UTDrive: Emotion and Cognitive Load Classification for In-Vehicle Scenarios

Hynek Bořil, Seyed Omid Sadjadi, John H. L. Hansen

Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, U.S.A.
E-mail: {hynek,sadjadi,john.hansen}@utdallas.edu

Abstract Emotions and non-driving related cognitive tasks affect a driver's control over a vehicle and may result in driving errors and traffic accidents. Presence of a monitoring device that would assess driver's state could help reduce such errors by providing the driver with alerts and directing other in-vehicle active safety devices. The focus of this study is on the evaluation of speech production-based and cepstral-based acoustic features for the task of emotion and cognitive load classification in real driving scenarios. The newly proposed classifiers utilize support vector machine (SVM) based fusion of raw features and Gaussian mixture model (GMM) scores and provide classification performance of 79 % and 95.2 % in the task of neutral vs. negative emotion classification and two cognitive tasks classification, respectively.

Keywords Driving scenarios, cognitive load, emotions, speech production variations, SVM fusion.

1. INTRODUCTION

Distraction due to increased, non-driving related cognitive load as well as emotions impact driver performance and are frequent causes of driving errors [1], [2]. Presence of an in-vehicle monitoring system that would assess driver's emotional state and cognitive load could considerably help reduce such errors by: (i) issuing alerts to the driver, (ii) directing other in-vehicle devices (e.g., decreasing frequency of navigation prompts, controlling loudness of the audio system, etc.), (iii) collaborating with other components of the active safety system. The focus of this study is on the assessment of drivers' emotional state and cognitive load from speech.

Emotion recognition has been receiving increased attention in the speech community [3, 4, 5, 6, 7]. The impact of stress in speech (including cognitive task stress) has also been widely studied [8, 9, 10]. In spite of these efforts, a limited body of literature has considered the impact of stress and emotions on drivers.

Studies on driver state assessment typically utilize driving simulators rather than real driving scenarios and prevalently focus on the analysis of physiological and EEG signals [11, 12], even though speech-based assessment has also been considered in several cases [13, 14, 15] in the context of cognitive tasks and emotional state classification. Clearly, driving simulators give researches good control over the induced scenarios without jeopardizing the driver's safety. On the other hand, it is not clear to what extent the simulator data reflect the real conditions where driving errors may have severe consequences.

The focus of this study is on the analysis and assessment of driver's speech acquired in real driving conditions. It extends our previous efforts presented in [16, 17]. Compared to [16], the number of analyzed subjects was increased from 15 to 68 and the scope was extended for emotion classification. The classification experiments of [17] are extended in this paper for the use of the state-of-the-art perceptually motivated minimum variance distortionless response (PMVDR) feature extraction front-end, which is shown to provide superior performance to the best system in [17] in the cognitive task classification.

Two cognitive load tasks considered in this study are represented by communication with a passenger versus interaction with two commercial dialog systems. Two emotional states, neutral and negative, where the negative state is induced by errors of the dialog system, are analyzed. Performance of selected speech production and cepstral acoustic features is compared in the classification of cognitive task and emotions.

The remainder of the paper is organized as follows. First, data sets used in the study are described. Subsequently, an approach to the cognitive task and emotion classification is proposed. Finally, experimental results are presented and discussed.

2. UTDRIVE CORPUS

This study is conducted on 68 driver sessions (33 females and 35 males) from the UTDrive database [18]. UTDrive captures recordings of real driving through urban areas in Richardson, TX. The driving routes comprise secondary, service, and main roads in residential and business districts. The vehicle used in the data acquisition was Toyota RAV4 equipped with microphones, CCD cameras monitoring the driver and the road scene, optical distance sensor, GPS, CAN-Bus OBD II port for speed measurement, steering wheel angle, gas and brake inputs from driver, and gas and brake pedal pressure sensors. Speech signal from the microphone mounted above the windshield is utilized in the experiments.

*This project was funded by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

During the driving sessions, the subjects were asked to perform a sequence of secondary (i.e., non-driving related) tasks such as sign reading, operating a radio and AC, talking to a passenger, and calling to two commercial automated dialog systems. The two dialog systems were: (i) American Airlines – online flight departure/arrival information system, and (ii) Tell ME – general information system (weather, game scores, movie theaters, etc.). In this paper, the driver interaction with the passenger and the dialog systems is studied.

3. PROPOSED APPROACH TO EMOTION/COGNITIVE LOAD TASK CLASSIFICATION

3.1. Emotional States and Cognitive Tasks

Following [16, 17], two emotional states and two cognitive tasks are studied in this paper. While calling the automated dialog systems, the users are frequently asked to repeat their queries. This is partly due to the confusion on the user's side about how to communicate with the system and in part due to the errors produced by the automatic speech recognition (ASR) within the dialog system. Intuitively, both types of errors are likely to occur more frequently when the user's attention is split between driving a vehicle and communicating with the system, and while the ASR engine is exposed to increased background noise.

Frequent requests of the dialog system for query repetitions may induce negative emotions in drivers. In our previous study [17], two broad *emotional classes* – neutral and negative, were introduced to categorize the UTDrive speech samples obtained during the interactions with the dialog systems. Emotion labels were assigned to each conversational turn based on subjective judgment of an expert annotator. The proportion of negative interactions with the dialog system, with respect to the number of requested repetitions per query, is shown in Table 1 (F and M denote female and male subjects respectively; Re0 – no repetition, Re1 – 1^{st} repetition, Re2-6 – 2^{nd}–6^{th} repetitions).

Table 1. Proportion of negative conversational turns in automated dialog system interactions.

Query	% of Negative Queries	
	Females	Males
All	38.1	17.6
Re0	23.9	8.8
Re1	65.9	47.2
Re2-6	78.3	60.0

It is difficult to quantify the absolute level of *cognitive load* in individuals, however, it may be argued that some types of tasks are likely to induce higher cognitive load than others. In our case, the focus is on the driver's interactions with a passenger and on the interactions with the automated dialog systems. In UTDrive, the passenger interactions were of a relaxed nature and the discussed topics did not require any extensive focus (discussing weather, etc.). On the other hand, the interactions with the dialog system through a cell phone comprised a sequence of steps that needed to be correctly carried out to fulfill the task, including holding the phone, dialing the number, and navigating through the menu of the dialog system. Further cognitive load was induced when the subject consciously tried to reformulate the query and altered the talking style to become more intelligible to the system. For the purpose of cognitive load labeling of the conversational turns, similar to [16, 17], we apply *cause-type* annotation [13] of the cognitive load and map cognitive load labels to individual tasks – co-driver interactions (low cognitive load) and dialog system interactions (high cognitive load).

3.2. Speech Production Model-Based Features

A number of speech production parameters is known to be sensitive to emotions and stress [9, 3, 19]. In [17], the following speech production factors were analyzed in the context of UTDrive sessions: mean utterance fundamental frequency F_0, first four formant center frequencies in voiced speech segments F_{1-4}, spectral slope, duration of voiced segments, spectral center of gravity (SCG), and spectral energy spread (SES). SCG is defined

$$SCG = \frac{\sum\limits_{k=1}^{N} X(k) \cdot k}{\sum\limits_{k=1}^{N} X(k)} \qquad (1)$$

where k is the discrete frequency and $X(k)$ amplitude of the corresponding spectral bin. Spectral energy spread is defined for the spectral energy distribution as a frequency interval of one standard deviation from SCG

$$SES = \sqrt{\frac{\sum\limits_{k=1}^{N} X(k) \cdot (k - SCG)^2}{\sum\limits_{k=1}^{N} X(k)}} \qquad (2)$$

and is expected to be sensitive to changes in energy distribution across the frequency axis. The following was found out in the statistical tests in [17] for the production parameter variation across cognitive tasks, emotions, and number of repeated queries:

- *Passenger (low cognitive load) vs. dialog system (high cognitive load) interactions*: F_0, F_1, F_4, SCG, SES, $duration$ – significant increase; $F_{2,3}$, *spectral slope* – no significant effects;

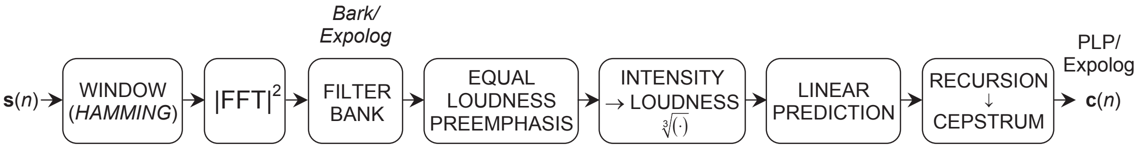

Fig. 1. Extraction of PLP and Expolog cepstral features.

- *Neutral vs. negative emotion interactions*: F_0, SCG, $duration$ – significant increase; F_{1-4}, SES, *spectral slope* – no significant effects;

- *Effect of requested query repetition interactions*: F_0 – significant interaction; F_{1-4}, SCG, SES, $duration$, *spectral slope* – no significant effects.

It is noted that spectral slope is typically reported in literature to be sensitive to emotions and stress [9,20]. In our study, the 'steadiness' of spectral slope is most probably caused by a prominent car noise energy content present at low frequencies of the amplitude spectrum (see details in [16]).

3.3. Cepstral Acoustic Features

In addition to speech production model-based features, suitability of several cepstral-based features, borrowed from the ASR domain, for the cognitive load and emotion classification is evaluated. In particular, mel frequency cepstral coefficients (MFCC) [21], perceptual linear prediction (PLP) [22], Expolog cepstral coefficients [23], and perceptually motivated minimum variance distortionless response (PMVDR) cepstral coefficients [24] are compared. Expolog and PMVDR were demonstrated to provide promising performance in stressed speech recognition. Expolog cepstra employ a triangular filter bank distributed on exponential-logarithmic frequency scale

$$\mathrm{Expolog}\,(f) = \begin{cases} 700 \cdot \left(10^{\frac{f}{3988}} - 1\right) & 0 \leqslant f \leqslant 2000Hz, \\ 2596 \cdot \log\left(1 + \frac{f}{100}\right) & f > 2000Hz. \end{cases} \qquad (3)$$

The Expolog filter bank is used in this study as a replacement of the trapezoid filter bank in the PLP feature extraction scheme (see Fig. 1). The PMVDR feature extraction utilizes a minimum variance distortionless response (MVDR) spectral estimator to represent the upper envelope of the speech signal. Unlike in other cepstral front-ends considered in this study, PMVDR (see block scheme in Fig. 2) does not employ any filter bank and performs frequency warping by directly interpolating the amplitude spectrum. In addition, two variants of MFCC and PLP, where discrete cosine transform cepstrum was replaced by linear prediction cepstrum (MFCC-LPC) and vice-versa (PLP-DCT) are considered in the experiments.

3.4. SVM-Based Feature Fusion

A Gaussian mixture model (GMM) based maximum *a posteriori* classifier is used in the baseline classification experiments. Here, selected features are used to parameterize speech signal. Separate GMM's are trained for each class (low/high cognitive load, neutral/negative emotions). Test samples are scored against each of the GMM's and a resulting class is assigned based to the ratio of the GMM likelihoods compared to the decision threshold. Cepstral features are extracted on the frame level (25 ms window/10 ms step); each frame is scored against the individual GMM's. The final class decision for a conversational turn is based on the overall likelihood across all frames of the turn. In the case of speech production features, mean feature values are extracted across the conversational turn, yielding a single feature vector for scoring.

Fusion of cepstral and speech production-based features is conducted using a support vector machine (SVM) classifier. Depending on the experimental setup, the input to the SVM classifier is formed by the combination of likelihood scores from the GMM's and raw speech production features (see Fig. 3). There are two parameters in the SVM framework that require to be tuned to optimize the classification performance. The first one is the penalty parameter C which determines the trade-off between margin maximization and training error minimization. The second parameter, γ, is inversely proportional to the Gaussian kernel width. These parameters are first optimized on the training data using a 5-fold cross validation strategy and a grid search in the intervals of $\log_2 \gamma \in \{-15, -14.5, ..., 3\}$, $\log_2 C \in \{15, 14.5, ..., -3\}$ and then used for open test set classification. Fig. 4 illustrates an example of such a parameter selection. It shows classification accuracy contours obtained from a grid search for several parameter sets. In this example, the inner-most contour represents the highest performance obtained for the pa-

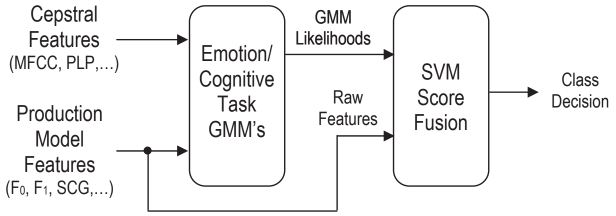
Cepstral Features (MFCC, PLP,...)
Production Model Features (F0, F1, SCG,...)
Emotion/ Cognitive Task GMM's
GMM Likelihoods
Raw Features
SVM Score Fusion
Class Decision

Fig. 3. SVM-based fusion of cepstral and speech production features.

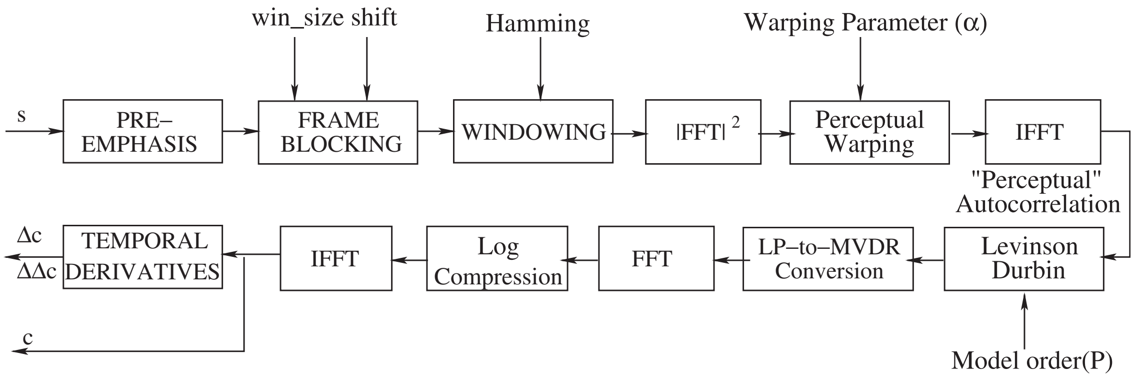

Fig. 2. PMVDR feature extraction scheme (*after* [24]).

rameter set $\log_2 C = 13$, $\log_2 \gamma = -6$. SVM has a potential to provide further performance gains compared to pure GMM/decision threshold classification since here, the decision hyperplane in the feature space is searched in a discriminative training process as opposed to the generative training of GMM models and the usage of a scalar decision threshold.

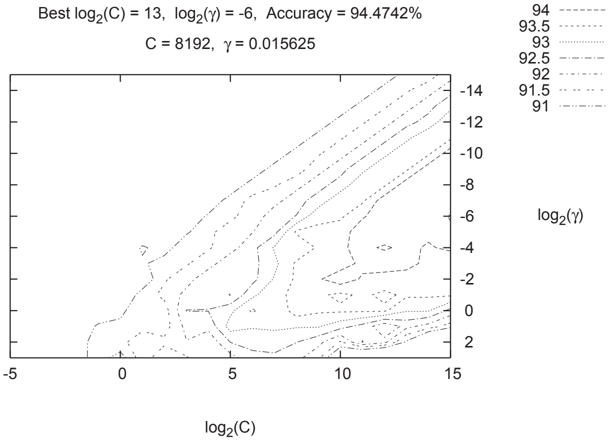

Fig. 4. Example of SVM parameter selection: classification accuracy contours; train set.

4. EXPERIMENTAL RESULTS

For both cognitive load and emotion classification experiments, samples from 40 speaker sessions (20 per gender) were used to train the GMM acoustic models. The remaining 28 sessions were used for open test set evaluations. The results of all experiments are reported by means of equal accuracy rate (EAR) where the class decision threshold is set to provide as close classification accuracy for both classes as possible and the class accuracies are subsequently averaged.

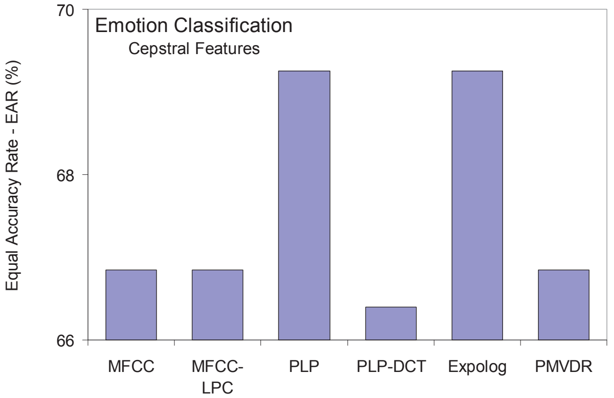

Fig. 5. Emotion classification – performance of cepstral-based GMM classifiers.

4.1. Emotion Classification

The results of neutral/negative emotion classification are summarized in Figures 5–7 and Table 2. Fig. 5 shows the performance of GMM classifiers utilizing cepstral-based feature extraction front-ends. It can be seen that the best performance is established by PLP and Expolog systems at EAR of 69.3 %.

Performance of the best production feature-based GMM classifiers (see Fig. 6) considerably exceeds the one of the cepstral-based systems. The combination of F_0, SCG, and *duration* yields 78.1 % EAR. Fig. 7 summarizes the performance of selected SVM-based classifiers. The best system utilizes likelihood scores of GMM's trained on a feature vector comprising F_0, SCG, and *duration* and provides EAR of 79 %. Note that here, SVM demonstrates its potential in finding a class-decision boundary that is more effective than the scalar decision threshold used in the GMM *a posteriori* classifier.

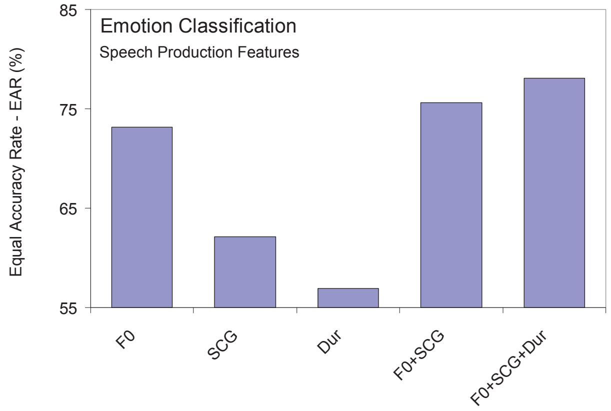

Fig. 6. Emotion classification – performance of production feature-based GMM classifiers.

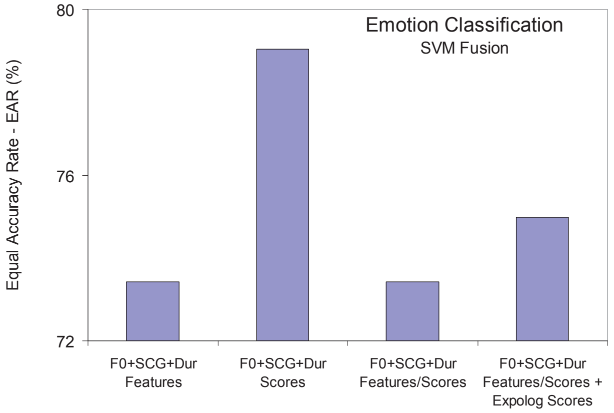

Fig. 7. Emotion classification – performance of SVM fusion systems.

4.2. Cognitive Task Classification

The results of passenger (low cognitive load) vs. dialog system (high cognitive load) classification are detailed in Figures 8–10 and summarized in Table 3. Fig. 8 presents performance of GMM classifiers utilizing cepstral-based features. The best performance is established by PMVDR with EAR of 94.3 %.

Unlike in the emotion classification task, performance of the best production feature-based GMM classifiers (see Fig. 9) is somewhat lower than that of cepstral-based systems. Using mean SCG as a single speech signal descriptor yields 90.0 % EAR. Combination of SCG with other parameters yields either similar or lower performance.

Fig. 10 summarizes performance of selected SVM-based classifiers. The best system utilizes likelihood scores of GMM's trained on PMVDR, likelihood scores of SCG, and raw SCG values (EAR of 95.2 %). It is noted that this system outperforms the best system in [17].

Table 2. Emotion classification – comparison of all systems.

Emotion Classification		
Domain	Features	EAR (%)
Cepstral Features	MFCC	66.9
	MFCC-LPC	66.9
	PLP	**69.3**
	PLP-DCT	66.4
	Expolog	**69.3**
	PMVDR	66.9
Speech Production Model Features	F0	73.2
	SCG	62.2
	Dur	57.0
	F0+SCG	75.7
	F0+SCG+Dur	**78.1**
SVM Fusion	F0+SCG+Dur Features	73.4
	F0+SCG+Dur Scores	**79.0**
	F0+SCG+Dur Features/Scores	73.4
	F0+SCG+Dur Features/Scores + Expolog Scores	75.0

5. CONCLUSIONS

This study compared efficiency of selected speech production and cepstral-based features for neutral/negative emotion and low/high cognitive load classification. The experiments utilized samples from 68 subjects acquired in real driving scenarios. In the emotion classification task, speech production features provided better class discriminability than cepstral features. The best performance was provided by an SVM classifier that utilizes class-specific GMM likelihoods as its input. This scheme yielded 79.0 % equal accuracy for a setup with F_0, SCG, and *duration* forming the input feature vector. In

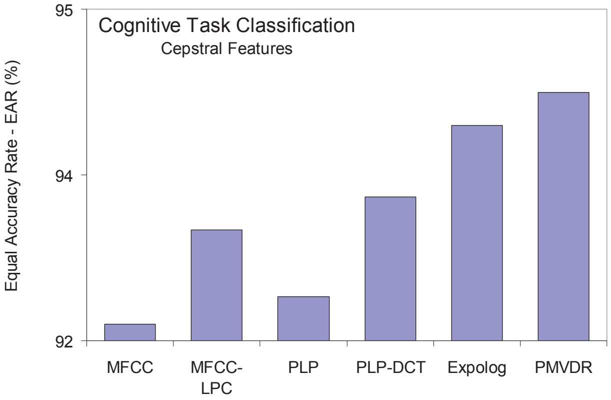

Fig. 8. Cognitive task classification – performance of cepstral-based GMM classifiers.

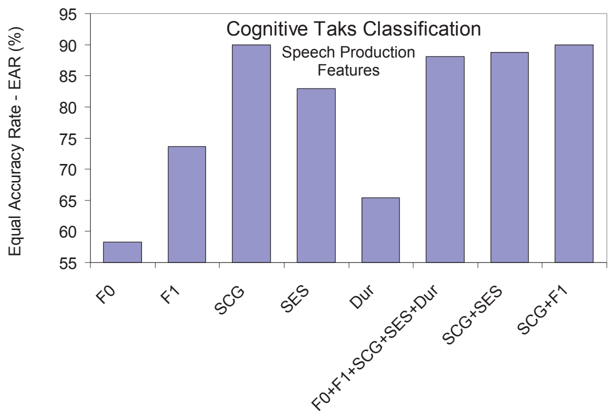

Fig. 9. Cognitive task classification – performance of production feature-based GMM classifiers.

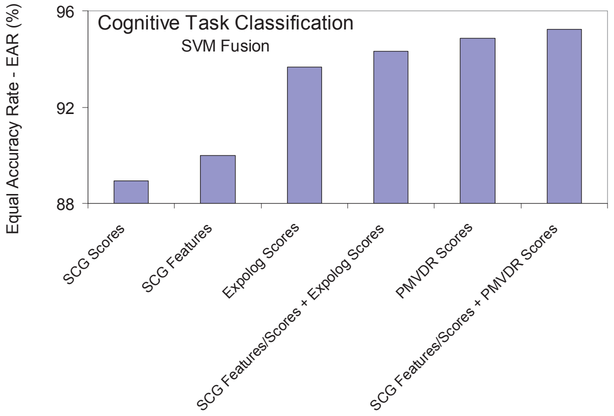

Fig. 10. Cognitive task classification – performance of SVM fusion systems.

Table 3. Cognitive task classification – comparison of all systems.

Cognitive Task Classification		
Domain	Features	EAR (%)
Cepstral Features	MFCC	92.2
	MFCC-LPC	93.0
	PLP	92.4
	PLP-DCT	93.3
	Expolog	94.0
	PMVDR	**94.3**
Speech Production Model Features	F0	58.3
	F1	73.7
	SCG	**90.0**
	SES	83.0
	Dur	65.5
	F0+F1+SCG+SES+Dur	88.2
	SCG+SES	88.8
	SCG+F1	**90.0**
SVM Fusion	SCG Scores	88.9
	SCG Features	90.0
	Expolog Scores	93.7
	SCG Features/Scores + Expolog Scores	94.3
	PMVDR Scores	94.9
	SCG Features/Scores + PMVDR Scores	**95.2**

the case of cognitive task classification, the best performance, EAR of 95.2 %, was provided by the fusion of cepstral (PMVDR) based GMM likelihoods and SCG raw values and GMM likelihoods.

6. REFERENCES

[1] B. Magladry and D. Bruce, *In-Vehicle Corpus and Signal Processing for Driver Behavior*. USA: Springer, 2008, ch. Improved Vehicle Safety and How Technology Will Get US There, Hopefully). K. Takeda, H. Erdogan, J. H. L. Hansen, H. Abut (Eds.), pp. 1–8.

[2] H. Cai, Y. Lin, and R. R. Mourant, “Study on driver emotion in driver-vehicle-environment systems using multiple networked driving simulators,” in *Proc. Driving Simulation Conference – North America 2009*, Iowa City, Iowa, 2007.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion recognition in human-computer interaction,” *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32 –80, jan 2001.

[4] D. Neiberg, K. Elenius, I. Karlsson, and K. Laskowski, “Emotion Recognition in Spontaneous Speech Using GMMs,” in *Proc. of ICSLP‘06*, Pittsburgh, PA, USA, 2006, pp. 809–812.

[5] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Transactions on Speech & Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[6] H. P. Espinosa, C. A. R. Garca, and L. V. Pineda, “Features selection for primitives estimation on emotional speech,” in *IEEE ICASSP’10*, Dallas, TX, 2010, pp. 5138–5141.

[7] W. Kim and J. H. L. Hansen, “Angry emotion detection

from real-life conversational speech by leveraging content structure," in *IEEE ICASSP'10*, Dallas, TX, 2010, pp. 5166–5169.

[8] K. Cummings and M. Clements, "Analysis of glottal waveforms across stress styles," in *Proc. of ICASSP'90*, vol. 1, Albuquerque, USA, 1990, pp. 369–372.

[9] 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.

[10] B. Yin, N. Ruiz, F. Chen, and M. A. Khawaja, "Automatic cognitive load detection from speech features," in *OZCHI '07: Proceedings of the 19th Australasian conference on Computer-Human Interaction*. New York, NY, USA: ACM, 2007, pp. 249–255.

[11] J. Wang and Y. Gong, "Normalizing multi-subject variation for drivers' emotion recognition," in *Proc. IEEE International Conference on Multimedia and Expo 2009*, New York, NY, USA, 2009, pp. 354–357.

[12] L. J. M. Rothkrantz, R. Horlings, and Z. Dharmawan, "Recognition of emotional states of car drivers by EEG analysis," *Neural Network World; International Journal on Neural and Mass - Parallel Computing and Information Systems*, vol. 19, no. 1, pp. 119–128, 2009.

[13] R. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech Communication*, vol. 40, no. 1-2, pp. 145–159, 2003.

[14] C. M. Jones and I. Jonsson, "Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses," in *Proc. 17th Australian Conference on Computer-Human Interaction*, Canberra, Australia, 2005, pp. 1–10.

[15] C. M. Jones and I.-M. Jonsson, *Universal Access in Human-Computer Interaction. Ambient Interaction*. Berlin/Heidelberg: Springer, 2007, ch. Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces). W.B. Kleijn and K.K. Paliwal (Eds.), pp. 411–420.

[16] H. Bořil, P. Boyraz, and J. H. L. Hansen, "Towards multi-modal driver's stress detection," in *Proc. of 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety*, Dallas, TX, 2009.

[17] H. Bořil, O. Sadjadi, T. Kleinschmidt, and J. H. L. Hansen, "Analysis and detection of cognitive load and frustration in drivers speech," in *Interspeech'10*, Makuhari, Chiba, Japan, September 2010, pp. 502–505.

[18] P. Angkititrakul, M. Petracca, A. Sathyanarayana, and J. Hansen, "Utdrive: Driver behavior and speech interactive systems for in-vehicle environments," June 2007, pp. 566–569.

[19] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416 – 433, 2008.

[20] H. Bořil, "Robust speech recognition: Analysis and equalization of lombard effect in czech corpora," Ph.D. dissertation, Czech Technical University in Prague, Czech Republic, http://www.utdallas.edu/∼hynek, 2008.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[22] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[23] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on Speech & Audio Processing*, vol. 8, no. 4, pp. 429–442, 2000.

[24] U. H. Yapanel and J. H. L. Hansen, "A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 2, pp. 142–152, 2008.