of 0–100). We ranked only video features to select the most interesting epochs for our presentation video: http://lexfridman.com/driverfrustration.

5 Conclusion

We presented a method for detecting driver frustration from 615 video epochs and 596 audio epochs captured during the driver's interaction with an in-vehicle voice-based navigation system. The data was captured in a natural driving context. Our method has been evaluated across 20 subjects that span over different demographic parameters and both cars that were used in our study. This method resulted in an accuracy of 81.2 % for detecting driver frustration from the video stream and 77.4 % from the audio stream. We then treated the video and audio streams as a sequence of interactions and achieved 88.5 % accuracy after 9 epochs by using decision fusion. Future work will include additional data streams (*i. e.*, heart rate, skin conductance) and affective annotation methods to augment the self-reported frustration measure.

Acknowledgments. Support for this work was provided by the New England University Transportation Center, and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs class counsel. Data was drawn from studies supported by the Insurance Institute for Highway Safety (IIHS) and Affectiva.

References

1. Boril, H., Sadjadi, S.O., Kleinschmidt, T., Hansen, J.H.L.: Analysis and detection of cognitive load and frustration in drivers' speech. In: Proceedings of INTERSPEECH 2010, pp. 502–505 (2010)
2. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K.: The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Trans. Affect. Comput. (2015)
3. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in openSMILE, the munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, Barcelona, Spain, October 2013, pp. 835–838. ACM
4. Fridman, L., Lee, J., Reimer, B., Victor, T.: Owl and lizard: patterns of head pose and eye pose in driver gaze classification. IET Comput. Vis. (2016, in Press)
5. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
6. Karegowda, A.G., Manjunath, A.S., Jayaram, M.A.: Comparative study of attribute selection using gain ratio, correlation based feature selection. Int. J. Inf. Technol. Knowl. Manage. **2**(2), 271–277 (2010)
7. Kuncheva, L.I.: A theoretical study on six classifier fusion strategies. IEEE Trans. Pattern Anal. Mach. Intell. **2**, 281–286 (2002)
8. Mahlke, S.: Understanding users' experience of interaction. In: Proceedings of the 2005 Annual Conference on European Association of Cognitive Ergonomics, pp. 251–254. University of Athens (2005)
9. Mehler, B., Kidd, D., Reimer, B., Reagan, I., Dobres, J., McCartt, A.: Multi-modal assessment of on-road demand of voice and manual phone calling and voice navigation entry across two embedded vehicle systems. Ergonomics, pp. 1–24 (2015). PMID: 26269281
10. NHTSA: Visual-manual nhtsa driver distraction guidelines for in-vehicle electronic devices (docket no. nhtsa-2010-0053). Washington, DC: US Department of Transportation National Highway Traffic Safety Administration (NHTSA) (2013)
11. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., Kim, S.: The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In: Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 2013. ISCA (2013) 5 pages
12. Senechal, T., McDuff, D., Kaliouby, R.: Facial action unit detection using active learning and an efficient non-linear kernel approximation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 10–18 (2015)
13. Stevens, A., Quimby, A., Board, A., Kersloot, T., Burns, P.: Design guidelines for safety of in-vehicle information systems. TRL Limited (2002)
14. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)