PREDICTING SECONDARY TASK INVOLVEMENT AND DIFFERENCES IN TASK MODALITY USING FIELD HIGHWAY DRIVING DATA

Alina Sinelnikova[1,2], Joonbum Lee[1], Bryan Reimer[1,*], Bruce Mehler[1],
and Joseph F. Coughlin[1]
[1]MIT AgeLab & New England University Transportation Center
Cambridge, Massachusetts, USA
[2] University of Augsburg, Germany
Corresponding author: reimer@mit.edu

Summary: This study examined differences in the impact of visual-manual and auditory-vocal based radio tuning tasks on field driving performance. Engagement in visual-manual tuning tasks were associated with higher steering wheel reversal rates than baseline driving. Both visual-manual and auditory-vocal based tuning tasks were associated with higher variances in speed maintenance compared to baseline driving. Models were built to utilize driving performance measurements as input to a classifier that aimed to distinguish between the three states (i.e., baseline driving, visual-manual tuning, and auditory-vocal tuning). Baseline driving could be classified from visual-manual tuning at an accuracy of over 99% and from auditory-vocal based tuning at an accuracy of 93.3%. Models could differentiate between the modalities at an accuracy of 75.2 % and between the three classes at an accuracy of 81.2%. Results suggest that changes in driving performance associated with visual-manual based tuning are statistically distinguishable from auditory-vocal based tuning. While not being free of visual-manual demand, tasks that involve auditory-vocal interactions appear to differ from visual-manual in how they impact driving performance.

INTRODUCTION

The understanding of how driver distraction influences motor vehicle crashes is evolving as new data continues to inform the underlying science to how drivers manage demand. It is well established that distractions to the driver may originate from vehicle systems, external objects and events, and activities of other occupants (Regan et al., 2009). The rapid proliferation of in-vehicle infotainment and other systems has heightened concerns about distraction. This may be due, in part, to the degree of off-road glances associated with many secondary tasks and the strong links between increased off-road glance behavior and increased risk of collision (Klauer et al., 2006; Victor et al., 2014). Voice-command interfaces have been promoted as a way of allowing drivers to keep their eyes on the road and hands on the wheel, decreasing driver distraction and allowing for safer operation of in-vehicle human-machine interfaces (Shutko et al., 2009). While some efforts have focused on the cognitive demands involved in voice control (e.g., Strayer et al., 2014), it is well established that production voice interfaces also often include visual-manual demands (Reimer et al., 2014) and are therefore best considered as drawing upon auditory-vocal-visual-manual and cognitive resources (Reimer et al., 2014).

Previous research has shown contradictory effects in comparing visual and cognitive secondary tasks (e.g., Horrey & Wickens, 2005). In some studies, visual tasks were associated with reduced velocity and increased lane keeping variability, while cognitive tasks were associated with

reductions in lane keeping variability (e.g., Engström, Johansson & Östlund, 2005). Other work has found that cognitive tasks can be as detrimental as visual tasks (e.g., Jamson & Merat, 2005). Many studies such as these rely on artificial tasks to induce demand and take traditional statistical approaches to analysis. More recent efforts investigate behavior at the individual's level and factor together multiple features through algorithms (e.g., Solovey et al., 2014).

The present study extends machine learning techniques to not only predict secondary task involvement, but also to differentiate between secondary task modalities using vehicle telemetry data. Predictive models were built using data from an on-road study that examined visual-manual and auditory-vocal radio tuning tasks with a production-level Human-Machine Interface (HMI).

METHODS

This study is a secondary analysis of a subset of data from a previously collected dataset (see Reimer et al., 2013 for complete details). A summary of key methods appear below.

Participants and Apparatus

The sample consisted of 60 active drivers equally distributed in two age groups (20-29 years and 60-69 years). Data collection was conducted in a 2010 Lincoln MKS equipped with a customized data acquisition system for time-synchronized recording of telemetry from the vehicle CAN bus. The study was approved by the local human subjects review board.

Tasks

Drivers were asked to complete a number of in-vehicle tasks using the vehicle's infotainment and communication system. This report focuses on an analysis of data from drivers' use of visual-manual controls and auditory-vocal controls for radio tuning. Comprehensive training was provided while stationary in a parking lot setting prior to assessment while driving. The visual-manual tuning task was equivalent to the "hard" tuning task employed in the Crash Avoidance Metrics Partnership (CAMP) Driver Workload Metrics project (Angell et al., 2006). It required four manual steps: (1) pressing the power / volume button to turn the radio on, (2) pressing the radio button, (3) selecting the FM1 or FM2 band button on a touch-sensitive display, and (4) rotating the tuning knob to locate a specified station. The auditory-vocal tuning task had the same functional goal as the visual-manual tuning task. However, it required only one manual operation (press a "push-to-talk" button on the steering wheel) to activate the voice-command system. The instructions for the four tuning tasks were:

***Visual-manual 1**: Your task is to turn on the radio, switch to FM2, and tune to 100.7*

***Visual-manual 2**: Your task is to turn on the radio, switch to FM1, and tune to 95.3*

***Auditory-vocal 1**: Your task is to turn the radio on using the push-to-talk button and requesting FM 100.7*

***Auditory-vocal 2**: Your task is to turn the radio on using the push-to-talk button and requesting FM 95.3*

Procedure

The driving portion of the study was conducted on RT I-495 in the greater Boston area. The portion of roadway utilized had three travel lanes in each direction and a speed limit of 65 mph.

The road was largely bordered by trees that also often divided traffic in the opposite direction. Half of the participants completed the voice tuning tasks while traveling south and the manual tuning tasks while traveling north. The other half started with the manual tuning tasks.

Data Preprocessing and Aggregation

Driver performance data (telemetry data), acquired directly from the vehicle's bus at a sampling rate of 10 Hz, were taken as input to create and evaluate predictive models. Subsequent sections describe details of feature generation for the models.

Steering wheel reversal rate. Steering wheel reversal rate measures the frequency of steering wheel reversals larger than a certain angle (so-called "gap") over a time period. A change of the steering wheel angle that is larger than the gap can be counted as one steering wheel reversal event. In this study, a gap size of three degrees was set as input parameter for the steering wheel reversal rate computation (see Östlund et al., 2005 for computational methods).

Normalized velocity. Normalized velocity is computed as the change in velocity relative to baseline driving. The baseline driving period is derived from a two minutes of data before task periods (i.e., the tuning tasks) (see Reimer et al., 2013 for details).

$$V_{normalized} = \frac{V_{raw} - \mu_{baseline}(V_{raw})}{\sigma_{baseline}(V_{raw})} \text{ (z-score function)} \quad (1)$$

Window. Windows with multiple aggregation functions were applied to the raw input signals (i.e., steering wheel reversal rate and normalized velocity) in order to create feature vectors. The window length was based on the completion time of each task period for each subject, thus allowing for the creation of one feature vector per subject and task period. Two task periods were related to each task modality (i.e., visual-manual and auditory-vocal). Three aggregation functions were implemented to create features based on the normalized velocity: maximum, mean, and the first derivative (i.e., acceleration). For the steering wheel reversal rate just the mean aggregation function was implemented.

Rescaling function. Windowing resulted in features with a disparate range of values (which can be problematic for machine learning algorithms). Therefore, a rescaling function was implemented to move the features to a more comparable scale.

$$x_{normalized} = \frac{x_{raw} - \min(x_{raw})}{\max(x_{raw}) - \min(x_{raw})} \text{ (rescaling function)} \quad (2)$$

Classifiers

Two supervised machine-learning algorithms, Hoeffding Tree (HT) and Naïve Bayes (NB), were implemented. The Hoeffding Tree is a tree-based model that requires fewer training instances compared to regular decision trees (Bifet et al., 2010). Naïve Bayes Classifiers performs Bayesian prediction with the "naïve" assumption that all features are independent (Bifet et al., 2010). Naïve Bayes Classifiers are also known to perform well with smaller training sets and converge faster to its asymptotic error than discriminative models like logistic regression (Ng & Jordan, 2002). Furthermore, both classifiers (Hoeffding Tree and Naïve Bayes) can be used for binary and multiclass classification.

Model Evaluation

Blockwise k-fold cross validation was used to evaluate the models (Rumpold, 2014). This method operates by assigning feature vectors for every subject to blocks. This approach controls for within-subject differences that classical approaches to cross validation ignore, i.e., just split data randomly into training and test data without consideration of subjects. With k = 5 in our sample, 20% of the blocks were used for testing in every evaluation step.

Accuracy (Acc) and Kappa (Kap) were used as performance statistics. Accuracy is simply the percentage agreement between the classifiers' decisions and the actual task category. Kappa normalizes percentage agreement by subtracting the amount of agreement expected by chance. Table 1 shows ranges of Kappa values and corresponding interpretive performance levels.

Table 1. Kappa value interpretation (adapted from Altman, 1990)

Kappa value	Performance
0 – 20 %	Bad
21 – 40 %	Fair
41 – 60 %	Moderate
61 – 80 %	Substantial
81 – 100 %	(Almost) Perfect

Four features (mean velocity score, max velocity score, mean acceleration, and mean steering wheel reversal rate) were used to build the models. The models' predictive abilities were tested with four conditions: (1) visual-manual vs. auditory-vocal tuning periods, (2) visual-manual tuning vs. auditory-vocal tuning vs. baseline driving periods (3 classes), (3) auditory-vocal tuning vs. baseline driving periods, and (4) visual-manual tuning vs. baseline driving periods.

RESULTS

Figure 1 shows group level data for the four features that were used as model inputs across baseline driving, visual-manual tuning, and auditory-vocal tuning periods. ANOVA (one-way) was applied to test the effect of task type on the four features.

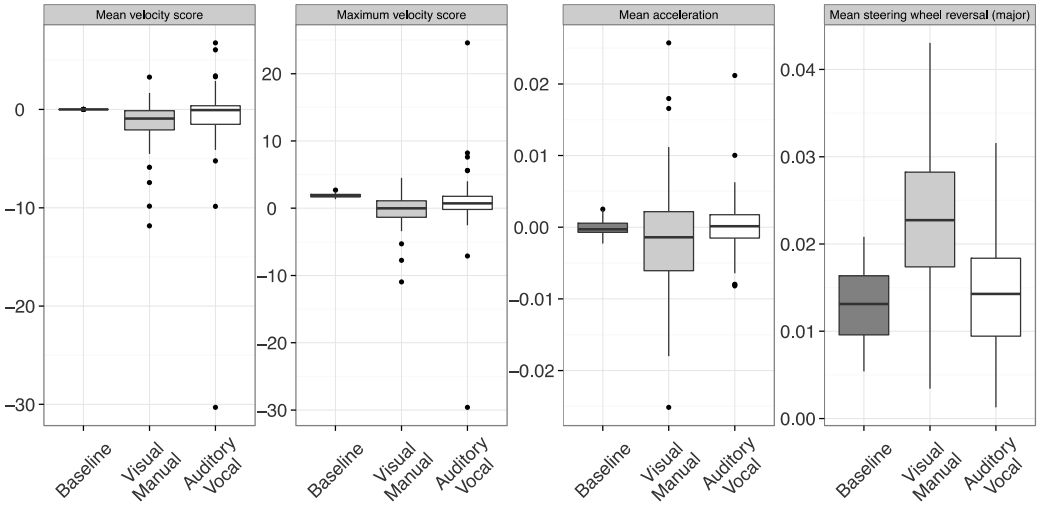

Figure 1. Feature distributions across task modalities and baseline driving

There were significant main effects of task types (e.g., baseline driving, visual-manual tuning, and auditory-vocal tuning), except on the mean acceleration [F (2, 174) = 3.38, $p < .05$ for the mean velocity score, F (2, 174) = 3.35, $p < .05$ for the maximum velocity score, and F (2, 174) = 11.6, $p < .001$ for the steering wheel reversal]. As a post-hoc test, paired t-test was applied to all combinations (e.g., baseline vs. VM, baseline vs. AV, VM vs. AV) from the four features. For the mean velocity and maximum velocity scores, there were significant effects of the task type between baseline driving and visual-manual tuning [t (59) = 4.39, $p < .001$, and t (59) = 6.51, $p < .001$]. For the mean steering wheel reversal, there were significant effects of the task type for visual-manual vs. auditory-vocal comparison [t (59) = -6.85, $p < .001$] and baseline driving vs. visual-manual tuning comparison [t (59) = -9.28, $p < .001$].

Figure 2 summarizes performance for each model and classification condition. All models performed better than chance and there was no remarkable difference in model performance between the two classifiers.

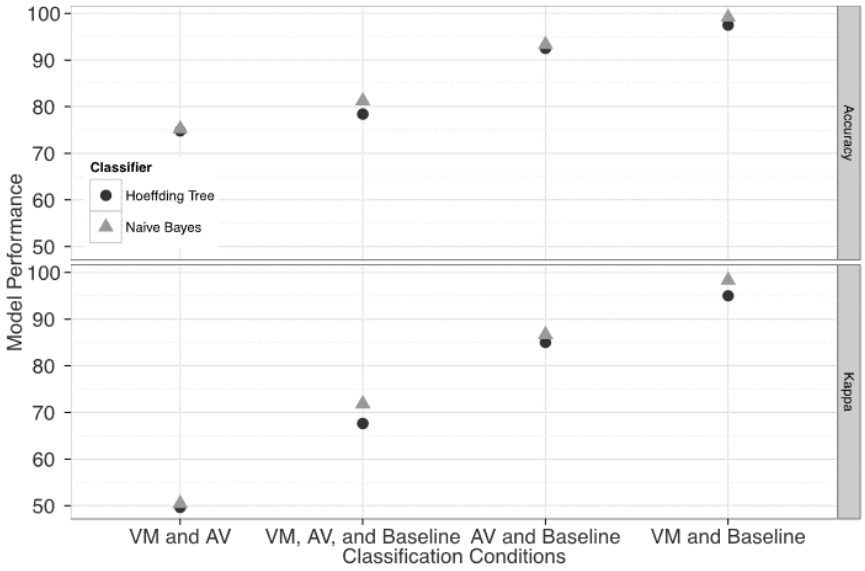

Figure 2. Model performance across all classification sets (note: upper panel shows accuracy measures and bottom panel shows Kappa measures)

Classification accuracy was the lowest when attempting to distinguish between visual-manual (VM) and auditory-vocal (AV) radio tuning periods (NB: Acc = 75.2%, Kap = 50.5%, HT: Acc = 74.8%, Kap = 49.6%), indicating moderate model performance. Classification performance was strongest when distinguishing between visual-manual tuning periods and baseline driving (NB: Acc = 99.2%, Kap = 98.3%, HT: Acc = 97.5%, Kap = 95.0%). This strong effect indicates that visual-manual tasks were easy to differentiate from driving without a secondary task (e.g., baseline driving). The auditory-vocal tuning pattern was also fairly distinct from baseline driving and resulted in high classification performance (NB: Acc = 93.3%, Kap = 86.7%, HT: Acc = 92.5%, Kap = 85.0%), though this is not as high as visual-manual vs. baseline. Interestingly, model performance for 3-class classification (e.g., VM, AV, and baseline driving) was also remarkably high (NB: Acc = 81.2%, Kap = 71.9%, HT: Acc = 78.4%, Kap = 67.6%).

DISCUSSION AND CONCLUSION

This study examined the impact of visual-manual and auditory-vocal radio tuning tasks on driving performance. Group level differences show that engagement in visual-manual tuning was associated with higher steering wheel reversal rates than baseline driving and auditory-vocal tuning. Visual-manual tasks led to decreased mean and maximum velocity relative to baseline driving, whereas auditory-vocal tasks did not show significant differences from baseline driving. The performance of two machine learning algorithms for classifying secondary task engagement from baseline driving and differences in task modality shows that differences in secondary task modalities were predicted with an accuracy of 75.2% and a Kappa value of 50.5%. Classification patterns suggest that visual-manual tuning has the most significant impact on driving behavior (based on the higher model accuracy compared to other conditions). Auditory-vocal tuning was fairly distinguishable from baseline driving. These results suggest that the impact of visual-manual tuning and auditory-vocal tuning on driving performance is statistically differentiable. Overall, these findings suggest that there are fundamental behavioral differences when drivers engage in visual-manual and auditory-vocal tasks (note that there might be other differences, such as task structure or task completion time, between visual-manual and auditory-vocal tuning tasks along with a task modality as the present study tested a production level HMI to secure face validity). While there is clear evidence for risk associated with visual demand/distraction during driving (Klauer et al., 2006; Victor et al., 2014) and strong support for the proposition that auditory-vocal interfaces can lower visual demand relative to visual-manual alternatives for accomplishing the same task (Reimer et al., 2014), it is unclear to what extent the reduction in visual demand associated with voice interfaces alleviates overall crash risk. Only naturalistic studies can establish a comprehensive understanding of crash risk, but with the absence of such data, voice-based interactions appear to have a lower degree of impact on overt driving performance than those that are more manual in nature. Furthermore, the results illustrate that additional research is needed to more fully consider the most salient parameters in which to measure the impact of multi-modal interfaces on driver behavior. While this study only tested one type of HMI task on one production-level system, future work aims to extend the types of tasks examined across a broader range of vehicles to assess generalizability.

ACKNOWLEDGMENT

Support for this work was provided by the New England University Transportation Center and the Toyota Class Action Settlement Safety Research and Education Program. The views and conclusions being expressed are those of the authors, and have not been sponsored, approved, or endorsed by Toyota or plaintiffs' class counsel. Data was drawn from studies supported by Toyota's Collaborative Safety Research Center (CSRC) and the Santos Family Foundation.

REFERENCES

Altman, D. G. (1990). *Practical Statistics for Medical Research*. CRC Press.

Angell, L., Auflick, J., Autria, P. A., Kochhar, D., Tijerina, L., Biever, W., … Kiger, S. (2006). *Driver workload metrics project task 2 final report*. Washington, DC.

Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (2010). MOA : Massive online analysis. *The Journal of Machine Learning Research*, *11*, 1601–1604.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 97–120.

Horrey, W. J., & Wickens, C. D. (2004). Driving and side task performance: The effects of display clutter, separation, and modality. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(4), 611-624.

Jamson, A. H., & Merat, N. (2005). Surrogate in-vehicle information systems and driver behaviour: Effects of visual and cognitive load in simulated rural driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, *8*(2), 79–96.

Klauer, S. G., Ramsey, D. J., Sudweeks, J. D., Neale, V. L., & Dingus, T. A. (2006). *The Impact of Driver Inattention on ear-crash/crash risk : An analysis using the 100-car naturalistic driving study data*.(No. HS-810 594).

Ng, A., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems, 14*, 841.

Östlund, J., Peters, B., Thorslund, B., Engström, J., Markkula, G., Keinath, A., … Foehl, U. (2005). *Driving performance assessment-methods and metrics (Report No. IST-1-507674-IP)*. Gothenburg, Sweden.

Regan, M. A., Young, K. L., Lee, J. D., & Gordon, C. P. (2009). Source of driver distraction. In *Driver Distraction: Theory, Effects, and Mitigation* (pp. 249–278). Boca Raton, FL: CRC.

Reimer, B., Mehler, B., Dobres, J., & Coughlin, J. F. (2013). The effects of a production level "voice-command" interface on driver behavior: reported workload, physiology, visual attention, and driving performance (Technical Report 2013-17A). MIT AgeLab.

Reimer, B., Mehler, B., Dobres, J., McAnulty, H., Mehler, A., Munger, D., & Rumpold, A. (2014). Effects of an “Expert Mode” voice command system on task performance, glance behavior & driver physiology. *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications,* pp. 1–9.

Rumpold, A. (2014). *Multi-dimensional classification of driver mental workload*. University of Augsburg. Thesis.

Shutko, J., Mayer, K., Laansoo, E., & Tijerina, L. (2009). *Driver workload effects of cell phone, music player, and text messaging tasks with the Ford SYNC voice interface versus handheld visual-manual interfaces* (No. 2009-01-0786). SAE Technical Paper.

Solovey, E. T., Zec, M., Garcia Perez, E. A., Reimer, B., & Mehler, B. (2014). Classifying driver workload using physiological and driving performance data: two field studies. *In Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 4057-4066). ACM.

Strayer, D. L., Turrill, J., Coleman, J. R., Ortiz, E. V., & Cooper, J. M. (2014). Measuring cognitive distraction in the automobile II: assessing in-vehicle voice-based interactive systems. Washington, D.C.

Victor, T., Bärgman, J., Boda, C., & Dozza, M. (2014). *Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk* (No.S2-S08A-RW-1).