

Chapter 1

Towards Multimodal Driver's Stress Detection

Hynek Bořil, Pinar Boyraz, and John H.L. Hansen

Abstract Non-driving-related cognitive load and variations of emotional state may impact the drivers' capability to control a vehicle and introduce driving errors. The availability of stress detection in drivers would benefit the design of active safety systems and other intelligent in-vehicle interfaces. In this chapter, we propose initial steps towards multimodal driver stress (distraction) detection in urban driving scenarios involving multitasking, dialog system conversation, and medium-level cognitive tasks. The goal is to obtain a continuous operation-mode detection employing driver's speech and CAN-Bus signals, with a direct application for an intelligent human-vehicle interface which will adapt to the actual state of the driver. First, the impact of various driving scenarios on speech production features is analyzed, followed by a design of a speech-based stress detector. In the driver-/maneuver-independent open test set task, the system reaches 88.2% accuracy in neutral/stress classification. Second, distraction detection exploiting CAN-Bus signals is introduced and evaluated in a driver-/maneuver-dependent closed test set task, reaching 98% and 84% distraction detection accuracy in lane keeping segments and curve negotiation segments, respectively. Performance of the autonomous classifiers suggests that future fusion of speech and CAN-Bus signal domains will yield an overall robust stress assessment framework.

Keywords Active safety • CAN-bus signal processing • Distraction detection • Stress

H. Bořil (✉) • P. Boyraz • J.H.L. Hansen
Center for Robust Speech Systems, Erik Jonsson School of Engineering
& Computer Science, University of Texas at Dallas, Richardson, TX, USA
e-mail: hynek@utdallas.edu; boyraz.pinar@gmail.com; john.hansen@utdallas.edu

1.1 Introduction

Recent advancements in the electronic industry have made access to information and entertainment easier than ever before. While undoubtedly benefiting many areas of our daily lives, there are situations where the presence of electronic gadgets has the opposite effect. In a current study, the Virginia Tech Transportation Institute (VTTI) reports that dialing on a handheld device while driving increases the risk of an accident by a factor of 3, and communicating via hands-free set increases the risk by one third. This suggests that performing secondary cognitive tasks while driving may severely impact driving performance. Besides cognitive load, drivers' emotions have also been shown to adversely affect driving performance, e.g., by the means of larger deviations of lane offset and steering wheel angle, and shorter lane crossing times in anger and excitation situations – signs of reduced lane control capability. Availability of an automated system assessing stress in drivers would benefit the design of active safety systems and other intelligent in-vehicle interfaces, making them capable of adapting to the driver's current state (e.g., by decreasing the frequency of navigation prompts when detecting high-cognitive-load situations).

A number of studies have analyzed the impact of emotions [1–4] and stress (including cognitive load) on speech parameters [5–9]. However, relatively limited attention has been paid to the impact of emotion, stress, or distraction on the speech of car drivers [10, 11]. In [10], speech from subjects driving a simulator was categorized into seven emotional states, using a classifier trained on a corpus of emotional speech from professional actors. The emotional states in drivers were evoked during conversation with a dialog system. Also, Jones and Jonsson [11] used speech data collected in a driving simulator and categorized them into four stress classes. Different stress levels were induced by requesting the driver to maintain a certain speed (60 mph or 120 mph) and solve simple math tasks prompted at slow and fast rates by a synthesizer over the phone. The obtained classification performance in the driver-independent task was relatively low (~51%). We note that both studies utilize simulated driving scenarios, and in the case of [10] also employ simulated emotions from actors to establish classification categories. Acted emotions represent exaggerated traits that are effective in convincing listeners of the individual speaker state, but are not accurate representatives of natural emotions. Using driving simulators also introduces differences from real driving scenarios since there is less or no consequence for making errors in the primary task. In addition, a significant drawback of approaches utilizing only speech is that the emotion or stress assessment can be conducted only in time intervals when the driver is engaged in conversation.

To address these issues, the present study is conducted on the database UTDive [12] collected in real driving conditions and aims at utilizing both speech and CAN-Bus signals in the stress assessment. The term stress here represents the modality of the driver's speech production or driving behavior conducted under cognitive load. In the course of this chapter, the terms stress and distraction are used interchangeably, where the primary task is driving.

The remainder of the chapter is organized as follows: First, the data acquisition procedure and distraction/stress scenarios in UTDive corpus are described.

Second, an analysis of speech production parameters in three cognitive load scenarios is conducted, and a speech-based stress classifier is introduced. Third, a classifier operating on CAN-Bus signals is proposed and evaluated.

1.2 UTDrive Corpus, Data Subsets, and Transcription Protocols

The data collection vehicle is a Toyota RAV4 equipped with the following sensors (illustrated in Fig. 1.1):

- Two CCD cameras for monitoring the driver and the road scene through front windshield
- Microphone array (five mics) to record driver's speech as well as noise conditions in the vehicle
- A close-talk microphone to obtain driver's speech with reduced noise content
- Optical distance sensor to obtain headway distance between equipped vehicle and other vehicles in traffic
- GPS for location tracking
- CAN-Bus OBD II port for collecting vehicle dynamics: vehicle speed, steering wheel angle, gas and brake inputs from driver
- Gas/brake pedal pressure sensors to collect information concerning pressure patterns in car-following and braking behavior

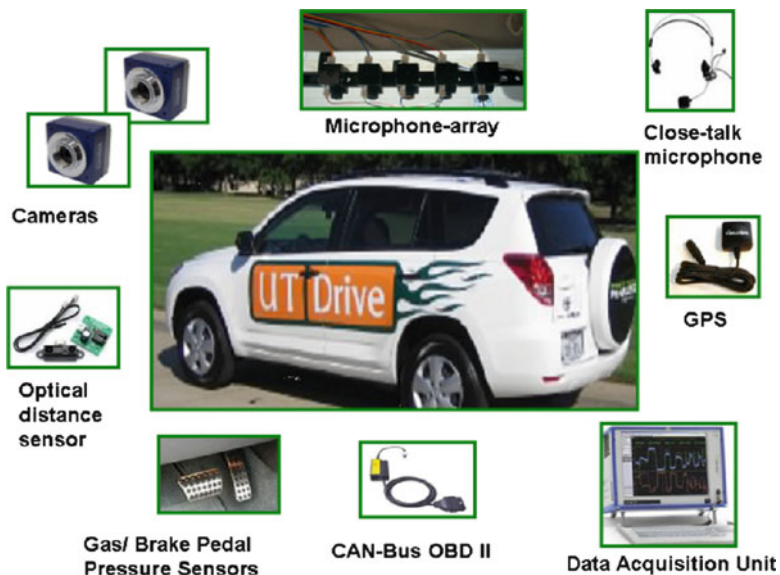


Fig. 1.1 Instrumented data collection vehicle: UTDrive



Fig. 1.2 Data collection: residential (left) and business (right) routes segmented according to assigned tasks

The UTDrive corpus includes data from the above-mentioned sensor channels (13 separate data streams: two video, six audio, one GPS, one optical distance, one CAN-Bus, two pressure sensors on gas/brake). The corpus is organized to have a balance in gender (37 males, 40 females), age (18–65), and different experience level (novice–expert) in driving. In order to examine the effect of distraction and secondary common tasks on these driver groups, a close-to-naturalistic data collection protocol is used.

The routes taken during data collection are given in Fig. 1.2, comprising a mixture of secondary, service, and main roads in residential (left-hand side map) and business (right-hand side map) districts in Richardson, TX. Each driver participating in the study is required to drive these two routes at least twice in each session to obtain a baseline and a distracted version of the same route. A session includes a mixture of several secondary tasks as listed in Table 1.1, taking place in road segments depicted in Fig. 1.2. According to this protocol, a participant performs 12 runs of data, with six being baselines for that day and that route, the other half featuring several distraction conditions. Each session is separated at least by 2 weeks in order to prevent driver complacency with the route and vehicle. Almost 60% of the data in the corpus have a full session profile from drivers. The remaining part contains incomplete sessions and data portions due to the consent of the participant not to continue data collection or several sensor failures. The secondary driver tasks are low to medium level of cognitive load while driving.

In this study, cell phone dialog parts including interaction speech with automated portals Tell-Me (information system) and American Airlines (reservation system) are utilized and analyzed using driver’s speech and CAN-Bus signals. The cell phone conversation takes place in route segment two which includes lane keeping and lane curvature negotiation tasks while the driver is engaged in cell phone dialog. In order to segment the data in terms of driving event and task timelines and find overlapping portions, two different transcription protocols are applied. First, using the audio and video, a task transcription is performed, having 13 labels to annotate the segments of the data in terms of where the driver and passenger talk and where other types of distractions occur. The second is called “event transcription” and performed

Table 1.1 UTDrive data collection protocol

		Secondary tasks		
Part		A	B	C
Route1	1	Lane changing	Common tasks (radio, AC etc.)	Sign reading
	2	Cell phone dialog	Cell phone dialog	Conversation
	3	Common tasks	Sign reading	Spontaneous
	4	Conversation	Spontaneous	Cell phone dialog
Route2	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	Conversation	Sign reading
Session		Route	Task	
1		1	Just drive	
		1	Secondary tasks A	
		2	Secondary tasks A	
		2	Just drive	
2		1	Just drive	
		1	Just drive	
		2	Secondary tasks B	
		2	Secondary tasks C	
3		2	Secondary tasks C	
		1	Secondary tasks C	
		2	Just drive	
		2	Just drive	

to have six labels to denote different maneuvers of the driver. A color-coded driving timeline is developed to observe aligned task and event transcriptions to obtain more insight into the data as well as to see the overlapping sections between tasks and events. A detailed explanation is given in [13] for transcription labels and color-coded driving timeline.

It should be noted that cell phone dialog includes different types of distractions: manual (dialing and holding), cognitive (interaction and processing), and auditory (listening). Therefore, the segment of the road containing the cell phone dialog can be considered as the highest possibility of observing high levels of distraction and divided attention. Although the cell phone in the car interfaces via a bluetooth device and the manual tasks from the driver minimized, the initial dialing might cause momentary distraction.

1.3 Stress Detection Using Speech Signal

This section focuses on the stress assessment from the driver's speech. First, it should be noted that the real level of stress in the driver caused by the cognitive load is not known. To define stress levels in the speech segments, we apply a *cause-type*

annotation of the data, as presented in [10]. Here, we hypothesize that a certain task the driver is asked to perform has a potential to cause a deviation of the driver’s speech production from neutral, and hence, represents a stress condition.

In particular, we expect that the interaction with the automated call centers *Tell-Me* and *American Airlines (AA)* puts an extensive cognitive load on the driver compared to the driver’s casual conversations with the passenger. This is expected partly due to the high demands of the automated call center on clear articulation, explicit formulation of the requests within a limited vocabulary of the system, and frequent requests for reentering the query due to the automatic speech recognition failure. For this reason, we denote spontaneous conversations with the passenger as *neutral* speech and calls to *Tell-Me* and *AA* as *stressed* speech. It is noted that even spontaneous communication with the passenger represents a certain level of cognitive load on the driver compared to silent segments and that due to the variable level of car noise, the driver is likely to exhibit various levels of Lombard effect [5, 14, 15].

In order to verify whether there are any measurable differences in the “neutral” and “stressed” portions of speech data and, hence, whether our hypothesis concerning the presence of stress in the higher-cognitive-load scenarios is reasonable, we first analyze the distributions of speech production parameters and compare them across hypothesized stress classes. Subsequently, we train separate Gaussian Mixture Models (GMMs) for neutral and stressed classes and evaluate the class discriminability using maximum likelihood classification. The gender-independent training and testing of the neutral/stress classifier is performed on disjunctive data sets from different speakers in order to evaluate the generalizing properties of the classification system.

1.3.1 *Speech Production Analysis*

Sessions from 15 drivers (seven females, eight males) are used in the speech analysis and stress classification experiments. An inspection of the close-talk microphone channel revealed a strong presence of “electric” noise completely masking the driver’s speech. For this reason, a middle microphone channel from the microphone array is used instead.

The following speech signal parameters are analyzed on the data down-sampled from 25 kHz to 16 kHz: signal-to-noise ratio (SNR), mean noise and speech power spectrum, fundamental frequency, first four formant frequencies and bandwidths, and spectral slope of voiced speech segments. SNR was estimated from (1) segmental SNR estimator [16], (2) average *noise* power spectrum, and (3) average *noisy speech* power spectrum. The SNR distribution obtained from the first method is shown in Fig. 1.3; the mean SNR reaches -2.7 dB, with the standard deviation of 4.4 dB. Note that the SNR values in the distribution are quite low due to the distant microphone placement from the driver.

To verify the estimate from the segmental detector, in the next step, SNR is estimated directly from the average *noise* power spectrum (N) extracted from all

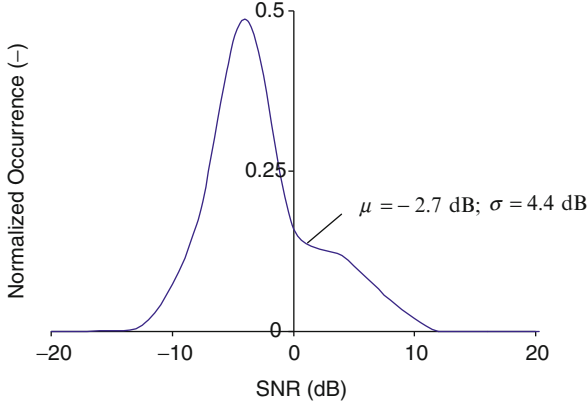


Fig. 1.3 Distribution of SNR across all sessions

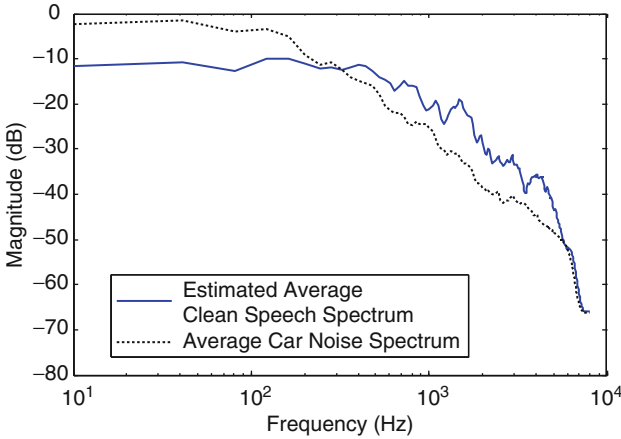


Fig. 1.4 Average amplitude spectrum of noise and clean speech – averaged across all sessions

nonspeech segments, and the average *noisy speech* power spectrum (SN) is estimated from all passenger conversation, Tell-Me and AA segments:

$$\widehat{SNR} = 10 \cdot \log \sum_k \frac{SN_k - N_k}{N_k}, \tag{1.1}$$

where k denotes the power spectrum frequency bin index. The SNR estimate obtained from the power spectra reaches -3.2 dB, confirming a reasonable accuracy of the segmental SNR estimation. The average power spectrum of noisy segments without speech and of clean speech estimated by subtracting N from SN is shown in Fig. 1.4. It can be seen that the car noise spectrum dominates over speech at low frequencies while speech becomes dominant, in spite of the low SNR, at frequencies higher than 300 Hz.

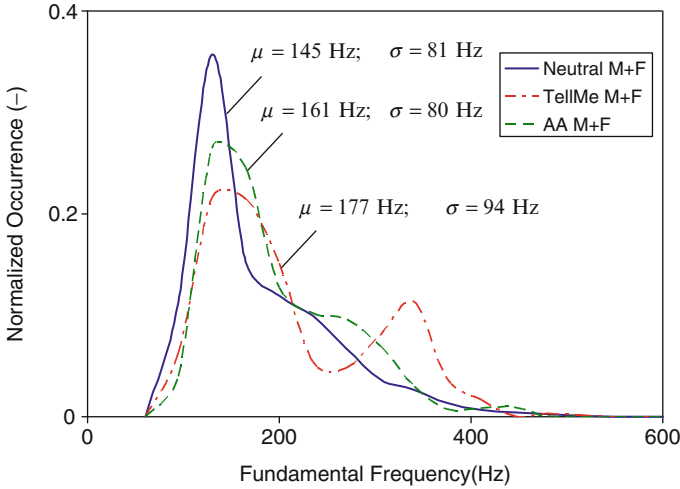


Fig. 1.5 Distribution of fundamental frequency in neutral, Tell-Me, and AA sessions

Table 1.2 Formant center frequencies and bandwidths (in parentheses)

Gender	Scenario	Formants and bandwidths (Hz)			
		F1	F2	F3	F4
F	Neutral	555 (219)	1,625 (247)	2,865 (312)	4,012 (327)
	Tell-Me	703 (308)	1,612 (276)	2,836 (375)	3,855 (346)
	AA	710 (244)	1,667 (243)	2,935 (325)	4,008 (329)
M	Neutral	450 (188)	1,495 (209)	2,530 (342)	3,763 (343)
	Tell-Me	472 (205)	1,498 (214)	2,525 (341)	3,648 (302)
	AA	503 (188)	1,526 (215)	2,656 (330)	3,654 (369)

In the next step, speech production parameters are analyzed. Distributions of fundamental frequency in passenger conversations (denoted *Neutral*), and Tell-Me and AA conversations are depicted in Fig. 1.5, where *M + F* stands for mixed-gender data sets. Both Tell-Me and AA samples display a consistent increase in mean fundamental frequency (177 Hz and 161 Hz) compared to neutral (145 Hz).

Mean center frequencies and bandwidths of the first four formants were extracted from voiced speech segments using WaveSurfer [17]. They are compared for neutral, Tell-Me, and AA conversations in Table 1.2. The voiced segments were identified based on the output of the pitch tracking algorithm implemented in [17] (RAPT [18]).

Mean center frequencies and standard deviations of F1 are displayed in Fig. 1.6. A consistent increase in F1 can be observed for Tell-Me and AA data. In AA,

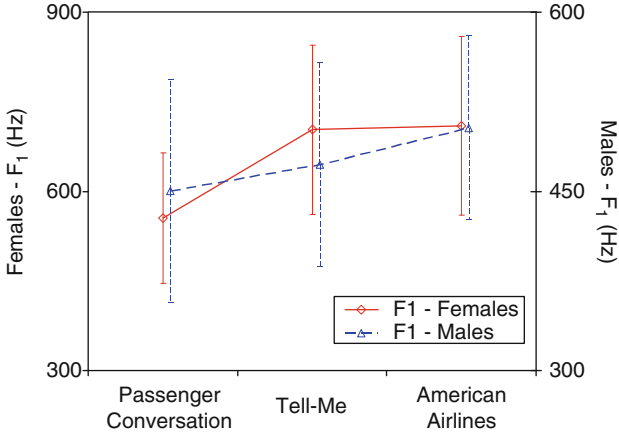


Fig. 1.6 Mean F₁ center frequency in neutral, Tell-Me, and AA sessions (accompanied by standard deviations in error plots)

also F₂ and F₃ increase in both genders while remaining relatively steady in Tell-Me. Note that F₁ and F₂ increases have been previously reported for stressed speech, including angry, loud, and Lombard speech modes [5, 14, 15]. Finally, spectral slopes of the voiced speech segments were extracted by fitting a straight line to the short-term power spectra in the log amplitude/log frequency plane by means of linear regression [14]. The mean spectral slope reaches values around -10.4 dB/Oct, displaying no significant differences across stress classes. Note that the average slope is somewhat higher than that reported in the literature for clean neutral speech, presumably due to the strong presence of background car noise, which introduces additional spectral tilt.

The analysis conducted in this section revealed differences in fundamental frequency, F₁, and F₂ center frequencies between the selected neutral and stressed classes, confirming that the initial hypothesis about the presence of stress in Tell-Me and AA segments due to increased cognitive load is valid.

1.3.2 Automatic Classification of Stress

In this section, speech-based neutral/stress classification is proposed and evaluated. For the purposes of classifier training and testing, the data from 15 drivers were split into a training set comprising of speech samples from two male and two female drivers, and test set comprising six male drivers and five female drivers.

Gaussian Mixture Models (GMMs) are chosen to represent probability density functions (PDFs) of the neutral and stressed classes. The probability of observation vector \mathbf{o}_t being generated by the j th GMM is calculated as

$$b_j(\mathbf{o}_t) = \sum_{m=1}^M \frac{c_{jm}}{\sqrt{(2\pi)^n |\Sigma_{jm}|}} \cdot e^{-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}_{jm})^T \Sigma_{jm}^{-1} (\mathbf{o}_t - \boldsymbol{\mu}_{jm})}, \quad (1.2)$$

where m is the index of the Gaussian mixture component, M is the total number of mixtures, c_{jm} is the mixture weight such that

$$\sum_{m=1}^M c_{jm} = 1, \quad (1.3)$$

n is the dimension of \mathbf{o}_t , Σ_{jm} is the mixture covariance matrix, and $\boldsymbol{\mu}_{jm}$ is the mixture mean vector. The GMM representing neutral speech was trained on the passenger conversations and the stressed speech GMM on joint Tell-Me and AA conversations from the training set. In the neutral/stress classification task, the winning model is selected using a maximum likelihood criterion:

$$j_{win} = \begin{cases} 1, & \sum_{t=1}^T \log(b_1(\mathbf{o}_t)) - \sum_{t=1}^T \log(b_2(\mathbf{o}_t)) \geq Th, \\ 2, & \sum_{t=1}^T \log(b_1(\mathbf{o}_t)) - \sum_{t=1}^T \log(b_2(\mathbf{o}_t)) < Th, \end{cases} \quad (1.4)$$

where t is the time frame index, T is the total number of frames in the classified utterance, and Th is the decision threshold.

In our experiments, the frame length was set to 25 ms, skip rate 10 ms, and the decision threshold to a fixed value $Th = 0$. Depending on the feature extraction scheme, the GMMs comprise 32–64 mixtures, and only diagonals are calculated in the covariance matrices. Unless otherwise specified, $c_{0-c_{12}}$ form the static observation feature vector. In all evaluation setups, delta and acceleration coefficients are extracted from the static coefficients and complete the feature vector. A variety of features, including Mel Frequency Cepstral Coefficients (MFCC), are considered.

In the UTDrive sessions, the amount of *neutral* spontaneous conversation data considerably exceeds the number of Tell-Me and AA samples. In this case, possible misclassification of small amount of stressed samples would have little effect on the overall classification accuracy, while classifying correctly only neutral data would assure high overall accuracy. To eliminate the impact of different sizes of the neutral and stressed sets, and to allow for accuracy-based selection of the optimal front-end for both AA and Tell-Me conversation scenarios, the overall classification accuracy is determined as

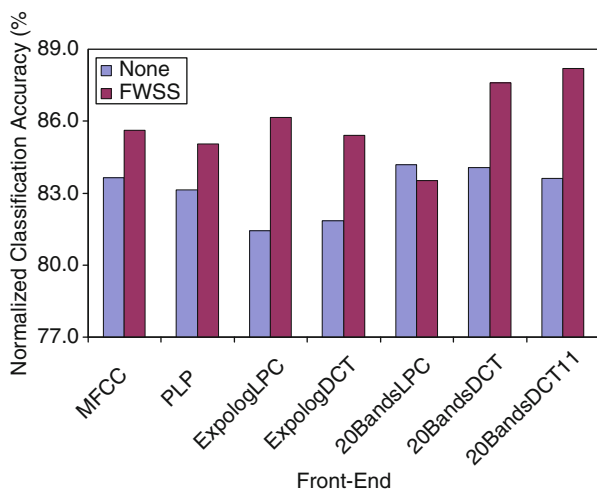
$$Acc = \frac{2Acc_{N-N} + Acc_{TellMe-S} + Acc_{AA-S}}{4} (\%), \quad (1.5)$$

where Acc_{N-N} is the accuracy of neutral samples being classified as neutral, $Acc_{TellMe-S}$ is the accuracy of Tell-Me samples being classified as stressed, and Acc_{AA-S} is the accuracy of AA samples being classified as stressed.

Efficiency of several feature extraction front-ends was evaluated in the neutral/stress classification task. In particular, Mel Frequency Cepstral Coefficients (MFCC [19]), Perceptual Linear Prediction (PLP) cepstral coefficients [20], Expolog cepstra [21],

Table 1.3 Classification performance; normalized accuracy (%)

NS	Front-end						
	MFCC	PLP	Expolog LPC	Expolog DCT	20Bands LPC	20Bands DCT	20Bands DCT11
None	83.7	83.1	81.4	81.9	84.2	84.1	83.6
FWSS	85.6	85.1	86.2	85.4	83.5	87.6	88.2

**Fig. 1.7** Front-end's classification performance

and cepstra extracted from a uniform filterbank of 20 non-overlapping rectangular filters distributed on a linear frequency scale (20Bands) [15] were compared. MFCC represent a common baseline front-end in speech/speaker recognition, and PLP has been shown by numerous studies to provide comparable or better performance to MFCC in various speech-related tasks [14].

Expolog is an outcome of studies on accent classification and stressed speech recognition, and features based on 20Bands filterbank have shown superior properties in noisy neutral and Lombard speech recognition [15].

In this study, Expolog and 20Bands filterbanks were used either as a replacement for the triangular Mel filterbank in MFCC, yielding front-ends denoted Expolog DCT and 20Bands DCT, or as a replacement for PLP trapezoid Bark filterbank, yielding setups denoted Expolog LPC and 20Bands LPC. In order to reduce the impact of strong background noise on classification, Full Wave Spectral Subtraction (FWSS) utilizing Burg's cepstral-based voice activity detector [14] was incorporated in the feature extraction. The classification results are summarized in Table 1.3 and Fig. 1.7. The first row of results in Table 1.3 represents the performance of a classifier without noise subtraction (NS), denoted "none."

It can be seen that in the majority of cases, FWSS considerably improves performance. Among front-ends employing 13 static coefficients and their first-and

second-order time derivatives, 20Bands DCT with FWSS provided the highest classification accuracy (87.6%). In addition, it was observed that decreasing the size of the static cepstral coefficients vector from 13 to 11 (c_0-c_{10}), denoted 20Bands DCT11, provides further accuracy increase to 88.2%. In this setup, the individual accuracies were $Acc_{N-N} = 91.4\%$, $Acc_{TeIlMe-S} = 70.0\%$, and $Acc_{AA-S} = 100.0\%$. Note that the accuracy and intraclass confusability can be further balanced by adjusting Th in Eq.1.4. However, for that, the availability of additional development data is required.

1.4 Distraction/Stress Detection Using CAN-Bus Signals

In this part of the study, we develop a distraction detection module based on a subset of CAN-Bus signals (mainly steering wheel angle and speed) using driver performance metrics, signal processing tools, and statistics. A generic distraction detection system without having the maneuver/context information and driver baselines for that particular maneuver is very difficult to design simply because the generic baseline for the nominal values of metrics/features has a wide range of variation due to driver characteristics and route/maneuver/context dependency.

CAN-Bus signals can reveal the distraction level of the driver when the variability due to maneuvers and driver characteristics are eliminated or dealt with so that they do not cause false alarms. Therefore, a methodology using a baseline for each individual driver and particular maneuver is proposed. A general flow diagram of the methodology is given in Fig. 1.8. The variation in the signals due to the maneuver/particular road segment is eliminated here by maneuver classification.

After the feature extraction process, distraction detection is performed by taking the driver's baseline for a given maneuver obtained from the same route segment (marked by two in Fig. 1.2) as when the conditions were neutral. Since UTDrive corpus includes multiple sessions collected from the same route and same driver under different conditions, hence, baselines can easily be obtained. The algorithm flow for distraction detection is shown in Fig. 1.9.

A normalized comparison ratio (α) is calculated for each element in the feature vector. The comparison ratio is used in multiple interval thresholds. Each threshold

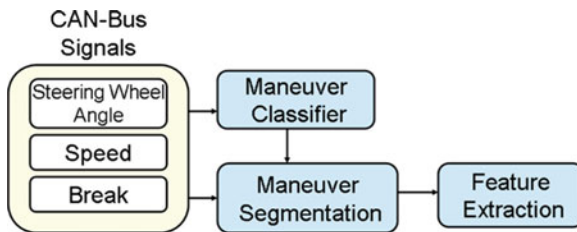


Fig. 1.8 Flow diagram of general methodology used for CAN-Bus-based analysis

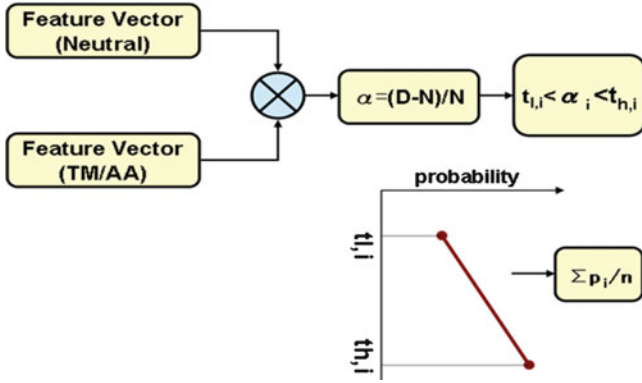


Fig. 1.9 Distraction detection algorithm flow based on features extracted from CAN-Bus signals

interval is assigned to a probability. For example, if the ratio is between 0.1 and 1, the probability of distraction is 0.7, and if the ratio is larger than 20, it is 1. This assignment approach allows for a probabilistic assessment of the distraction or can give an idea of the distraction level.

Comparison values larger than 0.1 in magnitude are considered to indicate a significant distraction. If the comparison value magnitude is below 0.1, the session is assumed to be close enough to baseline to be considered neutral. As the comparison ratio increases, the probability of being distracted increases, with the highest value being 1 as shown in Fig. 1.9. At the end of this probability mapping, the probabilities are summed along the feature vector (now comprised by comparison ratios) and normalized by dividing the resultant likelihood value in the feature vector dimension. The next section explains the feature extraction process and motivation behind the feature vector elements selected.

1.4.1 CAN-Bus-Based Features

The features are selected based on their relevance to distraction and definition of the maneuver. Using the color-coded driving timeline plots, it was observed that the route segment two contains lane keeping and curve negotiation tasks in terms of driving. For the lane keeping, several driver performance metrics are suggested in the literature mostly using steering wheel angle (SWA) to calculate a metric indicating the fluctuations or microcorrections in SWA input. Among these metrics, a widely accepted method is the sample entropy [22] and standard deviation. If available, the lane deviation measurements also give away if the driver is fully attentive and in control. The reversal rate of steering wheel is also considered to be a reliable metric to measure driver performance in a lane keeping task. Boer [23] recently updated his previous work and suggested some adjustments, taking high-frequency terms

Table 1.4 Feature vector and definitions

Notation	Definition
WDE_SWA	Wavelet decomposition detail signal energy for SWA
WDE_speed	Wavelet decomposition detail signal energy for speed
SampEnt_SWA	Sample entropy of SWA
SampEnt_speed	Sample entropy of SWA
STD_SWA	Standard deviation of SWA
STD_speed	Standard deviation of SWA
STD_SWAR	Standard deviation of SWA rate

into account. It was also pointed out in a thorough analysis [24] that the speed interval for which the SWA-dependent metric is being calculated is important since the lower speeds require more SWA inputs to achieve the same amount of lateral movement of the car compared to a higher speed. For the curve negotiation, a constant input of an angle required using the visual input of the road curvature.

The novice or distracted driver may have fluctuating inputs in the SWA, and the general trend is that the speed should be reduced while taking the curves to balance the centrifugal force. Although different in nature, lane keeping and curve negotiation can be seen as regulatory control tasks from the driver's point of view. Therefore, we selected a seven-dimensional feature vector 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 (R-STD). All features are extracted for SWA, and speed channels except R-STD are only applied to SWA. The time window length is taken as equal to the maneuver length, and the effect of the signal length is eliminated in the calculation of features. The entries of the feature vector are listed with their definitions in Table 1.4.

For the wavelet decomposition, Daubechies [25] wavelet kernel with fourth order is used, and detail signal is taken at the sixth level. Daubechies wavelet is chosen since it can approximate to signals with spikes and discontinuous attributes well. The level and order is adjusted to be able to extract the high-frequency content in the signal which is in the limitation of human control; the higher details are ignored since they might be caused by other disturbances in the measurement rather than driver. Scaling functions (a), wavelet function coefficients (b), scaling function (c), and wavelet function (d) for DB4 are given in equation group (1.6):

$$h_0 = \frac{1 + \sqrt{3}}{4\sqrt{2}}, h_1 = \frac{3 + \sqrt{3}}{4\sqrt{2}}, h_2 = \frac{3 - \sqrt{3}}{4\sqrt{2}}, h_3 = \frac{1 - \sqrt{3}}{4\sqrt{2}}, \quad (1.6a)$$

$$g_0 = h_3, g_1 = -h_2, g_2 = h_1, g_3 = -h_0, \quad (1.6b)$$

$$a_i = h_0 s_{2i} + h_1 s_{2i+1} + h_2 s_{2i+2} + h_3 s_{2i+3}, \quad (1.6c)$$

$$c_i = g_0 s_{2i} + g_1 s_{2i+1} + g_2 s_{2i+2} + g_3 s_{2i+3}. \quad (1.6d)$$

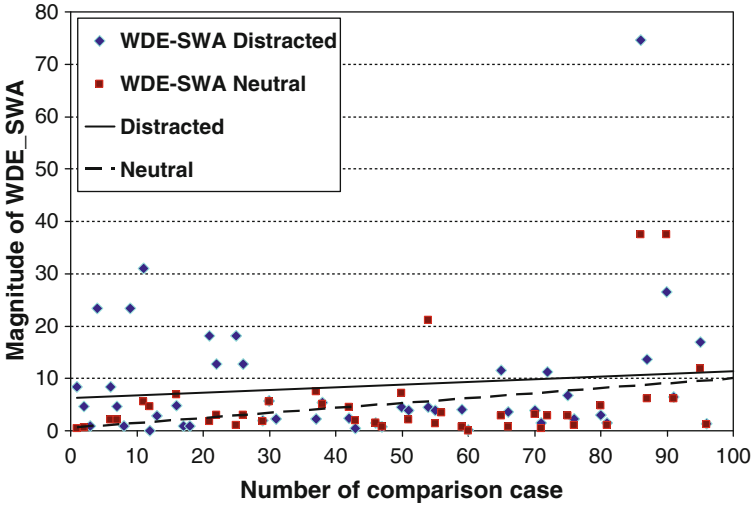


Fig. 1.10 Wavelet decomposition details signal energy for SWA calculated for 96 comparison cases of lane keeping

Sample entropy (SampEnt), which is used as a measure to quantify regularity and complexity of the signal, is a perfect match measuring the regularity of SWA signal. It is known that the measures based on entropy have long been employed in biosignal processing such as EEG, ECG, and EMG to measure regularity and detect abnormality. The method to calculate the sample entropy follows the work described in [26]. The standard deviation is calculated in a canonical form with statistics.

1.4.2 Distraction Detection Performance

Using the algorithm flow depicted in Fig. 1.9 and feature vectors explained in Table 1.4, 96 comparison cases for lane keeping and 113 cases for curve negotiation were examined using 14 drivers’ (20 sessions, seven female and seven male drivers) data. As an insight, WDE_SWA feature member is given for lane keeping maneuvers in Fig. 1.10. It can be easily seen that the distracted sessions are generally greater than the baseline for this metric. The accuracy of the distraction detection is given in Table 1.5 using seven-dimensional feature vector (LKS) and using four-dimensional feature vector subset containing only SWA-related features (LKC) with threshold values of 0.2, 0.1, and 0 for the final classification result.

From Table 1.5, it can be seen that if any probability value higher than zero is taken into account, the distraction can be detected with 98% accuracy using lane keeping segments (LKS) and by 84% accuracy using curve negotiation segments (LKC) during Tell-Me/AA conversations.

Table 1.5 Accuracy of distraction detection

Maneuver	Measure	Threshold					
		0.2	0.1	0	0 (Binary)		
LKS	Count	72/96	62/96	84/96	76/96	95/96	76/96
	Acc (%)	75	64	87	79	98	79
LKC	Count	65/113	64/113	82/113	79/113	95/113	79/113
	Acc (%)	57	56	72	69	84	69

The system offers a low-cost, driver-dependent, and reliable distraction detection submodule. Future work will focus on generic distraction detection using sums within the same feature space.

1.5 Conclusions

In this study, the impact of cognitive load on drivers was analyzed using the UTDrive database that comprises real-world driving recordings. In particular, driver's speech signal and CAN-Bus signals were studied and subsequently utilized in the design of autonomous speech and CAN-Bus domain neutral/stress (distraction) classifiers. The speech-based neutral/stress classification reached an accuracy of 88.2% in the driver-/maneuver-independent open test set task. The distraction detector exploiting CAN-Bus signals was evaluated in a driver-/maneuver-dependent closed test set task, providing 98% and 84% distraction detection accuracy in lane keeping segments and curve negotiation segments, respectively. The results suggest that future fusion of speech and CAN-Bus-based classifiers could yield a robust continuous stress (distraction) assessment framework.

References

1. Neiberg D, Elenius K, Karlsson I, Laskowski K (2006) Emotion recognition in spontaneous speech using GMMs. In: Proceedings of ICSLP'06, Pittsburgh, pp 809–812
2. Lee CM, Narayanan SS (2005) Toward detecting emotions in spoken dialogs. *IEEE Trans Speech Audio Proc* 13:293–303
3. Ijima Y, Tachibana M, Nose T, Kobayashi T (2009) Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM. In: Proceedings of IEEE ICASSP'09, Taipei, pp 4157–4160
4. Callejas Z, Lopez-Cozar R (2008) Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Comm* 50(5):416–433
5. Hansen JHL (1996) Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Comm* 20, 151–173
6. Cummings K, Clements M (1990) Analysis of glottal waveforms across stress styles. In: Proceedings of IEEE ICASSP'90, Albuquerque, vol 1, pp 369–372

7. Bou-Ghazale SE, Hansen J (1998) HMM-based stressed speech modeling with application to improved synthesis and recognition of isolated speech under stress. *IEEE Trans Speech Audio Proc* 6:201–216
8. Sarikaya R, Gowdy JN, (1998) Subband based classification of speech under stress. In: *Proceedings of ICASSP'98*, pp 569–572
9. Zhou G, Hansen J, Kaiser J (1998) Linear and nonlinear speech feature analysis for stress classification. In: *Proceedings of ICSLP'98*, Sydney, vol 3, pp 883–886
10. Fernandez Raul, Picard RW (2003) Modeling drivers' speech under stress. *Speech Comm* 40 (1–2):145–159
11. Jones CM, Jonsson I (2005) Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses. In: *Proceedings of 17th Australian conference on computer–human interaction*, Canberra, pp 1–10
12. Angkititrakul P, Petracca M, Sathyanarayana A, Hansen JHL (2007) UTDrive: driver behavior and speech interactive systems for in-vehicle environments. *IEEE intelligent vehicles symposium*, Istanbul, pp 566–569
13. Boyraz P, Sathyanarayana A, Hansen JHL (June 2009) CAN-bus signal modeling using stochastic methods and structural pattern recognition in time series for active safety. 4th biennial workshop on DSP for in-vehicle systems and safety, Dallas
14. Bořil H (2008) Robust speech recognition: analysis and equalization of Lombard effect in Czech corpora. Ph.D. dissertation, Czech Technical University in Prague <http://www.utdallas.edu/~hynek>
15. Bořil H, Hansen, JHL (2009) Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment. In: *Proceedings of IEEE ICASSP'09*, Taipei, pp 3937–3940
16. Vondrášek M, Pollák P (2005) Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency. *Radioengineering* 14:6–11
17. Sjolander K, Beskow J (2000) WaveSurfer – an open source speech tool. In: *Proceedings of ICSLP'00*, vol 4, Beijing, pp 464–467
18. Talkin D (1995) A robust algorithm for pitch tracking (RAPT). In: Kleijn WB, Paliwal KK (eds) *Speech coding and synthesis*. Elsevier, Amsterdam, pp 495–518
19. Davis SB, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans Acoust Speech Signal Proc* 28:357–366
20. Hermansky H (1990) Perceptual linear predictive (PLP) analysis of speech. *J Acoust Soc Am* 87:1738–1752
21. Bou-Ghazale SE, Hansen JHL (2000) A comparative study of traditional and newly proposed features for recognition of speech under stress. *IEEE Trans Speech Audio Proc* 8:429–442
22. Boer E (2001) Behavioral entropy as a measure of driving performance. In: *Proceedings of the first international driving symposium on human factors in driver assessment, training, and vehicle design*, Aspen, 14–17 August 2001
23. Boer E (2005) Steering entropy revisited. In: *Proceedings of the third international driving symposium on human factors in driver assessment, training, and vehicle design*, Rockport
24. Boyraz P, Sathyanarayana A, Hansen JHL (2009). Lane keeping metrics for assessment of auditory-cognitive distraction, [preprint] to appear in *SAE Book*, Chapter 10, Driver performance metrics
25. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Comm Pur Appl Math* 41:909–996
26. Xie HB, He WX, Liu H (2008) Measuring time series regularity using non-linear similarity-based sample entropy. *Phys Lett A* 372:7140–7146