Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1555–1559
Marseille, 20-25 June 2022

Multi-Speaker Audio-Visual Corpus RUSAVIC: Russian Audio-Visual Speech in Cars

Denis Ivanko, Dmitry Ryumin, Alexandr Axyonov, Alexey Kashevnik and Alexey Karpov
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
{ivanko.d, ryumin.d, axyonov.a, alexey.kashevnik, karpov}@iias.spb.su

Abstract

We present a new audio-visual speech corpus (RUSAVIC) recorded in-the-wild in a vehicle environment and designed for noise-robust speech recognition. Our goal was to produce a speech corpus which is natural (recorded in real driving conditions), controlled (providing different SNR levels by windows open/closed, moving/idling vehicle, etc.), and adequate size (the amount of data is enough to train state-of-the-art NN approaches). We focus on the problem of audio-visual speech recognition: with the use of automated lip-reading to improve the performance of audio-based speech recognition in the presence of severe acoustic noise caused by road traffic. We also describe the equipment and procedures used to create RUSAVIC corpus. Data are collected in a synchronous way through several smartphones located at different angles and equipped with FullHD video camera and microphone. The corpus includes the recordings of 20 drivers with minimum of 10 recording sessions for each. Besides providing a detailed description of the dataset and its collection pipeline, we evaluate several popular audio and visual speech recognition methods and present a set of baseline recognition results. At the moment RUSAVIC is a unique audio-visual corpus for the Russian language that is recorded in-the-wild condition and we make it publicly available.

Keywords: audio-visual corpus, automatic speech recognition, data collection, automated lip-reading, driver monitoring

1. Introduction

In recent years, along with the rapid development of artificial intelligence technologies, the trend to multimodality has become very important and significantly boosted machine perception. Audio and visual information represent the two main perceptual modalities that we use in our daily life. Thus, in the past decades they have been widely researched and developed by both academy and industry. Speech technology has been advanced over the last 15 years, however, despite a significant success achieved in automatic speech recognition there are still a lot of challenges when the training and test data have mismatched noise conditions such as SNR or speaking styles (Shillingford et al., 2018). This becomes especially noticeable in need of a reliable speech recognizer inside of a vehicle. Unfortunately, at the moment there is no noise-robust speech recognition system to be used in real-driving conditions. And use of a hand to control navigation system/air conditioner/smartphone may distract a driver and cause road accidents. Along with this, the acoustic noise itself is not the main challenge in this domain (Lin et al., 2018). More importantly, background noise affects not only the microphone but also it causes the speaker to increase vocal effort to overcome noise levels in his ears (so-called Lombard effect). Thus, it is not enough to simply add artificial noise to the lab-recorded data, because in the real-world scenarios the variation of speech production caused by noise exposure at the ear can damage the performance more than the acoustic noise itself. This phenomenon has been carefully analyzed by researchers in the work (Lee et al., 2004).

However, the most of existing audio-visual corpora were collected in laboratory conditions which greatly limits their practical use (Kagirov et al., 2020; Ivanko et al., 2021;). There are no known and publicly available audio-visual Russian speech corpora recorded in a vehicle environment. So, there is a need to create such database to develop reliable audio-visual speech recognition in this language. At the same time, most of the existing audio-visual datasets are subject to some license restrictions and it is difficult to compare the speech recognition accuracy of one recognition system to another, as there is no common benchmark dataset, especially for the Russian language. Our goal in releasing the RUSAVIC corpus is to provide such a benchmark. We used developed earlier methodology as well as a mobile application and cloud infrastructure to make the corpora recording in-the-wild vehicle environment more convenient for the driver and automate the recording procedures (Kashevnik et al., 2021). The multi-speaker audio-visual corpus RUSAVIC can be downloaded from: https://mobiledrivesafely.com/corpus-rusavic.

The contribution of this paper is summarized as follows:

Firstly, we present a new audio-visual Russian corpus in a vehicle environment. RUSAVIC consists of recordings of 20 drivers uttered the script of three categories: 62 most frequent requests from driver to a smartphone, 33 letters of the Russian alphabet and 39 digits (including tens and hundreds). At least for 10 recording sessions for each driver.

Secondly, we provide a detailed description of the recording pipeline and framework. The data are collected in a synchronous way through several smartphones located at different angles and mounted on the vehicle dashboard. Each smartphone is equipped with FullHD video camera (60 fps) and microphone (48 kHz frequency).

Thirdly, we evaluate several state-of-the-art audio and visual speech recognition methods and present a set of baseline recognition results. The results demonstrate the consistency and the challenges of proposed benchmark.

The paper is structured as follows: after the Introduction Section 2 provides an overview of research related to audio-visual speech corpora; Section 3 details the recording framework and describe corpus creation methodology; in Section 4 we present the RUSAVIC corpus and its main

characteristics; in Section 5 experimental results are shown and analyzed; conclusions from this study and proposed future research are presented in Section 6, followed by acknowledgements in Section 7.

2. Related Works

Nowadays the are many audio-visual (AV) speech datasets collected for different purposes and with different means. In order to develop noise-robust automatic speech recognition systems, high-quality training and testing corpora are crucial. The researchers in the works (Fernandez-Lopez et al., 2018) and (Ivanko, 2020) provide comprehensive analysis on existing audio-visual speech databases. In this paper we refrain from repeating existing research and refer readers to the aforementioned papers. It should be noted, that almost every of around 60 publicly available datasets are recorded in controlled laboratory conditions. However, as was proven by the researchers in (Lee et al., 2004) background noise affects not only the microphone but also it causes the speaker to increase vocal effort to overcome noise levels in his ears. So, it is almost impossible to model real-life data in laboratory conditions (Oghbaie et al, 2021). Combining video and audio information can improve speech recognition accuracy for low signal-to-noise ratio conditions (Ivanko et al., 2021b). It has been demonstrated, that for humans the presence of the visual information is roughly equal to a 12dB gain in acoustic signal-to-noise ratio (Lee et al., 2004).

Another modern trend that appeared recently is the web-based corpora: datasets collected from open sources such as youtube or TV shows (Ryumina et al., 2021). The most well known of them are discussed in the works: LRW dataset (Yang et al., 2019), LRS2-BBC, LRS3-TED datasets (Afouras et al., 2018; Afouras et al., 2019; Yu et al., 2020), VGG-SOUND dataset (Chen et al., 2020), Modality corpus (Czyzewski et al., 2017), Multilingual AVSD (Mandalapu et al., 2021). A survey (Zhu et al., 2021) regarding this topic provides essential knowledge of current state-of-the-art situation. However, despite the fact that all aforementioned corpora are collected in the wild we cannot just repeat their success to create speech corpus for the car environments – because no such data is available on the web.

It is obvious that when driving a car, the active head turns from side to side are often involved. This simple fact greatly complicates the task of automated lip-reading, because the driver is showed to the camera with different angles. On the other hand, heavy acoustic noise on the road significantly degrades the results of audio-based speech recognition (Fedotov et al., 2018). Thus, the real-life training and testing data is a prerequisite to build a noise-robust and reliable audio-visual speech recognition system. In our recent work (Kashevnik et al., 2021) we carefully analyzed all the possible challenges that we need to tackle and discussed the main differences between existing audio-visual corpora and the one we collected.

Along with this, we took advantage of the experience of researchers, who previously collected speech corpora in-vehicle environment. According to our knowledge, there were only three attempts to record audio-visual speech corpora in a car, namely AVICAR (Lee et al., 2004), AV@CAR (Ortega et al., 2014), and Czech AVSC (Milos et al., 2003) for English, Spanish and Czech languages. Thus, there are no Russian audio-visual datasets recorded in-vehicle environment available up to now. The most well-known Russian audio-visual corpus is HAVRUS (Verkhodanova et al., 2016), however it is also recorded in laboratory conditions. Therefore, we hope the multi-speaker audio-visual corpus RUSAVIC could fill a part of the gap for Russian.

3. Acquiring RUSAVIC corpus

We create the multi-speaker audio-visual speech corpus using the recording methodology recently proposed in our work (Kashevnik et al., 2021).

Three smartphones were mounted in the vehicle cabin. Basic data recording settings are shown in Figure 1, left. The angle of the smartphone in relation to the driver has not exceeded 30 degrees. In fact, on most records it was about 20 degrees. The main smartphone is responsible for synchronization and for establishing a connection with a secondary smartphones. It is also responsible for audio-based interaction with a driver, utilizing the smartphones' microphone. The application synthesizes phrases the driver should repeat. The system records the time when the phrases were generated and saves all information to the SQLite database for further analysis and processing. The detailed description of the developed application for audio-visual corpora recording can be found in the paper (Kashevnik et al., 2021). The secondary smartphones are located at an 20-30 degrees angle in a way that its camera successfully captures driver's face (see Figure 1, right). These smartphones mainly focused on recording video and audio information captured by the smartphones front-facing camera. It should be noted, that such locations are popular among drivers to set their mobile devices with the navigation system for vehicles.

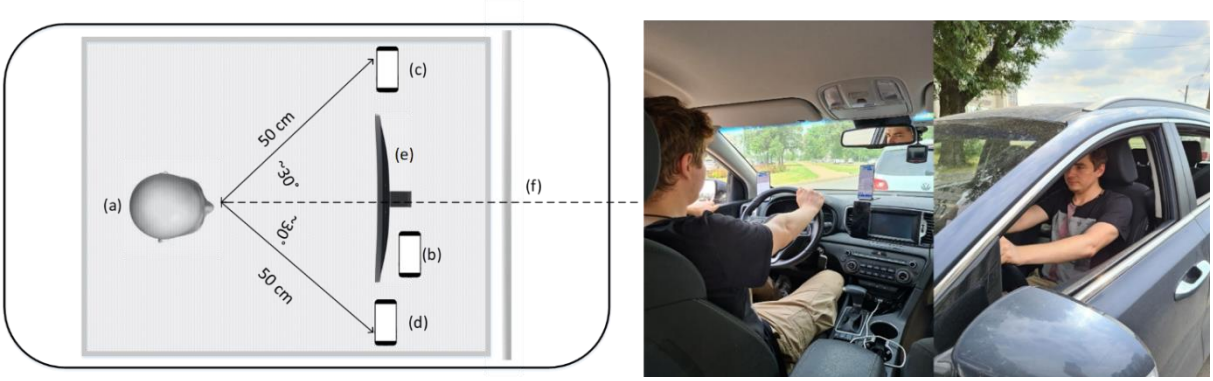


Figure 1. Data recording settings (left) and recording environment snapshots (right). Left: (a) driver; (b) main smartphone; (c) left-smartphone; (d) right-smartphone; (e) steering wheel; (f) windshield;

Figure 2. Snapshots of the drivers during recording session; top row - actual driving conditions; bottom row – vehicle parked near busy intersection.

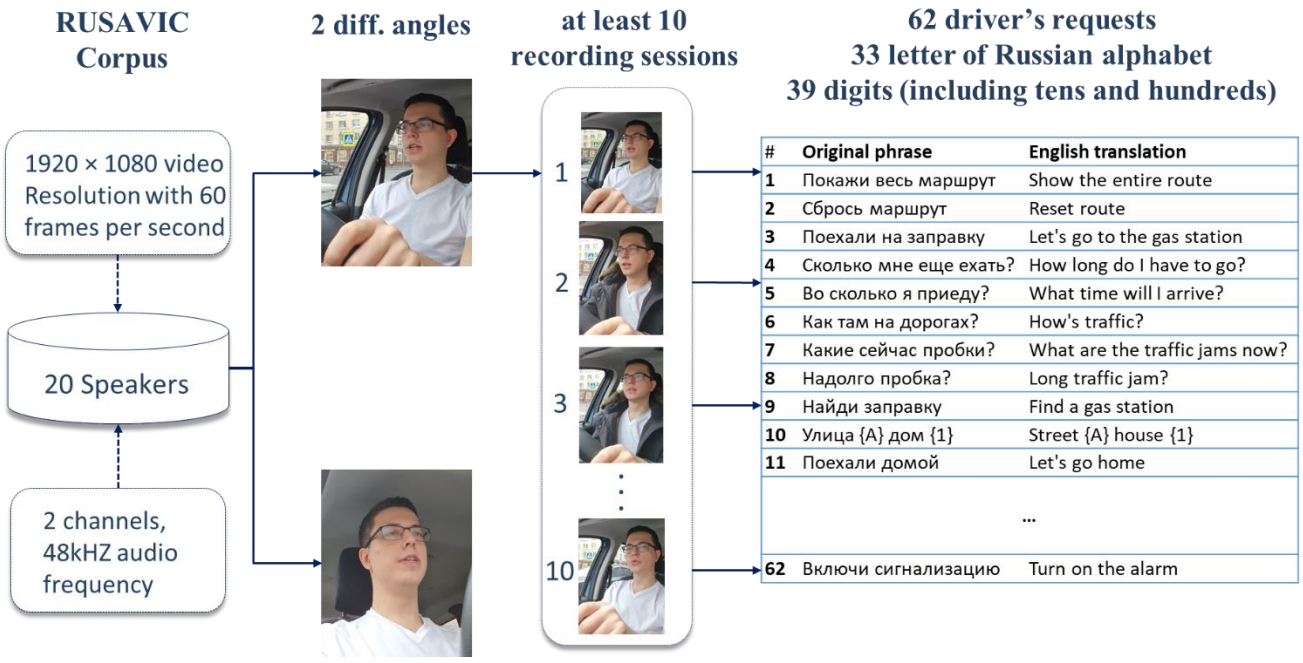
4. RUSAVIC corpus description

The audio-visual corpus RUSAVIC can be divided into two main parts. The first one is collected in actual driving conditions and the second one is collected in a vehicle parked near a busy intersection (Figure 2). Both parts of the database are composed of the recordings of 20 speakers. The main parametric characteristics of the recorded corpus are depicted in Figure 3. Each speaker uttered the script of three different dictionaries: 62 most frequent driver's requests to smartphones, 33 letters of the Russian alphabet and 39 digits (including tens and hundreds).

The first dictionary was chosen based on a market analysis of commercial driver assistance systems, such as AlexaAuto, YandexDrive, GoogleDrive, etc. Thus, the list of most frequently asked requests formed our main recognition dictionary. Two supplement dictionaries (letters and digits) were recorded to tackle out of vocabulary problems. It should be noted, that in Russian language we have special words for tens and hundreds, so we were obliged to record them as well.

At least 10 recording sessions have repeated each speaker (with maximum around 40 recording sessions). One recording session is a one repetition of three dictionaries. Since we record the corpus in-the-wild conditions the average SNR varies from 30 to 5 dB. The video resolution was FullHD 1920×1080 with 60 frames per second recording rate, mp4 format. The audio data was recorded with 48 kHz frequency. The current size of RUSAVIC corpus is about 250 Gb, mostly video data.

During postprocessing, segmentation and labeling are performed. Recording sessions metadata files are also



#	Original phrase	English translation
1	Покажи весь маршрут	Show the entire route
2	Сбрось маршрут	Reset route
3	Поехали на заправку	Let's go to the gas station
4	Сколько мне еще ехать?	How long do I have to go?
5	Во сколько я приеду?	What time will I arrive?
6	Как там на дорогах?	How's traffic?
7	Какие сейчас пробки?	What are the traffic jams now?
8	Надолго пробка?	Long traffic jam?
9	Найди заправку	Find a gas station
10	Улица {A} дом {1}	Street {A} house {1}
11	Поехали домой	Let's go home
	...	
62	Включи сигнализацию	Turn on the alarm

Figure 3. Main characteristics of the RUSAVIC audio-visual speech corpus

included. It contains such information as device description, driving hours, recording conditions, driver rotation angle, etc.

5. Evaluation experiments

In this section, we present the baseline evaluation results of popular lip-reading and audio-based recognition methods to illustrate the advantages and shortcomings of the created speech Corpus.

To answer the question of how well we can do automatic lip-reading in real driving conditions we train end-to-end neural network architecture, depicted in Figure 4.

The train and test sets were splitted 80 to 20 %. The input of the model is sequences of mouth images each 32 frames long with a resolution of 112×112 pixels, which pass through 3D convolutional layer (3D Conv) and modified residual blocks (Residual Blocks models ResNet-18) with attention modules (Squeeze-and-Attention, S.A.). Then the subsampling layer (Global Average Polling) transforms them into one-dimensional vectors that are fed to bidirectional networks with long short-term memory (BiLSTM) for subsequent recognition of phrases. Incoming video sequences are divided into segments of the same length into 32 frames with 50% overlap (16 frames). To reduce computational costs, the input images are transformed in grayscale and normalized to 112×112 pixels. To prevent overfitting MixUp augmentation technique is applied. The coefficient of combining two images and binary vectors ranged from 20 to 80%. For the remaining frames Label smoothing is applied.

A comparison of various lip-reading architectures on RUSAVIC corpus is presented in Table 1. As we can see from the results when applying several techniques, such as Cosine WR, MixUp, LS, and SA, the recognition accuracy of 62 voice commands of drivers increased from 46.45% to 64.09% (or by 17.64%). It can be seen that a significant contribution to the increase in accuracy is achieved by the SA module, and data augmentation techniques (MixUp and LS) give approximately the same increase in accuracy. However, despite the achieved result of accuracy (64.09%) there is still a lot of place for improvement. The next step is to boost the accuracy of the automated lip-reading by adding an audio modality.

Acoustic speech recognition generally performed better than the lip-reading. This fact is also proved by our results (see Table 1) Audio speech recognition results were obtained by the end-to-end 2D CNN spectrogram-based acoustic speech recognition system. We preprocess the raw acoustic data and obtain phrase-level spectrograms, followed by normalization and fed into pre-trained 2D CNN.

№	Neural network architecture	Recognition accuracy
1	3DResNet-18 + BiLSTM	46.45%
2	3DResNet-18 + BiLSTM + Cosine WR	48.28%
3	3DResNet-18 + MixUp + BiLSTM	49.14%
4	LS + 3DResNet-18 + BiLSTM	49.57%
5	SA + 3DResNet-18 + BiLSTM	55.59%
6	**LS + MixUp + SA + 3DResNet-18 + BiLSTM + Cosine WR**	**64.09%**
7	**Audio: Spectrogam + pre-trained 2D CNN / VGG19**	**87.26%**

Table 1: Speech recognition results on RUSAVIC corpus

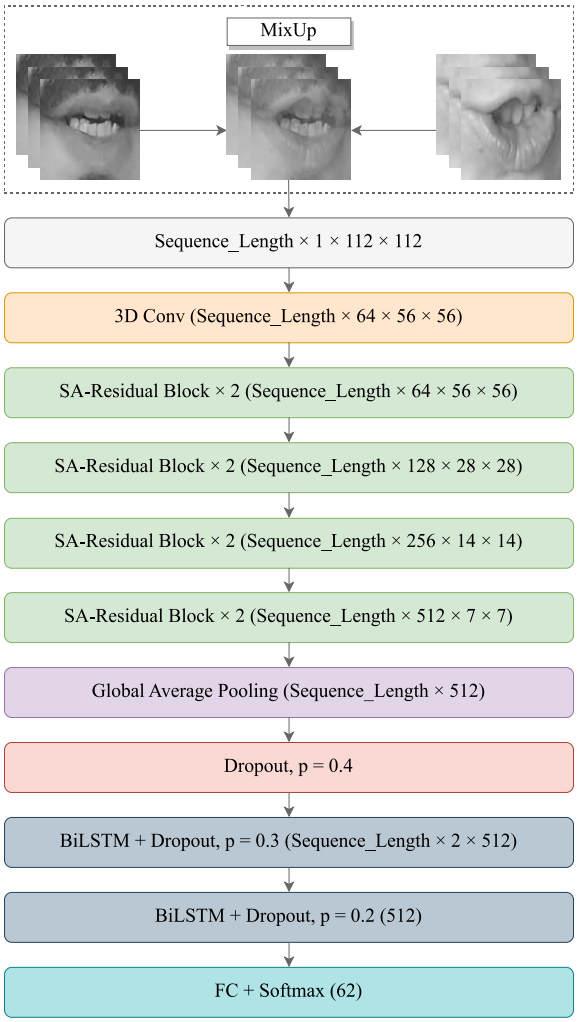


Figure 4. Visual speech recognition neural network model architecture

6. Conclusion

In this paper, we have created a multi-speaker audio-visual corpus RUSAVIC: Russian Audio-Visual Speech in Cars designed for noise-robust speech recognition.

The corpus includes the recordings of 20 drivers with the minimum 10 recording sessions for each (134 phrases in 3 dictionaries for each session). Besides providing a detailed description of the corpus and its collection pipeline, we evaluate several popular audio and visual speech recognition methods and present a set of baseline recognition results. At the moment RUSAVIC is a unique audio-visual corpus for the Russian language that is recorded in-the-wild condition and we make it publicly available. This database is available by request from < https://mobiledrivesafely.com/corpus-rusavic >.

With this new speech corpus, we wish to present the community with some challenges of the audio-visual speech recognition in-vehicle environment – acoustic noise, active head turns, pose, distance to recording devices, lightning conditions. These factors are encountered in many real-world applications and are very challenging for current state-of-the-art models. Our future work is related to new methods development for robust audio-visual speech recognition in a vehicle cabin based on RUSAVIC corpus.

7. Acknowledgements

This research is supported by the Russian Foundation for Basic Research (project No. 19-29-09081), as well as (Section V) by the Russian Science Foundation (project No. 21-71-00132).

8. Bibliographical References

Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. IEEE transactions on pattern analysis and machine intelligence.

Afouras, T., Chung, J. S., & Zisserman, A. (2019). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.

Chen, H., Xie, W., Vedaldi, A., & Zisserman, A. (2020). Vggsound: A large-scale audio-visual dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 721-725.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., & Szykulski, M. (2017). An audio-visual corpus for multimodal automatic speech recognition. Journal of Intelligent Information Systems, 49(2), 167-192.

Fedotov, D., Ivanko, D., Sidorov, M., & Minker, W. (2018). Contextual dependencies in time-continuous multidimensional affect recognition. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 78, 53-72.

Ivanko, D. (2020). Audio-Visual Russian Speech Recognition. PhD thesis. 399 p.

Ivanko, D., Ryumin, D., Axyonov, A., & Kashevnik, A. (2021). Speaker-Dependent Visual Command Recognition in Vehicle Cabin: Methodology and Evaluation. In International Conference on Speech and Computer, pp. 291-302.

Ivanko, D., Ryumin, D., & Karpov, A. (2021). An Experimental Analysis of Different Approaches to Audio–Visual Speech Recognition and Lip-Reading. In Proceedings of 15th International Conference on Electromechanics and Robotics" Zavalishin's Readings, pp. 197-209, Springer, Singapore.

Kagirov, I., Ivanko, D., Ryumin, D., Axyonov, A., & Karpov, A. (2020). TheRuSLan: Database of Russian Sign Language. In Proceedings of the LREC 2020, pp. 6079-6085.

Kashevnik, A., Lashkov, I., Axyonov, A., Ivanko, D., Ryumin, D., Kolchin, A., & Karpov, A. (2021). Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin. IEEE Access, 9, 34986-35003.

Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., & Huang, T. (2004). AVICAR: Audio-visual speech corpus in a car environment. In Eighth International Conference on Spoken Language Processing.

Lin, S. C., Hsu, C. H., Talamonti, W., Zhang, Y., Oney, S., Mars, J., & Tang, L. (2018). Adasa: A conversational in-vehicle digital assistant for advanced driver assistance features. In 31st Annual ACM Symposium on User Interface Software and Technology, pp. 531-542.

Mandalapu, H., Reddy, P. A., Ramachandra, R., Rao, K. S., Mitra, P., Prasanna, S. M., & Busch, C. (2021). Multilingual Audio-Visual Smartphone Dataset and Evaluation. IEEE Access, 9, 153240-153257.

Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). Advances and Challenges in Deep Lip Reading. arXiv preprint arXiv:2110.07879.

Ortega, A., Sukno, F., Lleida, E., Frangi, A. F., Miguel, A., Buera, L., & Zacur, E. (2004). AV@CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition. In LREC.

Rothkrantz, L. (2017). Lip-reading by surveillance cameras. In 2017 Smart City Symposium Prague (SCSP), pp. 1-6.

Ryumina, E., Ryumin, D., Ivanko, D., & Karpov, A. (2021). A novel method for protective face mask detection using convolutional neural networks and image histograms. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences. pp. 177-182.

Shillingford, B., Assael, Y., Hoffman, M. W., Paine, T., Hughes, C., Prabhu, U., de Freitas, N. (2018). Large-scale visual speech recognition. arXiv preprint arXiv:1807.05162.

Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., & Železný, M. (2016). HAVRUS corpus: high-speed recordings of audio-visual Russian speech. In SPECOM 2016, pp. 338-345.

Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., Chen, X. (2019). LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pp. 1-8.

Yu, J., Zhang, S. X., Wu, J., Ghorbani, S., Wu, B., Kang, S. & Yu, D. (2020). Audio-visual recognition of overlapped speech for the lrs2 dataset. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6984-6988

Zelezný, M., & Císar, P. (2003). Czech audio-visual speech corpus of a car driver for in-vehicle audio-visual speech recognition. In AVSP 2003-International Conference on Audio-Visual Speech Processing.

Zhu, H., Luo, M. D., Wang, R., Zheng, A. H., & He, R. (2021). Deep audio-visual learning: A survey. International Journal of Automation and Computing, 1-26.

9. Language Resource References

The multi-speaker audiovisual corpus RUSAVIC: RUSsian Audio-Visual speech In Cars: https://mobiledrivesafely.com/corpus-rusavic