IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 14, 2021, accepted February 23, 2021, date of publication February 26, 2021, date of current version March 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3062752

Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin

ALEXEY KASHEVNIK[1], IGOR LASHKOV[1], ALEXANDR AXYONOV[1], DENIS IVANKO[1], DMITRY RYUMIN[1], ARTEM KOLCHIN[2], AND ALEXEY KARPOV[1]

[1]St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 199178 Saint Petersburg, Russia
[2]Information Technologies and Programming Faculty, ITMO University, 197101 Saint Petersburg, Russia

Corresponding author: Alexey Kashevnik (alexey.kashevnik@iias.spb.su)

This work was supported in part by the Russian Foundation for Basic Research Project under Grant 19-29-09081, and in part by the Russian State Research developed mobile application for audio-visual corpus under Grant 0073-2019-0005.

ABSTRACT This paper introduces a new methodology aimed at comfort for the driver in-the-wild multimodal corpus creation for audio-visual speech recognition in driver monitoring systems. The presented methodology is universal and can be used for corpus recording for different languages. We present an analysis of speech recognition systems and voice interfaces for driver monitoring systems based on the analysis of both audio and video data. Multimodal speech recognition allows using audio data when video data are useless (e.g. at nighttime), as well as applying video data in acoustically noisy conditions (e.g., at highways). Our methodology identifies the main steps and requirements for multimodal corpus designing, including the development of a new framework for audio-visual corpus creation. We identify the main research questions related to the speech corpus creation task and discuss them in detail in this paper. We also consider some main cases of usage that require speech recognition in a vehicle cabin for interaction with a driver monitoring system. We also consider other important use cases when the system detects dangerous states of driver's drowsiness and starts a question-answer game to prevent dangerous situations. At the end based on the proposed methodology, we developed a mobile application that allows us to record a corpus for the Russian language. We created RUSAVIC corpus using the developed mobile application that at the moment a unique audiovisual corpus for the Russian language that is recorded in-the-wild condition.

INDEX TERMS Driver monitoring, automatic speech recognition, multimodal corpus, human–computer interaction.

I. INTRODUCTION

Last years, modern smartphones became perspective multi-functional powerful devices intended not only for calls and text messages, but also for a variety of different tasks including informational, multimedia, productivity, safety, lifestyle, accessibility related applications, and many others. Most modern smartphones already have a set of built-in sensors, sensing environment with readings of physical quantities. Essentially, they include a video-based camera, accelerometer, gyroscope, magnetometer, GPS, lightness, microphone, and proximity sensors. Because of their affordable low price, wide set of embedded sensors, and small sizes, smartphones are gaining popularity for building driver monitoring systems at a large scale.

Driver monitoring systems have become more and more popular in the last decades [1]. Such systems implement the functionality of dangerous states detection in vehicle cabin as well as driver style analysis based on the current situation and/or external factors [2]. Since such systems have to interact with the driver during the vehicle control the speech recognition based on human-computer interfaces has to be used to interact with the driver without distracting him/her from the road. In case the driver monitoring system detects drowsiness dangerous state we propose to use question/answer games to prevent the driver from sleeping. The smartphone that determines drowsiness dangerous state can be used for question/answer game implementation with the driver.

Modern speech recognition technologies [3] allow using of smartphone capabilities for the creation of human-computer interfaces that communicate with the human-based on speech commands. At the same time, some methods

The associate editor coordinating the review of this manuscript and approving it for publication was Razi Iqbal.

provide possibilities of emotional classification based on human speech analysis [4]. Modern smartphones have built-in graphics processing unit (GPU) modules and together with the sensors include all required functionality to implement these functions. We implemented a related work analysis in the topic of driver monitoring systems based on smartphone sensors to identify the main scenarios the designed speech recognition system should support.

In the paper, we review speech recognition approaches based on audio and video data that allow us to identify meta-parameters we should support our corpus creation. In recent years, automatic speech recognition by using both audio and video information streams has become a very active research topic, due to the ability of such approach to improving the accuracy and robustness of recognizing uttered speech. Multimodal speech recognition allows using of audio data at night-time when the video data cannot be used as well as video data in noisy conditions (e.g., highways). We analyze existing at the moment accessible corpora for audio-visual speech recognition and identify requirements for the corpora that can be used for speech recognition system development in vehicle cabin for the driver monitoring system. We identify the main use cases for a driver monitoring system that requires a speech recognition interface and discusses the vocabulary for the corpus. We discuss the mobile application design that we developed to record the corpora using several smartphones. Based on a related research review we concluded that we expected to record a corpus similar to AVICAR dataset [5]. This dataset is more related to the task of speech recognition based on human-computer interaction interface development in the vehicle cabin. Contrasting to AVICAR we identified the limited vocabulary and we will record the corpus in Russian language. AVICAR dataset supports only the English one. We expect that limited vocabulary will allow increasing the accuracy of speech recognition. In addition, we offer the use of various audio tools to prevent the driver from falling asleep. Based on the software and hardware that is available to us, we describe the prevention of falling asleep through word games.

We identify that the main contribution of the paper is to develop a methodology for corpus design as well as a software prototype for multimodal corpus creation for audio-visual speech recognition in driver monitoring systems.

We formulate the following research questions (RQ) that we answer in the paper.

- RQ1: What modern technologies are used for audio-visual speech recognition?
- RQ2: Which parameters are important to corpus creation for audio-visual speech recognition in a vehicle cabin?
- RQ3: How can the dialog-based interaction with the drowsy driver be used in the driver monitoring system to avoid sleepiness?
- RQ4: Which vocabulary should be supported for corpus creation?

The proposed methodology is convenient for the driver and provides for him/her a simple and effective way to record the data in-the-wild conditions. In the scope of the software prototype for the corpus recording system, one smartphone guides the driver what he/she needs to say and another one is synchronized with the first one and both of them are recorded the audio & video. Together with the audio/video data, we keep in the corpus driver monitoring data that allows us to recognize the context situation the driver pronounces the phrase.

The rest of the paper is organized as follows. Related research in the topics of driver monitoring systems based on smartphone sensors, speech recognition systems in the vehicle cabin, and corpora for audio-visual speech recognition are presented in Section 2. We divide them into several groups according to the data they use for monitoring driver behavior: visual-based data utilizing data obtained from a video camera; non-invasive data-based approach focused on measuring motion and position data; physiological data describing human behavior while driving (see subsection 2.1). In Subsection 2.2 we highlighted that more and more researchers are interested in a wide range of applications for automatic speech recognition (ASR) connecting humans and computers. In this regard, ASR in automotive voice navigation systems is highly popular and important. Modern reliable voice navigation systems allow providing additional safety for the driver by reducing the probability of driver distraction. Moreover, the dialogue system has the ability to interact with the driver in case of drowsiness to prevent sleepiness. In Subsection 2.3 we analyze a total of 50 databases accessible in modern research papers and highlight the top 10 for detailed review. The amount of acoustic-only speech datasets exceeds the number of audio-visual datasets multiple times. However, they are out of the scope of the presented study. Authors of the most considered papers record speech corpora in quiet office conditions. Such situation practically eliminates their usefulness for training models designed for deployment in noisy vehicle cabin conditions. We discuss the speech recognition vocabulary in Section 3. We define vocabulary both to support voice commands the driver uses to interact with the system and to support dialog-based question/answer games that the system proposes to the driver in order to detect the dangerous drowsiness state. We propose a corpus creation methodology in Section 4. It includes two smartphone and cloud services. The smartphone allows recording the audio and video data from different angles. We synchronize the data and send it to the cloud service for further processing. We discussed the corpus creation in Section 5 and shows the corpus structure as well as provide the link to the corpus portal. We discuss the research questions in Section 6. The conclusion summarizes the paper and contains the main discussion of the results.

II. RELATED WORK

We consider modern research has been done in the following main topics that are related to the problem domain: driver monitoring systems based on smartphone sensors, speech interfaces in the vehicle cabin, and available audio-visual speech recognition corpora recorded in the vehicle cabin.

A. SMARTPHONE-BASED DRIVER MONITORING

The research study [6] utilizes different sources of visual-based information to detect a driver's drowsiness state as well as, to predict when he/she reaches a given threshold of drowsiness. Authors explore a combination of measured behavioral and physiological indicators such as head and eyelid movements, percentage of time the driver's eyes remain closed (PERCLOS), heart rate, as well as recorded driving behavior, including speed of the vehicle, steering wheel angle, and position on the road lane. This kind of parameters may potentially affect dangerous state recognition and improve accuracy and prediction of drowsiness evaluation. Authors developed two models built on utilizing neural networks to determine and predict a person's drowsiness rate every minute. Data for these models were provided during the experiments conducted with participants who drove a car simulator under certain conditions. Summarizing study outputs, gaze, and head movements, including other kinds of information, including driving time, showed the best results in the prediction of the time when the driver will become distracted, as well as detection of the current drowsiness level.

In other researches, the authors of the paper [7] propose a method for recognizing a driver's drowsiness state based on analyzing driving actions utilizing audio devices integrated into smartphones. This method can detect a dangerous driver's behavior, including nodding, yawning, and abnormal operating of the steering wheel on the device in real-time. In this case, the proposed approach uses smartphone microphones to collect audio signals generated by speakers in real-time and splits audio-signals into separate frames. In order to identify different levels of drowsiness state, the authors analyzed unique patterns of Doppler shift with the aid of the sliding window algorithm. The outcome of the research study showed that different drivers demonstrate similar patterns based on the data collected in natural driving situations. This research proposed two separate classifiers, where the first one is responsible for short-term prediction, including nodding and yawning, while the other one is for long-term prediction, utilizing the steering wheel. One of the most distinctive features of this work is the usage of long short-term memory networks, which in its turn considers not only the current frame but also several previous frames to make a prediction.

A study [8] describes the monitoring of abnormal driving behavior and recognizing drowsiness, distraction, and gaze direction states. The presented approach utilizes facial landmarks to detect eye closure through eye aspect ratio and PERCLOS measurements as well as yawns with the aid of the ratio of the height of the mouth to its width. Authors propose to recognize the distraction state when the driver uses the smartphone while driving. They propose to use a pre-trained deep neural network YOLOv2 on the COCO dataset. This approach was tested using a smartphone camera solution.

Authors of the paper [9] researched the impact of the smartphone-based driver assistance system on driving behavior while the driver is impaired by the use of mobile social networking applications. The developed approach uses the driving simulator that replicates the car structure and its parts and includes the following modules. For a driver's behavior analysis, the researchers utilized the eye-tracking system Tobix X120 in order to show what people are looking at exactly. Authors use a smartphone to continuously analyze images from the front-facing camera to detect the presence of the head and eyes of the driver in the scene and alert him/her by a single beep of 1250 ms. Authors use a smartphone to track a driver's behavior, avoid distracting tasks and, thereby, increase driving performance. The data for the study describes a driver's behavior obtained with the aid of a video camera and eye-tracking system.

The study [10] considers the driver's behavior classification tasks for advanced driver assistance systems. In this research, the authors presented the neuro-fuzzy system to evaluate driving behavior based on their similarities to fuzzy patterns. The authors propose to use sensor fusion for determining types of driver maneuvers, including lane change, left/right turns, and U-turn. Accelerometer, gyroscope, and magnetometer provided initial data source for these maneuvers in a form of raw data readings, which are measurements of velocity, magnetic field, and rotation velocity, respectively. The output results of the proposed approach are two distinct scores, which are safe and aggressive driving scores. The results of this study demonstrate that the estimation of driving behavior plays an important role in increasing a driver's safety.

In other recent work [11] researchers propose an approach utilizing smartphone sensor data (acceleration in m/s2 with an accelerometer, angular velocity in rad/sec with gyroscope, and speed and vehicle position with GPS) recognizing unsafe driving styles based on a two-stage clustering approach and using the information on harsh events occurrence, acceleration profile, mobile usage, and speeding. This approach consists of the following steps: initially, clustering is applied in order to separate aggressive from non–aggressive trips; a second level clustering aided to distinguish normal trips from unsafe trips; thereby, trips were classified into six distinct groups ranked by the importance of driving safety: safe, aggressive, risky (speeding), distracted (mobile usage), aggressive/risky, and aggressive/distracted behavior. The further analysis of driver behavior in relation to the grouping of their trips indicated that drivers cannot maintain a stable driving profile through time, or, in other words, drivers behave differently every time.

One of the main causes of road accidents is drunk driving. Authors [12] propose a driver monitoring system intended for recognizing road dangerous situations and increasing driver's safety. The developed monitoring system of driver's health utilizes physiological parameters, including blood pressure, body temperature, heart rate, and other kinds of information obtained from vehicle sensors, and the ones that are integrated inside the smart band that the driver wears. Basically, the developed system includes the pressure sensor on the driver's seat, a non-contact infrared temperature sensor, and humidity sensor on the left side of the rearview mirror, a video

camera, and a MQ-3 alcohol sensor component integrated inside the steering wheel. It also includes a smart band worn by the driver that uses the photoplethysmography method focused on the reading of heartbeat information, and the ARM control board responsible for analyzing and uploading data to it. An Android-based application shows different types of the driver's health data, monitors conditions of dangerous situations and transmits information to the remote server. In the situation when a drunk or fatigue state is found, a driver receives audible alerts and reminders. The authors of the paper demonstrate low power consumption and low price of their system.

Another study [13] presents a non-invasive method for recognizing precise cues in the voice allowing to describe the state of a driver and measure sleepiness state. Experiments were conducted with the aid of patients having a suspicion of excessive daytime sleepiness who were required to read six different texts at a different time of the day. Along with it, the patients filled the Karolinska Sleepiness Scale [14] after reading texts. The audio files recorded during these sessions were divided into segments with length in a range of 50 seconds to 2 minutes. Following audio features were directly extracted from each recording to measure the patient's sleepiness: the duration of voiced parts, the percentage in the duration of voiced parts, the duration of vocalic segments, the percentage in the duration of vocalic segments. Other features were calculated on each voiced segment to characterize harmonic sounds and include descriptive values (frequency, power, bandwidth) of harmonics and formants; fundamental frequency and intensity; cepstral peak prominence; and Harmonics-to-noise ratio.

There are certain research studies leveraging cloud technologies in human behavior monitoring. One of them [15] presented a cloud-based vehicle data acquisition and analytic system intended for real-time driver behavior monitoring, trip analysis, and vehicle diagnostic. It consists of Bluetooth on-board diagnostics port, a mobile application on the driver's smartphone, and a cloud-based service. Authors use the developed complex event processor at both smartphones and the cloud platform to recognize unsafe driving and dangerous situations. Also, the system notifies a driver about such situations. Vehicle data are collected via OBD port and sent to the cloud using a smartphone connected to 3G/4G cellular network. The cloud platform is mainly responsible for recognizing reckless driving behavior based on the driver sensor data obtained from the vehicle OBD port. At the same time, it utilizes historical data to identify abnormal driving behavior. Underneath, it uses speed parameters to classify drivers into 3 groups: 20 km/h and below, 20-80 km/h, and greater than 80 km/h; and acceleration and de-acceleration counts. The developed mobile application is responsible for visualizing real-time sensor data, as well as alerting a driver about unsafe situations via textual and audible signals.

Vehicular Ad hoc Networks (VANETs) gained huge popularity due to the rapid development of mobile internet and Internet of Things applications. It utilizes dedicated short-range protocol either on-board units already integrated into some vehicles to transmit information messages between vehicles or infrastructure at a predefined rate. To provide safe driving, the authors of the study[16] proposed an intelligent Fuzzy-based Driver Monitoring System, integrating Cloud, Fog and Edge computing [17] in VANETs, and focused on the vehicle in-cabin information and driver's information to detect a potential traffic accident or a risky situation and alert the driver about the emergency.

TABLE 1. Comparison of functionality of driver monitoring systems.

Work	Video	Audio	Motion	Bio
[6]	+	–	–	+
[7]	–	+	–	–
[8]	+	–	–	–
[18]	+	–	–	–
[9]	+	–	–	–
[10]	–	–	+	–
[11]	–	–	+	–
[12]	–	–	–	+
[19]	–	–	+	–
[13]	+	–	–	–

The proposed system essentially focuses on increasing road safety and driving performance by analyzing and recognizing the driver's situation in real-time by considering different types of input driver and environment parameters, including the level of ambient noise measured in dB, heart rate of the driver calculated in bpm, respiratory rate of the driver, and ambient temperature measured in °C. In case the system indicates the situation as not safe, it may limit the vehicle's maximum speed, suggest a driver have a rest or call a doctor if he/she breathes abnormally.

Observed research studies leverage a set of different approaches providing driver and transport monitoring solutions. The comparison of the listed papers is shown in Table 1 and based on certain features, including the use of different kinds of trips and, in particular, driver-related information (video analysis obtained from the camera; motion activity tracked by position and orientation sensors, including accelerometer, gyroscope, GPS; audio signals are given by microphone; physiological activity measured through biological signals, including heart rate, blood pressure, and speech recognition support, etc.), and the use of cloud-based technologies to provide remote driver behavior analysis.

B. SPEECH RECOGNITION SYSTEMS IN VEHICLE CABIN

The paper [20] presents a method for adaptive audio classification using a smartphone. The proposed method is aimed at improving the performance of the speech recognition system in noisy conditions of the vehicle environment. The authors propose a classification algorithm based on the effective selection of features. Such feature selection helps to improve the accuracy of classification in driving conditions with various noise levels. An audio classification framework for mobile application classifies the input audio into four categories: music, speech, speech with music, and noise. The authors describe the possibilities for adjusting the framework

depending on various driving environments. The key idea of the proposed framework is the use of individual classification models generated for different driving scenarios and the adaptive application of these models based on real-time identification of the current scenario. More accurately, the proposed system contains a feature selection module for the identification of an optimal feature set and support vector machines (SVM) classification algorithm. SVM adapts to various driving environments. The paper discusses MIRToolbox framework [21] that implements features of extraction functionality. Authors consider the main 16 effective feature sets for audio classification into speech, music, and environmental sounds. Also, the paper provides a corpus of more than 420 minutes of real-world audio data (i.e. speech, music, speech with music, and noise). Audio data have been collected using a smartphone from different driving environments (i.e., local road, crowded city, and idling car).

Many vehicles currently have an Advanced Driver Assistance System (ADAS) integrated. Basically, drivers interact with the system using steering wheel controls and indicators on the dashboard. In the next recent study [22], the authors consider the use of the third modality: speech dialogue interface for ADAS. The paper presents the development of a speech dialogue system for ADAS called Adasa. The main advantage of the system is its informative features extracted from more than 9,000 conversations between drivers and the Ford customer service department. Also, they introduced an additional training dataset created by crowd workers.

The authors divide this number of conversations into 3 main groups [22]:

- division of driving responsibility between the driver and ADAS (∼50% conversations);
- interface to activate ADAS functions (∼25% conversations);
- meaning of instrument cluster iconography (∼15% conversations);
- other conversations (10%).

Adasa interface is based on the machine learning framework Lucida. Drivers access the Adasa using natural language in real-time without restrictions. The Adasa interface is accessed by pressing a button on the steering wheel, after which the driver can easily ask questions or give commands. Researchers have integrated the developed system into commercial vehicles. Also, the authors conducted a user study involving 15 drivers in real conditions. According to the authors, the accuracy of identifying user commands in the system is 77%. In addition, the user's feedbacks say about improved understanding and driving experience using ADAS features (feedback score is 8.9 / 10).

Voice control systems differ in the number of supported languages, the level of recognition of commands, the number of implemented management functions. At the same time, all these voice control systems share one thing in common - they don't work in conditions of strong external acoustic noise, which, nevertheless, is very typical for vehicles in real traffic conditions. Authors of the papers [23], [24] consider the active appearance model (AAM) based on features for multiple cameras visual and audio-visual speech recognition (VSR and AVSR) in the vehicle environment. In these papers, the authors conduct experiments on the AVICAR automotive database. Most of the existing VSR systems have been designed in controlled laboratory conditions and rarely addressed visual domain disturbance such as speaker's movement, bad illumination, low resolution, etc. Authors extend the research about multiple-camera VSR [25], [26] by applying AAM for feature extraction from visual articulators and single HMM (formed for each camera). Finally, they propose to merge four single visual stream HMMs by fusion technique [25] to form a single stream pose independent HMM. According to the researchers, AAM based VSR notably shows performance increasing across all driving conditions compared to discrete cosine transform (DCT) features baseline. For acoustic features, authors propose to use their own development system. Authors notice that multi-stream fusion of VSR with audio stream shows the improvement of multimodal over single modal speech recognition system.

The authors of the study [27] presented a smart and robust context-aware speaker recognition system for vehicle applications. Researchers proposed an algorithm for speaker detection and identification. This algorithm includes a preprocessing method using Voice Activity Detection (VAD). In addition, the presented algorithm is robust to noise and distance. The authors implemented an extensive performance analysis of the approaches that are used for speech processing in various conditions. The paper considers the effect of distance on the accuracy of speaker recognition. Moreover, researchers conducted experiments with additional noise in the environment. The study compares the performance of different classifiers with the proposed VAD algorithm. Following classifiers are used for comparison: SVM (one-against-all), SVM (one-against-one) and Gaussian Mixture Models (GMM) [27]. The performance of the proposed approach is determined by calculating the processing time required for the system to recognize a speaker. Table 2 shows the comparison of the considered systems.

As a rule, established technologies are used as the basis of the system. For video modality, AAM technology is usually used and for audio modality, HMM is widely applied. Most of the systems are designed for the English language. There are no available systems developed for the Russian language. All reviewed systems are aimed at improving speech recognition in the vehicle environment (i.e., the vehicle's cabin, in some cases a moving car). A future study needs to consider all environmental factors that affect speech recognition inside the vehicle (vehicle speed, ambient noise, camera angle, illumination, resolution, audio quality, etc.). Most of the studies consider only one modality (video or audio).

TABLE 2. Related speech recognition systems aimed for vehicle drivers.

Interface	Base Technology	Language	Problem Domain	Phrase Count	Environment	Modalities
Audio classification for Car Environment [20]	MIRToolbox for features extraction, SVM for classification	English	Speech recognition in-vehicle cabin during music or other noises	-	In-vehicle	Audio
Adasa: A Conversational In-Vehicle Digital Assistant [22]	Kaldi and GStreamer technology	English	Speech recognition in-vehicle cabin for ADAS function support	over 9000 real user verbatims (discussions)	In-vehicle	Audio
Multiple Cameras Visual Speech Recognition in Car Environment [23]	Active Appearance Model	English	Visual speech recognition in a vehicle environment	Phone number portion off all valid speaker across all driving conditions from AVICAR DB (~10 000)	In-vehicle	Video
Multiple cameras AVSR in car environment [24]	Video: Active Appearance Model Audio: Hidden Markov Model	English	Speech recognition for voice assistant for both: audio and video	Phone number portion off all valid speaker across all driving conditions from AVICAR DB (~10 000)	In-vehicle	Audio Video
Smart and Robust Speaker Recognition for Context-Aware In-Vehicle Applications [27]	SVM and GMM for classification. SmartVAD – for preprocessing	Italian	Human recognition based on speech, human number calculation, speech recognition	64 spontaneous sentences – Closed-set scenario. 80 spontaneous sentences – Open-set scenario (added unknown speakers).	In the vehicle on the road	Audio

C. MULTIMODAL CORPORA FOR AUDIO-VISUAL SPEECH RECOGNITION IN VEHICLE CABIN

A relatively large number of publicly available audio-visual databases exists in the scientific literature. They are created for different purposes and with different means. Works [28], [29] contain a comprehensive list and analysis of such databases from the audio-visual speech recognition point of view.

It is well known that driving a vehicle is accompanied by a rather active head turns from side to side. Thus, often, the driver is turned to the camera at different angles, which greatly complicates automatic lip-reading by using video information. At the same time, the presence of strong acoustic noise during driving significantly degrades the results of speech recognition by voice. Therefore, the existence of a database specially designed for such conditions is a prerequisite. Today, only three databases of audio-visual speech recorded directly in-vehicle environments are available in scientific literature, namely: AVICAR, AV@CAR, and Czech AVSC of a car driver (Table 3). While the use of other existing databases recorded in office conditions and with a frontal face capturing does not seem beneficial for the purpose of developing a reliable speech recognition system in the vehicle environment. However, since driving is often characterized by active head turns, we also examined existing databases that include audio-visual speech recordings at various angles.

They potentially could be used to create an automatic driver's speech recognition system. We examine the seven most representative multi-view datasets. Their main properties, such as the number of speakers, words, sentences, angle of rotation, video parameters are considered (Table 3). Thus, together with three datasets recorded in the car cabin, in total, we analyze 10 databases for the possibility of creating a reliable audio-visual speech recognition system in noisy traffic conditions.

AVICAR dataset [5], created by the University of Illinois, is the largest of existing audio-visual speech corpus in the car environment available. The authors collected the recordings of 100 (86 currently available) speakers by using an array of eight microphones and four video cameras integrated directly in a vehicle cabin. The dataset has recordings with five different noise conditions: idling, driving at 35 mph with closed windows, driving at 35 mph with open windows, driving at 55 mph with closed windows, and driving at 55 mph with open windows. Ten different script sets were used in the corpus and each set is for ten speakers. The vocabulary of the database consists of four categories: isolated digits, isolated letters, phone numbers, and sentences, all in English. Isolated digits are meant for automatic dialing purposes. Isolated letters are useful for automatic spelling tasks. The authors added phone numbers to tackle a connected digits issue and phonetically balanced sentences to train phoneme-based recognizers. For the equipment, the authors have three types of signal sources (array of microphones, cameras, and DTMF generator) and two types of recording devices (ADAT and MiniDV camcorder).

AV@CAR [30] is a Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. The audio part of the database is composed of seven types of recordings: clean speech (captured using a close-talk microphone), noisy speech from several microphones placed overhead the cabin, and noise only signals from the engine compartment. The video part of the database was recorded by one small video camera sensible to the visible and near-infrared bands. AV@CAR dataset consists of two main parts: the first one is collected inside a car in real driving condition and the second one is captured in a quiet environment. The recordings were made under different light conditions.

Czech AV corpus [31] of a car driver includes recordings of 12 speakers with all audio and video data captured using the single camcorder, mounted on the vehicle dashboard. This

TABLE 3. Comparison of related audio-visual speech corpora.

Corpus	Language	Context	Angle	# Speaker	Content	# Phrases	SNR	Video Parameters
AVICAR [5]	English	Driving at 35-55 mph	~0°	86	Letters (26), Digits (13), Sentences (1317)	59 000	~[15dB -> -10dB]	720 × 480 30 fps
AV@CAR [30]	Spanish	Driving, parking and studio conditions	~0°	20	Letters (26), Digits (10), Sentences (250)	7 400	vary	768 × 576 25 fps
Czech AVSC of a Car Driver [31]	Czech	Driving	~0°	12	Spelling (7-10) Words (45), Sentences (200)	>3 000	vary	720 × 576 25 fps
CUAVE [32]	English	Office	-90°, 0°, 90°	36	Digits (10)	7 000	~35-40 dB	720 × 480 30 fps
TCD-TIMIT [33]	English	Office	0°, 30°	62	Sentences (6 913)	6 913	~35-40 dB	1920 × 1080 30 fps
MV-LRS [34]	English	TV recordings	from 0° to 90°	3 783	Sentences (14 960)	74 564	~35-40 dB	160 × 160 25 fps
LILiR [35]	English	Office	0°, 30°, 45°, 60°, 90°	12	Sentences (200)	2 400	~35-40 dB	720 × 576 25 fps
CMU AVPFV [36]	English	Office	0°, 90°	10	Sentences (150)	15 000	~35-40 dB	640 × 480 30 fps
OuluVS2 [37]	English	Office	0°, 30°, 45°, 60°, 90°	53	Sentences (540)	2 120	~35-40 dB	1920 × 1080 30 fps
AV Digits [38]	English	Office	0°, 45°, 90°	92	Digits (10), Phrases (-)	-	~35-40 dB	1280 × 780 30 fps

is the first (recorded in 2003) speech dataset explicitly collected for in-vehicle audio-visual speech recognition; however, the amount and quality of data are rather low. The initial goal of collecting the corpus was the utilization of command control of car features by using audio-visual information.

Three types of data (isolated words, isolated digits, and sentences) are available in the corpus. All are recorded in real driving conditions with phrases prompted by the passenger to the driver for repeating the text.

Only these three AV speech databases recorded directly in cars are currently available in the scientific literature. Further, several multi-view speech datasets will be considered.

CUAVE [32] is one of the most cited databases for training automatic lip-reading systems. It contains recordings of 36 speakers with a large number of utterances for each. The authors record video data by both frontal and profile view of the speakers. Recorded scenarios include single and dual (conversation) speaker sessions. The inclusion of parts of simultaneous speaking is a very interesting feature of the database. It was the first document of this kind. CUAVE is designed to benefit the research in two areas: robust to a speaker's movement, audio-visual speech recognition, and distinguishing simultaneously speaking cases.

TCD-TIMIT [33] corpus is designed for continuous audio-visual speech recognition research and consists of high-quality audio and video footage. Authors record the video from two angles: straight on and at 30°. Researchers used two pairs of Sony PMW-EX3 cameras to record the database. The authors concluded that both types of information are equally important: a frontal view gives information about mouth height and width, but a profile view gives information about lip protrusion. The presented research is of great interest since a recognizer used by car drivers will have to work with angled views. Nonetheless, the authors of TCD-TMIT were unable to establish a ''sweet spot'' angle that provides the most representative information (angles between 0° and 30° were considered).

MV-LRS [34] is the largest multi-view audio-visual speech corpus that contains profile faces selected using a face pose regressor network. The authors collected the data from BBC programs (mainly news broadcasts) with different view angles, from frontal to profile. While building the multi-view model, researchers divided the data into five pose categories based on the rotation angle: the left profile, left three-quarter, frontal, right three-quarter, and right profile. MV-LRS corpus contains real-life conversational data and was made purely based on online data without any explicit recording sessions. The authors provide some experiments on frontal and profile view lip-reading. They concluded that it is possible to do profile view visual speech recognition, however, it is inferior to frontal one.

The authors of LILiR [35] collected a corpus for speaker-independent visual speech recognition. It contains recordings of five viewing angles with data captured simultaneously by five video cameras in both HD and SD formats. LILiR main dataset is yet under preparation, however, LILiR two talk corpus is already available for download. Among the main goals of creating a database, the authors also highlight facial feature tracking, comparing visual features for lip-reading, non-verbal communication, language identification, and automatic visemes identification.

Creating CMU AVPFV [36] dataset was the first attempt to make profile view lip-reading. This corpus consists of simultaneously recorded profile and frontal view audio and

video data of ten speakers. The authors used MRT (modified rhyme test) list of 150 words, commonly utilized in speech intelligibility testing.

Based on the conducted experiments, researchers prove that the profile view of lip-reading significantly outperforms frontal view lip-reading in terms of recognition accuracy. However, this statement does not coincide with the results obtained in [34], where, on the contrary, the accuracy of frontal view speech recognition is higher. The authors present the OuluVS2 [37] multi-view dataset for non-rigid mouth motion analysis. It contains recordings from five different viewing angles. Among 53 speakers who participated in the recordings, 40 were males and 13 were females.

An interesting feature is that there were no native English speakers among them. Most of the participants were university students and were grouped into five different appearance types: European, Chinese, Indian/Pakistani, Arabian, and African. The recordings were made in ordinary office conditions by using 6 cameras: facing HD camera and high-speed camera (0°), with other four HD cameras located in the following positions: 30°, 45°, 60°, and 90°. According to the authors, the best recognition performance was achieved from the camera installed at 60° and the worst from the camera installed at 45°. Initially, AV Digits [38] database was recorded for visual-only recognition of normal, whispered, and silent speech. It contains two parts: digits and short phrases. The database was recorded in a lab environment using three cameras. The three cameras record three different views of the participant's face. The digits and phrases were displayed on a laptop screen in front of the speaker. The database with annotation and transcription is publicly available. The authors reported significant differences in lip movements, according to the type of speech: normal, whispered, and silent. In particular, the silent speech recognition accuracy suffers most, when trained on the normal speech data.

We select several parameters for comparison, in our opinion, the most representative ones. Regarding the language, English dominates in the number of databases (8 out of 10 recorded in English, 1 in Spanish, and 1 in Czech). Regarding the context of recordings, 3 databases were collected in real driving conditions, 6 in control office conditions, and 1 (MV-LRS) was artificially made based on TV data. Angles of recordings vary in a wide range from near frontal view 0° to profile view 90°. The number of speakers started from ten in the CMU AVPFV dataset and is reaching a tremendous 3 783 in MV-LRS. The number of phrases in each dataset also changes dramatically depending on the task and considered scenario. E.g. isolated digits recognition and continuous speech recognition tasks have different demands on the quality and amount of data. As well as the video resolution and fps, with the lowest on MV-LRS (160 × 160, 25 fps) and highest on TCD-TIMIT (1920 × 1080, 30 fps).

Based on the conducted analysis we conclude that the number of currently available databases suitable for training recognition systems in-vehicle conditions is not sufficient (at the moment three databases exist: AVICAR, AV@CAR, Czech AV). Especially, minding other languages besides English. Even for English, the best available at present AVICAR dataset has a number of drawbacks - restricted vocabulary, low quality of video data. Furthermore, few available multi-view corpora are also not beneficial for this purpose since they are mostly recorded in a controlled office environment with artificial lighting. Thus, to create a driver's audio-visual Russian speech recognition system, it is necessary to collect our own database that corresponds to the actual conditions of use.

III. METHODOLOGY

In the previous section, we analyzed existing audio-visual speech corpora. We substantiated that creation of our own audio-visual speech corpus for the Russian language in-the-wild vehicle environment is an actual and important task. In turn, in this section, we describe the necessary steps that must be performed to record such a speech corpus, present resulting speech recognition vocabulary, and develop software for recording the corpus. Based on the conducted analysis of the research field and existing audio-visual datasets in-the-wild the vehicle environment, we propose a novel task-oriented methodology for speech corpus creation (see Fig. 1). The methodology is universal and can be used to record the corpus for any language. In the general case, the proposed methodology includes the sequential execution of the theoretical and practical steps.

A. ASSESSMENT OF THE SPEECH RECOGNITION USE-CASES

One of the most important steps in the creation of a speech corpus, since the definition of the recognition scenario, has a huge impact on the complexity and size of the corpus. E.g., ''continuous speech recognition'', ''keyword recognition'', or ''isolated letter/digit recognition'' tasks require fundamentally different approaches to forming vocabulary and collecting the data. In addition, the language itself also has a significant impact. Even for the same recognition task, the amount of data required for analytical English and inflectional Russian can differ significantly due to the linguistic and phonetic features of the languages.

B. CORPUS METAPARAMETERS

We divide the speech corpus metaparameters into two main groups: (1) input data parameters, i.e. audio sampling rate, video resolution, video frame rate, etc., which have to be predefined in advance based on the target recognition scenario and operating conditions, and (2) corpus structure parameters, where the size of the recognition vocabulary, the number of speakers, and the number of required repetitions per phrase are the main ones to be determined.

They are specified based on the selected recognition scenario, type of desired recognition system (speaker-dependent or speaker-independent), and the selected method of modeling audio-visual signal.

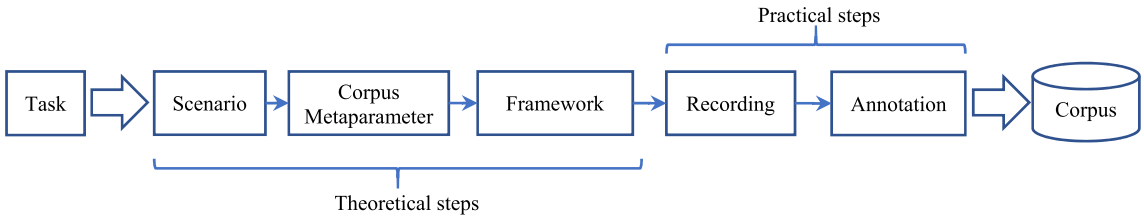

FIGURE 1. Corpus design methodology.

C. DEVELOPMENT OF THE CORPUS RECORDING FRAMEWORK

Firstly, we determine the basic requirements for the number of speakers, vocabulary size, quality of the data, etc. Then we proceed to framework development for the recording of the desired corpus. Audio-visual speech data synchronization and fusion from different sensors have to be performed. In the current research, we present the developed software for audio-visual speech corpus recording designed to work in a vehicle cabin environment.

D. CORPUS RECORDING

This step includes building a setup in accordance with the developed corpus recording framework, searching, and selecting native speakers with no speaking or hearing problems for participation in the recordings, and managing the recording process.

E. CORPUS ANNOTATION

Automatic, semi-automatic, or manual segmentation and labeling of the collected data. Organizing all the recordings into a logically structured database that comprises information about all the speakers and recording parameters, as well as a number of text, audio, and video files with temporal annotation for each recorded speaker.

IV. USE-CASES SCENARIO AND CORPUS STRUCTURE

We propose human-computer interaction that the driver monitoring system should support both: to interact with the driver in normal mode and to interact with the driver to prevent sleeping when the system detects the dangerous drowsiness state. For this purpose, we consider a vocabulary in the section that the corpus should support to meet the presented idea and describe the software prototype developed to create the corpus.

A. USE-CASES FOR SPEECH RECOGNITION IN DRIVER MONITORING SYSTEMS

We identify use-cases for human-computer interaction that requires audio and video-based speech recognition. We identify the main commands that the system should support to interact with the driver. The driver monitoring system allows detecting the drowsy driver behavior. Such behavior says that a driver can fall asleep in the nearest future. To prevent this situation, we propose question-answer games that allow a driver to distract from sleeping and drive the vehicle more safely.

Based on our previous work [39] we identify the following main commands that the system should support to interact with the driver in normal mode (see Table 4). In the training, the mode system asks if the driver is drowsy and distracted. When a user answers these questions, the system reads how the driver acts and looks in different conditions. The system learns on this data to predict dangerous situations in the future.

In the normal mode in case, the system automatically recognizes dangerous drowsiness situations (based on the algorithm proposed in [40]) it asks the driver about his/her condition. If the driver approves his/her dangerous drowsiness state, the system asks if a user wants to change a route or have a break. If the driver is driving on a highway and the rest of the trip is more than one hour, the system recommends staying at a hotel or have some caffeinated beverage. If a driver is driving in a city, the system recommends taking nap. If the rest of the trip is less than one hour, the system recommends playing a game or listen to music.

The system recognizes commands, listed in Table 9. Commands "Yes" and "No" are needed in both training and normal mode to receive an answer about drowsiness. The command "Change route" is needed in normal mode to inform the system of the selection of the proposed recommendations to make a stop. The commands "Turn on music" and "Play" are needed to inform the system of the selection to listen to music and to play the game.

We propose games to prevent a driver from sleeping in case the driver monitoring system detects drowsiness dangerous state. The authors of the research [41] clarified that playing games improve the drowsy driver state. This is a consequence of the fact that when playing word games, the driver actually engages in a dialogue with the passenger. According to various organizations, maintaining a non-monotonous dialogue with the passenger allows to delay falling asleep [42]. In the paper [43] authors clarify that music improves drivers' response times to accelerations and decelerations of a lead vehicle while following a car. In the paper experiments with the 47 participants have been considered.

The authors of the paper [44] showed based on the experiments in a simulator that verbal communication with a media device showed improved lane-keeping performance and had improvements in neurophysiological measures of alertness.

TABLE 4. Examples of the questions the system generates for the driver.

Question for a driver	Context situation	Commands
Are you drowsy?	Training mode	Yes / No
Are you distracted?	Training mode	Yes / No
Are you drowsy? Would you like to have a cup of coffee in a cafe/gas station (say change route)?	Drowsiness dangerous states have been determined several times, rest of the driver trip more than one hour, a cafe or gas station is found nearby.	Yes / No / Change route
Are you drowsy? Would you like to play with the system?	Drowsiness dangerous states have been determined several times.	Yes / No / Play
Are you drowsy? Would you like to listen to music?	Drowsiness dangerous states have been determined several times, the rest of the driver trip less than one hour.	Yes / No / Turn on / off the music
Are you drowsy? Would you like to take a nap?	Drowsiness dangerous states have been determined several times, driving in a city.	Yes / No / Change route

It can be clarified that the drowsiness level of the driver decreases.

Each of the games is a word-based game that requires a speech interface. In this case, the driver will not be distracted from the road. The games are distinguished by their complexity and, therefore, the load on the user's brain for concentration. The easiest game is ''rock-paper-scissors'' [45]. It allows a driver to slightly strain the brain and not particularly think about the answer.

We divide the drivers by the following: age (old, middle adulthood, young adulthood, young) and sex (women, man).

The Mid-core game is presented by the ''hangman'' game. It is harder and allows to strain the user's brain more, but it is still not hard, and it does not require analytics [46]. The hardest game is ''more-or-less'', which requires analytics and helps to concentrate as much, as possible by the word games.

We analyze a driver's portrait to make a decision on what type of game better fits a driver and a situation based on some statistical information. According to statistics, men and young people are more prone to analytical thinking than women and old people [47]. After receiving the information about a user's portrait and the situation context, the system counts points (presented in Table 5 and Table 6).

Due to the fact, that the context is more important than the decision, the total number of points will be calculated by the formula, where the context points will have a coefficient of 2. It allows the system to select the type of game relying primarily on the context. Each user can play any of the available games, and analytical thinking only adjusts the type of the game in some cases, when several games fit the same context. The game is selected according to the following formula:

$$P = P_p + 2P_c, \tag{1}$$

where P_p– portrait points (calculated according to Table 5), P_c– context points (calculated according to Table 6), P – total points to make a decision which game is more applicable for the driver in the current situation.

The game is selected by the total number of points and each game corresponds to a certain number of points. ''Rock-Paper-Scissors'' corresponds to points 0–7, ''Hangman'' corresponds to points ''8–15'', ''More-Or-Less'' corresponds to points ''16–22'' (see Table 7).

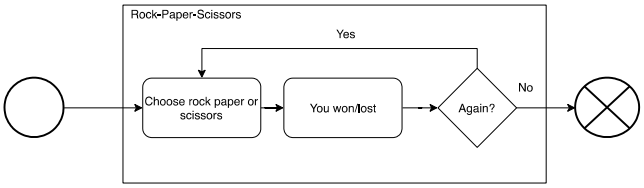

FIGURE 2. Rock-paper-scissors game.

After several sessions the system decides which game has a higher priority for the driver: it counts the average number of replays for each game and the percentage of success in ''hangman'' and ''More-Or-Less'' games. After receiving this data, the system changes the portrait points value with the statistic point value. The driver gets 1, 2, or 3 points if he prefers one of the games. It shows if the user prefers harder games, so he/she needs more analytical games to become cheerful. The system receives this information by comparing the number of replays. Also, a user earns 1, 2, or 3 points for the win rate of the ''hangman'' and ''more-or-less'' games, because they represent mid-core and hardcore word games.

B. VOCABULARY

Rock-paper-scissors is the most casual game, that does not strain the brain so much. This game uses small vocabulary and does not demand the resources or the level of training of the player. It fits users that do not need to increase the concentration level. Fig. 2 describes the game process within the system. The game takes place in the following way. The system suggests choosing 1 of 3 options: paper, rock, or scissors. After receiving an answer from the driver, the system randomly chooses 1 of 3 options. Then the system matches a user's and the system's values to determine if the user won or lost. The system compares values according to the rules of the game: rock defeats scissors but loses to paper. Scissors defeat paper but lose to rock. Paper defeats rock but loses to scissors.

We summarize all the words the system should recognize and pronounce to play the rock-paper-scissors game (see

TABLE 5. Human classification of synthetic or analytical mind based on statistics.

Criteria	Strong synthetic mind (0 points)	Synthetic mind (1 point)	Analytical mind (2 points)	Strong analytical mind (3 points)
Age	Old	Middle adulthood	Young adulthood	Young
Sex	Women			Man

TABLE 6. Human context related to game difficulty.

Criteria	Casual game (0 points)	Mid-core game (1 point)	Hard-core game (2 points)
Time	Day	Morning, Evening	Night
Road type	City		Highway
Traffic	Intensive	Average	Non-intensive
Speed	High speed	Low speed	Middle speed

TABLE 7. Game choice based on matchpoints.

Criteria	Rock-Paper-scissors	Hangman	More-Or-Less
Points	0-7	8-15	16-22

TABLE 8. Vocabulary for rock-paper-scissors question/answer game.

Synthesized Words	Recognized Words
Choose	Rock
or	Paper
Rock	Scissors
Paper	Yes
Scissors	No
You	
Won	
Lost	

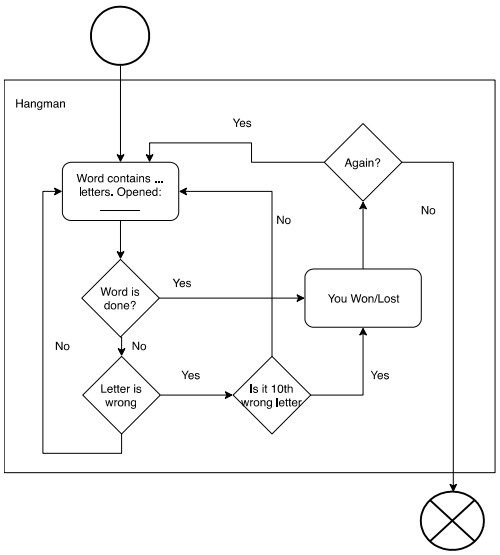

FIGURE 3. Hangman game.

Table 8). Words ''Yes'' and ''No'' are used to answer the question about replaying the game after finishing. Words ''Rock'', ''Paper'', and ''Scissors'' are recognized to understand what answer the driver chose.

TABLE 9. Vocabulary for hangman question/answer game.

Synthesized Words	Recognized Words
Word	33 Russian letters
Contains	Yes
Letters	No
Right	
Wrong	
Letter	
You	
Lost	
Won	
Letters (A-Z)	
Numbers (3-10)	

Hangman is a guessing-word game (see Fig. 3). The driver has 10 attempts to guess a word. This game requires a bigger user's engagement and it is good enough for a medium brain load, which promotes concentration. This game requires vocabulary and letter-detection. The game takes place in the following way.

The system suggests guessing the word. The word is selected from a pre-prepared vocabulary and consists of 3 to 10 letters. Then the system suggests guessing a letter from the selected word. If the word contains a letter that the player named, the system opens this letter in the word and suggests guessing the rest. If a player calls a letter that is not in the word, the system reports this and predicts a different letter. If a player names a wrong letter 10 times the player loses. If the player names all the letters in the word, the player wins. Vocabulary for the Hangman game is presented in Table 9.

The last game (more-or-less) requires maximum concentration from a driver. The system asks a question, where an answer is a number (see Fig. 4). The driver needs to think of the question and guess, after what he/she gets a hint if the right answer is more or less. The user has 5 attempts before he/she lost. This game effectively loads the brain, which helps to concentrate on a road. The game works as follows. The

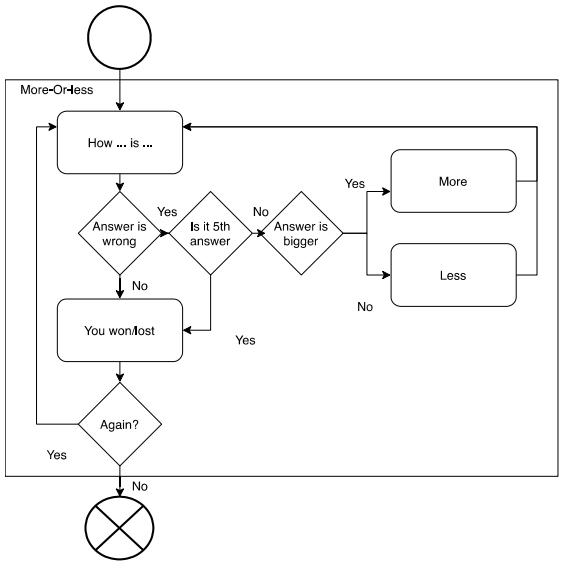

FIGURE 4. More-Or-Less game.

TABLE 10. Vocabulary for More-Or-Less question/answer game.

System should generate	System should recognize
How	Numbers (1-10)
Long	Thousand
Big	Million
More	Yes
Less	No
You	Hundreds
Won	Dozens
Lost	
Much	
Large	
Costs	
is	
River	
Lake	
Mountain	
Road	
Million	
Thousand	
Hundred	
Numbers(1-100)	
Pool of names[]	

system asks a user a question about some characteristics of some object. For example, a system might ask ''what size is Lake Baikal'' or ''how much is a Boeing 747?''.

After this question, the user might guess the answer and pronounce it to the system. The system compares the answer and the user's assumption and says if the answer is more or less. If the user gives the correct answer, he wins. If the user gives 5 wrong answers, he/she loses the game and the system shows the right answer.

Table 10 shows all the words the system should recognize and pronounce to play a more-or-less game. The system recognizes the numbers from 1 to 100 and the words ''thousand'', ''million'', ''hundred'' to get an answer from a driver. The system pronounces sentences from the words, listed in the table to construct questions for the driver. To fill the database with facts connected with the numbers the system uses Numbers API [48]. PHP-script gets facts with random numbers, generates a text file, after which the file is manually checked and cleared of the facts that do not fit the system.

C. CORPUS METAPARAMETERS

We have identified the following parameters the corpus should support according to the presented speech recognition scenario in the vehicle cabin.

- Realistic recording conditions. Since our task is to create a real-life recognition system and integrate it into the driver monitoring system, we need training data to be as close as possible to the actual operating conditions. This includes both the variability of SNR in a wide range and the variability of lighting conditions. The data should be recorded in the vehicle cabin, and not imitated in artificial conditions.
- The multiple number of speakers. To create a speaker-independent recognition system, it is necessary to record a large number of speakers (not less than a few dozens). Each of them should be recorded in all target scenarios.
- Size of the vocabulary. The size of the vocabulary is heavily dependent on the recognition task being solved. In the current research, we are primarily aimed at solving the problem of recognizing individual commands of the user, rather than continuous speech. Thus, a small-sized vocabulary of 140 words/short phrases (section 3.2) is suitable for our goal.
- Quality of the data. The quality of the recorded data, namely the video resolution and the audio sampling rate, has a significant impact on the resulting recognition accuracy. Therefore, given the complexity of the conditions of use, we need to have as high-quality data as possible: video resolution no less than 720 $\times$ 480, and audio sampling rate no less than 44.1 kHz. For the video part, it is also important that the face occupies at least 40% of the image, otherwise, the resolution of the video should be increased to at least 1280 $\times$ 720 pixels.
- Multi-view video data. We need to consider the specifics of driving conditions when the driver often has to make active head turns.

This fact greatly complicates the task of automated lip-reading and imposes a need to have the data recorded at different angles for reliable speech recognition. In the current research, we decided to use two cameras: one facing camera with a near-frontal view of the driver's face and one camera located at 30° to the driver's right-hand side.

V. CORPUS CREATION

A. FRAMEWORK

We consider the driver as a main source of information for the proposed driver behavior monitoring system. In order to conduct experiments and collect sensor data, obtained in natural driving conditions, we proposed a reference model

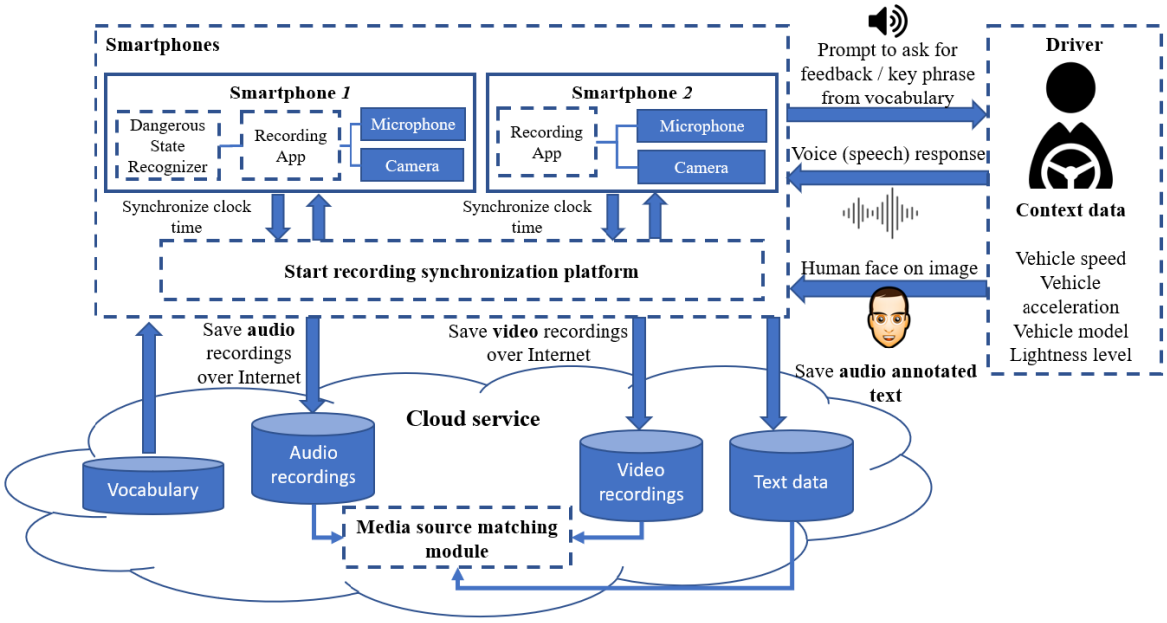

FIGURE 5. Reference model of the corpus creation framework.

for recording audio signals and visual-based information obtained from several smartphones, equipped with video-based camera and microphone, in a synchronous way that each obtained audio fragment matches a video frame (see Fig. 5). This approach consists of the following three components: smartphones, cloud service, and a driver.

The smartphone is installed inside the vehicle cabin in a way that its camera successfully captures and tracks a driver's face and position along the trip. Smartphones are responsible for recording audio signals, capturing human speech and noise level, obtained by a microphone sensor inside the vehicle cabin. This kind of audio information is passed through an Android-based automatic speech recognizer provided by the Google platform to extract words spoken by the driver.

We tackle the problem of synchronizing two smartphones (which can be potentially applied to a greater number of mobile devices) by introducing a start recording synchronization platform.

As soon as smartphones collect this kind of audio and video-based information, clients transmit it to cloud storage for further processing and analysis.

Cloud service is comprised of three main data storages, responsible for collecting audio recordings, video recordings, and text data, describing the media data, and a media source matching module. The latter component aids to join and match the collected data properly, process, and analyze it as a single source of collected data.

The first smartphone (main) is essentially responsible for establishing a connection with another smartphone (secondary) and employing a driver's monitoring. It runs dangerous state recognition by analyzing video and audio streams, obtained from a camera and a microphone of the smartphone,

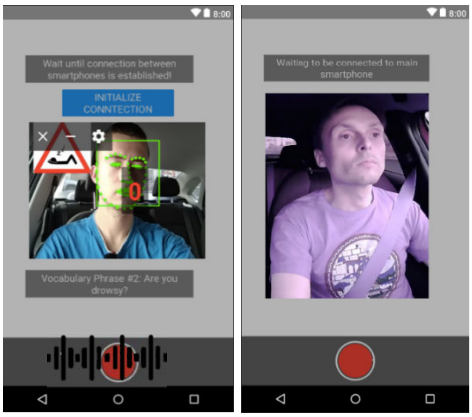

FIGURE 6. a) (left) presents a screenshot of mobile application taken from the main smartphone, and b) (right) shows a screenshot of the application acquired from the secondary smartphone.

respectively. The application runs drowsiness detection algorithms to recognize dangerous situations [40].

Also, it is responsible for audio-based interaction with a driver, utilizing the smartphones' microphone. The application plays an audio clip requesting a driver to repeat a certain audio phrase. The system will record the time when the phrases were generated and will save the details about the asked question and the answer in the SQLite database for further analysis and processing. The current phrase and its number are visible on the screen of this application (Fig 6, a).

As soon as the main smartphone initializes a connection with a secondary one, the latter starts to record video and audio signals produced inside the vehicle cabin to be later processed. Video recordings are saved on the external memory

FIGURE 7. Audio-visual corpus recording in vehicle cabin using two smartphones.

card of the smartphone. We accumulate audio and video data in smartphone SD card and send them to the cloud service as soon as the Internet connection is available. The screenshot of the application, responsible for recording media data, is shown in Fig. 6, b).

Both utilized smartphones are placed in the vehicle cabin (see Fig. 7). The first (main) smartphone is placed near the driver to the left of his view.

The second smartphone is located at an angle approximately 20 degrees to the right. The first smartphone performs a driver's monitoring using the front-facing camera directed to the driver's face, while the other (second) one is mainly focused on recording video and audio information captured by the smartphone's front-facing camera. We have chosen the smallest smartphones for data recording due to safety reasons. We do not place the smartphone in front of the driver. Such smartphone locations are popular for the drivers to set smartphones with the navigation system for the vehicles that do not have built-in.

VI. MULTIMODAL CORPUS RUSAVIC

RUSAVIC corpus contains the following data (see Fig. 8) recorded in-the-wild: audio and video data of the driver recorded from different angles and RUSAVIC database that includes sensor data from the main smartphone and driver condition detected by the driver monitoring system installed in the main smartphone [49]. RUSAVIC database includes the following information (see Table 12): date, time, vehicle speed, vehicle acceleration, noise level, and dangerous states detected by the driver monitoring system. Driver condition includes yaw and pitch head angles [50], eyes state, mouth state, PERCLOS (PERsantage of eyes CLOSure) [51], and detected dangerous state. So, the driver monitoring system provides possibilities to recognize the following states: drowsiness, distraction, belt unfasteness, smartphone usage, eating/drinking, and smoking.

The state of drowsiness characterizes the driver fatigue level as high which causes longer voiced duration as well as response time [52]. The state of distraction clarifies that driver attention is not concentrated on the road in contrast can cause faster speech but also longer response time. Other dangerous states also can cause the drive voice change.

We propose to consider the mentioned parameter as a context for the driver. Every driver trip is related to the audio-video file as well as mentioned parameters that provides possibilities for experiments with speech recognition in different surroundings. E.g. noisy environment vs silent one, drowsy driver, and the normal one.

TABLE 11. List of phrases in RUSAVIC corpus.

Original Phrase	English Translation
Покажи весь маршрут	Show the entire route
Сбрось маршрут	Reset route
Поехали на заправку	Let's go to the gas station
Сколько мне еще ехать?	How long do I have to go?
Во сколько я приеду?	What time will I arrive?
Как там на дорогах?	How's traffic?
Какие сейчас пробки?	What are the traffic jams now?
Надолго пробка?	Long traffic jam?
Найди заправку	Find a gas station
Улица {A} дом {1}	Street {A} house {1}
Поехали домой	Let's go home
Где находится ближайшее..?	Where is the nearest..?
Сколько стоит бензин сейчас?	How much is gasoline now?
Включи музыку	Turn on the music
Включи радио	Turn on the radio
Включи аудиокнигу	Play audiobook
Выключи музыку	Turn off the music
Заткнись	Shut up
Позвони по номеру	Call the number
Отправь сообщение	Send message
Отправь e-mail	Send e-mail
Ответь на звонок	Answer the call
Сбрось вызов	Drop the call
Увеличь яркость	Increase brightness
Уменьши яркость	Decrease brightness
Увеличь громкость	Turn up the volume
Уменьши громкость	Decrease the volume
Сколько градусов за окном?	How many degrees are outside?
Проложи маршрут	Get directions to
Избегай сборов	Avoid Fees
Что это за улица?	What is this street?
Найди ресторан поблизости	Find a restaurant nearby
Поставь таймер	Set a timer
Включи Wi-Fi	Turn on Wi-Fi
Включи мой плей лист	Play my playlist
Покажи парковку	Show me parking
Найди парковку	Find parking
Впереди авария	Accident ahead
Покажи мой календарь	Show my calendar
Перезагрузись	Reboot
Включи массаж спины	Turn on back massage
Смягчи подвеску	Soften the suspension
Выключи предупреждения о сходе с полосы	Turn off lane departure warnings
Включи свет в салоне	Turn on the interior light
Выключи свет в салоне	Turn off the interior light
Включи кондиционер	Turn on the air conditioner
Выключи кондиционер	Turn off the air conditioner
Включи печку	Turn on the heat
Выключи печку	Turn off the heat
Включи сигнализацию	Turn on the alarm

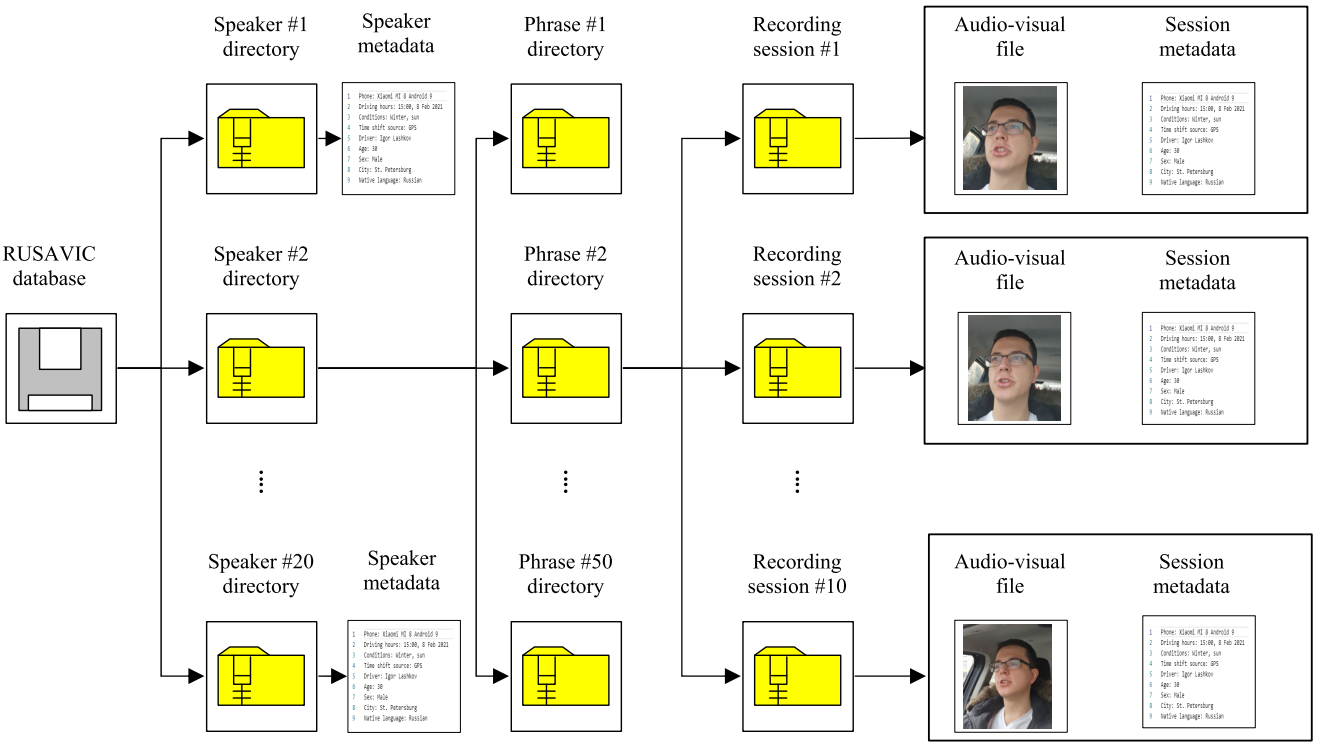

FIGURE 8. The logical structure of the RUSAVIC audio-visual speech database.

TABLE 12. RUSAVIC database description.

Driver	Link to AV file	Date & Time	Head angle (yaw, pitch)	Eyes state	Mouth state	PERCLOS	Speed (km/h)	Accele-ration (m/s^2)	Noise Level (dB)	Dangerous state
Driver 1	C:\...\file1	31.01.21 15:30	-12, 8	closed	closed	0.10	66	1.12	0.15	no belt
Driver 1	C:\...\file1	31.01.21 15:31	-2, 12	opened	closed	0.05	64	1.4	0.18	smoking
Driver 1	C:\...\file1	31.01.21 18:32	-9, 10	closed	closed	0.01	50	0.96	0.22	eating
Driver 1	C:\...\file2	31.01.21 18:33	-2, 5	closed	closed	0.05	55	1.01	0.14	fine
Driver 1	C:\...\file2	31.01.21 18:34	-1, 14	opened	opened	0.49	42	1.09	0.99	drowsiness
Driver 1	C:\...\file2	31.01.21 18:35	2, 12	closed	opened	0.50	39	1.03	0.95	drowsiness
Driver 2	C:\...\file1	08.02.21 10:00	-1, 3	closed	closed	0.11	36	0.40	0.15	fine
Driver 2	C:\...\file1	08.02.21 10:01	-25, 43	opened	closed	0.10	38	-1.32	0.22	distraction
Driver 2	C:\...\file1	08.02.21 10:02	-5, -2	closed	opened	0.30	26	-0.49	0.18	drowsiness
Driver 2	C:\...\file2	08.02.21 11:35	-7, 3	closed	opened	0.25	25	-0.45	0.17	no belt
Driver 2	C:\...\file2	08.02.21 11:36	-8, 4	closed	opened	0.10	15	-0.3	0.15	drowsiness

Table 13 shows the main parametrical characteristics of the created corpus. We have recorded 20 speakers each of them pronounce 50 phrases. We have repeated 10 sessions for each of the speaker. Every session has been recorded in own condition in-the-wild vehicle cabin (different speed, different acceleration, different nise level and driver condition). For the recordings we have used two smartphones mounted in the vehicle windshield (angles: $-20°$, $20°$). Since we record the corpus in the real conditions we got the difference signal to noise ratio: from 30 to 5. Examples of the video can be found in the RUSAVIC portal.[1] Context information is stored in PostgreSQL (general structure is presented in Table 12).

[1] https://mobiledrivesafely.com/corpus-rusavic

The logical structure of the RUSAVIC database is shown in Fig. 9. The database contains the number of directories equal to the number of speakers (20). Then, the directory of each speaker includes subdirectories equal to the number of uttered phrases (50 according to our dictionary) as well as metadata file that describes the speaker:

- Driver: Alexey Kashevnik;
- Vehicle: Geely Atlas;
- Age: 38;
- Sex: Male;
- City: St. Petersburg;
- Native language: Russian.

Each phrase subdirectory includes the number of folders according to the number of recording sessions per

TABLE 13. RUSAVIC corpus description.

Parameter	Value
The total amount of speakers	20
Phrases per speaker	50
Recording sessions per speaker	10
Recording angles	~ (-20°, 20°)
Amount of data per speaker (GB)	~ 10
Video resolution (pixels)	1920 × 1080
Video recording speed (FPS)	60
Audio sampling frequency (kHz)	48
Number of audio channels	2 (stereo)
Signal to noise ratio (dB)	Vary ~ (30 – 5)
The average age of the speakers (years)	26

speaker (8-12). Thus, at the bottom of our structure we have audio-visual files of the same speaker uttering the same phrase during different recording conditions. Recording sessions metadata files are also included. It contains such information as device description, driving hours, recording conditions, driver rotation angle, etc.

VII. DISCUSSION

In this section we summarize answers to research questions we specified in the introduction and studied in the paper. We remind research questions step by step and provide research answers (RA) got from the analysis.

RQ1: What modern technologies are used for audio-visual speech recognition?

RA1: According to the conducted analysis, the state-of-the-art methodology to tackle the problem of audio-visual speech recognition in a vehicle environment is usually based either on discrete cosine transform (DCT) coefficients-based or on active appearance models-based (AAM) features extraction, followed by Hidden Markov models (HMM), used for the classification. However, such solutions in some cases are ineffective, since they do not meet the requirements of practical systems of audio-visual speech recognition due to their inability to take into account many factors at the same time (various noisy conditions or occlusions in the video data). Nevertheless, the collection of a large amount of data, as well as the availability of computing resources, nowadays allows the researchers to use various architectures of neural networks to extract informative features, train models, and develop reliable audio-visual speech recognition systems. Using modern LSTM and CNN neural network architectures, both short-term and long-term spatio-temporal characteristics of audio-visual speech can be extracted. For our research, we plan to use different topologies of these neural networks. In turn, the spatial pyramid pooling allows us to normalize the formed spatio-temporal features for the subsequent hypothesis. In this case, the audio and video modalities are trained separately from each other and are combined only at the level of hypotheses to decide on the speech uttered by the speaker.

RQ2: Which parameters are important to corpus creation for audio-visual speech recognition in a vehicle cabin?

The main purpose of recording the database is to train a reliable and robust speech recognition system that is able to recognize a limited set of user's commands in noisy driving conditions. Based on the conducted research, we have identified corpus parameters specific for the considered scenario of speed recognition in the vehicle cabin.

Thus, having determined the basic requirements to the recording conditions, the number of speakers, vocabulary size, and quality of the data we proposed the reference model and developed a prototype of the mobile application for audio-visual corpus creation with the vocabulary presented in Section 3.2.

RQ3: How the dialog-based interaction with the drowsy driver can be used in the driver monitoring system to avoid sleepiness?

Dialogue-based games can help the driver stay awake. The advantages of this method are that the driver can play games at ease, which distracts him from sleep. In addition, games such as hangman or more-or-less develop analytical thinking and vocabulary. On the other hand, games can distract the driver from the road, which can lead to a dangerous situation. But all the negative aspects of using this system overlap with positive ones. In reality, the consequences of falling asleep are much more dangerous than the consequences of possible distraction from the road, so the risks are justified.

RQ4: Which vocabulary should be supported for corpus creation?

We analyze the main commands that the driver monitoring system should recognize to prevent the driver's distraction during the vehicle driving. At the same time, we propose question/answer games that the human-computer interaction interface can launch in the case of drowsiness dangerous state detection to prevent the sleeping condition. Games are distinguished by their complexity and, therefore, the load on the user's brain for concentration. We propose rock-paper-scissors, hangman, and more-or-less games. Our analysis shows that to support such functionality the system should recognize the following vocabulary: yes; no; 33 Russian letters; numbers (1-10); thousand; million; hundreds; dozens; and some special words (rock, paper, scissors, change the route, play, turn on/off music).

Recognition of the following vocabulary allows developing the human-computer interface that supports the presented in the paper functionality.

VIII. CONCLUSION

In this paper, we consider the problem of efficient audio-visual speech recognition for driver monitoring systems. We analyze related research works in the following topics: driver monitoring systems based on smartphone sensors, speech recognition systems in vehicle cabins, and multimodal corpus for audio-visual speech recognition in a vehicle cabin. We develop the task-oriented methodology for speech corpus creation and implement that was a goal of this paper. We formulate four research questions related to the task of collecting a representative audio-visual speech corpus. We identify the

main specifics of speech recognition for driver monitoring systems, research the main state-of-the-art techniques for audio-visual speech recognition, discuss important parameters for audio-visual corpus creation related to driver's speech recognition and develop our own recognition vocabulary for further corpus recording. Finally, we describe the mobile application developed for the corpus creation and record RUSAVIC corpus that includes 20 participants in-the-wild vehicle conditions.

Our future work is related to new algorithm development for effective speech recognition in a vehicle cabin based on RUSAVIC corpus presented in the paper.

REFERENCES

[1] A. Kashevnik, I. Lashkov, A. Ponomarev, N. Teslya, and A. Gurtov, "Cloud-based driver monitoring system using a smartphone," *IEEE Sensors J.*, vol. 20, no. 12, pp. 6701–6715, Feb. 2020, doi: 10.1109/jsen.2020.2975382.

[2] J. Kim, K. Sato, N. Hashimoto, A. Kashevnik, K. Tomita, S. Miyakoshi, Y. Takinami, O. Matsumoto, and A. Boyali, "Context-based rider assistant system for two wheeled self-balancing vehicles," *Inform. Automat.*, vol. 18, no. 3, pp. 583–614, Jun. 2019, doi: 10.15622/sp.2019.18.3.582-613.

[3] I. S. Kipyatkova and A. A. Karpov, "Variants of deep artificial neural networks for speech recognition systems," *Trudy SPIIRAS*, vol. 6, no. 49, pp. 80–103, Dec. 2016, doi: 10.15622/sp.49.5.

[4] O. Verkholyak, H. Kaya, and A. Karpov, "Modeling short-term and long-term dependencies of the speech signal for paralinguistic emotion classification," *Inform. Automat.*, vol. 18, no. 1, pp. 30–56, Feb. 2019, doi: 10.15622/sp.18.1.30-56.

[5] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, and T. Huang, "AVICAR: Audio-visual speech corpus in a car environment," in *Proc. 8th Int. Conf. Spoken Lang. Process. (ICSLP)*, 2004, pp. 1–4.

[6] C. J. de Naurois, C. Bourdin, A. Stratulat, E. Diaz, and J.-L. Vercher, "Detection and prediction of driver drowsiness using artificial neural network models," *Accident Anal. Prevention*, vol. 126, pp. 95–104, May 2019, doi: 10.1016/j.aap.2017.11.038.

[7] Y. Xie, F. Li, Y. Wu, S. Yang, and Y. Wang, "Real-time detection for drowsy driving via acoustic sensing on smartphones," *IEEE Trans. Mobile Comput.*, early access, Apr. 2, 2020, doi: 10.1109/tmc.2020.2984278.

[8] A. U. Nambi, S. Bannur, I. Mehta, H. Kalra, A. Virmani, V. N. Padmanabhan, R. Bhandari, and B. Raman, "Hams: Driver and driving monitoring using a smartphone," in *Proc. 24th Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 840–842, doi: 10.1145/3241539.3267723.

[9] A. I. Dumitru, T. Girbacia, R. G. Boboc, C.-C. Postelnicu, and G.-L. Mogan, "Effects of smartphone based advanced driver assistance system on distracted driving behavior: A simulator study," *Comput. Hum. Behav.*, vol. 83, pp. 1–7, Jun. 2018, doi: 10.1016/J.CHB.2018.01.011.

[10] H. R. Eftekhari and M. Ghatee, "A similarity-based neuro-fuzzy modeling for driving behavior recognition applying fusion of smartphone sensors," *J. Intell. Transp. Syst.*, vol. 23, no. 1, pp. 72–83, Jan. 2019, doi: 10.1080/15472450.2018.1506338.

[11] E. G. Mantouka, E. N. Barmpounakis, and E. I. Vlahogianni, "Identifying driving safety profiles from smartphone data using unsupervised learning," *Saf. Sci.*, vol. 119, pp. 84–90, Nov. 2019, doi: 10.1016/j.ssci.2019.01.025.

[12] X. Shen, J.-Y. Chen, X.-C. Lei, L. Huang, L.-S. Wu, and W.-Q. Yin, "Smart safety monitoring system for vehicles," in *Proc. 2nd Int. Conf. Big Data Internet Things (BDIOT)*, 2018, pp. 84–87, doi: 10.1145/3289430.3289467.

[13] V. P. Martin, "Towards automatic sleepiness measurement through speech," Univ. Bordeaux, Bordeaux, France, Tech. Rep. hal-02145255, Jun. 2019.

[14] A. Shahid, K. Wilkinson, S. Marcu, and C. M. Shapiro, "Karolinska sleepiness scale (KSS)," in *STOP, THAT and One Hundred Other Sleep Scales*. New York, NY, USA: Springer, 2011.

[15] M. Amarasinghe, S. Kottegoda, A. Liyana Arachchi, S. Muramudalige, H. M. N. D. Bandara, and A. Azeez, "Cloud-based driver monitoring and vehicle diagnostic with OBD2 telematics," in *Proc. 15th Int. Conf. Adv. ICT Emerg. Regions (ICTer)*, Aug. 2015, pp. 243–249, doi: 10.1109/ICTER.2015.7377695.

[16] K. Bylykbashi, E. Qafzezi, M. Ikeda, K. Matsuo, and L. Barolli, "Fuzzy-based driver monitoring system (FDMS): Implementation of two intelligent FDMSs and a testbed for safe driving in VANETs," *Future Gener. Comput. Syst.*, vol. 105, pp. 665–674, Apr. 2020, doi: 10.1016/j.future.2019.12.030.

[17] C. Avasalcai, I. Murturi, and S. Dustdar, "Edge and fog: A survey, use cases, and future challenges," in *Fog Computing: Theory and Practice*. Hoboken, NJ, USA: Wiley, 2020, pp. 43–65.

[18] R. Huang, J. Pedoeem, and C. Chen, "YOLO-LITE: A real-time object detection algorithm optimized for non-GPU computers," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2018, pp. 2503–2510, doi: 10.1109/BigData.2018.8621865.

[19] A. Moço and W. Verkruysse, "Pulse oximetry based on photoplethysmography imaging with red and green light," *J. Clin. Monitor. Comput.*, vol. 35, no. 1, pp. 123–133, Feb. 2021, doi: 10.1007/s10877-019-00449-y.

[20] M. Won, H. Alsaadan, and Y. Eun, "Adaptive audio classification for smartphone in noisy car environment," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1672–1679, doi: 10.1145/3123266.3123397.

[21] O. Lartillot, P. Toiviainen, and T. Eerola, *A MATLAB Toolbox for Music Information Retrieval* (Studies in Classification, Data Analysis, and Knowledge Organization). Berlin, Germany: Springer, 2008.

[22] S.-C. Lin, C.-H. Hsu, W. Talamonti, Y. Zhang, S. Oney, J. Mars, and L. Tang, "Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features," in *Proc. 31st Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, 2018, pp. 531–542, doi: 10.1145/3242587.3242593.

[23] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "AAM based features for multiple camera visual speech recognition in car environment," *Procedia Comput. Sci.*, vol. 57, pp. 614–621, Jan. 2015, doi: 10.1016/j.procs.2015.07.417.

[24] A. Biswas, P. K. Sahu, and M. Chandra, "Multiple cameras audio visual speech recognition using active appearance model visual features in car environment," *Int. J. Speech Technol.*, vol. 19, no. 1, pp. 159–171, Mar. 2016, doi: 10.1007/s10772-016-9332-x.

[25] R. Navarathna, D. Dean, S. Sridharan, and P. Lucey, "Multiple cameras for audio-visual speech recognition in an automotive environment," *Comput. Speech Lang.*, vol. 27, no. 4, pp. 911–927, Jun. 2013, doi: 10.1016/j.csl.2012.07.005.

[26] R. Navarathna, D. Dean, P. Lucey, S. Sridharan, and C. Fookes, "Recognising audio-visual speech in vehicles using the AVICAR database," in *Proc. 13th Australas. Int. Conf. Speech Sci. Technol.*, 2010, pp. 110–113.

[27] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Smart and robust speaker recognition for context-aware in-vehicle applications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8808–8821, Sep. 2018, doi: 10.1109/TVT.2018.2849577.

[28] A. Fernandez-Lopez and F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning," *Image Vis. Comput.*, vol. 78, pp. 53–72, Oct. 2018, doi: 10.1016/j.imavis.2018.07.002.

[29] D. Ivanko, A. Karpov, D. Fedotov, I. Kipyatkova, D. Ryumin, D. Ivanko, W. Minker, and M. Zelezny, "Multimodal speech recognition: Increasing accuracy using high speed video data," *J. Multimodal User Interfaces*, vol. 12, no. 4, pp. 319–328, Dec. 2018, doi: 10.1007/s12193-018-0267-1.

[30] A. Ortega, F. Sukno, E. Lleida, A. F. Frangi, A. Miguel, L. Buera, and E. Zacur, "AVCAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition," in *Proc. 4th Int. Conf. Lang. Resour. Eval. (LREC)*, 2004, pp. 1–4.

[31] M Zelezný and P Císar, "Czech audio-visual speech corpus of a car driver for in-vehicle audio-visual speech recognition," in *Proc. Int. Conf. Audio-Visual Speech Process. (AVSP)*, 2003, pp. 1–5. [Online]. Available: http://www.iscaaspeech.org/archive

[32] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human-computer interface research," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2002, pp. II-2017–II-2020, doi: 10.1109/icassp.2002.5745028.

[33] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015, doi: 10.1109/TMM.2015.2407694.

[34] J. S. Son and A. Zisserman, "Lip reading in profile," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2017, pp. 1–11, doi: 10.5244/c.31.155.

[35] Y. Lan, R. Harvey, B. Theobald, R. Bowden, and E. Ong, "Improving visual features for lip-reading," in *Proc. Int. Conf. Auditory-Vis. Speech Process.*, 2010, pp. 1–7.

[36] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007, pp. IV-429–IV-432, doi: 10.1109/ICASSP.2007.366941.

[37] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, May 2015, pp. 1–5, doi: 10.1109/FG.2015.7163155.

[38] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 6219–6223, doi: 10.1109/ICASSP.2018.8461596.

[39] A. Kashevnik, I. Lashkov, D. Ryumin, and A. Karpov, *Smartphone-Based Driver Support in Vehicle Cabin: Human-Computer Interaction Interface* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11659. Cham, Switzerland: Springer, Aug. 2019, pp. 129–138, doi: 10.1007/978-3-030-26118-4_13.

[40] A. Kashevnik, I. Lashkov, and A. Gurtov, "Methodology and mobile application for driver behavior analysis and accident prevention," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 6, pp. 2427–2436, Jun. 2019, doi: 10.1109/tits.2019.2918328.

[41] W. B. Verwey and D. M. Zaidel, "Preventing drowsiness accidents by an alertness maintenance device," *Accident Anal. Prevention*, vol. 31, no. 3, pp. 199–211, May 1999, doi: 10.1016/S0001-4575(98)00062-1.

[42] *Survive the Drive: A Guide to Keeping Everyone on the Road Alive*, Dingus, Tom, Buchanan-King, Mindy, Parrish, Alex, eBook Amazon.com. Accessed: Dec. 30, 2020. [Online]. Available: https://www.amazon.com/Survive-Drive-Guide-Keeping-Everyone-ebook/dp/B01DCLCIV2

[43] A. B. Ünal, D. de Waard, K. Epstude, and L. Steg, "Driving with music: Effects on arousal and performance," *Transp. Res. F, Traffic Psychol. Behaviour*, vol. 21, pp. 52–65, Nov. 2013, doi: 10.1016/j.trf.2013.09.004.

[44] P. Atchley, M. Chan, and S. Gregersen, "A strategically timed verbal task improves performance and neurophysiological alertness during fatiguing drives," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 56, no. 3, pp. 453–462, May 2014, doi: 10.1177/0018720813500305.

[45] S. Kikuchi, K. Iwata, Y. Onishi, F. Kubota, K. Nisijima, H. Tamai, Y. Koizumi, E. Watanabe, and S. Kato, "Prefrontal cerebral activity during a simple 'rock, paper, scissors' task measured by the noninvasive near-infrared spectroscopy method," *Psychiatry Res., Neuroimaging*, vol. 156, no. 3, pp. 199–208, Dec. 2007, doi: 10.1016/j.pscychresns.2007.01.002.

[46] E. Jensen, *Brain-Compatible Strategies*, 2nd ed. CA, USA: Corwin, 2007.

[47] M. Ingalhalikar, A. Smith, D. Parker, T. D. Satterthwaite, M. A. Elliott, K. Ruparel, H. Hakonarson, R. E. Gur, R. C. Gur, and R. Verma, "Sex differences in the structural connectome of the human brain," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 2, pp. 823–828, Jan. 2014, doi: 10.1073/pnas.1316909110.

[48] *Numbers API*. Accessed: Jul. 18, 2020. [Online]. Available: http://numbersapi.com/#42

[49] A. Kashevnik, I. Lashkov, A. Ponomarev, N. Teslya, and A. Gurtov, "Cloud-based driver monitoring system using a smartphone," *IEEE Sensors J.*, vol. 20, no. 12, pp. 6701–6715, Feb. 2020, doi: 10.1109/jsen.2020.2975382.

[50] A. Kashevnik, A. Ali, I. Lashkov, and D. Zubok, "Human head angle detection based on image analysis," in *Proc. Adv. Intell. Syst. Comput.*, vol. 1288, Nov 2021, pp. 233–242, doi: 10.1007/978-3-030-63128-4_18.

[51] A. Kashevnik, I. Lashkov, V. Parfenov, N. Mustafin, and O. Baraniuc, "Context-based driver support system development: Methodology and case study," in *Proc. 21st Conf. Open Innov. Assoc. (FRUCT)*, Nov. 2017, pp. 162–171, doi: 10.23919/FRUCT.2017.8250179.

[52] L. S. Dhupati, S. Kar, A. Rajaguru, and A. Routray, "A novel drowsiness detection scheme based on speech analysis with validation using simultaneous EEG recordings," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, Aug. 2010, pp. 917–921, doi: 10.1109/COASE.2010.5584246.