# Enhancing Speech Rehabilitation: Calibrating 3D-CNN Lip Reading Models for Higher Single User Accuracy to Improve Communication in Aphonia and Aphasia Cases

SUBMISSION DATE / POSTED DATE

06-09-2023 / 07-09-2023

CITATION

Kashyap, Jai (2023). Enhancing Speech Rehabilitation: Calibrating 3D-CNN Lip Reading Models for Higher Single User Accuracy to Improve Communication in Aphonia and Aphasia Cases. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.24086103.v1

DOI

10.36227/techrxiv.24086103.v1

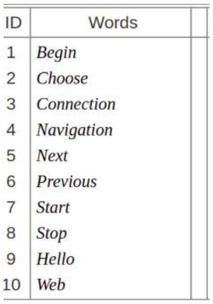
# Enhancing Speech Rehabilitation: Calibrating 3D-CNN Lip Reading Models for Higher Single User Accuracy to Improve Communication in Aphonia and Aphasia Cases

Jai Kashyap
*Academy of Engineering and Technology,*
Leesburg, Loudoun County
jaitkash@gmail.com
Orcid: 0009-0002-1181-9196

*Abstract*— **Automated lip-reading systems have the potential to greatly improve speech recognition and communication for individuals with speech or hearing disabilities. People with aphasia, aphonia, dysphonia, voice disorders and trouble swallowing have limited speaking ability and can be assisted by lipreading technology. Traditional approaches to lipreading have relied on hand-crafted features and statistical modeling techniques, which have limitations in capturing the complex spatiotemporal dynamics of lip movements. Deep learning approaches have shown promise in addressing these limitations by extracting features from data and have achieved state-of-the-art results in various speech-related tasks. In this paper, a 3D Convolutional Neural Network (3D-CNN) is proposed as an approach to automated lip reading. This system takes a video of a person speaking and processes it through a 3D-convolutional layer to extract spatial-temporal features from the video frames. The system uses deep learning algorithms to learn the mapping between lip-movements and corresponding phonemes, enabling it to recognize spoken words. The approach is evaluated on visual recordings of spoken words, the MIRACL-VC1 dataset. It contains 10 words and multiple instances of each. The proposed model achieves 99.0%training accuracy on the dataset. The testing accuracy achieved is 61.3%, indicating model overfitting and a high "speaker-dependency". A dataset was self-created using videos of one speaker. The model achieved an 89.0% training accuracy, and an 83.0% testing accuracy on this dataset. Both models are then evaluated on user input video. The proposed approach has applications in speech therapy, speech recognition, and translation for those with speech and voice disabilities.**

## *A. Introduction*

What if someone did not have the ability to speak or hear? People with aphonia, aphasia, or deafness are unable to communicate using speech. Some people who are born with rare genetic conditions that cause them to be mute from birth, can have difficulties communicating with others. Although the American Sign Language (ASL) helps, these individuals are limited to communicating with others who also know ASL. A program that can recognize lip movements and provide written text output (like subtitles for a movie) would be far more impactful and convenient for both parties. Computer vision is a subfield of Artificial Intelligence that deals with how computers gain an understanding of videos and pictures. The goals of this field are to essentially mimic the human visual system and automate the tasks it does [1]. Computer vision algorithms (or models) help computers learn by performing analysis on data. When enough data is provided, the model trains on this information and can identify different visual inputs. The data used is from the MIRACL-VC1 dataset. The dataset contains 15 speakers (10 women, 5 men) recorded saying words. Each speaker repeats the set of 10 words 10 times, for a total of 1500 instances of words [3]. The structure of the dataset and the words and IDs are shown below.



| ID | Words |
|---|---|
| 1 | *Begin* |
| 2 | *Choose* |
| 3 | *Connection* |
| 4 | *Navigation* |
| 5 | *Next* |
| 6 | *Previous* |
| 7 | *Start* |
| 8 | *Stop* |
| 9 | *Hello* |
| 10 | *Web* |

Fig 1. Words and IDs [3]

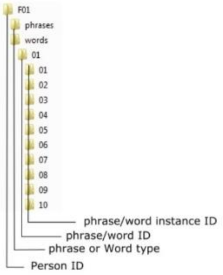


Fig 2. Dataset Structure [3]

## *B. Prediction*

To predict on this dataset a neural network (NN) is used [4]. A neural network is a fundamental component of machine learning. It is an algorithm that has a structure that imitates the human brain, using interconnected neurons (nodes) to form multiple layers of data processing. A specific type of neural network, called a convolutional neural network (CNN), is used for computer vision because of its ability to handle images and videos well [5]. CNNs use a mathematical operation called a convolution which allows them to look at specific regions of images to generate features. This allows CNNs to retain an image's spatial information while creating feature detectors that can generalize to different locations in an image. A specific type of CNN is the 3-Dimensional CNN. The advantages of 3D CNNs are many; unlike 2D-CNN's which can only process 2D images, 3D CNN's handle 3D data, such as videos. This allows them to capture both spatial and temporal information making them well suited for tasks like lipreading. Additionally, 3D CNN's can extract features specific to 3D data, such as shape and depth information. This enables them to produce more meaningful feature representations, leading to better performance. Tweaking, or fine-tuning this architecture can result in a varying level of accuracy [6]. For example, adding a layer to the model, or removing a layer can increase accuracy. Additionally, machine learning models use hyperparameters to control the learning process. Some of these hyperparameters include number of epochs (an epoch is one iteration through the training 3 dataset), batch size (the number of samples passed through the network at one time), learning rate (determines the size of the "step" at each iteration. It determines how fast the model moves towards optimal weights), and activation function (this determines whether a node should be activated or not). By adjusting these hyperparameters, the model can be optimized, and achieve greater accuracy on datasets.

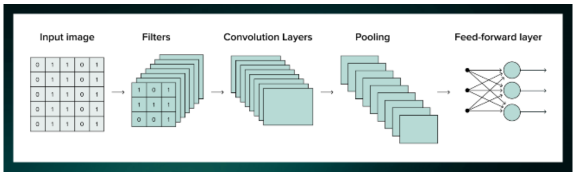


Fig 3. Example Layers of 1D-CNN [7]

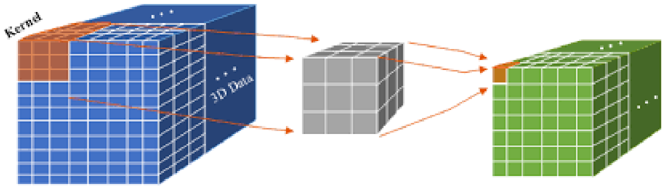


Fig 4. Convolution Operation in 3D-CNN Architecture [8]

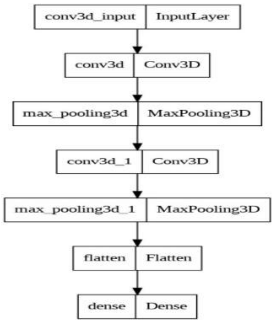


Fig 5. Model Architecture

The input shape of the model has to be specified as (10,100,100,1), as each sequence has a maximum length of 10 images, and each image is (100x100x1) in size. The model is then declared with the Keras sequential class. 3D Convolutional and 3D MaxPooling layers are added to the model, then a flattening layer to compress the data for output. The last layer is a Dense layer, with an output shape of 10, as there are 10 words. The model is trained using 45 epochs (iterations through the training data). The Matplotlib library is used to create the accuracy and loss graphs of the model, showing how the accuracy increases or decreases over time (as the number of epochs increases).

### *C. Dataset Lip Images*

The MIRACL-VC1 dataset contains sequences of images ranging from 10-15 in length. These image sequences correspond to one instance of a word being spoken. The sequences of images are a synchronized series of color and depth images, both being (640x480) pixels in size. Due to high computational costs, the images were scaled down to 100x100 pixels in size. Below is a 640x480 pixel image from a sequence.

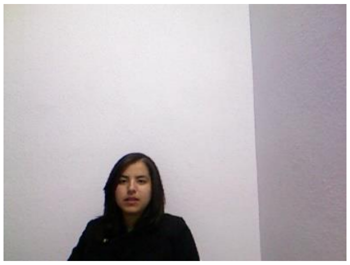


Fig 6. One Image from MIRACL-VC1 Dataset

Due to the nature of the dataset images, the location of the mouth and lips can vary from sequence to sequence, and even from image to image. This combat this issue a Haar Cascade Classifier was used to identify the necessary region of the image, so that it could be cropped for use by the convolutional neural network. Below, an image is shown with the Haar Cascade Classifier in use to find the lip region.

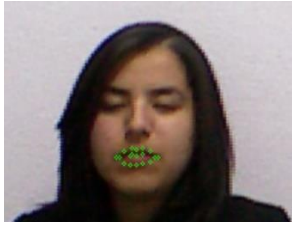


Fig 7. Haar Cascade Classifier Identifies Lip Region

Using lip landmark coordinates from the classifier, the image can be cropped to contain only the lips. Since computer vision lipreading is only concerned with the movement of lips from frame to frame, the images have been made grayscale to reduce computational power due to additional RGB data. Below is an image of the grayscale cropped image.

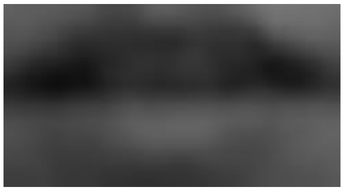


Fig 8. Cropped and Grayscale Lip Image

The process of identifying the necessary lip regions, converting the images to grayscale, and cropping the images was automated for the rest of the dataset. Then the cropped images are vectorized into NumPy arrays to further process the data. The data must be normalized to train the model on. Min Max Normalization is employed - Min Max Normalization is when the minimum values in an array are converted to zeros and maximum values converted to ones. All values in between are decimals between 0 and 1. This normalization is only for the x train, test, and validation sets. For the y train, test, and validation sets, one hot encoding must be used. One hot encoding is the process of converting categorical data to numerical data. For each category, or in this case word, the word ID is represented by zeros and ones. The index of the one in the array corresponds to the word ID. The new, processed data was  then resized so that every image was 100x100x1 in size.

## I. Method

### A. Training

The 3D-CNN shown in Fig 5. was trained on the MIRACL-VC1 dataset.

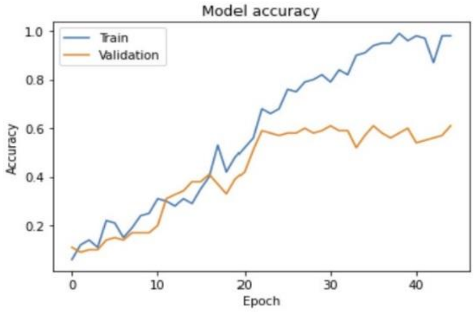


Fig 9. Model Accuracy vs Epoch for MIRACL-VC1 Dataset

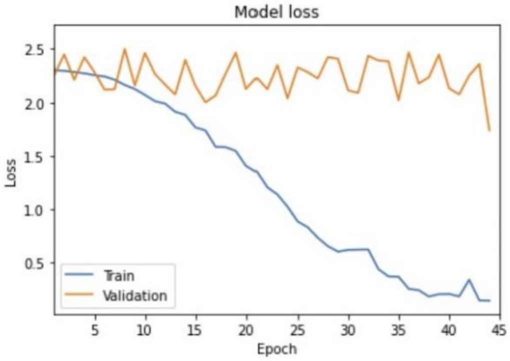


Fig 10. Model Loss vs Epoch for MIRACL-VC1 Dataset

Table 1. Train and Test Accuracy for MIRACL-VC1 Dataset

| *MIRACL-VC1 DATASET* | |
|---|---|
| Training Acc | 99.0% |
| Testing Accuracy | 61.3% |

The proposed model shows a 99% training accuracy and a 61.3% testing accuracy. These results indicated overfitting, a concept in data science which occurs when a model fits exactly against its training data [9].  When this occurs, the machine learning algorithm cannot perform accurately against unseen data, which defeats the very purpose of machine learning- which is for an algorithm to be able to predict and make choices on its own. When the model is constructed, it leverages a sample dataset (in this scenario, the MIRACL-VC1 dataset) to train on. When the model trains for too long (in terms of epochs) or is unable to learn, it can start to take in “noise”, or unnecessary information. When the model memorizes the noise, it can only fit the training set very well – it is unable to generalize new, unseen data. The high training accuracy and variant model loss for the validation data indicates overfitting.

But why is this happening? High speaker dependency in lip reading refers to a situation in which lip reading, as a method of understanding spoken language, heavily relies on specific characteristics and nuances of the individual speaker. In other words, when lip reading is highly speaker-dependent, the accuracy and effectiveness of lip reading can vary significantly from one person to another. This happens because different speakers may have distinct lip movements, accents, or speech patterns that make it more challenging for a lip reader (in this case the machine learning model) to accurately interpret their speech.

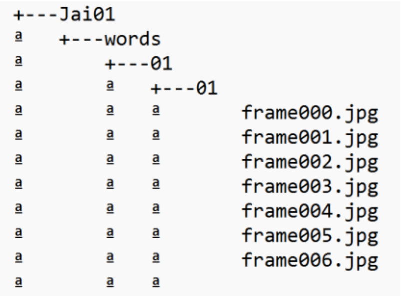
*B. Solution*

High speaker dependency is a major issue when trying to use computer vision for lipreading (see section 1A). The benefits of calibrating a model to a single user's unique lip movements are many. With only one speaker to focus on, the model can become more accustomed to the specific lip movements, facial expressions, and speech patterns of that individual. This familiarity can lead to increased accuracy in understanding the speaker's words over time. However, to do this, the model requires a large quantity of data that contains only the person that the model is intended to calibrate on. The MIRACL-VC1 dataset does not contain a vast amount of data for single users, as its image sequences are spread across a set of 15 different speakers. After unsuccessfully scouring various lip-reading datasets for a large amount of single speaker data, a new dataset was created to ensure consistent, high volume image sequences. The same model would then be trained on this new dataset, and the results compared with the original model trained on the MIRACL-VC1 dataset.

*C. Creating the Single Speaker Dataset*

The greatest challenge during this study was the creation of the single speaker dataset. To ensure a direct comparison with the model trained on the MIRACL-VC1 dataset, the same set of 10 words (see Fig. 1) had to be used for the single speaker dataset. To guarantee a sufficient amount of data, 100 videos were recorded for each word, in total 1000 videos (which make up 1000 image sequences).

The first step of creating the single speaker dataset was the build out the structure of the dataset. The figure below shows the structure of the new data.



```
+---Jai01
ª   +---words
ª       +---01
ª       ª   +---01
ª       ª   ª       frame000.jpg
ª       ª   ª       frame001.jpg
ª       ª   ª       frame002.jpg
ª       ª   ª       frame003.jpg
ª       ª   ª       frame004.jpg
ª       ª   ª       frame005.jpg
ª       ª   ª       frame006.jpg
ª       ª   ª
```

Fig 10. Single Speaker Dataset Structure

The image shown above contains only the first iteration of the first word of the first set, in the entire dataset. The Jai01 folder contains 10 instances of 10 words, which in total is 100 image sequences. This is then repeated for folders Jai02-Jai10, resulting in a thousand image sequences. 700 of the sequences are used for training data, 200 sequences used for validation data, and 100 sequences are used for testing data. The process of creating the data is shown below.

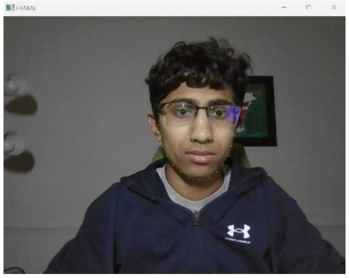


Fig 11. The user is recorded[1]

In Fig 11, the user is video recorded. The green circles that appear on the face is the Haar Cascade Classifier which is mapping out facial landmarks.

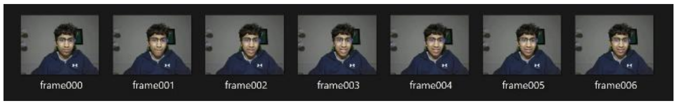


Fig 12. The Video is Split Into Frames

Here, the CV2 Python library is used to split the video file into separate frames.

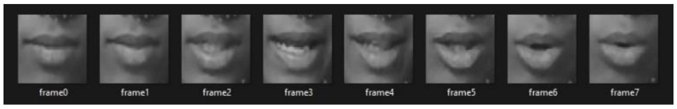


Fig 13. The Frames are Cropped and Made Grayscale

In the figure above, the Haar Cascade classifier is used to find *x* and *y* coordinates of lip landmarks so that each image can be cropped into only the lip. Below, a close-up image of one cropped frame is shown.

[1] The human subject in this study is the researcher and author.

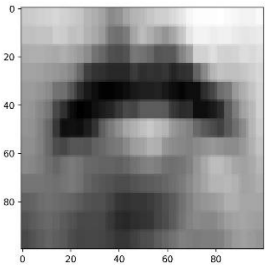


Fig 14. Close-Up of 100x100 Cropped Lip Image

The data creation process was repeated for 1000 iterations, recording 1000 videos of the same speaker saying 100 iterations of the 10 different words contained in the MIRACL-VC1 dataset.

*D. Training and Comparison*

The same 3D-CNN model architecture was used to train on the single speaker dataset.

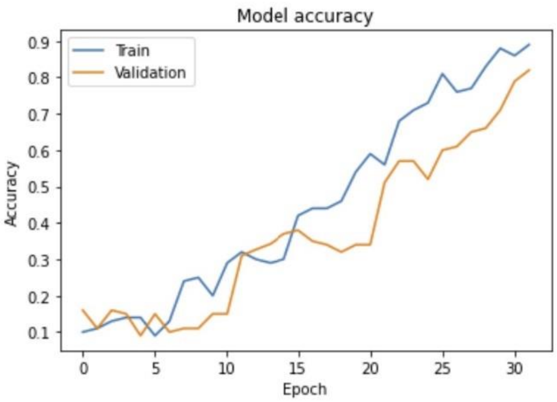


Fig 14. Model Accuracy vs Epoch for Single Speaker Dataset

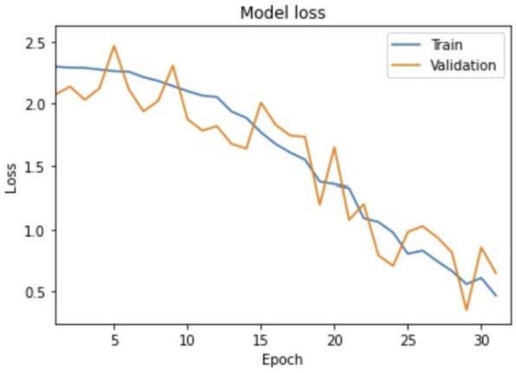


Fig 15. Model Loss vs Epoch for Single Speaker Dataset

Table 2. Train and Test Accuracy for Single Speaker Dataset

| *Single Speaker Dataset* | |
|---|---|
| Training Acc | 89.0% |
| Testing Accuracy | 83.0% |

The proposed 3D-CNN model calibrated to a single speaker shows an 89.0% training accuracy and an 83.0% testing accuracy on the single speaker dataset. The results demonstrate the difficulty of computer vision neural networks in predicting the lip movements of many different speakers, they are better suited to hone in on a specific user's tendencies and lip movements. Although the training accuracy for the single speaker dataset was lower than that of the MIRACL-VC1 dataset (see Table 1) the testing accuracy was much higher, indicating less overfitting and a better generalization to the data.

Both 3D-CNN models have been trained, and can be deployed to a software application which enables real-time lipreading. The design of the software is shown below.

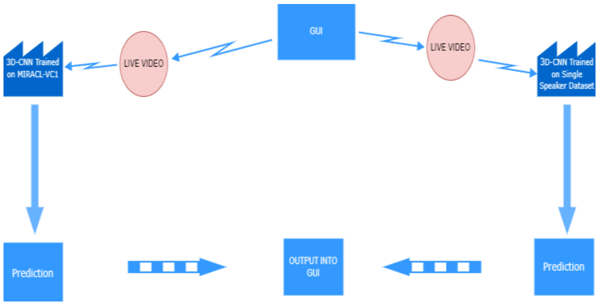

GUI
LIVE VIDEO
LIVE VIDEO
3D-CNN Trained on MIRACL-VC1
3D-CNN Trained on Single Speaker Dataset
Prediction
OUTPUT INTO GUI
Prediction

Fig 16. Diagram of Real-Time Lipreading Software

The diagram above shows the design of the real-time lipreading software. The user is presented with a graphical user interface where they can record a video of themselves, and have both models predict what word they said. Below is the implementation of the application.

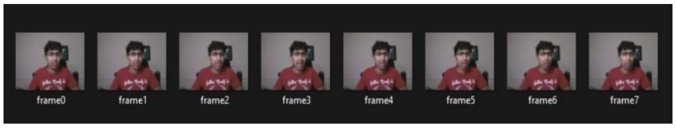


Fig 17. The Recorded Video is Split Into Frames

After the user records the video of themselves speaking, the video is split into frames as shown above. This process is similar to the data preprocessing for the single speaker dataset (see Fig 11). The frames are then cropped into a single lip sequence.

Fig 18. Frames are Cropped and made Grayscale

This sequence of images is then fed through to both models, which are already pretrained on their respective datasets. The models' prediction is then output onto the interface for the user to see.

This simulation of a user (the same user that the single speaker dataset was made with) recording a video and having the models predict was repeated for 25 iterations. Below are the results.

Table 3. User Prediction Accuracy by Both Models

| *User Prediction Accuracy* | |
|---|---|
| Model Trained on MIRACL-VC1 Dataset | 8.0% |
| Model Trained on Self-Created Dataset | 76% |

The results above show the vast discrepancy between the prediction of both models. The calibrated model trained on the Self-Created Dataset greatly outperformed the model trained on the 15-speaker wide MIRACL-VC1 dataset, highlighting the importance of calibrating a computer vision lipreading model.

## II. Conclusion

### A. Analysis

The proposed model achieved a training accuracy of 99% and a testing accuracy of 61.3% on the MIRACL-VC1 dataset. The graphs above show the model's training accuracy and validation accuracy, as well as the training loss and validation loss. The validation accuracy flatlines at around 20 epochs and makes no significant gain upward, whereas the training accuracy continues to rise to 99%. The loss graphs show a similar trend, with the validation loss being much higher than the training loss throughout the 45 epochs. This indicates overfitting, an undesirable machine learning behavior that indicates that the model fits to the training data too well. When this happens, the model is unable to generalize to unseen test data, resulting in a low-test accuracy, 61.3% as shown above. One reason for this is present in the science of lipreading: high speaker-dependency. Lip movements of the same words can vary greatly for different speakers, causing a lot more difficulty for machine learning algorithms to accurately predict on an unseen speaker. To test this issue, a brand-new dataset was created with videos of only one person. After recording 1,000 videos (100 videos per word), a machine learning model was trained on this single speaker data. The results were much better than on the MIRACL-VC1 dataset. A training accuracy of 89.0% and a testing accuracy of 83.0% was achieved on the single speaker dataset. This is around a 22% increase in testing accuracy. Analysis of training and loss graphs show much less overfitting, as the model was able to generalize to an unseen video (although the speaker was the same). The increase in model performance from the MIRACL-VC1 dataset to the single person dataset highlights the personal dependence of the technique. Lip shape and movement vary among individuals; different people have different lip shapes and movements when they speak. For example, some people have more prominent lips than others, and some open their mouths wider or move their lips more when they speak. Pronunciation and enunciation can also affect lip movements. For example, some consonant sounds, such as "p", "b", and "m" require the lips to come together, while other sounds such as "s" and "f" require the lips to be spread apart. Someone with an accent that affects their pronunciation can be very challenging to lipread. After training and testing both models on their respective datasets, they were evaluated on unseen user input videos. Each model predicted on 25 videos of a speaker saying one of the 10 words shown in Figure 2. The model trained on the MIRACL-VC1 dataset predicted correctly for 2/25 attempts, or 8%. The model that trained on the Self-Created dataset of the user themselves correctly predicted 19/25 attempts, or 76%. This stark difference in performance demonstrates the high speaker-dependency of computer vision based lipreading.

### B. Data Gaps

Data gaps in machine learning refer to situations where a model for computer vision lacks sufficient data or representative examples to accurately learn and generalize to new situations. To achieve higher accuracy and model performance, more data is needed. For computer vision tasks, an enormous amount of data is required. The MIRACL-VC1 dataset has only 1500 image sequences of words, which is 150 image sequences per class. This is not nearly enough to fit a model to and predict on. The single person, self-created dataset has even less, at only 100 image sequences per word. Standard computer vision tasks and datasets usually have upwards of 10,000 images to train on, far higher than what is available here.

### C. Moving Forward

The proposed 3D-Convolutional approach to an automated lip-reading system for the speech impaired shows great potential for improving the communication abilities of individuals who face challenges with spoken language. The results of this study demonstrate that the system can recognize spoken words with a moderate degree of accuracy, making it a promising tool for enhancing the quality of life of speech-impaired individuals. However, there are several avenues for further research and improvement of the system. First, the system's performance could be evaluated on a larger dataset, including more diverse speakers, accents, and languages. This would help to improve the system's robustness and generalizability across different populations and contexts. The next step would be to improve the model through hyperparameter finetuning. The incorporation of new deep learning technologies like Vision Transformers could also help to improve the accuracy of the system. Lastly, the current system has to record a video, split the video into frames, and then process the frames for prediction. Implementing a real-time lip-reading system could be possible using frameworks such as Google's MediaPipe.

# References

[1] Ashtari, H. (2022, May 13). What Is Computer Vision? Meaning, Examples,and Applications in 2022. Spiceworks. https://www.spiceworks.com/tech/artificial-intelligence/articles/what-iscomputer-vision/

[2] Assael, Y., Shillingford, B., Whiteson, S., & Nando De Freitas, amp; (n.d.). LIPNET: END-TO-END SENTENCE LEVEL LIPREADING. https://arxiv.org/pdf/1611.01599.pdf

[3] Ben-Hamadou, A. (2014). Achraf Ben-Hamadou - MIRACL-VC1. Sites.google.com. https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1R.

[4] Ma, P., Wang, Y., Petridis, S., Shen, J., & Pantic, M. (n.d.). Training Strategies For Improved Lip Reading https://arxiv.org/pdf/2209.01383.pdf

[5] O'shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. https://arxiv.org/pdf/1511.08458.pdf

[6] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing, 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

[7] Gavrilova Y. (2021, August). What are Convolutional Neural Networks? Serokell Software Development Company. https://serokell.io/blog/introduction-to-convolutional-neural-networks

[8] Vrskova, R., Hudec, R., Kamencay, P., & Sykora, P. (2022). Human Activity Classification Using the 3DCNN Architecture. Applied Sciences, 12(2), 931. https://doi.org/10.3390/app12020931

[9] IBM. (2023). What is Overfitting? | IBM. www.ibm.com https://www.ibm.com/topics/overfitting