Assessment of Speech Dialog Systems using Multi-Modal Cognitive Load Analysis and Driving Performance Metrics

Tristan Kleinschmidt*, Pinar Boyraz†, Hynek Bořil†, Sridha Sridharan*, and John H.L. Hansen†
*Speech & Audio Research Laboratory, School of Engineering Systems
Queensland University of Technology, GPO Box 2434, Brisbane, QLD, 4001, Australia
Email: {t.kleinschmidt, s.sridharan}@qut.edu.au
†Center for Robust Speech Systems, Erik Jonsson School of Engineering & Computer Science
The University of Texas at Dallas, USA
Email: {pinar.boyraz, hynek, john.hansen}@utdallas.edu

***Abstract*—In this paper, cognitive load analysis via acoustic- and CAN-Bus-based driver performance metrics is employed to assess two different commercial speech dialog systems (SDS) during in-vehicle use. Several metrics are proposed to measure increases in stress, distraction and cognitive load and we compare these measures with statistical analysis of the speech recognition component of each SDS. It is found that care must be taken when designing an SDS as it may increase cognitive load which can be observed through increased speech response delay (SRD), changes in speech production due to negative emotion towards the SDS, and decreased driving performance on lateral control tasks. From this study, guidelines are presented for designing systems which are to be used in vehicular environments.**

I. Introduction

The last 20 years have witnessed a dramatic increase in the number of human-machine interface applications for in-vehicle driver assistance, navigation and infotainment systems in order to meet the demands of modern life. Although these exciting developments bear the potential to help drivers travel more safely and efficiently, an often overlooked factor is the effect of such systems on drivers in terms of distraction and cognitive load. It has been reported that almost 90% of all accidents are caused by driver error [1]. Although not reported in full detail, factors which contribute to these errors are related to sleep deprivation, fatigue, distraction and inattention. If driver errors are to be reduced by the use of engineering systems and technology, each of these factors requires detailed study.

This study focuses on driver distraction and cognitive load; in particular, speech dialog systems (SDS) are studied to assess their effect on driver workload. The level of cognitive load is assessed using both speech and driving performance metrics. To indicate the delay in cognitive/auditory processing, we use speech response delay (SRD) [2], whilst acoustic analysis is used to measure changes in speech production which may result from stress when interacting with SDS. The use of these systems might also cause changes in driving performance; therefore driver-vehicle interaction signals – particularly steering wheel angle (SWA) and speed from the CAN-Bus – are used to assess the smoothness of lateral/longitudinal control applied by the driver during interaction with the SDS. Driving performance metrics are based on standard deviation (STD), sample entropy (SampEnt) and high-frequency signal energy.

The authors at Queensland University of Technology would like to acknowledge the contribution of the Co-operative Research Centre for Advanced Automotive Technology (AutoCRC) in supporting parts of this research.

TABLE I
SECONDARY TASK DEFINITIONS.

Part		Secondary Tasks		
		A	B	C
Route 1	1	Lane Changing	Common Tasks (Radio, AC etc.)	Sign Reading
	2	Cell Phone Dialog	Cell Phone Dialog	Conversation
	3	Common Tasks	Sign Reading	Spontaneous Speech
	4	Conversation	Spontaneous Speech	Cell Phone Dialog
Route 2	1	Sign Reading	Lane Changing	Common Tasks (Radio, AC etc.)
	2	Cell Phone Dialog	Cell Phone Dialog	Conversation
	3	Common Tasks (Radio, AC etc.)	Sign Reading	Lane changing
	4	Spontaneous Speech	Conversation	Sign Reading

In this paper, the experiment design and transcription protocols are first explained in Section II with details of the instrumented vehicle UTDrive, the driving routes, and characteristics of the speech dialog systems. In Section III, the metrics used in cognitive load assessment are presented with discussion of their reliability. This section also reports a comparison between the speech- and driving-based performance metrics on a subset of the UTDrive Corpus. Finally, in Section IV, guideline recommendations and conclusions based on cognitive load management strategies and their implications on SDS design are made based on the results of this study.

II. Experiment Design and Transcription Protocol

This study is based on the close-to-realistic data collection procedure used in the UTDrive project [3]. The data collection was performed using participants driving an instrumented vehicle in real traffic conditions. Although the instrumentation is visible to the subject which might shift behavior towards "precautious, self-aware driving", the risks in traffic and associated consequences are real. Details of the corpus relevant to the current study are given in the following sections.

TABLE II
SESSION-BY-SESSION EXPERIMENT DESIGN.

Session 1	Route 1	Neutral driving
	Route 1	Secondary Tasks A
	Route 2	Secondary Tasks A
	Route 2	Neutral driving
Session 2	Route 1	Secondary Tasks B
	Route 1	Neutral driving
	Route 2	Neutral driving
	Route 2	Secondary Tasks B
Session 3	Route 2	Secondary Tasks C
	Route 1	Secondary Tasks C
	Route 2	Neutral driving
	Route 1	Neutral driving

A. Instrumented Vehicle UTDrive and Experiment Design

A Toyota RAV4 was equipped with several sensors and a data acquisition unit to record 13 data streams. Sensors used in this study include:

- Two CCD cameras for monitoring driver and road scene;
- Microphone array (5 mics) to record driver's speech as well as noise conditions in the vehicle; and
- CAN-Bus OBD II port for collecting vehicle dynamics: speed, steering wheel angle, gas and brake driver inputs.

Data was collected in both residential and commercial traffic areas including a mixture of collector-arterial roads with up to 4 lanes. Both routes were driven while performing secondary tasks (shown in Table I), and without secondary tasks referred to as neutral driving. Table I also shows the tasks assigned to each road segment of each route. Each driver completed 3 ordered sessions as shown in Table II, resulting in 12 laps per driver. Each session was separated by at least one week.

B. Speech Dialog Systems

In the cell-phone segments in Table I, participants called stored telephone numbers to initiate spoken interaction with one of two commercial SDS whilst driving. This task involved both manual dialing and then responding to system prompts. A Bluetooth interface to the cell phone enabled hands-free interaction once the number was dialed.

Both SDS provide informational retrieval services, however Speech Dialog System A (SDSA) provides information on a much narrower scope than system B (SDSB). A comparison of the general characteristics of the two dialog systems assessed in this study is provided in Table III.

Characteristics of dialog interaction between humans and SDS which are important for this study include [4]:

- *Automatic Speech Recognition (ASR):* needs to be robust against the adverse effects of noise and stressed speech. A poorly performing system can increase user frustration which could also lead to decreased driving performance. Neither of the tested SDS was designed specifically for vehicular applications; therefore their recognition performance will be lower than in their target environments.
- *Grounding:* the process of confirming with the user what has been understood from previous communication(s). Grounding is required in order to detect recognition and other errors during the exchange prior to executing an action (in these cases information retrieval).
- *Barge-in:* the phenomenon where users respond to the SDS before the system prompt completes which is advantageous for experienced users to reduce transaction times. When the communication channel is duplex (as in hands-free environments), barge-in support requires echo cancelation to remove the effects of system prompts. If the SDS prompt cuts out whilst the user is speaking, the system must also deal with a user's tendency to stutter and repeat responses.

The effect of some of these characteristics are studied in Section III.

C. Multi-layer and Multi-modal Transcription Protocol

To assist this study, a multi-layered transcription protocol was developed to provide information relating to speech activity, driving maneuvers and other driving related tasks for the assessment of driver performance. For audio-based transcription, recordings were time labeled to indicate speech dialog prompts, silence and driver responses. A sub-layer of transcription for SRD analysis included the number of responses to each prompt, driver's emotion *towards* the SDS, as well as instances of barge-in, hesitancy or passenger assistance. Driver emotion was perceptually classified by a human transcriber based on the characteristics described in Table IV. Negative emotions were labelled based on any variation from what is considered as desirable for "natural" communication with a SDS – i.e. speaking in a neutral tone in order to maximise ASR performance. All other responses were regarded as non-negative in order to include speech with no perceived emotion. In addition to the audio-based transcription, video and CAN-Bus signals were used jointly to label driving maneuvers such as lane-keeping on straight or curved road segments. The two sets of transcriptions are used to assess the auditory and driving performance in well-defined scenarios in the following section.

III. COGNITIVE LOAD ANALYSIS

Cognitive load analysis in this study concerns the assessment of human-machine interfaces in terms of perception and processing burden placed on human subjects. The SDS might induce unacceptable cognitive and auditory load for the drivers or it may cause driver frustration if it is not designed specifically for in-vehicle interaction. Excessive cognitive load and frustration can both be considered as contributors to unsafe driving behavior. By assessing the systems from a cognitive load perspective, we aimed to obtain more in-depth information on driver workload/distraction and produce guidelines to design more effective in-car SDS. These two facets of the problem are reflected in terms of the metric definitions in the following sections. Auditory performance analysis is presented in Section III-A, followed by driving performance assessment using metrics indicating lateral control performance.

A. Analysis of Auditory Performance Metrics

The empirical distribution and means of the speech response delay (i.e. the time between the system prompt ending and the driver response starting) of the two SDS are shown in Fig. 1. It should be noted that barge-in responses were not included in this result as the SDS prompts typically cut out as soon as the driver speaks; therefore it was difficult to establish the true barge-in time as the prompt fails to complete. It can be seen that the distribution and mean of SDSB differs from those of the global population and SDSA. Using a two-sample Kolmogorov-Smirnov tests, Table V shows the

TABLE III
GENERAL ATTRIBUTES OF THE TWO COMMERCIAL SPEECH DIALOG SYSTEMS.

SDSA	**SDSB**
Information retrieval services on restricted domain.	General information retrieval service on several domains.
Deferred explicit grounding performed after several responses.	No grounding.
Error recovery handled by changing specific piece of information.	No error recovery.
Barge-in supported.	Barge-in supported, but not consistent.
System-directed dialogue.	System-directed dialogue.
Prompts have simple but varied linguistic structure.	Prompts follow same linguistic structure.
Robust against moving vehicle noise.	Overly sensitive to moving vehicle noise leading to 'false acceptances'.
Professionally recorded speech prompts.	Professionally recorded speech prompts.
Initial system prompt = 16.5sec.	Initial system prompt = 22.5sec.
Critical initial prompt length = 9sec.	Critical initial prompt length = 17.5sec.

TABLE IV
CLASSIFICATION OF DRIVER EMOTION TOWARDS SDS.

Negative	**Non-Negative**
Hesitancy / confusion	Neutral
Frustration / anger	Confident
Increased vocal effort	Happy
Decreased speaking rate	Humored
Altered pitch	Disinterested

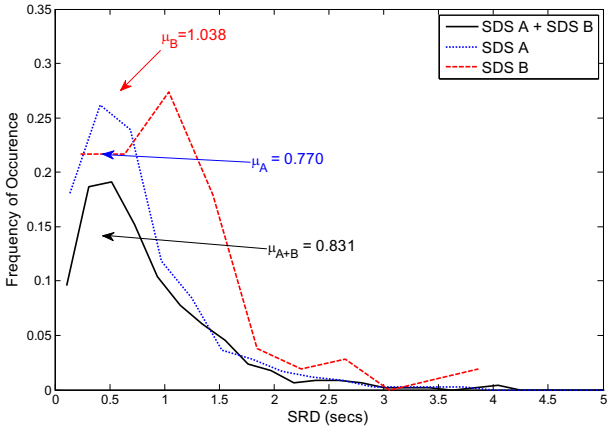

Fig. 1. Empirical distributions and means of SRD for two dialog systems.

distribution for SDSB was significantly different to the global and SDSA distributions (when $p < 0.05$). ANOVA test on the sample means also confirmed a difference at the 5% significance level ($p = 0.0007$). The observed distributions of SRD could be affected by the number of response repetitions or driver emotions towards the SDS, however further analysis showed neither of these factors produce empirical distributions significantly different from the global population.

Despite the fact that repeated responses and driver emotion towards the SDS did not affect the response delay, during the data collection it was noticeable that the ASR performance of the two systems was different. Since the number of repeated responses is a direct reflection of ASR performance and negative driver emotion towards the SDS includes frustration which could result from low ASR accuracy, these factors were assessed to compare the two systems. General observation of Figs. 2 and 3 shows a difference between the proportions of each factor. Statistical tests were then performed on each factor to test the hypothesis that the proportions were equal for both systems (i.e. $p_{SDSA} = p_{SDSB}$) at the 5% significance level (i.e. $z = 1.96$). Table VI summarizes these results.

TABLE V
TWO-SAMPLE KOLMOGOROV-SMIRNOV TESTS FOR SRD DISTRIBUTIONS.

Distribution 1	**Distribution 2**	***p*-score**
Global	SDSA	0.3933
Global	SDSB	7.55e-4
SDSA	SDSB	6.75e-6

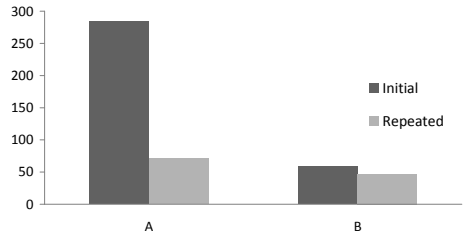

Fig. 2. Comparison of number of repeated utterances vs. initial responses.

Figure 2 shows the counts of initial and repeated responses for both SDS however it is the proportions of initial to repeated responses which are of interest. It can be seen that SDSA required fewer repeated responses (compared to total number of responses) than SDSB which had a ratio of initial to repeated responses of almost 1:1. This observation is supported by the first result in Table VI. The second and third rows confirm that SDSB led to a greater proportion of negative responses than SDSA as shown in Fig. 3.

Table VI also shows that System B exhibited a significantly larger proportion of barge-in responses than System A. This is attributed to two factors:

1) The long initial prompt of SDSB (22.5 sec) causes drivers to respond as soon as they hear the option they want. In this case it is vital for dialog systems to actively listen for responses whilst long prompts are delivered.
2) From listening to the responses it was observed that as drivers became increasingly frustrated, they tended to respond as soon as the system informed them of a misunderstanding.

Having established that SDSB induced more negative responses and that it required drivers to repeat themselves more regularly, it was necessary to confirm that negative emotions were more common as more repetitions were needed. Figure 4 shows a general increasing trend of proportions of negative utterances with respect to repeat number (where 0 = initial response). This trend is supported by correlation analysis which showed some form of linear relationship between these variables (correlation coefficient $r = 0.3432$ found to be

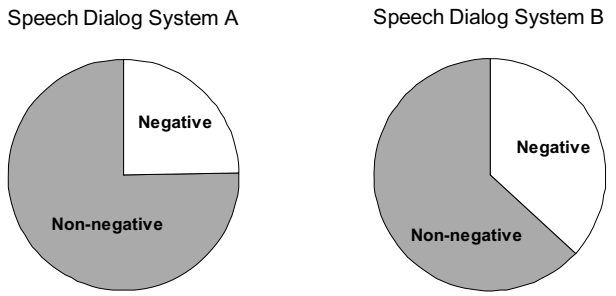

Fig. 3. Comparison of negative and non-negative response proportions.

TABLE VI
SUMMARY OF STATISTICAL TESTS OF PROPORTIONS.

Description	z-score	Result
Repeated responses	9.31	$p_{repSDSA} < p_{repSDSB}$
Driver's emotions towards SDS (exc. hesitancy)	2.424	$p_{negSDSA} < p_{negSDSB}$
Driver's emotions towards SDS (inc. hesitancy)	2.052	$p_{negSDSA} < p_{negSDSB}$
Barge-in responses	15.33	$p_{bargeSDSA} < p_{bargeSDSB}$

significant at the 95% confidence level). Despite these findings, it cannot be stated that a continual need for repeats *causes* drivers to be more negative towards the SDS, as it is also possible that negative emotions (i.e. stressed speech) *cause* the ASR performance of the SDS to reduce.

The following paragraphs analyze how speech production parameters are affected by the emotional state of the driver towards the SDS and the number of repetitions to SDS prompts. Using the same data as the above statistical analysis, we extracted fundamental frequency (F_0), short-term energy, formant center frequencies, and average duration of voiced speech segments. Figure 5 shows average fundamental frequencies for negative and non-negative emotions as well as F_0 as a function of the repetition number. It can be seen that negative emotions result in higher mean and variance of F_0 distribution than non-negative emotions. In addition, there is an increasing trend in F_0 as the repetition number increases, which reflects the greater proportions of negative responses as shown in Fig. 4. The observed increase in fundamental frequency for negative responses is consistent with previous studies, where F_0 increased when switching from neutral to stressed speech production [5].

The short-term energy analysis showed a 2.6% increase comparing negative to non-negative responses; further, 3^{rd} responses increased by 1.6% compared to initial responses and responses after the 3^{rd} increased by 1.8%. The joint increase in F_0 and short-term energy appears to confirm observations made in [6], [7] regarding the monotonic relationship between these two measures.

Formant frequencies were analyzed for voiced speech segments. Table VII presents the average formant frequencies in negative and non-negative responses. The results demonstrate noticeable differences between the two sets of emotions which are particularly pronounced for F_1, F_3 and F_4 taking into account both mean and standard deviation. Shift in formant frequencies between modalities of speech were observed previously [5]; these results further validate a change in speech production as driver's become frustrated with the SDS.

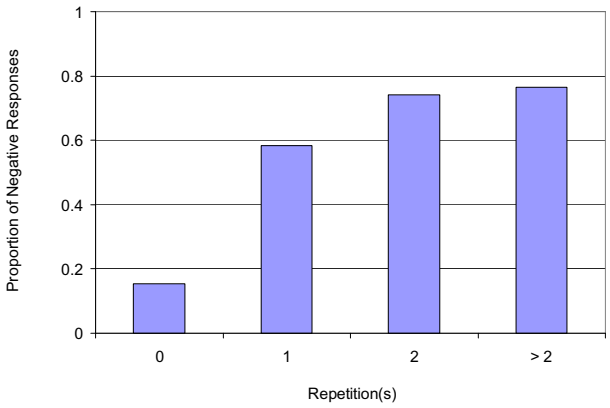

Fig. 4. Comparison of negative response proportions as repetitions increase.

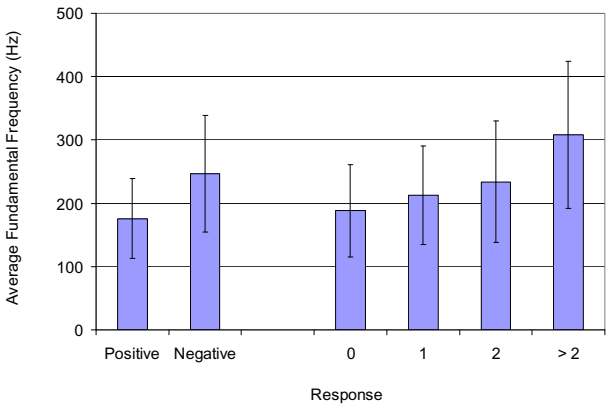

Fig. 5. Average fundamental frequency in repeated inputs and in positive and negative inputs. Vertical lines represent standard deviation intervals.

Finally, average duration of voiced speech segments was analyzed (see Fig. 6). Again, the durations were increased between negative and non-negative responses, and also showed an expanding trend as drivers were required to respond repeatedly to the same prompt. Increased duration was a characteristic previously observed for Lombard effect where speakers adjust their speech production in order to maintain intelligible communication over noise [8]. From these results, it appears that speakers believe their speech will become more intelligible (and therefore more easily recognized by the SDS) if they decrease their speaking rate.

The analysis in this section has demonstrated differences between the two assessed SDS. These differences could be attributed to the speech recognition accuracy of the two systems which is reflected in the number of repeated utterances required to complete a dialog. Low ASR performance can lead users to become negative towards the system which could further hinder the overall performance. The increased proportion of negative attitudes could be correlated to the observed lengthened speech response delay; further research is required to verify this statement.

The statistical evaluation was confirmed through analysis of speech production features which showed noticeable changes between negative and non-negative emotions towards the SDS, as well as elevated changes as drivers were required to repeat their responses to the same prompt. These speech production

TABLE VII
AVERAGE FORMANT FREQUENCIES EXTRACTED FROM VOICED SPEECH SEGMENTS. NUMBERS IN PARENTHESES REPRESENT STANDARD DEVIATION.

Response	T_{Total} (s)	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)	F_4 (Hz)
Positive	109.3	565 (126)	1545 (259)	2667 (314)	3776 (287)
Negative	58.2	632 (130)	1599 (264)	2840 (311)	3905 (241)

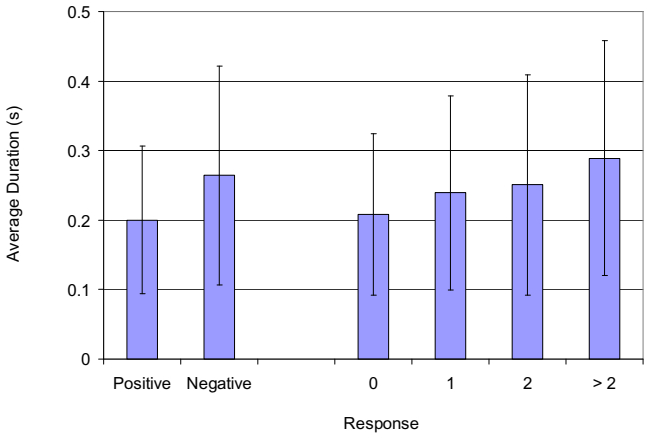

Fig. 6. Average duration of voiced speech segments. Vertical lines represent standard deviation intervals.

measures also validate the choice of the labelling procedure which was adopted in Section II-C to annotate the data.

B. Analysis of CAN-Bus-based Driver Performance Metrics

1) Metric Definitions: Segment 2 of both routes includes interaction with the SDS as a secondary task (see Table I). This segment comprises lane keeping and curve negotiation driving tasks. Therefore we use maneuver-based driver performance metrics which have proven to be good indicators of distraction. For lane keeping, several driver performance metrics have been suggested in the literature, mostly using steering wheel angle to calculate a metric indicating the fluctuations or micro-corrections in SWA. Amongst these metrics, a widely accepted one is the sample entropy [9] and standard deviation. Reversal rate of the steering wheel is also considered a reliable metric to measure driver performance in the lane keeping task. Boer [10] recently updated their previous work and suggested taking high frequency terms into account. In a thorough analysis in [11], it was also noted that the speed interval for which the SWA-dependent metric is calculated is important since lower speeds require more SWA input to achieve the same amount of lateral movement when compared to higher speeds. For curve negotiation, a constant input of some SWA is required using the visual input of the road curvature as reference. A novice or distracted driver will likely have fluctuating SWA inputs, and a general trend is for speed to be reduced while taking curves to balance the centrifugal force.

Although different in nature, lane keeping and curve negotiation can both be seen as regulatory control tasks from the driver's point of view. Therefore, we selected seven metrics using available information and observations about driver performance/behavior including energies of high frequency components Wavelet Decomposition (WD), sample entropy, standard deviation and standard deviation of rate of change (RSTD) – these are summarised in Table VIII. All features are extracted for SWA and speed channels except RSTD which only applies to SWA. The time window length is taken as equal to the maneuver length and the effect of signal length is eliminated in calculating the features.

TABLE VIII
CAN-BUS-BASED FEATURE VECTOR DEFINITION.

Notation	Definition
WDE_SWA	WD Detail Signal Energy for SWA
WDE_Speed	WD Detail Signal Energy for Speed
SampEnt_SWA	Sample Entropy of SWA
SampEnt_Speed	Sample Entropy of Speed
STD_SWA	Standard deviation of SWA
STD_Speed	Standard deviation of Speed
RSTD_SWA	Standard deviation of rate of change of SWA

For WD, Daubechies [12] wavelet kernel with 4^{th}-order is used and detail signal is taken at the 6^{th}-level. Daubechies wavelet is chosen over alternatives since it approximates well the spikes and discontinuities regularly seen in CAN-Bus signals. The level and order is adjusted to be able to extract the high frequency content in the signal within the limitation of human control; higher details are ignored since they might be caused by other measurement disturbances.

Wavelet decomposition is performed using the integral:

$$[W_\psi f](a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \psi(\frac{x-b}{a}) f(x) dx. \quad (1)$$

The wavelet function (ψ) is used to calculate the decomposition coefficient signals at all detail levels and a scaling function is used to calculate the approximation signal. Daubechies Wavelet at 4^{th}-order (DB4) wavelet can be expressed as a number sequence:

$$b_k = (-1)^k a_{N-1-k}, \quad (2)$$

where b is the wavelet number sequence:

$$b = \{0.1830127, -0.3169873, 1.1830127, -0.6830127\}, \quad (3)$$

and a are the scaling coefficients.

Sample entropy is used to quantify regularity and complexity of the signal and is a perfect match for measuring regularity of the SWA signal; it has long been used in bio-signals such as EEG, ECG and EMG to measure regularity. The SampEnt is calculated as per the work described in [13]. The standard deviation is calculated in statistical canonical form.

2) Analysis: The metrics explained in the previous section give an indication of irregularity in control which can be caused by increased cognitive load. Therefore, they are suitable to test the hypothesis that SDSB causes more frustration, negative emotions and possibly distraction compared to SDSA. In order to achieve this, all metrics are calculated over the maneuver. The same metrics are calculated for the same route segment, maneuver and driver using the CAN-Bus data from neutral driving to construct the baseline. A comparison rate is calculated using:

$$ComparsionRate = \frac{M_{distracted} - M_{baseline}}{M_{baseline}} \quad (4)$$

TABLE IX
INCREASE IN DRIVER PERFORMANCE COMPARED TO NEUTRAL DRIVING.

		SDSA	SDSB	Rel. Increase (SDSB vs. SDSA)
WDE	**SWA**	1.83	4.48	2.4
	Speed	0.36	0.59	1.6
SampEnt	**SWA**	0.11	2.60	23.6
	Speed	0.11	0.25	2.3
STD	**SWA**	0.25	0.50	2.0
	Speed	0.48	1.92	4.0
R-STD	**SWA**	1.73	0.35	0.2

so that the relative change for each metric (M) according to baseline driving can be measured independent of the driver. After obtaining comparison rates for 14 subjects (7 female, 7 male) over 100 maneuver segments, the average change caused by each system for each metric was calculated. The results shown in Table IX support the initial hypothesis that SDSB causes more distractions since most metrics show relative increases from SDSA to SDSB (from 1.6x to 23x). In other words, SDSB caused more irregularity compared to SDSA in steering angle and speed control on a lane keeping task.

From Table IX it can be observed that SDSA also caused increased in the irregularity of lateral control applied by the driver, since the numbers represent the comparison rate using the baseline as reference. However, SDSB has caused a larger irregularity which can be seen in all metrics except RSTD. The exception can be explained by the fact that SDSA caused quick but small corrections to SWA while SDSB engaged drivers' attention more, causing them to drift in the lane with larger errors – this is actually more dangerous driver behavior. If the driver allows the lateral control error to accumulate larger than a certain threshold it might occupy the adjacent lanes setting the conditions for an imminent accident.

IV. DISCUSSION AND CONCLUSIONS

In this study, speech- and driving-based performance metrics were used to assess 2 speech dialog systems. It was observed that dialog systems with sub-standard ASR performance caused driver emotion towards these systems to be proportionally more negative, a phenomenon which was elevated as they were required to repeat their responses. Speech response delay was shown to be greater in a system which causes greater negative reactions, and could therefore be an indicator of distraction related to these emotions.

Analysis of speech production measures such as fundamental and formant frequencies, and voiced speech durations showed trends which were consistent with previous research comparing these measures in neutral and stressed speaking styles. This acoustic analysis verified the data annotation process, and showed how speakers vary their speech production when exposed to misrecognition errors.

CAN-Bus-based driving performance metrics related to steering wheel angle and speed irregularities confirmed the hypothesis that a system which causes drivers to become more frustrated, also causes them to become more distracted while engaging in dialog with the SDS.

The analysis here clearly showed that only developing SDS with technical requirements is not enough. For SDS to be beneficial to the driver, it would be desirable for the SDS to use dynamically-estimated cognitive load level from CAN-Bus, speech response delays and acoustic-based metrics to adapt itself for a safer driving experience.

As there was no other data available to further analyze why Systems A and B perform differently, we can only hypothesize that certain characteristics of the SDS cause different human behaviors. Possible hypotheses which will be tested in future research on SDS for in-car systems are:

- Considerably long dialog prompts put extra strain on the driver to remember the options so they can respond appropriately at the end (i.e. they are unaware of support for barge-in). Any SDS where a list of options is provided rather than an either/or decision could induce greater cognitive load and also increase SRD.
- Drivers quickly learn when a system performs unsatisfactorily in hands-free mode, so they focus more on the road, rather than on dialog prompts. This could cause artificial delays to be included in the total response delay.
- Other features, such as background music, may confuse drivers as to when they are supposed to respond to the prompts, again producing an artificial delay.
- Systems which do not direct the dialog may cause drivers to concentrate more on the prompt, using considerable cognitive resources and increased driver distraction.

In our future studies we will be focusing on the effect of these systems on different age groups, and also native/non-native speakers of English.

REFERENCES

[1] J. R. Treat, N. S. Tumbus, S. T. McDonald, D. Shinar, R. D. Hume, R. E. Mayer, R. L. Stansifer, and N. J. Catellian, "Tri-level study of the causes of traffic accidents: Final report, vol. 1: Causal factor tabulations and assessments," Institute for Research in Public Safety, Indiana University, Tech. Rep., 1979.

[2] A. Ozaki, S. Hara, T. Kusakawa, C. Miyajima, T. Nishino, N. Kitaoka, K. Itou, and K. Takeda, "In-car speech data collection along with various multimodal signals," in *Proceedings of LREC*, 2008, pp. 1846–1851.

[3] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. H. L. Hansen, "UTDrive: Driver behavior and speech interactive systems for in-vehicle environments," in *IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, 2007, pp. 566–569.

[4] M. F. McTear, *Spoken Dialogue Technology*. London: Springer-Verlag, 2004.

[5] J. H. L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Comm.*, vol. 20, no. 1-2, pp. 151–173, 1996.

[6] P. Gramming, S. Sundberg, S. Ternstrm, and W. Perkins, "Relationship between changes in voice pitch and loudness," *STL-QPSR*, vol. 28, no. 1, pp. 39–55, 1987.

[7] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2936–2946, 1992.

[8] H. Bořil, "Robust speech recogniton: Analysis and equalization of Lombard effect in Czech corpora," Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, http://www.utdallas.edu/∼hxb076000, 2008.

[9] E. Boer, "Behavioral entropy as a measure of driving performance," in *Proceedings of 1st International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, August 14-17 2001.

[10] ——, "Steering entropy revisited," in *Proceedings of 3rd International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, Rockport, ME, USA, 2005.

[11] P. Boyraz, A. Sathyanarayana, and J. H. L. Hansen, "CAN-Bus signal modeling using stochastic methods and structural pattern recognition in time series for active safety," in *Proceedings of 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety*, June 25-27 2009.

[12] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pur. Appl. Math.*, vol. 41, pp. 909–996, 1988.

[13] H. B. Xie, W. X. He, and H. Liu, "Measuring time series regularity using non-linear similarity-based sample entropy," *Physics Letters A*, vol. 372, pp. 7140–7146, 2008.