INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Perceived usability and cognitive demand of secondary tasks in spoken versus visual-manual automotive interaction

Annika Silvervarg[1], *Sofia Lindvall*[1], *Jonatan Andersson*[1],
Ida Esberg[2], *Christian Jernberg*[2], *Filip Frumerie*[2], *Arne Jönsson*[1]

[1]Department of Computer and Information Science, Linköping University, Linköping, Sweden
[2]Volvo GTT, Göteborg, Sweden
annika.silvervarg@liu.se, ida.esberg@volvo.com

Abstract

We present results from a study of truck drivers' experience of using two different interfaces; spoken interaction and visual-manual interaction, to perform secondary tasks while driving. The instruments used to measure their experience are based on three popular questionnaires, measuring different aspects of usability and cognitive load: SASSI, SUS and DALI. Our results show that the speech interface is preferred both regarding usability and cognitive demand.
Index Terms: Evaluation of speech and multi-modal dialog systems, Multimodal human-machine interaction.

1. Introduction

Speech interfaces in cars were first introduced in 1996 when the Mercedez-Benz S-class car included a speech dialogue system for operation of the car's mobile phone, including number dialing (with connected digit dialog), number storing, user defined telephone directory entry name, name dialing, and directory editing [1]. Since then the development has continued and many large car manufacturers have speech interfaces in the cars for tasks like phone calls, navigation, infotainment and climate control.

The use of speech interfaces in vehicles is an important factor for improving safety. Many studies show that the use of cell phones or physical controls for infotainment systems or climate control are very distracting and often cause crashes. One investigation reports that 19% of crashes due to distraction were caused by the use of cell phones, adjustment of infotainment and climate control [2]. Other studies of real-world crashes and near-crashes have consistently demonstrated negative effects of visual distraction, for example, when dialing or texting on a cell phone [3, 4, 5, 6].

Although speech interfaces can increase safety [7], they are not unproblematic to introduce in vehicles. Bad usability can lead to worse performance, for instance, task completion time can increase for a speech interface since the system can misrecognise the input and the user may have to repeat himself or correct the system [2]. Bad usability can also lead to frustration and low user satisfaction.

User studies of how speech interfaces in cars are received show that although voice recognition is widely used it does not meet the customers expectations. Almost one-in-four U.S. motorists use voice recognition in their cars daily and 53% at least once a week. But audio, communication, entertainment and navigation (ACEN) systems are reported as the most problematic component category in today's new vehicles. The 2014 Multimedia Quality and Satisfaction Study conducted by J.D Power is based on responses from 86,118 new-vehicle owners surveyed between February 2014 and May 2014 [8]. The study measured the experiences and opinions of vehicle owners regarding the quality, design and features of their ACEN systems in the first 90 days of ownership. Problems with built-in voice recognition average 8.3 experienced problems per 100 vehicles.

Another aspect to consider is that even though speech often is more effective than visual-manual interaction in a vehicle, this interaction is a secondary task to driving and should affect the user as little as possible [9, 10]. Thus, it is very important to evaluate the usability of a speech interface for conducting secondary tasks in cars and consider its effect on the users primary task, which is safe driving, before adopting such systems. There have been a number of studies on safety and usability of speech interfaces for in-vehicle tasks while driving [11], but many of these have been conducted in simulators [12, 13, 14] or with Wizard of Oz set-ups [15, 16]. Very few have been conducted with real speech systems on roads, e.g. [17]. In this paper, we present results from a study of truck drivers' experience of using a state-of-the-art speech interface compared to a visual-manual interface for a variety of tasks in a naturalistic setting, when driving on a highway test track.

2. Evaluating automotive speech interfaces

When designing products in general, usability, i.e. how easy the interface is to use, is often stressed [18]. Learnability together with efficiency, memorability, errors and satisfaction are the five quality components of usability. For speech interfaces, another important factor is how to let the user know what voice commands are accepted by the system, i.e. the habitability of the system [19]. When evaluating a speech interface in the context of driving it also becomes very important to consider the cognitive demand of the interaction.

To investigate all theses issues we have chosen to combine three different questionnaires that address 1) various aspects of usability (SUS), 2) speech interfaces (SASSI) and 3) cognitive demand (DALI). DALI and SASSI have been combined before [20], but, as SASSI is primarily developed for speech interaction, c.f. [20], we also include the more generic usability oriented SUS questionnaire.

2.1. System Usability Scale (SUS)

The System Usability Scale (SUS) was initially designed to give usability practitioners a tool to quickly and easily assess the usability of a given system or product [21, 22]. The result was a questionnaire which nowadays has 10 items, see Table 1, an-

http://dx.doi.org/10.21437/Interspeech.2016-99

Table 1: The SUS questionnaire, with mean (M) and standard deviation (SD) for the speech (S) and the visual-manual interface (VM).

Question	S (M)	S (SD)	VM (M)	VM (SD)
1. The interaction with the system is consistent	4.09	1.24	3.91	0.90
2. It is clear how to interact with the system	5.45	1.08	3.45	1.08
3. It is easy to learn to use the system	5.82	1.19	4.45	0.89
4. I would use this system	6.64	0.64	3.91	1.83
5. I felt in control of the interaction with the system	5.45	0.99	3.73	1.60
6. I felt confident using the system	5.64	1.07	3.00	1.48
7. The system is easy to use	5.82	1.13	3.18	1.53
8. I always knew how to use the system	4.63	1.15	2.82	1.64
9. The system is simple	5.64	1.07	3.00	1.28
10. I found the various functions in the system were well integrated	5.27	0.86	3.36	1.07
SUS score (1-100)	74.09	13.32	42.36	16.17

Table 2: The SASSI questionnaire, with mean (M) and standard deviation (SD) for the speech (S) and the visual-manual interface (VM). Values for the factors are means for all questions including questions in SUS that coincide with questions in SASSI. The scale has been reversed for negative questions so that a high value always is positive.

Question	S (M)	S (SD)	VM (M)	VM (SD)
System Accuracy (incl SUS Q1)	3.82	0.96	3.91	0.77
1. The system makes few errors	3.55	1.16	3.91	1.38
Likeability (incl SUS Q2-Q5)	5.49	0.99	3.92	0.99
2. I was able to recover easily from errors	3.82	1.17	3.89	1.20
Cognitive demand (incl SUS Q6-Q7)	4.98	1.13	2.89	0.94
3. I felt tense using the system	5.55	1.67	2.82	1.03
4. I felt calm using the system	3.18	2.04	3.18	1.19
5. A high level of concentration is required when using the system	4.82	1.59	2.27	0.62
Annoyance	5.45	1.59	2.27	0.90
6. The interaction with the system is frustrating	5.45	1.59	2.27	0.90
Habitability (incl SUS Q8)	4.61	1.25	2.94	1.21
7. I sometimes wondered if I was using the right word	3.64	1.72	3.18	1.40
8. It is easy to lose track of where you are in an interaction with the system	5.55	1.44	2.82	1.03
Speed	4.45	1.51	3.09	1.38
9. The system responds too slowly	4.45	1.44	3.09	1.31
TOTAL	88.82	13.15	58.18	13.36

swered on a Likert scale, normally 1-5 (but 1-7 in this study). A SUS questionnaire has a score between 1-100 which will show how well a user appreciates a system's usability [23, 24]. The average score from 500 different studies [25] is 68 which is considered the median level of all SUS-scores. As such, anything under 68 is below average and anything over 68 is above average. A SUS score of 74 would fall into the B-grade interval, and the 70% percentile. To get the highest grade A, in the 90% percentile, you would need a score of at least 80.3. This is believed to be at the level where the user will recommend the system to a friend.

Despite only having 10 questions, research has shown that it is one of the most reliable questionnaires for assessing usability [22]. One of its strengths is that it is not limited to a specific product domain or area of use which makes it adaptable to any user-product relationship [26, 27]. Since its creation, the questionnaire has been validated and comes with a large database along with different ranking segments.

2.2. Subjective Assessment of Speech System Interfaces

Subjective assessment of speech system interfaces (SASSI) is a questionnaire for the purpose of accurately measuring a speech system's usability [28]. It strives to be discriminative, with many questions so that good and bad design aspects can be detected and used for further betterment of the system, and to be complete, capturing all important aspects of a user's experience with a speech system. It comprises a total of 34 questions divided in 6 factors [29]: System Response Accuracy (9 questions), Likeability (9 questions), Cognitive Demand (5 questions), Annoyance (4 questions), Habitability (5 questions), and Speed (2 questions). The participants rates their agreement on a Likert scale (7-points in this study).

The current state of SASSI shows promise, but is not yet a fully validated method for measuring usability of a speech system. Some prefer to instead use the PARADISE-framework [30] which combines subjective data with quantitative metrics such as task success. However, the PARADISE-framework's items chosen for user-satisfaction were not well-conducted or empirically based [31, 29] and PARADISE's way of summing all the test participants scores is questionable. Indeed, summing all participants scores into one would make it impossible to find differences between users, which may be an important factor.

An issue with SASSI is that there is no available database of reported SASSI results as the creators have not publicly released any data. This means that it is not possible to compare the SASSI results to a database, making it hard to know how good a system would be on the market. However, it is still usable to compare two different versions of a system as we do in our study. And using SASSI in conjunction with SUS adds the opportunity to compare our systems usability scores to validated scores.

Table 3: The DALI questionnaire, with mean (M) and standard deviation (SD) for the baseline (B) (only driving), the speech (S) and the visual-manual interface (VM).

Question	B (M)	B (SD)	S (M)	S (SD)	VM (M)	VM (SD)
1. The task required my attention	2.29	1.14	3.79	1.48	5.5	0.76
2. The task required visual demand	2.43	1.09	3.07	1.33	5.64	1.01
3. The task required auditory demand	1.43	0.65	3.79	1.58	2.64	1.22
4. The task required tactile demand	3.43	1.22	2.71	1.44	4.00	1.18
5. The task required temporal demand	1.93	0.99	3.07	1.38	4.86	1.75
6. It was hard to focus on driving while interacting with the system	2.64	1.39	3.79	1.76	5.71	0.91
7. I felt stressed using the system	2.00	1.11	2.71	1.64	5.07	1.44
TOTAL	17.00	8.51	22.93	7.85	33.43	5.16

A comparison of the questions in SUS and SASSI revealed that eight of the SUS items overlapped with items in SASSI (4 in the category Likeability, 2 in Cognitive Demand, and 1 each in System Response Accuracy and Habitability). Items 9 and 10 in SUS (see Table 1) had no comparable questions in SASSI. To complement SUS, but without using the complete SASSI questionnaire (due to avoiding questionnaire fatigue" and time constraints of the study), we therefore opted to use the SASSI questionnaire presented in Table 2. Notably, we added questions from the factors Annoyance and Speed, which did not have any corresponding questions in SUS. We also had to rephrase some questions to be more general and also applicable to visual-manual interaction, for example, using "interact with the system" instead of "speak to the system".

2.3. DALI

DALI (Driver Activity Load Index) derives from NASA TLX [32, 33] and is a questionnaire designed to measure subjective cognitive load when driving and performing a secondary task using an in-vehicle system. The DALI questionnaire focuses on task demands, effort of attention, interference and stress. Task demand is divided into visual, auditory, tactile and temporal demand. The questionnaire consists of seven statements, see Table 3, which are rated by the participants on a scale from 1 (Do not agree) to 7 (Agree completely). DALI is usually used to compare the effects of these factors while performing a secondary task while driving as compared to a baseline of just the primary task, i.e. driving.

3. Method

There were 14 participants in the study, all of which were men. The mean age was 46.6 *(SD=10.27)*. All of the participants were truck drivers, with C/CE driving licenses, employed by Volvo. Almost all of them used their smartphone several times a day, a navigation system once or twice a month, and about half of them never used a music player. Two of the test persons had previous experience of speech systems and used them once or twice a week.

The participants were tasked with performing secondary tasks through both a speech interface and the visual-manual counterpart (for example buttons and displays) while driving (with a randomised order of interfaces).

3.1. Secondary tasks

For both the speech interface and the visual-manual counterpart, the following tasks were carried out:

- Call your own phone number. Then call X from the phone book.
- Play Madonna, Like a prayer. Then ask the system to remind you to post the Declaration of income to the Tax Agency.
- Navigate to Vasagatan 15, Stockholm.
- Tell us the next time you need to take a break.
- Check your warning messages, vehicle message 2.

Task 1-3 are standard in-vehicle secondary tasks: phone, entertainment and navigation. Task 4 and 5 are tasks specific to truck drivers, since they have to take breaks within certain time intervals, and the warning messages were specific to trucks.

3.1.1. Speech interface

The voice recognition and text-to-speech interface uses Nuance VoCon/Nuance Vocalizer Expressive, see Figure 1. It can handle several languages but the system is programmed for English and Swedish. It consists of three programs that handle different aspects of the system. The speech recogniser used was online (which introduce a slight delay in processing) while the rest of the system was run in the vehicle. The system is integrated with the truck and uses the cluster display and speakers for interaction with the user.

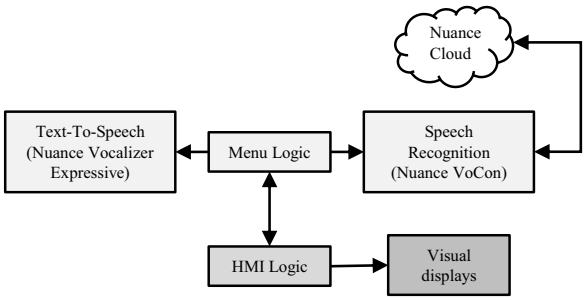

Figure 1: System overview. TTS, Menu logic and SR are run onboard the truck and linked through the HMI logic module with the trucks visual displays.

The user presses a Push-to-talk button (in this case the button was placed on the right-hand side of the arm rest). The system signals that it is listening via a chime and an indication on the cluster display. The user then issues a command and the truck responds through speech. Users can perform a task through a complete sentence, e.g. "U: Call Filip, S: Calling Filip" or stepwise, e.g. "I would like to make a phone call,

S: Who would you like to call?, U: Filip, S: Calling Filip". The system always ends a task by asking for confirmation. The available functionality included navigation, music, phone, information on warning messages that the car displays visually, access to the tachograph that tracks for how long the driver has been driving and when he needs to take a break.

3.1.2. Visual-Manual interface

For the Visual-Manual condition, the speech interface was replaced by a secondary visual display to the right of the steering wheel where the user could access his mobile phone, the entertainment system and the navigation system. For Task 2, the reminder was written down using paper and pencil. Task 4 and 5 required some usage of steering wheel buttons and visual readings of instruments, and a spoken answer to the test leader.

3.2. Procedure and instruments

Before the driving session, each participant executed a training session in the vehicle while standing still. The participants trained until they could perform the tasks without problems, thus ensuring that the study would not be about how easy it was to learn how to use the system, but rather to use a system you have already learned. The participants received oral information about the tasks from the test leader. The training tasks had the same complexity as the test tasks, but with different content.

To establish a baseline for the DALI questionnaire the participant started by driving on the test track for three minutes without executing any of the tasks. After the baseline drive, they were asked to stop and fill out the DALI questionnaire. Then the participants got to train on performing the secondary task using the speech and the visual-manual interfaces while driving. Both training and test was counterbalanced between the test drivers. The participants were then informed that the test started. They were instructed to carry out each of the five tasks using one of the interfaces (speech or visual-manual). Then, they stopped to fill out the three questionnaires (DALI, SASSI and SUS). Next, they were asked to drive and carry out five similar tasks with the other interface (visual-manual or speech), after which they filled out the three questionnaires again.

4. Results

The mean and standard deviation for the items in the questionnaires are summarised in Tables 1, 2, and 3. For the different factors in SASSI, the score is based on questions from both SASSI and SUS since overlapping questions were dropped from SASSI as they were included in SUS.

The SUS-score for the speech system ($M = 74.09, SD = 13.32$) was significantly higher ($t = 4.94, p < .001, r = 0.84$) than for the visual-manual system ($M = 42.36, SD = 16.17$). For results on each item see Table 1.

The overall SASSI score was also significantly higher ($t = 5.55, p < .001, r = 0.88$) for speech ($M = 88.82, SD = 13.15$) than for visual-manual ($M = 58.18, SD = 13.36$) interaction. There were significant differences ($p < 0.01$) for most factors, except Speed ($p = 0.09$) and System Accuracy ($p = 0.82$). For results on each item and factor see Table 2.

As for the DALI test, the Mauchly's test indicated that the assumption of sphericity had been met, $\chi^2 = 3.81, p > .05$. The results show that the raw DALI score was significantly affected by the type of task executed, $F(2, 26) = 26.05, p < .001, \eta_p^2 = .67$. Contrasts revealed that the DALI score for the baseline task ($M = 17, SD = 8.51$) was significantly lower than for the speech task ($M = 22.93, SD = 7.85$), $F(1, 13) = 11.53, p = .005, \eta_p^2 = .47$. The DALI score for the speech task was significantly lower than for the visual-manual task ($M = 33.43, SD = 5.16$), $F(1, 13) = 21.78, p < .001, \eta_p^2 = .63$.

5. Conclusions and discussion

We have presented results from a study where truck drivers drive a real truck, on a proving ground, using either a speech interface or a visual-manual interface to conduct a variety of secondary tasks normally carried out in trucks and cars.

The overall results show that the speech interface is preferred over the visual-manual on many accounts. For the selected tasks in this study, the speech interface has better overall usability, likeability and habitability. It requires lower cognitive demand and is considered less annoying.

The results of the SUS questionnaire show further that the speech interface has acceptable scores ($M \geq 5$) on most questions. The question "I would use this system" with a mean of 6.64 stand out, especially compared to the visual-manual interface with $M = 3.91$. The mean of 6.64 on a 7-grade scale, with a rather low SD, indicates that they really like to use the speech system. Other positive aspects are "It is easy to learn to use the system" ($M = 5.82$) and "The system is easy to use" ($M = 5.82$) which are central from a usability perspective and the safety critical domain. The visual-manual interface is perceived as more complex and not as clear how to interact with compared to the speech interface. Furthermore, the users are more confident using the speech interface.

Even if the speech interface overall is better than the visual-manual it only gets a grade B- on the SUS scale, which puts it in percentile rank of 70%. It is, thus, important to understand how the results of the study can be used to improve the design. Both the SUS and the SASSI items regarding habitability show that there is room for improvement in that aspect. The items in SASSI about error and error recovery also show that the speech system makes errors and that they are not always easy to recover from. Finally, the item regarding speed in SASSI shows that the system (at least sometimes) responds too slowly.

Both SASSI and DALI shows that the cognitive demand is much higher for the visual-manual interface (mean around or above 5 for most questions) than the spoken (means less than 4 for all questions). Especially items 1 and 6 in DALI regarding attention and distraction show that just driving, i.e. the baseline, scores around 2, the speech system has a mean of 3.79 for these, while the visual-manual interface scores means of 5.5 and 5.07. Thus, our results are in line with many previous studies on safety and speech interfaces.

The three questionnaires have been useful to compare the two interfaces and to get an overview of the usability of these. However, to use them in a formative way they can only be used to point out areas to investigate further, such as habitability and error recovery. They need to be complemented with interviews, observations and analysis of speech logs. The results from the DALI questionnaire should also be supplemented with a study of the drivers' behaviours, such as glances away from the road or other objective measures of cognitive load or distraction.

6. Acknowledgements

This research has been conducted in the RIVER FFI project, funded by Vinnova, Sweden's innovation agency, and Volvo. We want to thank all truck drivers that participated in the study.

7. References

[1] P. Heisterkamp, "Linguatronic product-level speech system for mercedes-benz cars," in *Proceedings of the First International Conference on Human Language Technology Research*, ser. HLT '01. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001, pp. 1–2. [Online]. Available: http://dx.doi.org/10.3115/1072133.1072199

[2] V. Ei-Wen Lo and P. A. Green, "Development and evaluation of automotive speech interfaces: Useful information from the human factors and the related literature," *International Journal of Vehicular Technology*, 2013.

[3] S. G. Klauer, T. A. Dingus, V. L. Neale, J. D. Sudweeks, and D. J. Ramsey, "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data." Washington, DC: National Highway Traffic Safety Administration., Tech. Rep., 2006.

[4] R. Olson, R. Hanowski, J. Hickman, and B. J., "Driver distraction in commercial vehicle operations." Department of Transportation, Federal Motor Carrier Safety Administration, Tech. Rep. Report No. FMCSA-RRR-09-042, 2009.

[5] J. Hickman, R. Hanowski, and J. Bocanegra, "Distraction in commercial trucks and buses: Assessing prevalence and risk in conjunction with crashes and near-crashes." Department of Transportation, Federal Motor Carrier Safety Administration, Tech. Rep. FMCSA-RRR-10-049, 2010.

[6] T. Victor, M. Dozza, and J. Bärgman, "Analysis of naturalistic driving study data: Safer glances, driver inattention, and crash risk," HRP 2 Safety Project S08A, Tech. Rep., 2014.

[7] L. Garay-Vega, A. Pradhan, G. Weinberg, B. Schmidt-Nielsen, B. Harsham, Y. Shen, G. Divekar, M. Romoser, M. Knodler, and D. Fisher, "Evaluation of different speech and touch interfaces to in-vehicle music retrieval systems," *Accident Analysis & Prevention*, vol. 42, no. 3, pp. 913 – 920, 2010, assessing Safety with Driving Simulators. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457509003364

[8] J. Power. (2014, August) Communication breakdown: Voice recognition No.1 problem with new vehicles. J.D. Power Reports, McGraw Hill Financial. [Online]. Available: http://www.jdpower.com/press-releases/2014-multimedia-quality-and-satisfaction-study

[9] N. Dahlbäck and A. Jönsson, "Dialogue systems when the dialogue is just a secondary task - some preliminaries to the development of in-car dialogue systems," in *Communication - Action - Meaning. A Festschrift to Jens Allwood*. Göteborg University, 2007.

[10] M. Blanco, W. J. Biever, J. P. Gallagher, and T. A. Dingus, "The impact of secondary task cognitive processing demand on driving performance," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 895 – 906, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0001457506000315

[11] A. Barón and P. A. Green, "Safety and usability of speech interfaces for in-vehicle tasks while driving: a brief literature review."

[12] C. Carter and R. Graham, "Experimental comparison of manual and voice controls for the operation of in-vehicle systems," in *Proceedings of the IEA 2000/HFES 2000 Congress, Santa Monica, CA: Human Factors and Ergonomics Society*.

[13] C. Forlines, B. Schmidt-Nielsen, B. Raj, P. Wittenburg, and P. Wolf, "A comparison between spoken queries and menu-based interfaces for in-car digital music selection."

[14] K. Itoh, Y. Miki, N. Yoshitsugu, N. Kubo, and S. Mashimo, "Evaluation of a voice-activated system using a driving simulator."

[15] B. Faerber, B. Faerber, and G. Meier-Arendt, "Speech control systems for handling of route guidance, radio and telephone in cars: Results of a field experiment," in *Vision in Vehicles - VII Proceedings, Amsterdam, Netherlands*.

[16] A. Gellatly and T. Dingus, "Speech recognition and automotive applications: Using speech to perform in-vehicle tasks," in *Proceedings of the Human Factors 32th Annual Meeting-1998, Santa Monica, CA: Human Factors and Ergonomics Society*.

[17] U. Gärtner, W. König, and T. Wittig, "Evaluation of manual vs. speech input when using a driver information system in real traffic," in *irst International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Aspen, CO*.

[18] J. Nielsen, *Usability engineering*. Elsevier, 1994.

[19] A. Jönsson, "A Model for Habitable and Efficient Dialogue Management for Natural Language Interaction," *Natural Language Engineering*, vol. 3, no. 2/3, pp. 103–122, 1997.

[20] H. Hofmann, V. Tobisch, U. Ehrlich, and A. Berton, "Evaluation of speech-based hmi concepts for information exchange tasks: A driving simulator study," *Computer Speech & Language*, vol. 33, no. 1, pp. 109–135, 2015.

[21] J. Brooke, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[22] T. S. Tullis and J. N. Stetson, "A comparison of questionnaires for assessing website usability," in *Usability Professional Association Conference*, 2004, pp. 1–12.

[23] A. Bangor, P. Kortum, and J. Miller, "Determining what individual sus scores mean: Adding an adjective rating scale," *Journal of usability studies*, vol. 4, no. 3, pp. 114–123, 2009.

[24] A. Bangor, P. T. Kortum, and J. T. Miller, "An empirical evaluation of the system usability scale," *Intl. Journal of Human–Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[25] J. Sauro and J. R. Lewis, "When designing usability questionnaires, does it hurt to be positive?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 2215–2224.

[26] P. Zaharias and A. Poylymenakou, "Developing a usability evaluation method for e-learning applications: Beyond functional usability," *Intl. Journal of Human–Computer Interaction*, vol. 25, no. 1, pp. 75–98, 2009.

[27] P. T. Kortum and A. Bangor, "Usability ratings for everyday products measured with the system usability scale," *International Journal of Human-Computer Interaction*, vol. 29, no. 2, pp. 67–76, 2013.

[28] K. Hone, "Usability measurement for speech systems: Sassi revisited," in *Designing Speech and Language Interactions Workshop, CHI 2014, Toronto, Canada*, 2014.

[29] K. S. Hone and R. Graham, "Subjective assessment of speech-system interface usability." in *INTERSPEECH*, 2001, pp. 2083–2086.

[30] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella, "Paradise: A framework for evaluating spoken dialogue agents," in *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997, pp. 271–280.

[31] M. Hajdinjak and F. Mihelič, "The paradise evaluation framework: Issues and findings," *Computational Linguistics*, vol. 32, no. 2, pp. 263–272, 2006.

[32] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage Publications, 2006, pp. 904–908.

[33] D. Kern, A. Mahr, S. Castronovo, A. Schmidt, and C. Müller, "Making use of drivers' glances onto the screen for explicit gaze-based interaction," in *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. ACM, 2010, pp. 110–116.