Ergonomics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/terg20

Patterns in transitions of visual attention during baseline driving and during interaction with visual–manual and voice-based interfaces

Bryan Reimer, Bruce Mehler, Mauricio Muñoz, Jonathan Dobres, David Kidd & Ian J. Reagan

To cite this article: Bryan Reimer, Bruce Mehler, Mauricio Muñoz, Jonathan Dobres, David Kidd & Ian J. Reagan (2021) Patterns in transitions of visual attention during baseline driving and during interaction with visual–manual and voice-based interfaces, Ergonomics, 64:11, 1429-1451, DOI: 10.1080/00140139.2021.1930197

To link to this article: https://doi.org/10.1080/00140139.2021.1930197

Published online: 01 Jun 2021.

ERGONOMICS
2021, VOL. 64, NO. 11, 1429–1451
https://doi.org/10.1080/00140139.2021.1930197

Taylor & Francis
Taylor & Francis Group

ARTICLE

Patterns in transitions of visual attention during baseline driving and during interaction with visual–manual and voice-based interfaces

Bryan Reimer[a], Bruce Mehler[a], Mauricio Muñoz[a], Jonathan Dobres[a], David Kidd[b] and Ian J. Reagan[b]

[a]AgeLab, Center for Transportation & Logistics, Massachusetts Institute of Technology Cambridge, MA, USA; [b]Insurance Institute for Highway Safety, Arlington, VA, USA

ABSTRACT

Voice interfaces reduce visual demand compared with visual-manual interfaces, but the extent depends on design. This study compared visual demand during baseline driving with driving while using voice or manual inputs to place calls with Chevrolet MyLink, Volvo Sensus, or a smartphone. Mean glance duration and total eyes-off-road-time increased when using manual input compared with baseline driving; only eyes off road time increased with voice input. Confusion matrices developed with hidden Markov modelling characterise the similarity of glance sequences during baseline driving and while making phone calls. Glance sequences with the MyLink voice interface were misclassified as baseline driving more frequently than the other voice interfaces. Conversely, glance sequences with the Sensus and smartphone voice interfaces were more often misclassified as manual phone calling. Thus, the MyLink voice interface not only reduced the overall visual demand of placing calls, but produced glance patterns more similar to driving without another task.

Practitioner Summary: The attention map and confusion matrix methodologies provide ways of characterising similarities and differences in glance behaviour across secondary task conditions, complementing traditional temporally based metrics (e.g. mean glance duration, long duration glances) while addressing some of the limitations of total-eyes-off-road-time (TEORT) for comparing secondary task behaviour to baseline driving.

ARTICLE HISTORY
Received 22 February 2021
Accepted 9 May 2021

KEYWORDS
Driving safety; smartphones; glance measures; total eyes-on-road time; hidden Markov models

1. Introduction

Naturalistic driving research suggests that the performance of visual–manual activities with an electronic device beyond the core driving task are associated with a significant 5-fold increase in crash risk relative to driving with no distractions (Kidd and McCartt 2015). This finding mirrors a growing literature of laboratory, test track, and on-road investigations into the dangers of driving while engaged in secondary activities (for a review, see Caird et al. 2014). Such elevated risk can be attributed, at a minimum, to a reduction in the ability of drivers to appropriately detect and respond to emergent on-road events that require timely decisions, as well as subsequent hysteresis (as in Jansen et al. 2016). One of the most readily measurable indicators of such multi-tasking engagement is the orientation of the eyes; that is, to or away from the road. Time and transition based measures of eye movements as an indicator of visual attention allocation have a relatively deep literature, even when constrained to the context of driving (for an overview, see Muñoz et al. 2016). The most commonly-used strategies include measuring the time visual gaze is directed away from the forward road (NHTSA 2013), and the time that visual gaze is directed to a driver-vehicle interface (see DFTWG 2006) such as the in-vehicle radio, climate controls, etc. As aggregate gaze measures (e.g. mean single glance duration, total eyes-off-road time across a task, etc.), these approaches abstract visual allocation over time to provide estimates of higher-order driver behaviour, such as gaze strategy and attention allocation.

Aggregate gaze measures inherently discard valuable variance related to operator behaviour. By collapsing across epochs of time and across the complex patterns of transitions between regions of interest that are indicative of active building of situational

CONTACT Bryan Reimer, reimer@mit.edu AgeLab, Center for Transportation & Logistics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, E40-275, Cambridge, MA 02139, USA

awareness, aggregate measures are unable to describe the distribution of attention over time and space. A growing body of literature suggests that time (Seppelt et al. 2017; Lee et al. 2017) and spatial (Muñoz, Reimer, and Mehler 2015; Muñoz et al. 2016) considerations may provide deeper insight into drivers' allocation of attention in-situ in a framework that supports more cohesive management of scarce human resources (see also Yuan, Liu and Fu 2018). Certainly, in pursuit of understanding the strategies deployed by drivers in managing perception and action between multiple tasks, analysis of allocation of glances across the information-bearing regions associated with both driving and driver-vehicle interfaces and other tasks may provide deeper insight into attentional strategies deployed in various driving conditions. The present work extends upon earlier efforts that explored two promising methods that supplement aggregate gaze measures. The first, Attention Map glance transition matrices (Muñoz, Reimer, and Mehler 2015), provides a quantitatively based, qualitative approach for visualising differences and similarities in how glances are distributed across a range of locations during a period of interest. The second uses machine learning and hidden Markov models (HMM), to infer driver attention allocation from sequences of glances (Muñoz et al. 2016). Both approaches use glance behaviour to provide insight into attentional strategies.

1.1. Background

The extrapolation of interpretable driver behaviour data from low level time-series telemetry data (as discussed in He et al. 2012) is a task well suited for data visualisation and machine learning. Decomposing raw data to discrete states and transitional relationships is a long-standing approach used to understand complex time-series data. These approaches have recently been used in on-road data (Muñoz et al. 2016). In other efforts, He et al. (2012) used a HMM to infer long-term driving strategy by directly sampling short-term samples of driving behaviour such as braking force, accelerator pedal position, and steering wheel angle during events such as negotiating a turn and obstacle avoidance (see Khan and Lee 2019 for a review). Muñoz et al. (2016) inferred the use of different in-vehicle interfaces by directly sampling from glance region and transitions during single task driving and when interacting with the vehicle's radio using either the visual–manual interface or through a voice-based interface. In both studies, the HMM approach was very accurate, which suggests that these models can infer complex driver behaviours to a degree which extends beyond individual differences, that is, a single HMM was trained across different drivers for a given task. The assumptions from using HMMs in this scenario are thus twofold: (a). there are common, basic glance patterns that can be predicted in the first place, and (b). the gaze dynamics have an ergodic nature, that is, learning a task at two different points in time has little or no impact on the quality of the trained model. It has been proposed that abstraction layers over raw data may be necessary in achieving high recognition rates (Pentland and Liu 1999). Abstraction of raw eye movements into glance regions is an approach used by Muñoz et al. (2016).

In-vehicle interfaces often draw on many of the same attentional resources as driving itself (Wickens 2002), and so use of both voice-based and visual–manual interfaces concurrent with driving place demands on attentional resources and lead to some degree of elevated workload. Properly implemented voice systems are not as demanding as their visual–manual counterparts (Chiang, Brooks and Weir 2005; Dobres et al. 2016; Reimer and Mehler 2013; Mehler et al. 2016; Sawyer et al. 2017), but the exact impact of this still increased demand on driving performance and crash risk is still a matter of contention (Hancock and Sawyer 2015; Strayer et al. 2015; Mehler et al. 2016; Reimer et al. 2016; Li et al. 2021). What is objectively evident is that vocal interaction is paired with varying degrees of visual and manual interaction in many current production voice interfaces (Reimer et al. 2014; Sawyer et al. 2014; Mehler et al. 2016), creating hybrid systems that are perhaps best characterised as multimodal (Reimer et al. 2016). This reality should highlight the qualified nature of the often stated assertion that voice-recognition systems per se allow drivers 'to keep their hands on the wheel and eyes on the road'. On the other hand, it is largely an open question as to what extent off-road glance behaviour, when it does occur during interactions with such systems, is qualitatively equivalent to that which takes place during interaction with visual–manual interfaces.

Gaze measures do reveal that the way attention is allocated when interacting with various interfaces are not identical to the way it is allocated when just driving (Dobres et al. 2016; Mehler et al. 2016). However, the ways in which these metrics might be best interpreted are not always clear when considering voice-based interfaces (DFTWG 2006; NHTSA 2013). While off-road glance measures such as mean single glance duration and the percentage of long duration glances ($>$ second seconds) suggest advantages for some

voice-based interactions over available visual–manual interfaces, the total time the eyes are off the forward road is harder to interpret, particularly if guidelines developed for assessing visual–manual interfaces are applied. This is highlighted in longer duration voice-based tasks such as address entry into a navigation system where the total time that a driver's eyes are directed at a location other than on the forward road can easily exceed the NHTSA 12 s guideline for pure visual-manual interfaces (Reimer and Mehler 2013; Reimer et al. 2016; Mehler et al. 2016;).

As previously mentioned, Muñoz, Reimer, and Mehler (2015) and Muñoz et al. (2016) used an on-road dataset that provided samples of single task driving and periods when drivers also interacted with the vehicle's embedded radio using either the visual-manual interface or a voice-based interface to explore driver behaviour by analysing glances to multiple locations inside and outside the vehicle and the transition patterns between them. The results demonstrated how patterns of glances between locations inside and outside the vehicle when using voice-based interfaces more closely resembled glance patterns when drivers were 'just driving' and could be distinguished from glance patterns during visual-manual interactions with the vehicle's radio (Muñoz, Reimer, and Mehler 2015). Examining how driver gaze moves between multiple locations inside and outside of the vehicle while interacting with an interface and driving may supplement aggregate measures to characterise important aspects of driver behaviour. First, by showing the degree to which visual behaviour while using an embedded infotainment interface (or portable device such as a smartphone brought into the vehicle) is distinguishable from driving alone, the methods may arguably assess the impact of the interface relative to 'baseline' allocation of attention when drivers are not interacting with the device. Further, the approaches may provide deeper insight into the impact of various systems on attention allocation. Finally, accounting for glances to mirrors and other regions off the forward road yet relevant to the driving task (other than to the interface under study) may better indicate the driver's ability to maintain situational awareness when using multi-modal interfaces. Building on previous work of Muñoz, Reimer, and Mehler (2015) and Muñoz et al. (2016), the present study extends the same analyses to consider a different on-road dataset for which extensive methodological details are provided in earlier publications (Mehler et al. 2016; Reimer et al. 2016). The dataset consists of two different vehicles, a different secondary task (initiating a phone call from a saved list of phone contacts using either a visual-manual or a voice-based interface), and the pattern of glances when drivers perform tasks using the vehicles' embedded interfaces or the voice and visual–manual interfaces of a smartphone mounted on the dashboard.

2. Methods

2.1. Participants

The analysis sample consists of 80 participants equally distributed by gender and age across two vehicles (40 each vehicle). The age distribution ranged from 20 to 66 years with an equal number of participants distributed across four age groups (18–24, 25–39, 40–54, 55 and older), conforming with NHTSA's *(2013)* recommendations for the assessment of embedded in-vehicle systems (see Mehler et al. 2016 for detailed breakdown). Cases were only included in the analysis sample if usable driving performance, glance, and physiological data were available and the drive did not include heavy traffic, adverse weather conditions, or other characteristics at variance with relatively typical steady-state driving.

Participants were recruited through online and newspaper advertisements in the greater Boston area. Participants were required to have been licenced for a minimum of 3 years, self-report driving at least 3 times a week, and be in relatively good health for their age. Also, individuals were excluded if they self-reported being involved as a driver in a police-reported crash in the past year, were positive for any of a range of serious medical conditions (e.g. a major illness resulting in hospitalisation in the past 6 months, a diagnosis of Parkinson's disease, a history of stroke) or were taking medications that might impair their ability to drive safely under the study conditions (e.g. anti-convulsants, anti-psychotics, medications causing drowsiness). Recruitment procedures and the overall experimental protocol were approved by MIT's institutional review board, and compensation of $75 was provided.

2.2. Apparatus

A 2013 Volvo XC60 equipped with the Sensus infotainment system and a 2013 Chevrolet Equinox equipped with the MyLink infotainment system were used as study vehicles. No modifications were made to these production user interfaces. A Samsung Galaxy S4 smartphone, model SCH-1545 (released March 2013) running Android 4.3 (Jelly Bean), was paired to each vehicle's embedded system via the vehicle's Bluetooth wireless interface. The smartphone was attached to the centre stack of each vehicle using a commercially

available mount. The distance and angle of reach to the smartphone varied somewhat between the two vehicles due to differences in the available mounting surfaces (see Reimer et al. 2016). Five video cameras mounted in the vehicle interior provided views of the driver's face for primary glance behaviour analysis, the driver's interactions with the vehicle's steering wheel and centre console, the forward road (narrow and wide-angle images), and the rear road. Video data were captured at 30 Hz for the face and narrow forward road cameras and 15 Hz for the remaining cameras. Both vehicles also were instrumented with a data acquisition system to record vehicle telemetry data, driver speech and audio, and signals from a physiological monitoring unit. Analyses of these data are reported elsewhere (Mehler et al. 2016; Reimer et al. 2016)

2.3. Tasks assessed

The full experimental protocol included extended periods of single task driving (highway driving when participants were not also engaged with the vehicle or smartphone-based infotainment systems), and periods when participants were asked to engage in secondary infotainment system interactions while driving. The secondary tasks included using voice-command based interfaces to enter addresses into the embedded vehicle and smartphone navigation applications, and using voice-commands or the visual-manual interface to place phone calls to contacts stored in the smartphone with the embedded vehicle interface and the smartphone. A set of manual radio tuning tasks were also presented during the final portion of the drive.

Only data from the phone contact calling task were considered in this analysis. The contact calling task was completed in a similar manner with the MyLink and Galaxy S4 voice interfaces. After enabling the speech recognition system, the driver stated the word 'Call' followed by the contact name and number type (e.g. mobile, work), if applicable. In contrast, when using the Sensus system the driver had to state a series of context-specific voice commands to navigate system menus until the contact and number type was selected. When completing the contact calling task using the visual-manual interface with MyLink, the driver had to use a rotary knob and inset push button to make selections from option lists presented in system menus. The driver had to first access the phone subsystem, then the correct alphanumeric bin containing the target contact, and finally the contact name and number type if necessary. With Sensus, the driver used a rotary knob to scroll through the full list of contacts until reaching the appropriate contact name. The number type was selected in a subsequent submenu if necessary. Finally, when calling a contact using the Galaxy S4 visual-manual interface, the driver began by touching a Contacts icon on the device's home screen to access the phone list. Then the driver swiped the phone's touchscreen to scroll to the appropriate contact. A number type was selected on a subsequent screen if necessary.

Full details on the embedded vehicle interfaces are provided in Mehler et al. (2016) and on the smartphone interface in Reimer, Mehler et al. (2016). In brief, a phone list of 108 contacts was used for all phone calling tasks. Participants were asked to place two calls for contacts having a single phone number, followed by two calls for contacts having multiple numbers (e.g. home and mobile), for a total of four calling trials per interface, 16 total calls to place per participant. The contacts were the same across the manual and voice interface interactions so that any aspects/characteristics of a particular contact name that might influence relative difficulty were constant (e.g. alphabetic location). Similarly, tasks were structured so that the start point in a contact list and other aspects of task initiation were consistent across participants and tasks. Presentation order of the four methods of placing calls was randomised across the sample to control for order effects of calling method.

2.4. Experimental protocol, driving environment and analysis periods

After a structured interview was conducted to confirm eligibility, participants reviewed and signed an informed consent and received training on either the embedded vehicle interface or the smartphone that took place in the facility's parking lot. Half of the participants were trained on and interacted with the embedded interface during the first half of the experiment and half during the second half. Similarly, presentation of the voice and manual versions of each phone contact calling task types were randomised across participants in a counterbalanced design (Figure 1). During training, participants were encouraged to repeat tasks until they felt comfortable to proceed. The initial orientation/training period typically ranged between 15 and 30 min, with a mean of approximately 20 min. Training for the second half of the study took place in a parking lot of a highway rest area.

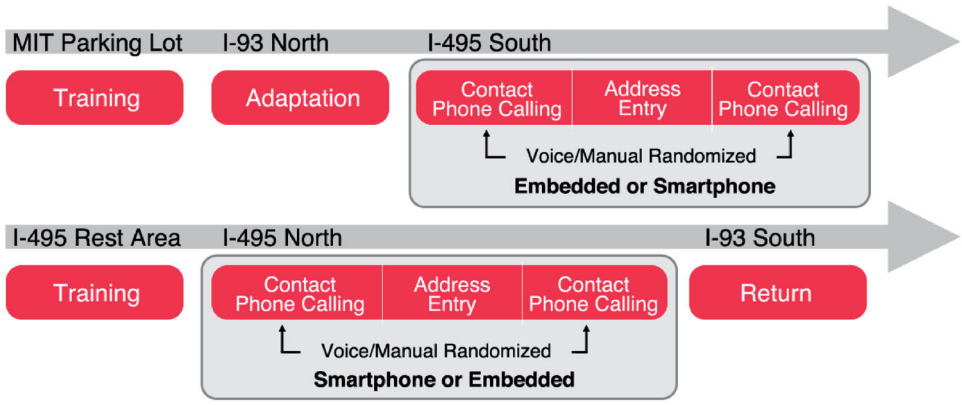


Figure 1. Schematic representation of the full experimental protocol. Half of the participants interacted with the embedded vehicle systems on I-495 South and half with the smartphone. Device type (embedded or smartphone) was reversed for the I–495 North segment so that all participants experienced both types (adapted from Reimer et al. 2016).

An adaptation period of approximately 30 min of driving took place prior to starting formal assessment, consisting of approximately 10 min of urban driving from MIT to interstate highway I–93 and 20 min driving north on I–93 to I–495. For the highway segments considered in the analysis, I–495 is a divided interstate surrounded largely by forest with three traffic lanes in each direction with lane widths of 15 feet (3.62 m). The posted speed limit is 65 mph (104.6 kph). The north and south segments of I–495 each took most participants approximately 35 to 40 min to drive (70 to 80 min total).

Baseline driving reference samples were collected during two minute periods immediately prior to a recorded audio message indicating that each new task period was about to start (Figure 1). Drivers were not given any secondary tasks during the baseline periods. Only the four baselines immediately prior to each phone contact calling period were included in the present analysis. For the phone contact calling periods, only intervals directly associated with placing a call were considered in the glance analysis; intervals during which recorded audio prompts cued participants on the contact to call and intervals between when one call was completed and the next prompt were excluded.

2.5. Annotation of visual behaviour and data processing

The glance annotation methodology used in this study is based on similar procedures developed by the Crash Avoidance Metrics Partnership (CAMP) Driver Workload Metrics project (Angell et al. 2006) and detailed in Smith et al. (2005). Specifically, two research assistants independently coded each driver's eye movements during phone contact calling trials and baseline driving periods by labelling a driver's glance targets according to one of the following glance regions: left blind spot, left (mirror/window), instrument cluster, forward road (road), rear-view mirror, centre stack, right rear passenger seat region (where an experimenter sat), right (mirror/window), right blind spot, other, and unknown. Annotators worked from a multi-image viewer that included the feed from the face camera, an over-the-shoulder view of the side of the driver's face, their hands, the steering wheel and the centre console region, as well as a forward view of the roadway for added context. Glance annotations were compared to check for discrepancies between the coders. A trial was considered discrepant if any of the following occurred: the coders started or ended their coding at different times; the coders described differing numbers of glances; the coders identified a different glance target for a glance; or the timing of a glance differed by more than 200 ms. A third coder resolved any discrepancies, making a 'final determination' as to which of the original two coders was correct.

A small percentage of task video was deemed unsuitable for coding, whether due to a momentary degradation of video quality (usually due to over- or under-exposure resulting from sudden changes in ambient lighting) or a transient movement of the participant's eyes beyond the frame of the video. Five percent of the task epochs under study contain at least some 'uncodable' segments, comprising 0.59% of the total video time coded for this analysis. Any such segments were removed from the data set prior to analysis.

2.6. Attention map visualizations

Muñoz, Reimer, and Mehler (2015) proposed a set of methods for considering and visualising the

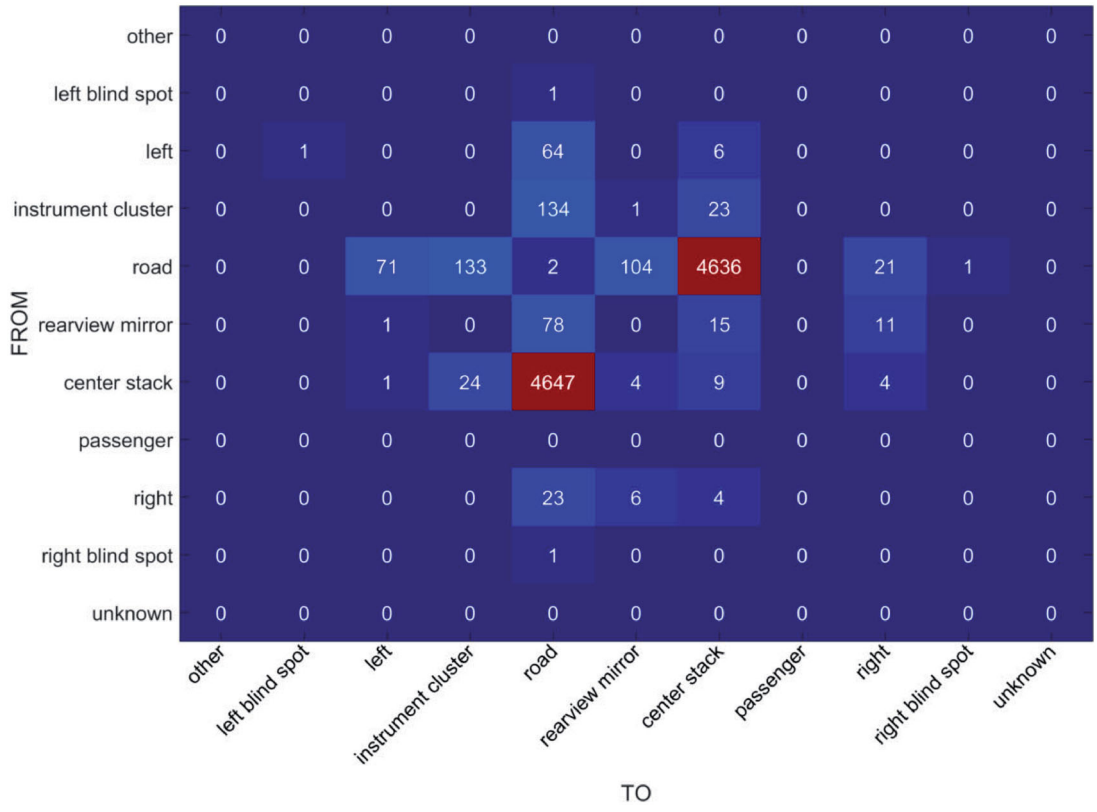


Figure 2. An attention map showing glance transition counts FROM (rows) and TO (columns) defined regions for the Chevrolet Embedded Manual Calling condition. Warm colours (e.g. red) indicate transition patterns of relatively high count, while cool colours (e.g. blue) are of lower count.

distribution of glance behaviour as flow of attention across defined regions of space. The most basic form of these 'attention maps' consists of a matrix representing simple *transition counts* (see Figure 2) of glances FROM one defined glance region (arranged as a row in the matrix) TO a different defined region (a column in the matrix). As described in the paper, the transition counts can be normalised into a *transition probability matrix* that indicates for any given glance region, the relative frequency that glances transition to another defined region. This is obtained by dividing the transition counts between regions (each cell) by its row sum (total number of transitions from a given location to all other locations). The resulting row sums in the new matrix all equal one. (See Appendix A for representative calculations and images of the primary new matrices described.)

One limitation of this probability matrix visualisation is that it fails to consider the overall significance of a transition from a particular region relative to the totality of glances observed. For example, in the dataset shown in Figure 2, the vast majority of glance transitions take place between the forward roadway and the centre stack and vice-versa. As detailed in Appendix A, if a glance originated on-road, the probability that it transitioned to the centre stack was very high (slightly greater than 93%). In contrast, in this dataset that considers glances in the Chevrolet Embedded Manual Calling task, one glance transition was observed to take place from the left blind spot to the road. As all the transitions from the left blind spot were to the road, the probability of transitioning from the left blind spot to the road was 100%. Without further normalisation across the entire matrix, this one transition gets outsized attention. To adjust for this, a *transition significance matrix* is computed. Starting again with the basic transition count matrix, this is accomplished by first dividing the transition counts between regions (each cell) by the maximum observed count (the cell with the largest number of transitions in the entire matrix) to create a *transition importance matrix*. Each cell in the transition importance matrix is then multiplied by the corresponding cell in the transition probability matrix to create its transition significance. In the example considered here (Figure 2), there are 4636 transitions from the road to the centre stack, a total of 4968 glances away from the road (row sum), and a total of 4647 glances from

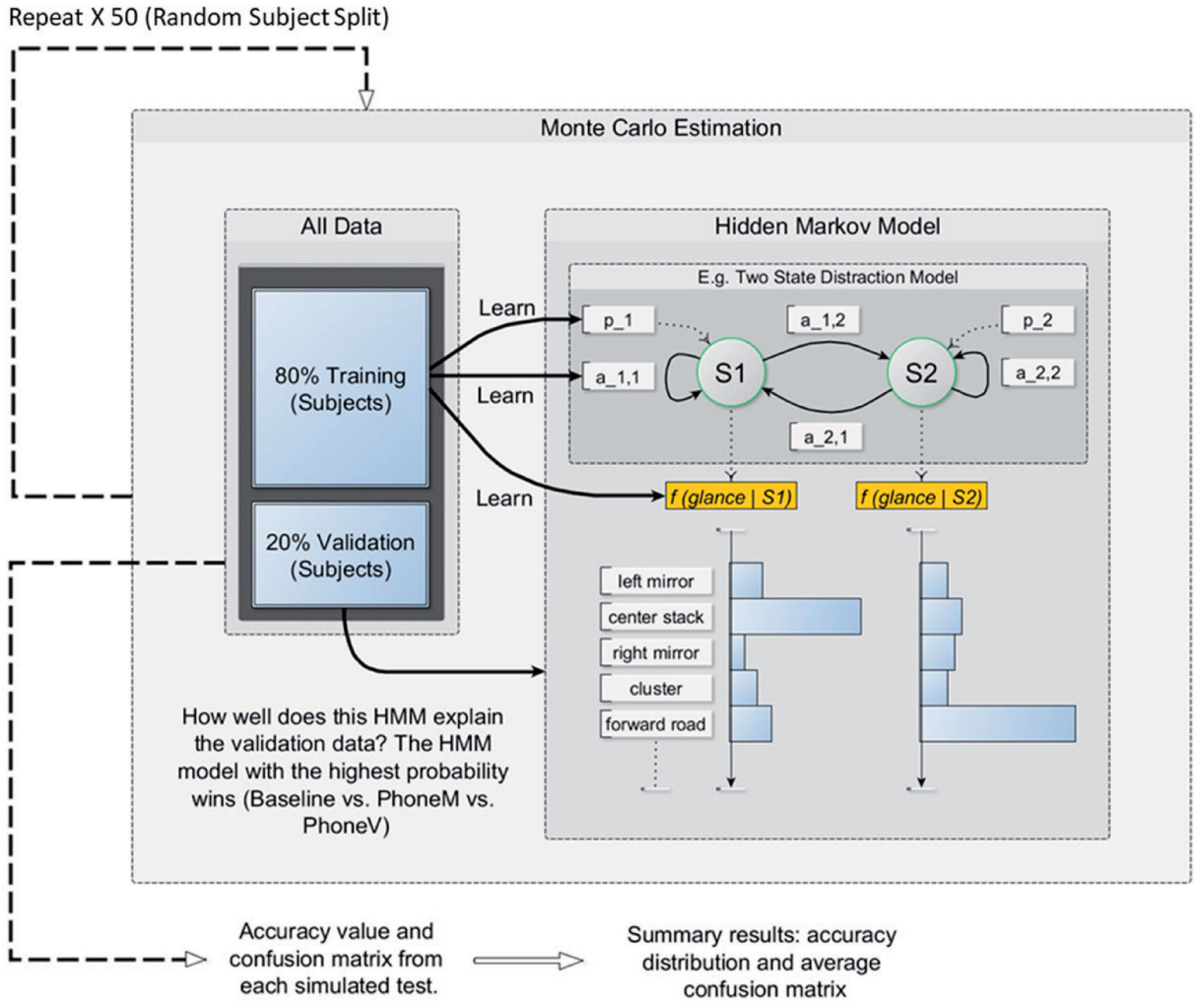


Figure 3. Monte Carlo accuracy approximation procedure. (See text for detailed explanation.).

the centre stack to the road (maximum number of transitions between any two regions). As such, the transition significance for movements from the road to the centre stack is computed as (4636/4968) * (4636/4647) = .93. In contrast, the significance value for the single glance between the left blind spot and the road becomes .00 in the resulting transition significance matrix (see Appendix A for a visual representation of this example and further consideration of the objective of the Transition Significance Matrix). These transition significance matrices are used for presentation in the results section. Across all matrices, colour codes (or shading) can then be used to visualise the most probable and frequent transitions (dark red) and least probable and frequent (dark blue).

2.7. Hidden Markov modelling

Hidden Markov models (HMMs) can be used to extract the statistical patterns of a sequence of glances to various glance regions for a set of participants performing a task. First, a model is trained for each task type (baseline driving, voice-based phone contact calling, and visual–manual phone contact calling). The trained models aim to learn how to describe a given type of task by the likelihood of glances between a set of regions by encoding probabilities of transition between a set of 'hidden' states. Note that the HMMs do not consider the duration of events and, thus, it is the sequence of glances between regions rather than the length of glances that is analysed. The likelihood that a separate sequence of glances (from a known type of task) is best associated with one of the trained models can then be assessed. The sequence is correctly classified when the trained HMM providing the best fit for 'new' data represents the same type of task. While statistical assessment of accuracy was not performed in the present work, model accuracy is here defined as the ratio of correct classifications, and confusion matrixes can be created to tabulate the number

of correctly classified sequences vs. incorrectly classified sequences for each task type.

The main parameter of the model is the number of hidden states with which to learn the behavioural transitions of visual demand. Muñoz, Reimer, and Mehler (2015) found two state HMMs maximised classification performance in prediction of baseline driving versus voice-based and visual-manual interactions with a radio and was employed in this study. A second model parameter is the minimum number of glances in a sequence (epoch of baseline period or individual task trial) necessary to be included in the analysis. Empirical assessment of the dataset suggested that sequences of 6 or more glances (minimum of 5 transitions) offered the best ratio of performance and quality-of-data, and was used in this analysis (see Appendix B for a detailed consideration). Sequences with fewer than six glances (5 transitions) were discarded from the analysis; otherwise, they were left at their natural length.

Figure 3 summarises the methodology used to develop the accuracy measure for the approach described above. From the pool of participants, 80% were randomly sampled as training cases, with the remaining 20% serving as validation (testing) cases. As an example, for a given vehicle's embedded interface, the maximum potential pool of 400 glance sequences (40 participants x 10 periods (i.e. 4 voice-based contact calling tasks + 4 manual contact calling tasks + 2 baseline periods)) results in a potential of 320 training sequences and 80 validation sequences. However, the number of actual sequences was limited by the minimum successive glance count discussed in the preceding paragraph. All glance sequences meeting the minimum length requirement (5 or more transitions) for the randomly selected training cases were used to train one HMM for each type of task (baseline driving, voice phone contact calling, and visual-manual phone contact calling) using the Chevy MyLink, Volvo Sensus, and the smartphone. Next, the trained HMMs were used to model the glance allocation sequences for the validation data. The number of correctly classified sequences were counted and used to compute accuracy. This procedure was repeated a total of 50 times to reduce bias associated with the random assignment of participants, resulting in an accuracy distribution (shown in the Results as Figure 5) rather than simply a single value. The selection of 50 repetitions was empirically chosen to ensure a fair amount of shuffling, while at the same time keeping total training and validation times within reason. This Monte Carlo accuracy approximation was repeated for every combination of vehicle and type of interface (embedded vs. smartphone), producing a total of 6 accuracy distributions for comparison.

3. Results

Statistical analyses were performed in R (R Core Team 2014) and an alpha level of 0.05 was used for assessing statistical significance. Owing to the non-normal distribution of the data and/or the use of ratio data (percentages) for several dependent measures, non-parametric statistics were calculated using the Friedman test (X^2) and Wilcoxon signed rank test (V), similar to the repeated-measures ANOVA and t-test, respectively. For multifactorial analyses, repeated-measures ANOVA by ranks are presented. These tests have been shown to be more robust against Type I error in cases where data are non-normal (Friedman 1937; Conover and Iman 1981).

3.1. Characterisation of glances in terms of duration off the forward road scene

NHTSA (2013) guidelines for assessing visual-manual in-vehicle electronic devices evaluate visual demand based on the mean duration of a single glance away from the forward road, the percentage of the total number of glances longer than 2.0 s, and the total duration of glances away from the forward road. These measures were used to examine the differences in glance behaviour between 'just driving' baseline behaviour and when drivers used voice commands or a visual-manual interface to select and place calls from a phone contacts list.

The average single glance duration and average percentage of glances away from the forward road that are longer than two seconds during baseline periods, voice-based phone calling, and visual-manual phone calling by vehicle and device are shown in Table 1. (Note that for the percentage of long duration glances columns, a value less than 1 indicate that, on average, less than 1% of the glances per participant had durations longer than two seconds.) Analyses revealed significant main effects across conditions (baseline, voice, and manual) for mean single glance duration (X^2 (2) = 119.28, $p < 0.001$) and the percentage of long duration glances off the forward road (X^2 (2) = 84.38, $p < 0.001$). *Post hoc* pairwise comparisons showed that mean single glance durations were, on average, significantly longer during interactions with visual-manual interfaces than during baseline driving (V = 0.00, $p < 0.001$) and during interaction

with voice-based interfaces ($V = 3239$, $p < 0.001$); no significant difference was observed between baseline driving and during interactions with voice-based interfaces ($V = 1262$, $p < 0.086$). Similarly, *post hoc* pairwise comparisons showed that the percentage of long duration glances were, on average, significantly greater during interactions with visual–manual interfaces than during baseline driving ($V = 7$, $p < 0.001$) and during interactions with voice-based interfaces ($V = 1696$, $p = 0.001$); no significant difference was observed between baseline driving and during interactions with voice-based interfaces ($V = 221$, $p = 0.603$).

Collapsing across embedded HMIs and smartphone interactions, there was a main effect across conditions (baseline, voice, and manual) on the total amount of time the driver's eyes were away from the forward road (X^2 (2) = 44.17, $p < 0.001$). Post-hoc pairwise comparisons indicated that the total eyes off road time (TEORT) was significantly longer when drivers used a visual–manual interface than a voice-based interaction ($V = 3011$, $p < 0.001$). However, comparison of the TEORT values for the two interface modes to baseline driving is problematic due to fact that the two-minute baseline period represents a somewhat arbitrarily selected duration relative to the variable length of the actual HMI task periods. Considering the baseline period as a task, pairwise comparisons show TEORT for manual interface interactions being significantly greater with a mean of approximately 14.8 s (for periods averaging 28.8 s) than the mean TEORT of approximately 12.7 s for the 120 s baseline periods ($V = 986$, $p = 0.002$). Conversely, the mean TEORT of approximately 10.1 s (for periods averaging 41.9 s) for the voice-based interactions is significantly less, on average, in a pairwise comparison than mean TEORT for the baseline driving periods ($V = 2329$, $p < 0.001$).

The right most columns in Table 2 represent one approach to dealing with the unequal time intervals, specifically by normalising off-road glance time as a percentage of the duration of individual task periods. Using these values shows an overall main effect of condition (X^2 (2) = 154.22, $p < 0.001$) and all pairwise comparisons are significantly different: manual vs. voice ($V = 3240$, $p < 0.001$), manual vs. baseline ($V = 0$, $p < 0.001$), and voice vs. baseline ($V = 36$, $p < 0.001$). These data highlight that the visual–manual HMI interactions, on average, drew glance orientation away from the forward road for approximately 50% or more of the duration of the task. The voice-based HMI interactions resulted in the eyes being directed off the forward road for approximately 23% of the task period vs. approximately 11% of the time during baseline driving. While this normalisation may, in some ways, provide a more appropriate metric for quantifying visual demand of each type of HMI interaction to baseline driving than TEORT values based on an arbitrary length sample of baseline driving, observation of drivers engaging with the HMIs suggested that there are additional aspects of how drivers allocate their glance behaviour during the task periods that are relevant in comparing these modes of interaction to baseline driving.

3.2. Attention map visualizations

Figure 4 shows a set of glance transition significance matrices comparing glance patterning across the two

Table 1. Mean and (standard error) of 'off-road' (away from the-forward-road) glance metrics for mean single glance duration (in seconds) and percentage of glances that are longer than two seconds during baseline (Base), voice-based (Voice), and visual-manual (Man.) phone contact calling.

	Mean Single Glance Duration			% Long Duration Glances		
Interface	Base	Voice	Man.	Base	Voice	Man.
Volvo Embedded	0.70 (0.2)	0.79 (0.1)	0.94 (0.2)	0.70 (2.3)	0.75 (1.9)	3.39 (5.6)
Chevrolet Embedded	0.66 (0.1)	0.60 (0.2)	0.92 (0.2)	0.22 (0.7)	0.29 (1.5)	2.46 (3.7)
Volvo Smartphone	0.67 (0.1)	0.71 (0.2)	1.01 (0.2)	0.30 (1.1)	0.52 (1.5)	4.18 (6.2)
Chevrolet Smartphone	0.65 (0.1)	0.67 (0.2)	1.00 (0.2)	0.31 (0.9)	0.36 (0.9)	3.81 (4.5)

Table 2. Mean and (standard error) of total eyes off road time (TEORT) ('off-road' = away from the forward road), task duration, and TEORT values normalised as a percentage of period under study: baseline (Base), voice-based (Voice), and visual-manual (Man.) phone contact calling.

	Total 'Off-Road' Glance Time (seconds)			Task Duration (seconds)			% 'Off-Road' Time		
Interface	Base	Voice	Man.	Base	Voice	Man.	Base	Voice	Man.
Volvo Embedded	12.64 (6.0)	10.22 (4.3)	16.39 (6.6)	120	38.17 (6.9)	32.90 (12.8)	10.56 (5.1)	26.23 (9.6)	49.71 (10.3)
Chevrolet Embedded	13.73 (7.8)	3.33 (2.6)	13.63 (5.3)	120	21.61 (8.3)	26.24 (7.6)	11.57 (6.5)	13.95 (7.2)	52.02 (10.9)
Volvo Smartphone	11.91 (7.5)	12.98 (7.4)	14.39 (5.8)	120	52.02 (12.9)	28.67 (9.8)	10.25 (6.3)	23.67 (9.9)	50.86 (11.3)
Chevrolet Smartphone	12.44 (6.4)	13.70 (7.1)	14.76 (5.7)	120	55.83 (16.0)	27.45 (9.7)	10.48 (5.4)	26.23 (9.6)	54.45 (11.0)

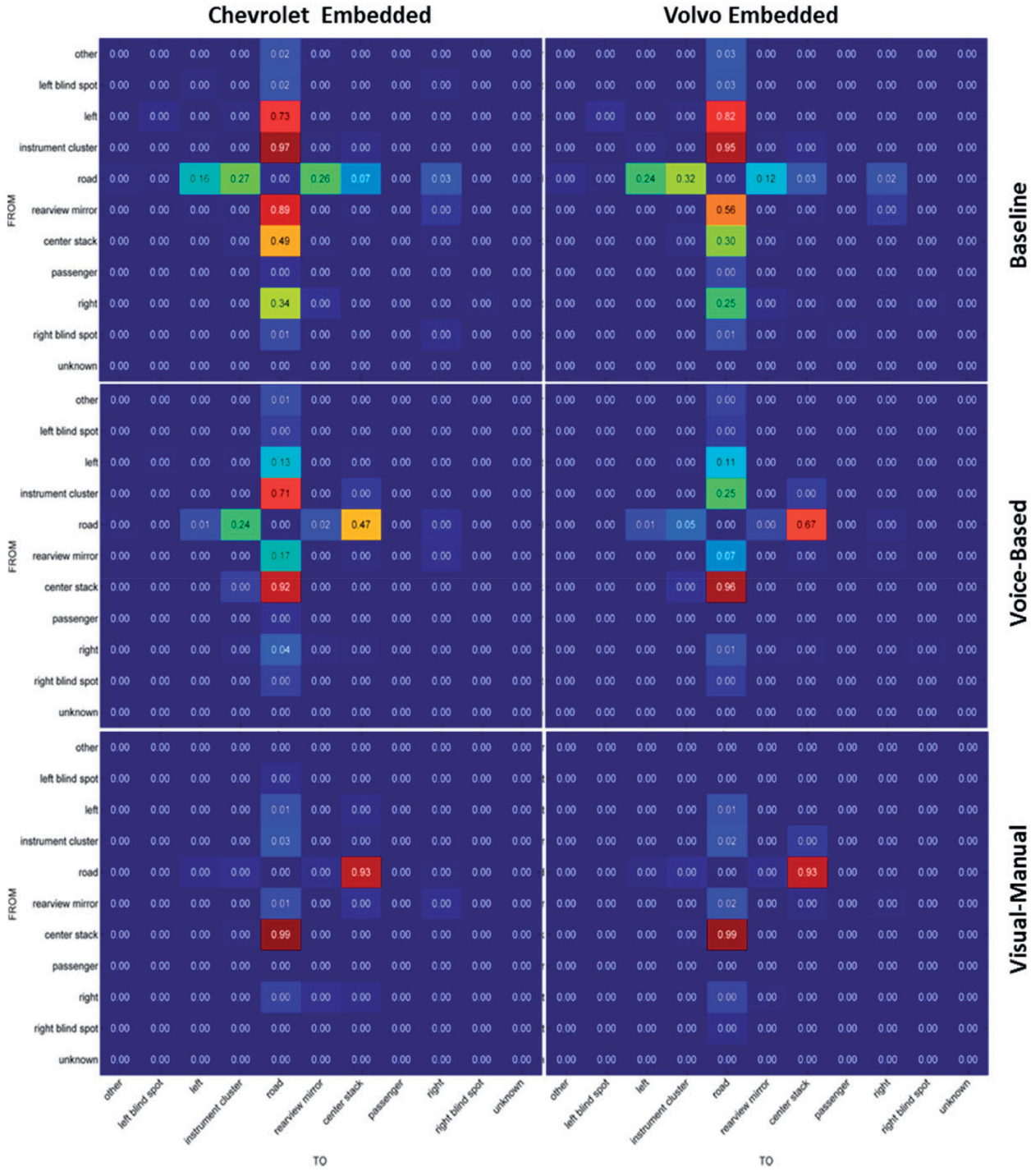


Figure 4. Glance transition significance matrices for baseline driving (top), voice-based phone contact calling tasks (middle), and visual-manual phone contact calling tasks (bottom) using the Chevrolet (left) and Volvo (right) embedded systems. Warm colours (e.g. red) indicate transition patterns of relatively high probability and frequency, while cool colours (e.g. blue) are of lower probability and frequency.

vehicles during baseline driving, embedded voice-based phone contact calling tasks, and embedded visual-manual phone contact calling tasks. As described in the methods, high-intensity cells (warm colours, e.g. red) indicate transition patterns of relatively high probability and high frequency in each condition, while lower intensity cells (cool colours, e.g. blue) are of lower probability and frequency. Considering all of the matrices together, it can be observed that, relatively speaking, there is a fair degree of similarity across the two vehicles in the pattern of glance transitions for each of the three conditions (baseline, voice

HMI, visual-manual HMI). Further, the pattern for the voice-based HMI tasks looks more similar to the pattern observed for baseline driving than the pattern observed for the visual-manual HMI tasks. Specifically, when using the visual-manual interfaces for phone contact calling, glance transitions are almost exclusively from the forward road to the centre stack and from the centre stack back to the forward road. In contrast, while drivers engaged with the voice-based phone contact calling task directed a high proportion of their off-road glances to the centre stack, a notable proportion of their off-road glances were to driving relevant locations (e.g. driver side window and mirror region, and rear-view mirror), more like what was observed during baseline driving periods.

The attention maps for the periods during which the smartphone visual-manual interface was used in each vehicle look very similar to those for the embedded vehicle visual-manual interfaces (Appendix Figure C1 and Figure C3). The attention maps for the periods during which the smartphone voice-based interface was used in each vehicle again look closer to those for baseline driving than is the case for the visual-manual interfaces (Figure C1), but the smartphone voice-based interactions do not show the same breadth of glances to other driving relevant locations as is seen with the embedded voice-based systems (Appendix Figure C2).

3.3. Hidden Markov modelling

Three Hidden Markov Models (baseline, manual, and voice) were individually trained on glance transition data following the procedure described in Section 2.7. Figure 5 presents a graphical view of HMM performance where specified datasets are aligned along the x-axis and the y-axis reflects the ratio of correctly classified glance sequences vs. total available sequences across 50 random splits of a dataset. The number under each dataset name indicates the average number of glance sequences per split that met the minimum established length for inclusion in the analysis (see Section 2.7); the maximum possible for each of the individual sets was 80 and was 160 for combined sets. In the plots for each dataset, each point represents the accuracy obtained for one of the 50 random splits. One standard deviation is represented by the larger rectangular box within each plot, the 95% confidence interval by the lighter internal box, and the mean by the horizontal line. A steady performance of about 70% may be observed (in particular for the smartphone interface), well above the random guessing expectation of 33%. More variation in correctness of classification is observed for the embedded interfaces, which is likely associated with the embedded interfaces being substantially different whereas the same smartphone was utilised in the two vehicles. In particular, glance sequences stemming from the Chevrolet interface seem, under visual inspection, to be more recognisable (i.e. having higher central tendency characteristics, although the range of values is also greater) than those associated with the Volvo interface.

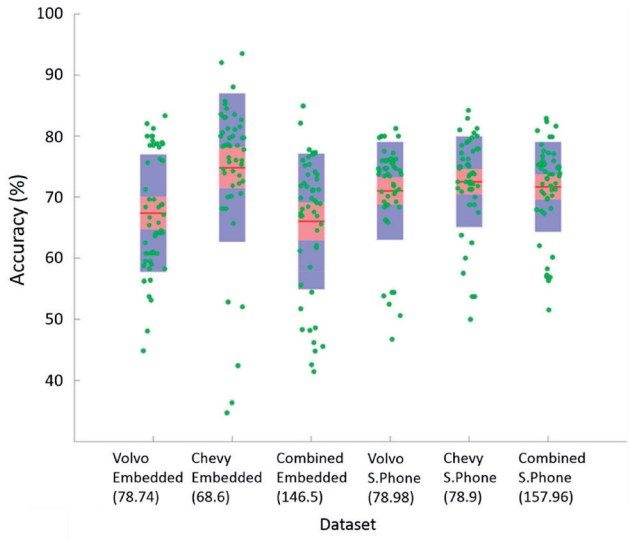


Figure 5. Accuracy distributions (y-axis) for a set of Hidden Markov Models trained individually on glance transitions for baseline driving, voice-based phone contact calling, and manual phone contact calling for the datasets arranged across the x-axis. The number under each dataset name indicates the average number of glance sequences over the 50 randomised splits that met minimum length criteria (see Section 2.7) where 80 was the maximum potential value for individual sets and 160 for combined sets. (See Section 3.3 for additional description of figure.).

As noted earlier, the number of sequences for each modelling run varied (maximum possible of 80 for single vehicle plots and 160 for two-vehicle plots) based upon the sequence length limitations. The average number of validation sequences considered across all random splits of each modelling case appear along the X-axis of Figure 5 (note that the values are fractional as they represent an average across the 50 iterations of random sampling). It can be observed that the average number of available sequences was close to 79; a loss on average of only one glance sequence from each validation set. The average number of available validation sequences was somewhat lower for the Chevrolet embedded system, with approximately 14% of the potential sequences not included in the

analysis, largely due to their length being less than the inclusion threshold (see Appendix B).

Figure 6 presents an approach for more closely examining the characteristics of the cases of misclassification in the HMMs. These confusion matrices give the average number of sequences that were either (1) correctly classified, i.e. lay on the diagonal, or (2) were known to belong to a specific class (baseline, voice-based phone contact calling, or visual-manual phone contact calling), but were misclassified as pertaining to one of the two other classes. As would be expected, the confusion matrices in the right column are quite similar as they represent baseline driving and interaction with the same smartphone interfaces in the Volvo (top row) and Chevrolet (middle row) datasets. Greater variation can be observed in

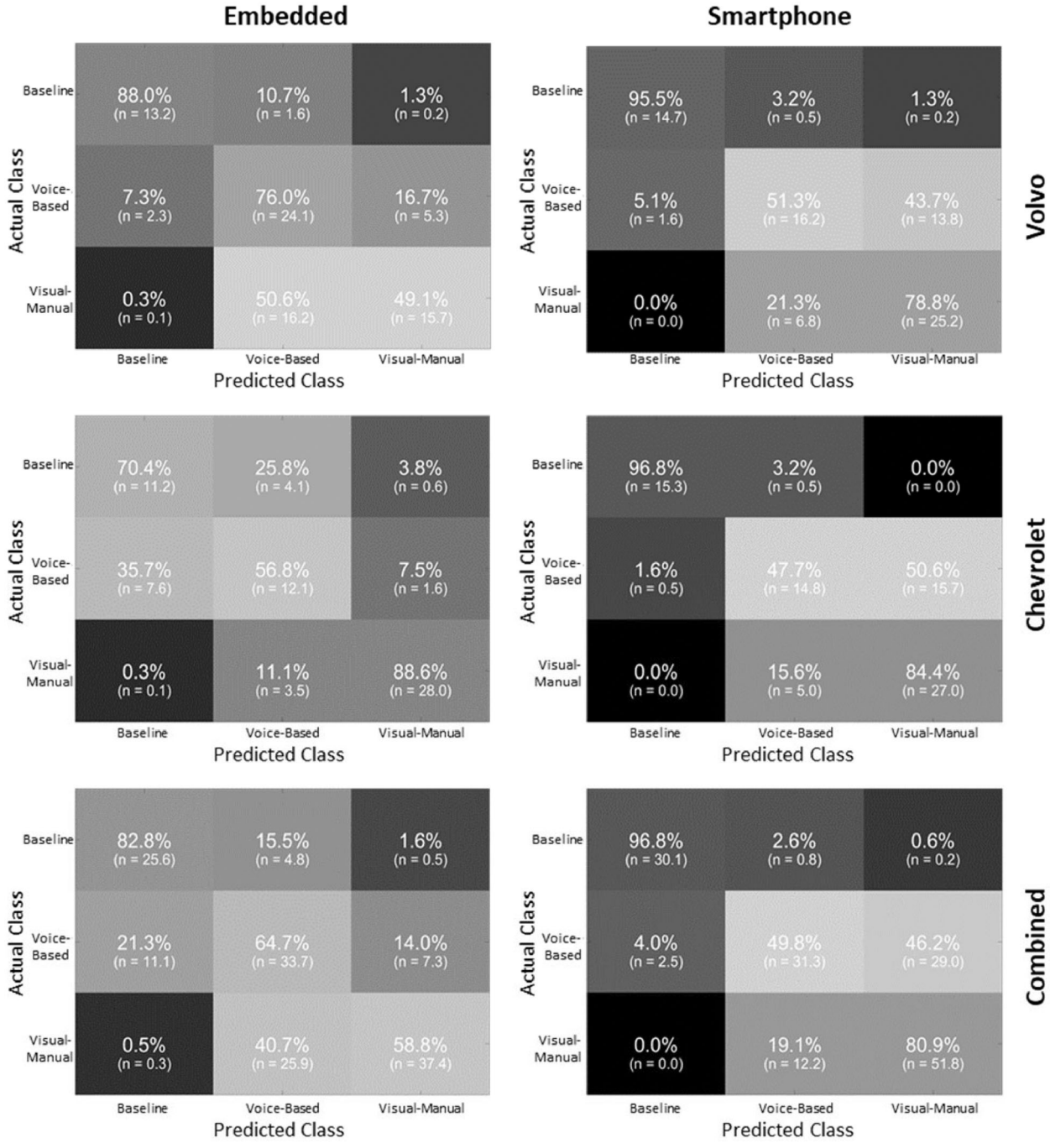


Figure 6. Confusion matrices for the classification of embedded system tasks (left column) and smartphone tasks (right column) in the Volvo (top row), Chevrolet (middle row), and both vehicles combined (bottom row). Lighter grey shading indicates greater degrees of misclassification where 100% corresponds to no confusion for the upper left to lower right diagonal and 0% for the remaining regions. These results correspond to the average over 50 runs.

the left column where baseline driving and interaction with the two different embedded vehicle interfaces are considered.

Across the top four matrices, correct classification of baseline driving ranges from an average 70.4% in the Chevrolet embedded interface dataset to 96.8% in the Chevrolet smartphone interface dataset. The very high correct classification average (96.8%) for baseline driving in both smartphone interface datasets indicates that interacting with smartphone interface for both the voice-based and the visual–manual based interface of this device resulted in a pattern of glance transitions on and off the forward road that was quite distinct from baseline driving. The source of the higher misclassification rates for baseline driving in the embedded interface datasets is quite informative. Note that misclassification of visual-manual phone contact calling as baseline driving across all datasets is very low (mean rate ranging from 0.0 for all smartphone cases to 0.5% for the combined embedded cases). Misclassification of voice-based interaction as baseline driving was lowest in the smartphone datasets, averaging only 4.0% across the two vehicles, and was only slightly higher for the Volvo embedded voice-based contact calling (7.3%). In contrast, voice-based contact calling was misclassified as baseline for 35.7% of glance sequences with the Chevrolet embedded voice-based interface. In other words, the sequence of glance transitions observed during voice-based interaction with this embedded interface was relatively often 'confused' with the observed pattern of glance transitions occurring during baseline driving.

4. Discussion

The attention maps and confusion matrices resulting from the hidden Markov modelling reported in this study illustrate differences in glance behaviours when drivers were using voice-based interfaces relative to visual-manual interfaces, and how glance behaviour during voice-based tasks were more similar to glance behaviour during baseline driving than that observed during visual-manual interface tasks. Specifically, in the attention maps, the pattern of glances away from the forward road towards situationally relevant in-vehicle information (mirrors, instrument cluster, etc.) when drivers completed a voice-based phone contact calling task shows similarities to the patterns observed during baseline driving. In contrast, glances observed during the visual-manual interface tasks were limited almost exclusively to transitions back and forth between the forward road and the physical interface. In the HMM based confusion matrices for the embedded systems, instances of drivers engaging in the voice-based interface were 'mistaken' for glance behaviour associated with 'just driving' approximately one-third of the time in the Chevrolet; this result was much less frequent for the voice-based interaction in the Volvo.

These findings extend upon the original work by Muñoz, Reimer, and Mehler (2015) and Muñoz et al. (2016) that drew upon glance data collected across 156 drivers in an unmodified production vehicle during baseline highway driving and during interactions with a visual-manual interface as well as an alternate voice-based interface for the vehicle radio. This work also describes an effective method for incorporating glance time as an input to the model. The present work addresses the question of replication and generalisability by evaluating on-road glance data in a new sample of 80 participants driving two different production vehicles, using the vehicles' embedded user interfaces as well as a dash mounted smartphone, and considering stored phone contact calling as the secondary task (Mehler et al. 2016; Reimer et al. 2016). By replicating and extending upon this earlier work, the present study further makes the case for both the basic findings and the utility of the methodology for characterising important differences in driver behaviour that arise from interaction with different HMI implementations.

Conventional glance metrics used to evaluate the visual-manual interfaces of embedded in-vehicle electronic devices (mean single glance duration, percentage of glances longer than two seconds, and total time the eyes are directed off the forward road) (NHTSA 2012, 2013) only consider whether the driver's eyes are on the forward road or off it. This simplification of glance allocations appears quite reasonable for the visual-manual interfaces evaluated since off-road glances were constrained almost entirely to the visual-manual interface. However, these metrics may not adequately characterise how drivers allocate visual attention when interacting with voice-based or other multi-modal interfaces. On the one hand, in the current study, mean single glance duration and the percentage of long duration glances off the forward road were significantly less when engaging with the voice-based interfaces compared with using the visual-manual interfaces. This was true in both vehicles, and for both the embedded interfaces and the mounted smartphone. Further, the mean values for these two metrics were largely indistinguishable for glances associated with interaction with the voice-based interfaces and off-road glances that occurred during single task,

baseline driving. These findings are in-line with previous work cited in the introduction showing reductions in visual demand associated with voice-based interfaces relative to visual-manual alternatives. It also extends upon much of that work by quantifying the similarity of the values obtained for these two metrics for the voice-based task periods and baseline driving.

On the other hand, interpreting the meaning of the values obtained for the total amount of time that the drivers' eyes were directed off the forward road is a more open question. As highlighted in Table 2, TEORT is immediately problematic for characterising glance behaviour during baseline driving since the value will inherently increase or decreasing depending on how long a sample is selected to consider as the baseline reference period. Attempting to normalise TEORT in some manner as represented by the percentage of off-road glance time in Table 2 provides an interesting perspective on the relative concentration of off the forward road glances during a period of interest, but is hard to interpret in an absolute measure sense as a safety relevant metric. For example, while a high percentage of off-road glance time is concerning for longer tasks, is it a problem if a high percentage of a short task interaction consists of off-road glances?

Setting aside for the moment the problem of applying TEORT to baseline data, for both vehicle's embedded interfaces and the smartphone, TEORT for the forward road were, on average, less than when interacting with the voice-based interface options than when using the visual-manual alternatives. This is particularly notable in the case of the Chevrolet embedded interface with a mean TEORT value of 3.3 s for the voice-based interface vs 13.6 s for the embedded visual-manual interface. The other vehicle and the smartphone interfaces were associated with longer mean total glance times, and also longer total task durations. Relatively 'long' TEORT is particularly evident when voice-based interfaces are used for more involved, multi-step tasks such entering destination addresses into a navigation system (Reimer and Mehler 2013; Reimer et al. 2014; Mehler et al. 2016; Mehler et al. 2016). These latter types of interfaces generally would not meet the NHTSA guidelines for TEORT if they were to be applied. In this regard, it should be emphasised that both the NHTSA guidelines and the earlier AAM guidelines (DFTWG 2006) make it clear that there may be issues in applying the respective guidelines to voice-involved interfaces.

The attention map analysis presents one line of evidence for why the TEORT guidance for visual-manual interfaces may not be applicable in the same manner for voice-based, and possibly other multi-modal interfaces. The transition significance matrices in Figure 3 highlight that glance allocation during the voice-based phone calling interactions with the two embedded vehicle interfaces share some of the features of how glances are distributed during single-task driving. Specifically, in contrast with glance allocation when interacting with a visual–manual interface, drivers engaged with the voice-based phone contact calling task were allocating some of their glances to driving relevant locations, e.g. the rear-view mirror and to the driver side window and mirror region. It would seem that glances allocated to locations such as the rear-view mirror, left window and left mirror region that directly support a driver's situational awareness of the overall driving scene should not logically be counted against the voice-based interface to the same extent as off-road glances to the centre stack as is done in a simple TEORT metric. The HMM classification of task type based on glance sequences and the resulting confusion matrices presented in Figure 5 provide a complementary quantitative analysis of the glance allocation patterns associated with baseline driving and driving while engaged with one of the visual–manual interfaces or one of the voice-based systems under study. It can be seen that glance allocation during the visual-manual task periods is quite distinct from that during single task driving and is almost never misclassified as such. In contrast, the sequential allocation of glances during voice-based multi-modal task periods across the two embedded vehicle interfaces was actually misclassified as single task, baseline driving a little over 20% of the time. While glance allocation patterns during the voice-based task periods were not, on average, indistinguishable from glance behaviour during single task driving, the confusion matrix results combined with the attention map plots indicate that they frequently share a number of similar characteristics. The confusion matrix data, in particular, suggests that some participants were able to interact with the voice-based interfaces in a manner that preserved a glance pattern much like that generally deployed when driving and not engaged in a secondary task. Thus, to the extent voice-based interfaces are intended to support a distribution of glances more like single task driving than is present with when interacting with a pure visual-manual interface, the present analysis along with the earlier work on radio tuning shows some evidence in-line with that goal.

As detailed in the results section, the confusion matrices provide objective values that may be useful in identifying differences between individual interface

implementations in addition to distinguishing between the general classification of voice-based vs. visual-manual. The matrices for the smartphone assessment in the two different vehicles (right column in Figure 5) generally show only relatively minor variations – in spite of the fact that these are drawn on independent samples of participants and the smartphone was mounted somewhat differently in the two vehicles (Reimer et al. 2016) to conform to a centre stack location within the constraints of differences in each vehicle's internal layout. Note in particular that the values across the smartphone baseline rows are almost identical. This increases the confidence that differences between matrices for the embedded interfaces of the two vehicles carry information about how differences in interface characteristics impact driver behaviour as opposed to reflecting variation between participants that make up the samples. Of the multiple comparisons that can be made across the embedded vehicle interfaces, perhaps the most interesting is the finding that the glance transition sequences for the voice-based phone contact calling in the Chevrolet were confused and misclassified as baseline driving 36% of the time vs. only 7% for the Volvo. Thus, the former interface, for this particular task, shows potential attentional advantages over the latter indicated by the greater difficulty discriminating the secondary task involved glance patterns from those observed during baseline driving. This finding complements the metrics showing shorter mean single glance duration, lower percentage of long duration glances, and lower TEORT. However, continued work is needed to assess whether the attentional advantages identified in this study reduce crashes associated with interaction with electronic devices.

It can be observed that the smartphone voice-based interface was identified as a voice interface about half the time and confused as a visual-manual interface about half the time in both vehicle samples. This contrasts with the results for the two embedded vehicle interfaces where, on average, the voice-based interfaces were classified as voice-based 65% of the time and confused as a visual-manual interface 14%. This appears to reflect a commonality of some aspects of glance behaviour across the two modes of the mounted smartphone interface, further highlighting that the extent to which a particular voice-based interface shows visual-manual characteristics in how glances are distributed can vary depending upon the overall implementation of the multi-modal design (e.g. task structure, display size, etc.). It is an open question as to whether a greater distinction between the voice-based and visual-manual based modes might be observed in hybrid implementations of a smartphone interface that are integrated with the embedded vehicle voice and visual-manual capabilities (e.g. Apple CarPlay).

5. Conclusions

In summary, the confusion matrices clearly indicate that the HMMs, trained upon condition-specific glance transition data, were able to distinguish among single-task baseline driving and secondary task conditions. The aggregate accuracy of these classifications, which range from 67% to 75% for the three condition classification, were based upon input sequences as short as only five glance transitions. It is important here to look beyond the present results, and understand that such findings reveal that individual patterns of glance behaviour contain information specific to each condition. That is to say that participants 'just driving', selecting a contact from a stored phone list and placing a call using a visual-manual interface, or placing the call using a voice-based interface are conditions that each rely on information-gathering strategies that produce discrete, identifiable glance patterns. These findings reflect upon the descriptive richness of the underlying transition matrices. In contrast to conventional techniques for identifying patterns of driver glance behaviour, these transitional matrices contain no temporal data; instead they distil patterns of glances between different locations. As such, the present work provides strong evidence that, as with glance duration, glance switching patterns are distinct in different task combinations (just driving vs driving while using an HMI) and between different interfaces (voice-based vs. visual-manual interfaces in two vehicles). Visual inspection of transition matrices, as presented here, clearly shows this informational richness. The present HMM-based method for abstracting this information into confusion matrices is an initial effort to quantify what is so visually salient in the attention heat maps. While not perfect, this numerical approximation of the ability of HMMs to identify specific patterns suggests that transition matrices may have both theoretical and applied use. In application, interpretation of transition matrices, both in the realm of machine learning and classical statistics, is surely possible. The present data suggesting that attentional allocation data contained in transitions can be used to inform machine learning of user activity in relatively short sequences of glance region-switches.

While the transition matrices employed in this analysis capture important information on the spatial distribution and sequencing of glances that is overlooked in TEORT, characterisations of how glances are distributed across time continue to be important and are seen as complementary to the results presented. As such, the assessment methods presented here are seen as complementary to more traditional aggregate measures (mean single glance duration, long duration glance percentages) and emerging measures (see Seppelt et al. 2017; also Lee et al. 2017; Seaman et al. 2017) in a holistic interpretation of DVI demands on driver attention. Further refinement of the methods presented is certainly possible and worthy of additional investigation. In addition, while the voice-based interfaces assessed in this study for voice-based phone contact calling showed a number of glance related advantages over the visual-manual methods, voice-based interfaces must be evaluated broadly and consideration given to both a specific task and implementation. For example, while the voice option in the 2013 Chevrolet Equinox MyLink infotainment system showed a number of apparent advantages over the 2013 Volvo XC60 Sensus interface for phone contact calling, the MyLink voice interface showed significantly higher error rates than the Sensus system for entering addresses into the navigation system and appeared to negatively impact trust in and willingness to recommend the system to others (Mehler et al. 2017).

As a final note, while it is reasonable to observe that the specific interfaces studied here date from 2013 and that the implementations in these particular vehicle lines and in smartphone interfaces have continued to evolve, the sensitivity of the methods presented here for identifying underlying differences in human engagement with overtly modest differences in design features have been demonstrated and are extensible to studying HMIs now coming on-line. It is an open question as to whether particular current generation HMIs have advanced the preservation of safety relevant glance distribution patterns or not. It is the hope of the authors that ergonomists and human factors specialists will find the approaches presented here add to the set of useful tools for such assessment.

6. Future work

The use of HMM confusion between the voice-based multi-modal interface and just driving as an indication of greater similarity between these conditions is an important avenue for further exploration. Because Markov processes cannot be observed, we are here unable to report certainty in presence or degree of similarity, or discern which features of the data drive similarity. As such, the present inference approach may not be taken as a proxy for statistical inference. Notably, there do exist approaches that allow the computation of false negatives and false positives within HMMs (Newberg 2009), and the application of such approaches should be explored in future work.

Acknowledgements

Support for this work was provided by the Insurance Institute for Highway Safety (IIHS). Mauricio Muñoz and Jonathan Dobres were affiliated with the MIT AgeLab at the time this work was developed. Mauricio Muñoz is currently a research engineer at the Bosch Center for Artificial Intelligence (a.m.munoz.delgado@gmail.com) and Jonathan Dobres is a senior data scientist at Sonos. David Kidd (David.Kidd@leidos.com) is currently a human factors program manager with Leidos.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Support for this work was provided by the Insurance Institute for Highway Safety (IIHS).

References

Angell, L., J. Auflick, P. A. Austria, D. Kochhar, L. Tijerina, W. Biever, T. Diptiman, J. Hogsett, and S. Kiger. 2006. "Driver Workload Metrics Task 2 Final Report & Appendices." Final Research Report: DOT HS 810 635. Washington, DC: NHTSA/USDOT.

Caird, J. K., K. A. Johnston, C. R. Willness, and M. Asbridge. 2014. "The Use of Meta-Analysis or Research Synthesis to Combine Driving Simulation or Naturalistic Study Results on Driver Distraction." *Journal of Safety Research* 49: 91.e1–46. doi:10.1016/j.jsr.2014.02.013.

Chiang, D.P., A.M. Brooks, and D.H. Weir. 2005. "Comparison of Visual-Manual and Voice Interaction with Contemporary Navigation System HMIs (Paper no. 2005-01-0433)." Paper presented at the SAE 2005 World Congress & Exhibition. SAE International: Warrendale, PA.

Conover, W. J., and R. L. Iman. 1981. "Rank Transformations as a Bridge between Parametric and Nonparametric Statistics." *American Statistician* 35 (3): 124–129.

Dobres, J., B. Reimer, B. Mehler, J. Foley, K. Ebe, B. Seppelt, and L. Angell. 2016. "The Influence of Driver's Age on Glance Allocations during Single-Task Driving and Voice vs. Visual-Manual Radio Tuning (No. 2016-01-1445)." SAE Technical Paper.

DFTWG - Driver Focus-Telematics Working Group. 2006. *Statement of Principles, Criteria, and Verification Procedures on Driver-Interactions with Advanced in-Vehicle Information and Communication Systems*. Washington, DC: Alliance of Automobile Manufacturers.

Friedman, M. 1937. "The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance." *Journal of the American Statistical Association* 32 (200): 675–701. doi:10.1080/01621459.1937.10503522.

Hancock, P. A., and B.D. Sawyer. 2015. "Judging Thieves of Attention: Commentary on "Assessing Cognitive Distraction in the Automobile," by Strayer, Turrill, Cooper, Coleman, Medeiros-Ward, and Biondi (2015)." *Human Factors* 57 (8): 1339–1342. doi:10.1177/0018720815578971.

He, L., C. F. Zong, and C. Wang. 2012. "Driving Intention Recognition and Behaviour Prediction Based on a Double-Layer Hidden Markov Model." *Journal of Zhejiang University SCIENCE C* 13 (3): 208–217. doi:10.1631/jzus.C11a0195.

Jansen, R. J., B. D. Sawyer, R. van Egmond, H. de Ridder, and P. A. Hancock. 2016. "Hysteresis in Mental Workload and Task Performance: The Influence of Demand Transitions and Task Prioritization." *Human Factors* 58(8): 1143–1157. doi:10.1177/0018720816669271.

Khan, M. Q., and S. Lee. 2019. "Gaze and Eye Tracking: Techniques and Applications in ADAS." *Sensors* 19 (24): 5540. doi:10.3390/s19245540.

Kidd, D.G., and A.T. McCartt. 2015. *The Relevance of Crash Type and Severity When Estimating Crash Risk Using the SHRP2 Naturalistic Driving Data. 4th International Driver Distraction and Inattention Conference*. Sydney, New South Wales: ARRB Group Ltd.

Lee, J.B., B.D. Sawyer, B. Mehler, L. Angell, B.D. Seppelt, L. Fridman, and B. Reimer. 2017. January). "Linking the Detection Response Task and the AttenD Algorithm through the Assessment of Human-Machine Interface Workload." Proceedings of the Transportation Research Board Annual Meeting, Washington, DC, January 8-12, 2017.

Li, J., Y. Dou, J. Wu, W. Su, and C. Wu. 2021. "Distracted Driving Caused by Voice Message Apps: A Series of Experimental Studies." *Transportation Research Part F: Traffic Psychology and Behaviour* 76: 1–13. doi:10.1016/j.trf.2020.10.008.

Mehler, B., D. Kidd, B. Reimer, I. Reagan, J. Dobres, and A. McCartt. 2016. "Multi-Modal Assessment of on-Road Demand of Voice and Manual Phone Calling and Voice Navigation Entry across Two Embedded Vehicle Systems." *Ergonomics* 59 (3): 344–367. doi:10.1080/00140139.2015.1081412

Mehler, B., B. Reimer, J. Dobres, J. Foley, and K. Ebe. 2016. "Additional Findings on the Multi-Modal Demands of Production Level 'Voice-Command' Interfaces." *SAE Technical Paper*: 2016–2001. doi:10.427/2016-01-1428.

Mehler, B., B. Reimer, C. Lee, D. Kidd, and I. Reagan. 2017. June). "Considering Self-Report in the Interpretation of Objective Performance Data in the Comparison of HMI Systems." Proceedings of the 9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, June 27-29, Manchester Village, Vermont, USA.

Muñoz, M.,. B. Reimer, J. Lee, B. Mehler, and L. Fridman. 2016. "Distinguishing Patterns in Drivers' Visual Attention Allocation Using Hidden Markov Models." *Transportation Research Part F: Traffic Psychology and Behaviour* 43: 90–103. doi:10.1016/j.trf.2016.09.015.

Muñoz, M.,. B. Reimer, and B. Mehler. 2015. "Exploring New Qualitative Methods to Support a Quantitative Analysis of Glance Behavior." Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '15), September 01-03, 2015, Nottingham, United Kingdom. http://dl.acm.org/citation.cfm?id=2799278

National Highway Traffic Safety Administration. 2012. *(Proposed Guidelines) Visual-Manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices (Docket No. NHTSA-2010-0053)*. Washington, DC: U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA).

National Highway Traffic Safety Administration. 2013. *Visual-Manual NHTSA Driver Distraction Guidelines for in-Vehicle Electronic Devices (Docket No. NHTSA-2010-0053)*. U.S. Department of Transportation National Highway Traffic Safety Administration (NHTSA), Washington, DC.

Newberg, L. A. 2009. "Error Statistics of Hidden Markov Model and Hidden Boltzmann Model Results." *BMC Bioinformatics* 10 (1): 212. doi:10.1186/1471-2105-10-212.

Pentland, A., and A. Liu. 1999. "Modeling and Prediction of Human Behavior." *Neural Computation* 11 (1): 229–242. doi:10.1162/089976699300016890.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria. http://www.R-project.org/.

Reimer, B., L. Angell, D. Strayer, L. Tijerina, and B. Mehler. 2016. "Evaluating demands associated with the use of voice-based in-vehicle interfaces." Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Washington, DC, September 19–23, 2016, pp. 2083–2087. doi:10.1177/1541931213601472.

Reimer, B., and B. Mehler. 2013. *The Effects of a Production Level 'Voice-Command' Interface on Driver Behavior: Summary Findings on Reported Workload, Physiology, Visual Attention, and Driving Performance. MIT AgeLab White Paper No. 2013-18A*. Cambridge, MA: Massachusetts Institute of Technology. (Link)

Reimer, B., B. Mehler, J. Dobres, H. McAnulty, A. Mehler, D. Munger, and A. Rumpold. 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). ACM. doi:10.1145/2667317.2667320.

Reimer, B., B. Mehler, I. Reagan, D. Kidd, and J. Dobres. 2016. "Multi-Modal Demands of a Smartphone Used to Place Calls and Enter Addresses during Highway Driving Relative to Two Embedded Systems." *Ergonomics* 59 (12): 1565–1585. doi:10.1080/00140139.2016.1154189.

Sawyer, B. D., V. S. Finomore, A. A. Calvo, and P. A. Hancock. 2014. "Google Glass: A Driver Distraction Cause or Cure?" *Human Factors* 56 (7): 1307–1321. doi:10.1177/0018720814555723.

Sawyer, B. D., J. Lee, J. Dobres, B. Mehler, J. Coughlin, and B. Reimer. 2016. "Effects of a Voice Interface on Mirror Check

Decrements in Older and Younger Multitasking Drivers." Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Washington, D.C., September, 2015.

Seaman, S.,. J. Lee, B. Seppelt, L. Angell, B. Mehler, and B. Reimer. 2017. "It's All in the Timing: Using the AttenD Algorithm to Assess Texting in the NEST Naturalistic Driving Database." Proceedings of the 9th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, June 27–29, Manchester Village, Vermont, USA. doi:10.17077/drivingassessment.1665.

Seppelt, B., S. Seaman, J. Lee, L.S. Angell, B. Mehler, and B. Reimer. 2017. "Glass half-full: On-road glance metrics differentiate crashes from near-crashes in the 100-Car data." *Accident; Analysis and Prevention* 107: 48–62. doi:10.1016/j.aap.2017.07.021

Smith, D., J. Chang, R. Glassco, J. Foley, and D. Cohen. 2005. "Methodology for Capturing Driver Eye Glance Behavior during in-Vehicle Secondary Tasks." *Transportation Research Record: Journal of the Transportation Research Board* 1937 (1): 61–65. doi:10.1177/0361198105193700109.

Strayer, D. L., J. M. Cooper, J. Turrill, J. R. Coleman, and R. J. Hopman. 2015. *Measuring Cognitive Distraction in the Automobile III: A Comparison of Ten 2015 in-Vehicle Information Systems*. Washington, DC: AAA Foundation for Traffic Safety.

Wickens, C. D. 2002. "Multiple Resources and Performance Prediction." *Theoretical Issues in Ergonomics Science* 3 (2): 159–177. doi:10.1080/14639220210123806.

Yuan, W., Z. Liu, and R. Fu. 2018. "Predicting Drivers' Eyes-off-Road Duration in Different Driving Scenarios." *Discrete Dynamics in Nature and Society* 2018: 1–9. doi:10.1155/2018/3481628.

Appendix A: from transition counts to a transition significance matrix

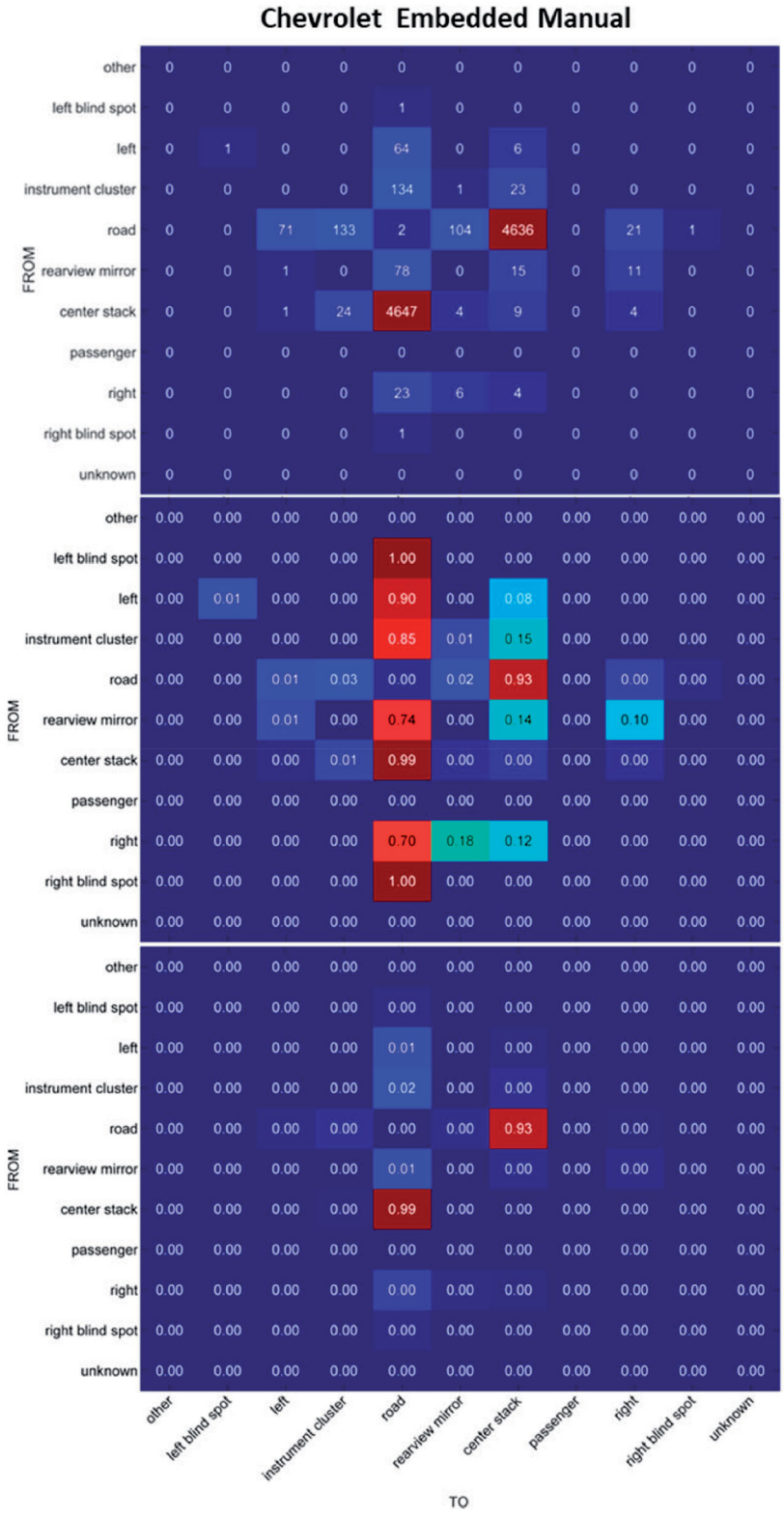
The differences between and the reasons for generating the different transition matrix types can seem subtle. Section 2.6 provides a description of each and how it is calculated. The reasoning for calculating the Transition Significance Matrix is further described here for readers interested in additional background.

The motivation for calculating the Transition Significance Matrix is that the more typical transition matrix (Transition Probability), which yields the probability of transitioning from a given location J to another location K as a number between 0 and 1, does NOT consider how representative the probability is in real-world data. For example, during a radio tuning task, transitions between the centre stack and the forward roadway are frequent, while transitions to the passenger seat are likely to not be as frequent. Going from the source count matrix to a basic transition probability matrix loses this information, yielding say 90% probability of switching back to the centre stack if currently at the forward roadway, and say 90% probability of switching to the forward roadway if currently looking at the passenger seat. The transition probability matrix summarises both behaviours using the same number, irrespective of how frequent the individual transition pairs were.

While the transition probability matrix can reveal interesting information, it only provides half the picture. In order to address this, the goal is to provide a measure for the extent to which transition pairs are "important" in the overall glance behaviour of the driver, ranking them on the scale of 0 to 1, with 1 being the "most important". This can be done by dividing all elements of the transition probability matrix by the maximum element in the matrix. Note, this is just one possible metric for importance - dividing by the sum of the entire matrix would work as well, as more frequent transition pairs will take up a higher percentage of this total sum than pairs that were not as common. The same goal is achieved, to differentiate them from each other.

The transition significance calculation is thus nothing other than the element wise multiplication of the transition probabilities with these importance values. The values are also restricted between 0 and 1, and signal a relative importance score that can now be compared across tasks and experiments (i.e. we can now compare two of these matrices with each other as the transition probability bias discussed above is gone). Values that are high on the 0 to 1 scale reflect transition pairs that are not just likely, but that also appear frequently in the data. In the lower end of the importance ranking are pairs that either are not very likely, or simply did not appear often enough (proportionally speaking with respect to the most frequent transitions) to ensure a fair comparison.

Figure A1 details the individual matrices described in Section 2.6 on *Attention Map Visualisations*



Basic Matrix: Simple transition counts of glances FROM one region (row) TO another region (column).

Transition Probability from any single region to another region: Divide the transition counts between regions (each cell) by its row sum (total number of transitions from a given location to all other locations). The resulting row sums in the new matrix all equal one (allowing for precision rounding errors).

Transition Significance: Divide the transition counts between regions (each cell) by the maximum observed count (cell with the largest number of transitions in the entire matrix) to create a *transition importance matrix* (not shown). Each cell in this matrix is then multiplied by the corresponding cell in the transition probability matrix to create its transition significance. The transition significance for glances from road to center stack thus equals (4636/4968) * (4636/4647) = 0.93.

Figure A1. Representative matrices for the Chevrolet Embedded Manual Calling condition.

Appendix B: Setting a minimum length for usable glance sequences

As summarised in the main body of the paper, a potentially important parameter of the modelling framework considers the conditions under which a given glance allocation sequence is considered valid or useful. In the context of the current work, this reduces to question of whether a minimum number of glance sequences should be present in a given epoch (baseline period or individual task trial) before including it when using an HMM approach to attempt to distinguish the patterns of behaviour of interest. HMMs can be characterised as 'data-hungry', typically classifying long, expressive sequences more accurately than shorter ones. *Muñoz, Reimer and Mehler (2015)* evaluated the correlation between the length of the individual sequences and classification performance for visual-manual and voice radio tuning, noting the trade-off between the length of individual sequences and the expressivity of the final accuracy measure as an artefact of the size of the available data set. Intuitively, using longer sequences, though beneficial during training as HMMs are able to profit from more expressive information content, leaves fewer sequences for actual validation. In an extreme case, this could make the accuracy metric uninformative. On the other end of the spectrum, allowing shorter sequences allows for more validation points, but potentially feeds the models with less-descriptive samples during training.

Muñoz, Reimer and Mehler (2015) set an empirically-determined threshold of a minimum of 11 glances (a sequence of at least 10 transitions from one region to another) for all training and validation sequences. For the present analysis, optimisation of a minimum threshold was explored for the dataset, seeking to maximise the number data samples available for comparison without harming model performance. Starting with the minimum threshold of 11 glances sequences per epoch, Table B1 shows the number of additional sequences that became available for inclusion in the modelling as the minimum threshold was decreased. As this number of additional sequences represents both training and validation sequences, multiplying by a factor of 0.2 gives the additional number of sequences available for validation.

In order to examine the sensitivity of the models in terms of their discriminative potential as applied to sequences of glance allocations for the phone contact calling tasks, the approach described above was applied and repeated for a subset of these thresholds. Average prediction accuracies from each distribution were sampled for each dataset component, and a final accuracy was produced by averaging these scores across all components (Table B2). Overall, classification performance seems to be rather insensitive to sequence length for classification of baseline, manual and voice phone contact calling profiles down to the minimum sequence length evaluated (4 glances; 3 transitions). Since using sequences with lengths of 6 or more glances (5 transitions) offers the best performance to quantity-of-data ratio, all subsequent HMM results use this threshold.

Table B1. Total number of additional sequences available in the dataset as a function of the threshold selected for a minimum number of glances, where 1 is equivalent to allowing all sequences in the modelling and validation procedures.

Threshold	11	10	9	8	7	6	5	4	3	1
Additional Sequences	–	9	22	33	36	57	62	86	89	90

A threshold of 11 or more glances was used in Muñoz et al. (2015).

Table B2. Prediction accuracies for different values of a minimum acceptable number of glances (11, 8, 6, or 4) for including an observed set of glance transitions in the analysis.

	Minimum Acceptable Number of Glances in a Sequence			
Dataset	11	8	6	4
Volvo Embedded	0.69	0.66	0.67	0.66
Chevrolet Embedded	0.75	0.75	0.75	0.7
Volvo & Chevrolet Embedded	0.67	0.66	0.66	0.66
Volvo S.Phone	0.7	0.7	0.71	0.7
Chevrolet S.Phone	0.7	0.72	0.73	0.71
Volvo & Chevrolet S.Phone	0.7	0.69	0.72	0.68
Average	0.7016	0.6966	0.7066	0.685

(Note: a sequence of 6 glances corresponds to a sequence of 5 transitions.).

Appendix C: Additional transition significance matrices visualizations

Transition significance matrices considering smartphone conditions

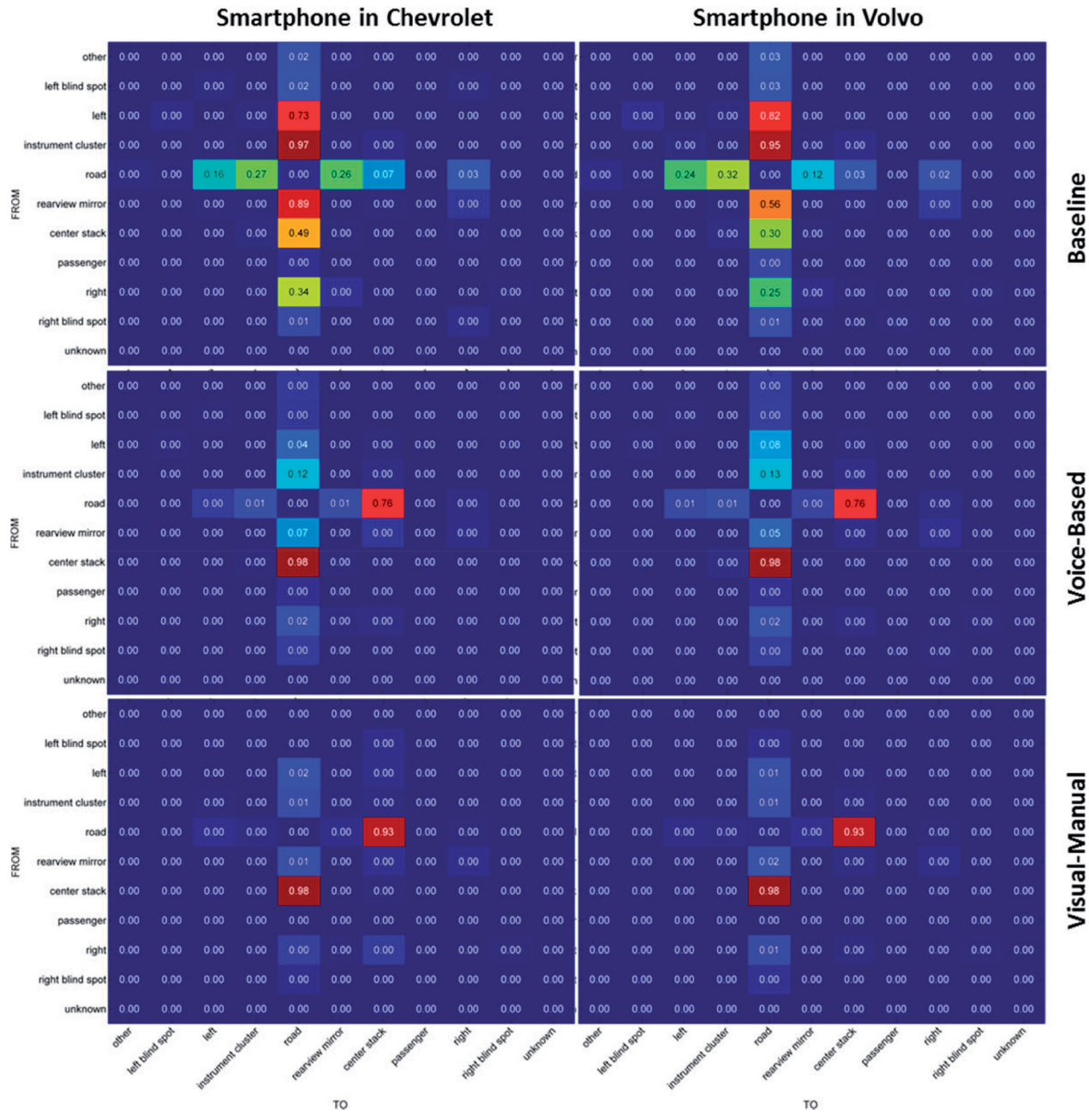


Figure C1. Glance transition significance matrices for baseline driving (top), voice-based phone contact calling tasks (middle), and visual-manual phone contact calling tasks (bottom) using the smartphone interface while driving in the Chevrolet (left) and Volvo (right). Warm colours (e.g. red) indicate transition patterns of relatively high probability and frequency, while cool colours (e.g. blue) are of lower probability and frequency.

Transition significance matrices for all Voice-Based interaction periods

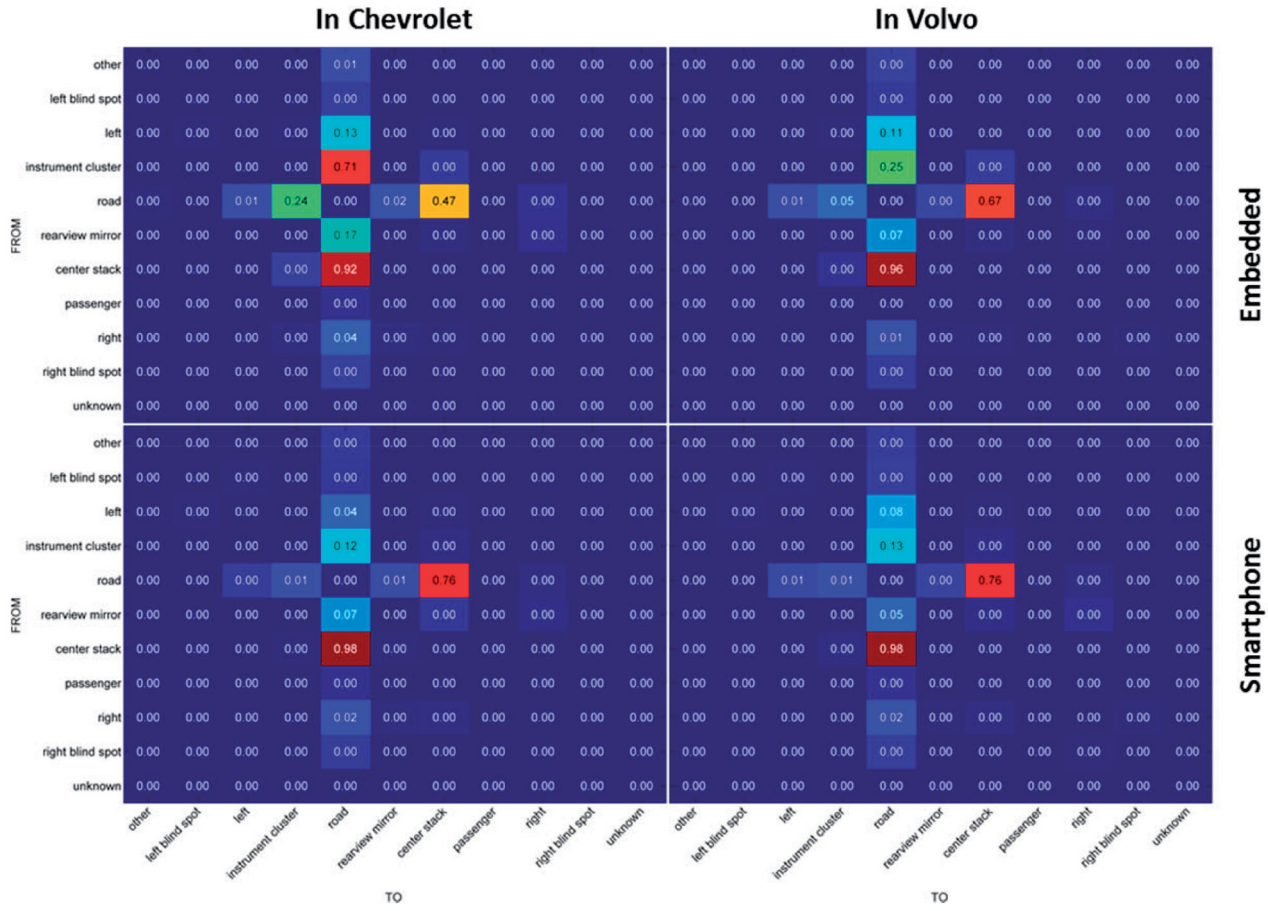


Figure C2. Glance transition significance matrices for voice-based phone contact calling tasks in the Chevrolet and in the Volvo using the embedded interfaces (top) and smartphone interface (bottom). Warm colours (e.g. red) indicate transition patterns of relatively high probability and frequency, while cool colours (e.g. blue) are of lower probability and frequency.

Transition significance matrices all Visual-Manual interaction periods

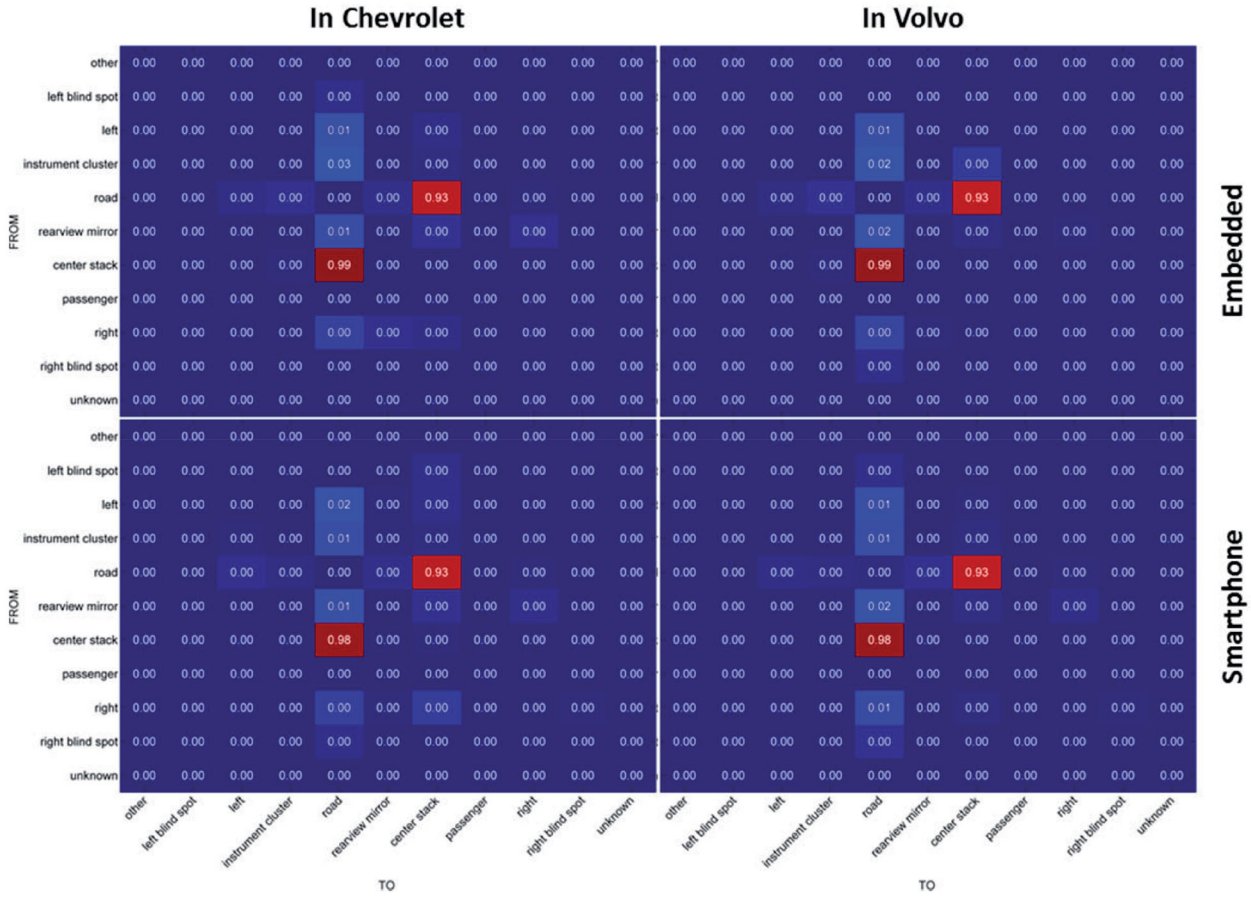


Figure C3. Glance transition significance matrices for visual-manual phone contact calling tasks in the Chevrolet and in the Volvo using the embedded interfaces (top) and smartphone interface (bottom). Warm colours (e.g. red) indicate transition patterns of relatively high probability and frequency, while cool colours (e.g. blue) are of lower probability and frequency.