Analysis of real-world driver's frustration

Lucas Malta, Chiyomi Miyajima, Norihide Kitaoka, and Kazuya Takeda

***Abstract*—This study investigates a method for estimating a driver's spontaneous frustration in the real world. In line with a specific definition of emotion, the proposed method integrates information about the environment, the driver's emotional state, and the driver's responses in a single model. Driving data are recorded using an instrumented vehicle on which multiple sensors are mounted. While driving, drivers also interact with an automatic speech recognition (ASR) system to retrieve and play music. Using a Bayesian network, we combine knowledge on the driving environment, assessed through data annotation, speech recognition errors, driver's emotional state (frustration), and driver's responses measured through facial expressions, physiological condition, and gas- and brake-pedal actuation. Experiments are performed with data from 20 drivers. We discuss the relevance of the proposed model and features of frustration estimation. When all of the available information is used, the overall estimation achieves a true positive rate of 80% and a false positive rate of 9% (i.e., the system correctly estimates 80% of the frustration and, when drivers are not frustrated, makes mistakes 9% of the time).**

***Index Terms*—**

I. Introduction

A. The Role of Emotions in Traffic

In efforts to improve overall safety and comfort, a number of vehicular technologies have been developed and deployed in the market over the last few decades. Nowadays, it is possible to buy an automobile with such systems as pedestrian detection [1], cruise control [2], and collision mitigation brake system [3], which predicts rear-end collisions and assists brake operation to reduce the impact to occupants. A major drawback of available vehicular systems, however, is that they do not include the driver in the loop of decision-making processes. For example, even if the driver is heavily cognitively loaded or distracted, the decision threshold of safety systems as well as the human-computer interface's information exchange protocol remain the same. In the interest of implementing necessary changes, a number of studies have been conducted focusing on the internal state (physical and emotional) of a driver [4], [5].

Car driving is a complex cognitive process in which even a small disruption of attention can have disastrous consequences. Emotion is a key factor that is likely to affect cognitive functioning and therefore to increase the demand on drivers [6]. According to Lazarus [7], emotions can be considered a process that promotes adaptation to the environment and prepares the person for adaptive action. Emotions are usually accompanied by an altered physiological state, such as increased heart rate, and by behavioral changes, such as in voice, face or gestures. Consistent with current approaches, emotions are caused by a person evaluating an event or encounter based on his/her personal goals. This notion that emotional behavior results from an interaction with an event implies that, in the driving context, emotions may arise frequently.

Studies on the emotional state of a driver have been conducted from different perspectives. From a traffic-psychology viewpoint, researchers have focused on the determinants and consequences of emotions in traffic [6]. In addition, automatic detection of the emotional state of a driver has been addressed using physiological signals [8], [9], speech [10], [11], and both visual and acoustic cues [12].

Although most automatic detection methods focus on the six basic emotions (i.e., happiness, sadness, fear, anger, surprise, and disgust [13]) in the driving context, frustration plays a unique role. Frustration, which is defined as an interference with goal-directed behavior, generates aggressive inclinations toward another person or object primarily perceived as its cause, as stated in the classical frustration-aggression hypothesis [14]. When applied to the driving context, the frustration-aggression model implies that the driver's goal is to achieve mobility with minimum interruption (and possibly some pleasure). Negative emotions caused by impeded progress can then escalate in sequence from frustration to hostility to hatred. The situational sources of frustration are the same that cause congestion: red-light signals, slow moving vehicles, blocked path of travel by other cars or pedestrians, and so on. One of the most relevant studies recently published in this field was conducted by Shinar [15], which through a series of experiments examined the effect of frustrating environmental factors on driver's aggression.

As the number of in-vechicle devices increases, the need for intelligent interfaces also calls attention to the importance of frustration in driving. Together with interest, puzzlement, and boredom, frustration plays an important role in human-computer interaction, and recently it has been considered by many research works in this field [16]. Kapoor *et al.* [17] proposed a method for automatic prediction of frustration of 24 middle-school students who interacted with a learning companion. An overall accuracy of 79% was reported. Potential disadvantages of systems devoted to the automatic detection of frustration and other emotions in general include: (1) conduction of experiments under controlled environments (lack of generality); (2) use of acted or carefully elicited data rather than spontaneous emotions; and (3) disregard for the context in which emotions were elicited, as will be explained in the following section.

B. Proposed Approach

In this study, we propose a method for estimating a driver's frustration that integrates features of a different nature. The

L. Malta, C. Miyajima, N. Kitaoka, and K. Takeda are with the Grad. School of Information Science, Nagoya University, Japan. Furo-cho, Chikusa-ku, Nagoya, 464-8603. This research was supported by Toyota Motor Corporation and the Strategic Information and Communications R&D Promotion Program of MIC Japan.

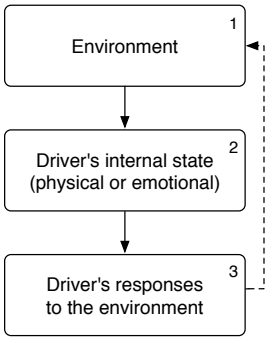

Fig. 1. Simple model of driver-environment interaction.

designed model is based on the assumption that emotions are the result of an interaction with the environment and are usually accompanied by physiological changes, facial expressions or actions, as described in the schematics of Fig. 1. Methods designed for the estimation of the internal state of a driver tend to oversimplify this model by disregarding one of its boxes, usually box 1. Ignoring box 1 may lead to inconsistent internal state inference based solely on responses measured through, for example, facial expressions, speech or physiological signals [8], [17]. Responses in this case are erroneously assumed to be absolute and independent of the environment. For example, a smile can be a reaction to different emotional states—even of different valences—depending on the context.

Emotion recognition could greatly benefit from a more comprehensive strategy that takes into account a person's emotional state, his/her responses, and the environment in a single model. In the present study, we investigate such a model while focusing on the following questions: (1) How can knowledge of different types be acquired and meaningfully represented considering the driving context? (2) How can this knowledge be combined efficiently? (3) What are the most relevant cues for estimating spontaneous frustration in real driving?

Once frustration is detected it can be used as a feedback to in-vehicle systems, allowing them to adapt accordingly. Possible targets for this adaptation are the thresholds of safety systems that can be made more sensitive, type of music playing, color of the dashboard, and both speaking style and timing of ASR systems. The same information can be presented to the drivers as well. When drivers are aware of their own emotional state it becomes easier for them to adopt optimal strategies to cope with it in a safe manner. In addition, a mean level of frustration experienced by fleet drivers over a long period could be used by fleet owners. Targeting and treating employees who might need a psychological intervention could help in saving costs, providing, at the same time, more pleasant working conditions for drivers.

II. Materials and Methods

A. Real-World Driving Database Annotation

Our first attempt to collect data on spontaneous frustration in real driving was a manual annotation of a large database of 492 drivers, recorded from 2000 to 2002 [18]. The Center for Integrated Acoustic Information Research (CIAIR) database is composed of image, driving behavior, and location signals recorded synchronously with speech. Each driver drove for about 15 minutes on city streets in Nagoya, Japan, and interacted with a human operator, an automatic speech recognition (ASR) system, and a Wizard of Oz system.

Six graduate students, who served as annotators, preselected 259 possible frustration scenes from the database using audio and video footage. The pre-selected scenes were then analyzed by professional annotators who found concrete evidence of frustration in only 16 of them. Although the final number of scenes was insufficient for training a classifier, since most of them were only a few seconds long, two important lessons were learned: (1) The ASR system, which is particularly error-prone in a noisy traffic environment and relies on a synthesized voice that is sometimes difficult to understand, was a frequent cause of discomfort to drivers; (2) Except for a few situations when the driver was highly frustrated, manual annotation of spontaneous frustration in real traffic was extremely difficult, due to its high ambiguity, individuality, and context-dependency.

The lack of physiological information on drivers was another drawback of the CIAIR database that led to our decision to collect new multimodal data in an environment where drivers would have a greater tendency to get frustrated. Moreover, in order to avoid having to again annotate frustration, we decided to rely on a self-reported assessment.

B. Data Collection Vehicle

A data collection vehicle was designed for synchronously recording audio with other multimedia data. Various external sensors were mounted on a Toyota Hybrid Estima with 2,360-cc displacement and automatic transmission. Figure 2 shows the data collection vehicle. All of the sensors used for recording are commercially available.

C. Participants

In all, 30 participants (20 male, 10 female) took part in the experiment. They were, on average, 31 years old (range 20-58 years) and had held a driver's license for a mean period of

Fig. 2. Data collection vehicle.

11.4 years (range 1.4-39 years). They received 5,000 Japanese yen as compensation for their participation.

D. Procedure

When the participant arrived at the university, the experimenter took him/her to a meeting room to explain the procedure of the experiment, as well as the sensors used and the measurement method. Then, the participant was brought to the instrumented vehicle. After settling into the vehicle, the participant left the parking area. The experimenter monitored the experiment from the rear seat and indicated the route to the driver. The first few minutes were used to let the participant get used to the car and the sensors. Signals recorded during this initial period were not used in the study. All experiments were performed on the city streets of Nagoya, Japan. The data collection vehicle, experimental route, equipment, and operational conditions were the same for all participants.

The experimental route and time of day were pre-selected in order to increase the number of frustrating environmental factors. Throughout the drive, participants encountered various events that interrupted their driving journey and created stop-and-slow traffic conditions in which vehicles were unable to regain full speed before reaching the next obstruction. These events included pedestrians (especially students) and bicycles crossing the road, oncoming vehicles moving into the driver's lane, red-light signals, and slow moving vehicles blocking the driver's path. Furthermore, using an automatic speech recognition (ASR) system [19], drivers retrieved and played songs while driving from a list of 635 titles from 248 artists. Music could be retrieved by artist name or song title. The experimenter instructed the participant to retrieve as many songs as possible; accordingly, within around 30 seconds of successfully retrieving each song, the participant had to retrieve another song. This secondary task further increased the likelihood of frustration, since due to environmental noise, speech recognition errors were a frequent cause of impeded progress. After the experiment, the participant and experimenter returned to the university. The participant was then left alone in a room to complete his/her assessment of the experienced frustration. Finally, compensation for their participation was paid.

E. Measures

1) Pedal Actuation: Force sensors (Kyowa Electronics Instruments Co., LPR-A-03KNS1 and LPR-R-05KNS1) were mounted on the gas and brake pedals, respectively, so as to record the pedal actuation through force signals. The signals were originally acquired at a sample frequency of 16 kHz and further down-sampled to 10 Hz. The effects of different emotional states on the way we drive remains an open and very interesting question. This study tackled this problem by trying to show that gas- and brake-pedal actuation is affected by frustration. A controlled area network (CAN-Bus) was not used for our recordings because we decided to adopt sensors as similar as possible to those used during the CIAIR database recordings.

2) Electrodermal Activity: Electrodermal activity (EDA) was obtained with a skin potential sensor (SkinosSK-SPA) placed on the driver's left hand. The signals were originally acquired at a sample frequency of 16 kHz and further down-sampled at 10 Hz. Electrodermal activity is one of the most widely used response systems in the literature. It is caused by activation of the sympathetic nervous system, which changes the level of sweat in the eccrine sweat glands and has been shown to be linked to psychological concepts of arousal and attention [20]. High levels of arousal tend to indicate new, significant, or attention-getting events.

3) Speech Recognition Errors: Speech was recorded using a headset microphone. ASR systems, which are particularly error-prone in a noisy traffic environment, may present adverse effects such as taxing a driver's short-term memory. In order to identify speech recognition errors that may have led to frustration, speech recognition results, i.e., the machine's real-time transcription of the driver's utterances, were recorded.

4) Video Recordings: Videos were captured by three cameras with set focal points: the driver's face (x2) and the road view ahead of the vehicle.

5) Self-Reported Frustration: After the experiment, which lasted for around 15 minutes, participants were asked to assess their subjective level of frustration by referring to the front-view and facial videos as well as the corresponding audio. A user interface, shown in Fig. 3, was designed for such assessment. The strength of experienced frustration was indicated using a continuous intensity scale ranging from neutral to extremely frustrated. Participant were instructed to keep the button over "0" if no frustration was experienced, or indicate by how much they got frustrated by sliding this button. In case participants were in a certain mood already from the beginning of the experiment, they were asked not to incorporate such feelings in their rating. This was required to ensure that authentic emotions related to traffic events or to human-machine interactions were reported instead of general moods [6]. The level of frustration was recorded at fixed intervals. Figure 4 shows the assessments of three different

Fig. 3. Interface designed for frustration assessment.

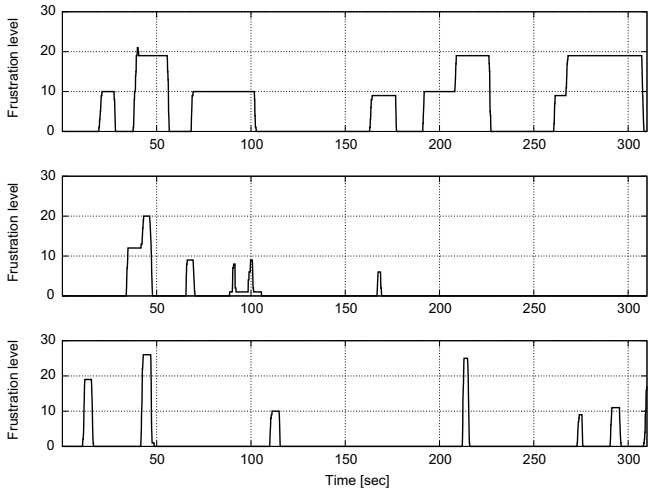

Fig. 4. Frustration levels assessed by three different drivers.

participants.

F. Feature Extraction

Recorded data were processed so that relevant information could be extracted from raw signals. In this section we describe the extracted features and related methods.

1) Pedal Actuation: Gas- and brake-pedal signals were first divided into frames having a length of M points. Features were then extracted through cepstral analysis of the gas- and brake-pedal signals. Ceptrum (cepstral coefficients) is a widely used spectral feature for speech and speaker recognition, and, more recently, it has proven to be effective in driver modeling [21]. Ceptrum is defined as the inverse Fourier transform of the short-term log-power spectrum and is obtained as follows:

$$c(m) = \frac{1}{M} \sum_{k=0}^{M-1} \log |X(k)| e^{j2\pi km/M}, \quad m = 0, 1, ..., M-1, \quad (1)$$

where $X(k)$ denotes the M-point discrete Fourier transform of the windowed signal $x(n)$. Cepstral analysis is a source-filter separation process [22]. By keeping only the first several coefficients in the lower "quefrency" and setting others to zero, we can obtain a spectral envelope as a filter that represents the process of acceleration or braking. On the other hand, a fine structure of the spectrum, the source, which works as the command signal for hitting a pedal, can be obtained by maintaining a higher "quefrency" range and setting the lower "quefrency" coefficients to zero. It is important to note that information on the spectral envelop is lost when all coefficients but $c(0)$ are set to zero. Furthermore, before calculating the cepstrum, gas and brake were combined by setting the pedal actuation signal to

$$x(n) = F_G(n) - F_B(n), \quad (2)$$

where $F_G(n)$ and $F_B(n)$ denote the gas and brake force signals, respectively. In order to take into account the dynamics of pedal actuation, the time derivative of the cepstral coefficients was also used as a feature. As shown in (3), for example, the time derivative of a discrete-time signal can be calculated by using linear regression coefficients for signal $y(n)$ with a window of length $2K$:

$$\dot{y}(n) = \frac{\sum_{k=-K}^{K} k \cdot y(n+k)}{\sum_{k=-K}^{K} k^2}. \quad (3)$$

2) Electrodermal Activity: The skin potential signal, represented as S, was first low-pass filtered using a second-order Savitzky-Golay smoothing filter with a length of 40.1 seconds, forming a smoothed skin potential G. Filter characteristics satisfactorily removed high-frequency noise from the raw signal. The G signal was normalized by subtracting its long-term mean and dividing the result by its maximum. G was then further divided into frames of length L points. Let $\tilde{G}(n)$ represent the value of the n^{th} sample of a given data frame, where $n = 1, ..., L$. We calculated the following two statistical features:

1) Local mean of normalized signal (mean skin potential):

$$f_1 = \frac{1}{L} \sum_{n=1}^{L} \tilde{G}(n) \quad (4)$$

2) Absolute value of the first-order difference of the normalized signal (Δ skin potential):

$$f_2 = \sum_{n=1}^{L} \left| \tilde{G}(n+1) - \tilde{G}(n) \right|. \quad (5)$$

f_1 was further uniformly quantized into four levels and f_2 into two. The threshold used to quantize f_2 was defined experimentally, being the one that provided the best overall estimation.

3) Speech Recognition Errors: The failure of the ASR system to correctly recognize the name of artists or songs was the most common type of recognition error. Participants were instructed to say *No* when reacting to such errors so that they could repeat the desired input until the machine got it right, as exemplified in the following dialogue:

Driver: Bee Gees.
Machine: Do you want to search for Britney Spears?
Driver: *No...*
Machine: I am sorry. One more time, please.
Driver: Bee Gees.

As a possible indicator of speech recognition errors, we used the instances when the ASR system recognized a participant's utterance as *No*. This indicator was selected due to its consistency across different drivers and required calculation time, which is negligible. Since this is a very short event that lasts around 1 second, an enlargement of its boundaries was necessary. Based on a preliminary analysis of frustration videos, we added 5 seconds before and 15 seconds after each utterance recognized as *No*, as shown in Fig. 5. Here, 20 seconds is the time span in which significant verbal or gestural reactions

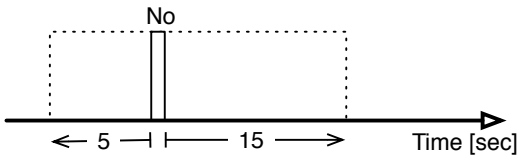

Fig. 5. Enlargement of boundaries of an utterance recognized as *No*.

still occurred; accordingly, we encoded speech recognition errors as a binary signal, in which errors were indicated by a 20-second window of "1s." The enlargement of boundaries partially solved the problem of different timings between ASR errors and other reactions, such as facial expressions.

4) Self-Reported Frustration: The frustration level was quantized into two levels: frustrated (frustration $>$ threshold) and not-frustrated (frustration $\leq$ threshold). The optimal quantization threshold was determined experimentally, being the one that provided the best overall estimation. This quantization generated a more consistent signal than the original one (0-30).

G. Data Annotation

We designed a data annotation protocol that covers a number of environmental factors that might impede a driver's progress and cause frustration. Facial expression was also included in the annotation, since it is one of the possible responses to frustrating events. Currently, there seems to be no consensus on the best way of annotating driving data, and thus different protocols have been adopted [4], [6]. In this study, based on scenes of high levels of frustration, selected annotation labels and possible states were

1) Traffic density (light/medium OR high);
2) Obstructions caused by pedestrians, bicycles, and parked vehicles (non-obstructed/obstructed);
3) Stops at red-light signals (non-stopped/stopped);
4) Turn (not turning/turning);
5) Curve (not a curve/curve);
6) Overall face (neutral/non-neutral).

Labels 1-5 provide us with a simple description of the driving environment that may affect drivers' internal state. Data annotation from all 30 participants was manually carried out by seven annotators who were allowed to utilize only frontal video. No audio was provided to avoid bias when labeling overall face data from the video. Annotators, who volunteered for the task, coded the time span of each state, so that results could be seen as a multi-stream of binary information, as show in Fig. 6. Annotation results were checked individually by the authors to ensure a high reliability. On average, ten minutes of data took 50 minutes to annotate. To speed up the process, we are now developing a dedicated annotation interface. We focused on designing simple yet informative annotation labels, making it easier for a real application to automatically and effectively annotate them in the future.

H. Analysis

A method for combining all of the different features and annotation results in an efficient language was needed, and a

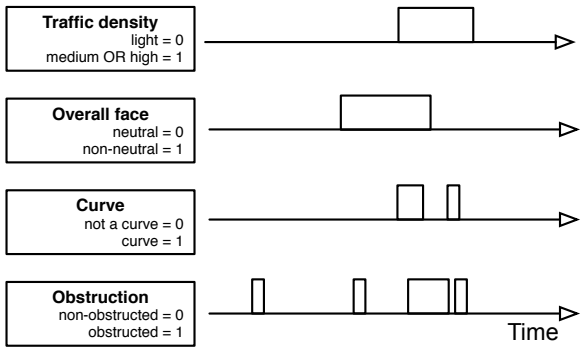

Fig. 6. Example of annotation.

Bayesian network (BN) was the natural choice to deal with such a task. One of the important characteristics of a BN is the capability to infer the state of an unobserved variable, given the state of observed ones. In our case, we wanted to infer a participant's frustration given the driving environment, speech recognition errors (communication environment), and the participant's responses measured through his/her physiological state, overall face, and pedal actuation.

The graph structure proposed to integrate all of the available information is shown in Fig. 7. This model was based on the following assumptions: (1) environmental factors that may have an impact on goal-directed behavior (traffic density, stops at red-light signals, obstructions, turn or curve, and speech recognition errors) may also have a direct effect on frustration; (2) a frustrated driver is likely to present changes in his/her facial expression, physiological state, and gas- and brake-pedal actuation. In Fig. 7, squares represent discrete (tabular) nodes and the circle represents a continuous (Gaussian) node. The number inside each node represents the number of mutually

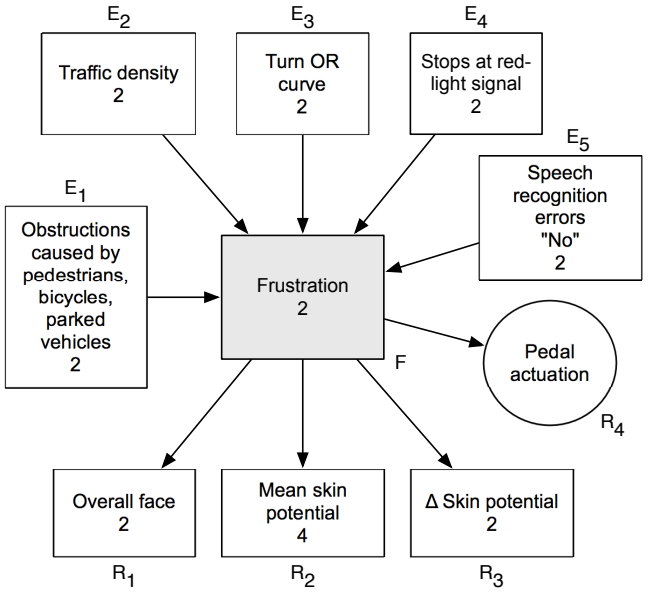

Fig. 7. Proposed Bayesian network structure. Squares represent discrete (tabular) nodes and the circle represents a continuous (Gaussian) node. The number inside each node represents the number of mutually exclusive states that the node can assume. Labels outside nodes identify random variables.

exclusive states that the node can assume. Random variables were identified by a label outside each node: "F" for frustration, "E" for environment, and "R" for responses.

In addition to the graph structure, it is necessary to specify the parameters of the model, obtained here using a training set. During parameterization, we calculate the Conditional Probability Distribution (CPD) at each node. If the variables are discrete, this can be represented as a table (CPT), which lists the probability that the child node takes on each of its different values for each combination of values of its parents. On the other hand, if the variable is continuous, the CPD is assumed as a Gaussian distribution. For example, the continuous node *Pedal actuation*, which has only one binary parent, was represented by two different multivariate Gaussians, one for each emotional state: frustrated and not frustrated. For each observed environment (driving and communication) and the corresponding driver responses, we can use Bayes' rule to compute the posterior probability of frustration, as described in (6):

$$P(F|E_1,E_2,E_3,E_4,E_5,R_1,R_2,R_3,R_4) = P(F|E_1,E_2,E_3,E_4,E_5)\cdot \frac{P(R_1|F)\cdot P(R_2|F)\cdot P(R_3|F)\cdot P(R_4|F)}{P(E_1,E_2,E_3,E_4,E_5,R_1,R_2,R_3,R_4)} \cdot \prod_{j=1}^{5} P(E_j), \tag{6}$$

The denominator in (6) was calculated by summing out F from the joint probability of $P(F,E,R)$. In addition, in this study we set a uniform Dirichlet prior to every discrete node in the network. This was done in order to avoid over-fitted results due to the Maximum Likelihood approach used for calculating the conditional probability tables. Without a prior, patterns that were not observed in the training set would be assigned zero probability, compromising the estimation [23]. Further details on Bayesian networks can be found elsewhere [24], [23]. In experiments, we used the Bayes Net Toolbox for Matlab, freely available [25].

The network data input scheme is shown in Fig. 8. All of the available data—pedal actuation, skin potential and other binary signals—were entered concurrently. At a given time step t, frames of sizes L and M were used to extract features from the skin potential and pedal actuation signals, respectively. Results served as network inputs. The value of each binary label at the current time step was directly entered in the network without further processing. Frame shift was kept fixed at 0.5 seconds. As described in Fig. 8 for two consecutive frames, the value of, for example, current traffic density has an effect on future skin potential and pedal actuation signals in order to account for delayed physiological and behavioral reactions. In addition, frustration was estimated continuously, i.e., we did not preselect segments where we were certain about frustration or neutrality and then ignore ambiguous regions.

III. Experimental Setup and Evaluation

Experiments were performed with data from 13 male and 7 female participants. They were, on average, 29 years old (range 20-46 years) and had held a driver's license for a mean period of 10.4 years (range 1.4-27 years). The frustration of 10 from the original 30 participants was very short (total duration of frustration scenes less than 20 seconds), so their data were not used. Since possible causes and consequences of frustration depend on personal characteristics, we trained one network for each participant.

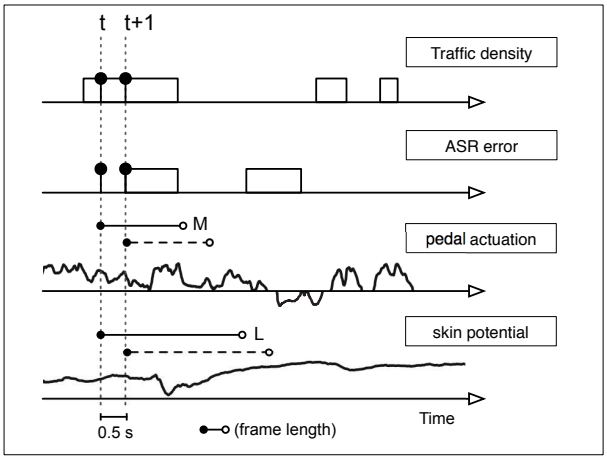

Fig. 8. Scheme for entering data into the Bayesian network. Frame shift is kept to 0.5 seconds. Two consecutive time steps, t and $t+1$, are shown. M and L denote the frame length for pedal actuation and skin potential, respectively.

The optimal configuration of estimation parameters was achieved experimentally by, first, selecting possible ranges. We then changed one parameter at a time—keeping the others fixed—in the following order: skin potential signal frame length L (16, 32, 64, 128, 160), threshold for quantizing the Δ skin potential (f_2) into two steps (0.1, 0.2, 0.3, 0.4, 0.5), Dirichlet prior hyperparameter (1, 5, 10, 15, 20, 25, 30, 40, 50), and threshold for quantizing the frustration signal into two steps (0, 1, 2, 3, 4, 5, 6, 7, 8 , 9, 10). At each step, we selected the optimal value for the parameter being changed. Data frames were divided into training (60%) and test (40%) sets.

In order to determine the usefulness of different types of information as indicators of frustration, we compared the results of five distinct variations of the network shown in Fig. 7. The direction of arrows was kept fixed and the number of nodes was modified, depending on the network variation—all other parameters were kept fixed at their optimal value. Table I summarizes the network variations. The *Basic* network was used to verify the role of physiological state features and facial expressions on frustration estimation; the *Driving env* network added information on the driving environment to the *Basic* network; the *Communication env* network focused on evaluating the importance of speech recognition errors; the *Pedals* network focused on evaluating the importance of pedal actuation; and the *Full* network (Fig. 7) had all nodes.

In order to achieve the optimal parameters for the Pedal actuation node, we conducted experiments using the *Pedals* network. Tested parameters were driving behavior signal frame length M (8, 16, 32, 64, 128) and number of cepstral coefficients (0, 1, 2, 4). When calculating the time derivative of pedal actuation, $2K = 800$ ms was used.

TABLE I
NETWORK VARIATIONS

Nodes	Network variations				
	Basic	Driving env	Communication env	Pedals	Full (Fig. 7)
frustration	√	√	√	√	√
mean skin potential	√	√	√	√	√
Δskin potential	√	√	√	√	√
overall face	√	√	√	√	√
traffic density		√	√	√	√
turn or curve		√	√	√	√
stops at red-light signal		√	√	√	√
obstructions caused by pedestrians, bicycles, and parked vehicles		√	√	√	√
speech recognition errors			√		√
pedal actuation				√	√

We evaluated the capacity of the proposed system to detect frustration. After calculating the estimation signal for each participant, i.e., the posterior probability of the frustration node, it was quantized into two steps: frustrated (probability $>$ 0.5) and not frustrated (probability $\leq$ 0.5). The result was then filtered using a median filter of 12 seconds so that spikes could be removed. In order to estimate the overall effectiveness of detection, we summed true (T)/false (F) positives (P)/negatives (N) from all drivers so that we could calculate overall true and false positive rates, represented by a single point in the receiver operating characteristic (ROC) space. Overall true and false positive rates were calculated as follows:

$$\text{overall TP rate} = \frac{\sum_i^I TP_i}{\sum_i^I (TP_i + FN_i)}, \text{ and} \tag{7}$$

$$\text{overall FP rate} = \frac{\sum_i^I FP_i}{\sum_i^I (TN_i + FP_i)}, \tag{8}$$

where I is the total number of participants. The ROC space provides a ratio indicating the system's ability to correctly estimate frustration versus its transparency, i.e., the system's ability to suppress false alarms and avoid driver annoyance.

IV. RESULTS

Within the data used in experiments, 129 scenes of frustration (segments with original scale above 0) were found. On average, participants got frustrated 6.5 times while driving. The mean strength of frustration scenes was 10.5, and the mean duration was 11.8 seconds.

Figure 9 shows estimation results for individual drivers arranged side by side: actual frustration from all 20 participants (top); the posterior probability of the frustration node calculated using the *Basic* network (center); and quantized posterior probability using a threshold of 0.5 (bottom). The quantized probability of each driver was further median-filtered to remove spikes. Figure 10 shows the same data calculated using the *Full* network. These results suggest that the estimation benefited from the introduction of additional information.

Overall quantitative results for all five network variations are shown in Fig. 11. In the ROC space, the point (0,1) represents the perfect estimation. The closer the result gets to this point, the better. Circles centered at (0,1) are plotted so that different results can be easily compared. The *Basic* network, which relied on physiological state and facial expressions, actually achieved the worst overall result. The best overall result was achieved by the *Full* network: a true positive (TP) rate of 80% and a false positive (FP) rate of 9% (i.e., the system correctly estimated 80% of the frustration and, when drivers were not frustrated, made mistakes 9% of the time). Furthermore, the results suggest that information on the driving and communication environment, as well as pedal actuation, was effective in improving the model accuracy, adding new information to the *Basic* network.

Figure 12 shows the above influence in the results of different thresholds for quantizing the frustration level into two steps. Automatic selection of the optimal threshold provided a better estimation than simply setting values above zero in the original scale as "frustrated" (threshold = 0).

Figure 13 shows the results of the *Pedals* network trained with different numbers of cepstral coefficients. The worst performance was achieved when the network relied only on $c(0)$, i.e., when no spectral information was provided. This finding suggests that not only the intensity of the pedal actuation signal but also its spectral envelope contains important information on the driver's emotional state.

Optimal estimation parameters, which provided the best overall estimation results, were skin potential signal frame length L of 128 points, threshold of 0.3 for quantizing the Δ skin potential (f_2) into two steps, Dirichlet prior hyperparameter of 30, and threshold of 5 for quantizing the frustration signal into two steps. The optimal driving behavior signal frame length M was 16 points, and the best cepstral coefficients were $c(0)$ + $c(1)$.

V. DISCUSSION

The results of this study suggest that the estimation of a driver's emotional state would gain from a more meaningful modeling strategy that combines information on the environment, the driver's internal state, and his/her responses in a single system. We believe that the same type of modeling discussed here could be applied to a number of driver-monitoring problems other than frustration estimation, such as stress, attention and drowsiness detection. In this section we discuss the results and their implications.

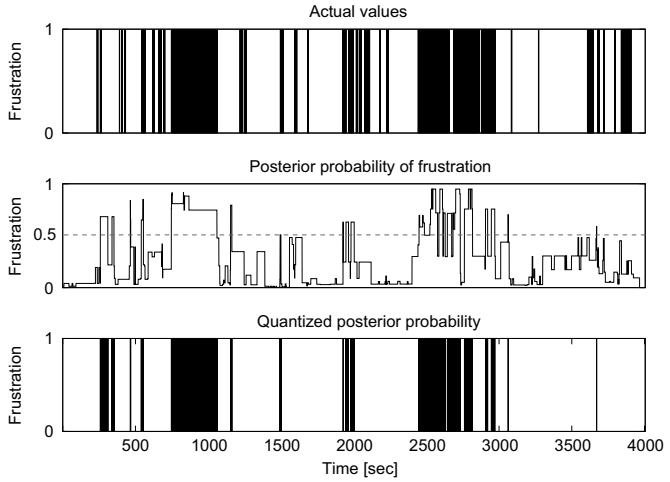

Fig. 9. Results for individual drivers (arranged side by side) calculated using the *Basic* network. Comparison between actual frustration (top), posterior probability of the frustration node (center), and its quantized version using a threshold of 0.5 (bottom).

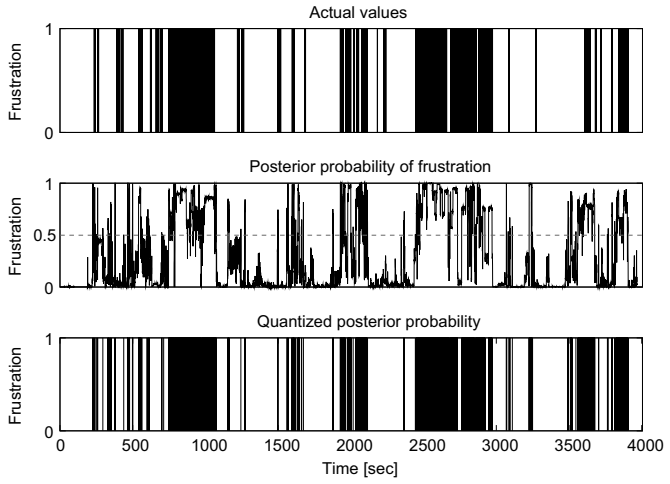

Fig. 10. Results for individual drivers (arranged side by side) calculated using the *Full* network. Comparison between actual frustration from all drivers (top), posterior probability of the frustration node (center), and its quantized version using a threshold of 0.5 (bottom).

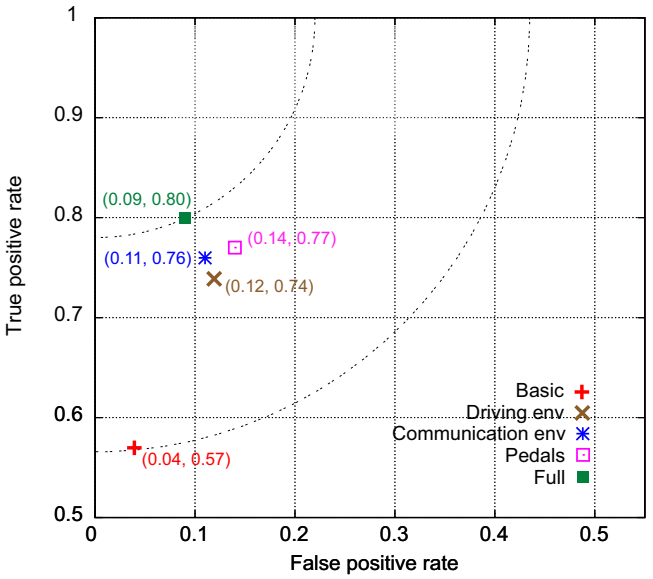

Fig. 11. Overall results achieved by the five different network variations. Dashed lines are part of circles centered at (0,1).

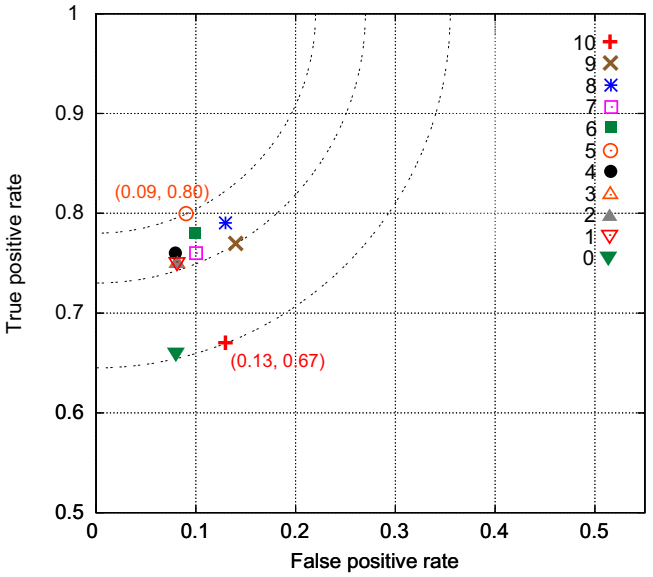

Fig. 12. Overall results achieved using different thresholds for quantizing frustration level into two steps.

Experiments were performed with five different network variations. The *Basic* network presented the worst performance, being unable to satisfactorily identify frustration based solely on facial expressions and physiological condition. One of the reasons for such poor results is that drivers may respond similarly to both frustrating and certain non-frustrating situations. When the driving environment was introduced, a more accurate estimation was obtained. The newly added information provided a context to drivers' responses, allowing a meaningful interpretation of events. The driving environment in the present study was represented as a series of binary signals manually coded; however, we believe that the same knowledge could be obtained automatically by integrating information from the infrastructure, navigation system, and in-car cameras.

Experiments with the *Communication env* network indicated that information on speech recognition errors was effective in increasing the overall estimation accuracy, as expected. This result, together with the lessons learned during the CIAIR database annotation, indicates the relevance of considering communication quality as an indicator of emotions. Although in a more realistic scenario the driver could simply refrain from using speech commands, interaction with ASR systems are becoming part of everyday life as such systems become ubiquitous. Therefore, the effects of this interaction must be carefully studied.

Results from the *Pedals* network also allowed us to draw important conclusions: (1) the way gas and brake pedals are used might be affected by frustration; and (2) the effect of frustration can be observed not just in the intensity of the pedal actuation signal but also in its spectral envelope. Therefore,

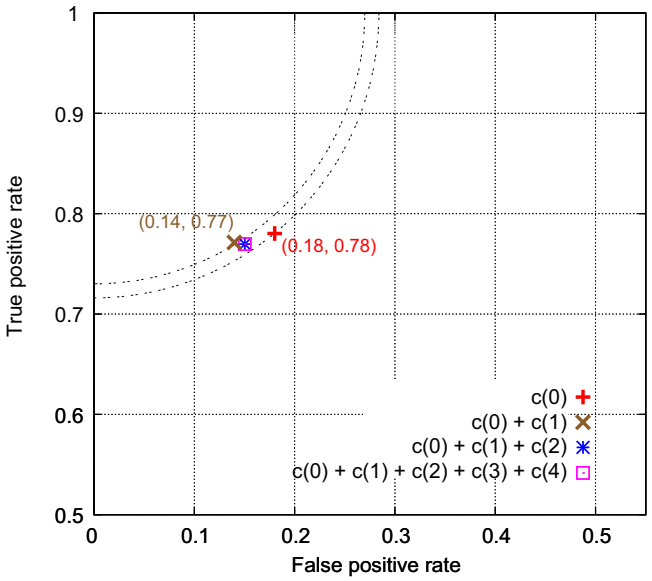

Fig. 13. Overall results achieved using a network trained with different numbers of cepstral coefficients. Dashed lines are part of circles centered at (0,1).

cepstral coefficients of the pedal actuation signal and their time derivative ought to be considered in future research in emotion recognition. Current studies that investigate pedal actuation tend to use only such features as mean and standard deviation.

The threshold for quantizing the frustration level into two steps was the same for all drivers, and it was selected experimentally. Results show that, compared to a simple selection of values above zero as "frustrated," automatic search for an optimal threshold provides a better overall estimation. Results for other values close to the optimal threshold of five were similar. We believe that selecting this threshold individually may further improve the estimation. This topic will be studied as a future work.

In this study, frustration was reported quite often. One possible explanation is that the unfamiliarity of experimental conditions, instrumented car, experimenter, and ASR system may have had an impact on drivers, making them more sensitive. The higher frequency can also be explained by the assessment method. Usually, driving-log or questionnaire-based assessment leads to an under-registration of mild emotions. People tend to remember only more extreme cases and forget about mild ones. The fact that in the present study most of the reported emotions were mild (average strength of 10.5) supports this hypothesis, which is also in line with the work done by Mesken [6].

The present study had some limitations, the first being related to the frustration assessment method. Data reported here were based solely on self-assessment, thus social desirability might have biased the data. It is possible that some participants may have embellished their answers. Nevertheless, participants could not gain any benefit by disguising their behavior since confidentiality of their personal data was assured to them and the assessment was conducted in a separate room, without interference from the experimenter. In fact, classical studies on human-computer interaction suggest that, concerning the reporting of self-perceived negative information, subjects may be more economical with the truth with humans than with computer interrogators [26]. Moreover, self-assessment may have allowed each participant to have his/her own standards. In an effort to cope with this issue, we only trained individual models tailored to each driver. The personalization of vehicles is a current tendency [21]; thus the adoption of individual models was not an unreasonable choice. Moreover, the raw frustration signal (0-30) was automatically quantized into two levels, generating more consistent information.

Other limitations were related to the experimental design. Previous studies on the occurrence of emotions [15] suggested that the inclusion of a sense of time pressure in the scenario resulted in higher levels of aroused anger and stress. Time urgency could have been used in this study to further provoke frustration. Nevertheless, although the only form of pressure considered here had to do with the number of songs that should be retrieved using the ASR system, the results indicate that the scenario experienced by the participants was sufficient to elicit frustration in most of them.

It is impossible to simultaneously consider more than a small part of the available information when modeling the driver. Our selection of features inevitably constrained our interpretation of frustration in ways difficult to predict. Bradley and Lang [27] summarized emotional cues into three categories: behavioral sequences, physiological reactions, and emotional language. In this study, although behavioral and physiological clues were combined, language was analyzed only to identify speech recognition errors. Other cues from the speech signal and speech recognition results, such as prosody and semantics, should be investigated in future work. Information on the vehicle (e.g., velocity or lane deviation) could also be included, depending on the goal of the application. The developed system is very scalable; therefore, new features can be easily combined. Finally, we intend to assess the generality of our method regarding both individual and overall tendencies after collecting a larger dataset.

VI. Conclusion

This study showed that a more comprehensive modeling method, which takes the environment into account, provides a better estimation of frustration and, possibly, of other emotions. The inclusion of speech recognition errors was effective in increasing the overall accuracy of results, suggesting that ASR systems have an impact on the driver's emotional state and, thus, should be considered in future studies on emotion recognition and interface design. Information on gas- and brake-pedal actuation positively contributed to the overall results, which gives further indication that the way we drive might be affected by our emotional state. Not just the energy of the pedal actuation signal but also its spectral envelope played a role in improving estimation. In addition, automatic quantization of frustration levels proved to be more satisfactory than a simple manual selection.

Further research may be directed at the inclusion of a model of speech prosody or semantics-related features. In addition, the evaluation of a dynamical structure that considers the

temporal relationship between nodes is an interesting topic and ought to be investigated. More insights could be gained as well by further analyzing the ASR system and how different drivers react to different types of errors.

REFERENCES

[1] T. Gandhi and M. M. Trivedi, "Pedestrian protection systems: Issues, survey, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 3, pp. 413–430, 2007.

[2] S. K. Gehrig and F. J. Stein, "Collision avoidance for vehicle-following systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 233–244, 2007.

[3] N. Kaempchen, B. Schiele, and K. Dietmayer, "Situation assessment of an autonomous emergency brake for arbitrary vehicle-to-vehicle collision scenarios," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 4, pp. 678–687, 2009.

[4] J. Healey and R. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[5] Y. Liang, M. L. Reyes, and J. D. Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 340–350, June 2007.

[6] J. Mesken, "Determinants and consequences of drivers' emotions," Ph.D. dissertation, University of Groningen, 2006.

[7] R. S. Lazarus, *Appraisal processes in emotion: Theory, methods, research*. New York: Oxford University Press, 2001, ch. Relational meaning and discrete emotions, pp. 37–67.

[8] C. Katsis *et al.*, "Toward emotion recognition in car-racing drivers: A biosignal processing approach," *IEEE Transactions on Systems, Man and Cybernetics, Part A*, vol. 38, no. 3, pp. 502–512, 2008.

[9] C. L. Lisetti and F. Nasoz, "Affective intelligent car interfaces with emotion recognition," *11th International Conference on Human Computer Interaction*, pp. 1–10, 2005.

[10] B. Schuller, M. Lang, and G. Rigoll, "Recognition of spontaneous emotions by speech within automotive environment," *Jahrestagung für Akustik (DAGA)*, vol. 32, pp. 57–58, 2006.

[11] C. Jones and I.-M. Jonsson, "Detecting emotions in conversations between driver and in-car information systems," *Proc. 1st International Conference on Affective Computing and Intelligent Interaction*, pp. 780–787, 2005.

[12] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal fusion of emotional data in an automotive environment," *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 2, pp. 1085–1088, 2005.

[13] P. Ekman, *Handbook of Cognition and Emotion*. Sussex, U.K. John Wiley & Sons, Ltd., 1999.

[14] J. Dollard, N. E. Mille *et al.*, *Frustration and aggression*. Yale University Press, New Haven, 1939.

[15] D. Shinar, "Aggressive driving: the contribution of the drivers and the situation," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 1, no. 2, pp. 137–160, 1998.

[16] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 39–58, 2009.

[17] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 724–736, 2007.

[18] N. Kawaguchi, K. Takeda, and F. Itakura, "Multimedia corpus of in-car speech communication," *The Journal of VLSI Signal Processing*, vol. 36, no. 2, pp. 153–159, 2004.

[19] S. Hara, C. Miyajima, K. Itou, and K. Takeda, "An online customizable music retrieval system with a spoken dialogue interface," *The Journal of the Acoustical Society of America*, vol. 1, no. 5, pp. 3378–3379, 2006.

[20] J. T. Cacioppo and L. G. Tassinary, *Principles of Psychophysiology: Physical, Social and Inferential Element*, J. T. Cacioppo and L. G. Tassinary, Eds. Cambridge University Press, 1990.

[21] C. Miyajima *et al.*, "Driver modeling based on driving behavior and its evaluation in driver identification," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 427–437, 2007.

[22] A. V. Oppenheim, "Superposition in a class of nonlinear systems," Ph.D. dissertation, Res. Lab. Electronics, Massachusetts Institute of Technology, Cambridge, MA, 1965.

[23] C. M. Bishop, *Pattern recognition and machine learning*, M. Jordan, J. Kleinberg, and B. Schölkopf, Eds. Springer, 2006.

[24] K. P. Murphy, "Inference and learning in hybrid Bayesian networks," University of California, Tech. Rep. CSD-98-990, 1998. [Online]. Available: citeseer.ist.psu.edu/murphy98inference.html

[25] K. Murphy, "Bayes Net Toolbox for Matlab," 2007. [Online]. Available: http://bnt.sourceforge.net/

[26] R. W. Lucas, P. J. Mullin, C. B. Luna, and D. C. McInroy, "Psychiatrists and a computer as interrogators of patients with alcohol-related illnesses: a comparison." *Br J Psychiatry*, vol. 131, pp. 160–167, 1977.

[27] M. Bradley and P. Lang, *Cognitive neuroscience of emotion*. New York: Oxford University Press, 2000, ch. Measuring emotion: Behavior, feeling and physiology, pp. 242–276.