Visual Speech Recognition in a Driver Assistance System

Denis Ivanko
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
denis.ivanko11@gmail.com

Dmitry Ryumin
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
ryumin.d@iias.spb.su

Alexey Kashevnik
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
alexey.kashevnik@iias.spb.su

Alexandr Axyonov
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
axyonov.a@iias.spb.su

Alexey Karpov
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)
St. Petersburg, Russia
karpov@iias.spb.su

***Abstract*—Visual speech recognition or automated lip-reading is a field of growing attention. Video data proved its usefulness in multimodal speech recognition, especially when acoustic data is heavily noised or even inaccessible. In this paper, we present a novel method for visual speech recognition. We benchmark it on the famous LRW lip-reading dataset by outperforming the existing approaches. After a comprehensive evaluation, we adapt the developed method and test it on the collected RUSAVIC corpus we recorded in-the-wild for vehicle driver. The results obtained demonstrate not only the high performance of the proposed method, but also the fundamental possibility of recognizing speech only by using video modality, even in such difficult natural conditions as driving.**

Keywords—visual speech recognition, automated lip-reading, end-to-end, speech recognition, computer vision

I. Introduction

Multi-modal human-machine spoken language communication gained increasing attention in recent years. Due to the prosperous growth of artificial intelligence and deep neural networks automated lip-reading become an appealing tool for intelligent human-machine interaction. Visual information is beneficial to the automatic speech recognition systems, especially when the audio itself is noisy or even inaccessible. Benefiting from the emergence of several large-scale datasets, such as GRID [1], LRW [2], LRS [3], etc. visual speech recognition has made great progress over the last five years. However, despite significant successes achieved by researchers, there is still a lot of place for improvement. Especially for visual speech recognition in natural conditions of use, e.g. driving a car. Speech recognition accuracy in these cases remains terribly low and often not suitable for practical applications.

To further push the boundary of this research area, in this work we focus on improving the performance of the state-of-the-art lip-reading models. We propose a method to solve the problem of robust visual speech recognition. On the one hand, we perform a comprehensive comparative analysis of the proposed method with the existing state-of-the-art, benchmarking on the well-known LRW dataset. We achieved better results than the current state-of-the-art on the largest publicly available word-level dataset. On the other hand, we demonstrate the efficiency of the proposed method and model architecture by testing it on the collected RUSAVIC corpus we recorded in-the-wild for the vehicle driver assistance system. Thus, we demonstrated that modern lip-reading systems can provide reasonable speech recognition accuracy based purely on video data even in acoustically and visually noisy conditions, such as driving.

The structure of the paper is as follows. Section II consider related work in the topic of visual speech recognition. In Section III we present datasets we used. Section IV considers the proposed methodology and benchmarking results. Section V discusses experimental results for the driver assistance system. Conclusion summarize the paper.

II. Related Works

Research on visual speech recognition has a long history. A comprehensive study of earlier approaches can be found in the work [4]. In the recent years, with the rapid developments of machine learning approaches and artificial intelligence, deep neural networks were introduced to this area. The first breakthrough on large scale lip-reading datasets was achieved in [5] by introducing multi-layer CNN architecture based on VGG-M. Soon after, deep Residual networks were proposed by [6, 7] as the front-end of the visual speech recognition model. The current state-of-the-art, temporal convolutional neural networks were proposed for lip-reading by researchers in [8, 9]. The first end-to-end sentence lip-reading model LipNet was proposed by researchers in [10].

According to the design of the front end network, the modern lip-reading methods can be divided into three categories: 2-dimensional (2D) convolutional neural networks (CNN), such in [11], 3-dimensional convolutional neural networks (3D CNNs), such in [12], or a combination of 2D and 3D convolutions, which inherit the advantages of both [13]. Recently, the method of the third type has become widely used in visual speech recognition due to its ability to simultaneously capture temporal dynamics of lips movements and extract discriminative features. For sequence modeling LSTMs or its variations are often used [14]. When temporal modeling is required, LSTMs usually lead to better performance and are commonly used in NLP, video prediction, automated lip-reading, etc. [15, 16].

With these impressive methods, state-of-the-art visual speech recognition accuracy has been raised from 61.1% [2]

to 88.5% [9] on the largest English dataset LRW during the last five years.

The researchers in the work [17] provide analysis on existing audio-visual and visual-only speech databases. The most well-known of them are: LRW dataset [2], LRS2-BBC [18], LRS3-TED [19], VGG-SOUND [20], Modality corpus [21]. A survey [22] regarding this topic provides essential knowledge of the current state-of-the-art situation.

Driver assistance systems have gained remarkable progress recently. It allows drivers to use short commands to handle complex operations, which is a current industry demand [23]. In the past several years, many multimodal and visual speech datasets have been released to facilitate the research of in-vehicle speech recognition [24]. However, none of the previous works focuses on in-vehicle command recognition, especially, for languages other than English. The lack of data is one of the largest issues in building such systems. To the best of our knowledge, RUSAVIC corpus [25] is one of its kind for the Russian language.

The combination of state-of-the-art deep learning approaches and large-scale audio-visual datasets has been highly successful, achieving significant recognition accuracy results and even surpassing human performance [26]. However, there is still a long journey for practical visual speech recognition applications to meet the performance requirements of real-life scenarios and deal with various road environments and noise conditions for various driver languages and abilities.

III. Data

Two different visual speech datasets were used in the current research. For benchmarking, we test our methodology on a well-known Lip-Reading in the Wild (LRW) dataset [2], collected in 2016 based on BBC TV shows. The second dataset was collected specifically for use in drivers' assistive systems and is called RUSAVIC: Russian Audio-Visual Speech in Cars.

A. LRW dataset

Lip-Reading in the Wild dataset combines the recordings of hundreds of English language speakers. Dataset dictionary includes 500 words, forming up to 1000 difference utterances. All videos have the same frame rate. The LRW dataset parameters are presented in Table I.

Some snapshots of the speakers of the LRW dataset are given in Figure 1. Since the data in LRW are taken in natural conditions co-articulation of the lips from adjacent words is present.

B. RUSAVIC dataset

Russian Audio-Visual Speech in cars (RUSAVIC) is a multi-speaker and multi-modal corpus created based on the methodology proposed in the work [24] and described in detail in our recent paper [25].

TABLE I. LRW Dataset Parameters

Dataset	Classes	Samples for each class	Number of frames
Train	500 (words)	800-1000	29
Valid		50	
Test		50	

RUSAVIC is designed specifically to tackle the speech recognition of the most frequent driver's requests and is meant to be used in the creation of driver's assistive systems. The main parametric characteristics of the RUSAVIC corpus regarding automated lip-reading tasks are shown in Table II. Each of the 20 speakers uttered 62 most frequent requests at least 10 times during several recording sessions (including actual driving conditions and a vehicle parked near a busy intersection). The video resolution is FullHD 1920×1080 with 60 frames per second frame rate. Some snapshots of the speakers of RUSAVIC are shown in Figure 2.

IV. Proposed Methodology and Benchmarking Results

In this section, we describe the proposed methodology to automated visual speech recognition and present the benchmarking results on the LRW dataset. According to our evaluation, the present approach clearly outperforms the existing methods known in the scientific literature to date.

A. Proposed Methodology

The functional diagram of the proposed visual speech recognition method consists of two stages and is shown in Figure 3.

TABLE II. RUSAVIC Dataset Characteristics

Parameter	Value
Number of speakers	20
Video Resolution	1920 × 1080
Frame Rate	60
Classes	62
Samples for each class	>400

Fig. 1. Snapshots of speakers of LRW dataset

Fig. 2. Snapshots of speakers of RUSAVIC dataset

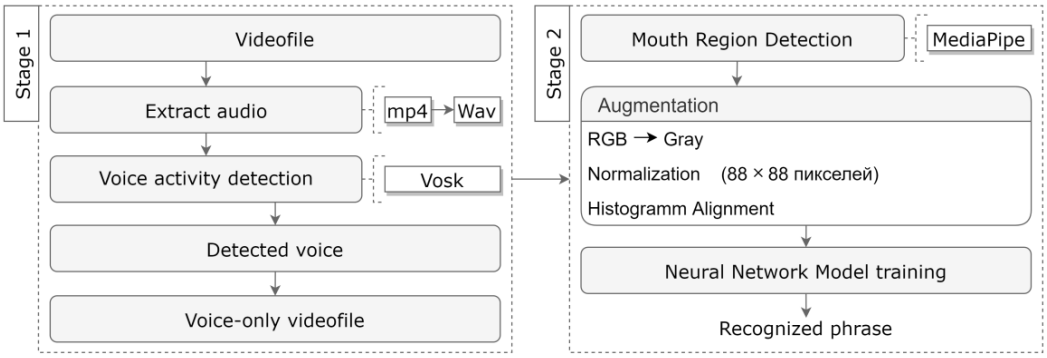

Fig. 3. Functional diagram of the proposed visual speech recognition method

The first stage involves sequential execution of the several steps with the main idea to extract the voiced part of the speech and get rid of the silence (on video data). To achieve this goal, we use the Vosk voice activity detection model (https://github.com/alphacep/vosk-api), which is able to confidently detect speech regions even in heavy acoustically noisy conditions. Thus, after applying this simple preprocessing step we get proper video files, without redundant data.

At the second processing stage, we detect the mouth region on each frame using the MediaPipe FaceMesh algorithm [27]. The region-of-interest (mouth region) detection process is described in detail in our previous work [28]. After cropping region-of-interest several procedures are applied, namely: (1) grayscaling, (2) image normalization (in case of LRW dataset to 88 × 88 pixels), and (3) histogram alignment. This followed by the MixUp augmentation technique in order to reduce overfitting while training the neural network model.

The proposed end-to-end neural network architecture used for visual speech recognition on the LRW dataset is shown in Figure 4. As already mentioned, during the training procedure MixUp data augmentation technique was applied to images with a probability of 40% to reduce overfitting. The merging ratio of the two images varied from 30 to 70% so that the sum was always 100% (zero transparency). During this process, two samples $A: (x_A, y_A)$ and $B: (x_B, y_B)$ are selected to generate a new sample $(\hat{x}, \hat{y})$ by a weighted linear interpolation as:

$$\hat{x} = \lambda x_A + (1 - \lambda) x_B, \hat{y} = \lambda y_A + (1 - \lambda) y_B \qquad (1)$$

where x_i, y_i denotes the training sample and the word label of data $i \in \{A, B\}$ respectively.

Label Smoothing (LS) was applied to the labels of those frames that did not have MixUp. The resulting images were formed into batches and fed into convolutional layers for visual features extraction. Given an input sample belonging to word class i, we denote p_i as the prediction logits and y as the annotated word label, as was done in [29]. Let N be the number of classes. Then the cross-entropy loss is computed as follows:

$$L = \sum_{i=1}^{N} q_i \log(p_i) \begin{cases} q_i = 0, y \neq i \\ q_i = 1, y = i \end{cases} \qquad (2)$$

When applying LS q_i is changed (ϵ is a small constant) to:

$$q_i = \begin{cases} \epsilon \,/\, N, y \neq i \\ 1 - \frac{N-1}{N} \epsilon, y \neq i \end{cases} \qquad (3)$$

A modified 3DResNet-18 neural network [30] was used in order to extract informative features. Useful representation for semantic segmentation appears at both global and local levels of each frame. At the pixel level, convolution layers generate feature maps conditional on local information, as convolution is computed locally around each pixel. At the global image level, context can be exploited to determine which parts of feature maps are activated, because the contextual features indicate which classes are likely to appear together in the image. To get the advantage of both, local and global information we use a squeeze-and-attention (SA) module.

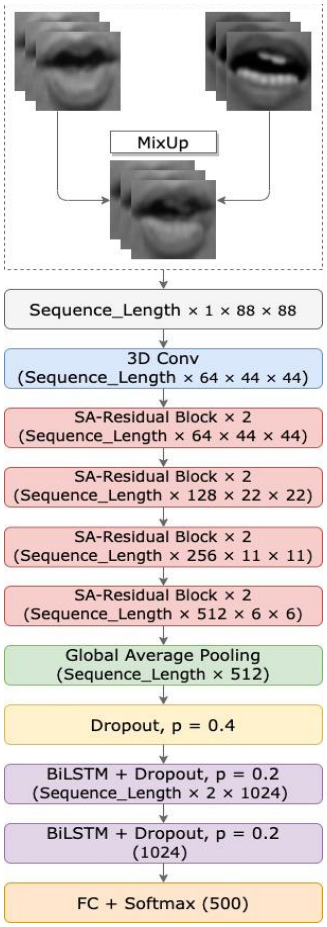

Fig. 4. Proposed NN model architecture for visual speech recognition

TABLE III. RECOGNITION RESULTS ON LRW DATASET

No.	Method	Recognition accuracy	Paper
1	3D Conv + ResNet-34 + Bi-LSTM	83.0 %	[6]
2	Multi-grained + Bi-ConvLSTM	83.34 %	[37]
3	3D Conv + ResNet-34 + Bi-GRU	83.39 %	[36]
4	PCPG	83.5 %	[35]
5	DFTN	84.13 %	[34]
6	SpotFast + Transformer + Product-Key memory	84.4 %	[33]
7	3D Conv + ResNet-18 + Bi-GRU	84.41 %	[32]
8	3D Conv + P3D-ResNet50 + TCN	84.8 %	[31]
9	3D Conv + ResNet-18 + Bi-GRU (Face Cutout)	85.02 %	[25]
10	3D Conv + ResNet-18 + MS-TCN	85.3 %	[8]
11	3D Conv + ResNet-18 + Bi-GRU + Visual-Audio Memory	85.4 %	[30]
12	3D-ResNet + Bi-GRU + MixUp + Label Smoothing + Cosine LR	85.5 %	[29]
13	3D-ResNet + Bi-GRU + MixUp + Label Smoothing + Cosine LR (Word Boundary)	88.4 %	[29]
14	3D Conv + ResNet-18 + MS-TCN + KD (Ensemble)	88.5%	[9]
15	**Proposed Method**	**88.7 %**	

B. Benchmarking Results on LRW dataset

The back-end of the model is a Bidirectional Long-Short Term Memory (LSTM) network. The extracted features were fed to 2 layers of BiLSTM. The output of the first BiLSTM layer is sequence-to-sequence. The output of the second BiLSTM layer is sequence-to-one. The last fully-connected layer determines the most probable recognition result from 500 classes.

A comparison of recognition results of several state-of-the-art approaches with the proposed methodology is presented in Table III. As can be seen from the table, our model outperforms all recent state-of-the-art approaches up to 5.7 % absolute (in comparison with the work [6]). The closest result was obtained by researchers in the work [9]. They rely on the same core idea of using 3D CNNs for features extraction, however, they used MS-TCN + KD for recognition, resulting in 0,2 % less accuracy than our approach.

As we can see from Table III, all modern approaches (11 out of 15) rely on the use of 3D Convolution Neural Networks for visual features extraction from the video. The same consent can be observed at the back-end of the model: all the researchers try to use some kind of Recurrent Neural Network, such as Bi-LSTM or Bi-GRU for the recognition part. The main advantages of our approach are: (1) better preprocessing by using Vosk to get rid of redundant silence on the video data, (2) applying MixUp augmentation technique to reduce overfitting, (3) using Squeeze-and-Attention module as heads to extract features and fully exploit their multi-scale. After a comprehensive comparative evaluation of the proposed method, we assert its full viability. In the following section, we implement it for the task of visual speech recognition in our driver assistance system based on the collected RUSAVIC dataset.

TABLE IV. RECOGNITION RESULTS ON RUSAVIC DATASET

№	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	**Vosk + MediaPipe + LS + MixUp + SA + 3DResNet-18 + BiLSTM + Cosine WR (Our Method)**	**64.09%**

V. EXPERIMENTAL RESULTS FOR DRIVER ASSISTANCE SYSTEM

The train and test sets of the RUSAVIC dataset were split from 80% to 20%. Since the video resolution of RUSAVIC differs from LRW the input of the NN model was sequences of mouth images every 32 frames long with a resolution of 112×112 pixels. The rest of the end-to-end model architecture was the same as in Figure 4, except the final fully-connected layer (62 neurons with SoftMax activation in case of the RUSAVIC).

A comparison of various visual speech recognition architectures on RUSAVIC dataset is shown in Table IV. We can see that the recognition accuracy of 62 visual speech commands of drivers increased from 46.45% to 64.09% (or by 17.64%) by using our method in comparison with baseline, where 3D CNNs used for features extraction followed by Bi-LSTM for recognition.

The absolute recognition values on LRW and RUSAVIC datasets are differs by about 20%. However, it should be noted, that we cannot compare these results directly due to huge differences between datasets: amount of data (LRW dataset much bigger), language (English and Russian), recording conditions (much noisy environment in vehicles, a lot of head turns, etc.), etc. Thus, we can conclude with confidence that current state-of-the-art approaches and our method specifically can provide high speech recognition accuracy based purely on video data even in acoustically and visually noisy environments, such as driving conditions or TV shows.

VI. CONCLUSION

In this paper, we present state-of-the-art results on visual speech recognition. We propose a method for automated lip-reading. We benchmark it on the well-known LRW lip-reading dataset by outperforming the existing approaches. After a comprehensive evaluation, we adapt the developed method and test it on the collected RUSAVIC corpus we recorded in-the-wild for the vehicle driver assistance system. The results obtained demonstrate the high performance of the proposed method and the fundamental possibility of recognizing speech by using video modality only, even in such difficult natural conditions as driving.

ACKNOWLEDGMENT

This research is financially supported by Russian Foundation for Basic Research (project No. 19-29-09081), Grant (No. MK-42.2022.4) and the Leading scientific school (NSH-17.2022.1.6). Section IV is supported by the Russian Science Foundation (project No. 21-71-00132).

REFERENCES

[1] M. Cooke, J. Barker, S. Cunningham, en X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition", *The Journal of the Acoustical Society of America*, vol 120, no 5, pp. 2421–2424, 2006.

[2] J. S. Chung en A. Zisserman, "Lip reading in the wild", in *Asian conference on computer vision*, 2016, bll 87–103.

[3] J. S. Chung, A. Senior, O. Vinyals, en A. Zisserman, "Lip reading sentences in the wild", in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 3444–3453.

[4] Z. Zhou, G. Zhao, X. Hong, en M. Pietikäinen, "A review of recent advances in visual speech decoding", Image and vision computing, vol 32, no 9, pp. 590–605, 2014.

[5] J. S. Chung en A. P. Zisserman, "Lip reading in profile", 2017.

[6] T. Stafylakis en G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading", arXiv preprint arXiv:1703. 04105, 2017.

[7] T. Stafylakis, M. H. Khan, en G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs", Computer Vision and Image Understanding, vol 176, pp. 22–32, 2018.

[8] B. Martinez, P. Ma, S. Petridis, en M. Pantic, "Lipreading using temporal convolutional networks", in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6319–6323.

[9] P. Ma, B. Martinez, S. Petridis, en M. Pantic, "Towards practical lipreading with distilled and efficient models", in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 7608–7612.

[10] Y. M. Assael, B. Shillingford, S. Whiteson, en N. De Freitas, "Lipnet: End-to-end sentence-level lipreading", arXiv preprint arXiv:1611. 01599, 2016.

[11] D. Ivanko, D. Ryumin, A. Axyonov, en A. Kashevnik, "Speaker-Dependent Visual Command Recognition in Vehicle Cabin: Methodology and Evaluation", in *International Conference on Speech and Computer*, 2021, pp. 291–302.

[12] S. Yang *et al.*, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild", in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019, pp. 1–8.

[13] B. Shillingford *et al.*, "Large-scale visual speech recognition", *arXiv preprint arXiv:1807. 05162*, 2018.

[14] T. Afouras, J. S. Chung, en A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition", *arXiv preprint arXiv:1809. 00496*, 2018.

[15] E. Ryumina, D. Ryumin, D. Ivanko, en A. Karpov, "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*, 2021.

[16] D. Ivanko, D. Ryumin, A. Axyonov, en M. Železny, "Designing advanced geometric features for automatic Russian visual speech recognition", in *International Conference on Speech and Computer*, 2018, pp. 245–254.

[17] A. Fernandez-Lopez en F. M. Sukno, "Survey on automatic lip-reading in the era of deep learning", *Image and Vision Computing*, vol 78, pp. 53–72, 2018.

[18] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, en A. Zisserman, "Deep audio-visual speech recognition", *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[19] T. Afouras, J. S. Chung, en A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition", *arXiv preprint arXiv:1809. 00496*, 2018.

[20] H. Chen, W. Xie, A. Vedaldi, en A. Zisserman, "Vggsound: A large-scale audio-visual dataset", in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[21] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, en M. Szykulski, "An audio-visual corpus for multimodal automatic speech recognition", *Journal of Intelligent Information Systems*, vol 49, no 2, pp. 167–192, 2017.

[22] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, en R. He, "Deep audio-visual learning: A survey", *International Journal of Automation and Computing*, vol 18, no 3, pp. 351–376, 2021.

[23] M. Zhou, Z. Qin, X. Lin, S. Hu, Q. Wang, en K. Ren, "Hidden voice commands: Attacks and defenses on the VCS of autonomous driving cars", *IEEE Wireless Communications*, vol 26, no 5, pp. 128–133, 2019.

[24] A. Kashevnik *et al.*, "Multimodal corpus design for audio-visual speech recognition in vehicle cabin", *IEEE Access*, vol 9, pp. 34986–35003, 2021.

[25] D. Ivanko, Axyonov A., Ryumin D., Kashevnik A., Karpov A., "Multi-Speaker Audio-Visual Corpus RUSAVIC: Russian Audio-Visual Speech in Cars", in LREC 2022 Conference, 2022, pp. 1-5. In press.

[26] Y. Zhang, S. Yang, J. Xiao, S. Shan, en X. Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition", in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 356–363.

[27] Y. Kartynnik, A. Ablavatski, I. Grishchenko, en M. Grundmann, "Real-time facial surface geometry from monocular video on mobile GPUs", *arXiv preprint arXiv:1907. 06724*, 2019.

[28] D. Ivanko en D. Ryumin, "Development of Visual and Audio Speech Recognition Systems Using Deep Neural Networks".

[29] D. Feng, S. Yang, S. Shan, en X. Chen, "Learn an effective lip reading model without pains", arXiv preprint arXiv:2011. 07557, 2020.

[30] M. Kim, J. Hong, S. J. Park, en Y. M. Ro, "Multi-modality associative bridging through memory: Speech sound recollected from face video", in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 296–306.

[31] B. Xu, C. Lu, Y. Guo, en J. Wang, "Discriminative multi-modality speech recognition", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14433–14442.

[32] X. Zhao, S. Yang, S. Shan, en X. Chen, "Mutual information maximization for effective lip reading", in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 420–427.

[33] P. Wiriyathammabhum, "SpotFast networks with memory augmented lateral transformers for lipreading", in International Conference on Neural Information Processing, 2020, pp. 554–561.

[34] J. Xiao, S. Yang, Y. Zhang, S. Shan, en X. Chen, "Deformation flow based two-stream network for lip reading", in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 364–370.

[35] M. Luo, S. Yang, S. Shan, en X. Chen, "Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading", in 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), 2020, pp. 273–280.

[36] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, en M. Pantic, "End-to-end audiovisual speech recognition", in 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2018, pp. 6548–6552.

[37] C. Wang, "Multi-grained spatio-temporal modeling for lip-reading", arXiv preprint arXiv:1908. 11618, 2019.