Journal of Education for Pure Science- University of Thi-Qar
Vol.12, No2 (December, 2022)
Website: jceps.utq.edu.iq Email: jceps@eps.utq.edu.iq

DOI: http://doi.org/10.32792/utq.jceps.12.02.03

Comparison of DenseNet201 and ResNet50 for lip reading of decimal Digits

Kwakb Saadun Naif[1] **Prof. Dr. Kadhim Mahdi Hashim**[2]
[1]Computer Sciences Department, College of Education for pure Sciences, University of Thi-Qar Iraq.
[2]Imam Ja,afar AI-Sadiq University.

Received 18/4/2022 Accepted 2/6/2022 Published 01/12/2022

Abstract:

Lip reading is a technology supportive of humanity. It a process that interprets the movement of the lips to understand speech by means of visual interpretation. Where understanding speech is difficult for some groups of people, especially the hearing impaired or people who are in noisy environments such as the airport or factories lip reading is the alternative source for understanding what people are saying.

In the proposal the work starts with inserting the video into the Viola Jones algorithm and taking a sequential frame of the face image, then face detection, mouth detection and ROI cropping, then inserting the mouth frame into a convolutional neural network (DenseNet201) or ReNet50 neural network where features are extracted and then the test frames are categorized. In this research, a database consisting of 35 videos of seven people (5 males and 2 females) was used to pronounce decimal numbers (0, 1, 2, ..., 9). The test results indicate that the accuracy in DenseNet20 network is 90%, and in ResNet50 network we got an accuracy of 86%.

Keyword: Lip reading, Recognition, CNN, Densenet201 and ResNet50.

Introduction:

The research paper aims to read a set of decimal digits for personal speech in the absence of sound across two neural networks (DenseNet201 and ResNet50) and compare them. Lip reading is a very important way for people with hearing impairments to recognize or interpret speech by visually observing the movement of the lips [1]. It is a way to predict words and phrases just by watching a video without any audio signal[2]. It is difficult to teach lip reading because one has to learn the language and context of conversation in order to read properly[3]. Lip reading is not easy as it faces many challenges due to differences in inputs such as white and brown skin colors, facial features, languages, ability to speak and intensity, these difficulties can be reduced by using more visual input data collected. Automatic lip reading mainly includes face detection, lip positioning, feature extraction, classifier training, and decimal recognition through lip movement [4].

The lip-reading process is applied to the levels of sentences, numbers, alphabets and words where the success rate of lip-reading is directly related to the correct segmentation of the picture frame for the region of interest[5].

Automatic Lip reading is used to interpret or understand speech without the hearing it, a technique mastered by human with hearing difficulties. In this paragraph, we review previous work that is related to our work. Zhao et al. (2009) in [6]. They used method local spatiotemporal descriptors an automatic lip reading, they represented the isolated words by extraction spatiotemporal Local Binary Patterns (LBP) from mouth area, this method obtained an accuracy of 58.85%. R. Bowden et al. (2009) in [2]. They proposed compared various features for automatic lip reading, including four types of Active Appearance Model (AAM) features, 2D DCT features, sieve features and Eigen lip features in general, AAM features with appearance outperform other types of feature, which means that appearance is more useful than the shape .The proposed achieved accuracy rate 65%. Ngiam et al. (2011) in [7]. They used a deep learning method to recognize speech using video and audio, where they used two data set Cuave to digits and AVletters where their accuracy was 68.7%and 64.4%. A method was proposed to a process the problem of non rigid mouth movement analysis using a database OuluVS2 and they got a recognition of 47%. C.Tian and W. Ji (2016) in [8]. They proposed method an auxiliary multimodal LSTM that aims to merge visual and audio data at the same time. It learns from both visual and audio methods and uses PCA to reduce dimensional and VGG16 model to extract features the performance obtained was 88.83%. Befkadu Belete (2019) in [9]. He suggested the audio-visual method for lip reading, where Viola Jones algorithm used to detect from the mouth, and to extract the features, it used Discrete Wavelet Transformation (DWT), and the database used was AAVC, and it obtained an accuracy of 72%, and the discrimination was 67.08%.

1. Materials and methods:

2. The proposed system

The proposed system consists of three stages; preprocessing, convolution neural network and classification. The input to the proposed system is a video that was recorded by installing a camera in front of persons with different lighting conditions. We used the Viola-Jones algorithm to detect the face and mouth area, then cropped the mouth area from each frame. Then fed frames into CNN models, which include DenseNet20 and ResNet50, and observed the difference in accuracy results. Our proposed system is shown in Figure 1

3.1 The preprocessing stage

Pre-processing begins by cutting the video into a sequential frame then defining a region of interest (RIO) is mouth, which is the region that we closely investigate from a specific region within an image, where it needs engineering operations that modify the spatial coordinates of the image such as cropping, resizing, translate and rotation . Then mouth frames stored in temporary folder**.** Then we change the frame size to 224 * 224 to match the work of the network. The preprocessing stage includes three steps as in the figure 2.

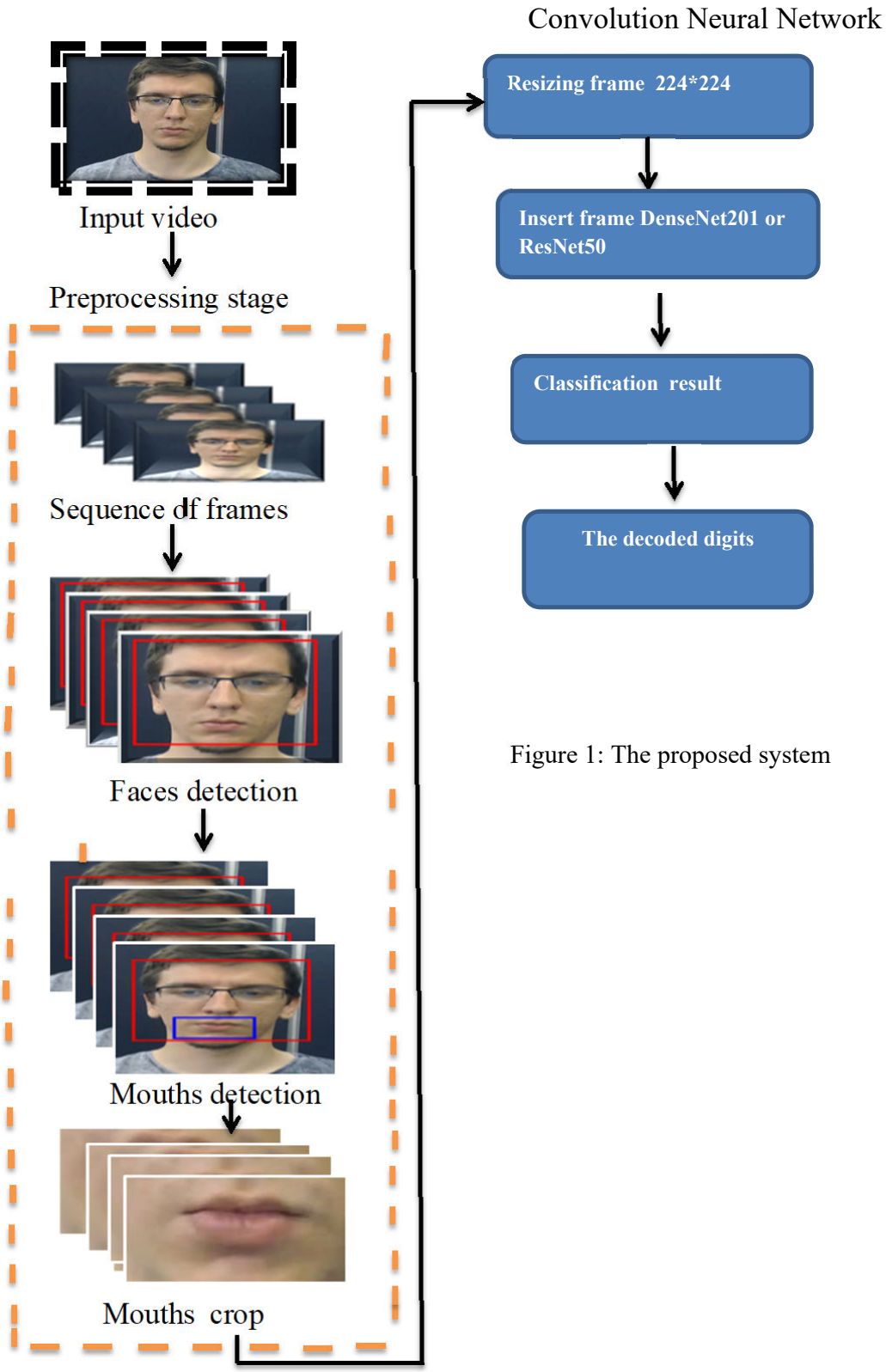

Figure 1: The proposed system

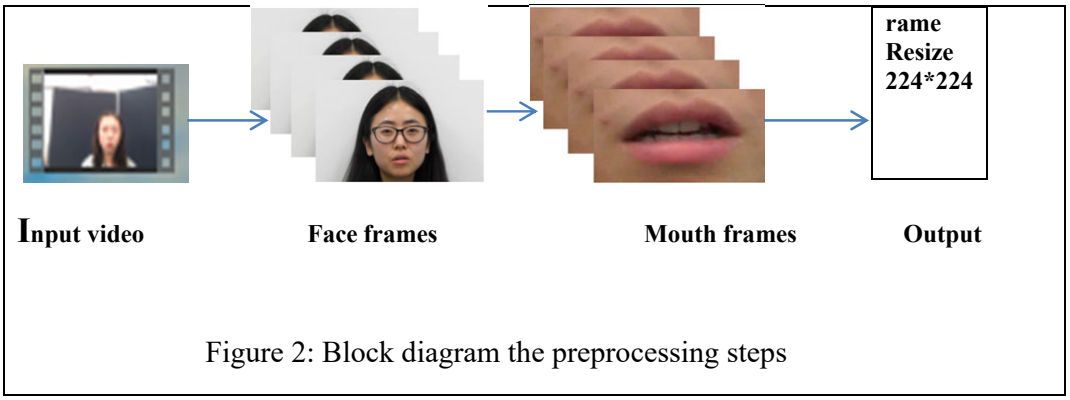

Figure 2: Block diagram the preprocessing steps

3.1.1 Face and mouth detection

After capturing a video of the person speaking the digits, it is stored in a temporary folder and the face and mouth region are detected, and this is done by Viola Jones algorithm [10]. As shown in Figure (3) Detecting the face and mouth region

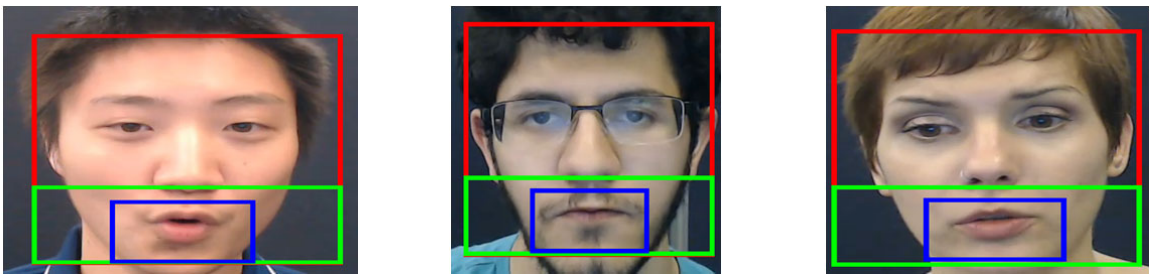

Figure 3: Sample for face and mouth region detection by viola jones algorithm .

3.1.2 Crop mouth image from the frame

It is the process of cutting out areas of the image that are of interest called a sub-image and that benefit us in the analysis. At this stage the mouth is identified and cropping.

Figure 4: Examples for cropping the mouth.

3.1.3 Resize mouth image to 224×224.

In ResNet50 network, the size of the input image must be 224 * 224, and this means that we need to modify the size of the input image of the CNN to reduce or enlarge it to reach the required size and the mouth image is the cutout area

3.2 Convolution neural network stage

In this stage, we use the Convolutional Neural network, after completing the pre-processing of the specified image, where the movements of the lips must be matched with the spoken number and to know whether the pronunciation is correct or not.

3.2.1 convolution neural network layers

It is a specific type of artificial neural network that uses sensory perception [11]. Also known as (conNet) [12]. It contains sequential layers, and each layer has its own task, as the CNN consists of three layers [13]:

A. Convolution layer

It is the most important component of the CCN architecture and it is a set of automatic kernels also called filters that are executed with the input image to create an output feature map [12]. Also called feature maps [14]. And it is a generalized linear model of the basic local image [15]. Where the convolution layer Creates feature maps that highlight the unique features of the original image. Inserts the image through convolution filters into the feature map.

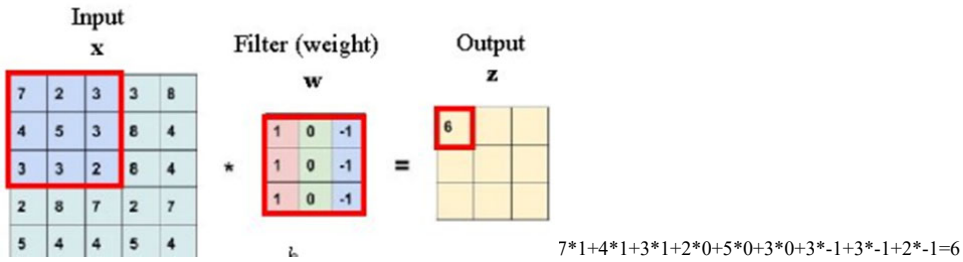

Figure5: Example convolution operation [16].

B. Pooling layer:

Pooling layer is the second layer in CNN and it comes after the convolutional layer and it is called the implementation of the reduction process and this is called pooling and it means the process of reducing the size of each dimension of the output maps and this reduces the number of parameters that shorten the training time, this layer preserves the maps Input and Output it also merges pixels adjacent to a specific area of the image [17].

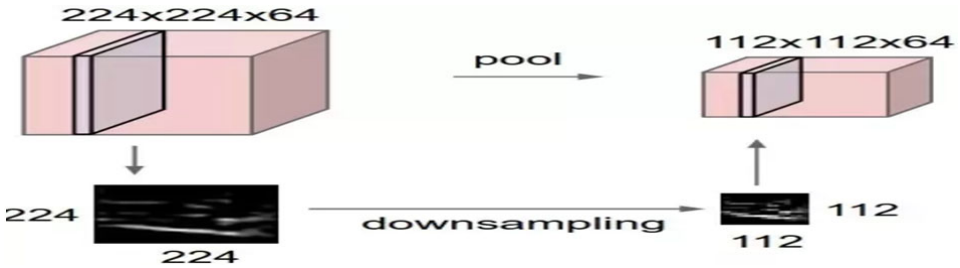

Figure 6:Pooling operations [18].

C. Fully connected layer:

It is the output of the final pooling layer or is the output of the final layer of a CNN [18]. Used as a classification layer where the result of each feature class extracted from the convolutional layers in the previous steps is calculated, the FC layer usually takes advantage of the activation function Sotfmax to classify the input appropriately [17].

3.2.2 DensentNet201:

It is a network that provides us with sequences of all feature maps from previous applications, which means that all feature maps are spread to later applications and are connected to the newly created feature maps one of its advantages is to reuse the features and reduce the problem of scaling or fading [19].

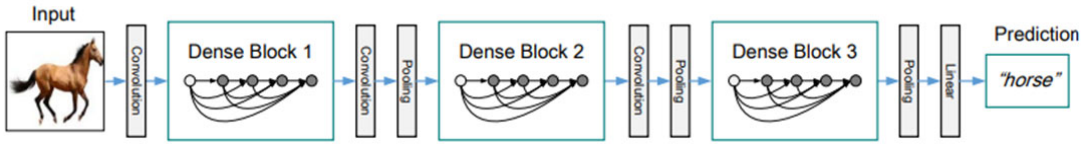

Figure 7: Densenet201 architecture[19].

3.2.3 ResNet50:

Acronym Residual network is a specific type An important convolutional neural network that was introduced in 2015 by kaiming and join sun and others consists of fifty layers that can receive more than one million trained images from a database and can also classify the trained images into 1000 object classes [11] .

The goal is to solve a complex problem by stacking some extra features in the network and this leads to an increase in performance and accuracy because the increase in the number of layers leads to gradual learning of more complex feature [20].

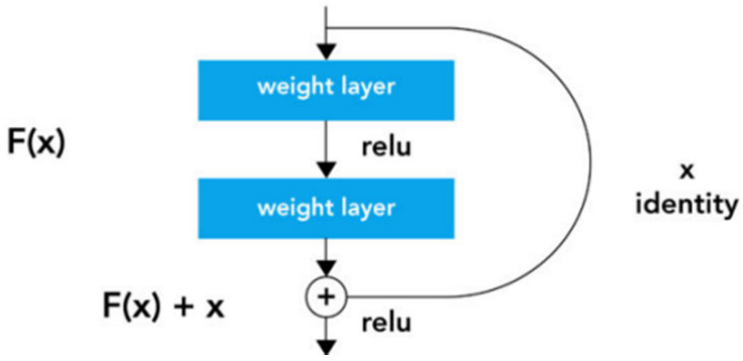

Figure 8: A residual learning building block [20].

3.3 The classification Stage:

The video frame entered into a convolution neural network DenseNet201 and ResNet50 in order to obtain the classification result. As the features were extracted and classified correctly in the classification process, if the test frame was correctly classified and matched the training frame as in Figure (9), this

means that the network work was done correctly But if the test frame is classified into a frame that does not match the training frame as in the figure (10), this means that the network is classified incorrectly.

Figure9:Example test frame matching training frame Figure10: test frame not matching training

4. Dataset Description.:

The database consists of seven personal, each person has five videos, meaning we have 35 videos that speak digits from 0 to 9 randomly and in different languages individuals to 5 males and 2 females, there are about 21500 frame (lip image) in our database. by placing in front of the camera and the location from which we captured the videos is https://ibug-avs.eu/ [21] . figure(10) shows examples of training set and test set to various people.

4.1 Experimental Details:

We inserted the lip frame into the convolutional neural network DenseNet201 and ResNet50, which numbered 21500 frames. It will pass through the training phase, which was 80% (17200) images from the training images samples randomly. Then the testing phase starts after the training process which was 20% (4300) images were selected from The sum of randomly selected samples. and achieved an accuracy of 90% and 86%, respectively.

Training and test of DenseNet201 and ResNet201 based on MATLAB R2020a language to easily implement the code of CNN. A computer Lenovo that has specifications such as Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz for CPU, 8.00 GB Windows 10 Pro of RAM, and 64-bit operating system, x64-based processor.

The following table show the experimental results to compare the two networks DenseNet201 and ResNet50.

Table1: Illustrates the main classification criteria To compare the two networks.

Metrics / Model	Accuracy	Precision	Recall	F-measure	Total time
DenseNet201	90%	90%	90.7%	90.3%	432
ResNet50	86%	85.3%	85.9%	86.7%	366

5. conclusion:

Our proposed system went through several stages to reach the recognition of the pronunciation of digits by the movement of the lips, the first of which is our use of the Viola Jones algorithm to detect the face, then the mouth area, and then cut out the region of interest (the mouth) and store the lip frame in a temporary folder and then Then we added the frames size 224 * 224 to one of the two networks DenseNet201 or ResNet50 the to extract the features Then classify it. One of the advantages of our proposed system is that it gave us a correct classification, despite the differences in the structure and design of the lips from one person to another**.**

REFRENCES:

[1] A. Nagzkshay Chandra Aarkar, "ROI EXTRACTION AND FEATURE EXTRACTION FOR LIP READING OF," vol. 7, no. 1, pp. 484–487, 2020.

[2] R. Bowden, "Comparing Visual Features for Lipreading," no. September 2016.

[3] A. Garg and J. Noyola, "Lip reading using CNN and LSTM," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Jan, p. 3450, 2017.

[4] J. S. Chung and A. Zisserman, "Learning to lip read words by watching videos," *Comput. Vis. Image Underst.*, vol. 173, pp. 76–85, 2018, doi: 10.1016/j.cviu.2018.02.001.

[5] T. As for training and classification, this is done with the help of artificial neural networks and A. Basturk, "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models," vol. 7, no. 2, pp. 195–201, 2019, doi: 10.17694/bajece.479891.

[6] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans. Multimed.*, vol. 11, no. 7, pp. 1254–1265, 2009, doi: 10.1109/TMM.2009.2030637.

[7] J. Ngiam and A. Y. Ng, "Multimodal Deep Learning," 2011.

[8] C. Tian and W. Ji, "Auxiliary Multimodal LSTM for Audio-visual Speech Recognition and Lipreading," no. 1, pp. 1–9, 2017, [Online]. Available: http://arxiv.org/abs/1701.04224.

[9] Mr. Befkadu Belete Frew, "Audio-Visual Speech Recognition using LIP Movement for Amharic Language," *Int. J. Eng. Res.*, vol. V8, no. 08, pp. 594–604, 2019, doi: 10.17577/ijertv8is080217.

[10] H. VAIBHAV, "Face Identification using Haar cascade classifier," *Medium*, pp. 1–5, 2020, [Online]. Available: https://medium.com/geeky-bawa/face-identification-using-haar-cascade-classifier-af3468a44814.

[11] T. Bezdan and N. Bačanin Džakula, "Convolutional Neural Network Layers and Architectures," no. July, pp. 445–451, 2019, doi: 10.15308/sinteza-2019-445-451.

[12] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, *Fundamental concepts of convolutional neural network*, vol. 172, no. January. 2019.

[13] J. Wu, "Introduction to Convolutional Neural Networks," *Introd. to Convolutional Neural Networks*, pp. 1–31, 2017, [Online]. Available: https://web.archive.org/web/20180928011532/https://cs.nju.edu.cn/wujx/teaching/15_CNN.pdf.

[14] W. M. Learning and P. Kim, *MATLAB Deep Learning*. .

[15] Q. Zhang, M. Zhang, T. Chen, Z. Sun, Y. Ma, and B. Yu, "Recent advances in convolutional neural network acceleration," *Neurocomputing*, vol. 323, pp. 37–51, 2019, doi: 10.1016/j.neucom.2018.09.038.

[16] "Convolutional Layer." [Online]. Available: https://qiita.com/icoxfog417/items/5fd55fad152231d706c2.

[17] M. Z. Alom *et al.*, "The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches," 2018, [Online]. Available: http://arxiv.org/abs/1803.01164.

[18] H. J. Jie and P. Wanda, "Runpool: A dynamic pooling layer for convolution neural network," *Int. J. Comput. Intell. Syst.*, vol. 13, no. 1, pp. 66–76, 2020, doi: 10.2991/ijcis.d.200120.002.

[19] S. Wang, "DenseNet-201-Based Deep Neural Network with Composite Learning Factor and Precomputation for Multiple Sclerosis Classification DenseNet-201 based deep neural network with composite learning factor and precomputation for multiple sclerosis classification," no. September, 2020, doi: 10.1145/3341095.

[20] C. Vision, H. Resnet, R. Block, H. Resnet, and U. Resnet, "What is Resnet or Residual Network | How Resnet Helps? Introduction to Resnet or Residual Network," pp. 1–8, 2020, [Online]. Available: https://www.mygreatlearning.com/blog/resnet/.

[21] S. Petridis, J. Shen, and D. Cetin, "Visual-only recognition of normal, whispered and silent speech," 2018.