IMPROVED AUTOMATIC LIP-READING BASED ON THE EVALUATION OF INTENSITY LEVEL OF SPEAKER’S EMOTION.

D. Ivanko, E. Ryumina, D. Ryumin

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation – denis.ivanko11@gmail.com, ryumina_ev@mail.ru, ryumin.d@ iias.spb.su

Commission II, WG II/8

KEY WORDS: Automatic lip-reading, Emotion recognition, Intelligent Video Analytics, Computer Vision, Human-Machine Interaction.

ABSTRACT:

Automatic audio-visual speech recognition systems (AVSRs) have recently achieved tremendous success, especially in limited vocabulary tasks by far surpassing human abilities to recognize speech, especially in acoustically noisy conditions. Automatic speech recognition systems based on processing of audio and video information are being actively researched and developed all over the world. However, scientific studies aimed at analyzing the influence of the speaker's emotional state (anger, disgust, fear, happy, neutral, and sad), and, most importantly, intensity level of emotion (low - LO, medium - MD, high - HI) on automatic lip-reading have not been conducted. In this regard, the relevance of this research topic cannot be overestimated and requires detailed study. In this paper, we present a novel approach for emotional speech lip-reading, that includes evaluation of a speaker’s emotion and its intensity level. The proposed approach uses visual speech data to detect a person’s emotion type and its intensity level and based on this information assigns it to one of the trained emotional lip-reading models. This essentially resolves the multi-emotional lip-reading issue associated with most real-life scenarios. The proposed approach improves the state-of-the-art results due to the consideration of the intensity of the pronounced audio-visual speech up to 8.2% in terms of the accuracy. Current research is the first step in the creation of emotion-robust speech recognition systems and leaves open a wide field for further research.

1. INTRODUCTION

Speech changes can be observed both in audio and video modalities. For example, with emotions “happy” and “anger”, words are pronounced with a more open mouth than with emotions “neutral” and “sad”, in addition with the emotion "anger" the pauses between words are shorter than with the emotion "sadness". All these micro-changes lead to the fact that speech is not recognized correctly by modern automatic recognition system, although it is often very important for a person to be understood in the emotional state by devices/medical equipment.

Currently, leading scientific institutions and global industrial corporations working in the field of artificial intelligence are actively conducting research aimed at creating highly efficient visual speech recognition systems. Automatic audio-visual speech recognition systems (AVSRs) have recently achieved tremendous success, especially in limited vocabulary tasks by far surpassing human abilities to recognize speech, especially in acoustically noisy conditions (Ivanko et al., 2019). However, the efficiency of state-of-the-art AVSRs is significantly deteriorating due to a number of factors, one of which is the speaker's emotions (Ryumin et al., 2023). Depending on the speaker's emotion, changes: timbre, pitch and loudness of the voice, duration of sounds, duration of pauses, articulation, etc.

Despite the progress of digital technologies achieved in recent years, automatic speech recognition systems are not always able to function with high performance (accuracy, recognition speed). In the presence of controlled office conditions, a limited vocabulary and a controlled grammar of recognized commands, the accuracy of modern speech recognition systems in terms of audio modality (sounding speech) can approach 100%. However, in the case of a complex dynamic acoustic environment (external noise, reverberation, interference in the microphone channel, etc.), the accuracy of automatic speech recognition is significantly reduced. Based on the foregoing, we can assume that the acoustic speech signal is the main modality in automatic speech recognition systems, but in addition to it, it also makes sense to use visual information about speech (the movements of the speaker's lips), to make it is possible to improve accuracy of speech recognition, especially in conditions where the acoustic signal is noisy or unavailable.

Despite the fact that audio signals are generally much more informative than video signals, numerous experiments have shown that most people use lip reading to better understand the interlocutor's speech. However, this often happens unconsciously and depends to varying degrees on aspects such as the person's auditory abilities or acoustic conditions (for example, the visual channel becomes more important in noisy environments). In addition, the visual channel is the only source of information for people with hearing impairments to understand oral speech.

Visual speech recognition (lip reading) is a rather difficult skill for a person, however, in acoustically noisy conditions and when a large number of people are talking, the interlocutors themselves begin to pay attention to each other's lips in order to better understand the meaning of statements. Human speech perception is a multi-modal process, and based on this, in recent years, it has been possible to improve the efficiency of automatic speech recognition systems, due to the availability of representative audio-visual corpora and the improvement of neural network architectures. Automatic speech recognition, especially in noisy environments, is a rather difficult task. One of the most important steps in speech recognition is the correct determination of speech boundaries in the incoming audio stream. For isolated words, this problem comes down to finding the correct boundary between words, but if we are talking about continuous speech, then this task is much more difficult, due to the fact that it is a continuous stream, usually with a minimum pause.

In this paper, we present a novel approach for emotional speech lip-reading, that includes evaluation of a speaker’s emotion and its intensity level. The proposed approach uses visual speech data to detect a person’s emotion type and its intensity level and based on this information assigns it to one of the trained emotional lip-reading models. This essentially resolves the multi-emotional lip-reading issue associated with most real-life scenarios.

2. RELATED WORK

With recent developments of neural network models, and more specifically with the introduction of such deep neural network architectures such as VGG (Chung et al., 2017) and ResNet-like (Stafylakis et al., 2017) that are able to consume raw data without a feature extraction phase, modern emotion speech recognition approaches started to shine. For the last five years numerous research works have been published, e.g. (Feng et al., 2020, Martinez et al., 2020, Ivanko et al., 2022b, Ivanko et al., 2022c). In existing deep learning emotion recognition models for recognizing spatial-temporal input, there are three common topologies: CNN-RNN (Deng et al., 2020), 3DCNN (Huang et al., 2020), and Two-Stream Network (Schoneveld et al., 2023).

Traditionally, automatic lip-reading systems were based on the extraction of visual features, classification and modeling of speech sequences. Thus, traditional systems mainly consist of image preprocessing and feature extraction combined with hidden markov models (HMMs) that use concise context information to model the temporal dynamics of a signal. The first lip-reading systems of the speaker solved simple tasks, such as recognizing individual letters or numbers. However, over time, researchers gradually moved on to more complex and realistic scenarios, which eventually led to the emergence of modern systems designed for automatic lip reading of continuous speech. To a large extent, these advances were made possible by the creation of powerful systems based on deep learning architectures, which were rapidly replacing traditional systems.

Recognition of audio-visual speech is gradually being replaced by an End-to-End integral approach, i.e. cascade of neural networks. In the first approximation, the End-to-End approach is close to traditional methods: a sequence of mouth images is fed into a convolutional neural network to extract features (Xu et al., 2022), which are then transferred to an internal model (RNN, LSTM, GRU, etc.) to take into account time-dependence and classification. The conducted studies demonstrate that the features extracted in this way are more suitable for automatic lip reading than those calculated by traditional methods.

The main advantage of the modern approach is that the entire system consists of a single neural network. Thus, the extracted features are better related to the data on which the network is trained. In (Noda et al., 2014), it was first proposed to use CNN to replace the feature extraction block. In turn, in (Hochreiter et al., 1997), it was first proposed to use LSTM for the classification problem. Later, researchers in (Petridis et al., 2018) proposed a neural network for extracting acoustic features and tried to combine them with video information.

CNN-RNN combines the advantages of both transferred knowledge of a pre-trained convolutional network and a temporal modeling capability. The input features to the RNN are usually abstract and global features represented by higher layers. It makes this architecture able to extract larger and more sustained changes in facial appearance (macro-motion). The 3DCNN combines information over both space and time using convolutional filters starting from the lowest layers. This enables it to capture both macro and micro motions. However, it cannot incorporate transferred knowledge as conveniently as CNN-RNN. Two-Stream Network contains two parallel convolutional networks: a network that processes images and a temporal network that processes motions

Nowadays there are a lot of corpora containing emotionally-colored speech data (Ivanko et al., 2022a). At the same time, there are a lot of datasets aimed for lip-reading (Kashevnik et al., 2021). However, at the moment there are almost no combined emotional lip-reading databases suitable for model training in the scope of deep learning approaches. Despite the variety of existing emotional datasets, there are at least four corpora suitable for automatic reading visual speech by lips: CREMA-D (Cao et al., 2014), RAVDESS (Livingstone et al., 2018), and eNTERFACE’05 (Martin et al., 2006). CREMA-D is the more promising due to the number of various speakers and amount of data available. The authors highlight the following problems of automatic recognition of visual speech:
- stable detection of the area of interest (mouth area);
- extraction of the most informative features from visual speech;
- effective modeling and recognition of the speaker's visual speech (both isolated words and continuous speech).

A number of recent works (Zhang et al., 2020) are devoted to methods for extracting visual features from a previously detected area of interest by combining different architectures of neural networks with linear classifiers. It should be noted that the stage of detecting the speaker's mouth area has a significant impact on the final efficiency of visual speech recognition. The most common solutions for this problem (basic approaches) include methods based on Haar primitives and methods based on active appearance models (Viola et al., 2001).

Several training strategies and deep neural networks have been recently proposed for lip-reading of isolated words (Shi et al., 2022). It received a lot of attention due to the availability of large publicly available corpora, e.g., LRW and LRS (Chung et al., 2017). The majority of state-of-the-art approaches follow the similar lip-reading strategy that consists of a visual encoder, followed by a temporal model and a classification layer (Ma et al., 2022). A visual encoder was initially proposed in (Stafylakis et al., 2017), and since then has been widely used and improved in following works (Martinez et al., 2020). At the same time, the most recent advances include the temporal model and the training strategy. Bidirectional (Bi) GRUs and LSTMs, and Multi-Scale Temporal Convolution Networks have been the most popular temporal models (Vaswani et al., 2017).

3. DATASET

In current research we use the CREMA-D dataset (Cao et al., 2014). The corpus contains 7440 audio-visual recordings from 91 speakers, age range 20-60+. The average duration of one recording is 2.5 seconds with approximately 75 frames. The database contains people of different races and ethnicities (African American, Asian, Caucasian, etc.). Speakers uttered from a selection of 12 sentences (Table 1). The sentences were pronounced using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and with 3 levels of intensity of emotions: Low (LO), Medium (MD) and High (HI).

Abbreviation	Phrase
1. IEO	It's eleven o'clock
2. TIE	That is exactly what happened
3. IOM	I'm on my way to the meeting
4. IWW	I wonder what this is about
5. TAI	The airplane is almost full
6. MTI	Maybe tomorrow it will be cold
7. IWL	I would like a new alarm clock
8. ITH	I think I have a doctor's appointment
9. DFA	Don't forget a jacket
10. ITS	I think I've seen this before
11. TSI	The surface is slick
12. WSI	We'll stop in a couple of minutes

Table 1: Phrases of CREMA-D dataset.

We split the corpus into training, validation and test sets in the ratio of 70, 10 and 20, respectively, considering speaker independence, gender and age.

We divide the training set into 6 parts according to emotions. At first, we train the model only on records with neutral emotions. We do not divide the validation and test sets according to emotions. Initially, we trained the model on neutral emotions. We fine-tuned the training parameters, such as: sequence length, image size, batch size, number of image channels, prior to training the model on other emotions

4. METHODOLOGY

Since there is no out-of-the-box solution to process emotional speech, in current research we use the well-known 3D ResNet-18 to tackle emotional speech recognition. 3D ResNet is a type of model for video that employs 3D convolutions. ResNet includes 17 3D convolutional layers that make it relatively easy to increase accuracy by increasing depth, which is more difficult to achieve with other networks. The input of the CNN is an image that passes through the first 3D convolution layer and the pooling layer, then 4 residual blocks with 3D convolution layers follow, each of which is re-peated 2. The global average pooling layer is next, and a fully connected layer of 12 neurons completes the CNN. The last fully connected layer determines the most probable hypothesis from 12 recognition classes.

We used the Mediapipe open-source library to detect lips areas. To train the models, a window size of 30/60 frames with a step two times smaller than the specified window size was used. Frames were selected sequentially from the video with-out thinning.

We apply an pre-processing pipeline similar that includes: (1) video downsampling to five frames per seconds (FPS) to ensure the same processing conditions for recurrent neural networks in terms of temporality; (2) channel normalization of the image; (3) resizing the image to 224×224 pixels. Lip-reading pre-processing includes detection of a mouth region-of-interest with the same approach (MediaPipe) on each frame and cropping it. We resized the mouth image to 88×88×3. In order to maintain mouth proportions, we pad the image to the desired size with the average pixel value along the vertical of the image itself. We then use a linear normalization technique. We do not apply any FPS downsampling at this stage.

As a performance indicator we used Recall, which is calculated as the proportion of correctly predicted phrases to the total number of phrases, related, for example, to phrases reproduced with an emotion of anger with high intensity.

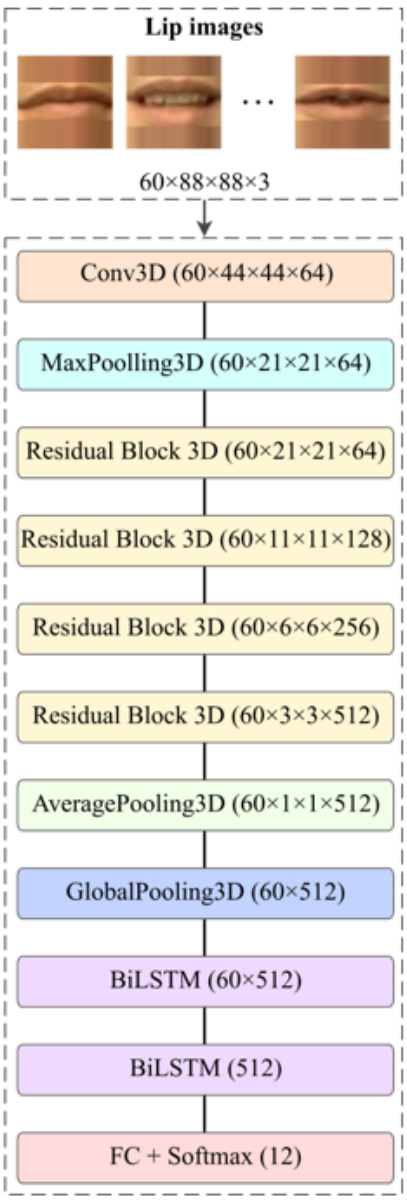

Figure 1: Proposed lip-reading architecture

Emo-tion	ANG			DIS			FEA			HAP			SAD			NEU	mean Recall
	LO	MD	HI	LO	MD	HI	LO	MD	HI	LO	MD	HI	LO	MD	HI		
NEU	0	11	0	33	11	22	0	11	11	22	11	33	44	44	55	77	24
ANG	88	100	100	77	100	88	66	66	100	55	66	66	100	66	100	100	84
DIS	22	66	44	66	66	55	22	22	22	44	66	66	55	55	66	88	52
FEA	44	88	77	55	77	77	55	100	100	55	44	88	100	77	100	100	77
HAP	11	44	11	44	55	55	22	11	11	66	66	77	22	22	33	55	38
SAD	55	55	22	66	66	55	66	77	55	77	66	66	100	77	100	100	69
mean Recall	37	61	42	57	62	59	38	48	50	53	53	66	70	57	75	87	-
ALL	100	100	100	77	100	100	100	88	100	100	88	100	100	100	100	100	-

Table 2**:** IEO phrase recognition results in terms of intensities for the validation set.

We use the 3DCNN and BiLSTM to tackle emotional speech lip-reading. 3DCNN is a type of model for video that employs 3D residual blocks. We show the general architecture and layers dimensions of the model in Figure 1. 3DCNN includes seventeen 3D convolutional layers that makes it relatively easy to increase accuracy by increasing depth, which is more difficult to achieve with other networks. The BiLSTM network consists of two BiLSTM layers with 512 consecutive neurons in each layer. 3DCNN inputs an image that passes through the first 3D convolution layer and the pooling layer, then four residual blocks with 3D convolution layers follow, each of which is repeated twice. The average and global pooling layers are next, two LSTM layers, and a fully connected layer of twelve neurons completes the 3DCNN-BiLSTM network. The last fully connected layer determines the most probable hypothesis from twelve recognition classes.

5. EXPERIMENTAL EVALUATION

To analyze the recognition accuracy of the phrase "IEO": "It's eleven o'clock" within three intensities LO - Low, MD - Medium, HI - High. And for the neutral state, the intensity of XX is Unspecified. There are 144 IEO phrases in the validation set. Phrases reproduced with intensity XX (for the neutral state) - 9, LO - 45, MD - 45, HI - 45.

In Tables 2 and 3, the performance indicator is Recall, which is calculated as the proportion of correctly predicted phrases (IEO) to the total number of phrases (IEO), related, for example, to phrases reproduced with an emotion of anger intensity LO, etc.

The training was carried out on 100 epochs, the learning rate is 0.001 and it is constant throughout the training process; the optimizer is SGD. Training stops if UAR does not increase during 6 epochs on validation. We chose the unweighted average recall (UAR) metric because it is better suited for unbalanced classes, e.g. we have IEO class three times larger than the others

The results of the mean Recall experiments (vertical) show that when learning on phrases reproduced with the emotion of anger, the maximum value of mean Recall is 84. This is well explained by the fact that in a state of anger a person wants to be understood as best as possible, therefore articulation and speech become clearer. Whereas the minimum value of mean Recall equal to 24 is achieved when training on phrases reproduced in the neutral state. This is well explained by the fact that when training on phrases in the neutral state, there is no intensity, which means that the amount of training data for the IEO phrase is halved and it is more difficult to recognize them.

The results of the mean Recall experiments (horizontal) show that the IEO phrase is best recognized at LO and HI intensities and with sad emotion. Phrase recognition IEO at MD intensity is best recognized with emotions of anger and disgust.

According to the results of experiments on the selection of training parameters for the neural network, it was revealed that the highest UAR value for recognizing 12 phrases is achieved when: 1) the image size is 88×88×3, where 3 is the number of channels; 2) sequence length equal to 60; 3) batch size - 2.

If we look at mean UAR metrics, it turns out that it is best to train on phrases played with the Fear emotion (FEA). And worst of all, they learn from phrases played with the Disgust emote (DIS). At the same time, if we look at mean UAR, we will see that no matter what emotion the phrases are reproduced in the

Emo -tion	ANG			DIS			FEA			HAP			SAD			NEU	mean Recall
	LO	MD	HI	LO	MD	HI	LO	MD	HI	LO	MD	HI	LO	MD	HI		
NEU	26	31	0	42	21	26	42	42	36	26	5	15	57	47	42	47	31
ANG	94	89	94	47	52	73	84	84	89	68	57	68	84	68	78	89	76
DIS	63	52	36	47	42	57	42	52	63	73	68	68	57	57	68	63	57
FEA	68	68	63	47	68	73	73	63	84	73	57	63	73	78	78	68	69
HAP	42	57	57	57	57	57	52	42	42	63	84	78	68	31	52	36	55
SAD	84	84	63	63	78	78	78	78	57	63	57	57	100	78	73	89	74
mean Recall	63	64	52	50	53	61	62	60	62	61	55	58	73	60	65	65	-
ALL	100	100	100	89	100	94	94	100	94	100	94	94	89	98	84	100	-

Table 3**:** IEO phrase recognition results in terms of intensities for the test set.

training sample, then for the test sample, the accuracy is higher for phrases reproduced with the emotion of Anger, least of all - with the emotion of Disgust. For the validation set, the highest is in the neutral state, and the least is in the Happiness emotion.

In the test set, the number of IEO phrases equals to 304. Phrases reproduced with intensity XX (for the neutral state) are 19, LO are 95, MD are 95, and HI are 95. The results of the mean Recall experiments (vertical) show that when training on phrases reproduced with the emotion of anger, the maximum value of mean Recall is 76. Whereas the minimum value of mean Recall equal to 31 is achieved when learning on phrases reproduced in the neutral state. The results of the mean Recall experiments (horizontal) show that the IEO phrase is best recognized at LO and HI intensities and with sad emotion. IEO phrase recognition at MD intensity is best recognized with anger emotions.

6. CONCLUSIONS

We presented and studied the novel approach for automatic lip-reading based on the evaluation of the intensity level of the speaker's emotion and compared the results with the classical approaches. We conducted experimental investigations that demonstrated how different classes of emotions and the intensity of emotions affect automatic lip-reading. The proposed approach improves the state-of-the-art results due to the consideration of the intensity of the pronounced audio-visual speech up to 8.2% in terms of the accuracy. Current research is the first step in the creation of emotion-robust speech recognition systems and leaves open a wide field for further research.

ACKNOWLEDGEMENTS

This research is financially supported by the Russian Science Foundation (https://rscf.ru/en/project/21-71-00132/, No. 21-71-00132)

REFERENCES

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE transactions on affective computing, 5(4), 377-390.

Chung, J. S., & Zisserman, A. 2017. Lip reading in the wild. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, 87-103.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. 2017. Lip reading sentences in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition, 6447-6456.

Deng, D., Chen, Z., Zhou, Y., & Shi, B. 2020. Mimamo net: Integrating micro-and macro-motion for video emotion recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 03, 2621-2628.

Feng, D., Yang, S., Shan, S., & Chen, X. 2020. Learn an effective lip reading model without pains. arXiv preprint arXiv:2011.07557.

Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. 2020. Multimodal transformer fusion for continuous emotion recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 3507-3511.

Ivanko, D., Axyonov, A., Ryumin, D., Kashevnik, A., & Karpov, A. 2022a. RUSAVIC Corpus: Russian audio-visual speech in cars. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 1555-1559.

Ivanko, D., Ryumin, D., Kashevnik, A., Axyonov, A., Kitenko, A., Lashkov, I., & Karpov, A. 2022b. DAVIS: Driver's Audio-Visual Speech Recognition. In ISCA Annual Conference Interspeech, 1141-1142.

Ivanko, D., Ryumin, D., & Karpov, A. 2019. AUTOMATIC LIP-READING OF HEARING IMPAIRED PEOPLE. International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences.

Ivanko, D., Ryumin, D., Kashevnik, A., Axyonov, A., & Karnov, A. 2022c. Visual Speech Recognition in a Driver Assistance System. In 2022 30th European Signal Processing Conference (EUSIPCO), pp. 1131-1135.

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.

Livingstone, S. R., & Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), 0196391.

Luna-Jiménez, C., Kleinlein, R., Griol, D., Callejas, Z., Montero, J. M., & Fernández-Martínez, F. 2021. A proposal for multimodal emotion recognition using aural transformers and action units on RAVDESS dataset. Applied Sciences, 12(1), 327.

Ma, P., Wang, Y., Petridis, S., Shen, J., & Pantic, M. 2022. Training strategies for improved lip-reading. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8472-8476.

Martin, O., Kotsia, I., Macq, B., & Pitas, I. 2006. The eNTERFACE'05 audio-visual emotion database. In 22nd international conference on data engineering workshops (ICDEW'06), 1-8.

Martinez, B., Ma, P., Petridis, S., & Pantic, M. 2020, May. Lipreading using temporal convolutional networks. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6319-6323.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H. G., & Ogata, T. 2014. Lipreading using convolutional neural network. In fifteenth annual conference of the international speech communication association.

Pan, X., Ying, G., Chen, G., Li, H., & Li, W. 2019. A deep spatial and temporal aggregation framework for video-based facial expression recognition. IEEE Access, 7, 48807-48815.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. 2018. End-to-end audio-visual speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), 6548-6552.

Ryumin, D., Ivanko, D., & Ryumina, E. 2023. Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices. Sensors, 23(4), 2284.

Schoneveld, L., Othmani, A., & Abdelkawy, H. 2021. Leveraging recent advances in deep learning for audio-visual emotion recognition. Pattern Recognition Letters, 146, 1-7.

Shi, B., Hsu, W. N., & Mohamed, A. 2022. Robust self-supervised audio-visual speech recognition. arXiv preprint arXiv:2201.01763.

Stafylakis, T., & Tzimiropoulos, G. 2017. Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems, 30.

Viola, P., & Jones, M. 2001. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001.

Xu, B., Wang, J., Lu, C., & Guo, Y. 2020. 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, 1637-1646.

Zhang, Y., Yang, S., Xiao, J., Shan, S., & Chen, X. (2020, November). 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), 356-363.