Development of Visual and Audio Speech Recognition Systems Using Deep Neural Networks

Denis Ivanko [1] and Dmitry Ryumin [1]

[1] *St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 14th lin. Vasilievsky Island, 39, St. Petersburg, 199178, Russia*

Abstract

In this paper we design end-to-end neural network for the low-resource lip-reading task and audio speech recognition task using 3D CNNs, pre-trained CNN weights of several state-of-the-art models (e.g. VGG19, InceptionV3, MobileNetV2, etc.) and LSTMs. We present two phrase-level speech recognition pipelines: for lip-reading and acoustic speech recognition. We evaluate different combinations of front-end and back-end modules on the RUSAVIC dataset. We compare our results with traditional 2D CNN approach and demonstrate the increase in recognition accuracy up to 14%. Moreover, we carefully studied existing state-of-the-art models to be use for augmentation. Based on the conducted analysis we have chosen 5 most promising model's architectures and evaluated them on own data. We have tested our systems on a real-word data of two different scenarios: recorded in idling vehicle and during actual driving. Our independently trained systems demonstrated acoustic speech accuracy up to 90% and lip-reading accuracy up to 61%. Future work will focus on the fusion of visual and audio speech modalities and on speaker adaptation. We expect that fused multi-modal information will help to further improve recognition performance compared to a single modality. Another possible direction could be the research of different NN-based architectures to better tackle end-to-end lip-reading task.

Keywords

Computer vision, automated lip-reading, speech recognition, end-to-end, CNN, LSTM

1. Introduction

The ability to use a natural-to-human way of communication greatly improve the interaction quality of modern computer vision-based assistive systems. Speech is the usual way for humans to communicate. At the same time, the accuracy and robustness of automatic speech recognition (ASR) systems is not satisfactory in many practical conditions of use (e.g. in acoustically noisy conditions, while driving a car or being in a crowded place, etc.). In these cases, the advantage of using visual information about speech (lip-movements) in addition to audio is undeniable and is used in a number of state-of-the-art systems.

In current research, we tried to approach the problem of automatic audio-visual speech recognition from a computer vision and machine learning perspective. We developed and research two independent integral (end-to-end) systems for automatic recognition of Russian speech with limited vocabulary using CNN-based deep neural networks architectures. Moreover, we tried to consider the problem of acoustic speech recognition as a purely computer vision task by using images of speech spectrograms in order to train the networks.

There is no doubt that in recent years the active development of machine learning field has pushed the results in many other areas with automated lip-reading is no exception. However, despite all the achieved progress, the development of end-to-end speech recognition systems based on audio and visual

GraphiCon 2021: 31st International Conference on Computer Graphics and Vision, September 27-30, 2021, Nizhny Novgorod, Russia
EMAILdenis.ivanko11@gmail.com (D. Ivanko); ryumin.d@iias.spb.su (D. Ryumin);
ORCID: 0000-0003-0412-7765 (D. Ivanko); 0000-0002-7935-0569 (D. Ryumin);

information is still a new direction. Practically no research has been carried out in this field for the Russian language. There is no out-of-the-box solution accepted by researchers to the development of such systems. There are no representative open-access datasets for training NN models that have the required parameters, such as a sufficient number of speakers, phone-viseme labelling, vocabulary size adequate for the task, etc. (there are almost no public datasets available for languages other than English). The combination of these factors allows us to state a significant gap in the field of research. One of the main goals of this study is to bring the recognition efficiency of automatic systems closer to the level of human speech perception in noisy conditions, which is an extremely important task.

In this paper, we present lip-reading pipeline and acoustic speech recognition pipeline with the use of deep 3D CNNs. We trained and evaluated our models using RUSAVIC [1] dataset on a limited vocabulary of 50 phrases. To handle the over-fitting problem due to the increased number of parameters from the 3D kernels, we applied the idea from [2] to inflate the pre-trained weights of the several state-of-the-art models, such as MobileNetV2 [3], DenseNet121 [4], NASNetMobile [5], etc.

2. Related works

The classical approach towards audio-visual speech recognition involves a two-stage pipeline, including informative features extraction and classification using the sequence model. Usually the processes in two stages are independent. The most popular feature extraction approaches were based on dimension reduction and compression, such as Discrete Cosine Transform [6, 7]. Followed in the second stage by a sequence model (e.g. Hidden Markov Model) to tackle the temporal dependency from the extracted features for classification [8-10].

Although this two-stage pipeline methods have made significant progress over the decades, all such methods directly separate the feature extraction process from the classifier's training process, resulting in the extracted features might not be the optimal for classification. In the recent years, deep learning approaches have been proposed and achieved the state-of-the-art performance [11-16]. To date, the classical approach to AV speech recognition have been gradually replaced by the end-to-end trainable neural networks. In a raw approximation they behave somewhat similar to the traditional methods: a sequence of the mouth images is fed into the convolutional network to extract the features [17,18], which a further passed to a back-end model (RNN, LSTM, GRU or other) to account for the temporal dependency for classification [19 - 21]. Since the calculated gradient can be send back from back-end model to the front-end, the entire network is end-to-end trainable. Recent advances demonstrated that the learned features are more suitable for speech recognition and lip-reading than the standalone features calculated by traditional methods [22].

The major advantage of modern approach is that entire system consists of an end-to-end trainable front-end and back-end neural network (so two-stage process no longer exists). Thus, the learned features are more connected to the task that the network is trained on. The first work which proposed to use the CNNs to replace the independent features extractor was [23]. In turn, the first work that proposed to use the LSTM for classification and achieved a significant improvement was [24]. Other researchers in [11] proposed to take advantages of large-scale lip-reading dataset to train a front-end followed by the LSTM module at the end for classification. The researchers in [15] proposed a neural network to extract the audio features and tried to fuse them with video information.

It is generally accepted that visual features extracted from images by 2D CNNs are suitable for some computer vision tasks (e.g. image classification, lip-reading, gesture recognition etc.) [25,26]. However, it is more natural to learn spatio-temporal features by using a 3D CNNs as the front-end for feature extraction. Nonetheless, according to our knowledge only a few works have researched the use of 3D CNN for lip-reading [15, 16, 27, 28]. In addition, these works usually apply only a shallow version of 3D CNNs with no more than 3 convolutional layers. Obviously, this approach contradicts to the common rule that a deep network is expected to do better than a shallow one. Thus, the question of how to train a deep neural network without over-fitting on standard audio-visual datasets is still open and practically not studied.

Convolutional neural network architectures have been developed for image and video processing for a long time. In the work [29] an approach to extract features independently from each frame using a 2D CNN have been proposed to re-use the pre-trained weights of ImageNet [30] model. In the work [31]

the 3D CNNs for video action recognition have been introduced as a natural extension of the 2D convolution. The researchers in the work [32] went beyond using the shallow CNN version and explored the deep 3D CNN version with replacing all 2D operations with their 3D counterparts.

3. Data & Preprocessing

Almost no publicly-accessible audio-visual Russian speech datasets are available and suitable for NN training. The most recent one was introduced in the work [1] and was specifically designed for the task of robust speech recognition in acoustically-noisy car environment.

The multi-speaker audio-visual corpus RUSAVIC (RUSsian Audio-Visual speech In Cars) includes a continuous Russian speech with multi-angle video and audio data. It contains recordings of 20 native Russian speakers. The database stores audio and video recordings of Russian speech, as well as labelling information. The recording and labelling of audio-visual data was carried out using the created software package [33] designed to capture, synchronize and combine audio and video data from two or more smartphones located in the vehicle cabin. Recording of the speech corpus was carried out both in traffic conditions and in the idling of a vehicle, i.e. in full-scale and semi-natural conditions, as close as possible to the real conditions of functioning.

Each speaker performed 10 recording sessions and was captured by three smartphones from three different angles with FullHD 1920 × 1080 video resolution and 60 fps recording rate. During each recording session speaker uttered 50 phrases, which are the most frequent driver requests for smartphones (according to open source data of several state-of-the-art speech recognition engines, such as AlexaAuto, YandexDrive, GoogleDrive, etc.). The basic structure of the corpus is depicted in the Figure 1, and some snapshots of the speakers during the recordings are shown in Figure 2.

3.1. Visual data

Detecting a region-of-interest (ROI) that contains the mouth motion is the first and very important step in building a reliable automated lip-reading system. Thus, our first target is to crop this ROI (mouth region) from each frame of the video. To this end we applied the state-of-the-art solution of MediaPipe Face Mesh [34] that is able to estimate 468 3D face landmarks.

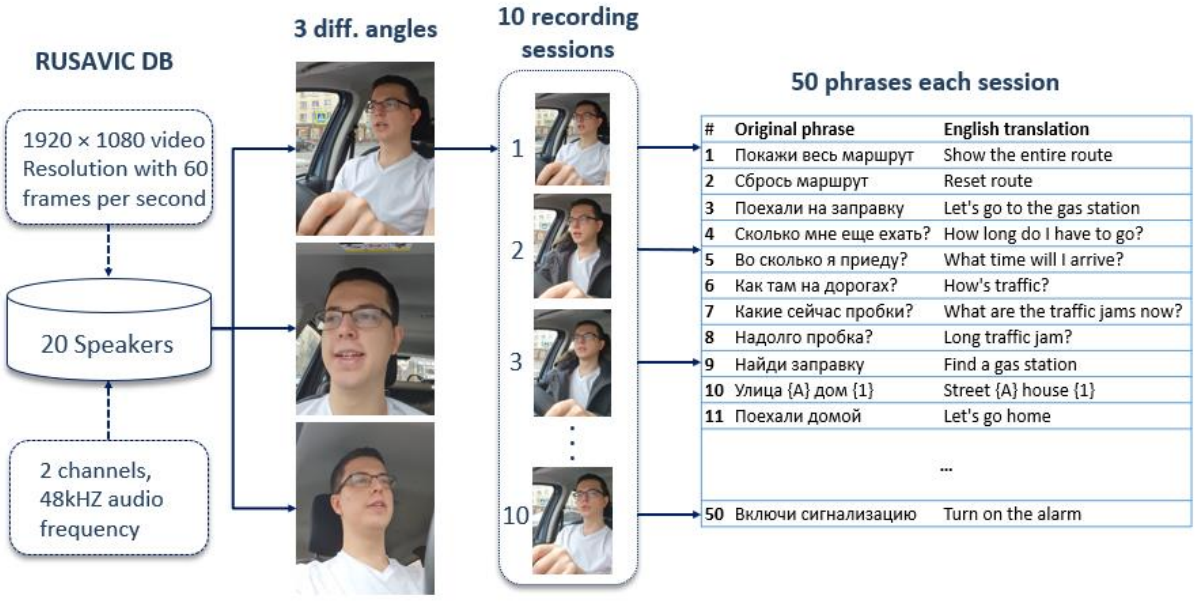

Figure 1: Basic structure of the RUSAVIC dataset

Figure 2: Snapshots of the RUSAVIC [1] speakers during the recording session

Face Mesh employs machine learning to infer the 3D surface geometry and provides real-time performance critical for real-life speech recognition scenarios. The general pipeline consists of two deep neural network models working together: (1) A detector that computes face locations (operates on the full image) and (2) a 3D face landmark model (operates on the detected locations) that predicts the approximate surface geometry via regression. The basic structure of this pipeline depicted in the Figure 3.

In addition, the mouth region crops can also be generated based on the face landmarks identified in the previous frame and only when the Model could no longer detect face presence the face detector is invoked to relocalize the face region. The pipeline is implemented as a graph that uses face landmark subgraph from the Face detection Module (Fig. 3, left) and renders using face renderer subgraph. We used the same BlazeFace detector as in original work [35]. The 3D face landmark model employed transfer learning and was trained with several objectives: it simultaneously predicts 3D landmark coordinates on synthetic rendered data and 2D semantic contours on annotated real-word data.

The 3D landmark network (Fig. 3, stage 2) receives as input a cropped frame and outputs the positions of the 3D points, as well as the probability of a face being present and aligned in the input. The Face landmark module performs a face landmark detection in the screen coordinate space, where the X- and Y- coordinates are normalized screen coordinates. Example of the detected 468 face landmarks is shown in the Figure 3, right.

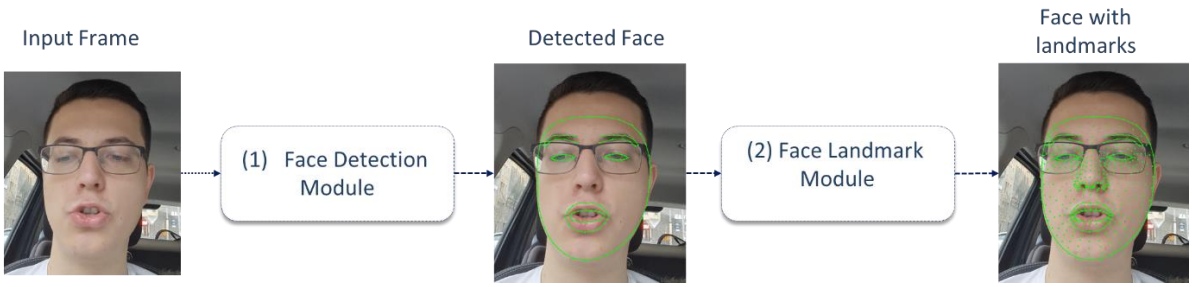

Figure 3: Applying Face Mesh ROI landmark localization pipeline to the RUSAVIC dataset

3.2. Acoustic data

One of the first works, that treated raw acoustic signal as an image was [36]. The authors proved that between the first two convolutional layers, the CNN learns (in parts) and models the phone-specific spectral envelope information of 2-4 ms speech. They demonstrated advantages of using the CNN-based approach to yield ASR performance.

In current research we handle acoustic speech processing by obtaining spectrograms of the uttered phrases from the raw audio data with its further processing by the integral CNN-LSTM network. We implement this using librosa library [37]. It is a python package for music and audio analysis. It is structured as collection of submodules. A spectrogram is calculated by computing the fast fourier transform (FFT) over a series of overlapping windows extracted from the raw audio signal. The process of dividing the signal in short term sequences of fixed size and applying FFT on those independently is called Short-time Fourier transform (STFT). The spectrogram is then calculated as the squared complex magnitude of the STFT. The general process of calculating spectrogram from the raw acoustic signal is depicted in the Figure 4.

4. Proposed methodology

End-to-end approach to automatic speech recognition assumes the training of only one neural network that combines all the stages of the traditional approach. At the same time, this presupposes the presence of certain structural blocks of the network, which we divide into four sequential processing stages:

1. Inputs, which is a sequence of cropped mouth images in case of lip-reading or a spectrogram images in case of acoustic speech recognition.
2. Front-end-module, to extract features from the inputs. We used a 3-4 3D CNN layers for the visual features extraction in the lip-reading system and a number of pre-trained CNNs for acoustic speech recognition.
3. Back-end module, to model the temporal dependency and summarize the features into a single vector that represents the score for each phrase.
4. Classification module, to compute the probabilities of each phrase. In both systems represented by a softmax layer.

The most of the existing end-to-end systems fall into this structure. In this paper we focused on the inputs (visual and audio data preprocessing) described in Section 3 and front-end modules which will be introduced in the following sections.

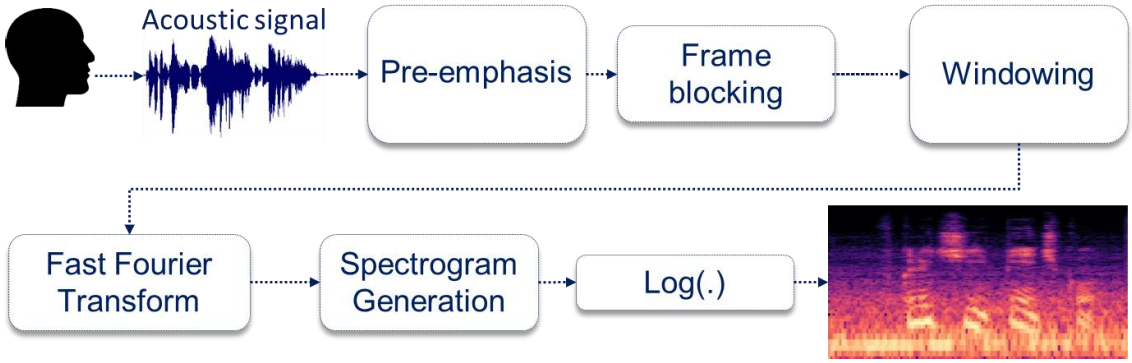

Figure 4: Spectrogram generation from raw acoustic signal

4.1. Building automated lip-reading system

General network architecture of our 3D CNN-based automated lip-reading system is presented in the figure 5. We compare our lipreading results with the recent work [33] that used deep 2D CNNs, which were originally proposed to solve image-base tasks. The general approach with applying 2D

CNNs on lip-reading data is to concatenate the features independently extracted from each frame. On the other hand, 3D convolution can process the dynamics (at least short-term dynamics) and is proven to be useful in many other computer vision-related tasks followed by the recurrent network at the back-end. However, due to the difficulty of training a vast number of parameters introduced by the three-dimensional kernel in current research we explore only 3 to 4 layers network with 3D convolution.

Cropped mouth frames sequences are first normalized to the size of 224×224 and then split into batches of 30 frames with 50% overlap (15 frames) before fed into the network. On all 3D CNN layers, we use three-dimensional kernel, followed by the batch normalization, Rectified linear units and 3D max-pooling. Specifically, in case of 3-layer network the number of kernels were 32, 64 and 128 respectively for each layer. In case of 4-year network the number of kernels were 32, 32, 64 and 128 respectively for each layer. The front-end visual features extraction part of our model ends with one densely connected layer with 512 neurons in it.

The back-end of the model consists of 2 (Long-short term memory) LSTM layers. LSTM is a type of recurrent neural networks, which are well-known for the ability to model temporal dependency and are typical back-end modules used in many computer vision and speech recognition tasks. Among RNNs, LSTM is proven to be useful when dealing with the exploding and vanishing gradient problem [38]. Specifically, we use a two-layer LSTM with a hidden state dimension of 512 for each cell in the first layer and 256 in the second, followed by 50 phrases classificatory represented by densely connected linear layer.

We trained and evaluated the proposed network on the phrase-level lip-reading dataset RUSAVIC. The number of target phrases is 50. We took 8 repetition of each phrase for the training and 2 for the testing for each speaker. Hence, the network has learned to discriminate between 50 target phrases based purely on the lip movements information.

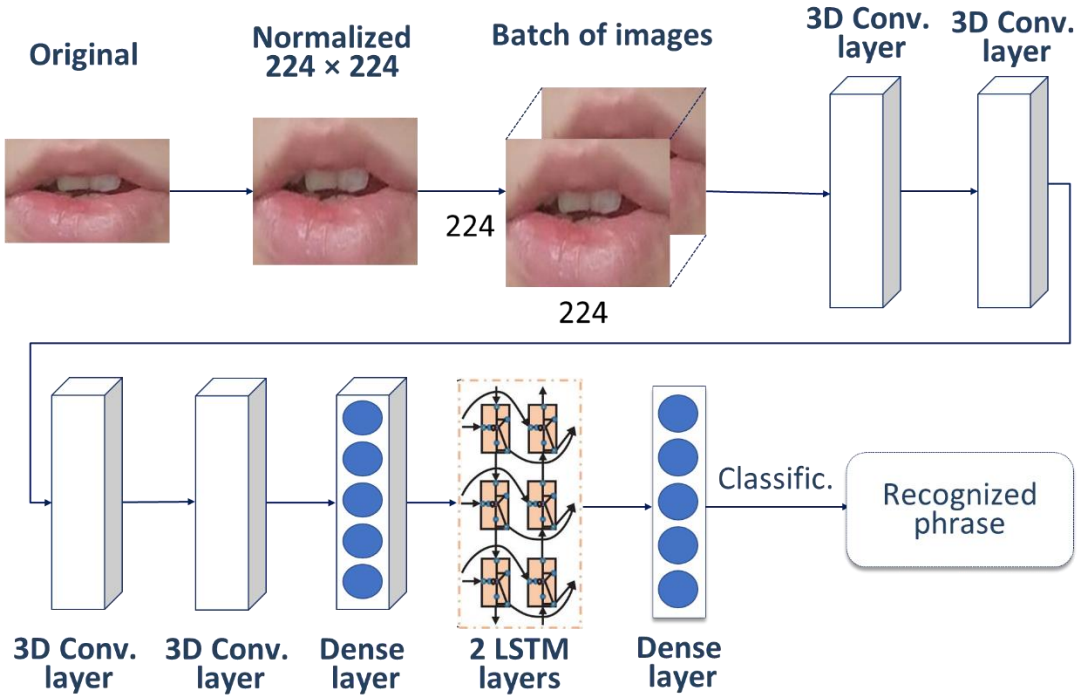

Figure 5: Network architecture of the end-to-end 3D CNN-LSTM lip-reading system

4.2. Building acoustic speech recognition system

Basic architecture of the end-to-end 2D CNN spectrogram-based acoustic speech recognition system is depicted in the Figure 6. We preprocess the raw acoustic data and obtain phrase-level spectrograms in accordance with the pipeline presented in the section 3.2. This step is followed by spectrogram normalization (we tested 2 types of input dimensions 224×224 and 299×299 depending on the pre-trained model).

Pre-trained weights are proven to be useful in many image-based tasks. Therefore, we tried to get the best use of modern transfer learning approaches and applied five different pre-trained deep CNN architectures, namely VGG19 [39], InceptionV3 [40], MobileNetV2 [3], DenseNet121 [4] and NASNetMobile [5].

VGG19 is 26-layer deep convolutional neural network with >143 million of trainable parameters. The default input size for this model is 224×224. The model is trained for the large-scale image recognition scenarios.

InceptionV3 is 159-layer deep network with >23 million of trainable parameters, developed specifically for mobile vision scenarios and big-data scenarios.

MobileNetV2 is 88-layer CNN with >3.5 million of trainable parameters. It uses inverted residual blocks with bottlenecking features and has a drastically lower parameter count than the original MobileNet. MobileNets support any input size greater than 32×32, with larger image sizes offering better performance.

DenseNet121 is 121-layer network with >8 million of trainable parameters. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters.

NASNetMobile has >5 million of trainable parameters. It is a scalable architecture for image classification and consist of two repeated building blocks termed Normal Cell and Reduction Cell. In current research we applied the latest 769 layers architecture.

The layers with pre-trained weights are followed by a 50 neuron softmax classification layers, that provides final recognition result.

5. Evaluation experiments

In this section we evaluate and compare the proposed architectures on different speakers of the RUSAVIC dataset. Maximum number of epochs was 50 and training was interrupted if the accuracy does not increase for 5 epochs. For each speaker the train and test data were splitted into 80 : 20 percent ratio. In total we trained a three speaker-dependent lip-reading system and three acoustic speech recognition systems, based on the available amount of data.

We summarize the lip-reading recognition results in the Table 1 and acoustic (spectrogram-based) speech recognition results in the Table 2. The first two systems (ID1 and ID2) were trained on the data, recorded in the idle vehicle, parked on the busy crossroads. The system #3 was trained with the actual driving data.

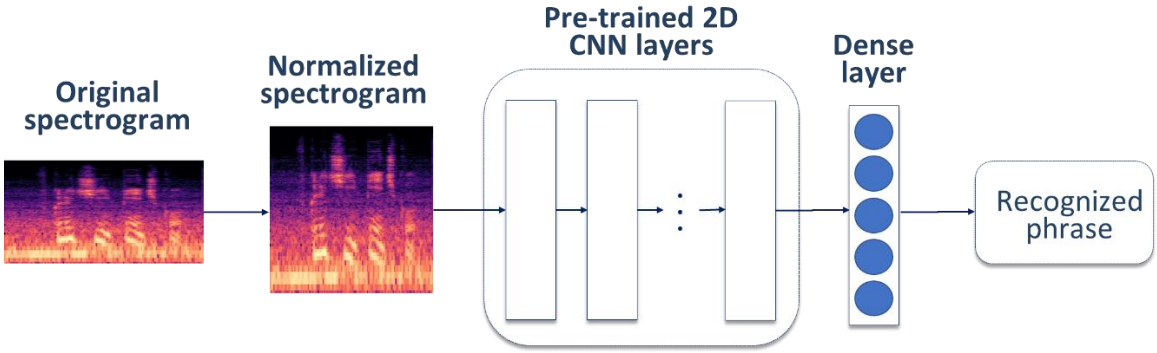

Figure 6: Network architecture of the end-to-end acoustic speech recognition system

According to the table 1, the 3D CNN-based architecture clearly outperforms the traditional CNN for lip-reading both: in vehicle idling conditions 61% vs (47 to 55 %) and 57% vs (46 to 53 %) accuracy on the phrase-level recognition, and driving conditions 59% vs (51 to 54 %) accuracy. The 2D CNN results were from the recent paper [33] evaluating speaker-dependent recognition systems on RUSAVIC dataset. Thus, in all systems 3D CNNs demonstrated significant improvements in the terms of recognition accuracy. Interestingly, despite the fact that driving requires rather active head turns we

did not found much difference in recognition accuracy between the models trained on driving data versus models trained on data recorded in a parked vehicle.

Another interesting finding was that using a slightly deeper 3D CNN (from 3 to 4 spatio-convolutional layers) results in increasing recognition accuracy up to 3% absolute. However, due to the limited amount of Russian lip-reading data available, further increase in network's depth does not lead to further improvements of recognition accuracy.

It can be seen from the Table 2, that spectrogram-based acoustic speech recognition generally performed better than the lip-reading. These results are within expectation range since acoustic information usually convey much more speech-related information that the lips movements. We achieved the maximum result of 90% recognition accuracy on the speaker #2 using pre-trained weights of the VGG19 model. In turn, the lowest recognition results were demonstrated by the model with NASNetMobile pre-trained weights (59%), which was trained on the driving data.

In addition to that, we perform experimental study and assess several state-of-the-art model architectures in order to research which of them provides better pre-trained weights for the task of automated speech recognition with using spectrograms as the network input. According to the obtained results, the most suitable for this task was VGG19 model, that achieved from 79 to 90% recognition accuracy. On the other hand, the lowest recognition results demonstrated NASNetMobile architecture, with only 59 to 61% speech recognition accuracy on all three systems. These results are almost the same that the one achieved by the 3D CNN lip-reading system with four spatio-temporal layers.

However, the advantage of VGG19 model is easily explained by fact that it has more than 143 million of trainable parameters, when the NASNetMobile only provides slightly more than 5 million trainable parameters. Thus, it is natural that VGG19 generalized better on the provided lip-reading data, since it was initially trained on much bigger amount of visual data. However, the disadvantage of using this architecture might be its resource-costly to the computational power of the device. E.g. it is not optimal to use it on smartphones or similar resource-dependent devices.

Table 1

Lip-reading visual recognition results comparison: 3D CNN vs traditional 2D CNN [33]

ID	Architecture	Number of 3DCNN layers	LSTM layers	Recognized classes	Accuracy, % (epoch)
1	3DCNN	3	1) 512 neurons with L2 regularization = 0,001 2) 256 neurons with L2 regularization = 0,001	50	58 (21)
	3DCNN	4			**61** (17)
1	2DCNN	-	-		47-55
2	3DCNN	3	1) 512 neurons with L2 regularization = 0,001 2) 256 neurons with L2 regularization = 0,001		55 (20)
	3DCNN	4			**57** (15)
2	2DCNN	-	-		46-53
3	3DCNN	3	1) 512 neurons with L2 regularization = 0,001 2) 256 neurons with L2 regularization = 0,001		56 (39)
	3DCNN	4			**59** (34)
3	2DCNN	-	-		51-54

Table 2
Spectrogram-based audio speech recognition results and comparison of several CNN architectures

ID	CNN architecture	Optimizer	Input size	Recognized classes	Accuracy, % (epoch)
1	InceptionV3	Adam=0,0001	299×299	50	79 (25)
	VGG19		224×224		**87 (26)**
	MobileNetV2		224×224		74 (27)
	InceptionResNetV2		299×299		79 (25)
	NASNetMobile		224×224		61 (28)
2	InceptionV3		299×299		81 (27)
	VGG19		224×224		**90 (23)**
	MobileNetV2		224×224		73 (21)
	InceptionResNetV2		299×299		81 (27)
	NASNetMobile		224×224		64 (29)
3	InceptionV3		299×299		73 (37)
	VGG19		224×224		**79 (33)**
	MobileNetV2		224×224		66 (35)
	InceptionResNetV2		299×299		73 (36)
	NASNetMobile		224×224		59 (38)

6. Conclusions

In this paper we have successfully demonstrated the capability and feasibility of designing an end-to-end neural network for the low-resource lip-reading and audio speech recognition task using 3D CNNs, pre-trained CNN weights of several state-of-the-art models (e.g. VGG19, InceptionV3, MobileNetV2, etc.) and LSTMs. We were able to achieve a state-of-the-art accuracy of 90 % for acoustic speech and 61% for lip-reading with 50 recognizable classes. To the best of our knowledge current research is one of the first attempts to work with Russian audio-visual speech.

We presented two phrase-level speech recognition pipelines: for lip-reading and acoustic speech recognition. We evaluated different combinations of front-end and back-end modules on the RUSAVIC dataset. We compared our results with traditional 2D CNN approach and demonstrated that even shallow 3 to 4 spatio-convolution layer network can outperform traditional approach up to 14 % recognition accuracy. Moreover, we carefully studied existing state-of-the-art models to be used for augmentation and transfer learning in the field of image processing and computer vision. Based on the conducted analysis we have chosen 5 most promising model's architectures and provided recognition results for each.

In the current research, we have studied Russian audio-visual speech from a computer vision perspective. We have tested our systems on real-word data of two different scenarios: idling vehicle and actual driving. Our independently trained systems demonstrated acoustic speech accuracy up to 90% and lip-reading accuracy up to 61%. Future work will focus on the fusion of visual and audio speech recognition systems and on speaker adaptation. We expect that fused multi-modal information will help to further improve recognition performance compared to a single modality.

7. Acknowledgements

This research is financially supported by the Russian Science Foundation (project No. 21-71-00132).

8. References

[1] A. Kashevnik et al.: Multimodal Corpus Design for Audio-Visual Speech Recognition in Vehicle Cabin. In IEEE Access, vol. 9 (2021) 34986-35003. doi: 10.1109/ACCESS.2021.3062752.

[2] J. Carreira, A. Zisserman: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 6299-6308. doi: 10.1109/CVPR.2017.502.

[3] A., Howard, M. Zhu, B. Chen, et al.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. In arXiv:1704.04861, pp. 1-9 (2017). arXiv:1704.04861.

[4] G. Huang, Z. Liu, L. Van Der Maaten, K. Weinberger: Densely Connected Convolutional Networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 2261-2269. doi: 10.1109/CVPR.2017.243.

[5] Z. Barret, V. Vijay, S. Jonathon, L. Quoc: Learning Transferable Architectures for Scalable Image Recognition. In Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 8697-8710. doi: 10.1109/CVPR.2018.00907.

[6] A. Pass, J. Zhang, D. Stewart: An Investigation into Features for Multi-View Lipreading. In 2010 IEEE International Conference on Image Processing (2010) 2417-2420. doi: 10.1109/ICIP.2010.5650963.

[7] X. Hong, H. Yao, Y. Wan, R. Chen: A PCA Based Visual DCT Feature Extraction Method for Lipreading. International Conference on Intelligent Information Hiding and Multimedia Signal Processing, (2006) 321-326. doi: 10.1109/IIH-MSP.2006.265008.

[8] V. Estellers, M. Gurban, J. Thiran: On Dynamic Stream Weighting for Audio-Visual Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20, 4 (2012) 1145-1157. doi: 10.1109/TASL.2011.2172427.

[9] D. Ivanko, A. Karpov, D. Ryumin, et al.: Using a high-speed video Camera for robust audio-visual speech recognition in acoustically noisy conditions. In International Conference on Speech and Computer, (2017) 757-766. doi: 10.1007/978-3-319-66429-3_76.

[10] D. Stewart, R. Seymour, A. Pass, J. Ming: Robust Audio-Visual Speech Recognition under Noisy Audio-Video Conditions. IEEE transactions on cybernetics, 44, 2 (2014) 175-184. doi: 10.1109/TCYB.2013.2250954.

[11] A. Zisserman, J. Chung. Lip Reading in the Wild. In Asian conference on computer vision, (2016) 87-103. doi: 10.1007/978-3-319-54184-6_6.

[12] H. Ninomiya, N. Kitaoka, S.Tamura, Y. Iribe, K. Takeda. Integration of Deep Bottleneck Features for Audio-Visual Speech Recognition. In 16th annual conference of the international speech communication association, (2015) 575-582. doi: 10.1109/APSIPA.2015.7415335.

[13] D. Ivanko, A. Karpov, D. Fedotov et al.: Multimodal speech recognition: increasing accuracy using high speed video data. Journal of Multimodal User Interfaces, vol. 12, (2018) 319–328. doi: 10.1007/s12193-018-0267-1.

[14] D. Ivanko, D. Ryumin, A. Karpov: An Experimental Analysis of Different Approaches to Audio–Visual Speech Recognition and Lip-Reading. In Proceedings of 15th International Conference on Electromechanics and Robotics, Singapore, (2021) 197-209. doi: 10.1007/978-981-15-5580-0_16.

[15] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos M. Pantic: End-to-End Audiovisual Speech Recognition. In IEEE international conference on acoustics, speech and signal processing (ICASSP) (2018) 6548-6552. doi: 10.1109/ICASSP.2018.8461326.

[16] G. Tzimiropoulos, T. Stafylakis: Combining Residual Networks with LSTMs for Lipreading. In arXiv preprint arXiv:1703.04105. (2017). doi: 10.21437/INTERSPEECH.2017-85.

[17] B. Xu, J. Wang, C. Lu, Y. Guo: Watch to Listen Clearly: Visual Speech Enhancement Driven Multi-modality Speech Recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, (2020) 1637-1646. doi: 10.1109/WACV45572.2020.9093314.

[18] E. Ryumina et al.: A Novel Method for Protective Face Mask Detection Using Convolutional Neural Networks and Image Histograms. In ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 4421 (2021) pp. 177–182. doi: 10.5194/isprs-archives-XLIV-2-W1-2021-177-2021.

[19] X. Chen, J. Du., H. Zhang: Lipreading with DenseNet and resBi-LSTM, Signal, Image and Video Processing, 14(5) (2020) 981–989. doi: 10.1007/s11760-019-01630-1.
[20] E. Ryumina, A. Karpov: Facial expression recognition using distance importance scores between facial landmarks. CEUR Workshop Proceedings, vol. 2744. (2020) 1–10. doi: 10.51130/graphicon-2020-2-3-32.
[21] D. Ivanko, D. Ryumin, I. Kipyatkova et al.: Lip-Reading Using pixel-based and geometry-based features for multimodal human–robot interfaces. In Proceedings of 14th International Conference on Electromechanics and Robotics, Singapore, (2020) 477-486. doi: 10.1007/978-981-13-9267-2_39.
[22] I. Fung, B. Mak: End-to-end low-resource lip-reading with maxout CNN and LSTM. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2018) 2511-2515. doi: 10.1109/ICASSP.2018.8462280.
[23] K. Noda, Y. Yamaguchi, K. Nakadai, H. Okuno, T. Ogata. Lipreading Using Convolutional Neural Network. In proceedings of the Annual Conference of the International Speech Communication Association, (2014) 1149-1153.
[24] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory. Neural computation, 9(8) (1997) 1735-1780. doi: 10.1162/neco.1997.9.8.1735.
[25] D. Ivanko, D. Ryumin, A. Axyonov, M. Železný: Designing Advanced Geometric Features for Automatic Russian Visual Speech Recognition. In International Conference on Speech and Computer, pp. 245-254 (2018). doi: 10.1007/978-3-319-99579-3_26.
[26] D. Ivanko, D. Ryumin: A Novel Task-Oriented Approach Toward Automated Lip-Reading System Implementation. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 44(2) (2021) 85-89. doi: 10.5194/isprs-archives-XLIV-2-W1-2021-85-2021.
[27] T. Afouras, J. Chung, A. Senior, O. Vinyals, A. Zisserman: Deep Audio-Visual Speech Recognition. IEEE transactions on pattern analysis and machine intelligence (2018). doi: 10.1109/TPAMI.2018.2889052.
[28] Y. M. Assael, B. Shillingford, S. Whiteson, N. De Freitas. LipNet: End-to-End Sentence-Level Lipreading. arXiv preprint arXiv:1611.01599, 2(4) (2017). arXiv:1611.01599
[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, F. Li: Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (2014) 1725-1732. doi: 10.1109/CVPR.2014.223.
[30] A. Krizhevsky, I. Sutskever, G.E. Hinton.: Imagenet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems, 25, (2012) 1097-1105. doi: 10.1145/3065386.
[31] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri: Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the IEEE international conference on computer vision, (2015) 4489-4497. doi: 10.1109/ICCV.2015.510.
[32] K. Liu, W. Liu, C. Gan, M. Tan, H. Ma: T-C3D: Temporal Convolutional 3D Network for Real-time Action Recognition. In Proceedings of the AAAI conference on artificial intelligence, 32(1), (2018) 7138-7145.
[33] D. Ivanko, D. Ryumin, A. Axyonov, A. Kashevnik: Speaker-Dependent Visual Command Recognition in Vehicle Cabin: Methodology and Evaluation. In International Conference on Speech and Computer (2021). To appear.
[34] Y. Kartynnik, A. Ablavatski, I. Grishchenko, M. Grundmann: Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. In: CVPR Workshop on Computer Vision for Augmented and Virtual Reality, IEEE (2019) 1-4. arXiv:1907.06724.
[35] V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, M. Grundmann: Blazeface: Sub-millisecond neural face detection on mobile gpus. arXiv preprint (2019). arXiv:1907.05047
[36] D. Palaz, M. Magimai-Doss, R. Collobert: Analysis of cnn-based speech recognition system using raw speech as input. In 16th annual conference of the international speech communication association, (2015), pp. 686-692.
[37] B. McFee, R. Colin, L. Dawen, E. Daniel, M. McVicar, E. Battenberg, O. Nieto: librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference (2015) 18-25. doi: 10.25080/Majora-7b98e3ed-003.

[38] X. Weng, K. Kitani: Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading. arXiv preprint (2019). arXiv:1905.02540.
[39] K. Simonyan, A. Zisserman: Very deep convolutional networks for large-scale image recognition. In arXiv preprint (2019). arXiv:1409.1556.
[40] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna: Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 2818-2826. doi: 10.1109/CVPR.2016.308.