Jurnal Kejuruteraan 36(1) 2024: 147-154
https://doi.org/10.17576/jkukm-2024-36(1)-14

Deep Learning-Based Audio-Visual Speech Recognition for Bosnian Digits

Husein Fazlić[a], Ali Abd Almisreb[a] & Nooritawati Md Tahir[b,c,d*]

[a]*Department of Computer Science and Engineering, Faculty of Engineering and Natural Sciences, International University of Sarajevo, Sarajevo, Bosnia and Herzegovina*

[b]*Applied College, Princess Nourah bint Abdulrahman University (PNU), Riyadh, Saudi Arabia*

[c]*College of Engineering, Universiti Teknologi MARA, Selangor, Malaysia.*

[d]*Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Universiti Teknologi MARA, Selangor, Malaysia.*

[*]*Corresponding author: nooritawati@ieee.org*

Received 10 April 2023, Received in revised form 2 June 2023
Accepted 2 July 2023, Available online 30 January 2024

ABSTRACT

This study presents a deep learning-based solution for audio-visual speech recognition of Bosnian digits. The task posed a challenge due to the lack of an appropriate Bosnian language dataset, and this study outlines the approach to building a new dataset. The proposed solution includes two components: visual speech recognition, which involves lip reading, and audio speech recognition. For visual speech recognition, a combined CNN-RNN architecture was utilised, consisting of two CNN variants namely Google Net and ResNet-50. These architectures were compared based on their performance, with ResNet-50 achieving 72% accuracy and Google Net achieving 63% accuracy. The RNN component used LSTM. For audio speech recognition, FFT is applied to obtain spectrograms from the input speech signal, which are then classified using a CNN architecture. This component achieved an accuracy of 100%. The dataset was split into three parts namely for training, validation, and testing purposes such that 80%, 10% and 10% of data is allocated to each part, respectively. Furthermore, the predictions from the visual and audio models were combined that yielded 100% accuracy based on the developed dataset. The findings from this study demonstrate that deep learning-based methods show promising results for audio-visual speech recognition of Bosnian digits, despite the challenge of limited Bosnian language datasets.

Keywords: Audio-visual speech recognition; deep learning; convolutional neural networks (CNN); recurrent neural networks (RNN); lip reading; Bosnian language

INTRODUCTION

Human communication is predominantly carried out through speech, which varies across different linguistic groups. Bosnian, a member of the Indo-European language family and the Slavic language group, is composed of thirty letters, each with a corresponding sound or phoneme. While the Latin script is more commonly used to write Bosnian, Cyrillic is also accepted (Jahić Dževad et al. 2000). Additionally, audio-visual speech recognition requires the recognition of speech through both audio and visual means. One such visual approach is lip reading, which is also used to identify speech (Fenghour et al. 2021). While the ears are usually relied upon to receive speech information, audio speech recognition may be more comfortable for some individuals.

Conversely, lip reading is a widely used technique nowadays, particularly during major soccer matches when coaches and players cover their mouths to keep their tactics undisclosed from their opponents. Skilled individuals can decipher spoken words by observing the movements of a person's lips. This method has also been adopted by computers, which can be trained to differentiate between different words after being fed large amounts of data. However, labelling of data is necessary to ensure that the machine learns the different classes correctly.

The history of audio-visual datasets dates back to the 1990s (Fenghour et al. 2021). Most of these datasets are in the English language, with the most famous one being the Lip Reading in the Wild (LRW) dataset, which was obtained from various clips shown on the BBC (Chung & Zisserman, 2017). This dataset has been used in conjunction with different architectures (Afouras et al. 2018a; Makino et al. 2019; Ma et al. 2021; Serdyuk et al. 2021; Sterpu & Harte 2018; Yu et al. 2020, Asghar A et al. 2022), with speech in the dataset being either isolated or continuous. Continuous speech datasets may contain partial occurrences of previous words or variations in mouth shape. Speakers in these datasets may be captured from frontal, angled, and side-view perspectives. The most commonly used orientation is a frontal view, often rotated at a slight angle. In the recent years, there has been a shift from recording in laboratory settings towards building pipelines for automatic dataset creation, cropping, and labelling (Chung & Zisserman 2017; Makino et al. 2019). Several datasets have been created in Arabic, Polish, Russian, French, and other languages.

A novel approach to speech recognition involves the use of deep learning, which utilizes deep neural networks to solve problems in machine learning. These neural networks can be quite large and typically consist of an input layer, one or more hidden layers, and an output layer. The defining characteristic of deep neural networks is the number of hidden layers they incorporate. With advances in processing power, researchers have been able to propose increasingly larger architectures. This technique has been used in many studies (Abedalla A et al. 2021; Alzubaidi et al. 2021; Afouras et al. 2018b; Michelucci, 2019; Shashidhar et al. 2020; Szegedy et al. 2014), with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) as the most preferred architectures. CNN is used for feature extraction, followed by a variation of RNN that can capture the time-based nature of data. In audio classification, Fast Fourier Transform (FFT) is used to convert the problem into the realm of image classification (Afouras et al. 2018b; Kashevnik et al. 2021).

The classical approach to modelling sequential data is through Markov Chains, with Hidden Markov Models (HMM) being used for speech recognition (Jurafsky & Martin, 2020). HMMs assume that the probability of a chain being in a certain state depends only on a defined number of previous states. Many problems involve hidden states that are not directly observable, and only their corresponding outputs can be observed. Such sequences of hidden states are referred to as HMM. For instance, HMM has been applied to speech recognition, as evidenced in these studies (Jadczyk 2018; Yu et al. 2020).

METHODOLOGY

PARTICIPANTS

To the extent of our knowledge, there is currently no publicly available audio-visual speech Bosnian language dataset, thus, it is necessary to acquire and develop one. The dataset comprises isolated speech sounds of digits in the Bosnian language. To create this dataset, we recruited five volunteers as subjects, four of whom were male and one of whom was female. Four of the volunteers were native speakers of Bosnian, and one was not. Their ages ranged from 10 to 55 years. For each digit class in the dataset, fifteen utterances were collected, resulting in a total of one hundred and fifty samples. The classes in the dataset are presented in Table 1. During the preprocessing phase, the dataset was augmented by adding different types of noise, expanding the total number of samples to seven hundred and fifty. As a result, each class consisted of seventy-five samples.

Each speech clip in the dataset is approximately two seconds long, with speakers positioned in front of a white background. The video captured both the speaker's head and the background (wall), with a resolution of 1280x720 (HD standard) and a frame rate of 30 frames per second. Audio is recorded at a frequency of 48 kHz. The speech videos are stored on a hard drive in separate folders labelled by class (digit). Figure 1 shows several frames of the first author, as one of the subjects during data acquisition session.

TABLE 1. Classes in the dataset

Bosnian	English	Bosnian	English
nula	zero	*pet*	five
jedan	one	šest	six
dva	two	*sedam*	seven
tri	three	*osam*	eight
četiri	four	*devet*	nine

FIGURE 1. Sample of lips movement acquired as database during recording sessions

The proposed architecture of the system is presented in Figure 2. The video frames served as inputs to the CNN for feature extraction along with Google Net and ResNet-50. The output from this stage are sequences that served as inputs to the RNN network. Google Net and ResNet-50 are chosen in this study because these two models have proven to be effective in extracting meaningful features from input data and also these networks are able to learn complex representations while minimizing the number of parameters (Szegedy et al. 2014; He et al. 2015).

Further, the Long Short-Term Memory (LSTM) network is used based on its output probability predictions. In audio speech classification, we will convert the speech signal into spectrograms using the FFT. The resulting spectrograms will then be inputs to the CNN for classification.

EXPERIMENT DESIGN AND PROCEDURE

VISUAL SPEECH CLASSIFICATION

The pre-processing steps is to be described in detail here, as depicted in Figure 3. Firstly, the original video is often too large and contains unnecessary information. Therefore, the Viola-Jones algorithm is used to perform face detection in the videos (Atiqur et al. 2018). Then, in a series of steps, the mouth region is extracted, which is define as the region of interest (ROI). Finally, this ROI is resized to 224 x 224, which is the input size of the pre-defined CNN architectures.

To extract features from the videos, two state-of-the-art CNNs will be used: GoogLeNet and ResNet-50. Recall that these networks can be trained on large datasets and are capable of extracting numerous sets of features (He et al. 2015; Szegedy et al. 2014). The choice of CNN impacted the performance of the system, so both versions are implemented and their results are compared. Once the CNNs extract features, sequences that describe the video features will be obtained. These sequences are fed into the LSTM network, which is a type of RNN that can learn time-based dependencies in the videos. The final layer of the LSTM network is the softmax layer that assigns probability scores to different classes. This enabled the use of visual speech classification solely or combination with audio classification to produce a combined result.

AUDIO SPEECH CLASSIFICATION

As previously mentioned, the recordings include both audio and video components. The audio component consists of speech segments, which are not useful in the original time-domain form. Instead, these signals are converted into spectrograms using FFT, which utilises Fourier analysis to transform speech segments from a time-based representation to a frequency-based representation. Further, the resulting spectrograms are used for spectrogram (image) classification, as shown in Figure 4. The similarity of the shapes within a class can be observed in the top row, while the differences among classes are evident in the bottom row.

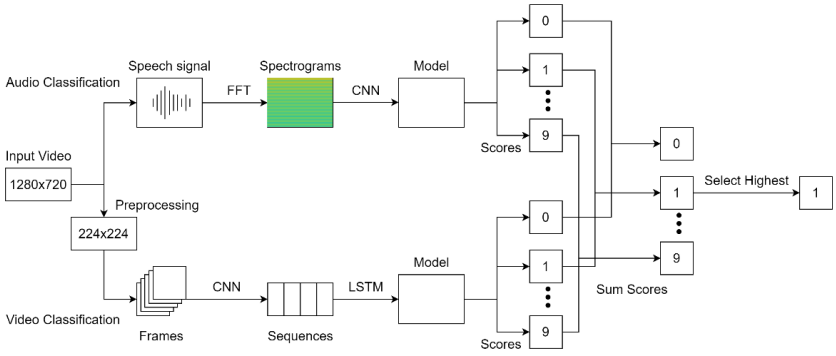

FIGURE 2. Architecture of audio-visual speech recognition system.

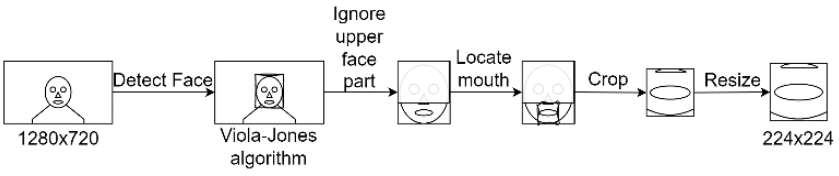

FIGURE 3. Sequence of pre-processing steps applied to original video.

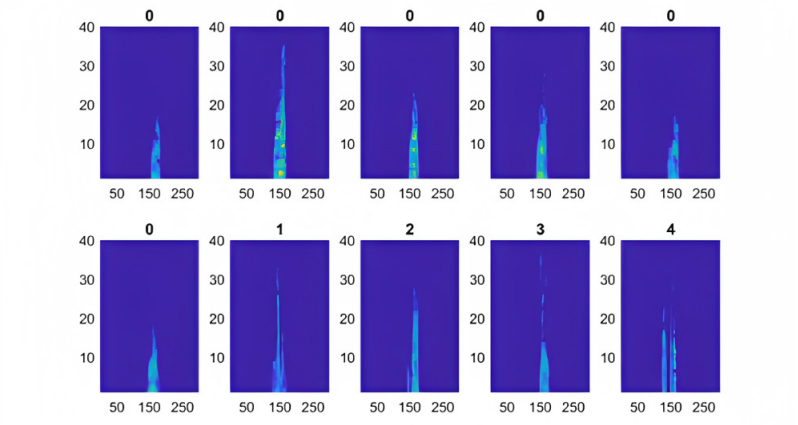

FIGURE 4. Spectrogram in one class (top) and in different classes (bottom)

RESULTS AND DISCUSSION

The performance of the dataset is based on the holdout method. The dataset is divided into three parts: training (80%), validation (10%), and testing (10%). The training and validation parts are used during the training phase, while the testing part is used to evaluate the performance of the model on unseen data. The model is trained for 300 epochs and validated once per epoch. However, it is important to note that due to the small size of the dataset, there may be an imbalance among the classes based on the random process of dividing the data into these three parts.

The basic performance metric used for evaluation include accuracy (Acc), which is defined as the ratio of correctly classified examples to the total number of examples. However, accuracy alone may not provide enough information. Therefore, confusion matrices are also used to display more complex relations among the classes. This helps to understand which classes are easily recognizable or otherwise, in terms of true positive (TP), false positive (FP), false negative (FN) and true negative (TN). In addition, specificity (Spec) and sensitivity (Sens) is also used as performance measures.

$$Accuracy = \frac{TN + TP}{TP + FN + TN + FP} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (4)$$

Visual speech recognition is found to perform worse than audio speech classification, which is not surprising based on previous studies (Ivanko & Ryumin, 2021; Shashidhar et al. 2020). The accuracy for audio speech classification is 100%, while for visual speech classification, the situation is not as optimistic. The two proposed variants of visual speech classification, namely Google Net and ResNet-50, do not perform as well as the audio counterpart. The accuracy rate for Google Net is 63% and for ResNet-50 is 72%.

In Figure 5, it is observed the difference between accuracy results on the training set (blue line) versus validation set (black line), that more accurately represents the model's true performance. These plots represent the model obtained for visual speech recognition using Google Net as a feature extractor and LSTM architecture as classifier.

The results of the visual speech recognition (VSR) using both Google Net and ResNet-50 are shown in Figure 6. Figure 6(a) displays the results of the Google Net, which has an accuracy of 63% in visual speech recognition. On the other hand, Figure 6(b) shows the results of the ResNet-50 variant, with 72% Acc in visual speech recognition. The situation is different for audio classification, as the results obtained were much higher compared to video classification. Furthermore, as tabulated in Table 2, results based on Spec and Sens showed that Res Net performed better as compared to Google Net. This is illustrated in Figure 7. In Figure 8, the results of audio speech recognition on a test set containing 75 samples in ten different classes are displayed. It is apparent that the model achieved a perfect classification based on all samples are correctly classified.

Performing audio-visual speech recognition (AVSR) involves combining predictions from both audio and visual speech recognitions. One way to achieve this is by adding the prediction scores from both techniques and select the highest value. The assumption is assigning equal weight to both predictions, resulted in better performance

compared to visual speech recognition, that further achieved accuracy equivalent to that of audio speech recognition. In both the GoogLeNet and ResNet variants, accuracy increased to 100%, matching that of audio predictions. Results for audio-visual speech recognition (AVSR) are displayed in Figure 9, with the performance of both GoogLeNet and ResNet-50 on the left and right, respectively. Both achieved comparable performance to audio speech classification. These results are comparable to those obtained from a similarly sized English dataset with a CNN-LSTM architecture similar to ours (Shashidhar et al. 2020).

TABLE 2. Performance Measure for visual speech recognition using GoogLeNet and ResNet

Class	Google Net		Res Net	
	Spec.	Sens.	Spec	Sens
0	0.99	0.20	0.96	0.80
1	1.00	0.33	0.94	0.89
2	0.95	0.77	0.98	0.69
3	0.96	0.57	1.00	0.57
4	0.88	1.00	0.97	0.83
5	0.99	0.71	0.97	0.86
6	1.00	0.42	0.98	0.75
7	0.87	0.75	0.97	0.50
8	0.89	0.40	0.97	0.60
9	1.00	0.43	0.94	0.57

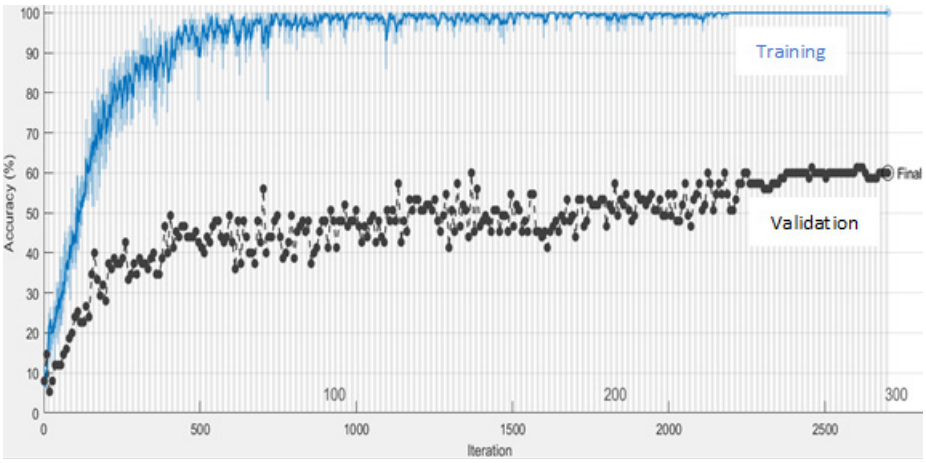

FIGURE 5. Network learning curve of Google Net in visual speech classification

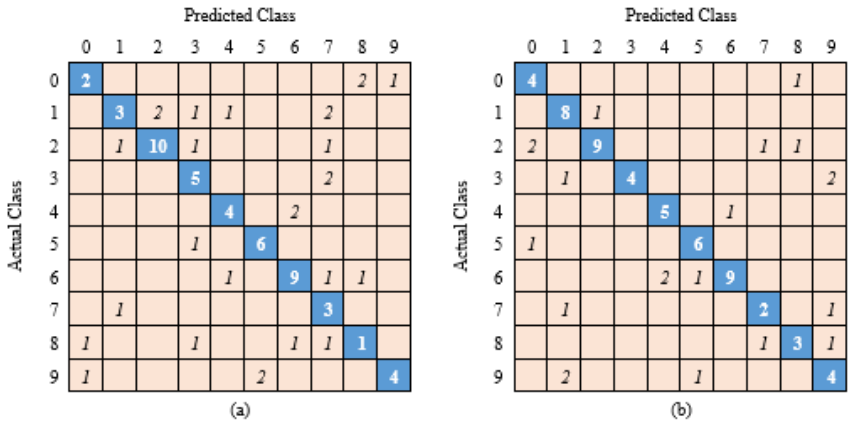

FIGURE 6. Visual speech recognition results: (a) Google Net; (b) ResNet-50

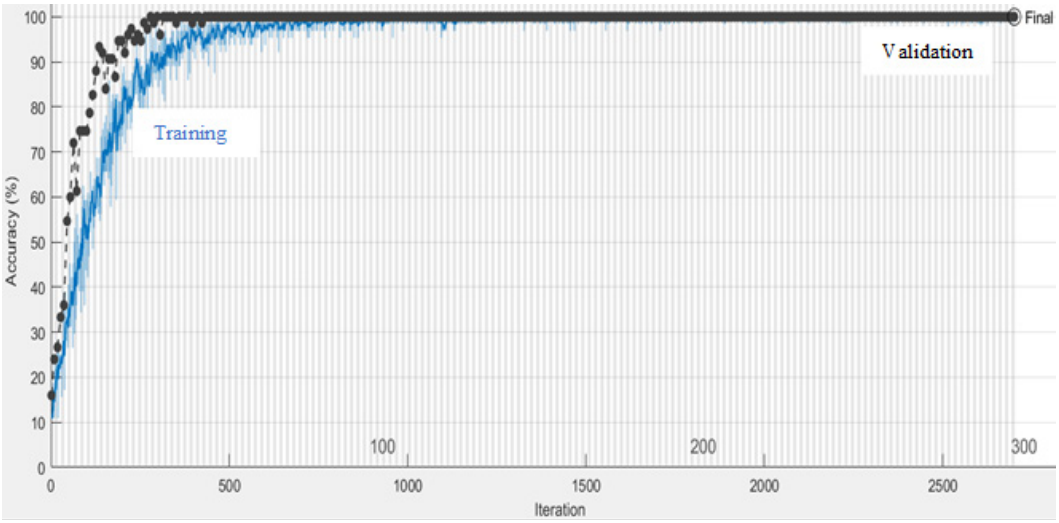

FIGURE 7. Network learning curve for audio speech classification

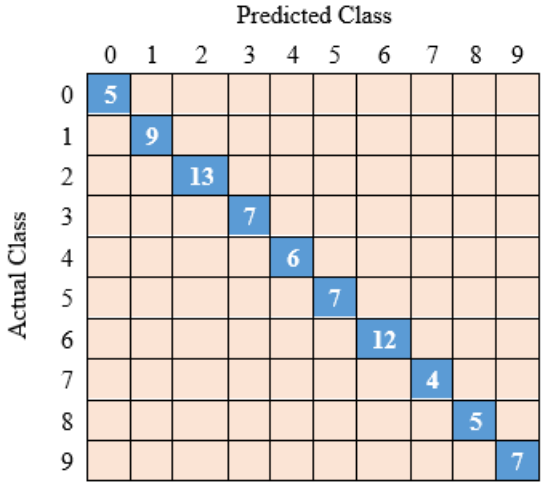

FIGURE 8. Audio speech recognition results

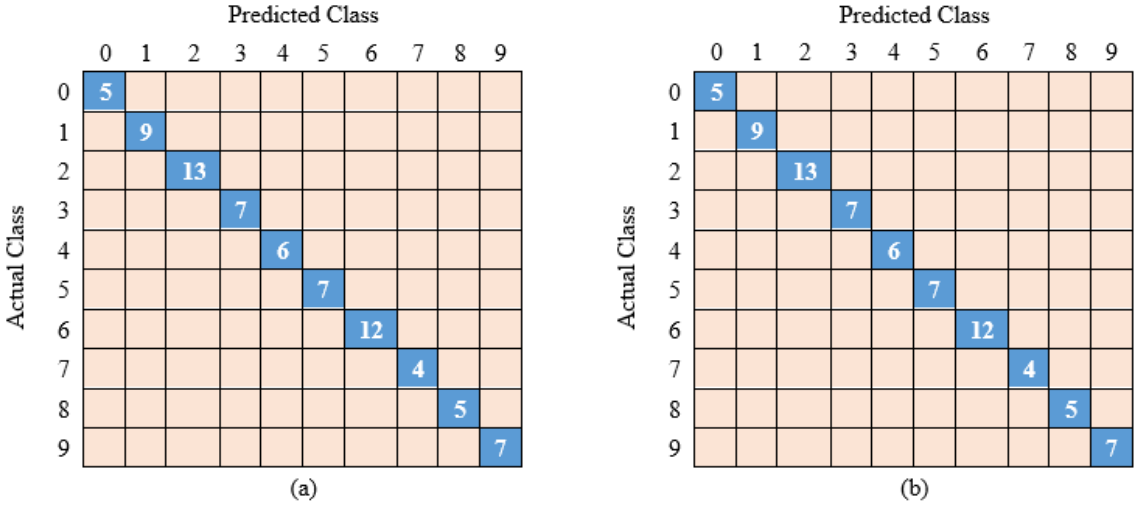

FIGURE 9. AVSR results (a) GoogLeNet; (b) ResNet-50

CONCLUSION

As a conclusion, this study investigated an extension of audio-visual speech recognition for digits in the Bosnian language. The approach used is based on deep learning, as opposed to the more traditional statistical approach. The first step in solving the problem was to create a speech dataset for Bosnian language digits, as there were no publicly available datasets.

To accomplish this task, both visual speech recognition and audio speech recognition were utilized. Both recognition methods were organized separately in the proposed architecture. Visual speech recognition used a CNN-RNN architecture for feature extraction and classification, while audio speech recognition used a CNN architecture on spectrograms. The results obtained were promising, with visual speech recognition achieving an accuracy of 72% with ResNet-50 architecture, and audio speech recognition achieving 100% accuracy. The combination of both predictions using audio-visual speech recognition was successful in recognising digits in the Bosnian language. The future work for this project involves acquiring data for a broader range of words to expand the vocabulary and achieve greater robustness in recognition. This will also involve including more subjects as speakers in the dataset and yielding more realistic real-world results. It is important to emphasize that the results presented in this paper depend on the nature of the dataset. In this particular case, the dataset was specifically designed for audio-visual speech recognition purposes. As a result, the speakers were instructed to maintain a stationary position, and the video quality was ensured to be of high standard.

ACKNOWLEDGEMENT

The authors would like to thank the International University of Sarajevo, Sarajevo, Bosnia and Herzegovina and College of Engineering, Universiti Teknologi MARA, Selangor, Malaysia for the support to conduct this research.

DECLARATION OF COMPETING INTEREST

None

REFERENCES

Abedalla A, Abdullah M, Al-Ayyoub M, Benkhelifa E. 2021. Chest X-ray pneumothorax segmentation using U-Net with EfficientNet and ResNet architectures. *PeerJ Computer Science* 7: e607 https://doi.org/10.7717/peerj-cs.607

Afouras, T., Chung, J. S., & Zisserman, A. 2018a. Deep Lip Reading: A comparison of models and an online application, *Proc. Interspeech 2018, 3514-3518.*

http://dx.doi.org/10.21437/Interspeech.2018-1943.

Afouras, T., Chung, J. S., & Zisserman, A. 2018b. The Conversation: Deep Audio-Visual Speech Enhancement*, Proc. Interspeech 2018, 3244-3248.* http://arxiv.org/abs/1804.04121

Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. 2021. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8: 53. https://doi.org/10.1186/s40537-021-00444-8

Asghar A, Sohaib S, Iftikhar S, Shafi M, Fatima K. 2022. An Urdu speech corpus for emotion recognition. *PeerJ Computer Science* 8: e954 https://doi.org/10.7717/peerj-cs.954

Atiqur, M., Ahad, R., Shammi, U. A., & Kobashi, S. 2018. A study on face detection using viola-jones algorithm for various backgrounds, angels and distances fall detection-Human Activity Analysis (HAA) view project human activity recognition view project, 2018, *Biomedical Soft Computing and Human Sciences* 23(1): 27-36.

https://www.researchgate.net/publication/327415969

Chung, J. S., & Zisserman, A. 2017. Lip reading in the wild*. Lecture Notes in Computer Science, vol 10112. Springer, 2017.* doi: https://doi.org/10.1007/978-3-319-54184-6_6

Fenghour, S., Chen, D., Guo, K., Li, B., & Xiao, P. 2021. Deep Learning-Based Automated Lip-Reading: A Survey. *IEEE Access* 9: 121184 – 121205.

doi:https://doi.org/10.1109/ACCESS.2021.3107946

He, K., Zhang, X., Ren, S., & Sun, J. 2016. Deep Residual Learning for Image Recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778.*

doi: https://doi.org/10.1109/CVPR.2016.90

Ivanko, D., & Ryumin, D. 2021. Development of Visual and Audio Speech Recognition Systems Using Deep Neural Networks*, Proceedings of the 31th International Conference on Computer Graphics and Vision.* doi: https://doi.org/10.20948/graphicon-2021-3027-905-916

Jadczyk, T. 2018. Audio-visual speech-processing system for Polish applicable to human-computer interaction. *Computer Science* 19(1): 41–64.
https://doi.org/10.7494/csci.2018.19.1.2398

Jahić Dževad, Halilović, S., & Palić Ismail 2000. Gramatika bosanskoga jezika. Dom štampe, ISBN:9958420465, 9789958420467

Jurafsky, D., Martin, J. 2020. Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

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. https://doi.org/10.1109/ACCESS.2021.3062752

Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., & Siohan, O. 2019. Recurrent neural network transducer for audio-visual speech recognition. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 2019, 905-912.*
https://doi.org/10.1109/ASRU46091.2019.9004036

Ma, P., Petridis, S., & Pantic, M. 2021, *End-to-end* Audio-visual Speech Recognition with Conformers*, ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7613-7617* http://arxiv.org/abs/2102.06657

Michelucci, U. 2019. Advanced applied deep learning: Convolutional neural networks and object detection, First Edition, ISBN: 978-1-4842-4975-8, Apress Berkeley, CA. https://doi.org/10.1007/978-1-4842-4976-5

Serdyuk, D., Braga, O., & Siohan, O. 2021. Audio-Visual Speech Recognition is Worth 32x32x8 Voxels, *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),2021.* http://arxiv.org/abs/2109.09536

Shashidhar, R., Patilkulkarni, S., & Puneeth, S. B. 2020. Audio Visual Speech Recognition using Feed Forward Neural Network Architecture, *2020 IEEE International Conference for Innovation in Technology, INOCON 2020*. https://doi.org/10.1109/INOCON50539.2020.9298429

Sterpu, G., & Harte, N. 2018. Towards Lipreading Sentences with Active Appearance Models*, 14th International Conference on Auditory-Visual Speech Processing (AVSP 2017)* http://arxiv.org/abs/1805.11688

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. 2015, Going Deeper with Convolutions, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 1-9,* doi:https://doi.org/10.1109/CVPR.2015.7298594

Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., & Yu, D. 2020. Audio-visual recognition of overlapped speech for the LRS2 dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 6984-6988,* http://arxiv.org/abs/2001.01656