Information Society Technologies (IST) Programme

AIDE
IST-1-507674-IP

Evaluation of the AIDE demonstrators

Dissemination Level PP

Deliverable No. (use the number indicated on technical annex)		D2.4.1	
SubProject No.	SP2	SubProject Title	Evaluation and assessment methodology
Workpackage No.	WP2.4	Workpackage Title	Prototype Evaluation
Activity No.	A2.4.1	Activity Title	Comparison of results and final evaluation report
Authors (per company, if more than one company provide it together)		Björn Peters (VTI), Emma Johansson (VTEC), Rino Brouwer (TNO), Enrica Deregibus, Elisabetta Nodari (CRF), Cristiana Gambera (KITE), Maria Alonso, Juan Plaza, Henar Vega (CIDAUT)	
Status (D: draft, S: Submitted to EC, F: Final accepted by EC):		S	
File Name:		AIDE D2.4.1 v4.doc	
Date		31 October 2008	
Project start date and duration		01 March 2004, 50 Months	

Executive summary

The three demonstrators developed in AIDE (City Car (SEAT), Luxury Car (CRF) and Heavy Truck (VTEC)) were evaluated in WP2.4. The evaluation was carried out by CIDAUT, TNO and VTI in collaboration with SEAT, CRF and VTEC. Three different test sites were used namely Valladolid (Spain), Torino (Italy) and Gothenburg (Sweden). The evaluations were based on recommendations in the AIDE Cookbook (see D2.1.4, Jansen et al., 2007) and an overall evaluation plan (Peters, Brouwer & Markkula, 2008). The implementations of the AIDE system were done somewhat different in the individual demonstrators even if the basic functions were the same, i.e., the tested HMIs were both integrated and adaptive according to the DVE (Driver Vehicle Environment) state. The functional difference between AIDE and Non-AIDE was that the Non-AIDE was neither adaptive nor integrated. The main hypothesis to be tested was that driving with the AIDE system should be safer and better accepted by the drivers compared to the Non-AIDE condition. The evaluation results were also used to provide feedback to the AIDE Cookbook.

The three evaluations were harmonised as much as possible. Thus, a within subject design with three driving conditions (Baseline, AIDE and Non-AIDE) and common inclusion/exclusion criteria for subjects were used. The recommendation was to use a minimum of 18 subjects. Both objective and subjective dependent measurements were applied according to the AIDE Cookbook. However, the used measures differed between the three test sites mainly due to technical reasons (e.g. the lack of instruments).

Common objective measures were: Steering wheel Reversal Rate (SRR), and speed (mean, max and standard deviation (std)). Subjective measures consisted of a rating scale (RSME, Rating Scale Mental Effort) applied during driving and two questionnaires after each driving condition (DALI, (Driver Activity Load Index) and the CRF questionnaire - performance, usability, adequacy, aesthetics and comparison between AIDE vs. Non-AIDE).

Some additional measures were used in the truck evaluation: brake jerks, a modified version of SDLP (standard deviation of lateral position), mean TLC (time-to-line-crossings), TDT (tactile detection task) and eye gaze measures. These were applied based on the recommendation from the work performed in Workpackage 2.2.

Questions on car image and willingness to pay were included in the two passenger cars evaluations and subjective driving performance (in real time) in the evaluation of the luxury car. Different Use Cases (UC; specific tasks in certain situations with the objective to evaluate functions of the AIDE system) were specified in collaboration with the developers of the demonstrators. Tasks were repeated 2 – 3 times during the two experimental driving conditions in the truck and city car evaluation. Data were analysed in three ways: all UCs, UC by UC (with 2-3 repetitions[1]) and finally according to UC type according to a functional classification specified in the Overall evaluation plan. The drivers did not perform any tasks during Baseline condition. This condition was only used to collect reference date. Manual annotations by the test leaders were used to extract relevant driving behaviour data. The following results were achieved.

[1] The evaluation of the Luxury car did not include repetitions but all UC were unique.

The evaluation of the City Car was done with 18 subjects (9 female, 9 male) on urban, extra-urban and rural roads in Valladolid. Five different types of UCs were used in the test. In short, the UCs included incoming phone call which were postponed (in the AIDE condition) due to various reasons and enhanced warnings. Over all UC's the only significant difference found was that the mean and maximum speed was higher for Baseline compared to the two experimental conditions. When the UCs were analysed individually it was found that the enhanced warnings did not result in any differences in driving behaviour between AIDE and Non-AIDE. However, postponing the phone calls resulted in a significant difference in standard deviation of speed with Baseline>AIDE>Non-AIDE. Furthermore, the reaction time to warnings messages was shorter in AIDE compared to Non-AIDE. RSME and DALI revealed few differences between AIDE and Non-AIDE, only that time demands were lower in AIDE.

The evaluation of the Luxury Car was carried out with 18 drivers (8 female and 10 males) on urban, extra-urban and rural roads in Orbassano near Torino. Seventeen UCs were presented on different locations along a predefined route. In both AIDE and Non-AIDE conditions the HMI aspect and way of interaction was the same. The difference was that in AIDE the information exchange was managed by ICA (Information Communication Assistant), while in Non-AIDE all information was provided immediately to the driver without any prioritization, filtering, modality and channel adaptation and independently by the DVE conditions. Both objective and subjective measures were used. There were hardly any effects with respect to driving behaviour measures. Most of the effects found referred to RSME and the self-rated driving performance (which was done after each use case). About equally often there was a difference between the AIDE condition and the Baseline condition and between the Non-AIDE condition and the Baseline condition. Even the differences between AIDE and Non-AIDE were equally divided. The DALI questionnaire provided the same result with the exception that the drivers found the auditory demands to be higher for the AIDE condition compared to Non-AIDE. The replies to the questions related to perceived usability seemed to favour the AIDE system. The Non-AIDE system more often showed a deviation to the negative side of the scale then the AIDE system while the AIDE system more often showed a deviation to the positive side of the scale. The adequacy of the HMI in terms of acoustic, visual, and input (i.e. haptic barrel key, HBK) was assessed with a number of questions and it was found that the drivers were quite positive to the HMI. In general, most participants did not really report a difference between AIDE and Non-AIDE. As there were only few differences found it was difficult to translate the results into risk. There were only differences found with respect to workload while generally the workload was still low in the AIDE and Non-AIDE condition. However, it seemed that risk went up with both AIDE and Non-AIDE compared to Baseline which is to be expected as they performed the UCs. Anyhow, it should be noted that measures like headway and lateral control were not included.

The evaluation of the Heavy Truck demonstrator developed by VTEC was performed with 21 professional truck drivers on urban and inter-urban roads in Gothenburg. Eight different Use Cases with two repetitions were carried out during the AIDE and Non-AIDE conditions. Three Use Cases included the use of Nomadic devices, two concerned rescheduling of messages and three concerned enhanced warnings. A more extensive set of objective measures were used e.g. lateral control, gaze, and TDT (Tactile Detection Task). However, no headway measures were included. The same set of subjective measures as for the other evaluations were used. In general it was found as expected that Baseline induced less workload and performance was better compared to both AIDE and Non-AIDE. For SRR (Steering Wheel Reversal Rate) it was found that performing the Use Cases increased visual workload in total and for 5 out of 8 Use Cases.

However, a clear distinction between AIDE and non-AIDE was only found when "making a voice controlled phone call vs. using a hand-held phone". Lateral control (MSDLP) was improved when using voice controlled phone in the AIDE condition compared to hand-held in non-AIDE condition. Analysis of speed data showed that drivers tended to decrease speed during task performance. It was also found that receiving an incoming phone call with the AIDE solution affected speed control less than with a nomadic phone due to added visual load. The analysis of the TDT data showed that e.g. delaying a phone call so that it did not interfere with the driving task (intersection) caused lower workload. Analysis of eye gaze data showed a clear benefit of the AIDE system as the drivers kept their eyes more on the road compared to driving with the Non-AIDE system. Subjective workload was assessed with RSME and DALI questionnaire. It was found that workload (RSME) was lower when an incoming phone call was delayed. However, DALI did not reveal any differences between AIDE and Non-AIDE. Analyses of the questionnaires showed no clear indications for a preference for AIDE. Most likely due to that the questions were asked after each drive and not after each UC. In a final set of questions the drivers were e.g. asked to state which system they preferred (i.e. which of the two drives). The drivers were of course not aware which drive was the AIDE or non-AIDE condition). More than 60% of the drivers preferred AIDE.

In summary, several objective measures clearly showed higher workload level for AIDE and Non-AIDE conditions compared to Baseline. Furthermore, AIDE compared to Non-AIDE contributed to improve the situation generally (eyes on the road) and in specific situations (handling phone calls). However, the subjective measures did not contribute very much to identify differences between AIDE and Non-AIDE. This might have been due to small differences between the experimental conditions (e.g. warning messages – display/sound vs. display/voice). However, it seems like the drivers in general were positive to the AIDE concept. Furthermore, it should be noted that the evaluation could only address short-term effects which could have contributed to the lack of very clear results. Even if the drivers were allowed to train with the system until they felt comfortable with the new functions they were more or less beginners. Thus, the results achieved should not be considered as a failure of the AIDE concept rather there are some indications that it rather contributes to improve safety and was in general received positively by the test drivers. However, it is clear that there is a need for further development and evaluation during an extended period.

The following recommendations were made for the Cookbook based on the results from the three evaluations:

- Consider the need for objective measures in traffic situations when ordinary driving behaviour measures are not applicable.
- When messages are delayed due to e.g. high workload it is important to be able to trace after the evaluation exactly what has happened, when a task starts and ends, when messages are presented. This should be addressed in the cookbook
- Make Gaze measures mandatory or at least highly recommended
- Make TDT mandatory and include guidelines as presented in previous AIDE work (Merat, 2004)
- Extend the description of use and interpretation for RSME and DALI in the cookbook and consider a more simple rating scale such as the one used in HASTE (Östlund, 2004)
- Consider the possibility of partial loss of data when specifying recommendations for number of subjects

- Consider the possible problems with too many tasks (Use Cases) in a test situation and how to administrate questionnaires in the guidelines – possibly provide at least tentative recommendations on how many should be included
- There might be a need to consider possible problems when evaluating HMI with rather subtle functional differences, e.g. spoken vs. sound warnings. It could be problematic to identify possible (highly relevant) differences in limited field tests.