A Multimedia Corpus of Driving Behaviors

Lucas Malta [#1], Akira Ozaki [#2], Chiyomi Miyajima [#3], Norihide Kitaoka [#4], Kazuya Takeda [#5]

[#] *Grad. School of Information Science, Nagoya University*
464-8603 Furo-cho, Chikusa-ku, Nagoya, Japan
[1,2,3] {malta, ozaki, miyajima}@sp.m.is.nagoya-u.ac.jp
[4,5] {kitaoka, kazuya.takeda}@nagoya-u.jp

***Abstract*—In this paper we present our multimedia corpus of real-world driving data (NUDrive), built with the primary objective of firming foundations for applying digital signal processing technologies in the vehicular environment. NUDrive is a content rich corpus composed of driving, speech, video, and physiological signals. So far, we have collected data from 250 drivers, who drove an instrumented vehicle under very similar conditions. In order to provide a more meaningful description of the situations drivers experience, a comprehensive data annotation protocol is proposed. We also briefly present a multimedia processing system, which uses information from various sources in NUDrive to implement a context-dependent estimation of a driver's spontaneous frustration. Results are encouraging and stress the relevance of content rich driving corpora to driver behavior modeling.**

I. Introduction

The enormous impact on society of traffic-related problems has, for many decades now, led academia and industry to extensively study driver behavior. As a major step towards a firmer research foundation, on which technologies for safety and comfort while driving can be studied and evaluated, the collection of human activities related to driving has proven to be indispensable [1][2].

Laboratory-based approaches, especially driving simulators, are the most common alternative to driving data collection. Although less costly and more controllable, results based solely on simulator data cannot be directly applicable to real-world, since it has been shown that, among other things, drivers' reactions might be different in the lab than in real vehicles [1][3]. Under real-world conditions, there are many challenges in both implementing measurement systems and carrying out the data collection experiments. On-road and naturalistic approaches are by far more expensive and time consuming, and require expertise in different research fields. Nevertheless, data collected under real-world conditions provide a unique insight into driver behaviors and open a wide range of research possibilities [4][5].

Although the use of instrumented vehicles for collecting real-world driving data dates back to the 70s [6], only recently has the recording of large-scale information become possible. The Center for Integrated Acoustic Information Research (CIAIR) at Nagoya University (NU), a pioneer in this work, recorded real-world data from 500 drivers from 2000 to 2002 [7]. The CIAIR database is primarily focused on speech processing; thus, precious informations on driver status (e.g., physiological signals) and vehicle environment (e.g., following distance) are not available.

To overcome this drawback and provide more detailed information on drivers and environmental conditions, we have been constructing a large-scale corpus (NUDrive) since 2006. In this paper, we describe this research project devoted to capturing exhaustively the human activity with multimedia signals from many sensors in real traffic. Our main goal is to, based on collected data, develop firm foundations for applying digital signal processing technologies in the vehicular environment. So far, driving, speech, video, and physiological signals have been recorded from 250 drivers, who drove the same instrumented vehicle under very similar conditions. We plan to continue collecting these data from 250 additional drivers until 2011. Data collection apparatus is presented in section II-A, collection process in section II-B, and collected data is overviewed in section II-C.

One of the major challenges in the driver behavior analysis field is the development of context-sensitive methods that also take into account the situation in which the behavior is elicited, rather then relying solely on responses (e.g., force on pedals, facial expressions, physiological changes). While driving, our actions are, most of the time, carefully planned after a complex cognitive decision-making process based on different variables such as weather condition, road design, and the presence or absence of pedestrians. Therefore, an effective labeling of multimedia information is critical for providing a more meaningful description of the situations drivers experience. In this work we also proposed a data annotation protocol, presented in section III.

In order to study how different cultural traits and traffic conditions affect the behavior of drivers, part of the data collection is conducted under international collaboration with the University of Texas at Dallas (USA) and with Sabanci University (Istanbul, Turkey). Data are collected and annotated in a similar fashion in all three sites. Such collaboration is crucial for not only research analysis and consistent technology evaluation, but also for establishing international standards in this area.

Finally, in section IV we present an example of a multimedia processing system devoted to the context-dependent estimation of a driver's spontaneous frustration. This system combines both collected data and annotations, and has a potential to

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00 ©2009 IEEE.

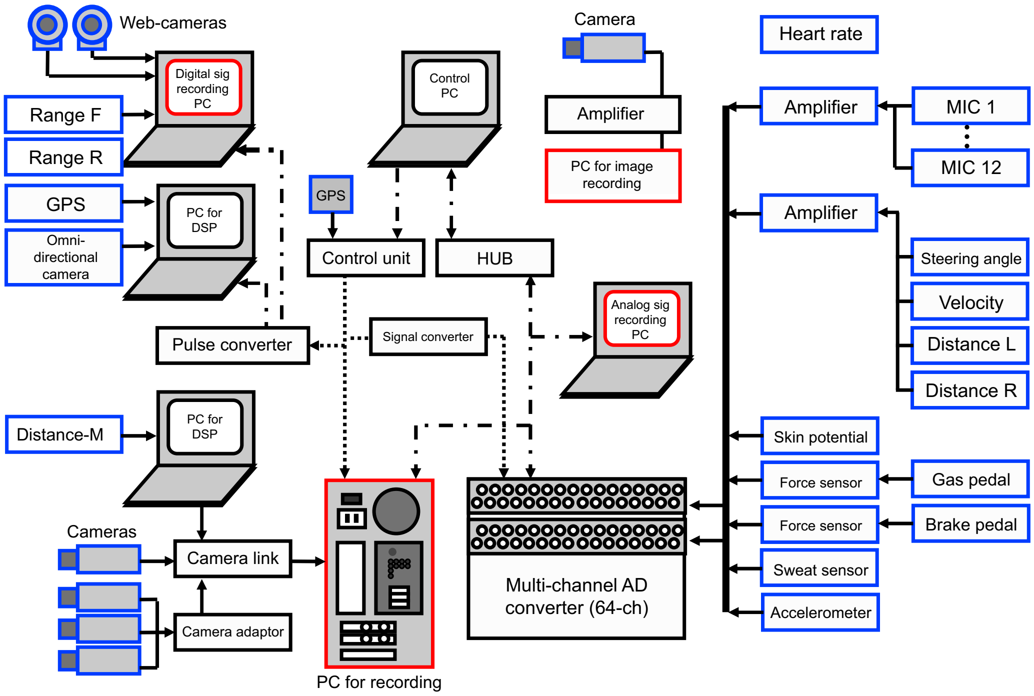

Fig. 1. Sensor and designed recording system utilized in data collection. Devices are divided into: input (blue), processing (black), and storage (red).

enhance the interaction between driver and vehicular systems.

II. On-Road Data Collection

A. Data collection apparatus

A data collection vehicle was designed for synchronously recording audio with other multimedia data. Various sensors were mounted on a Toyota Hybrid Estima with 2,360 cc displacement and automatic transmission. Data collection was conducted using the system described in Fig. 1. Blue, black, and red boxes indicate input, processing, and storage devices, respectively. Videos were captured by five cameras with set focal points: the drivers face (x3 different views), drivers feet, and the road view ahead of the vehicle. An omnidirectional camera was also mounted on the roof. A potentiometer (Copal M-22E10-050-50K) was used to measure steering angles, and force sensors (Kyowa Electronics Instruments CO. LPR-A-03KNS1 and LPR-R-05KNS1) were mounted on the gas and brake pedals, respectively. Vehicle velocity was measured based on the output of the JIS5601 pulse generator. Distance per 100 ms was obtained by multiplying pulse intervals and tire circumference. All digital signals were converted to analog by a D/A converter, so as to be sampled synchronously with other analog signals. Two kinds of distance sensors (Sick DMT-51111 and Mitsubishi MR3685) were mounted in front of the vehicle to measure short and long ranges, respectively. A differential GPS was used for recording the vehicle's position. In addition, 3D acceleration was acquired using a three-axial low-power accelerometer from Crossbow Technology, Inc. (CXL04LP3). Eleven omnidirectional condenser microphones (Sony ECM-77B) and a close-talking headset microphone were mounted on the vehicle to record driver speech. As for the physiological signals, driver heart rate was acquired using a chest belt sensor (Polar S810i), electrodermal activity (EDA) was obtained with a skin potential sensor (SkinosSK-SPA), placed on drivers left hand, and sweat levels using a sweat sensor (Skinos, SKN 2000).

All sensors used in recordings are commercially available. Driving data collection is also possible using an increasingly common onboard communication protocol called Controlled Area Network (CAN). The CAN-Bus signals contain real-time vehicle information in the form of messages, which greatly facilitates the acquisition process. CAN was not used in NUDrive recordings because we decided to adopt sensors as similar as possible to what we had in a previous data collection project [7].

B. Data collection process

Participants drove the instrumented vehicle on city streets and expressways in the city of Nagoya, Japan. During the experiment, drivers performed secondary tasks carefully designed to provide activities that were most likely to occur during everyday driving. Detailed instructions on how to perform each task were provided prior to the start of the experiment. Data collection vehicle, route, equipments, and treatment conditions were the same for all drivers. Drivers performed the same secondary tasks in the same order at very similar locations, so data from different drivers can be readily compared. An experimenter monitored the experiments from the rear seat.

Secondary tasks have been widely used as a way of increasing workload, so driver performance under different circumstances can be evaluated. Secondary tasks proposed in

TABLE I
SECONDARY TASKS.

ID	Description
SR	*Signboard reading task.* Drivers read aloud signboards containing, for example, names of shops seen from the drivers seat while driving.
ALS	*Alphanumeric strings reading.* Drivers repeated random four-character strings spoken by a machine.
ND	*Cellular phone navigation dialogue.* Drivers were guided to an unfamiliar place by a human navigator through a hands-free cellular phone.
MR	*Music retrieval task.* Using an automatic speech recognition (ASR) system, drivers retrieved and played songs from a list of 635 titles from 248 artists. Music could be retrieved by artist name or song title, e.g., "Beatles" or "Yesterday."
NT	*No Task Baseline.* Just driving without any task.

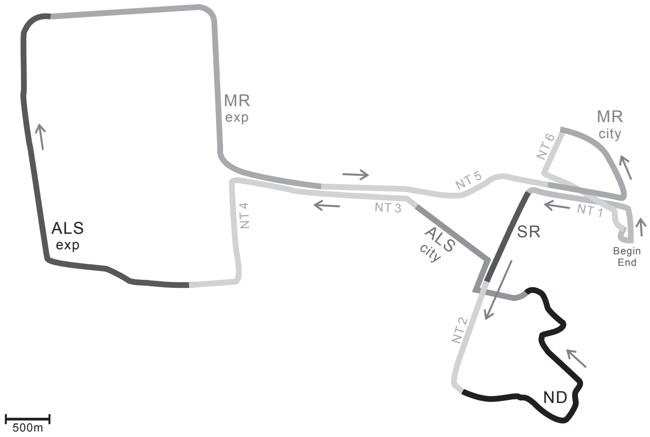

Fig. 2. Route participants followed. Secondary tasks and travel direction are indicated. Music retrieval (MR) and alphanumeric strings reading (ALS) were performed on both city streets (city), and on expressways (exp). Periods of no secondary task are labeled NT.

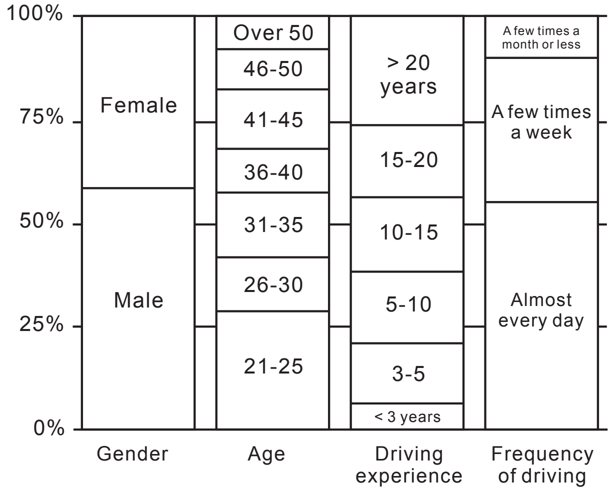

Fig. 3. Basic demographics for the 250 participants.

NUDrive focused on information exchange with in-vehicle interfaces. Especially in Japan, interaction with navigation systems and information retrieval tasks are part of everyday driving and so require careful study. Proposed secondary tasks are described in Tab. I. Figure 2 shows the route that participants followed. Letters indicate where the different types of tasks were performed.

Two questionnaires, one before and one after the experiment were filled out. The first questionnaire collected information on driver demographics such as age, driving experience, and frequency of driving. In the second one, drivers were asked to express their opinion about secondary tasks and equipment used during the experiment, such as the navigation and the speech recognition system.

After completing the route, the participant was also asked to assess his/her subjective level of frustration by referring to the front-view and facial videos as well as the corresponding audio. A user interface for such assessment was designed so that drivers used a continuous intensity scale and slid a bar from neutral to extremely frustrated. The assessment was done using data recorded when subjects drove on city streets while interacting with the ASR system to retrieve and play music (MR city). For this specific part of the experiment, route and time were pre-selected in order to increase the number of frustrating environmental factors, such as pedestrians, especially students, and bicycles crossing the road, oncoming vehicles moving across into the driver's lane, red-light signals, and slow moving vehicles blocking driver's path. While interacting with the machine, errors in speech recognition were also a frequent cause of discomfort.

C. Collected data

Video footage, driving, audio, and physiological signals were continuously recorded under both driving and idling conditions from 250 subjects (about 250 hours). Figure 3 shows demographic data on gender, age, driving experience, and frequency of driving. Data are representative of the Japanese population. Figure 4 shows examples of recorded signals. Data from other 200 drivers were collected in Dallas, USA (100 drivers) and Istanbul, Turkey (100 drivers) in a similar fashion as part of an international collaboration. Preliminarily analysis showed that there is no significant difference in the global distribution of velocity, gas, and brake pedal force signals across countries.

III. DATA ANNOTATION

An effective annotation of multimedia information is crucial for providing a more meaningful description of the situations drivers experience. In this study, we proposed a data annotation protocol that covers many of the factors that might affect drivers and the drivers' responses. The annotation labels are comprised of four major groups: driver actions (e.g. facial expression, head position), driving environment (e.g. type of road, traffic density), vehicle status (e.g. turning, stopped), and speech / background noise. The annotation protocol designed in this research is comprehensive, and can be used in a wide

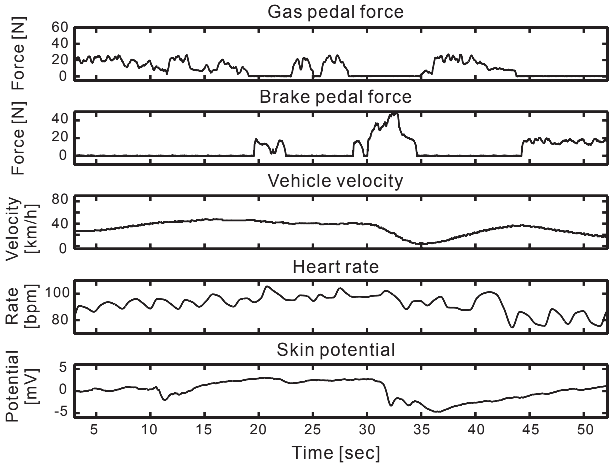

Fig. 4. Examples of collected signals.

range of research fields. We are currently annotating data from all drivers in our database. Data are being annotated in a similar fashion at Dallas (USA) and Istanbul (Turkey), what facilitates the consistent evaluation of technologies.

IV. Example of Multimedia Driving Signals Processing: Driver Frustration Estimation

In previous sections we described our multimedia corpus, collected data, and annotation protocol. In this section we briefly present an example of multimedia processing system designed to integrate information from different sources in NUDrive. The proposed system is devoted to the estimation of a driver's spontaneous frustration based on a context-dependent multimedia data fusion technique. Frustration, which is defined as the outcome of interferences with a goal-directed behavior, plays an important role in the driving context, since it is one the major sources of aggression [8][9]. As the number of in-vechicle devices increases, the need for intelligent interfaces also stress the relevance of frustration in driving. Together with interest, puzzlement, and boredom, frustration is critical in human-computer interaction, and recently it has been considered by many researches in this field [10]. An accurate estimation of drivers' emotional state can be used to increase safety and comfort, acting as a feedback for intelligent in-vehicle interfaces and adaptive safety systems.

A few attempts have been made to automatic recognize affect displays in in-car environments [11][12][13]. In the present study, the proposed 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. Methods on the estimation of a driver's emotions tend to oversimplify this model by, usually, disregarding the environment. The present approach is an extension of a pilot study described in [14], where the frustration estimation framework was outlined and results for three different drivers were presented.

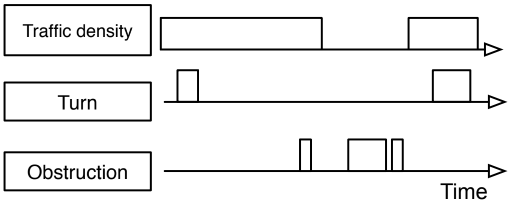

Fig. 5. Example of annotation labels.

A. Materials and Methods

Data recorded while drivers interacted with a speech recognizer (MR city) were utilized, so that not only the traffic but also the man-machine interaction could be regarded as natural sources of frustration. The annotation protocol was used to manually code data from 20 drivers. The following labels and possible states were used:

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).

Coders annotated the time span of labels, so annotation results can be seen as multiple streams of binary information, as shown in Fig. 5. Labels 1-5 were used as a way of describing the driving environment. In order to generate more consistent information, the frustration level—which was self-assessed as described in section II-B—was automatically quantized into two levels: frustrated and not frustrated. The quantization threshold was defined experimentally, being the one which provide the best overall estimation.

When recognizing emotions, the use of complementary information from various channels has proved to be superior to its single-modal counterpart, since the uncertainty due to one channel can be decreased by adding new information. In this study, together with the information obtained from annotation labels, the following data/features were combined:

- *Electodermal activity (EDA).* Electrodermal activity is one of the most widely used response systems in the literature. It is linked with psychological concepts of arousal and attention. Mean of normalized skin potential signal (mean skin potential) and the absolute value of the first-order difference of the normalized signal (Δ skin potential) were used as EDA features for each data frame.
- *Pedal actuation.* The investigation on the effects of different emotional states on the way we drive is an open and very interesting question. This study tackles this problem by trying to show that actuation is also affected by frustration. The force signal from gas minus brake was used as the pedal actuation signal. Features were extracted through spectral analysis of this signal by using a special feature called "cepstrum" (cepstral coefficients): a widely used spectral feature for speech and speaker recognition,

and, more recently, it proved to be effective in driver modeling [4]. Features are calculated for each data frame.

- *Speech recognition errors.* The incapacity 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 gets it right. As a possible indicator of speech recognition errors, we used the instants 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. Nevertheless, since this is a pinpoint feature that indicates an instant, an enlargement of its boundaries was necessary. An analysis of frustration videos showed that adding five seconds before and 15 seconds after each utterance recognized as "No" was adequate. 20 seconds is the time span in which significant verbal or gestural reactions still occured; accordingly, we encoded speech recognition errors as a binary signal, in which errors were indicated by 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.

The process that causes frustration is complex. Several uncertainties might be present in this process, and while driving, frustration can be regarded as the result of a wide range of contextual variables. To effectively estimate an emotional state, a system that integrates evidence from multiple sources in an efficient language is needed, and a Bayesian network (BN) is the natural choice to deal with such task. A BN is a state-of-art knowledge representation that creates a very efficient language for building models of domains with inherent uncertainty. Joint probabilities of a set of continuous or discrete random variables (nodes) are represented in a BN, which also explicitly encodes conditional independence assumption in its structure. The details on Bayesian networks can be found in [15], [16], and [17]. Figure 6 shows the proposed Bayesian Network. 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 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.

B. Experiments and Evaluation

In order to verify the effectiveness of the proposed method and features, experiments were performed with four different variations of the network in Fig. 6. The direction of arrows was kept fixed and the number of nodes was different depending on the variation: (1) *Basic:* without neither driving behavior nor speech recognition nodes; (2) *Pedal actuation:* the basic plus pedal actuation node; (3) *Speech:* the basic plus speech recognition errors; and (4) *Full:* with all nodes (Fig. 6).

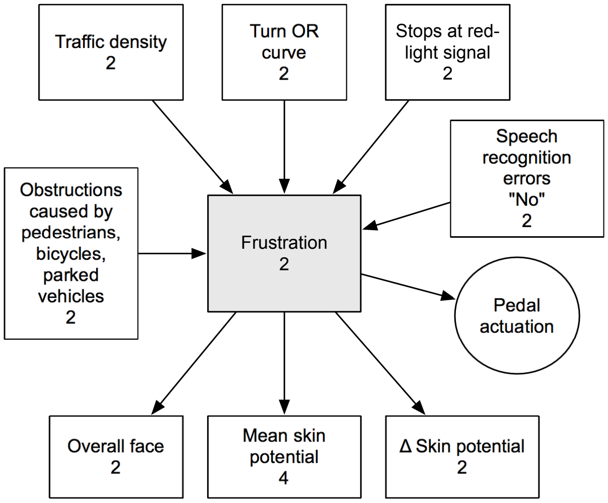
Traffic density
2
Turn OR curve
2
Stops at red-light signal
2
Speech recognition errors "No"
2
Obstructions caused by pedestrians, bicycles, parked vehicles
2
Frustration
2
Pedal actuation
Overall face
2
Mean skin potential
4
Δ Skin potential
2

Fig. 6. Proposed Bayesian network structure. Squares represent discrete (tabular) nodes and circle represents a continuous (Gaussian) node. Numbers represent the number of mutually exclusive states each node can assume. Experiments were conducted with four different variations of this structure.

Detailed experimental conditions and parameter values can be found in [14]. Individual networks were trained using 60% of data from each participant. The rest was used for test. During training, all nodes were filled with information, while during test, the *Frustration* node was empty and its posterior probability was inferred using data from all other nodes. Inferred probability was used as estimated frustration.

We evaluated the capacity of the proposed system to detect frustration. After calculating the estimation signal from each driver, it was filtered using a median filter of twelve seconds so that spikes and short gaps could be removed. In order to estimate the overall detection effectiveness, we added together true/false positives/negatives 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.

C. Results

Overall results for all four network variations are shown in Fig. 7. In the ROC space, the point (0,1) represents the perfect estimation. The closer the result gets to this point, the better. Circles centered in (0,1) are also shown so that different results can be easily compared. The *Full* network, in which both driving behavior and speech recognition results were used, achieved the best result: a true positive (TP) rate of 80% and a false positive (FP) rate of 9%, i.e., the system correctly detected 80% of the frustration, and, when drivers were not frustrated, it made mistakes 9% of the time. The *Basic* variation achieved the worse result, with a TP rate of 74% and a FP rate of 12%. Both of the proposed features, driving behavior and speech recognition errors, were effective in boosting the estimation. These are a encouraging results.

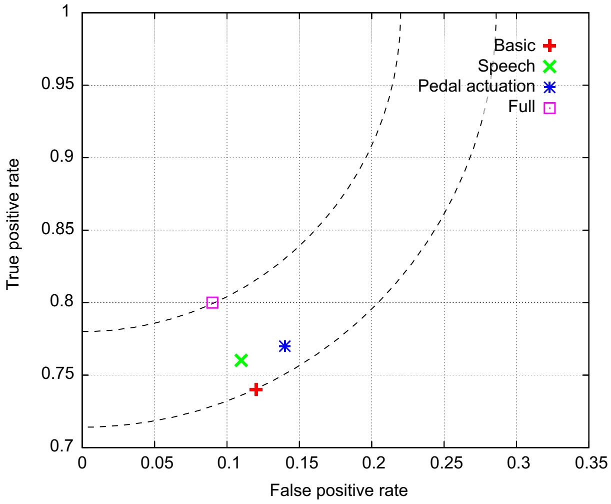

Fig. 7. Overall results achieved by the network in all for different configurations.

V. Summary and Conclusions

In this paper we described our multimedia corpus of driving behaviors (NUDrive). Data from 250 drivers have already been collected. We overviewed the data collection apparatus, data annotation, and offered basic statistics on collected data. With NUDrive we are able to test theoretical concepts and evaluate anticipated systems using genuine human signals under real-world conditions, that is, real instrumented car on real city streets and highway roads. Part of the data collection is performed under international collaboration with universities in USA and Turkey. All collaborating partners use similar sensors and a coherent data collection scenario—an important step toward more general models of driver behavior. A sample of collected data in Japan, USA, and Turkey can be downloaded from the DriveBest website[1].

We also presented a multimedia processing system, which uses information from various sources in NUDrive to implement a context-dependent estimation of a driver's spontaneous frustration. Results stressed the importance of using multimedia data in order to effectively model a driver.

Acknowledgment

This research is supported by Strategic Information and Communications R&D Promotion Program of MIC Japan.

References

[1] L. Angell *et al.*, "Driver workload metrics project task 2, final report," National Highway Traffic Safety Administration (NHTSA), Tech. Rep. DOT HS 810 635, 2006.

[2] S. Y. Cheng and M. M. Trivedi, "Turn-intent analysis using body pose for intelligent driver assistance," *IEEE Pervasive Computing*, vol. 5, pp. 28–37, 4 2006.

[3] J. Dressel and P. Atchley, "Cellular phone use while driving: A methodological checklist for investigating dual-task costs," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 11, no. 5, pp. 347–361, 2008.

[4] 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.

[5] J. McCall and M. Trivedi, "Driver behavior and situation aware brake assistance for intelligent vehicles," *Proceedings of the IEEE*, vol. 95, no. 2, pp. 374–387, 2007.

[6] M. Helander and B. Hagvall, "An instrumented vehicle for studies of driver behaviour," *Accident Analysis & Prevention*, vol. 8, no. 4, pp. 271–277, Dec. 1976.

[7] 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, Feb. 2004.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] L. Malta, P. Angkititrakul, C. Miyajima, and K. Takeda, "Multi-modal real-world driving data collection, transcription, and integration using Bayesian network," in *IEEE Intelligent Vehicles Symposium*, 2008, pp. 150–155.

[15] 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

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

[17] F. V. Jensen, *Bayesian networks and decision graphs*, M. Jordan, Ed. Springer, 2001.

[1] http://www.sp.m.is.nagoya-u.ac.jp/NEDO/Drive-Best/