THE USE OF AUTOMATIC SPEECH RECOGNITION TO REDUCE THE INTERFERENCE BETWEEN CONCURRENT TASKS OF DRIVING AND PHONING

Robert Graham, Chris Carter* & Brian Mellor***

* HUSAT Research Institute, Loughborough University, UK
tel: +44-1509-611088, fax: +44-1509-234651, email: r.graham@Lboro.ac.uk

** Speech Research Unit, Defence Evaluation and Research Agency (DERA) Malvern, UK

ABSTRACT

Previous research has found that using manually-operated mobile phones while driving significantly increases the risk of a collision. It has been suggested that automatic speech recognition (ASR) interfaces may reduce the interference between the tasks of phoning and driving. A laboratory experiment was designed to examine this hypothesis, and also to investigate the optimal design for in-car ASR systems. Forty-eight participants dialled phone numbers from memory while carrying out a concurrent tracking task. Tracking performance was found to be adversely affected while using a manual phone. This effect was significantly reduced, although not eliminated, with a speech phone. Participants also perceived the mental workload of manual dialling while driving to be greater than speech dialling. A system of audio feedback was found to be marginally preferable to combined audio plus visual feedback. The recognition accuracy of the ASR device did not appear to have any significant bearing on driving performance nor acceptance. The results are encouraging for the use of speech interfaces in the car for phone and other functions.

1. INTRODUCTION

A recent study by Redelmeier & Tibshirani [1], reported in the New England Journal of Medicine, found that the use of cellular telephones while driving significantly increased the risk of a vehicle collision. The authors studied 699 drivers who had cellular phones and who were involved in accidents which involved property damage but no personal injury. By obtaining detailed billing records, they were able to analyse those calls made close to the accident; for calls made within ten minutes of the collision, the relative risk was quadrupled. This study represents probably the first direct evidence of a link between phone use and collisions. Previous research had found other negative effects of phoning on driving behaviour, including increased lateral lane deviations, impaired judgment of gaps, and increased reaction times to speed changes or brake lights of the vehicle in front (for a review, see [2]).

The interference between the tasks of phoning and driving is not surprising when one considers that to operate a standard phone interface places extra loading on the overburdened visual-manual modality. The use of ASR for phone functions has the potential to significantly reduce this interference. ASR operation is eyes-free and hands-free, allowing drivers to keep their visual attention on the road, and their hands on the steering wheel. As well as improving driving safety, ASR could increase system acceptability by simplifying the dialogues between the user and system. ASR could be particularly effective for those phoning tasks which currently require significant visual and manual attention to be directed away from the primary driving task, such as dialling a number.

Although many of the larger telecommunications companies are investigating speech-operated mobile phones, and their technical feasibility has been reported in the literature, little research has been carried out to test their proposed benefits within a car environment. A limited study by Serafin et al [3] is one exception. Their driving simulator experiment suggested that voice operation could improve lane-keeping and allow faster dialling than manual operation in certain circumstances. However, the study only tested 12 drivers and interface modality was only one of 8 independent variables.

The main hypothesis examined in the present study, therefore, was that speech-operated phoning would result in lower levels of interference with the driving task than manual phoning.

In order to optimise the design of future in-car speech recognition systems, two other important human factors variables were included in the experiment. The first was the sensory modality of the feedback provided by the recognition software. In the car environment, auditory feedback (speech or tones) allows the driver to keep their eyes on the road, but is transient, meaning that it cannot easily be re-checked. It is also impossible to ignore, and therefore potentially irritating for the user. Visual feedback, probably via a dashboard text display, has the advantage that it can be scanned as and when required, but it requires the eyes to be taken off the road, and may not be monitored with the same degree of accuracy as speech [4]. A redundant combination of both modalities is often suggested as the optimal solution.

The second issue was the accuracy of the ASR system. Although speech technology vendors generally like to deny it, recognition systems always make errors. This is particularly the case in the car environment which is

characterised by high levels of noise and user stress/ workload. Perhaps the most important question is what degree of accuracy is required for the system to be acceptable by users. The present study used the approach of artificially degrading the performance of a high-performance speech recogniser such that three levels of accuracy were tested.

2. METHOD

2.1. Experimental Design

Three independent variables were included in the experiment, as follows:

• phone interface modality - standard button phone ('manual'), speech recognition with auditory feedback ('speech audio') or speech recognition with auditory feedback plus a visual display ('speech combined').

• concurrent task - phoning at the same time as driving ('phoning + driving'), phoning only or driving only.

• for the speech phone conditions, recognition accuracy - 0%, 3%, or 6% additional errors.

The variables of interface modality and driving load were within-subjects, whereas the variable of recognition accuracy was between-subjects.

A variety of objective and subjective measures were taken. The dependent variables that are reported within this paper include driving/tracking performance, subjective self-ratings of mental workload, and subjective preference ratings.

2.2. Participants

Forty eight participants were recruited for the study, via the HUSAT Research Institute's subject database or advertisements in local shops. They were made up of 27 males and 21 females, and all were aged between 20 and 50 years with a mean age of 35.2. All were regular drivers, and the majority did not regularly use a mobile phone. They were randomly allocated into three treatment groups (see 'recognition accuracy' above), while ensuring that each group was matched for gender (9 males, 7 females) and age (mean age of each group between 34.9 and 35.5 years). Participants were paid UK£15 for their time.

2.3. Apparatus

'Driving' Task: Because of the ethical difficulties of real-road studies, an artificial tracking task was designed to mimic aspects of driving. Participants were seated in front of a PC screen on a desk. A steering wheel input device was mounted on the front of the desk, and brake and accelerator pedals positioned below the desk. Software on the PC showed a white rectangle ('the car') moving within a larger horizontal blue strip ('the lane'). The lateral velocity of the car was randomly varied, and participants were required to compensate for this movement, using the steering wheel, to keep the car at the centre of the lane. Piloting ensured that the perceived difficulty of the driving task was comparable to the difficulty of negotiating a real road with little traffic and few bends. At the same time as the main driving task, occasional visual stimuli were presented in the upper part of the PC screen; participants had to respond with a movement of their foot from the accelerator pedal to depress the brake pedal. The software logged the tracking performance, measured as root-mean-squared (RMS) error in the number of pixels away from the centre of the lane. The responses to each peripheral target were also logged, but these data are not presented here.

Speech recogniser: Improvements in automatic speech recognition (ASR) performance encouraged the use of a real recogniser rather than a Wizard-of-Oz simulation. The recogniser used was DERA's AURIX recognition unit which comprises a stand-alone processor running a sub-word hidden Markov model based, fully continuous connected word algorithm. The model set was composed by extracting suitable context sensitive sub-word units from a data-set trained on a phonetically balanced corpus. This allowed a rapidly reconfigurable vocabulary for development and, with addition of extra digit data, high accuracy recognition. To provide a controllable error rate, recognition errors were inserted into the system at 0%, 3% or 6% word error rates. The types of errors were based on real recognition results in order to provide consistent recogniser characteristics.

Speech interface: The speech-based interface was developed using an iterative model of prototyping, evaluation, improvement and re-evaluation. The rapid prototyping environment used was DERA's GUIDE tool-kit for Visual Basic [5]. Dialling was initiated using the command word "Phone" and the call made using the word "Dial", to mimic a mobile phone user interface. A press-to-talk switch (PTT) mounted on the steering wheel was used to activate the AURIX recogniser. Digit entry could be either continuous or in chunks segmented by PTT action. Error correction was carried out using the word "Correction" to delete the previous chunk or "Cancel" to delete the whole entry. The word "Zero" was used rather than "Oh" to represent the digit '0', in order to optimise recognition accuracy. The word "Double" (e.g. "three double-four two") was also a valid input. Audio feedback was provided on release of the PTT by voice output of the last digit chunk using a recorded female voice. A short 'ping' was also sounded after each command word was recognised. In the 'combined modality' condition, a visual display was also provided on a PC monitor to the left of the driving task monitor. This printed the word "Phone" followed by the digits, with chunks separated by a space.

Manual mobile phone: A Nokia Orange (model NHK-1XA) phone hand-set was chosen for the experiment. This had the attributes of being one of the most common hand-sets used in the UK, with an interface which was simple and representative of a number of other hand-sets.

2.4. Procedure

The trial took approximately two hours, roughly divided into a 40-minute training period, a 70-minute experimental period, and a 10-minute debrief. During the training period, the experimenter demonstrated the use of the driving game

and each type of phone to the participant. Participants practiced using each of the phones, first without then with concurrent driving, until their performance reached a minimum criterion. If at the end of training, their speech recognition rate did not reach 90% (without added artificial errors), they were excluded from the experiment.

The experiment proper was divided into six blocks; manual only, manual + driving, speech audio only, speech audio + driving, speech combined only, speech combined + driving. The presentation order of each block was balanced between subjects. During each block, the participant was required to dial five phone numbers from memory (on arrival, the experimenter had been provided with 5 numbers the participant knew from memory, with a name tag to associate with each number). The dialling of the numbers was prompted by the experimenter (e.g. "now call Mum", "now call Bob"). The three blocks which involved concurrent driving each lasted 8 minutes, allowing periods while using the phone to be compared with periods of driving only. Immediately after each block, participants completed a NASA-RTLX mental workload questionnaire [6]. In this standard questionnaire, the workload experience is self-rated on six sub-scales - mental demand, physical demand, temporal demand, performance, effort and frustration level - and an overall mean rating calculated on a 0-100 scale. At the end of the trial, participants completed a final questionnaire in which they rated their preferences for each of the 3 types of phone interface. They responded to a series of statements (e.g. "The system was easy to use while driving", "I would like to have this system in my car", etc.) on a linear scale marked from 'strongly agree' to 'strongly disagree'.

3. RESULTS

Within each of the 3 driving blocks, mean RMS tracking error was calculated for periods of driving only compared with driving while using one of the 3 types of mobile phone. These data were subjected to a 2x3x3 ANOVA involving the variables of concurrent task (driving only, driving + phoning), phone modality (manual, speech audio, or speech combined) and ASR accuracy (0% errors, 3% errors, 6% errors) respectively. As expected, the main effect of concurrent task was highly significant ($F(1,45)=54.3$, $p<0.0001$) showing that tracking performance while phoning was poorer than driving only. The main effect of modality was also highly significant ($F(2,90)=83.3$, $p<0.0001$), indicating that tracking performance in the manual phone condition was poorer than either of the speech phone conditions. The significant interaction effect between the variables of concurrent task and modality ($F(2,90)=43.7$, $p<0.0001$) is illustrated in figure 1. A number of interesting features can be noted. Driving performance while using the manual phone was substantially worse than the speech phones, and slightly worse in the speech combined condition than the speech audio condition (contrast: $F(1,47)=3.8$, $p=0.54$). In all conditions, driving while phoning was significantly worse than driving alone, but particularly so in the manual condition (contrast: $F(1,47)=199.7$, $p<0.0001$). There also seemed to be some 'carry-over' of poor tracking performance from the periods of using the manual phone to the periods between calls, as indicated by the manual/ driving only performance being poorer than the speech/ driving only performance. All main effects and interaction effects involving the variable of ASR accuracy on RMS tracking error were non-significant.

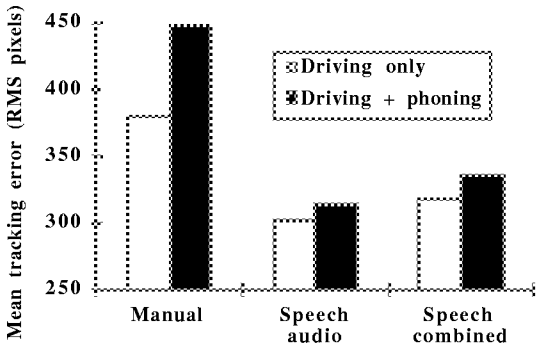

Figure 1: Interaction between concurrent task and phone modality on driving performance

NASA-RTLX workload ratings were obtained from each of the 6 experimental blocks and subjected to a 3 (modality) x 2 (concurrent task - phoning only vs. phoning while driving) x 3 (ASR accuracy) ANOVA. Phoning while driving was rated as significantly more demanding than phoning alone ($F(1,45)=162.8$, $p<0.0001$). There was also a significant main effect for modality ($F(2,90)=4.04$, $p=0.02$). The interaction effect between modality and concurrent task, illustrated in figure 2 below, was highly significant ($F(2,90)=57.9$, $p<0.0001$). This showed that driving while using the manual phone was rated the most demanding of all the conditions, but using the manual phone alone was least demanding. Using the speech phones while driving were found to be more demanding than using the speech phones alone. There were no differences between the speech audio and speech combined conditions. Again, no effects were found for the ASR accuracy variable.

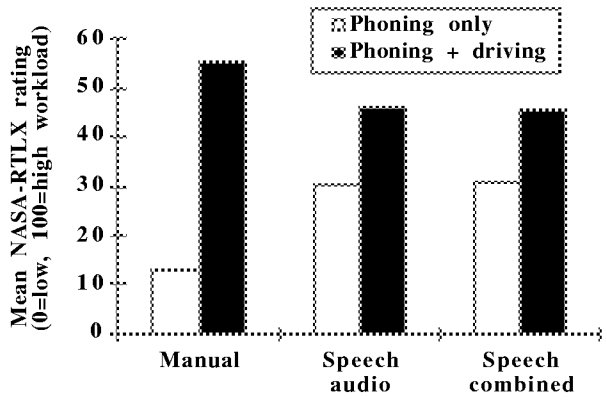

Figure 2: Self-assessed mental workload ratings for the six conditions

Data from the final subjective preference questionnaire mirrored the workload results. Participants felt that the speech phones were easier to use while driving at the same time than the manual phone, but that the manual phone was easiest to use when operated by itself. They were aware that their driving performance was affected more by using the manual phone than the speech phones. However, in response to the statement "my driving/ tracking performance was affected by using the system", all phones were rated towards the 'agree' end of the scale. The statement "I would like to have this system in my car" showed a preference for the speech phones over the manual version. There were no significant differences between the two feedback modalities of speech phone for any of the questions. Similarly, the recognition accuracy variable did not have any affect on the ratings.

4. DISCUSSION

The experimental results indicated that the use of a standard manual mobile phone while driving adversely affected driving performance. This interference was significantly reduced through the use of an ASR interface. The objective data were backed up by subjective data showing that manual dialling while driving was perceived to be more demanding than speech-dialling while driving.

The results have important implications for future legislation regarding driving safety and interface design. Although a number of countries have now banned the use of mobile phones while driving, there remains a large part of the world in which such phones may be contributing to accident causation. The present experiment suggests that speech operation may be one way of improving safety. It is important to note, however, that the use of speech interfaces for phone functions cannot completely eliminate the effects on driving, and the present study demonstrated a significant, albeit reduced, interference effect.

There are a number of other in-car systems which currently rely on manual controls and visual displays, and which could benefit from speech interfaces. Most current attention is being given to the new range of Intelligent Transport Systems (ITS), such as those which aid the driver in route navigation or obtaining travel and traffic information. Many of these have interfaces which are significantly more complex than a mobile phone and therefore have the potential to impact even more on concurrent driving performance.

With regards to the design of in-car speech interfaces, the results showed a marginal advantage for a system of audio only feedback over a system of combined visual and auditory feedback. The presence of a visual display seemed to distract participants from the driving task and most felt that it was superfluous. The recognition accuracy of the ASR device did not appear to have any bearing on objective driving performance or subjective responses. On one hand, this would seem to show that participants were prepared to cope with a recognition error rate of at least 6% (note that the actual error rate, taking into account real errors as well as artificially-added errors was more than this). On the other hand, the recognition accuracy variable may simply have been 'drowned out' by the strong preference for the speech phones over the manual phone and/or the wide natural variation in individuals' recognition performance.

5. ACKNOWLEDGMENTS

This work was carried out as part of the SPEECH IDEAS project under the UK government's LINK Inland Surface Transport collaborative research programme, funded by the Economic and Social Research Council and the Department of the Environment, Transport and the Regions. The ASR application was provided and supported by DERA, and the manual phone hand-set provided by Orange PCS Ltd. For more information on the project and/or further details of this study, please contact the first author.

6. REFERENCES

1. Redelmeier, D. A., and Tibshirani, R. J. "Association between cellular-telephone calls and motor vehicle collisions", *New England Journal of Medicine, 336(7)*: 453-458, 1997.

2. Parkes, A. M. "Voice communications in vehicles", In A. M. Parkes & S. Franzen (Eds.), *Driving Future Vehicles* (pp. 219-228), Taylor & Francis, London, 1993.

3. Serafin, C., Wen, C., Paelke, G., and Green, P. "Car phone usability: a human factors laboratory test", *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (pp. 220-224), Human Factors Society, Santa Monica, CA, 1993.

4. Noyes, J. M., and Frankish, C. R. "Errors and error correction in automatic speech recognition systems", *Ergonomics, 37(11)*: 1943-1957, 1994.

5. Mellor, B. A., Tomlinson, M. J., and Coleman, N. J. "The generic user interface development environment GUIDE: overview and features", *Proceedings of the ESCA Workshop on Spoken Dialogue Systems* (pp. 117-120), ESCA, Grenoble, France, 1995.

6. Byers, J. C., Bittner, A. C., and Hill, S. G. "Traditional and raw task load index (TLX) correlations: are paired comparisons necessary?" In A. Mital (Ed.), *Advances in Industrial Ergonomics and Safety I* (pp. 481-485), Taylor & Francis, London, 1989.