Long Text Reading in a Car

Ladislav Kunc[1], Martin Labsky[1], Tomas Macek[1], Jan Vystrcil[1], Jan Kleindienst[1], Tereza Kasparova[1], David Luksch[1], and Zeljko Medenica[2]

[1] IBM Prague Research and Development Lab, Prague, Czech Republic
{ladislav_kunc1,martin.labsky,tomas_macek,jan_vystrcil,
jankle,tereza.kasparova,david.luksch}@cz.ibm.com
[2] Nuance Communications, Inc., Burlington, United States
Zeljko.medenica@gmail.com

Abstract. We present here the results of a study focused on text reading in a car. The purpose of this work is to explore how machine synthesized reading is perceived by users. Are the users willing to tolerate deficiencies of machine synthesized speech and trade it off for more current content? What is the impact of listening to it on driver's distraction? How do the answers to the questions above differ for various types of text content? Those are the questions we try to answer in the presented study. We conducted the study with 12 participants, each facing three types of tasks. The tasks differed in the length and structure of the presented text. Reading out a fable represented an unstructured pleasure reading text. The news represented more structured short texts. Browsing a car manual was an example of working with structured text where the user looks for particular information without much focusing on surrounding content. The results indicate relatively good user acceptance for the presented tasks. Distraction of the driver was related to the amount of interaction with the system. Users opted for controlling the system by buttons on the steering wheel and made little use of the system's display.

Keywords: Architectures for interaction, CUI, SUI ad GUI, HCI methods and theories, Interaction design, Speech and natural language interfaces, Long text reading, car, UI, LCT.

1 Introduction

Drivers are well accustomed to listening to radio, music or audio books. The quality of machine synthesized speech is however still inferior to performance of a professional speaker reading out a text tailored for audio presentation. However, it is much slower, less flexible and more expensive to create such content.

The purpose of the study presented in this text is to learn to what extent the user is willing to cope with the deficiencies of text to speech synthesis (TTS).

Text processing is one of the activities humans do frequently. It ranges from passive reading to text creation, error correction and team collaboration. Users tend to shift most of their activities conducted previously on desktop to mobile environment. They even want to perform certain tasks in a car while driving. User interfaces for

M. Kurosu (Ed.): Human-Computer Interaction, Part II, HCII 2014, LNCS 8511, pp. 440–449, 2014.

mobile devices however have to respect a smaller form factor, less efficient input methods and distraction caused by using the system in a car. We addressed the tasks of text creation and correction in our previous work [2]. In this paper we focus on an apparently less difficult but important task of text reading.

2 Related Work

Significant attention was devoted in the past to assessing the impact of various in-car activities [1]. The Lane Change Test (LCT) [9] and subjective tests using questionnaires such as NASA TLX [5] and DALI [6], [7] are examples of popular methods used to assess the impact of various secondary in-car tasks on the primary task of driving.

Although electronic systems are more and more abundant in cars, which rightfully causes worries about their impact on driving, communication between the driver and passengers is frequent and hardly can be regulated [4]. The negative impact on driving performance due to having conversation with someone while driving was assessed by various studies [15], [16].

Several approaches to designing speech-based UIs for in-car usage including menu-based and search-based UIs were described [8], [10].

General quality of various TTS systems can be effectively measured only on the basis of reliable and valid listening tests, e.g. using mean opinion scale [18]. TTS quality was also assessed in terms of its suitability for various tasks such as computer assisted learning of foreign languages [17]. In this study we try to show that the quality of today's state-of-the-art TTS systems is sufficient for reading out texts in a car.

3 Research Goals and Experiment Design

The purpose of this study is to analyze the usability and distraction aspects of text reading in a car in general. The research questions that we search answers for are of three categories: usability, distraction and performance.

- **Usability:** Is the TTS quality sufficient for this kind of task? What part of the implemented functionality is actually used by the user? What are the preferred control mechanisms (buttons vs. swipe gestures, audio vs. visual feedback)? What are the preferred usage patterns (auto-playback vs. manual browsing through the text)? Is there correlation between the results and personal information about the subjects?
- **Distraction:** What levels of distraction can we observe for each of the tasks? How is distraction perceived subjectively? How often and for how long do the users look at the screen?
- **Performance:** Does the user remember what has been read?

We decided to carry out tests using three scenarios: 'Fable', 'News' and 'Car Manual'. They differ in the complexity of information, in the structure of the presented text and in the ways the user is allowed to interact with the text being read.

- **The fable scenario** represents a task of reading a plain unstructured text such as a short book chapter or an article. The user is only able to navigate within the text and may navigate by sentences and paragraphs.
- **The news scenario** involves reading multiple shorter texts (news articles). It demands more interactivity. The user can navigate between the articles or within the text of an article.
- **The car manual scenario** represents a complex task of looking for specific information in a car owner's user manual. It requires formulation of a query by the user, navigation in multiple search results and finally navigation in the retrieved user manual section to find the relevant piece of information. The user manual text was presented without modifications as extracted from a standard PDF car owner's manual.

Testing procedure consisted of the following steps. Initially, the whole procedure was explained to the participants. All training and evaluated drives were conducted at a constant speed of 60km/h on a standard straight 3-lane road in a Lane Change Test Simulator [9]. All drives were approximately 3.5 km long and took 3.5 minutes. First, our subjects trained the primary task of driving during a single drive and filled in a pre-test questionnaire. Prior to driving with secondary tasks, participants conducted one undistracted ride which was used to estimate an ideal LCT track adapted to each participant's style of driving. Another undistracted ride was conducted at the end of the testing session and was used as a reference to compare against distracted rides.

Training for each reading task was done shortly before evaluating it. The order of tasks was counterbalanced to compensate for a possible learning effect. Three distracted rides were conducted and each was followed by filling in the DALI [6] and SUS [13] questionnaires. In addition, for the car manual task, participants first searched by voice for a pre-specified topic, such as "turning fog lamps on", and only then they navigated through the retrieved set of articles to locate the relevant piece of information. For this task, participants also filled in an additional SASSI [14] form at the end of the drive.

Tests were conducted in a laboratory environment. The drivers were using a low-fidelity driving simulator to mimic driving on a highway. The primary task was performed using the standard LCT [9] used according to ISO 26022:2010 [12]. Fig.1 depicts the physical location of the devices during the experiment.

As a test bed, we used a prototype of an in-car infotainment system with a dedicated component for text reading (right part of Fig.1). We used the Nuance Vocalizer TTS system with a Premium US English voice named Ava.

The tested system presented text primarily through the audio channel via TTS playback. It allowed both for passive listening and for active navigation in the text using steering wheel buttons and touch screen swipe gestures. Participants could make use of up to 6 steering wheel buttons in a layout depicted in Fig.2, which allowed for advanced navigation in the presented text, including navigation between articles and within an article at the level of individual sentences and paragraphs.

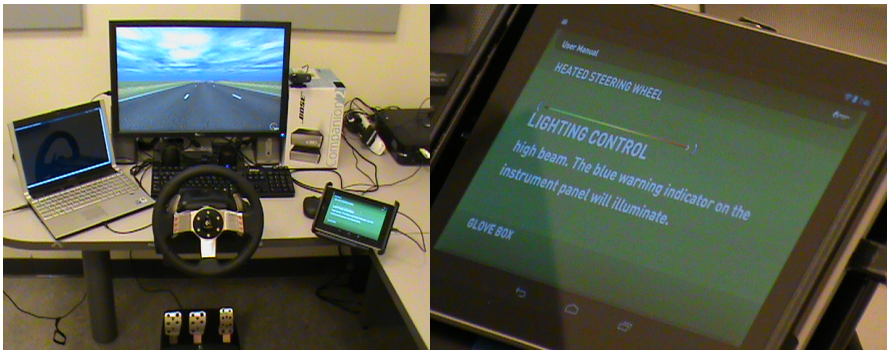

Fig. 1. Testing setup (left) and sample text shown on the system's display (right)

Control using swipe gestures was limited to navigation between articles only, using vertical swipe gestures. The double tap gesture activated speech recognition with automatic end-of-speech detection. Horizontal swipe gestures were reserved for swiping between different applications and thus were not used in this test.

Fig. 2. Steering wheel buttons layout

The visual presentation of the text on a display was intended as complementary information only. The display showed three lines of text in large fonts. The word currently being read was underlined. A progress indicator above the text showed the current reading position within a text block (e.g. news article).
Speech recognition (Nuance Vocon Hybrid) was used to search for relevant Car Manual articles.

4 Testing Results

The study was piloted with one subject and then conducted with 14 subjects in Burlington, USA. All participants were US English native speakers. Two subjects were

excluded due to an error in recording of the data. Half of the test subjects were females; all of them were driver's license holders, age varied from 20 to 55 with 7 participants under 29 years. 11 subjects drove and used radio daily; all had at least high school education.

We collected both usability feedback and objective distraction statistics, and also evaluated performance of test subjects on the reading task using simple reading comprehension tests.

Distraction. We measured driving distraction both objectively [9] and subjectively [6]. Fig.3 depicts objective measurements using LCT driving logs. We report SDLP (Standard Deviation of Lateral Position) and MDev (Mean Deviation) calculated both for the whole evaluated drive and for lane keeping segments only (excluding lane change segments).

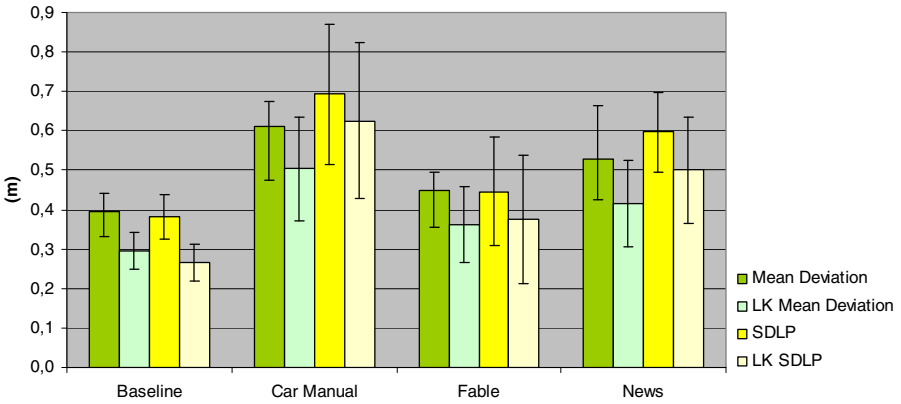

Fig. 3. Distraction measured objectively using mean deviation and SDLP; LK denotes lane-keeping versions of the statistics. 95% confidence intervals are shown.

There are statistically significant differences in the distraction for the news and car manual tests when compared to the undistracted ride. The fable was found to be the least distracting task with impact on driving that did not reach statistical significance with α=0.05.

Fig.4 depicts distraction measured subjectively using the DALI [6] test. The distraction ranking of tasks for all of the observed domains is the same as for the objective statistics. Fig.5 shows numbers of glances that each user made at the application screen. The counts vary. Some participants did not look at the screen at all, while others used the screen more frequently. Overall, the observed distraction results confirm the hypothesis that the more interactive tasks (car manual and news) cause more distraction.

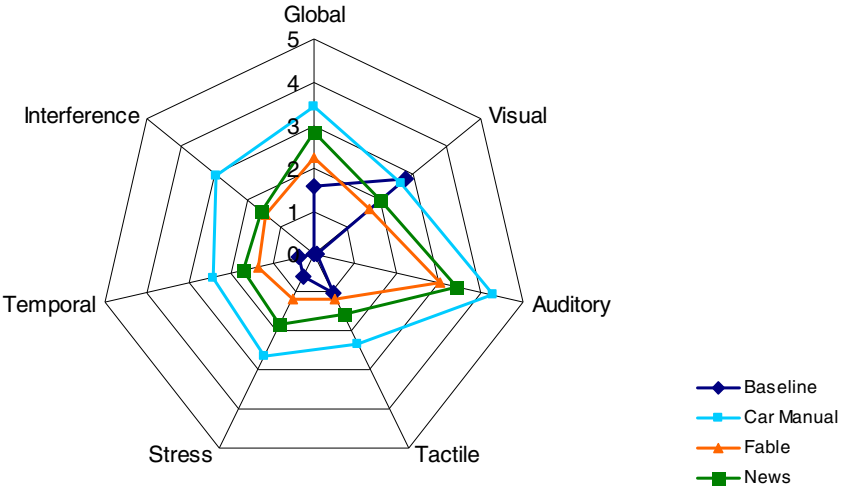

Fig. 4. Subjective distraction using DALI

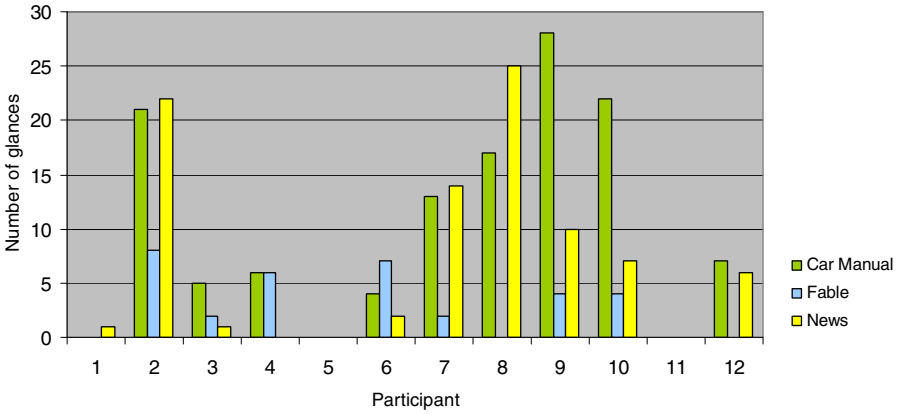

Fig. 5. Number of glances at the screen

The application screen was subjectively perceived as distracting. Most of the participants preferred to use the system without a screen or would move the screen to a position closer to the windshield. This finding is similar to the results presented in [11]. Most of the glances at the application screen occurred during the car manual task as the participants often checked the results of their voice search commands. The number of glances tended to be higher in the case when the retrieved content included irrelevant search results.

Performance. We evaluated efficiency of the system by asking participants several questions regarding the presented content at the end of each task, to verify that the content was understood and remembered.

The **car manual** test consisted of three tasks, each assigned immediately after the previous one was completed. Each task was rated successful or unsuccessful based on whether the user was able to find the requested information. Success rate was calculated as the number of successful tasks normalized by the total number of tasks (3).

The **fable** test was followed by asking three questions to the participant. The success rate was calculated as the number of right answers normalized by the number of questions (3).

The **news** test was evaluated as follows. For each article, three important facts were chosen that were expected to be remembered by the participants. The subjects were asked to repeat what they remembered and the experimenter could ask complementary questions. The success rate was the number of facts correctly remembered normalized by the overall number of facts (9).

The results in Fig.6 indicate that the tasks were reasonably complex. It may however still be problematic to compare the difficulty of the tasks using the achieved average success scores as they depend on the complexity of questions that were constructed subjectively. Overall, the mean values of success rate for the fable task were the highest (mean 89, deviation 0.16) followed by the news task (mean 75, deviation 0.18) and the car manual task (mean 67, deviation 0.24).

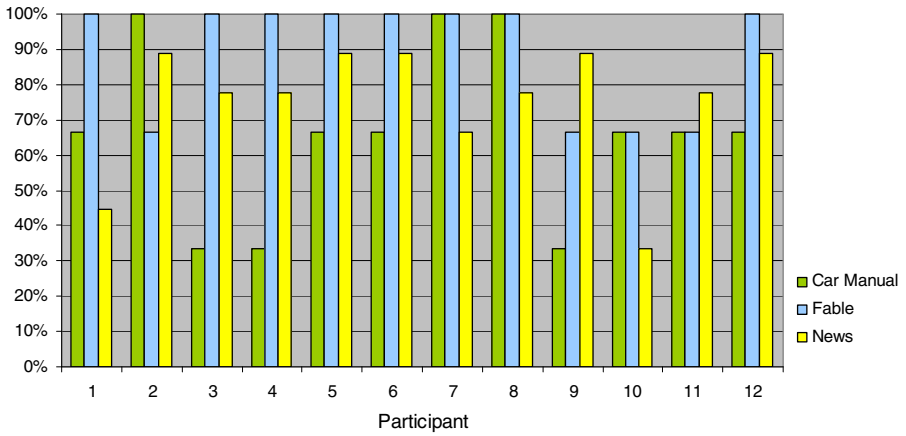

Fig. 6. Success rates for all tasks measured for all participants

Usability. The important part of the study was to collect usability feedback from participants; both about the text reading task in general and also concerning the utilized prototype. We collected feedback by analyzing video recordings, by interviewing the participants and by asking them to fill in several questions that were specific to each task.

The Car Manual task was also evaluated by collecting the SASI factors [14] as it was the only task that included search functionality that could be evaluated for accuracy. The scaled SASI factors are shown in Fig.7, indicating that part of the users considered it difficult to find a specific piece of information in a list of retrieved user manual sections.

In general, the subjects found the system useful. It was clear how to use it and easy to learn. Younger and more educated users liked the system more, and they performed better. The participants who were used to process information audibly (preferred radio to TV or newspaper) also performed better than others. We observed the reading process and analyzed how the users handled the related tasks of browsing and searching for specific content. We wanted to understand the degree to which the users exploit some of the advanced features of the prototype such as multiple browsing granularities or the way of displaying text on the screen.

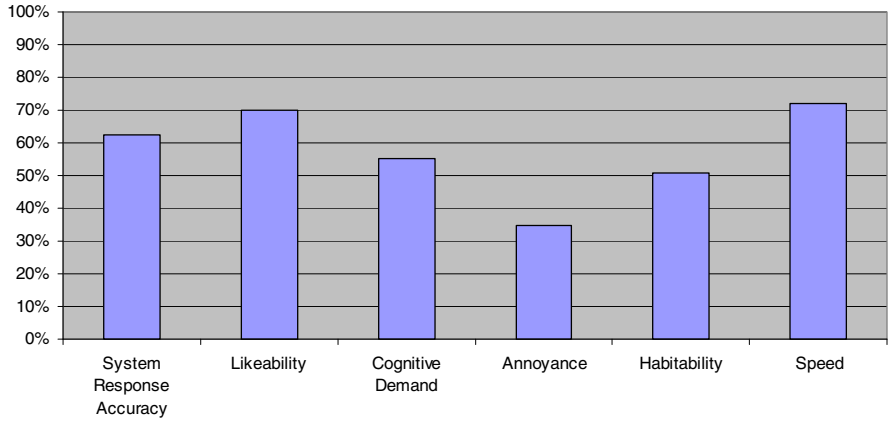

Fig. 7. Scaled SASI factors for the car manual task

Although some participants complained about the quality of the synthesized voice, they declared that they would like to use the system for reading of a wide spectrum of texts (news, emails, books and instant messages). They preferred not to use it for browsing car manual content in its original form. The reasons for that included distraction caused by navigating technical text, the limited suitability of the original user manual for presentation by voice and limited need to perform the task while driving. The users opted for controlling the UI by buttons on the steering wheel instead of using swipe gestures on a touch screen.

5 Conclusion

We presented here the results of a long text reading study performed on three types of texts. Although the number of tested subjects is relatively small to make a major

quantitative evaluation, the study provides useful qualitative observations and distraction estimates. A longitudinal study should be carried out to observe adoption of long text reading components in a daily driving scenario. We paid special attention to the impact of the content type presented and to the acceptance of TTS for each of the tasks. The study showed that most of the participants would use the system for reading various kinds of texts in spite of some complaints about the quality of TTS. The results suggest that car manual content in its original form may be appropriate for browsing and searching while parked, but other forms of presentation such as question answering should be considered for use while driving. The answers should be specially tailored for in-car presentation by voice. The application GUI was perceived as distracting. However some participants still used the GUI during the car manual tasks, mainly to verify the correctness of the retrieved search results.

Acknowledgment. This study was done jointly by Nuance and IBM as part of their joint research and development agreement.

References

1. Brostrom, R., Bengtsson, P., Axelsson, J.: Correlation between safety assessments in the driver-car interaction design process. Applied Ergonomics 42(4), 575–582 (2011)
2. Cuřín, J., Labský, M., Macek, T., Kleindienst, J., Young, H., Thyme-Gobbel, A., Quast, H., Koenig, L.: Dictating and editing short texts while driving: Distraction and task completion. In: Proceedings of the AutomotiveUI Conference. ACM, New York (2011)
3. Karat, J., Horn, H., Karat, C.: Overcoming unusability: Developing efficient strategies in speech recognition systems. In: Proceedings of CHI 2000 Conference, pp. 141–142. ACM, New York (2000)
4. Kun, A.L., Schmidt, A., Dey, A., Boll, S.: Automotive user interfaces and interactive applications in the car. In: Personal and Ubiquitous Computing, pp. 1–2 (2012)
5. Hart, S.G., Stayeland, L.E.: Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (eds.) Human Mental Workload. North Holland Press, Amsterdam (1988)
6. Pauzié, A.: A method to assess the driver mental workload: The driving activity load index (DALI). IET Intelligent Transport Systems 2(4), 315–322 (2008)
7. Pauzie, A.: Evaluation of Driver Mental Workload Facing New In-vehicle Information and Communication technology. IET Intelligent Transport Systems, Special Issue – selected papers from HCD (2008)
8. Yun-Cheng, J., Paek, T.: A Voice Search Approach to Replying to SMS Messages. In: Proc: INTERSPEECH 2009, 10th Annual Conference of the Intl. Speech Communication Association, Brighton, United Kingdom (2009)
9. Stefan, M.: The lane-change-task as a tool for driver distraction evaluation. In: Proceedings of the Annual Spring Conference of the GFA/ISOES, vol. 2003 (2003)
10. Labsky, M., Kunc, L., Macek, T., Kleindienst, J., Vystrcil, J.: Recipes for building voice search UIs for automotive. Submitted to EACL 2014 - Dialogue in Motion Workshop, Sweden (2014)

11. Vystrcil, J., Macek, T., Luksch, D., Labsky, M., Kunc, L., Kleindienst, J., Kasparova, T.: Mostly Passive Information Delivery - A Prototype. Submitted to EACL 2014 - Dialogue in Motion Workshop, Sweden (2014)
12. Road vehicles-Ergonomic aspects of transport information and control systems-Simulated lane change test to assess invehicle secondary task demand, International Standard ISO/DIS 26022:2010
13. Brooke, J.: SUS-A quick and dirty usability scale. In: Jordan, P.W., Thomas, B., Weerdmeester, B.A., McClelland, A.L. (eds.) Usability Evaluation in Industry, pp. 189–194. Taylor and Francis, London (1996)
14. Hone, K.S., Graham, R.: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). Natural Language Engineering 6(3-4), 287–303 (2000)
15. Kubose, T.T., Bock, K., Dell, G.S., Garney, S.M., Kramer, A.F., Mayhugh, J.: The effects of speech production and speech comprehension on simulated driving performance. Applied Cognitive Psychology 20(1), 43–63 (2006)
16. Drews, F.A., Pasupathi, M., Strayer, D.L.: Passenger and cell phone conversations in simulated driving. Journal of Experimental Psychology: Applied 14(4), 392 (2008)
17. Handley, Z.: Is text-to-speech synthesis ready for use in computer-assisted language learning? Speech Communication 51(10), 906–919 (2009)
18. Viswanathan, M., Viswanathan, M.: Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. Computer Speech & Language Iss. 19(1), 55–83 (2005)