Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces

Christian Martyn Jones[1] and Ing-Marie Jonsson[2]

[1] University of the Sunshine Coast, Queensland, 4558, Australia
[2] Department of Communication, Stanford University, California 94305, USA
cmjones@usc.edu.au, ingmarie@csli.stanford.edu

Abstract. The automotive industry are integrating more technologies into the standard new car kit. New cars often provide speech enabled communications such as voice-dial, as well as control over the car cockpit including entertainment systems, climate and satellite navigation. In addition there is the potential for a richer interaction between driver and car by automatically recognising the emotional state of the driver and responding intelligently and appropriately. Driver emotion and driving performance are often intrinsically linked and knowledge of the driver emotion can enable to the car to support the driving experience and encourage better driving. Automatically recognising driver emotion is a challenge and this paper presents a performance analysis of our in-car acoustic emotion recognition system.

Keywords: In-car systems, emotion recognition, emotional responses, driving simulator, affective computing, speech recognition.

1 Introduction

Cars are becoming feature rich with numerous interactive technologies embedded such as audio/video players, satellite navigation, hands-free mobile communications, climate controls and performance settings (suspension, semi-automatic gearbox etc). Attention theory suggests that speech-based interactions are less distracting to the driver than interactions with a visual display [1]. The introduction of speech-based interaction and conversational systems into the car highlights the potential influence of linguistic cues (such as word choice and sentence structure) and paralinguistic cues (such as pitch, frequency, accent, and speech rate which provide acoustic indicators of emotion). Driving presents a context in which a user's emotional state plays a significant role. The road-rage phenomenon [2] provides one undeniable example of the impact that emotion can have on the safety of the roadways. Considering the effects of emotion, and in particular that positive affect leads to better performance and less risk-taking, it is not surprising that research and experience demonstrate that happy drivers are better drivers [3]. The emotion of the car-voice has also been found to impact driving performance. Results from a study pairing the emotion of the car-voice with the emotion of the driver showed that matched emotions positively

C. Stephanidis (Ed.): Universal Access in HCI, Part II, HCII 2007, LNCS 4555, pp. 411–420, 2007.

impacted driving performance [4]. With a focus on driving safety and driving performance, these results motivate our research to investigate the design of an emotionally responsive car.

2 Affective Computing and Acoustic Emotion Recognition

The field of affective computing is growing with considerable research interest in automatically detecting and recognising human emotions [5], [6], [7]. Emotional information can be obtained by tracking facial motion, gestures and body language using image capture and processing [8]; tracking facial expressions using thermal imaging [9]; monitoring physiological changes using biometric measurements taken from the steering wheel and seat/seat-belt [10]; and also analysing the acoustic cues contained in speech [7]. Our research considers in-car affective computing where the car can recognise driver emotion and respond intelligently to support the driving experience such as offering empathy or providing useful information and conversation. In-car speech controlled systems are already commonplace with voice-controlled satellite navigation, voice-dial mobile phones and voice-controlled multimedia systems. Therefore we have adopted a speech based emotion recognition system for the in-car device.

Most emotions in speech are associated with acoustic properties such as fundamental frequency, pitch, loudness, speech-rate and frequency range [11]. The emotion recognition (ER) system presented in this paper uses 10 acoustic features including pitch, volume, rate of speech and other spectral coefficients. The system then maps these features to emotions such as boredom, sadness, grief, frustration, extreme anger, happiness and surprise, using statistical and neutral network classifiers. The emotion recognition system uses changes in acoustic features representative of emotional state whilst suppressing what is said and by whom. It is therefore speaker independent and utterance independent and can be readily adapted to other languages. Using a test set of previously unseen emotive speech, the overall performance of the emotion recognition system is greater than 70% for five emotional groups of boredom, sadness/grief, frustration/extreme anger, happiness and surprise. The emotion recogniser can track changes in emotional state over time and present its emotion decisions as a numerical indicator of the degree of emotional cues present in the speech.

3 The Emotional In-Car Conversation Interface

The emotive driver project builds on previous research which assessed the feasibility of automatically detecting driver emotions using speech [12] and developed a conversational interaction between the driver and car using driver emotion recognition and intelligent car response [13]. With knowledge of driver emotion the car can modify its response both in the words it uses but also the presentation of the message by stressing particular words in the message and speaking in an appropriate emotional state. As the car alters its ‘standard’ voice response it will be able to empathise with the driver and ultimately improve the wellbeing and driving performance. A previous study on pairing the emotional state of the driver with the emotional colouring of the car voice shows that pairing the emotions has an enormous positive influence on driving performance [4]. The same study reported that the engagement and amount of conversations was significantly higher when the emotions were paired. To pair emotions between the driver and car voice requires accurate automated emotion recognition. In this paper we extend the emotive driver project to provide quantitative analysis of the performance of our acoustic emotion recognition engine within a simulated car setting.

4 Experimental Method

The experimental study involved recording conversations between the driver and the car which we analysed to test the performance accuracy of the automatic acoustic emotion recognition. The experiment consisted of an 8 day study at Oxford Brookes University, UK using 41 participants, 20 male and 21 female. Participants were all in the age group 18 – 25, all driving with a mixed informational and conversational in-car system.

The experimental study used the STISIM driving simulator [14] for the driving session with the in-car information system. The driving simulator displays the road and controls such as speedometer and rev counter, and provides a rear-view mirror and control buttons for side-views left and right. Participants are seated in a real car seat and control the driving simulator using an accelerator pedal, a brake pedal, and a force-feedback steering wheel. All participants experienced the same pre-defined route and properties for both driving conditions and the car. The drive lasted approximately 20 minutes for each participant.

Engine noise, brake screech, indicators, sirens etc together with the output from the in-car information system was played through stereo speakers. The in-car information system was described as a system that would present two types of information to the drivers, informational and conversational. The informational part of the system related to road conditions, traffic and driving conditions such as ’the police often use radar here so make sure to keep to the speed limit’. The second part focuses on engaging the driver into conversation and was based on self-disclosures. Self-disclosure is elicited by reciprocity; the system will disclose something about itself and then ask the driver a question about the same (or similar) situation such as ‘I like driving on mountain roads, what's your favorite road to drive on?’. Engaging drivers in conversation can be useful for reasons such as, detecting the driver’s emotional state, gathering information on driver preferences to personalize the in-car information system, and as a potential aid for drowsy drivers.

Speech from the participants was recorded using an Andrea directional bean with 4 microphones placed in front and about 1.5 meters away from the driver. This microphone is typical of those used in the cars of today and provided a clean acoustic recording without overly sampling the car noise. The driving sessions were also videotaped from the front left of the driver to show driver hands, arms, upper body, head and eye motion, and facial expressions. Drivers self-reported emotions via questionnaires collected in the experiment. The questionnaires were based on DES (Differential Emotional Scale) [15], a scale based on the theory and existence of ten basic emotions.

The participants exhibit a range of emotions including boredom, sadness, anger, happiness and surprise; however for most of the drive the participants have a neutral/natural. In the absence of a distinct neutral/natural emotional state in the emotion recognition system we record average driver emotion as somewhere in between bored and sad. When challenged in the drive by obstacles in the road, other drivers, difficult road conditions and pedestrians, we observe strong emotions (aroused emotional states) from the drivers.

The performance of the automatic emotion recognition system was assessed by comparing the human emotion transcript against the output from the automatic acoustic emotion recognition system using a 2 second window analysis. The human transcripts were created by trained experts in recognising affective cues in speech. These experts did not take part in the driver study and are familiar with the emotion classifications of boredom, sadness/grief, frustration/extreme anger, happiness and surprise used by the automatic emotion recognition system. The experts were asked to listen to the speech soundtrack for each drive and report on the perceived emotional state of the driver.

The human evaluators used to transcribe the soundtracks for the rolling visual analysis and to report on the 2 second window analysis are employees of Affective Media Limited, UK. They are researchers in the product development team and have at least 2 years experience working in acoustic affective computing. They complete listening studies blind and are given unlabeled speech soundtracks. They have no knowledge of the driving experiment or the emotion recognition output. Although listener studies are subjective and the potentially the emotional transcripts could vary from one listener to another, we report that there is commonality in the classifications across human subjects for the five emotional groups of boredom, happiness, surprise, anger/frustration and sadness/grief.

5 Experimental Results

The 2 second window analysis is a detailed comparison between the automatically generated output of the emotion recognition system and the emotional transcript created by the human listener. The human listener considers in sequence each two second window of recording from the drive. For each two second window they report on the emotional state of the driver speech in terms of boredom, happiness, surprise, anger/frustration, sadness/grief (when the driver speaks); presence of the in-car voice, background car noise including engine noise, brake screech, doors and crashes; windows of no sound and other information such as other voices, non-speech sounds, Table 1.

In parallel to the human study, the emotion recognition system is presented with the driver soundtrack. The emotion recognition system is configurable and can classify any size of sound window. For this comparison the ER system will output an emotional classification for every 2 second window to allow for direct correlation

Table 1. Example human listener transcription (for participant 19) showing start and end time in seconds for the 2 second analysis speech segment; listener decision (1) on the emotional content of the 2 second window using acoustic emotion cues (not emotional cues from words spoken); whether the speech is that of the driver (blank) or the in-car conversational system (1); in-car noise such as tyre screech, indicator ticks, wheel knocks and engine noise; no sound present (1); and additional information such as what the driver says

start	end	bored	happy	surprise	anger	sadness	car voice	car noise	no sound	comments
…	…									
476	478							knocks		
478	480						1			what types
480	482						1			
482	484								1	
484	486	1				1				
486	488	1				1				neutral
488	490								1	uh
490	492	1								erm
492	494	1			1					old people
494	496	1				1				
496	498	1				1				
498	500	1				1				
500	502	1				1		screech		
502	504								1	
504	506	1				1				
506	508								1	
508	510	1				1				
510	512	1				1				

with the human transcript. The ER system outputs a continuous measure between 0 and 1 for each emotion including boredom, happiness, surprise, anger/frustration and sadness/grief, Table 2. Note multiple emotions can be present within a 2 second window and therefore it is possible to have high emotion levels for happiness and surprise, surprise and anger, boredom and sadness etc.

The ER system classifies all acoustic sounds as emotions. This can be useful for non-speech sounds such as sighs, non-words and hums/whistles [13]. However the ER system will also classify brake screech, crashes and the in-car voice. In the 2 second window study we assume that the ER system processes speech only and we use only those windows which are not corrupt by high levels of non-speech.

Each two second window from the human listener transcript is compared directly with the ER output by an additional human subject. This subject then rates the correlation between the ER output and the listener using the following criteria: ‘exact’ = both ER and human report same emotion; ‘equivalent’ = ER and human report equivalent emotions such as happy and surprise, boredom and sadness; ‘mismatch’ =

Table 2. Example acoustic emotion recognition output (for participant 19) showing start and end time in seconds for the 2 second analysis speech segment; automatic decision on the emotional content of the 2 second window (note '---' indicates insufficient speech for classification or no speech present) and additional information such as what the driver says (not recognised by the system but presented here to assist correlation with the human listener transcription)

start	end	bored	happy	surprise	anger	sadness	comments
…	…						
476	478	---	---	---	---	---	0
478	480	---	---	---	---	---	what types
480	482	---	---	---	---	---	0
482	484	---	---	---	---	---	0
484	486	0	0	0	0	1	0
486	488	1	0	0	0	0	neutral
488	490	---	---	---	---	---	uh
490	492	1	0	0	0	0.349	erm
492	494	0	0	0	0	1	old people
494	496	1	0	0	0	0.217	0
496	498	0	0	0	0	1	0
498	500	1	0	0	0	0	0
500	502	---	---	---	---	---	0
502	504	---	---	---	---	---	0
504	506	1	1	0	0	0	0
506	508	---	---	---	---	---	0
508	510	1	0	0	0	0	0
510	512	1	0	0	0	0	0

the ER and human differ in their judgement of emotion; 'no emotion correct' = ER returns a no-emotion present output ('0' for all emotions) which is confirmed by human transcript; 'no emotion mismatch' = ER returns a no-emotion present output ('0' for all emotions) which is in disagreement with human transcript; 'no classification' = ER does not detect speech in the window ('---' for all emotions) whereas human listener has report speech (this can be the case when there is little speech or speech is very quiet), Table 3.

The totals for each criteria for every 2 second window of the driver soundtrack are calculated and form the performance accuracy of the ER system. This provides a quantitative analysis of the correlation between the classified emotions of the emotion recognition system and that of the human listener. However it should be noted that any mismatches between the ER system and human may not indicate an error with the ER output but rather a misclassification by the human listener.

Table 3. Example comparison between the human listener transcription and the automatic emotion recognition output (for participant 19) for each 2 second window. 'Exact' = both ER and human report same emotion; 'Equivalent' = ER and human report similar emotions; 'mismatch' = the ER and human disagree; 'no emotion correct' = both ER and human report no emotion; 'no emotion mismatch' = ER and human disagree on whether emotion is present; 'no classification' = ER and human disagree on whether speech is present

		Emotion Recognised			No Emotion Recognised			
start	End	Exact	Equiv	Mismatch	Correct	Mismatch	No Class	comments
…	…							
476	478							0
478	480							what types
480	482							0
482	484							0
484	486	1						0
486	488	1						neutral
488	490							uh
490	492	1						erm
492	494			1				old people
494	496	1						0
496	498	1						0
498	500	1						0
500	502							0
502	504							0
504	506		1					0
506	508							0
508	510	1						0
510	512	1						0

6 Discussion

The project is ongoing and we continue to process all 41 participants of the in-car emotional conversation experiment to provide an overall performance score for the emotion recognition system. We are able at this stage to present an example performance measure for one male driver and one female driver. Early indications suggest that the emotion recognition performance with these two drivers is representative of the overall performance for all 41 participants. The results are obtained from the 2 second window analysis.

Performance Evaluation (for female participant 19). The human listener report 81 frames of driver speech from the 696 two second windows in the soundtrack. Of these 44 are 'exact' = both ER and human report same emotion; 13 are 'equivalent' = ER and human report equivalent emotions such as happy and surprise, or boredom and sadness; 12 are ‘mismatch’ = the ER and human differ in their judgement of emotion; 0 are ‘no emotion correct’ = ER returns a no-emotion present output which is confirmed by human transcript; 0 are ‘no emotion mismatch’ = ER returns a no-emotion present output which is in disagreement with human transcript; 12 are ‘no classification’ = ER does not detect speech in window whereas human listener has report speech (this can be the case when there is little speech or speech is very quiet).

83% two second windows were recognised by the ER system as the appropriately matching emotion to the human listener. 64% are exact matches with the transcript from the human listener and another 19% are equivalent recognitions eg happy for surprise, boredom for sadness etc. The error rate is 17%, however note that this is a comparison between the ER system output and the opinion of the human listener. The human listener may on occasion be wrong with the emotional classification.

Out of a total of 81 two second windows in which human listener has reported driver speech, the ER system detects 85% as speech windows, missing 15% of speech windows. Of the 69 windows detected by the ER system, 64% correspond exactly with the human listener emotion recognition. A further 19% are equivalent recognitions between human listener and ER system eg happy for surprise and boredom for sadness. The error rate is 17%, however note again that this is a comparison between the ER system output and the opinion of the human listener.

Performance Evaluation (for male participant 19). The human listener report 97 frames of driver speech from the 698 two second windows in the soundtrack. Of these 56 are ‘exact’ = both ER and human report same emotion; 10 are ‘equivalent’ = ER and human report equivalent emotions such as happy and surprise, boredom and sadness; 14 are ‘mismatch’ = the ER and human differ in their judgement of emotion; 0 are ‘no emotion correct’ = ER returns a no-emotion present output which is confirmed by human transcript; 4 is ‘no emotion mismatch’ = ER returns a no-emotion present output which is in disagreement with human transcript; 13 are ‘no classification’ = ER does not detect speech in window whereas human listener has report speech (this can be the case when there is little speech or speech is very quiet).

83% two second windows recognised by the ER system as the appropriately matching emotion to the human listener. 70% are exact matches with the transcript from the human listener and another 13% are equivalent recognitions eg happy for surprise, boredom for sadness etc. The error rate is 17%, however note that this is a comparison between the ER system output and the opinion of the human listener.

Out of a total of 97 two second windows in which the human listener has reported driver speech, the ER system detects 87% as speech windows, missing 13% of speech windows. Of the 84 windows detected by the ER system, 67% correspond exactly with the human listener emotion recognition. A further 12% are equivalent recognitions between human listener and ER system eg happy for surprise and boredom for sadness. The error rate is 21%, however note that this is a comparison between the ER system output and the opinion of the human listener.

There is a strong correlation between the emotional transcript created by the human listener and the emotion output returned automatically by the acoustic emotion recognition system. However there are occasions where the speech is masked by car noise (such as engine noise, sirens and brakes). Other times, the automatic system could not disambiguate between emotional states so that the driver was assessed to be in one of two emotional states - bored or sad (negative emotions with low arousal), or - happy or surprised (positive emotions with moderate arousal).

7 Conclusions

In-car speech recognition systems are becoming commonplace, enabling drivers to navigate using GPS, speed-dial on mobile communications and control audio and video. There is the opportunity for richer driver-car interaction if the car can recognise and respond to driver emotion. Our earlier studies have shown that driver emotion and driving performance can be correlated and matching the emotional tones of the car voice with the emotion of the driver can increase driver safety. Here we report a quantitative performance analysis of our acoustic emotion recognition system. The emotion recognition system can classify into 5 emotional groups of boredom, happiness, surprise, frustration/anger, sadness/grief with an average accuracy of 65% when comparing the automated classification with that of human listeners. An average of 15% of additional recognitions are equivalent, for example happiness is confused with surprise, or boredom with sadness, and the emotional response by the car remains appropriate. Less than 20% of driver speech is incorrectly recognised for emotion state. The performance is sufficient to allow the in-car system to respond to driver emotion with empathy in order to support and improve driving.

8 Future Work

The quantitative performance analysis is ongoing however a number of other research challenges remain. Some drivers did not converse with the car. We need to consider why drivers do not talk to the car. Are they too engaged in the task of driving? Are the questions posed by the car inappropriate? Are they uncomfortable talking to the car? Do they not like the car voice? Using automatic emotion recognition we hope that the car can detect driver emotion and adapt voice, conversation and vehicle parameters to support the driving experience. However there are additional questions to answer. Previous studies consider varying the paralinguistic cues only [16], however should the content of the response also change, and how? Should the car become less or more talkative depending on driver emotion? Should the car alter the telemetry, climate, music in the car in response the mood of the driver? How fast should the system change? Further research will consider automatically adapting the emotion of the car-voice to match the user. Empathy communicates support, caring, and concern for the welfare of another [17]. A voice which expresses happiness in situations where the user is happy and sounds subdued or sad in situations where the user is upset would strongly increase the connection between the user and the voice [4]. Mood must be taken into account to make the car-voice an effective interaction partner. Drivers in a good mood when entering a car are more likely to experience positive emotion during an interaction with a car-voice than drivers in a bad mood. Therefore it seems that emotion in technology-based voices must balance responsiveness and inertia by orienting to both emotion and mood. The research continues towards developing an intelligent conversational in-car system which can recognise and respond to driver emotion to support the driving experience.

Acknowledgments. We wish to thank Affective Media, UK for assistance in developing the ER technologies and help with the analysis of the driver speech data.

References

1. Lunenfeld, H.: Human factor considerations of motorist navigation and information systems. In: Vehicle Navigation and Information Systems Conference Proceedings, pp. 35–42 (1989)
2. Galovski, T., Blanchard, E.: Road rage: a domain for psychological intervention? Aggressive Violent Behavior 9(2), 105–127 (2004)
3. Groeger, J.A.: Understanding driving: Applying cognitive psychology to a complex everyday task. U.K. Psychology Press, Hove (2000)
4. Nass, C., Jonsson, I.-M., Harris, H., Reaves, B., Endo, J., Brave, S., Takayama, L.: Increasing safety in cars by matching driver emotion and car voice emotion. In: Proceedings of CHI 2005, Portland, Oregon, USA (2005)
5. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G.: Emotion recognition in human-computer interaction. IEEE Signal Proc, pp. 32–80 (2001)
6. Humaine Portal.: Research on affective computing http://www.emotion-research.net/
7. Jones, C.: Project to develop voice-driven emotive technologies, Scottish Executive, Enterprise transport and lifelong learning department, UK (2004)
8. Kapoor, A., Qi, Y., Picard, R.: Fully automatic upper facial action recognition, IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003) held in conjunction with ICCV 2003. Nice, France (2003)
9. Khan, M., Ward, R., Ingleby, M.: Distinguishing facial expressions by thermal imaging using facial thermal feature points. In: Proceedings of HCI 2005, pp. 5–9. Edinburgh, UK (2005)
10. Healey, J., Picard, R.: SmartCar: Detecting driver stress. In: Proceedings of ICPR 2000, Barcelona, Spain (2000)
11. Nass, C., Brave, S.: Wired for speech: How voice activates and advances the human-computer relationship. MIT Press, Cambridge, MA (2005)
12. Jones, C., Jonsson, I.-M.: Speech patterns for older adults while driving. In: Proceedings of HCI International 2005, Las Vegas, Nevada, USA (2005)
13. Jones, C., Jonsson, I.-M.: Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In: Proceedings of OZCHI, Canberra, Australia (2005)
14. STISIM drive system: Systems technology, Inc. California http://www.systemstech.com/
15. Izard, C.: Human Emotions. Plenum Press, New York (1977)
16. Isen, A.M.: Positive affect and decision making. In: Lewis, M., Haviland-Jones, J.M. (eds.) Handbook of emotions, pp. 417–435. The Guilford Press, New York (2000)
17. Brave, S.: Agents that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent, Doctoral dissertation. Stanford University, CA (2003)