Multimodal Inference for Driver-Vehicle Interaction

Tevfik Metin Sezgin
College of Engineering
Koç University
Istanbul, Turkey
mtsezgin@ku.edu.tr

Ian Davies
Computer Laboratory
University of Cambridge
Cambridge, CB3 0FD, UK
Ian.Davies@cl.cam.ac.uk

Peter Robinson
Computer Laboratory
University of Cambridge
Cambridge, CB3 0FD, UK
Peter.Robinson@cl.cam.ac.uk

ABSTRACT

In this paper we present a novel system for driver-vehicle interaction which combines speech recognition with facial-expression recognition to increase intention recognition accuracy in the presence of engine- and road-noise. Our system would allow drivers to interact with in-car devices such as satellite navigation and other telematic or control systems. We describe a pilot study and experiment in which we tested the system, and show that multimodal fusion of speech and facial expression recognition provides higher accuracy than either would do alone.

Categories and Subject Descriptors

H5.2 [**User Interfaces**]

General Terms

Design, Human Factors.

Keywords

Driver monitoring, facial-expression recognition, speech recognition, multimodal inference.

1. INTRODUCTION

Accurate measurement of drivers' intentions and responses is an important requirement for effective human-vehicle interaction. Detecting user response reliably is especially important in interaction scenarios where feedback is expected in response to a question (e.g., posed by an in-car navigation system). So far, in-car interaction modalities have been restricted to traditional graphical dialog-box representations and speech-based input. Traditional graphical representations usually require interacting with small touch-sensitive displays, and can be distracting because of the visual attention required for the interaction. Speech-based interfaces, on the other hand, offer a more natural modality for interaction, although their usefulness is subject to a number of limitations.

We have developed a framework for automatic analysis of drivers' facial expressions with the goal of adding facial displays to the list of modalities available for human-vehicle interaction. Specifically, we have investigated the feasibility of combining head-based displays with speech in order to achieve higher recognition results in the presence of noise. We studied the effects of noise in an interaction scenario that required responses to a series of "yes/no" questions, which are typical in interacting with a navigation system (e.g. "The gas is running low. Would you like directions to the nearest gas station?").

Because vehicle-noise and the willingness of the driver to express themselves clearly through spoken dialogue are the primary causes of misrecognized speech, we focused our investigation on intelligent fusion of head-display and speech information for varying noise levels and varying speaker volumes. Using our in-house driving simulator, we conducted a pilot study where we recorded a participant answering a series of "yes/no" questions while driving. We recorded separate audio and video streams that captured the driver's speech and facial displays. Based on promising results from the pilot study, we conducted a larger controlled experiment with 4 further subjects (age 22–50) to verify our findings. Clearly four subjects is still only a small sample and we would need to run a much larger study in order to draw stronger quantitative conclusions from our results.

We implemented a speech recognition application for processing the audio and used our own facial expression recognition software [[5]] to interpret the video stream. We used Support Vector Machines to fuse audio- and video-based inference results and constructed a multimodal recognition engine that outperforms the individual modalities.

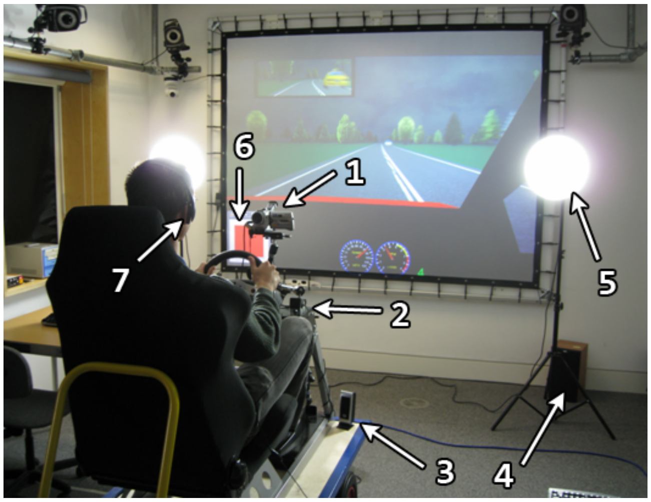

Figure 1: The physical setup of our experiment

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
Copyright 2009 ACM 978-1-60558-772-1/09/11...$10.00.

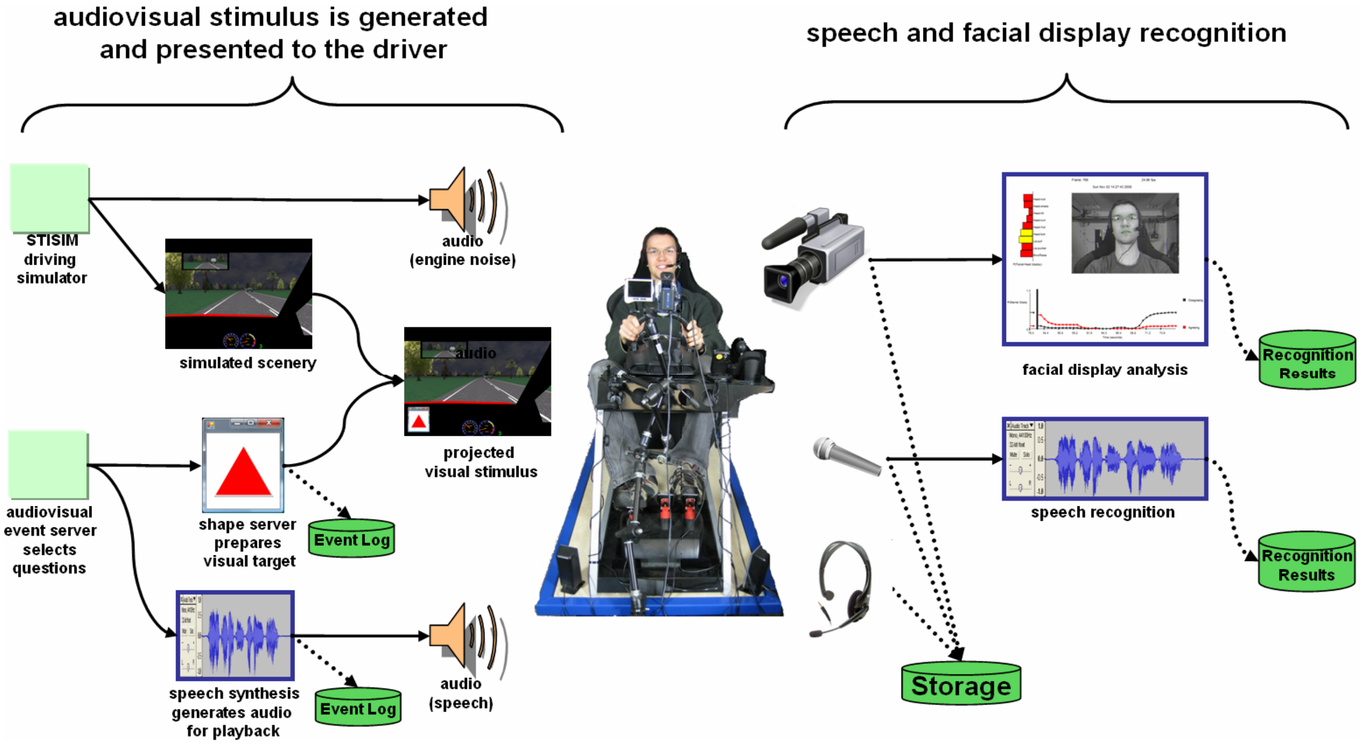
audiovisual stimulus is generated and presented to the driver
speech and facial display recognition
STISIM driving simulator
audio (engine noise)
simulated scenery
audio
projected visual stimulus
audiovisual event server selects questions
shape server prepares visual target
Event Log
speech synthesis generates audio for playback
Event Log
audio (speech)
facial display analysis
Recognition Results
speech recognition
Recognition Results
Storage

Figure 2: System Structure

2. EXPERIMENTAL SETUP

Figure 1 shows the physical setup of our experiment. We had a driving simulator equipped with a digital video camera for recording facial displays (1). A microphone secured to the frame in front of the driver (2) was used to record audio for speech recognition. Two pairs of speakers were used for playing audio stimuli consisting of questions for the driver (3), and simulated engine noise (based on vehicle speed and road conditions) generated by the STISIM driving simulator software (4). A pair of studio lights (5) was used to control the lighting conditions. During the experiment, the driver was asked a series of questions referring to a target shape displayed next to the virtual dashboard projected on a screen (6). The driver also wore a headset with microphone to record higher quality speech for further analysis (7).

The audio stream used for speech recognition was captured by a microphone on the dashboard (#2 in Figure 1). The speech was processed by a Microsoft SAPI 5 application customized to recognize 29 words that a driver would be likely to use in the context of driving including "yes," "no," and five variants (such as "yeah", "nope"). The video used for recognizing head displays was captured by a digital camcorder positioned directly in front of the driver (#1 in Figure 1) and was processed by our mind-reading software [[5]] trained to recognize facial displays for agreement and disagreement.

The pilot study was designed to investigate the effects of noise in an interaction scenario requiring responses to a series of "yes/no" questions. Such situations frequently arise while interacting with a navigation system. In order to elicit a verbal or non-verbal agreement/disagreement response, we designed a task where the user is first presented a geometric shape (like the red triangle as shown in Figure 2), and then asked a question about the shape, which can be answered verbally ("yes" or "no"), and non-verbally (head-nod or head-shake). We refer to each round of shape presentation as a trial. In order to collect representative data for varying noise levels, we had a total of 60 trials consisting of 30 questions requiring an affirmative response, and 30 requiring a negative response. For the subsequent larger experiment, we ran 100 trials with each of the 4 participants. The order of trials was randomized. Each trial consisted of playing an audio clip which asked a question (e.g. "Is the shape a red triangle?"), displaying the target shape in a particular colour for 2 seconds, then expecting an appropriate verbal and/or non-verbal response from the driver. We waited 5 seconds between trials.

The primary source of noise in the study was the engine noise generated by the STISIM driving simulator based on the vehicle speed and road conditions. We started the experiment with the noise volume set to zero, and gradually increased the noise level over the course of the drive. An appropriate range of noise levels was determined by measuring the noise level in a transit van at 60mph (100kph) with a digital sound level meter. The loudest noise level experienced by the driver was 75 dBA, so we used this as the maximum in the experiment. For the final 40 trials in the large controlled experiment, the driver was asked to speak louder or quieter in order to provide additional data relating to variation in speaker loudness.

3. RESULTS

3.1 Audio-Based Recognition Results

Speech recognition events were attributed to particular questions by nearest-neighbour matching. Speech events in the audio were identified by clustering. The RMS value of the audio data for each cluster provided a measure of speaker loudness. The background noise was measured by averaging the absolute value of the audio signal in the time between the question finishing and the driver speaking.

Figure 3 presents speech recognition results from the pilot study for increasing levels of noise. As seen in this graph, speech recognition works reasonably well when the noise is low. The recognition accuracy is very poor for high noise levels, and there is a transitional gray zone between the high-noise and low-noise segments where the recognition results are unpredictable. The recognition accuracy based on speech alone was 57% in the pilot study.

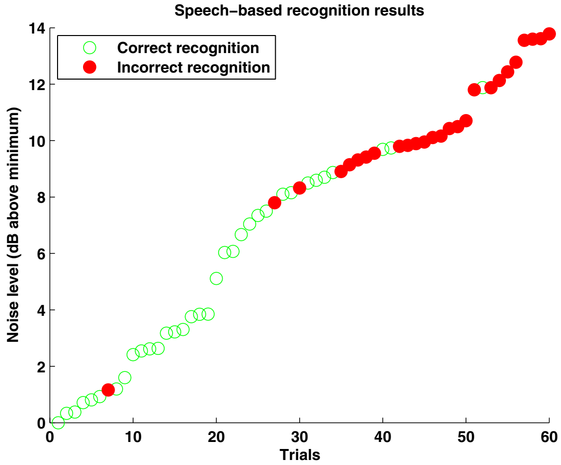

Figure 3: Speech-Based Recognition Results

3.2 Video-Based Recognition Results

To compute the head display hypotheses for each trial, we compared the average probabilities of agreement and disagreement for a short video segment taken briefly after the completion of the question. Although our facial display analysis software computes probability scores for agreement and disagreement, these do not represent actual probabilities in the Bayesian sense and this prohibits direct comparison of the values. Therefore we treat these numbers as scores and train a linear SVM classifier that maps a pair of agreement/disagreement scores to an agreement or disagreement decision. Figure 4 shows the video-based recognition results from the pilot study for increasing levels of noise. As seen in this graph, video recognition generally works well, and its accuracy does not appear to depend on the noise level. The recognition accuracy based on video alone was 78% in the pilot study.

3.3 Multimodal Fusion Results

Our framework for multimodal fusion is based on the observation that speech recognition works remarkably well for low-noise conditions, but performs quite badly in high noise conditions, while the video-based recognition performance is reasonably accurate regardless of the noise level. We fuse the audio and video information at the decision level by treating the results of our speech and head-display analysis as inputs to a classifier along with the noise level of the environment. More specifically we consider a classification problem where the inputs are 3-tuples $<a_i, v_i, n_i>$, which respectively represent the class assigned by the speech recognizers (a_i: yes/no/other, where 'other' means nothing was recognised), head-display recognizers (v_i: agreement/ disagreement), and the noise level (n_i) for trial i. For the subsequent experiment, we also included speaker loudness. Although this appears to be a simple construction, the high dimensional space representing the decision problem is sufficiently complex that it is not linearly separable. Although some of the categorical data inputs could be re-ordered to achieve a better space, it is highly likely that the non-linearity would still not be avoided when more features (such as a measure of the complexity of the speech recognition grammar) were added to the input space. In order to deal with this non-linearity, we trained Support Vector Machines (SVMs) with Radial Basis Function kernels for multimodal fusion.

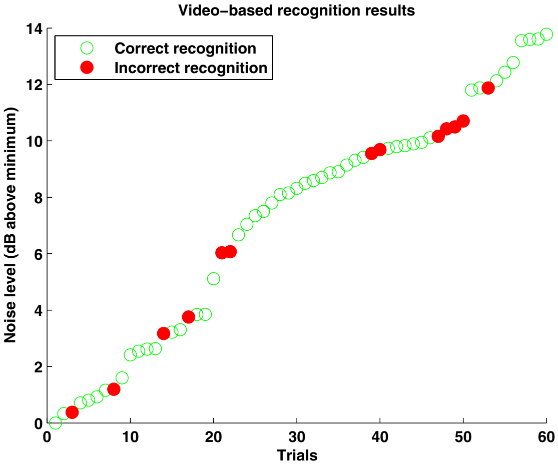

Figure 4: Video-Based Recognition Results

We measured the performance of our SVM using 10-fold cross validation, and ran it 10 times while randomly shuffling the training and testing sets, thus performing bootstrapping. Representative results of one cross-validated run are shown in Figure 5. The average recognition rate for the multimodal classifier was 89% in the pilot study, with a standard deviation of 3.1. This is substantially higher than either of the individual modalities considered separately.

The results obtained in the larger controlled experiment (summarised in Table 1) supported those from the pilot study. In all cases multimodal inference outperformed each individual mode. Note that our system was particularly valuable for subject 4, for whom speech recognition alone was entirely inadequate.

4. DISCUSSION

These results suggest that multimodal approaches where the individual modalities complement each other and compensate for their shortcomings have the potential to reduce recognition errors, substantially improve driver-vehicle interaction, and enhance the overall driving experience.

It is worth noting that the speech recognition accuracy depends on factors other than just the noise level, such as the speech

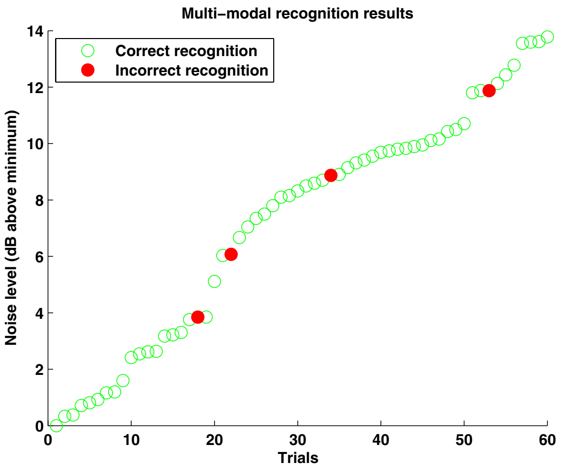

Figure 5: Multi-Modal Recognition Results

recognition engine in use, and the complexity of the grammar that guides the parsing process. In particular, the number and choice of terminals in the grammar (i.e., the vocabulary) affects accuracy.

For example, our later analysis with a more recent version of the Microsoft Speech Recognizer improved the average speech recognition accuracy by 11 points, even though the increase was not found to be statistically significant.

Using a simpler grammar also resulted in improvements, albeit in smaller quantities. For example, removing the top two words that caused misrecognitions one at a time resulted in 4.7% and 5.2% increases in the average recognition accuracies, compared to a baseline computed using the original grammar. Recognition accuracy for speech and video based inference systems could easily be increased with further refinement, but it seems likely that multimodal fusion would always yield a better result.

The architecture of our system was specifically designed to allow further experiments to be carried out using other modalities. The modular design allows any number of input modalities to be used, and the multimodal fusion engine can be easily modified to accept them and produce appropriate outputs. This system will now allow new hypotheses to be tested quickly and easily.

Subject	Speech-Based Recognition Accuracy (%)	Video-Based Recognition Accuracy (%)	Multimodal Accuracy (Mean, %)	Multimodal Accuracy (S.D.)
1 (Pilot)	57	78	88.7	3.1
2	46	67	72.3	2.5
3	54	69	75.8	1.5
4	31	65	70.4	3.4
5	39	69	73.9	2.6

Table 1: Controlled Experiment Recognition Results

5. RELATED WORK

It has long been recognised that current driver-vehicle interaction techniques are inadequate for safe and effective use of increasingly complex in-car devices. People have begun to investigate alternative methods, such as speech recognition [[1]]. A considerable amount of work has been done to tackle the problem of speech recognition in noisy environments [[2],[3]], with good results. However, recognition accuracy always decreases as noise increases and there is a limit to how much it can be improved.

Automatic facial expression recognition has also been dealt with previously [[4],[5]], and Rong and Tan implemented explicit head-nod and shake detection [[6]]. Although this experiment only deals with agreement and disagreement, our facial expression recognition software [[5]] uses more than just nod and shake detection and is capable of distinguishing several more mental states.

Several people have combined vision-based approaches with speech recognition - usually in the context of broader affective inference for emotion recognition. Busso et al. used multi-modal fusion of speech and facial expression to identify the six archetypal emotions of surprise, fear, disgust, anger, happiness, and sadness [[8]], while Jaimes and Sebe provide a survey of many uses of multi-modal fusion in the field of human-computer interaction [[9]]. Work has also been done on supporting speech recognition specifically; Cooke et al. have collected a large audio-visual corpus designed for both automatic speech recognition-based and perceptual studies of speech processing [[7]] and Oviatt has shown that multi-modal approaches can support significant levels of mutual disambiguation of errors in speech processing [[10]]. Erzin et al. have developed a multilevel Bayesian decision fusion scheme, combining vision and speech in automotive environments for identification and authentication [[9]].

6. FUTURE WORK

These results suggest that it would be worth pursuing further investigations of affective inference as a component in the dialogue between a driver and an in-car telematic system.

The next step would be to move from detection of simple agreement and disagreement to a more elaborate dialogue involving a broader range of options in a larger and more realistic task, such as interacting with a satellite navigation device. This could involve programming a destination as well as responding to incorrect turns and providing routing options. Understanding the driver’s concentration level so as to avoid distraction from more critical driving tasks would also be important and is a potential application of affective computing techniques to this domain. Another interesting study would be to compare decision-level with feature-level fusion.

7. SUMMARY

We have demonstrated, tested and validated a system for driver-vehicle interaction which uses multimodal fusion of speech and facial expression recognition. We have shown that combining these inference techniques gives a level of accuracy unattainable when using either system on its own. The architecture of the inference system we built provides a more general framework in which new techniques can be tested.

8. REFERENCES

[1] Lee, J.D. et al. Speech-based Interaction with In-vehicle Computers: The Effect of Speech-based E-mail on Drivers’ Attention to the Roadway. *Human Factors 43*, 2001, 631–640.

[2] Gong, Y. Speech recognition in noisy environments: A survey. *Speech Communication,* 16 (1995), 261–291.

[3] Frey, B.J. et al. Learning dynamic noise models from noisy speech for robust speech recognition. In *NIPS*, 2002.

[4] Fasel, B., Luettin, J. Automatic facial expression analysis: a survey. *Pattern Recognition 36*, 1 (2003), 259–275.

[5] El Kaliouby, R., Robinson, P. Real-Time Inference of Complex Mental States from Facial Expressions and Head Gestures. In *Real-Time Vision for Human-Computer Interaction*, 2005, 181–200.

[6] Rong, G., Tan, W. A real-time head nod and shake detector using HMMs. *Expert Systems with Applications 25*, 3 (2003), 461–466.

[7] Cooke, M. et al. An audio-visual corpus for speech perception and automatic speech recognition. *The Journal of the Acoustical Society of America 120*, 5 (2006) , 2421–2424.

[8] Busso, C. et al. Analysis of emotion recognition using facial expressions, speech and multimodal information. *Proc. ICMI 2004*, 205–211.

[9] Jaimes, A., Sebe, N. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding 108*, 1–2 (2007), 116–134.

[10] Erzin, E. et al. Joint Audio-Video Processing For Robust Biometric Speaker Identification In Car. *DSP for In-Vehicle and Mobile Systems*, 2005. 237–256

[11] Oviatt, S. L. Mutual disambiguation of recognition errors in a multimodal architecture. *Proceedings of the Conference on Human Factors in Computing Systems (CHI'99)*, 576-583. New York: ACM Press.