Analyzing crowdsourced ratings of speech-based take-over requests for automated driving

Bazilinskyy, P.; de Winter, J. C.F.

DOI
10.1016/j.apergo.2017.05.001

Publication date
2017

Document Version
Final published version

Published in
Applied Ergonomics: human factors in technology and society

Applied Ergonomics 64 (2017) 56–64

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

Analyzing crowdsourced ratings of speech-based take-over requests for automated driving

P. Bazilinskyy[*, 1], J.C.F. de Winter [1]

Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 13 February 2017
Received in revised form
18 April 2017
Accepted 2 May 2017
Available online 22 May 2017

Keywords:
Auditory displays
Autonomous driving
Speech perception
Human-automation interaction

ABSTRACT

Take-over requests in automated driving should fit the urgency of the traffic situation. The robustness of various published research findings on the valuations of speech-based warning messages is unclear. This research aimed to establish how people value speech-based take-over requests as a function of speech rate, background noise, spoken phrase, and speaker's gender and emotional tone. By means of crowdsourcing, 2669 participants from 95 countries listened to a random 10 out of 140 take-over requests, and rated each take-over request on urgency, commandingness, pleasantness, and ease of understanding. Our results replicate several published findings, in particular that an increase in speech rate results in a monotonic increase of perceived urgency. The female voice was easier to understand than a male voice when there was a high level of background noise, a finding that contradicts the literature. Moreover, a take-over request spoken with Indian accent was found to be easier to understand by participants from India than by participants from other countries. Our results replicate effects in the literature regarding speech-based warnings, and shed new light on effects of background noise, gender, and nationality. The results may have implications for the selection of appropriate take-over requests in automated driving. Additionally, our study demonstrates the promise of crowdsourcing for testing human factors and ergonomics theories with large sample sizes.

1. Introduction

1.1. Take-over requests

Until cars can drive autonomously, there will be situations where the driver has to resume manual control. Prior to such control transition, the automation may issue a take-over request to the driver (SAE International, 2016; Zeeb et al., 2015). How to provide a take-over request is a widely studied topic in human factors and ergonomics (Hergeth et al., 2015; Naujoks et al., 2014; Petermeijer et al., 2016; Pfromm et al., 2015).

A take-over request can be provided through pre-recorded voice (Gold et al., 2015; Mok et al., 2015; Politis et al., 2015), which may be an effective approach because humans are able to perceive sounds irrespective of head or eye orientation (Bazilinskyy and De Winter 2015). In aviation, a similar approach is used: traffic alert and collision avoidance systems (TCAS), which are mandatory in today's aircraft, apply voice commands (Kuchar and Yang, 2000).

Take-over situations may be of different urgency. Several studies have measured driver behavior in highly urgent situations, such as Mok et al. (2015), who found that 50% of the drivers veered off the road when a critical lane-closure event followed only 2 s after a take-over request ("Emergency, Automation off"). Other studies have been concerned with larger lead times of 5 or 7 s (Gold et al., 2013; see Eriksson and Stanton, 2017; for an overview) or with discretionary transitions having a low urgency (Damböck et al., 2013; Merat and Jamson, 2009; Nilsson et al., 2013). Politis et al. (2015) found that participants reacted 1.3 s faster to urgent take-over requests ("Danger! Collision imminent; You have control!") than to non-urgent ones (e.g., "Warning! GPS signal weak; Want to take over?"). In sum, how to convey the right sense of urgency is regarded as an important topic in automated driving research.

* Corresponding author. Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands.
E-mail address: p.bazilinskyy@tudelft.nl (P. Bazilinskyy).
[1] The authors contributed equally.

http://dx.doi.org/10.1016/j.apergo.2017.05.001
0003-6870/

1.2. Speech warnings

Previous research has shown that semantics have an effect on urgency, in that a word such as 'danger' is perceived as more urgent than 'attention' (Arrabito, 2009; Baldwin, 2011; Wogalter and Silver, 1995; Wogalter et al., 2002). Second, emotional tone has important effects: phrases are considered more urgent if spoken in an urgently intoned style (Edworthy et al., 2003a; Ljungberg et al., 2012). Third, it has been found that the greater the speech rate, the higher the perceived urgency (Hollander and Wogalter, 2000; Jang, 2007; Park and Jang, 1999). No clear gender effects seem to exist: words spoken by a female typically yield similar urgency ratings as the same words spoken by a male (e.g., Hellier et al., 2002; Wogalter et al., 2002). However, Jang (2007) and Park and Jang (1999) found that a male voice yielded higher urgency ratings than a female voice. Furthermore, interaction effects have been observed, where the word "Note" received a higher urgency rating when spoken by a male instead of a female (Hellier et al., 2002). Differences in the degree of smoothness, pitch, and timbre may explain these gender differences (Edworthy et al., 2003a,b; Jang, 2007).

In addition to urgency, it is important to consider whether the message is comprehensible and pleasant. If people become displeased with a warning, they may ignore or disable the warning system, potentially causing unsafe situations (Eichelberger and McCartt, 2014; Parasuraman and Riley, 1997). A female voice has been regarded as more pleasant (Bazilinskyy and De Winter 2015; Machado et al., 2012) and is more often used in route navigation devices (Large and Burnett, 2013) than a male voice. The female and male voice are supposedly equal in terms in intelligibility, but it has been reported that the male voice is easier to understand in a noisy environment such as an aircraft cockpit (Nixon et al., 1998; Noyes et al., 2006). However, it is unknown whether this effect is replicable. Arrabito (2009) stated that "further research is required to study the effects of speech parameters and word semantics across multiple talkers of each sex for variations of urgency under different background noise sources" (p. 18).

There is currently an irony in automated driving, because the technologies are deployed in the highest-income countries, which already have commendable road safety statistics, while low-income countries account for the vast majority of fatal road traffic accidents (Gururaj, 2008; World Health Organization, 2015). At present, car manufacturers are exploring cross-national perceptions of warnings (Langlois et al., 2008), but it is unknown whether speech-based take-over requests should be differentially developed per country. Research has shown that there are national differences in how people perform at basic visual perception tasks (Henrich et al., 2010). Regarding the appraisal of sounds, similar differences may exist. For example, it has been found that the sound of a bell was rated as pleasant among German listeners (possibly because it yielded connotations to a church bell), whereas this sound was rated as dangerous and unpleasant among Japanese listeners (Fastl, 2006). One specific question is whether a speech-based warning should be tailored to the language and accent of the host country. For example, it is possible that drivers from the UK prefer a British accent, and drivers from the US prefer an American accident. It has been found that a foreign English accent does not reduce the intelligibility and comprehensibility of speech (Munro and Derwing, 1995; Munro, 2008; Smith and Rafiqzad, 1979), but these findings deserve further investigation.

1.3. Aim of the study

This paper assesses how different speech-based take-over requests are perceived. Specifically, in line with the above research gaps, we assessed (1) the effects of speech rate on perceived urgency, commandingness, pleasantness, and ease of understanding, for speakers that differ in gender and emotional tone. Additionally, we investigated (2) the effects of spoken phrase (semantic content) on perceived urgency for a male and female speaker, (3) the effects of noise on the ease of understanding, for a male and female speaker, and (4) the effect of participants' (i.e., listeners') gender on pleasantness. Finally, we explored (5) the relationship between the participants' country and the ease of understanding of the messages. To acquire a large sample, we used crowdsourcing, an approach that is gaining popularity (Bazilinskyy and De Winter 2015; Behrend et al., 2011; Buhrmester et al., 2011; Crump et al., 2013; Kyriakidis et al. 2015; Rand, 2012).

2. Methods

This research was approved by the Human Research Ethics Committee at the TU Delft under the ethics approval application titled "Rating audio messages by means of crowdsourcing" on May 24, 2016. Informed consent was obtained from each participant via a dedicated survey item.

2.1. Speech-based messages

Speech-based messages "Take over, please" were created using the online tool Acapela-Box (https://acapela-box.com). Acapela-Box reproduces the natural sound of language based on voice of human speakers, and was selected because it offers high-quality speech and adjustability of speech rate. Two male voices (Will: US English accent; Graham: UK English accent) and two female voices (Karen: US English accent; Deepa: Indian English accent) were used. These three English accents represent highly populated countries with a strong automotive industry where English is either the first language (US and UK) or one of the official languages (India). The tool offered the option for speech to be generated with an emotional tone. We created recordings for two emotional tones by selecting speakers Will Happy and Will FromAfar. We expected that Will FromAfar, in which the speaker shouts the words from a distance, would be interpreted as urgent. Will Happy was expected to sound pleasant among listeners. Note that Acapela-Box offered a limited number of speakers and emotional tones: there was no male voice with Indian English accent, and among the US English speakers, the Happy and Afar emotional tones were only available for Will. Furthermore, different voices exhibited different speech rates (e.g., Deepa spoke relatively fast).

Using Acapela-Box, each of the six speakers was recorded at eight additional settings of speech rate: −60, −45, −30, −15, +15, +30, +45, and +60, which altered the duration of the sample to approximately 151%, 131%, 119%, 109%, 90%, 85%, 79%, and 76% of its nominal value, respectively. In addition, for each speaker and speech rate, background noise was added, extracted from a YouTube video showing a Tesla Model S in Autopilot mode (Oedegaarde, 2015). For Will and Karen, noise with three extra levels of volume was added (Table 1).

Moreover, 13 phrases were recorded using Will and Karen at a nominal speech rate and without added noise: (1) "Take over please?", (2) "Take over", (3) "Please take over", (4) "Could you please take over", (5) "Could you please take over?", (6) "Take over now", (7) "Take over immediately", (8) "Hazard: take over", (9) "Danger: take over", (10) "Warning: take over", (11) "Caution: take over", (12) "Attention: take over", and (13) "Note: take over".

In summary, the number of recordings was 140, consisting of 108 recordings where speech rate and noise were varied for each of the six speakers (6 speakers x 9 speech rate levels x 2 noise levels) plus 32 recordings (3 noise levels and 13 additional phrases, for

Table 1
Overview of the sound samples for the phrase "Take over, please" at the nominal speech rate. Shown in parentheses is the sound volume when background noise was added to the original sample, for noise levels 1, 2, 3, and 4.

Speaker	Duration (s)	Maximum volume (0–1)	Mean volume (0–1)
Will	1.67	0.332 (0.384, 0.541, 0.816, 1.000)	0.038 (0.053, 0.108, 0.180, 0.311)
Karen	1.62	0.335 (0.435, 0.660, 0.919, 1.000)	0.051 (0.062, 0.113, 0.183, 0.312)
Graham	1.59	0.333 (0.357)	0.031 (0.047)
Deepa	1.28	0.266 (0.326)	0.029 (0.048)
Will Happy	1.78	0.337 (0.405)	0.044 (0.058)
Will FromAfar	1.96	0.046 (0.160)	0.009 (0.033)

Will and Karen).

2.2. Survey

A survey was developed using CrowdFlower (http://www.crowdflower.com). At the beginning of the survey, contact information of the researchers was provided, and the purpose of the survey was described as "to determine the public opinion on auditory messages that may be used in automated driving". Participants were informed that the survey would take 5 min of their time. The participants were also informed that they had to be at least 18 years old. Information about anonymity and voluntarily participation was provided as well.

The survey started with a question about whether the participant had read and understood the instructions, and contained questions on the participant's age, gender, driving experience, and opinion on automated driving. The main part of the survey focused on the voice recordings. Each participant was given a random selection of 10 out of the 140 voice recordings. The participants were asked to click on the recordings to listen to them. The filenames of the recordings were masked as voiceXXX.mp3 (with XXX being a number between 1 and 140).

Below each recording, five questions were provided: (1) "Did you listen to the recording of a female or male voice in recording XX?" (this was a test question), (2) "The message in recording XX is urgent.", (3) "The message in recording XX is pleasant.", (4) "The message in recording XX is commanding.", (5) "The message in recording X is easy to understand.", where XX denotes a number between 1 and 10. For questions 2–5, the response options were "Disagree strongly", "Disagree a little", "Neither agree nor disagree", "Agree a little", "Agree strongly", and "I prefer not to respond". The participants had to answer all questions in order to complete the survey. The survey did not explain the notions of urgency, pleasantness, commandingness, and ease of understanding to the participants.

2.3. CrowdFlower configuration

We allowed contributors from all countries to participate in the survey. Completing the survey more than once from the same CrowdFlower worker ID was not permitted. A payment of $0.14 was offered for the completion of the survey. We collected responses from 3061 participants, at a total cost of $524.

2.4. Analyses

Five analyses were conducted. First, we determined the effect of speech rate on the degree to which the message was regarded as urgent, pleasant, commanding, and easy to understand. Second, we evaluated the effect of the 14 different phrases on perceived urgency, and whether there were differences between the speakers of different gender (Will vs. Karen). Third, we assessed the effect of noise level on whether the message was easy to understand for each of the six speakers, and whether the male voice (Will) was easier to understand than the female voice (Karen) as a function of noise level. Fourth, we determined the effect of the participants' (i.e., listeners') gender on pleasantness, for each of the six speakers. Finally, we assessed whether the six speakers had different levels of comprehensibility for participants from different countries, with the participants' country being automatically identified by CrowdFlower. In order to arrive at statistically reliable conclusions, we included only those countries with 100 or more participants in the cross-national analyses.

All analyses were conducted at the level of participants. If multiple responses per condition were available per participant (e.g., responses to recordings with and without background noise for the same speaker and speech rate), then these responses were averaged per participant. The mean scores on a scale from 1 (*Disagree strongly*) to 5 (*Agree strongly*) were calculated and visualized in bar graphs. For each depicted mean, the 95% confidence interval was provided, defined as the mean ± 1.96 times the standard deviation divided by the square root of the sample size. Comparisons between selected pairs of conditions were conducted by means of independent-samples t tests. A previous simulation study showed that for five-point Likert data, the t-test provides appropriate statistical power and protection against false positives (De Winter and Dodou, 2010). In principle our experiment has elements of a within-subject design, because each participant rated multiple auditory samples. However, because each participant rated only 7.1% (=10/140) of random auditory samples, the probability was low that a participant rated a reference sample that could be used in a paired comparison. Therefore, we conducted between-subjects statistical analyses.

No corrections for multiple comparisons were applied, because our interest was not only in detecting whether effects are statistically significant, but also in showing whether effects are not statistically significant despite large sample sizes. In other words, if we had reduced the significance level to a value smaller than the nominal 0.05, then a finding of 'no statistically significant differences' would not be compelling.

3. Results

The responses were collected between 29 May 2016, 13:30 and 5 June 2016, 21:35 (GMT). Each of the 3061 participants answered four queries (urgency, pleasantness, commandingness, ease of understanding) regarding 10 voice recordings. 337 participants completed the optional user satisfaction survey. The satisfaction survey received an overall satisfaction score of 4.4 out of 5.0 (1 = *very dissatisfied*, 5 = *very satisfied*), with "instructions clear", "test questions fair", "ease of job", and "pay" receiving ratings of 4.6, 4.3, 4.3, and 4.1, respectively.

Participants who indicated they had not read the instructions ($N = 25$), were 17 or younger ($N = 3$), or whose country was not identified ($N = 3$) were excluded. As a data quality filter, participants who made one or more mistakes in the question 'did you

listen to the recording of a female or male voice?' were excluded ($N = 375$). Regarding this latter exclusion criterion, Deepa was not taken into consideration because we ourselves had difficulty identifying whether Deepa was male or female. A sizeable portion of participants also seemed to have difficulty distinguishing whether Deepa was male or female (there were 10% errors for Deepa versus 4% error for the other speakers). The results were hardly affected by the decision not to include Deepa in this filtering process. In total 392 participants were excluded, leaving 2669 participants from 95 countries. The mean survey completion time was 580 s ($SD = 285$ s). The mean age was 33.7 years ($SD = 10.6$) and the sample consisted of 1777 males, 884 females, and 8 participants with unknown gender. These 8 participants selected 'I prefer not to respond' and were retained in the analysis.

The effects of speaker and speech rate are shown in Fig. 1. The higher the speech rate, the higher the ratings of urgency and commandingness. Averaged across the nine speech rates, Will FromAfar received the highest urgency ratings ($M = 3.34$) and Will the lowest ($M = 2.95$). Speech rate had non-monotonic effects on pleasure, showing different inverted U-shapes per speaker. The female speaker Karen was rated as most pleasant at low speech rates. Although Will FromAfar was rated as urgent, he was rated as least pleasant ($M = 2.44$) and least well understood ($M = 3.27$). It is possible that the low intelligibility of Will FromAfar was caused by its low volume (Table 1). The speaker with Indian accent (Deepa) did not receive high pleasure ratings either ($M = 2.77$); Deepa had a high speech rate in its nominal condition (Table 1), and higher speech rates were considered to be unpleasant.

Fig. 2 confirms that the spoken phrase has an effect on urgency, with "Take over immediately" and "Danger: take over" yielding the highest urgency and "Could you please take over?" the lowest. There were statistically significant gender differences, with "Take over, please", "Take over please?", and "Please take over" being perceived as more urgent when spoken by Karen than when spoken by Will, while the opposite was observed for "Take over now". It is worth noting that for the 14 spoken phrases, there was a positive correlation between the mean urgency and the mean commandingness ($r = 0.92$, $n = 14$), but a negative correlation between the mean urgency and the mean pleasantness ($r = -0.72$, $n = 14$). The highest commandingness was found for "Take over now" ($M = 4.40$) and the highest pleasantness was found for "Take over, please" ($M = 4.06$).

The results regarding noise are shown in Fig. 3. The t tests show that a mild noise level (Level 1 noise) has only minor effects on ease of understanding, except for Will FromAfar, which had a low volume without noise (Table 1). The ease of understanding dropped with increasing noise level (see Will and Karen in Fig. 3). In contrast to Nixon et al. (1998), the female voice (Karen) was easier to understand than the male voice (Will), especially at higher noise levels (Fig. 3). These gender differences were statistically significant; No noise: $t(2690) = -1.84$, $p = 0.065$; Level 1 noise: $t(2597) = -3.22$, $p = 0.001$; Level 2 noise: $t(378) = -2.79$, $p = 0.005$; Level 3 noise: $t(367) = -4.20$, $p < 0.001$; Level 4 noise: $t(359) = -2.90$, $p = 0.004$.

The results regarding the participants' gender revealed no statistically significant differences on pleasantness, for four of the six speakers, despite the fact that statistical power was high, with about 2000 degrees of freedom (Fig. 4). Deepa and Will FromAfar were rated as slightly more pleasant by male participants than by female participants.

Finally, we assessed national differences. The mean ease-of-understanding scores for the eight countries with 100 or more participants are shown in Fig. 5. The effects of speaker are consistent across these geographically diverse countries, with Will FromAfar being rated as difficult to understand, and Will, Karen and Will Happy receiving high scores. Deepa, who had an Indian accent, received higher ratings from Indian participants than from participants from other countries. To illustrate, the mean ease-of-understanding rating for Deepa was significantly greater for participants from India ($M = 4.20$, $N = 123$) than for participants from the USA ($M = 3.23$, $N = 138$; $t(259) = 6.88$, $p < 0.001$) and Venezuela ($M = 3.41$, $N = 239$; $t(360) = 5.84$, $p < 0.001$).

4. Discussion

This study determined how people value speech-based takeover requests as a function of speech rate, background noise, speaker (gender and emotional tone), and spoken phrase, by means of a crowdsourcing study with a large sample size. A total of 2669 participants completed the task over the course of 7 days.

There are several advantages to using a large sample size. First, a larger sample size increases statistical power, which means that if a research finding is true, it is more likely to be detected. Second, a larger sample size increases the probability that a research finding is in fact true. Third, if the sample size is larger, the results are less susceptible to bias (Gadbury and Allison, 2012; Ioannidis, 2005; Wagenmakers et al., 2015). In recent years, psychology has been said to be in a replication crisis (Maxwell et al., 2015). This concern was recently confirmed by the Open Science Collaboration (2015), showing that from 97 published significant effects, only 35 replicated. Small samples are a prime cause of poor replicability, a message that has now transpired to many fields, including medicine (Arrowsmith, 2011; Begley and Ellis, 2012; Freedman et al., 2015), economics (Ioannidis and Doucouliagos, 2013), and neuroscience (Button et al., 2013). Asendorpf et al. (2013) argued that "it cannot be stressed enough that researchers should collect bigger sample sizes, and editors, reviewers and readers should insist on them" (p. 110).

Our research replicated several published effects. In agreement with Park and Jang (1999), an increase of speech rate yielded an increase of self-reported urgency, an effect that held regardless of the gender or emotional tone of the speaker. In agreement with Hellier et al. (2002), amongst others, the spoken phrase (e.g., "Danger" versus "Note") had an important impact on perceived urgency as well. Overall, our results point to the robustness of published human factors and ergonomics research, and are in line with the idea that psychological effects generalize well across different research settings (Klein et al., 2014).

Several of our findings are in disagreement with the literature. First, Hellier et al. (2002) found that the word "Note" received higher urgency ratings when spoken by a male than when spoken by a female, whereas we found no statistically significant gender effect for "Note: take over". This discrepancy may be a consequence of the specific phrase and intonation. Perhaps our findings represent a social-psychological phenomenon in which direct utterances ("now") are deemed urgent when spoken by a male, whereas suggestive utterances ("please") are deemed urgent when spoken by a female (cf. Fig. 2). Second, Nixon et al. (1998) found that the intelligibility of female speech was lower than that of male speech, especially for strong cockpit noise, whereas our results showed the opposite, with the female voice being easier to understand under strong background noise. Third, the fact that a speaker with Indian accent was relatively easy to understand by listeners from India is in line with the 'interlanguage speech intelligibility benefit' (Bent and Bradlow, 2003; Podlipský et al., 2016), but appears to contradict published literature stating that "listeners did not consistently exhibit an intelligibility benefit for speech produced in their own accent" (Munro et al., 2006, p. 111). It is noted that we did not perform a direct replication of past research, but rather a conceptual replication (Stroebe, 2016). Our findings therefore do not refute

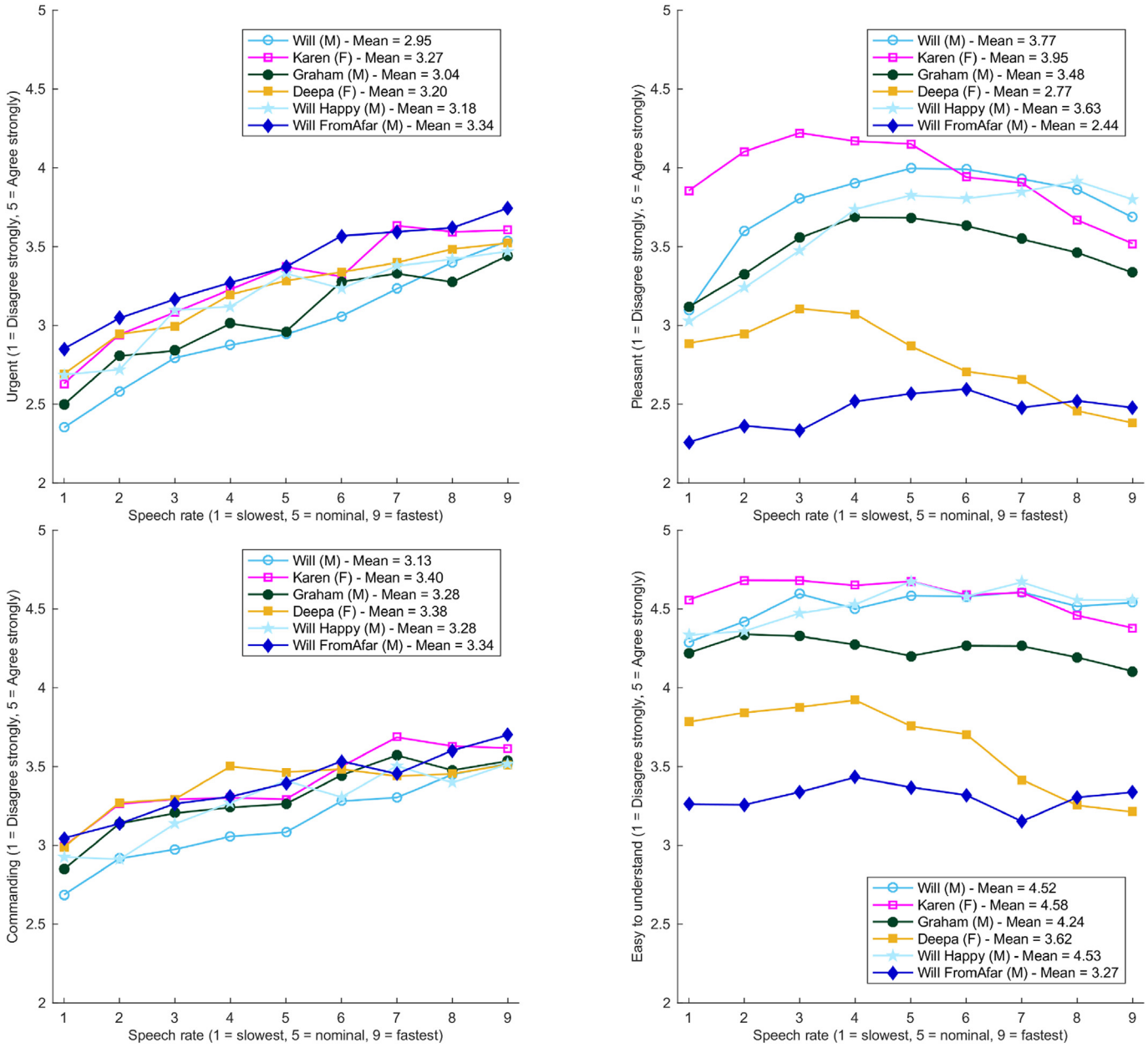

Fig. 1. Participants' ratings as a function of speech rate and speaker. Left top = mean urgency, Right top = mean pleasure, Left bottom = mean commandingness, Right bottom = mean ease-of-understanding. Each individual point in these four graphs represents the average across 2 of 140 sound recordings (i.e., noise trials and no-noise trials averaged), from an average of 366 participants (min = 320, max = 402; the corresponding 95% confidence intervals per point ranges between 0.12 and 0.31). The overall mean across the nine speech rates per speaker is indicated in the legend box. The corresponding sample size per speaker (all nine speech rates aggregated) in the four graphs is on average 2019 (min = 1992, max = 2046) and the 95% confidence interval per speaker ranges between 0.060 and 0.119.

the original findings, but rather suggest that there may be various unknown moderators at play. Possibly, specific features in the speaker's voice relative to the background noise (e.g., vehicle vs. aircraft noise) may have made the female voice stand out more (see also Cooke et al., 2013; Lerner et al., 2015).

One limitation of crowdsourcing is that the participant pool is limited in size, encompassing several thousands of people (Chandler et al., 2014; Stewart et al., 2015). Participants in our study were from 95 different countries, and previous research has found that there are national income-related differences in driving culture, traffic violations, and opinion about automated cars (De Winter and Dodou, 2016; Kyriakidis et al., 2015; Özkan and Lajunen, 2007). Furthermore, it has been found that people from non-English speaking countries take longer to complete CrowdFlower surveys than people from English speaking countries, which may signal difficulties with reading and interpreting the questions (De Winter et al., 2015). Considering the heterogeneity of our participant pool, it remains to be seen how well the observed effects apply to a specific target population of prospective users of automated cars. However, a similar limitation applies to lab-based research often conducted at universities with students as research participants.

Second, our study was not performed in a realistic driving context. The participants were not shown any automated driving scenarios, and we did not measure the response times of participants (see Arrabito, 2009; Ljungberg and Parmentier, 2012, in

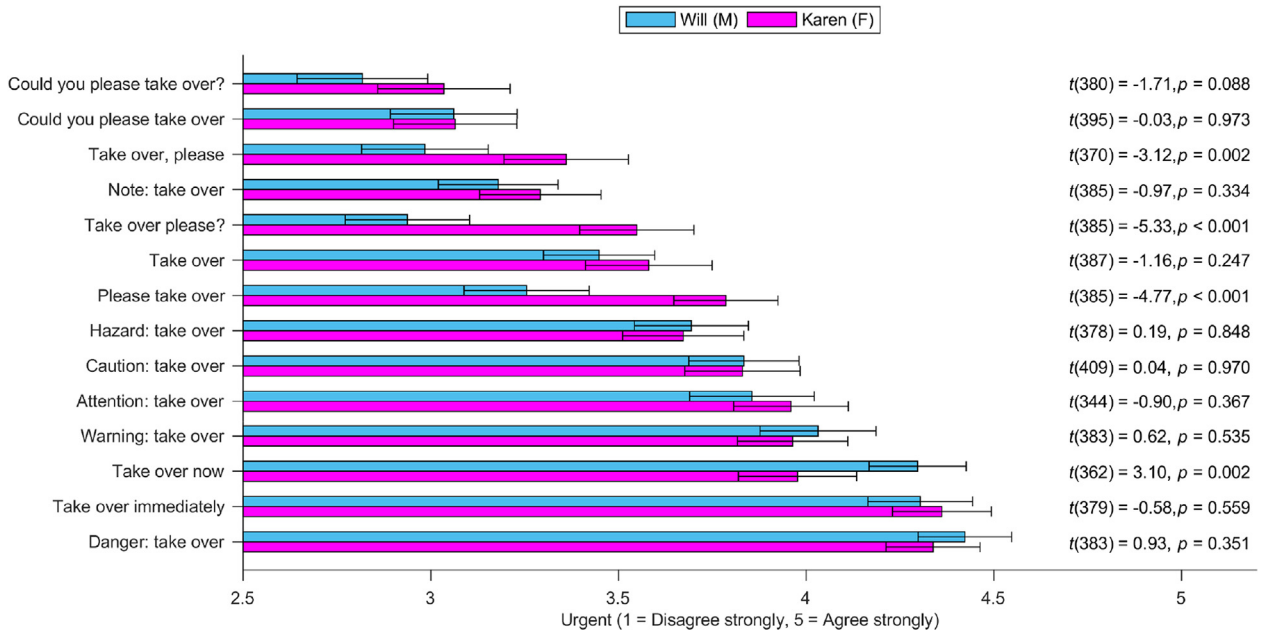

Fig. 2. Mean urgency scores for different phrases expressed by Will and Karen. The figure also depicts the results of a comparison between Will and Karen by means of independent-samples *t* tests. Each individual bar in the graph represents the average for 1 of 140 sound recordings. The error bars represent 95% confidence intervals. The 14 phrases are sorted on the mean urgency level.

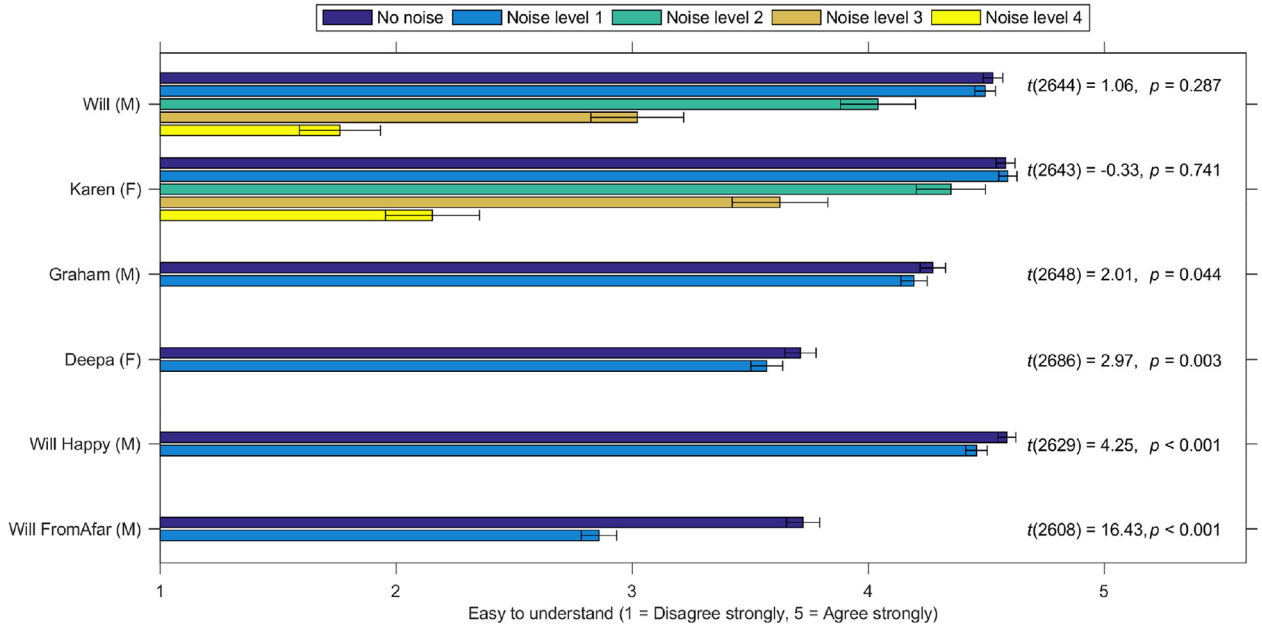

Fig. 3. Mean ease-of-understanding scores for the six speakers as a function of background noise level. The figure also depicts the results of a comparison between noise level 0 and noise level 1 by means of independent-samples *t* tests. Noise and volume levels are described in Table 1. The bars for No noise and Noise level 1 represent the average across 9 of 140 sound recordings (i.e., across the 9 levels of speech rate). The bars for Noise levels 2, 3, and 4 are the average for 1 of 140 recordings. The error bars represent 95% confidence intervals.

which participants' responses to speech were measured). It remains to be investigated how actual drivers would respond to speech-based take-over requests. It is possible that in demanding real-life traffic scenarios, a driver may be confused by the message "Take over please", especially if other warning sounds can be heard simultaneously. In addition, we learned through a discussion with fellow researchers that the phrase "Take over" (i.e., to reclaim manual control) may be suboptimal because it can be confused with "Overtake" (i.e., to pass a vehicle in front). Driving simulator studies in which drivers are exposed to different driving contexts are recommended in order to resolve these uncertainties.

Third, even though we used as many as 140 different auditory samples, our results may still be limited because the auditory samples reflect only a snapshot of the types of male and female voices and their emotional tones. To further explore whether the female voice is easier to understand than the male voice under

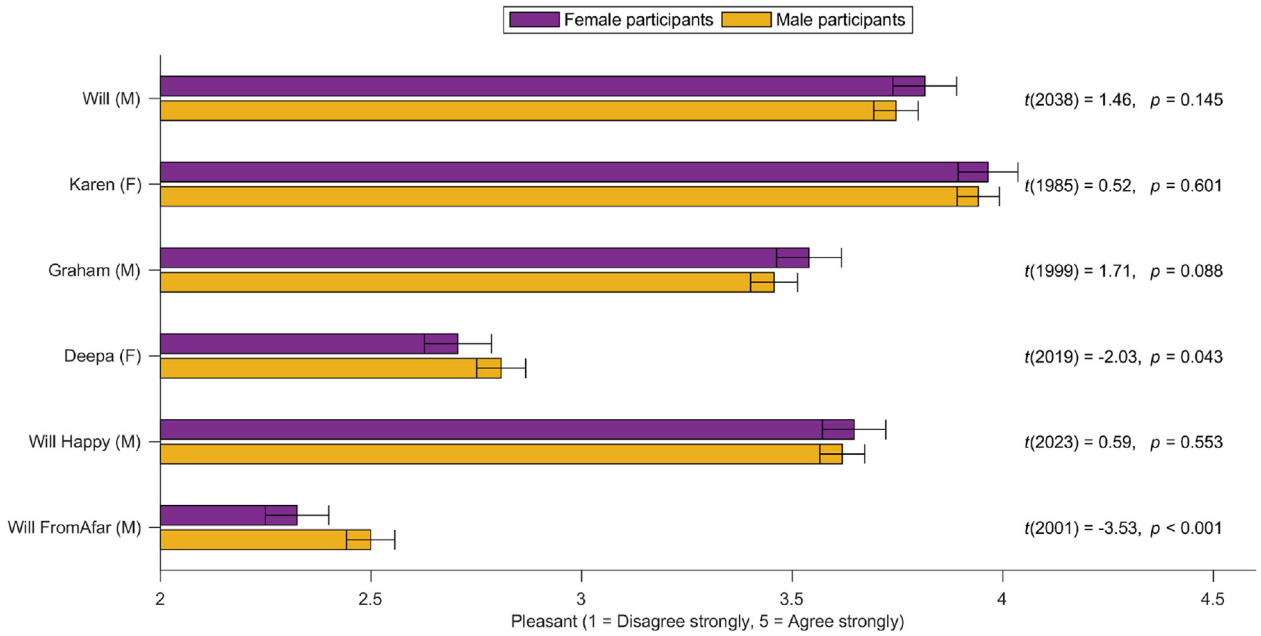

Fig. 4. Mean pleasure scores for the six speakers as a function of participant's gender. The figure also depicts the results of a comparison between female and male participants by means of independent-samples *t* tests. The bars represent the average across 18 of 140 sound recordings (i.e., across the 9 levels of speech rate, and for both noise levels). The error bars represent 95% confidence intervals.

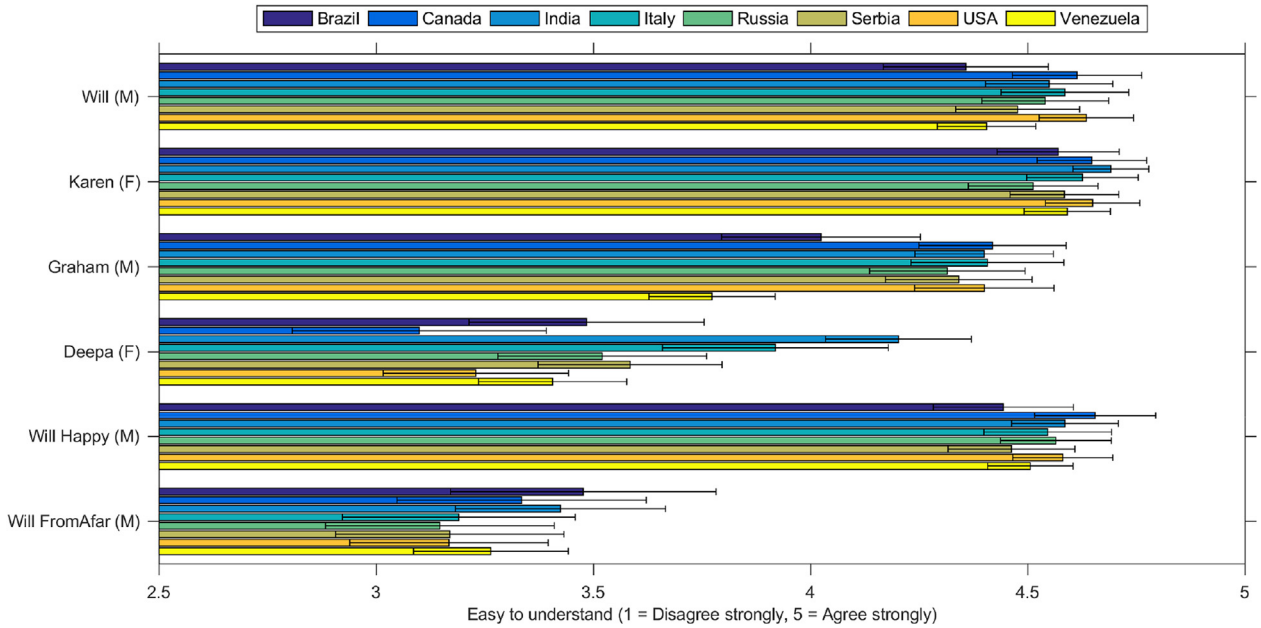

Fig. 5. Mean ease-of-understanding scores per participant's country and speaker. Only countries with 100 or more participants were shown. The bars represent the average across 18 of 140 sound recordings (i.e., across the 9 levels of speech rate, and for both noise levels). The sample size per bar is 70–79 for Brazil, 80–84 for Canada, 114–131 for India, 84–97 for Italy, 83–92 for Russia, 104–113 for Serbia, 131–142 for USA, and 232–254 for Venezuela. The error bars represent 95% confidence intervals.

background noise, multiple male and female voices and different types of noise spectra could be tested. Additionally, to better understand the interaction between listeners' gender and the speaker's accent, different accents could be included (other than the present US, UK, and Indian accents). The text-to-speech tool that we used offered a limited number of English accents and emotional tones. Recent developments in artificial intelligence give rise to increasingly flexible text-to-speech systems (e.g., Arik et al., 2017). The development of new software that offers a high range of choices for the customization of synthesized voice will be beneficial for future research on the valuation of speech. The number of auditory samples that can be tested depends on the researchers' financial resources and on the size of the participant pool on the crowdsourcing platform.

Because we used computerized speech, automotive researchers can readily reproduce the same speech warnings as used in this research. The results in Fig. 1 can be used to select a take-over request by considering each of the four dimensions. Our study

also has implications for human factors research in general. We showed that robust knowledge can be generated via the Internet, confirming earlier claims that crowdsourcing is a viable research tool (Crump et al., 2013). An important strength of this research is that it is effectively a between-subjects design, with participants listening to only 10 out of 140 take-over requests. In lab-based research, one usually has to resort to within-subjects designs to generate sufficient statistical power. Within-subject designs introduce carryover effects, which counterbalancing does not perfectly resolve (Greenwald, 1976; Keren, 1993). Crowdsourcing may be especially worthwhile when the experiment requires no special apparatus, as is often the case in usability research. Examples of research suited for crowdsourcing are perceptual tasks, cognitive tasks, and questionnaires (Buhrmester et al., 2011; De Winter et al., 2015; Rand, 2012).

Acknowledgements

The research presented in this paper is being conducted in the project HFAuto – Human Factors of Automated Driving (PITN-GA-2013-605817).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apergo.2017.05.001.

References

Arik, S.Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., et al., 2017. Deep voice: Real-time neural text-to-speech. Retrieved from. https://arxiv.org/pdf/1702.07825.pdf.

Arrabito, G.R., 2009. Effects of talker sex and voice style of verbal cockpit warnings on performance. Hum. Factors J. Hum. Factors Ergonomics Soc. 51, 3–20.

Arrowsmith, J., 2011. Trial watch: phase II failures: 2008–2010. Nat. Rev. Drug Discov. 10, 328–329.

Asendorpf, J.B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J.J., Fiedler, K., Perugini, M., 2013. Recommendations for increasing replicability in psychology. Eur. J. Personality 27, 108–119.

Baldwin, C.L., 2011. Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance. Ergonomics 54, 328–337.

Bazilinskyy, P., De Winter, J.C.F., 2015. Auditory interfaces in automated driving: an international survey. PeerJ Comput. Sci. 1, e13.

Begley, C.G., Ellis, L.M., 2012. Drug development: raise standards for preclinical cancer research. Nature 483, 531–533.

Behrend, T.S., Sharek, D.J., Meade, A.W., Wiebe, E.N., 2011. The viability of crowdsourcing for survey research. Behav. Res. Methods 43, 800–813.

Bent, T., Bradlow, A.R., 2003. The interlanguage speech intelligibility benefit. J. Acoust. Soc. Am. 114, 1600–1610.

Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? Perspect. Psychol. Sci. 6, 3–5.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376.

Chandler, J., Mueller, P., Paolacci, G., 2014. Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. Behav. Res. Methods 46, 112–130.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y., 2013. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Commun. 55, 572–585.

Crump, M.J., McDonnell, J.V., Gureckis, T.M., 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. Plos One 8, e57410.

Damböck, D., Weißgerber, T., Kienle, M., Bengler, K., 2013. Requirements for cooperative vehicle guidance. Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems, The Hague, the Netherlands, 1656–1661.

De Winter, J.C.F., Dodou, D., 2010. Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. Pract. Assess. Res. Eval. 15, 11.

De Winter, J.C.F., Dodou, D., 2016. National correlates of self-reported traffic violations across 41 countries. Personality Individ. Differ. 98, 145–152.

De Winter, J. C. F., Kyriakidis, M., Dodou, D., Happee, R., 2015. Using CrowdFlower to study the relationship between self-reported violations and traffic accidents. Proceedings of the 6th International Conference on Applied Human Factors and Ergonomics (AHFE), Las Vegas, NV, 2518–2525.

Edworthy, J., Hellier, E., Walters, K., Clift-Mathews, W., Crowther, M., 2003a. Acoustic, semantic and phonetic influences in spoken warning signal words. Appl. Cogn. Psychol. 17, 915–933.

Edworthy, J., Hellier, E., Rivers, J., 2003b. The use of male or female voices in warnings systems: a question of acoustics. Noise Health 6, 39–50.

Eichelberger, A.H., McCartt, A.T., 2014. Volvo drivers' experiences with advanced crash avoidance and related technologies. Traffic Inj. Prev. 15, 187–195.

Eriksson, A., Stanton, N.A., 2017. Takeover time in highly automated vehicles: noncritical transitions to and from manual control. Hum. Factors. http://dx.doi.org/10.1177/0018720816685832.

Fastl, H., 2006. Psychoacoustic basis of sound quality evaluation and sound engineering. Proceedings of the Thirteenth International Congress on Sound and Vibration, Vienna, Austria.

Freedman, L.P., Cockburn, I.M., Simcoe, T.S., 2015. The economics of reproducibility in preclinical research. PLOS Biol. 13, e1002165.

Gadbury, G.L., Allison, D.B., 2012. Inappropriate fiddling with statistical analyses to obtain a desirable p-value: tests to detect its presence in published literature. Plos One 7, e46363.

Gold, C., Damböck, D., Lorenz, L., Bengler, K., 2013. "Take over!" How long does it take to get the driver back into the loop? Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 57, 1938–1942.

Gold, C., Berisha, I., Bengler, K., 2015. Utilization of drivetime – Performing non-driving related tasks while driving highly automated. Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting, 59, 1666–1670.

Greenwald, A.G., 1976. Within-subject designs: to use or not to use? Psychol. Bull. 83, 314–320.

Gururaj, G., 2008. Road traffic deaths, injuries and disabilities in India: current scenario. Natl. Med. J. India 21, 14–20.

Hellier, E., Edworthy, J., Weedon, B., Walters, K., Adams, A., 2002. The perceived urgency of speech warnings: semantics versus acoustics. Hum. Factors J. Hum. Factors Ergonomics Soc. 44, 1–17.

Henrich, J., Heine, S.J., Norenzayan, A., 2010. The weirdest people in the world? Behav. Brain Sci. 33, 61–135.

Hergeth, S., Lorenz, L., Krems, J. F., Toenert, L., 2015. Effects of take-over requests and cultural background on automation trust in highly automated driving. Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Salt Lake City, UT, 331–337.

Hollander, T. D., Wogalter, M. S., 2000. Connoted hazard of voiced warning signal words: An examination of auditory components. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44, 702–705.

Ioannidis, J.P., 2005. Why most published research findings are false. PLOS Med. 2, e124.

Ioannidis, J., Doucouliagos, C., 2013. What's to know about the credibility of empirical economics? J. Econ. Surv. 27, 997–1004.

Jang, P.S., 2007. Designing acoustic and non-acoustic parameters of synthesized speech warnings to control perceived urgency. Int. J. Industrial Ergonomics 37, 213–223.

Keren, G., 1993. Between-or within-subjects design: a methodological dilemma. In: Keren, G., Lewis, C. (Eds.), A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues. Erlbaum, Hillsdale, NJ, pp. 257–272.

Klein, R.A., Ratliff, K.A., Vianello, M., Adams Jr., R.B., Bahník, Š., Bernstein, M.J., Cemalcilar, Z., 2014. Investigating variation in replicability: a "many labs" replication project. Soc. Psychol. 45, 142–152.

Kuchar, J.K., Yang, L.C., 2000. A review of conflict detection and resolution modeling methods. IEEE Trans. Intelligent Transp. Syst. 1, 179–189.

Kyriakidis, M., Happee, R., De Winter, J.C.F., 2015. Public opinion on automated driving: results of an international questionnaire among 5,000 respondents. Transp. Res. Part F Traffic Psychol. Behav. 32, 127–140.

Langlois, S., Suied, C., Lageat, T., Charbonneau, A., 2008. Cross cultural study of auditory warnings. Proceedings of the 14th International Conference on Auditory Display (ICAD2008), Paris, France.

Large, D.R., Burnett, G.E., 2013. Drivers' preferences and emotional responses to satellite navigation voices. Int. J. Veh. Noise Vib. 9, 28–46.

Lerner, N., Singer, J., Kellman, D., Traube, E., 2015. In-vehicle noise alters the perceived meaning of auditory signals. Proceedings of the Eighth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Salt Lake City, UT, 401–407.

Ljungberg, J.K., Parmentier, F.B., Hughes, R.W., Macken, W.J., Jones, D.M., 2012. Listen out! Behavioural and subjective responses to verbal warnings. Appl. Cogn. Psychol. 26, 451–461.

Ljungberg, J.K., Parmentier, F., 2012. The impact of intonation and valence on objective and subjective attention capture by auditory alarms. Hum. Factors J. Hum. Factors Ergonomics Soc. 54, 826–837.

Machado, S., Duarte, E., Teles, J., Reis, L., Rebelo, F., 2012. Selection of a voice for a speech signal for personalized warnings: the effect of speaker's gender and voice pitch. Work 41, 3592–3598.

Maxwell, S.E., Lau, M.Y., Howard, G.S., 2015. Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? Am. Psychol. 70, 487–498.

Merat, N., Jamson, A. H., 2009. How do drivers behave in a highly automated car. Proceedings of the 5th International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, Big Sky, MT, 514–521.

Mok, B., Johns, M., Lee, K. J., Miller, D., Sirkin, D., Ive, P., Ju, W., 2015. Emergency, automation off: Unstructured transition timing for distracted drivers of automated vehicles. Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Canary Islands, Spain, 2458–2464.

Munro, M.J., Derwing, T.M., 1995. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. Lang. Learn. 45, 73–97.
Munro, M.J., Derwing, T.M., Morton, S.L., 2006. The mutual intelligibility of L2 speech. Stud. Second Lang. Acquis. 28, 111–131.
Munro, M.J., 2008. Foreign accent and speech intelligibility. In: Hansen Edwards, J.G., Zampini, M.L. (Eds.), Phonology and Second Language Acquisition. John Benjamins, Philadelphia, pp. 193–218.
Naujoks, F., Mai, C., Neukum, A., 2014. The effect of urgency of take-over requests during highly automated driving under distraction conditions. In T. Ahram, W. Karwowski, & T. Marek (Eds.), Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics (AHFE), Kraków, Poland.
Nilsson, J., Strand, N., Falcone, P., Vinter, J., 2013. Driver performance in the presence of adaptive cruise control related failures: Implications for safety analysis and fault tolerance. Proceedings of the 43rd Annual IEEE/IFIP Conference on Dependable Systems and Networks Workshop, Budapest, Hungary.
Nixon, C.W., Morris, L.J., McCavitt, A.R., McKinley, R.L., Anderson, T.R., McDaniel, M.P., Yeager, D.G., 1998. Female voice communications in high levels of aircraft cockpit noises–part I: spectra, levels, and microphones. Aviat. Space, Environ. Med. 69, 675–683.
Noyes, J.M., Hellier, E., Edworthy, J., 2006. Speech warnings: a review. Theor. Issues Ergonomics Sci. 7, 551–571.
Oedegaarde, 2015. Tesla Autopilot Test (EU Version) - on Country Road, Highway and in Town. https://www.youtube.com/watch?v=7YVNW7-lFx8.
Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349, aac4716.
Özkan, T., Lajunen, T., 2007. The role of personality, culture, and economy in unintentional fatalities: an aggregated level analysis. Personality Individ. Differ. 43, 519–530.
Parasuraman, R., Riley, V., 1997. Humans and automation: use, misuse, disuse, abuse. Hum. Factors 39, 230–253.
Park, K.S., Jang, P.S., 1999. Effects of synthesized voice warning parameters on perceived urgency. Int. J. Occup. Saf. Ergonomics 5, 73–95.
Petermeijer, S., De Winter, J.C.F., Bengler, K., 2016. Vibrotactile displays: a survey with a view on highly automated driving. IEEE Trans. Intelligent Transp. Syst. 17, 897–907.
Pfromm, M., Khan, R., Oppelt, S., Abendroth, B., Brudera, R., 2015. Investigation of take-over performance of driving tasks by the driver due to system failure of semi-automated and assisted driving. Proceedings 19th Triennial Congress of the IEA, Melbourne, Australia.
Podlipský, V.J., Šimáčková, Š., Petráž, D., 2016. Is there an interlanguage speech credibility benefit? Top. Linguist. 17, 30–44.
Politis, I., Brewster, S., Pollick, F., 2015. Language-based multimodal displays for the handover of control in autonomous cars. Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Application, Nottingham, UK.
Rand, D.G., 2012. The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments. J. Theor. Biol. 299, 172–179.
Smith, L.E., Rafiqzad, K., 1979. English for cross-cultural communication: the question of intelligibility. TESOL Q. 13, 371–380.
SAE International, 2016. Taxonomy and Definitions for Terms Related To Driving Automation Systems for On-Road Motor Vehicles (Standard No. J3016).
Stewart, N., Ungemach, C., Harris, A.J., Bartels, D.M., Newell, B.R., Paolacci, G., Chandler, J., 2015. The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. Judgm. Decis. Mak. 10, 479–491.
Stroebe, W., 2016. Are most published social psychological findings false? J. Exp. Soc. Psychol. 66, 134–144.
Wagenmakers, E.J., Verhagen, J., Ly, A., Bakker, M., Lee, M.D., Matzke, D., Morey, R.D., 2015. A power fallacy. Behav. Res. Methods 47, 913–917.
Wogalter, M.S., Silver, N.C., 1995. Warning signal words: connoted strength and understandability by children, elders, and non-native English speakers. Ergonomics 38, 2188–2206.
Wogalter, M.S., Conzola, V.C., Smith-Jackson, T.L., 2002. Research-based guidelines for warning design and evaluation. Appl. Ergon. 33, 219–230.
World Health Organization, 2015. WHO Global Status Report on Road Safety 2015. World Health Organization, Geneva, Switzerland.
Zeeb, K., Buchner, A., Schrauf, M., 2015. What determines the take-over time? An integrated model approach of driver take-over after automated driving. Accid. Analysis Prev. 78, 212–221.

Pavlo Bazilinskyy is a Marie Curie Fellow in the BioMechanical Engineering Department at the Delft University of Technology. He received his double MSc degree (with distinction) from the University of St Andrews and Maynooth University in 2014.

Joost C. F. de Winter is associate professor in the BioMechanical Engineering Department at the Delft University of Technology, where he received his PhD degree (cum laude) in 2009.