

Dictating and Editing Short Texts while Driving: Distraction and Task Completion

Jan Cuřín, Martin Labský, Tomáš Macek, Jan Kleindienst
IBM Česká Republika, spol. s r.o.
V Parku 2294/4, 148 00 Praha 4
Czech Republic
+420 234 341 111
{jan_curin, martin.labsky, tomas_macek, jankle}@cz.ibm.com

Hoi Young, Ann Thyme-Gobbel
Nuance Communications Inc
1198 E Arques Ave
Sunnyvale, CA 94085
+1 408 245-5358
{Hoi.Young, Ann.Thyme-Gobbel}@nuance.com

Holger Quast
Nuance Communications International BVBA
Guldensporenpark 32
9820 Merelbeke, Belgium
+32 (0) 9 239 8000
Holger.Quast@nuance.com

Lars König
Nuance Communications Aachen GmbH
Soeflingerstr. 100
D-89077 Ulm, Germany
+49 160 93974905
Lars.Koenig@nuance.com

ABSTRACT

This paper presents a multi-modal automotive dictation editor (codenamed *ECOR*) used to compose and correct text messages while driving. The goals are to keep driver's distraction minimal while achieving good task completion rates and times as well as acceptance by users. We report test results for a set of 28 native US-English speakers using the system while driving a standard lane-change-test (LCT) car simulator. The dictation editor was tested (1) without any display, (2) with a display showing the full edited text, and (3) with just the “active” part of text being shown. In all cases, the system provided extensive text-to-speech feedback in order to prevent the driver from having to look at the display. In addition, cell phone messaging and GPS destination entry were evaluated as reference tasks. The test subjects were instructed to send text messages containing prescribed semantic information, and were given a list of destinations for the GPS task. The levels of driver distraction (evaluated by car's deviation from an ideal track, reaction times, number of missed lane change signs, eye gaze information etc.) were compared between the 3 *ECOR* and the 2 reference tasks, and also to undistracted driving. Task completion was measured by the number and quality of messages sent out during a 4 minute LCT ride, and subjective feedback was collected via questionnaires. Results indicate that the eyes-free version keeps the distraction level acceptable while achieving good task completion rate. Both multi-modal versions caused more distraction than the eyes-free version and were comparable to the GPS entry task. For native speakers, the missing display for the eyes-free version did not impact quality of dictated text. By far, the cell phone texting task was the most distracting one. Text composition speed using dictation was faster than cell phone typing.

Categories and Subject Descriptors

H5.2. Information interfaces and presentation: User Interfaces.

Copyright held by author(s)

AutomotiveUI'11, November 29-December 2, 2011, Salzburg, Austria

General Terms

Design, Experimentation, Human Factors.

Keywords

Automotive dictation and messaging, speech recognition, lane change test, driving distraction.

1. INTRODUCTION

The popularity of using various communication, navigation, entertainment and driver assistance systems in cars is steadily increasing as more of these systems enter the market at competitive prices. Text entry is one domain not yet covered by current production systems. According to [6], about 30% of drivers surveyed in Australia sometimes entered a text message using their mobile phone while driving and 1 out of 6 drivers did so regularly. At the same time, receiving and especially sending messages is perceived by drivers as one of the most distracting tasks [20]. As a consequence, the number of distraction-related crashes is thought to be increasing. The primary constraint when developing automotive UIs is thus to keep driver's distraction minimal. The secondary aims are to minimize task completion time and maximize task completion quality.

In this paper, we evaluate a prototype of the text dictation UI, codenamed *ECOR*. The *ECOR* system has been developed with the aim to act as a test bed for evaluation of multiple error correction techniques. It implements several variants of multimodal user interface. We report here results of the tests conducted to evaluate impact of the UI. By conducting a mix of objective and subjective tests we attempt to capture system performance and find the answers for questions related to optimal UI design.

We report objective measures based on driving performance during LCT trips [10] and on road attention annotation of video recordings. The subjective measures include System Usability Scale (SUS) rating [2], NASA Task Load Index (NASA-TLX) rating [5], and four additional questions related to our task.

Section 2 discusses related research, Section 3 introduces the *ECOR* UI and Section 4 describes the experimental setup. The evaluation procedure is outlined in Section 5 and evaluation results for both objective and subjective measures are presented in Section 6. Finally, Section 7 concludes and outlines future work.

2. RELATED WORK

Impact of driver distraction on driving has been the subject of numerous studies in the past. According to the US National Highway Traffic and Safety Administration [15], about 25% to 30% of police-reported crashes involve some form of driver inattention. More than half of this amount corresponds to driver distraction. An interesting study reporting the impact of several types of distraction was published by the American AAA Foundation for Traffic Safety [16]: various types of driver distraction were observed and quantified. The study showed that when excluding simple conversing with passengers, drivers were engaged in some form of potentially distracting activity up to 16% of the total time that their vehicles were moving. Cell phone usage amounted for 1.3% of the total driving time. About 30% of users used the phone while driving.

The effect of using cell phones in a car is studied for example in [21]. Users tend to use phones in spite of law restrictions. The study reports significant negative impacts of cell phone use on driving performance.

Comparisons of driving performance degradation due to using conventional and speech-enabled UIs has been addressed in several works, see for example [11]; a good summary can be found in [1]. The general conclusion is that while speech UIs still impact driving quality, they do so significantly less than conventional UIs. Most distraction caused by conventional systems seems to be due to drivers looking away from the road, which can be measured e.g. by the number and duration of eye gazes. In addition, using speech was observed to be faster for most evaluated tasks.

A number of approaches were described to perform dictation in hands-busy environments [18]. Previous work has also been done on email messaging in a car [7]. In particular, hands-free text navigation and error correction were addressed by [17]. The impact of the most prevalent correction method, re-speaking, was evaluated by [19]. Microsoft described a prototype system [8] that allowed responding to incoming text messages by matching message templates.

We have presented evaluation results of a predecessor to this system on a smaller set of non-native speakers in [9].

Use of the LCT simulator for the primary (driving) task is a de facto standard in the area of automotive usability tests and enables us to compare with many other studies. On the other hand, it has certain shortcomings such as absence of sudden events, high-way scenario only, absence of traffic simulation, and more [13]. We also chose the LCT simulator since it enabled us to complete a single test drive in a short time (we had a limitation of one hour per subject).

3. APPLICATION INTERFACE

The *ECOR* dictation editor allows for entering short texts primarily using open-domain dictation. Alternate input modalities include spelling by voice and handwriting (e.g. to input out-of-vocabulary words); in this paper we however focus on the “mainstream” dictation usage.

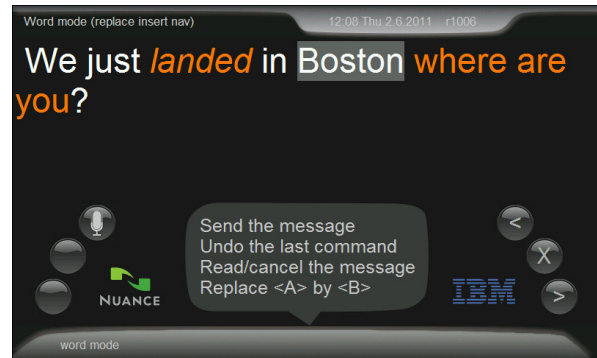


Figure 1. Full message view.



Figure 2. Strip view with text position indicator.

The prototype can be used with or without a display (multi-modal and eyes-free modes). The user initiates dictation by pressing the speech button (Figure 3). Recording ends automatically after the user has stopped speaking or after the speech button has been pressed again. After dictating a phrase, the recognized text is echoed back using text-to-speech (TTS). We are using Nuance Vocalizer for Automotive, version 5. The driver may navigate the text using previous/next buttons or a rotary knob while TTS plays back the active word(s). Text can be navigated by whole recognized phrases, by individual words, and by letters (chunk, word and letter modes).

The active text item is always the one last spoken by the TTS. It is subject to contextual editing operations, which include deletion, replacement by the next or previous n-best alternate, and several voice commands. Context-free editing operations include undo and redo, and corrective voice commands.

Besides the eyes-free setup, there are two kinds of GUI available: the **Full view** (Figure 1), always showing the full dictated text, and the **Strip view** (Figure 2), only showing the active word(s) and, optionally, near context.

Besides dictation input, several voice commands listed in **Table 1** are recognized by the system in order to carry out actions not mapped to physical controls such as buttons or knobs.

Table 1: Types of voice commands

Always available	After some text has been entered
Help	Send it
Chunk / word / spell mode	Read the whole message
Undo	Delete the whole message
	Capitalize / To uppercase / To lowercase
	Replace <wrong> by <correct>
	I said <correct>

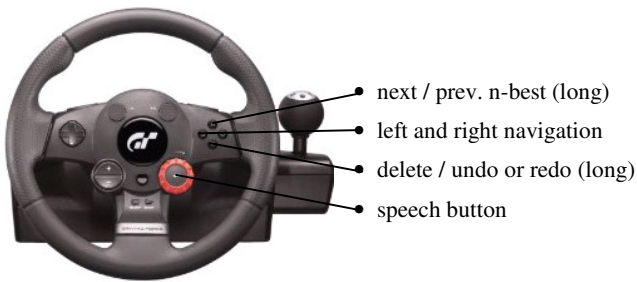


Figure 3. Steering wheel button functionality.



Figure 4. Experiment setup.

4. EXPERIMENT SETUP

A standard LCT car simulator [10] was used to simulate driving in an office environment shown in **Figure 4**. The simulator was showing on a 40" screen and the *ECOR* screen showed on a separate 8", 800x600 touch-screen, positioned on the right side of the simulator screen. A *Logitech MOMO* steering wheel and pedals were used to control the simulator and 5 buttons (incl. push-to-talk) on the steering wheel controlled the prototype. Our setup was very similar to that used by [4], [12] and to that described as "PC condition" by [3].

One LCT trip consisted of a 3km straight 3-lane road with 18 irregularly distributed lane change signs (lane changes of both 1 and 2 lanes were included). The evaluated segment started with an extra "START" sign and ended 50m after the last change lane sign. Drivers kept a fixed speed of 40km/h (11.1m/s) during the whole trip. One LCT trip took approximately 5 minutes.

5. EVALUATION PROCEDURE

All user studies described in this paper were conducted by the Nuance Usability Testing Lab in Sunnyvale, California, with native US-English speakers, who had no previous experience in using voice-controlled dictation systems and who used their cell phones regularly to send at least 10 text messages or emails per day.

A group of 28 novice test subjects (14 female, 14 male, age between 18 and 55) was used to measure objective statistics as well as to record subjective feedback using questionnaires. The Eyes-free, Full view and Strip view versions of *ECOR* were evaluated as well as cell phone typing and GPS address entry, and compared to undistracted driving. Each test subject was evaluated on the undistracted driving task, one of the three *ECOR* tasks, and one of the two reference tasks (cell phone texting or GPS). **Table 2** uncovers details on the number of participants performing individual tasks (the last sum row) and task order for each participant (numbers in the table denote task order).

Participants with the same task ordering are grouped into single rows. Exceptions (some participants were not evaluated for particular measurements) are denoted by * and ^. Minor imbalances between the groups were caused by participants not showing up or technical problems in some of the captured data.

First, each subject was allowed to train driving until they mastered the LCT and their driving performance did not further improve. Then, two **undistracted** LCT trips were collected. The second one was used to compute an adapted model of the driver's ideal path and the first was used to compute driving performance statistics using the adapted ideal path.

The **ideal path** was modeled using a linear poly-line. Because each driver had a slightly different driving style, we adapted several parameters of the ideal path to accommodate for individual driving styles, in order to make the driving performance statistics more comparable among drivers. The major differences among drivers during their undistracted drives are listed below along with the corresponding adapted parameters of the ideal path:

- Different steering angles and durations of lane changes (maneuver lengths changed 1 or 2 lanes in both directions).
- Different reaction times to lane change signs (distance before the lane change sign where the maneuver starts).
- Different standard driving positions within each of the 3 lanes (lateral car position offset for each lane).

After the undistracted LCT trips, approximately half (15) of test subjects was first introduced to the selected *ECOR* prototype and then evaluated while using it while driving. The same was then done for the selected reference task (cell phone texting or GPS). The order of the two distracted driving tasks was swapped for the remaining test subjects (13).

For the selected *ECOR* task (eyes-free, full message view, or strip view) each subject had sufficient time to practice dictating arbitrary text without driving the LCT simulator (with the car parked). After practicing *ECOR* for 10-20 minutes, each subject conducted a single LCT trip, during which s/he was instructed to use *ECOR* to enter a sequence of text messages with pre-defined semantic content (e.g. "instruct your partner to buy oranges, wine and chocolate" or "tell your secretary to set up a meeting, at the library, tomorrow, at 5pm.").

Table 2. The task order for individual participants; only showing for tasks that the participant(s) actually performed. Several measurements could not be performed for some participants: * denotes missing visual road attention information for a single participant (caused by an unusable video recording) and ^ denotes a single missing questionnaire.

Subject count	undistracted LCT ride	undistracted (adapted on)	Eyes-free <i>ECOR</i>	Full msg view <i>ECOR</i>	Strip view <i>ECOR</i>	Cell phone texting (SMS)	GPS address entry
4	1	2		3			4*
1	1	2		3		4	
3	1	2		4			3^
1	1	2		4		3	
3	1	2	4				3
4	1	2	3				4
1	1	2	3			4	
5	1	2			3**	4*	
6	1	2			4	3	
sum	28	28	8	9	11	14	14

For the selected *ECOR* task (eyes-free, full message view, or strip view) each subject had sufficient time to practice dictating arbitrary text without driving the LCT simulator (with the car parked). After practicing *ECOR* for 10-20 minutes, each subject conducted a single LCT trip, during which s/he was instructed to use *ECOR* to enter a sequence of text messages with pre-defined semantic content (e.g. “instruct your partner to buy oranges, wine and chocolate” or “tell your secretary to set up a meeting, at the library, tomorrow, at 5pm.”).

For the **cell phone texting task**, the subjects were instructed to enter a sequence of text messages with the same semantic content as for the *ECOR* tasks. The subjects were using their own cell phones, so they were familiar with its UI. Use of predictive typing was left up to the choice of the users. As stated already, the subjects were claiming themselves to be texting quite often.

For the **task of entering an address using in-car GPS navigation**, the subjects were asked to enter addresses from a pre-defined list. The same GPS device (TomTom XXL) was used by all subjects performing this reference task. As for the *ECOR* tasks, subjects were given enough training time to practice using the GPS device while parked before performing the evaluated LCT trip. The GPS device used auto-complete so users did not typically need to enter the full address.

6. EVALUATION RESULTS

6.1 Driving Quality Evaluation

Driving quality was evaluated using the LCT and associated tools, which enabled us to collect the following statistics:

- The car’s mean deviation (**Overall MDev**) from the ideal track in meters. This measures how much, on average, the driver drove off his/her ideal track, and is computed simply as the absolute value of subtracting the actual and ideal lateral car positions at each sampled point, and by averaging over the entire distance of the trip.
- The standard deviation of lateral position (**Overall SDLP**) of the car in meters. SDLP measures how much the driver “weaves” within the lane and is computed as the standard deviation of the car’s absolute actual deviation from its ideal path at each sampled point and averaged over the entire distance of the trip.
- **Lane-keeping MDev** is computed as Overall MDev with exclusion of the lane change maneuver parts of the LCT trip. I.e., it indicates how well the subject keeps the lane between the lane change signs.
- **Lane-keeping SDLP** is again similar to Overall SDLP, but excluding the lane change maneuver parts.
- **Reaction time** is the delay between the moment when the lane change sign becomes visible and the moment when the driver starts responding by an observable turn of the steering wheel in the correct direction.

Figure 5 shows a short section of an LCT trip highlighting the absolute value of the difference between the ideal and actual tracks.

Averaged values of **MDev** and **SDLP** (for the whole trip and for its lane keeping segments) are shown in **Table 3** and graphically compared in **Figure 8**. Detailed graphs for overall and lane-keeping driving statistics are shown in **Figure 6** and **Figure 7**, respectively, showing 95% confidence intervals.

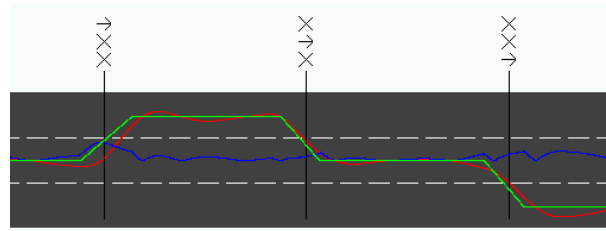


Figure 5. Visualization of (LCT) output. Red ~ actual track, Green ~ optimal track, Blue ~ absolute value of the difference.

Driver’s **reaction times** to lane change signs were measured as follows: the actual car path was examined in the range between 35m before the sign (sign visibility) and 20m after the sign. The first noticeable steering wheel movement of angle greater than 1.5° was identified and its start was considered as the time of driver’s reaction to the sign. Reaction times shown in **Table 3** and in **Figure 8** were computed by subtracting the time the sign became visible from the time of driver’s reaction.

Table 3: Detailed values of MDev, SDLP, and reaction time.

Task	Overall Mdev (m)	Overall SDLP (m)	Lane keeping MDev (m)	Lane keeping SDLP (m)	Reaction time (sec)
Undistracted	0.28	0.25	0.22	0.10	0.47
Eyes-free	0.38	0.44	0.26	0.12	0.60
Full msg view	0.45	0.62	0.32	0.13	0.67
Strip view	0.45	0.60	0.28	0.13	0.72
Cell (SMS)	0.64	0.68	0.56	0.28	0.74
GPS (address)	0.52	0.57	0.42	0.23	0.70

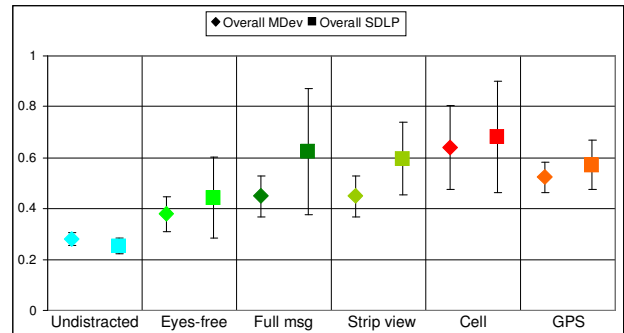


Figure 6: Overall MDev and SDLP



Figure 7: Lane keeping MDev and SDLP.

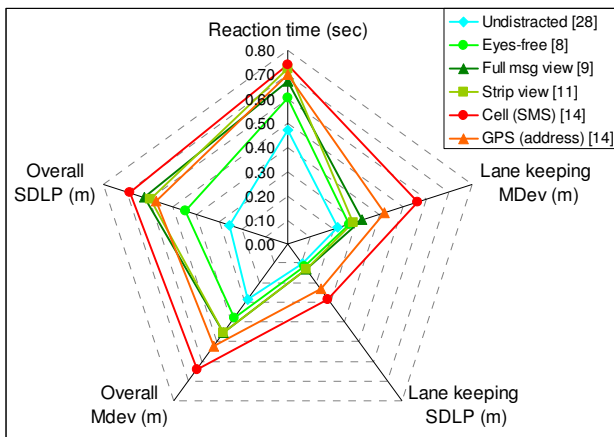


Figure 8. Radar graph for driving performance statistics. *ECOR* tasks (in green) are in most cases between undistracted (light blue) and reference tasks (red and orange). Numbers in brackets indicate number of subjects per task.

During the evaluation, no lane change signs were missed for the Undistracted (28 subjects) and *ECOR* Eyes-free tasks (8 subjects). One sign was missed for the *ECOR* Full View (9 subjects), and also for the GPS task (14 subjects). For the Strip view, we recorded 3 missed signs in total (11 subjects) and for the Cell phone task, there were 4 missed signs (14 subjects). Each LCT trip contained 18 signs. There were no out-of-the-road excursions throughout all evaluated drives.

Statistical significance in this paper was determined by the two-sample unequal variance Student's t-test for one-tailed distribution. Two sample means were considered significantly different for $p < 0.05$ (significantly different numbers are in green color in the following text).

For the **Driving quality** all *ECOR* setups were significantly less distracting than cell phone usage in most of the measures:

- **MDEV = 0.38-0.45m vs. 0.64m** ($0.006 < p \leq 0.029$)
- **SDLP = 0.44-0.62m vs. 0.68m**
- **LK.MDEV = 0.26-0.32 vs. 0.56** ($0.002 < p \leq 0.008$)
- **LK.SDLP = 0.12-0.13m vs. 0.28m** ($p < 0.001$)
- **RT = 0.60-0.72s vs. 0.74s** ($p = 0.047$ for 0.60s)

The Eyes-free *ECOR* setup was significantly less distracting than both cell phone and GPS usage, except for overall SDLP for eyes-free *ECOR* vs. GPS:

- **MDEV = 0.38m vs. 0.64 and 0.52m** ($0.003 < p \leq 0.006$)
- **SDLP = 0.44m vs. 0.68 and 0.57m** ($p = 0.047$ vs. cell)
- **LK.MDEV = 0.26 vs. 0.56 and 0.42** ($0.001 < p \leq 0.002$)
- **LK.SDLP = 0.12m vs. 0.28 and 0.23m** ($p < 0.001$)
- **RT = 0.60s vs. 0.74 and 0.70s** ($0.029 < p \leq 0.047$)

All *ECOR* setups with GUI were (insignificantly) less distracting than GPS tasks in most of the measures:

- **MDEV = 0.45m vs. 0.52m**
- **SDLP = 0.60-0.62m vs. 0.57**
- **LK.MDEV = 0.28-0.32m vs. 0.42** ($0.000 < p \leq 0.025$)
- **LK.SDLP = 0.13m vs. 0.23m** ($p < 0.001$)
- **RT = 0.67-0.72s vs. 0.70s**

All secondary tasks led to higher distraction than during undistracted driving:

- **MDEV = 0.38-0.64m vs. 0.28m** ($0.000 < p \leq 0.014$)
- **SDLP = 0.44-0.68m vs. 0.25m** ($0.000 < p \leq 0.026$)
- **LK.MDEV = 0.26-0.56m vs. 0.22m** ($0.000 < p \leq 0.015$, except for 0.26m for the eyes-free *ECOR*)
- **LK.SDLP = 0.12-0.28m vs. 0.10m** ($0.000 < p \leq 0.006$, except for 0.12m for the eyes-free *ECOR*)
- **RT = 0.60-0.74s vs. 0.47s** ($0.000 < p \leq 0.001$)

The radar plot in **Figure 8** shows data from **Table 3** in an easily comparable way. The best driving performance (closest to the center of the graph) was observed for the Undistracted task (light blue), followed by the Eyes-free *ECOR* task (light green), in most cases followed by the two *ECOR* tasks with display (in other two green colors). The GPS task (orange) is overlapping with both *ECOR* tasks with display for the Overall SDLP and for reaction time. The Cell phone task (in red) was associated with the worst driving performance for all statistics.

6.2 Road Attention Evaluation

As another objective measure, we analyzed subjects' eye gazes during all distracted LCT trips by manually annotating their videos. **Figure 9** shows highest road attention for the Eyes-free *ECOR* task. Subjects spent 95% of their driving time looking at the road. For the reference tasks, cell phone and GPS, the average time spent looking at the road was as low as 56% and 57%, respectively, whereas the subjects performing *ECOR* tasks with display had their eyes on the road for 78% and 79% of the time. Apart from the text entry UI, the steering wheel attracted eyes for about 5% of the time for all *ECOR* tasks, mainly due to the novice subjects checking positions of navigation buttons. The remaining attention (not shown in the graph) was directed elsewhere.

As shown in **Table 4**, the average number of gazes at the device display (application) was about 70% higher for the reference tasks (166 and 170 gazes per LCT trip) than for the *ECOR* tasks with display (94 and 97 gazes). If we count also the gazes at the steering wheel buttons, we get similar numbers (144 and 157 gazes) for *ECOR* with GUI and reference tasks. But note the huge difference in the average duration of the gazes: it is more than half a second (643 and 613 msec) for the reference tasks in comparison to 367 and 346 msec of "on-application" gazes and 255 and 230 msec of "on-steering-wheel" gazes for *ECOR* with GUI tasks.

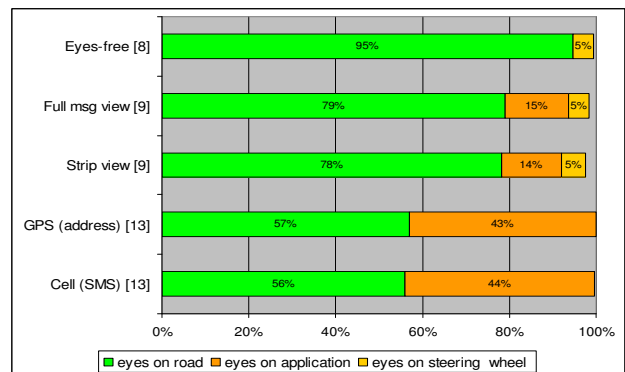


Figure 9. Road attention evaluation, manual annotation total duration of eye-gazes (road, application, steering wheel). Numbers in brackets indicate number of subjects per task.

Table 4: Number of gazes and average gaze duration on application and on steering wheel buttons.

Task	Gazes at the application		Gazes at the steering wheel	
	Avg. number of gazes	Avg. gaze duration (ms)	Avg. number of gazes	Avg. gaze duration (ms)
Eyes-free	-	-	51	255
Full msg view	94	367	50	230
Strip View	97	346	60	221
Cell (SMS)	166	643	-	-
GPS (address)	170	613	-	-

The intuition is that shorter gazes are less distracting than the longer ones. Based on gaze evaluation, we conclude that the visual distraction caused by reading text appearing on the display had significant adverse effects on driving. When available, subjects looked at the display even though all information was available via acoustic feedback. The t-test of statistical significance for road attention clearly ($0.000 < p \leq 0.005$) determined the following three groups: the eye-free ECOR task in the first one, full message and strip view ECOR tasks in the second group, and both reference tasks (cell phone and GPS) in the third group.

6.3 Message Composition Evaluation

In addition to driving performance, we also evaluated the subjects' performance of composing messages. For distracted driving trips except GPS, we collected texts of composed messages. These texts were analyzed manually by a single test conductor and each message was scored on a 0-1 scale. High scores were assigned to messages that contained all prescribed semantic content and did not contain undesired text. The major scoring factor was the semantic understandability of the message. Typos that could be easily decoded had minimal impact on the scores. For example if requested to buy a chocolate, wine and oranges, the message: "Please by chocolate and bag of oranges" had a rating of 0.75 since one information item ("wine") was missing out of a total of 4 prescribed items: "buy something", "wine", "oranges", and "chocolate". The confusion of "by" for "buy" was disregarded.

Table 5: Text entry speed and quality, numbers of editing operations and dictations needed to compose a message, and average word error rate (WER) per task. Numbers in brackets indicate number of subjects per task.

Task [subjects]	Avg. msgs sent	Operations per message	Dictations per message	Msg. quality	Avg. WER
Eyes-free [8]	5.5	3.2	1.8	96%	14.1%
Full msg view [9]	5.0	4.2	2.8	92%	14.7%
Strip View [11]	4.1	4.0	2.6	97%	16.2%
Cell (SMS) [11]	4.4	-	-	92%	-

Table 5 shows the average number of messages sent out during one LCT trip. We can see that using voice to dictate messages was on average faster than typing using cell phone. At the same time, the quality of messages did not significantly differ across all tested setups. This is a different finding from our previous study [9] with non-native speakers, who had significantly higher dictation error rates. In that experiment, subjects achieved lower text quality for all dictation tasks compared to cell phone texting, and the eyes-free task showed slightly lower text quality, perhaps due to some

errors not being noticed from the audio feedback. On the other hand, the subjects in our previous study were not selected as frequent message senders and were thus able to send significantly less messages from their cell phones than by dictating.

In the present study as well as in the previous one, subjects were able to send slightly more messages without the display, which we attribute to the subjects not spending extra time checking the display (perhaps both the text and visual UI elements).

The types of editing operations collected included dictation inputs, voice commands including "send it", deletion, undo, browsing mode selection, and n-best application. The amount of editing operations was least for the eyes-free setup. One reason for this could be that in the eyes-free setup, subjects mostly re-dictated the whole last utterance (1 operation) instead of navigating and replacing individual words (multiple operations). Another reason could be that people did not tend to fine-tune text details when they lacked visual feedback.

The number of sent messages did not statistically differ for the three ECOR setups (eyes-free, full message, and strip view) and there were no significant differences in semantic quality of text across all ECOR setups and cell phone for native speakers. The average word error rate (14.1-16.2%) for individual ECOR setups did not differ with statistical significance and even though we did not ask the subjects explicitly they seemed to be satisfied or even positively surprised with the ASR performance.

6.4 Subjective Evaluation

After performing each measured distracted task during an LCT trip, the subjects were asked to fill-in the System Usability Scale (SUS) questionnaire [2], the NASA Task Load Index (NASA-TLX) questionnaire [5], and to answer four additional questions.

The additional task-related questions are shown at the bottom of Figure 10. Users answered each question with a score between 0 and 6 points (more points corresponded to a more usable, accurate, or appealing UI). The graph shows average summed scores to these questions. The question being most relevant to driving performance is question Q2, which shows a significant difference between all three ECOR tasks and the two reference tasks (GPS and Cell phone). There was no significant difference between all three ECOR tasks for all the questions, the subjects however felt the ECOR Eyes-free task was slightly more accurate (Q3) than the other ECOR tasks with display. One reason for this could be that some ASR errors might not have been noticed by the users just from the audio playback when display was not present; on the other hand, message quality was not lower for the eyes-free version of ECOR as shown in Section 6.3.

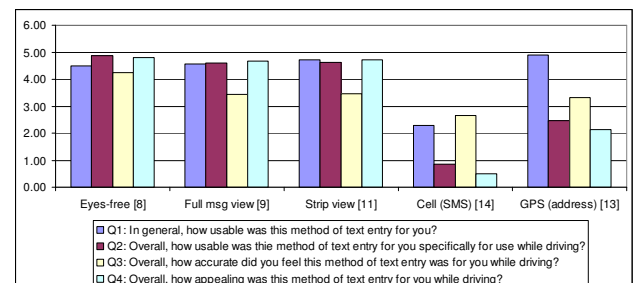


Figure 10: Average summed scores collected for 4 subjective questions about usability and accuracy. Range for each question was 0-6, 0=less, 6=more usable/accurate/appealing. Numbers in brackets indicate number of subjects per task.

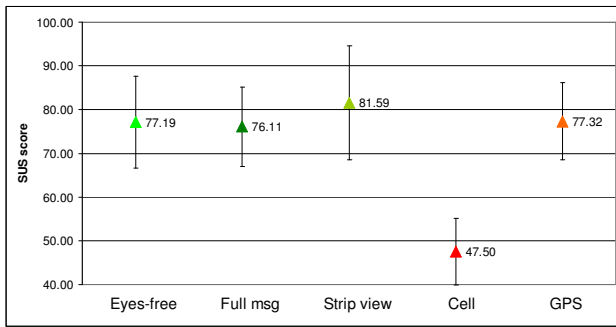


Figure 11: Average System usability scale (SUS) rating for individual tasks. Higher is better.

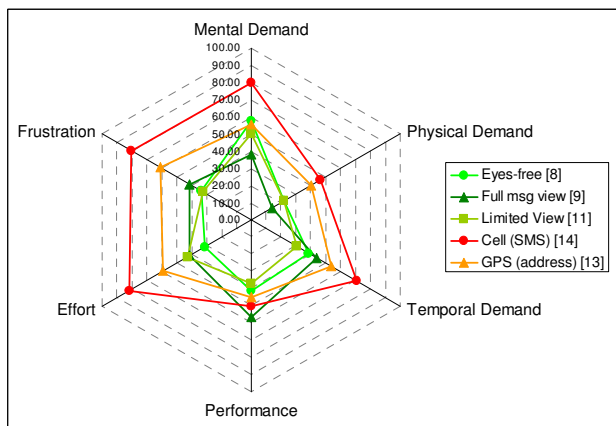


Figure 12: Subjective evaluation using NASA-TLX methodology; closer to the center of the graph means better. Numbers in brackets indicate number of subjects per task.

Figure 11 presents the average System Usability Scale (SUS) ratings for the individual tasks. Cell phone texting was the only significantly different outlier in the SUS rating. Using an A through F usability scale described in [14], the *ECOR* systems and GPS scored A- or B+ levels whereas the Cell phone texting task fell close to the worst F level ranking. The SUS score also correlated well with the additional notes and verbal comments made by the subjects; i.e. they liked the *ECOR* system and they were quite confident with the GPS task. In terms of statistical significance, only the cell phone received significantly worse rating than all other UIs ($0.000 < p < 0.002$).

The last subjective measure evaluation to be presented is the NASA Task load Index (NASA-TLX) rating shown in **Figure 12**.

This radar graph shows averaged ratings of the individual NASA-TLX dimensions for all distracted tasks (using the same color scheme as for driving performance evaluation). Again, the order of the ratings is the same as for the objective measures: all *ECOR* tasks were slightly better than GPS and significantly better than Cell phone texting.

Note that among all *ECOR* tasks, the Eyes-free version received the highest (worst) rating for Mental Demand. This means that not seeing what is dictated is mentally demanding, but it seems to “pay off” in most of the other indicators measured by both subjective and objective measures reported in this paper.

When comparing all *ECOR* setups against the cell phone and GPS, most dimensions of the NASA-TLX reached statistical

significance with the exception of Performance and Mental demand for both reference tasks, and the Temporal demand for GPS.

7. CONCLUSION AND FUTURE WORK

We presented a system for dictation and error correction of short messages intended for in-car use (codenamed *ECOR*). We evaluated three different *ECOR* settings: (1) Eyes-free system without any GUI, (2) Full message view system showing the whole dictated message on a screen, and (3) Strip view system showing only the last dictated, browsed, or corrected word; along with its immediate context. We evaluated these three *ECOR* systems against an undistracted LCT driving task and against two reference tasks: composing text messages using a cell phone and using a GPS car navigation device to enter an address.

Unlike [8], we consider the eyes-free use of an automotive dictation system possible by solving the problems of error detection and correction. To explore the space between *ECOR* Eyes-free and the *ECOR* Full message view systems, we introduced the *ECOR* Strip view version (GUI shown in **Figure 2**) with the hope that limiting the amount of displayed text and accompanying animations can be less distracting to the user than showing the whole dictated message. Even though the *ECOR* Strip view was rated slightly better in the subjective evaluation than the *ECOR* Full message view (for SUS rating and some of the additional task-related questions), the objective measures fell in the same range for both *ECOR* tasks with display.

We also compared the results based on a subjects’ gender and on the order of the tasks. Despite [12], we observed no significant differences between female and male subjects. There was also no significant difference between the tasks that were performed as first and as second.

We reported objective statistics describing driving performance during LCT trips and road attention statistics based on annotated video recordings, as well as results for subjective measures including the System Usability Scale (SUS) rating, NASA Task Load Index (NASA-TLX), and four additional questions related to the text entry task.

Also note that all reported tests included novice users only; experienced users are expected to perform significantly better.

The conclusions from our evaluation study are that (a) all *ECOR* settings were far below cell phone usage with regards to distraction, (b) the Eyes-free version of *ECOR* was less distracting than the GPS task, (c) all *ECOR* settings with display were in the same range or better than the GPS task with regards to distraction, (d) both message throughput and the quality of composed text were comparable between all *ECOR* settings and the cell phone (subjects were selected as frequent cell phone texters).

The experiments showed that the best *ECOR* setup with respect to driving performance was the eye-free version (no graphical UI at all). On the other hand, many users expected some form of feedback on the screen, as they were used to it from other in-car devices and appliances. They also asked for a possibility to double-check what exactly was recognized in addition to hearing it via playback, as well as for contextual information about their position in the current text. However, as the duration of eye gazes needs to be minimized, we recommend to provide a simple GUI that only displays the information which is most likely to be of interest to the user. Good support for orientation in the overall dictated text can either be provided by showing the strip view

together with a graphical indicator depicting the text length and the current position in it. An alternative could be to display the strip and full text views in parallel. Another frequent request from our test subjects was to have contextual help at their disposal at any point. In our case this was done by displaying appropriate hints in the teleprompter area.

Future work will include implementation and evaluation of additional designs of the *ECOR* application with the aim of finding a less-attention demanding GUI and more intuitive means to control of the system. We also plan to include other languages. Finally, we consider use of a more sensitive driving simulator such as the STISIM/PDT test discussed in [13] and the use of an automatic eye-tracker.

8. ACKNOWLEDGMENTS

We would like to thank to Jindřiška Hinge who helped us annotate video recordings with eye gaze information and our colleagues from IBM and Nuance for their support.

9. REFERENCES

- [1] Barón, A., Green, P.: Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review. Technical Report UMTRI-2006-5. University of Michigan Transportation Research Institute, 2006.
- [2] Brooke, J.: SUS: a “quick and dirty” usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & A. L. McClelland. *Usability Evaluation in Industry*. London: Taylor and Francis, 1996.
- [3] Bruyas, M.P., Brusque, C., Auriault, A., Tattegrain, H., Aillerie, I., Duraz, M.: Impairment of Lane Change Performance due to Distraction: Effect of Experimental Contexts. Proc. European Conf. on Human Centered Design for Intelligent Transport Systems, 2008.
- [4] Harbluk, J.L., Mitroi, J.S., Burns, P.C.: Three Navigation Systems with Three Tasks: Using the Lane-change test to Assess Distraction Demand. Proc. 5th Intl. Driving Symp. on Human Factors in Driver Assessment, Training and Vehicle Design, 2009.
- [5] Hart, S., Staveland, L.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock & N. Meshkati (Eds.), *Human mental workload*, pp. 139-183. Amsterdam: North Holland, 1988.
- [6] Hosking, S., Young, K., Regan, M.: The effects of text messaging on young novice driver performance. In: I. J. Faulks et al. (Eds.) *Distracted driving*, pp. 155-187. NSW: Australasian College of Road Safety, 2007.
- [7] Jamson, AH., Westerman, SJ., Hockey, GR., Carsten, OM.: Speech-based E-mail and driver behavior: effects of an in-vehicle message system interface. *Hum Factors*. 2004 Winter; 46(4): 625-39.
- [8] Ju, Y.C., Paek, T.: A Voice Search Approach to Replying to SMS Messages in Automobiles. *International Speech Communication Association*, 2009.
- [9] Labsky, M., Macek, T., Kleindienst, J., Quast, H., Couvreur, C.: In-car Dictation and Driver's Distraction: a Case Study. In: J. Jacko (Ed.) *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, Lecture Notes in Computer Science, Volume 6763/2011, pp. 418-425, Springer Berlin / Heidelberg, 2011.
- [10] Mattes, S.: The Lane-Change-Task as a Tool for Driver Distraction Evaluation. Proc. Annual Spring Conference of the GFA/ISOES 2003.
- [11] Medenica, Z., Kun, A.: Comparing the Influence of Two User Interfaces for Mobile Radios on Driving Performance, Proceedings of the 4th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Stevenson, Washington, July 9-12, 2007.
- [12] Petzoldt, T., Bär, N., Krems, J.F.: Gender Effects on Lane Change Test Performance. Proc. 5th Intl. Driving Symp. on Human Factors in Driver Assessment, Training and Vehicle Design, 2009.
- [13] Ranney, T.A. et al.: Measuring Distraction Potential of Operating In-Vehicle Devices. Technical Report DOT HS 811 231. U.S. Dept. of Transportation, National Highway Traffic Safety Administration, Dec. 2009.
- [14] Sauro, J.: Measuring Usability with the System Usability Scale, <http://www.measuringusability.com/sus.php>, 2011.
- [15] Shelton, L.: Statement before the Subcommittee on Highways and Transit, Committee on Transportation and Infrastructure, U.S. House of Representatives, May 9, 2001. <http://www.nhtsa.dot.gov/nhtsa/announce/testimony/distracti%20testimony.html>
- [16] Stutts, J., Reinfurt, D., Staplin, L., Rodgman, E.: The Role of Driver Distraction in Traffic Crashes, May 2001. <http://www.aaafoundation.org/pdf/distractio%20pdf>
- [17] Suhm, B., Myers, B., Waibel, A.: Multimodal error correction for speech user interfaces. Proc. ACM Transactions on Computer-Human Interaction (TOCHI) 2001.
- [18] Oviatt, S. et al.: Designing the user interface for multimodal speech and pen-based gesture applications: State-of-the-art systems and future research directions. *HCI*, Vol. 15 (4), p.263-322, 2000.
- [19] Vertanen, K.: Speech and Speech Recognition during Dictation Corrections. Proc. Interspeech 2006.
- [20] Young, K.L., Regan, M., Hammer, M.: Driver Distraction: A Review of the Literature. Monash University Accident Research Centre. Report no. 206, 2003.
- [21] Wang, Y., Chen, Ch., Cao, J., Zhang W.: Measurement of Degradation Effects of Mobile Phone Use on Driving Performance with Driving Simulation, Proceedings of the Eighth International Conference of Chinese Logistics and Transportation Professionals, 2008.