Speech Tactons Improve Speech Warnings for Drivers

Ioannis Politis[1], Stephen Brewster[2]
Glasgow Interactive Systems Group
School of Computing Science
University of Glasgow
Glasgow, G12 8QQ, UK
[1]I.Politis.1@research.gla.ac.uk
[2]Stephen.Brewster@glasgow.ac.uk

Frank Pollick
School of Psychology
University of Glasgow
Glasgow, G12 8QB, UK
Frank.Pollick@glasgow.ac.uk

ABSTRACT
This paper describes two experiments evaluating a set of speech and tactile driver warnings. Six speech messages of three urgency levels were designed, along with their tactile equivalents, Speech Tactons. These new tactile warnings retained the rhythm of speech and used different levels of roughness and intensity to convey urgency. The perceived urgency, annoyance and alerting effectiveness of these warnings were evaluated. Results showed that bimodal (audio and tactile) warnings were rated as more urgent, more annoying and more effective compared to unimodal ones (audio or tactile). Perceived urgency and alerting effectiveness decreased along with the designed urgency, while perceived annoyance was lowest for warnings of medium designed urgency. In the tactile modality, ratings varied less as compared to the audio and audiotactile modalities. Roughness decreased and intensity increased ratings for Speech Tactons in all the measures used. Finally, Speech Tactons produced acceptable recognition accuracy when tested without their speech counterparts. These results demonstrate the utility of Speech Tactons as a new form of tactile alert while driving, especially when synchronized with speech.

Author Keywords
Multimodal interaction; warnings; audio; tactile; speech; Tactons; urgency; annoyance; alerting effectiveness; recognition accuracy.

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces. - Haptic I/O ; Voice I/O.

INTRODUCTION
As technology progresses, vehicles become increasingly able to provide multimodal warnings to drivers. There have been numerous studies investigating the efficacy of auditory and tactile warnings in the car, signifying events of varying urgency, e.g. [10,19,23]. Auditory warnings used can either be abstract signals [19], sounds associated with the events [16] or speech [11]. However the potential of tactile messages to convey some aspects of speech has been much less investigated in general, and never before in the context of driving. This is interesting to investigate, since transferring features of speech to vibration has provided good results in the past [22,24]. This paper presents a first investigation of how tactile messages based on speech, called Speech Tactons, perform as warnings for drivers. A set of audio and tactile messages related to different driving events was designed and evaluated in terms of subjective responses and recognition accuracy. The goal was to investigate the potential of using them as warnings and how Speech Tactons might aid responses to speech warnings.

AutomotiveUI '14, September 17 - 19 2014, Seattle, WA, USA

http://dx.doi.org/10.1145/2667317.2667318

Speech Warnings
McKeown & Isherwood [17] evaluated a set of auditory warnings with different semantic associations to driving events. Four suites of warning sounds were designed, with nine sounds in each suite. Abstract sounds (e.g. tones or a siren), environmental sounds with no relation to driving (e.g. a baby sneezing or footsteps), environmental sounds related to driving events (e.g. car speeding past) and speech messages (e.g. *"Exceeding speed limit"* or *"Petrol is low"*) were used. It was found that abstract sounds had the highest response times and the lowest identification accuracy. Speech and auditory icons had the lowest response times and the highest accuracies. Speech was perceived as more pleasant and less urgent compared to abstract sounds. This is evidence of the potential of speech when alerting drivers of various events, which also motivated our study.

Ho & Spence [11] investigated the use of car horn sounds and speech cues as warning signals. Participants responded quicker to a critical event when the cues were coming from the direction of the event (front or back) and when their attention was directed to the correct direction through a speech cue (*"front"* or *"back"*). Serrano *et al.* [23] also presented a set of speech messages to drivers (*"Look out on the left / ... on the right /... on the road!"*). They were followed by pictures of either hazardous or non-hazardous road scenes, asking drivers to identify whether there was a hazard. Reaction times to this task were shorter and responses more accurate when the speech messages were presented from the direction of the hazard as opposed to a random direction. Messages presented from the correct direction created even shorter reaction times, when that direction was uttered in the message as opposed to not specified. The above studies demonstrate the effectiveness of short speech warnings when delivered from the appropriate direction relative to the threat. However, the subjective responses of the warnings as well as the recognition accuracy were not assessed. This is essential for evaluating perception of the cues, as well as when drivers need to interpret the meaning of messages and act appropriately.

Cao *et al.* [6,7] investigated the use of speech, abstract audio cues and visuals for presenting road obstacle warnings in a simulated driving task. Speech messages along with pictures led to good recall of the signified events but low reaction times. Thus, the authors suggested the use of speech along with pictures of signified events for tasks not requiring quick responses, such as navigation. Speech along with pictures was also perceived as most useful in various driving contexts, i.e. low visibility, under fatigue and high demand. The speech cues used in these studies were relatively long, e.g. *"Broken vehicle in 180 meters on the right roadside"*, resulting to longer utterances. In our study, we used shorter speech warnings to investigate whether they will still lead to satisfactory ratings when combined with tactile instead of visual cues. Previous studies, e.g. [18], have provided evidence of an overload in the visual modality when driving, therefore richer non-visual ways to warn drivers may be useful. Further, the positive effect of audio combined with vibration observed in [10] was also a motivation, so as to investigate if the benefits will hold for recognition accuracy and subjective responses.

Designing Urgency in Speech

There have been studies on how speech is perceived in warnings in terms of its urgency, annoyance and alerting effectiveness. Baldwin & Moore [2] investigated the signal words *"Danger"*, *"Warning"*, *"Caution"* and *"Notice"* when used together with different collision avoidance related messages. They found that the signal word *"Danger"* was perceived as more urgent compared to the words *"Warning"* and *"Caution"*, which in turn were perceived as more urgent compared to *"Notice"*. It was also found that a higher S/N ratio when presenting warnings positively impacted ratings of urgency and alerting effectiveness, without strong impact in annoyance. Higher S/N ratio also positively affected the ratings of urgency, regardless of the semantic content of the collision avoidance messages. Similar effects were observed later by Baldwin [3] in terms of reaction times, where participants responded quicker to urgent warnings, created by using urgent words and high signal intensity. In our study, we use a subset of these signal words at the beginning of our speech messages and investigate whether these effects hold for the resulting cues.

Hellier *et al.* [9] also found an influence of acoustics and speaking style in the ratings of urgency. Signal words spoken urgently created higher ratings compared to non-urgently, which in turn were higher compared to words spoken in a monotone manner. Female speakers induced higher urgency ratings and a higher range in these ratings compared to males. Finally, the word *"Danger"* was perceived as highly urgent, matched only by the word *"Deadly"*. Additionally, it was found that an urgent utterance of the messages resulted in louder sounds, with higher pitch and pitch range. Edworthy *et al.* [8] extended these findings, observing that signal words spoken urgently are perceived as more urgent, believable and appropriate as well. In our study, we use the above guidelines in the warnings, matching the urgency of the utterances to the urgency of the signified events.

Tactile Messages

Investigating the design of tactile warnings, Pratt *et al.* [20] and Baldwin & Lewis [1] suggested increasing interpulse interval as a means to increase perceived urgency. However, the resulting messages are repeated vibrations with relatively low semantic content. Such messages have also produced higher annoyance ratings and lower recognition rates compared to messages in audio or visual modalities in earlier studies, e.g. [19]. Brown, Brewster & Purchase [4,5] investigated the use of Tactons as a means to convey more complex information with the tactile modality. Varying parameters such as rhythm and roughness (amplitude modulation on the original waveform that provides the vibration) of such messages enabled the design of richer cues without cost in their recognition accuracy, as long as a reduced number of different levels for roughness was used. Hoggan and Brewster [12] extended this work and evaluated parameters of Tactons when used in conjunction with audio messages. They suggested rhythm and roughness of vibration as a means to convey information effectively when used along with audio. Further, the use of intensity was regarded as requiring further investigation in the cues. In our study, we use tactile cues along with audio, varying the rhythm of the cues, by imitating the rhythm of speech. We also vary roughness in only two levels (no roughness present and roughness present), so as to enrich the information conveyed without cost in recognition accuracy. We finally vary intensity in the cues, to address the open questions on how effectively it can be used in Tactons.

The potential for transferring some speech features into vibration has been investigated. Spens *et al.* [24] developed a handheld tactile aid to lip reading, which vibrated synchronously to speech and was designed for people with hearing impairments. Li [15] investigated the use of synthesized Tactons, consisting of repeated pulses of varying duration and intensity. They were mapped through forced choice responses to simple speech messages that frequently occur when texting (e.g. *"hello?"*, *"goodbye."*, *"where are you?"*). The number of syllables, intonation and stress of the spoken messages were identified as important features to be mapped through vibration. The number of syllables was mapped to number of pulses, while the intonation and stress were mapped to vibration intensity. Salminen *et al.* [22] investigated the use of audiotactile messages, where the vibration mimicked the amplitude changes of speech. The audiotactile messages were presented through a handheld device of form factor similar to [24]. Participants rated the audiotactile messages as more arousing and dominant compared to the audio ones. Tuuri, Eerola & Pirhonen [25] used intonation and rhythm of speech messages to create pure tones that were then delivered either through audio or through vibration. The messages *"Slow"*, *"Urge"* and *"Ok"* showed high recognition rates in both modalities. The authors conclude that the two modalities can be used interchangeably for interface design. In our study, we developed a technique similar to [22] and [25] to transform speech messages into vibration and evaluated the responses to the resulting cues when used alone or together with speech. This has not been attempted in driving and can provide insights on designing novel driver displays.

To summarise, speech is a promising means of conveying information to drivers. When used together with vibration it could have an improved effect on responses. The mapping of some speech features to vibration has been shown to provide good recognition rates and subjective ratings but has not been attempted in the driving context. Therefore, we designed a set of speech messages for drivers and created a technique to construct their tactile equivalents. We also evaluated responses to the cues to provide insights on their applicability for driving.

WARNING DESIGN

Six speech messages relating to various driving events were recorded, designed to convey three different urgency levels, Level High (L_H), Level Medium (L_M) and Level Low (L_L). Using three different levels of urgency has shown good results in previous studies [18,19]. The messages used were chosen from [13], where a set of in-vehicle messages were prioritized according to the SAE J2395 standard [21]. Messages of highest priority in [13] were mapped to L_H in our study, messages of intermediate priority to L_M and messages of lowest priority to L_L. We also added the word *"Danger!"* before each L_H message, *"Warning!"* before each L_M and *"Notice!"* before each L_L, since these words have shown to provide distinctively different urgency ratings in previous studies [2,9]. The resulting messages were *"Danger! Collision Imminent"* (D_1) and *"Danger! Tire pressure falling"* (D_2) for L_H, *"Warning! Activate fog lamps"* (W_1) and *"Warning! Left side headlamp out"* (W_2) for L_M and *"Notice! Rest area 17 miles"* (N_1) and *"Notice! Call and win free tickets"* (N_2) for L_L. W_1, N_1 and N_2 were slightly adjusted from their original text in [13] so that no messages would resemble each other in terms of rhythm and number of syllables. All messages were recorded by a female voice actor using a Rode NT2-A[1] condenser microphone. Female speakers have been found to produce messages with higher variation in ratings of urgency [9]. In line with [9], the actor was instructed to speak messages of L_H in an urgent manner, as if a loved one was in imminent danger. Accordingly, L_M messages were spoken non-urgently, as if in a friendly conversation with nothing interesting about the situation and L_L messages were spoken in a monotone, deadpan manner. Both L_H messages were 1.7 *sec* long and had a peak of -1.9 *dBFS* and an average frequency of 377 *Hz* (D_1) and 372 *Hz* (D_2). L_M were 2.6 *sec* (W_1) and 2.7 *sec* (W_2) long, had a peak of -9.5 *dBFS* (W_1) and -11.1 *dBFS* (W_2) and an average frequency of 310 *Hz* (W_1) and 285 *Hz* (W_2). Finally, L_L messages were 3.4 *sec* (N_1) and 3.7 *sec* (N_2) long, had a peak of -15.2 *dBFS* (N_1) and -16.5 *dBFS* (N_2) and an average frequency of 198 *Hz* (N_1) and 202 *Hz* (N_2). We stress that the above values were acquired by only presenting the actor with the verbal instructions described and using no other intervention[2].

For the Speech Tactons, all stimuli used were auditory, designed for a C2 tactor[3]. Initially, the fundamental frequency F_0 (pitch) of each sample of the speech recordings was obtained, resulting in alternating pure tones for each utterance. Then, the changes in intensity of the original sound files were used in the tones. This resulted in tactile design *P* (Pitch). In order to investigate the effect of roughness in the resulting cues, an amplitude modulation of 30 *Hz* was added over the P messages, as in [5]. This resulted in design *PR* (Pitch-Roughness). Designs P and PR maintained the intensity levels of the original audio recording, i.e. they had the same peak levels as the respective audio cues. Finally, to investigate the use of intensity in the cues, two more tactile designs were created, where the maximum possible intensity was used in the cues, while still avoiding clipping. This modification to design P provided design *PI* (Pitch and maximum Intensity) and the same modification to PR provided *PRI* (Pitch-Roughness and maximum Intensity). Designs PI and PRI had peak levels of 0.0 *dBFS.* All tactile cues retained the rhythm and intensity variations of the original utterances. Further, the resulting values of average frequency of all tactile cues never differed to the average frequency of the audio more than ±10*Hz*. Overall, 54 different cues were created, 6 Audio (A), 24 Tactile (T), i.e. 6 cues × 4 designs and 24 audio and tactile (AT), i.e. A cues together with the equivalent T ones. As an example, see Figure 1.a for the waveforms of N_2. For all modifications, Praat[4] and Audacity[5] software were used.

---

[1] http://www.rodemic.com/microphones/nt2-a

[2] All the resulting messages can be found in http://soundcloud.com/idpolitis

[3] http://www.atactech.com/PR_tactors.html

[4] http://www.fon.hum.uva.nl/praat/

[5] http://audacity.sourceforge.net/

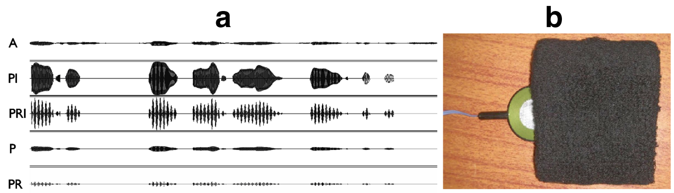

Figure 1.(a) The waveforms of message N_2: "Notice! Call and win free tickets." (a) and the equivalent tactile messages. (b) The wristband and the C2 tactor used in the experiments.

EXPERIMENT 1

Two experiments were designed, evaluating the cues described above. The first experiment investigated the subjective responses provided by participants when exposed to the warnings. A 6×3×4 within subjects design was used with Message, Modality and Design as the independent variables and Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) as the dependent ones. Message had 6 levels (D_1, D_2, W_1, W_2, N_1, N_2), Modality had 3 levels (A, T, AT) and Design had 4 levels (P, PR, PI, PRI).

Procedure

Twenty two participants (9 female) aged between 18 and 44 years (M = 25.04, SD = 5.95) took part in this experiment. They all held a valid driving licence and had between 1 and 27 years of driving experience (M = 5.79, SD = 5.85). Participants were all right handed and reported normal hearing. They were either University students or employees. The experiment took place in a University room, where participants sat in front of 27-inch Dell 2709W monitor and a PC running the experimental software. They wore a set of Sennheiser HD 25-1 headphones and a wristband on their left hand with a C2 tactor attached on the inside of the band (see Figure 1.b), in line with [20,25]. Participants provided all responses using a mouse with their right hand and were asked to rest their left hand on the desk. To cover the tactor noise, car sound was played throughout the experiment.

After being welcomed and explained the experimental procedure, participants were exposed to the 54 cues (6 A, 24 T, 24 AT) in a random order, to familiarize them with the signals. For each cue, they had the option to repeat it or to proceed to the next one when they felt familiar with it. Afterwards, they were again presented with the cues and asked to rate them all in terms of PA, PU and PAE, by completing a 5-point Likert scale for each rating, in line with [3]. In all ratings the scale was: Not at all (1), Slightly (2), Moderately (3), Very (4) and Extremely (5). Participants were asked to imagine they were driving and wearing a wrist mounted device like a smart watch for vibration, while also listening to their car speakers for sound. The wrist was selected, since previous studies [20,25] have shown good recognition of vibration on this area, while using the abdomen has resulted to higher ratings of perceived annoyance [19]. Each cue was presented twice, resulting to 108 trials. The experiment lasted about 30 minutes and participants were then prepared for Experiment 2 in the same session.

Results

Perceived Urgency (PU)

Data for PU were analysed using a two-way repeated measures ANOVA, with Modality and Message as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Modality and Modality × Message, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates. There was a significant main effect of Modality ($F(1.4,60.26) = 6.26$, $p < 0.05$). Contrasts revealed that modality AT created higher ratings of PU compared to A and T ($F(1,43) = 15.34$, $r = 0.51$, $p < 0.001$). There was also a significant effect of Message ($F(5,215) = 223.21$, $p < 0.001$). Contrasts revealed that D_1 was perceived as more urgent that D_2 ($F(1,43) = 7.36$, $r = 0.38$, $p < 0.05$), D_2 more urgent than W_1 ($F(1,43) = 124.39$, $r = 0.86$, $p < 0.001$), W_2 more urgent than N_1 ($F(1,43) = 112.37$, $r = 0.85$, $p < 0.001$) and N_1 more urgent than N_2 ($F(1,43) = 9.67$, $r = 0.43$, $p < 0.05$). There was a significant interaction between Modality and Message ($F(6.34,272.84) = 68.25$, $p < 0.001$). Contrasts revealed that the significant differences in ratings of PU described above were not present in modality T.

Data for Modalities T and AT, where there was a Design present, were analysed in terms of their PU using a three-way repeated measures ANOVA, with Modality, Message and Design as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Design, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates. Effects of Modality, Message and their interaction were similar to above and are omitted. There was a significant main effect of Design ($F(2.38,102.36) = 17.98$, $p < 0.001$). Contrasts revealed that design PI created higher ratings of PU compared to P ($F(1,43) = 7.27$, $r = 0.38$, $p < 0.05$), P higher ratings compared to PRI ($F(1,43) = 4.28$, $r = 0.30$, $p < 0.05$) and PRI higher ratings compared to PR ($F(1,43) = 10.08$, $r = 0.44$, $p < 0.05$). See Figure 2 for ratings of PU for modalities, messages and designs.

Perceived Annoyance (PA)

Data for PA were analysed using a two-way repeated measures ANOVA, with Modality and Message as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Message and Modality × Message, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates. There was a significant main effect of Modality ($F(2,86) = 8.42$, $p < 0.001$). Contrasts revealed that modality AT created higher ratings of PA compared to A and T ($F(1,43) = 13.53$, $r = 0.49$, $p < 0.001$). There was also a significant effect of Message ($F(2.46,105.94) = 15.03$, $p < 0.001$). Contrasts revealed that D_2 and N_2 had higher PA than D_1 ($F(1,43) = 17.52$, $r = 0.54$, $p < 0.001$), which in turn had higher PA than N_1, W_1 and W_2 ($F(1,43) = 5.11$, $r = 0.32$, $p < 0.05$). There was a significant interaction between Modality and Message ($F(5.84,251.30) = 9.52$, $p < 0.001$), indicating that the above differences in PA were not present in modality T.

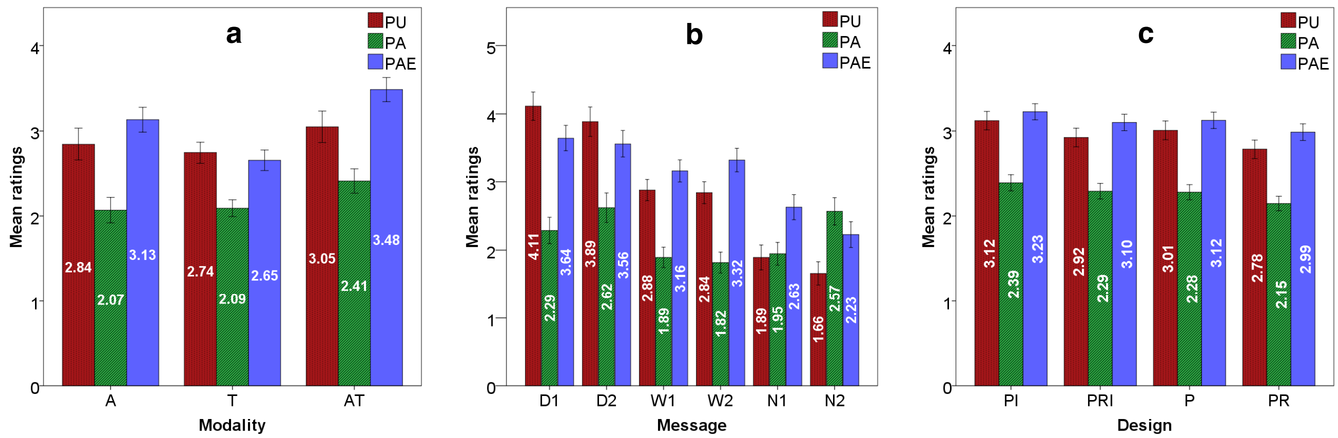

Figure 2.(a) Mean ratings of Perceived Urgency (PU), Perceived Annoyance (PA) and Perceived Alerting Effectiveness (PAE) across modalities (a), messages (b) and designs (c) for Experiment 1. Error bars indicate 95% confidence intervals.

Data for Modalities T and AT, where there was a Design present, were analysed for PA using a three-way repeated measures ANOVA, with Modality, Message and Design as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Design and Modality × Message × Design, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates. Effects of Modality, Message and their interaction were similar to above and are omitted. There was a significant main effect of Design ($F(2.46,105.95) = 13.31$, $p < 0.001$). Contrasts revealed that design PI created higher ratings of PA compared to PRI and P ($F(1,43) = 9.49$, $r = 0.42$, $p < 0.05$) and the latter created higher ratings of PA compared to PR ($F(1,43) = 13.50$, $r = 0.49$, $p < 0.001$). See Figure 2 for mean ratings of PA for modalities, messages and designs.

Perceived Alerting Effectiveness (PAE)

Data for PAE were analysed using a two-way repeated measures ANOVA, with Modality and Message as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Modality, Message and Modality × Message, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates. There was a significant main effect of Modality ($F(1.63,70.28) = 28.48$, $p < 0.001$). Contrasts revealed that modality AT created higher ratings of PAE compared to A ($F(1,43) = 13.99$, $r = 0.49$, $p = 0.001$) and A higher ratings compered to T ($F(1,43) = 18.44$, $r = 0.55$, $p < 0.001$). There was also a significant effect of Message ($F(3.48,149.83) = 55.23$, $p < 0.001$). Contrasts revealed that D_1 and D_2 were rated higher in PAE compared to W_2 ($F(1,43) = 6.23$, $r = 0.35$, $p < 0.05$), W_1 higher compared to N_1 ($F(1,43) = 32.66$, $r = 0.66$, $p < 0.001$) and N_1 higher compared to N_2 ($F(1,43) = 19.73$, $r = 0.56$, $p < 0.001$). There was a significant interaction between Modality and Message ($F(6.68,287.05) = 25.20$, $p < 0.001$), indicating that the differences in ratings of PAE described above were not present in modality T.

Data for Modalities T and AT, where there was a Design present, were analysed for PAE using a three-way repeated measures ANOVA, with Modality, Message and Design as factors. Mauchly's test revealed that the assumption of sphericity had been violated for Message × Design and Modality × Message × Design, therefore degrees of freedom were corrected using Greenhouse–Geisser estimates. Effects of Modality, Message and their interaction were similar to above and are omitted. There was a significant main effect of Design ($F(3,129) = 12.90$, $p < 0.001$). Contrasts revealed that design PI created higher ratings of PAE compared to P and PRI ($F(1,43) = 6.98$, $r = 0.37$, $p < 0.05$) and the latter higher ratings compared to PR ($F(1,43) = 9.11$, $r = 0.42$, $p < 0.001$). See Figure 2 for mean ratings of PAE across modalities, messages and designs.

From Experiment 1 it was clear that AT was rated higher in all measures compared to A and T, showing clear evidence of the usefulness of the designed cues when modalities were combined. Further, it was evident that the urgency designed in the warnings was reflected in their PU. PAE escalated according to PU, indicating that messages signifying situations of higher importance were regarded as more useful. PA was higher for messages of both L_H and L_L and lower for L_M, allowing for several interpretations which will be discussed later. Finally, all ratings were not as responsive to messages in the T modality, a further indication of the higher utility of the cues when presented multimodally. In order to investigate the ability of participants to recognise the tactile cues without the audio present, Experiment 2 was performed immediately after Experiment 1, investigating the recognition accuracy of the messages.

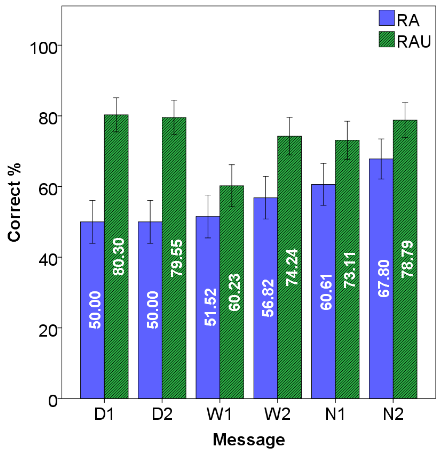

Figure 3: RA and RAU across messages for Experiment 2.

EXPERIMENT 2

Experiment 2 investigated the recognition accuracy of participants when exposed to the T warnings. To investigate participants' performance in identifying individual cues, as well as recognizing their urgency, two measures were used. Recognition Accuracy (RA) was 1 when participants recognized the exact message, e.g. responded "N_2" when the message was indeed N_2, and 0 in all other cases. Recognition Accuracy of Urgency (RAU) was 1 when participants recognized the urgency of the message, e.g. responded "D_1" or "D_2" when the message was D_2, and 0 in all other cases. A 6×4 within subjects design was used with Message and Design as the independent variables and RA and RAU as the dependent ones.

Procedure

Participants and equipment were identical to Experiment 1. After completing Experiment 1 participants were again exposed to the 54 cues (6 A, 24 T, 24 AT) in a random order, to further familiarize themselves with the signals and the mapping between T and AT. Then, they were presented with only the 24 T warnings, repeated three times each in a random order, resulting in 72 trials. Participants were asked to map each warning to one of the A messages (D_1, D_2, W_1, W_2, N_1 or N_2). They were able to do this by selecting one option out of six available, each displaying the full text of the speech warning. They were also free to feel the T cues as many times as needed before responding. The experiment lasted about 30 minutes and finally participants were debriefed and paid £6 for participating to both experiments.

Results

Data for RA and RAU were treated as dichotomous and analysed with Cochran's Q tests. It was found that N_1 had higher RA than D_1 ($Q(1) = 6.03$, $p < 0.05$), D_2 ($Q(1) = 9.56$, $p < 0.05$) and W_1 ($Q(1) = 5.65$, $p < 0.05$). Further, N_2 had higher RA than D_1 ($Q(1) = 16.12$, $p < 0.001$), D_2 ($Q(1) = 23.25$, $p < 0.001$), W_1 ($Q(1) = 18.31$, $p < 0.001$), W_2 ($Q(1) = 10.38$, $p < 0.05$) and N_1 ($Q(1) = 5.23$, $p < 0.05$). In terms of RAU, it was found that W_1 had lower values compared to D_1 ($Q(1) = 25.31$, $p < 0.001$), D_2 ($Q(1) = 24.31$, $p < 0.001$), W_2 ($Q(1) = 12.79$, $p < 0.001$), N_1 ($Q(1) = 11.11$, $p < 0.001$) and N_2 ($Q(1) = 24.25$, $p < 0.001$). Also, N_1 had lower RAU compared to D_1 ($Q(1) = 5.08$, $p < 0.05$) and D_2 ($Q(1) = 4.07$, $p < 0.05$). Additionally, design P had higher RAU compared to PR (78% vs. 71%, $Q(1) = 7.86$, $p < 0.05$). Designs PI (73%) and PRI (75%) did not show significant differences in RAU compared to any other designs. Finally, no significant differences were found in RA between designs. See Figure 3 for mean ratings of RA and RAU across messages.

DISCUSSION

The results of PU are a clear indication that the messages designed conveyed the desired urgency. Participants rated the messages as expected, highlighting that available guidelines in [2,3,9] are also valid for multimodal messages. It is interesting how a set of simple guidelines to a voice actor succeeded in producing messages of distinct differences in average frequency and peak. This is an indication of the applicability of guidelines presented in [9]. Even messages of the same level of designed urgency presented different ratings, which still were not similar to messages of different levels. This provides potential for selecting different cues for one level and enriching the interaction. AT messages had higher ratings of PU, which is an improvement compared to A or T messages, when a situation of high criticality needs to be conveyed. Similar improvements of speech warnings when combined with visuals have been observed in studies like [6,7]. In our study, we show that also tactile cues can improve responses to speech warnings. This adds to the existing body of work, suggesting enhanced responses to multimodal signals versus unimodal ones, e.g. [14,19]. Further, T messages did not present highly different ratings of PU, adding to the argument that such cues work better when used multimodally.

In terms of tactile designs, it was clear that intensity was the main factor that led to higher PU (PI was rated higher than P and PRI higher than PR), while roughness led to lower PU (PRI and PR were rated lower than PI and P). This strengthens the evidence that intensity of the tactile part is useful to create more urgent messages and can be compared with [2,3,9], where high intensity of audio affected PU ratings. Roughness seems to produce the opposite effect, unlike some prior studies such as [5]. However, it needs to be noted that in [5] roughness was not used to design urgency per se, but to signify more or less important scheduling events. Also, the tactile cues were not speech based. The above results are promising when designing AT cues based on Speech Tactons and provide a variety of ways to do this.

Data for PA, while presenting interesting variations, had low values overall, since all average values across all factors observed were below 3 (moderately annoying). This is an improvement compared to previous studies using such cues, where annoyance was higher for T, e.g. [19]. Consistent with prior work, e.g. [18,19], multimodal messages were rated higher in PA. The T modality again created less variation in ratings of annoyance across messages, adding to our suggestion that T cues can be better used to enhance the responses rather than used standalone. Interestingly, both cues of high and low designed urgency were rated higher in PA (D_2 and N_2). This can be explained by several participants' comments indicating firstly that they would not like warnings for non-important events and secondly that although L_H cues were more annoying, this was desired by them since it would increase their alertness. Looking at the results of PAE, which will be discussed later, this interpretation is further supported. We suggest that on one hand more modalities create more annoyance but this is not necessarily a flaw when the event signified is critical. Since intermediate designed urgency is less annoying to participants, unimodal signals seem to be a good option to choose in this case. In terms of tactile design, cues using intensity were again rated higher (PI was rated highest). This adds to the suggestion that intensity can be used for urgent events, where higher PA can be tolerated. Roughness did not lead to as high ratings (PR was rated lowest), making this feature a better candidate for low urgency cues, also considering the similarly low PU ratings for roughness.

Results for PAE are encouraging, since they are similar to PU, with often higher average values. As generally observed in this study, more modalities increase ratings in all measures and PAE is no exception. Also in [22], participants rated the audiotactile messages as more arousing and dominant compared to the audio ones. Our results show that Speech Tactons are also rated as more effective when combined with audio. With the technique we suggest for designing these Tactons, we hope to provide a simple and easily implementable way to derive such messages from speech. As mentioned earlier, various techniques have been suggested in the past to map speech to vibration and with mixed results. Our study is a comprehensive examination of how speech and tactile cues are perceived by drivers and shows positive results not previously observed. These results also further support the claim that participants valued more being warned about important events and can further justify why higher ratings of annoyance for L_H messages were acceptable according to participants' comments, while for L_L they were not. This also relates to [8], where highly urgent messages were perceived as more appropriate. As a guideline, warnings of high criticality can be more alerting even at the cost of more annoyance. In terms of tactile design, intensity was again preferred to roughness (PI rated highest and PR lowest), addressing open questions of [12] on utilizing intensity, and further suggesting that it is a good feature to design effective cues.

Finally, recognition accuracy produced acceptable values overall, but especially good values of RAU. For values of RA, we note that a random response, indicating that participants were just guessing the messages, would provide RA percentages of $100\% \div 6 = 16.7\%$. In our study, the lowest RA observed was 50%, well above that value. Looking at RAU, the results are even more encouraging. Interestingly, high urgency messages performing poorer in RA performed best in RAU. This is also backed up by participants' comments, mentioning that it was easy for them to recognise which level a message belonged to, but not as easy to tell which message it was, especially when messages were short. For longer messages, RA performance was also high, since participants had more time to distinguish the different properties of the T cues. This is also an indication that fewer individual T cues could be better recognised compared to more, since the cues would be more different. It is also in line with [5], where the number of Tactons needed to be reduced to achieve better recognition results. In our study, we had not intended to suggest Speech Tactons to be presented on their own, but only along with audio. This is especially true for L_H cues, where ambiguity of message meanings cannot be tolerated. However, our results can even support individual presentation, if cues are limited in number and not urgent. Finally, P showed better RAU values compared to PR, indicating that roughness may also hinder recognition and should be avoided.

To summarize, the following guidelines can be derived from this work:

- Speech Tactons improve warnings in all measures used, so they are suggested as an addition to speech warnings;
- Perceived Urgency and Effectiveness escalate similarly in ratings, indicating that multimodal warnings are more appreciated in urgent situations;
- Annoyance is higher but more acceptable for high urgency warnings. Low urgency warnings are perceived as less effective and more annoying. In all cases though, annoyance is kept at low levels with our cues;
- Speech Tactons can be recognised well in terms of their urgency even if they are presented alone. We expect their performance to improve further when fewer cues are used.

CONCLUSIONS

This paper presented two experiments evaluating Speech Tactons, the tactile counterparts of speech warnings for drivers. Results showed that the addition of these new cues improved subjective responses of drivers to speech warnings. The warnings were clearly distinguished in terms of urgency, their annoyance was low and their alerting effectiveness changed similarly to urgency, increasing for more urgent messages and for multimodal cues. Recognition accuracy of the tactile cues' urgency was high overall and recognition accuracy of individual messages was higher for longer cues. This provides potential for using the tactile cues even alone for non-critical events, if their number is limited. We suggest the use of Speech Tactons to accompany speech warnings, so as to make use of the observed advantages of multimodal cues, but not for low urgency situations, to avoid annoyance. With the technique we provide, these tactile cues can be easily designed and added to warnings that will improve drivers' responses.

ACKNOWLEDGMENTS

The studies presented in this paper were partly funded by Freescale Semiconductor Inc., Automotive Microcontroller Product Group.

REFERENCES

1. Baldwin, C.L. and Lewis, B.A. Perceived urgency mapping across modalities within a driving context. *Applied Ergonomics*, (2013), 1–8.
2. Baldwin, C.L. and Moore, C. Perceived Urgency, Alerting Effectiveness and Annoyance of Verbal Collision Avoidance System Messages. *HFES Annual Meeting 46*, 22 (2002), 1848–1852.
3. Baldwin, C.L. Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance. *Ergonomics 54*, 4 (2011), 328–337.
4. Brown, L.M., Brewster, S.A., and Purchase, H.C. A First Investigation into the Effectiveness of Tactons. *Eurohaptics*, IEEE (2005), 167–176.
5. Brown, L.M., Brewster, S.A., and Purchase, H.C. Multidimensional tactons for non-visual information presentation in mobile devices. *MobileHCI '06*, ACM Press (2006), 231 – 238.
6. Cao, Y., Mahr, A., Castronovo, S., Theune, M., Stahl, C., and Müller, C.A. Local danger warnings for drivers: The effect of modality and level of assistance on driver reaction. *Intelligent User Interfaces*, ACM Press (2010), 239–248.
7. Cao, Y., Theune, M., and Müller, C. Multimodal Presentation of Local Danger Warnings for Drivers : A Situation-dependent Assessment of Usability. *Professional Communication Conference (IPCC)*, IEEE (2010), 226–229.
8. Edworthy, J., Hellier, E., Walters, K., Clift-Mathews, W., and Crowther, M. Acoustic, semantic and phonetic influences in spoken warning signal words. *Applied Cognitive Psychology 17*, 8 (2003), 915–933.
9. Hellier, E., Edworthy, J., Weedon, B., Walters, K., and Adams, A. The Perceived Urgency of Speech Warnings: Semantics versus Acoustics. *Human Factors 44*, 1 (2002), 1–17.
10. Ho, C., Reed, N., and Spence, C. Multisensory In-Car Warning Signals for Collision Avoidance. *Human Factors 49*, 6 (2007), 1107–1114.
11. Ho, C. and Spence, C. Assessing the effectiveness of various auditory cues in capturing a driver's visual attention. *Journal of experimental psychology. Applied 11*, 3 (2005), 157–74.
12. Hoggan, E.E. and Brewster, S.A. Crossmodal icons for information display. *CHI 2006*, ACM Press (2006), 857 – 862.
13. Lee, J.D., Bricker, D.L., and Hoffman, J.D. A Dynamic Programming Algorithm for Scheduling In-Vehicle Messages. *IEEE Transactions on Intelligent Transportation Systems 9*, 2 (2008), 226–234.
14. Lewis, B.A., Penaranda, B.N., Roberts, D.M., and Baldwin, C.L. Effectiveness of Bimodal Versus Unimodal Alerts for Distracted Drivers. *Human Factors in Driver Assessment, Training and Vehicle Design*, (2013), 376–382.
15. Li, K.A. Enabling eyes-free interaction with tactile messages based on human experience. PhD Dissertation. 2009, 107 – 116.
16. McKeown, D., Isherwood, S., and Conway, G. Auditory Displays as Occasion Setters. *Human Factors 52*, 1 (2010), 54–62.
17. McKeown, D. and Isherwood, S. Mapping Candidate Within-Vehicle Auditory Displays to Their Referents. *Human Factors 49*, 3 (2007), 417–428.
18. Politis, I., Brewster, S.A., and Pollick, F. Evaluating multimodal driver displays under varying situational urgency. *CHI 2014*, ACM Press (2014), 4067 – 4076.
19. Politis, I., Brewster, S.A., and Pollick, F. Evaluating Multimodal Driver Displays of Varying Urgency. *Automotive UI 2013*, ACM Press (2013), 92 – 99.
20. Pratt, S.M., Lewis, B.A., Penaranda, B.N., Roberts, D.M., Gonzalez, C., and Baldwin, C.L. Perceived Urgency Scaling in Tactile Alerts. *HFES Annual Meeting 56*, 1 (2012), 1303–1306.
21. SAE. Its In-Vehicle Message Priority Standard. *J2395*, (2002), 1–7.
22. Salminen, K., Surakka, V., Lylykangas, J., et al. Tactile Modulation of Emotional Speech Samples. *Advances in Human-Computer Interaction 2012*, (2012), 1–13.
23. Serrano, J., Di Stasi, L.L., Megías, A., and Catena, A. Effect of directional speech warnings on road hazard detection. *Traffic injury prevention 12*, 6 (2011), 630–635.
24. Spens, K.E., Huss, C., Dahlgvist, M., and Agelfors, E. A hand held two-channel vibro-tactile speech communication aid for the deaf : characteristics and results. *Scandinavian Audiology 47*, (1997), 7–13.
25. Tuuri, K., Eerola, T., and Pirhonen, A. Leaping across modalities: Speed regulation messages in audio and tactile domains. *Haptic and Audio Interaction Design*, Springer (2010), 10 – 19.