ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation Journal

Vol. 4, No. 2, June, 2023

Lip-Reading with Visual Form Classification using Residual Networks and Bidirectional Gated Recurrent Units

Anni [1*], Suharjito [1]

[1] *Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia.*

Received 21 January 2023; Revised 09 May 2023; Accepted 17 May 2023; Published 01 June 2023

Abstract

Lip-reading is a method that focuses on the observation and interpretation of lip movements to understand spoken language. Previous studies have exclusively concentrated on a single variation of residual networks (ResNets). This study primarily aimed to conduct a comparative analysis of several types of ResNets. This study additionally calculates metrics for several word structures included in the GRID dataset, encompassing verbs, colors, prepositions, letters, and numerals. This component has not been previously investigated in other studies. The proposed approach encompasses several stages, namely pre-processing, which involves face detection and mouth location, feature extraction, and classification. The architecture for feature extraction comprises a 3-dimensional convolutional neural network (3D-CNN) integrated with ResNets. The management of temporal sequences during the classification phase is accomplished through the utilization of the bidirectional gated recurrent units (Bi-GRU) model. The experimental results demonstrated a character error rate (CER) of 14.09% and a word error rate (WER) of 28.51%. The combination of 3D-CNN ResNet-34 and Bi-GRU yielded superior outcomes in comparison to ResNet-18 and ResNet-50. The correlation between increased network depth and enhanced performance in lip-reading models was not consistently observed. Nevertheless, the incorporation of additional trained parameters offers certain benefits. Moreover, it has demonstrated superior levels of precision in comparison to human professionals in the task of distinguishing diverse word structures.

Keywords: Deep Learning; Neural Networks; Residual Network; Speech Recognition; Viseme.

1. Introduction

Language is the ability of humans to communicate with each other. Verbal communication will be disrupted if there is hearing loss or noisy environmental conditions. Another alternative that can be used to communicate is by using sign language or lip-reading. However, both require special training as they are challenging to learn. Lip-reading is a technique that relies on visual interpretation to comprehend spoken words or sentences. The use of lip-reading arises as a must in situations when the auditory perception of the speaker's speech is hindered by ambient noise or when the comprehension of dialogue in a video is impeded due to the unavailability of audio. Moreover, it has the potential to be incorporated into biometric security systems designed for mobile devices [1].

As technology has advanced, lip-reading has been widely researched. One of the primary difficulties encountered in the practice of lip-reading is the limitation of visual representation for several phonemes, resulting in potential ambiguity in word interpretation. Viseme, which stands for a visual phoneme, is the shape of the lips to represent a specific sound. The word viseme was introduced by Fisher as a visual form of a phoneme [2]. For example, /s/ and /r/ are phonemes because they differentiate the meanings of the words: "sing" and "ring". One viseme can represent more than one phoneme [3]. For instance, the phonemes /b/, /p/, and /m/ have the same viseme.

[*] Corresponding author: anni@binus.ac.id

http://dx.doi.org/10.28991/HIJ-2023-04-02-010

Multiple classification schemas have been developed to categorize lip movements due to their potential interpretations, such as visemes [4], phonemes [5], and ASCII characters [6]. A viseme classification schema benefits over other methods because it can predict words that are not in the training phase. It is because a viseme can be categorized to match all possible spoken words. Since several languages share the same viseme, it can also be employed in numerous languages [7]. Lip movements in different languages have a similar pattern due to similarities in the development of human vocal organs, even though each language has its own unique grammar and pronunciation norms [8]. Lip-reading research has been carried out using various classification segments such as letters [9], numbers [10], syllables [11], words [12], and sentences [13].

The datasets used as training data are diverse. The commonly used datasets were lip-reading in the wild (LRW) [14] and lip-reading sentences 2 (LRS2) [15], both of which consist of news or event programs. Other datasets were created for research audio-visual speech recognition purposes, namely OuluVS2 [16] and GRID [17]. Other studies used custom datasets to fit their models, as their focus was mostly on classifying short speech segments.

In their study, Lu & Li [10] constructed an in-house dataset for the purpose of predicting numerical values within the range of 0 to 9. The integration of the visual geometry group (VGG) network and the long short-term memory (LSTM) with attention mechanisms resulted in the development of a feature extraction model. This model has demonstrated a high level of fault tolerance in the domain of image recognition. The VGG network was used for other micro-content [9]. The dataset is based on 2700 recordings of the letters being pronounced by 11 different people. Short speech segments with syllable-level models were developed for the purpose of recognizing novel words that were not included in the training phase [11]. The model architecture was built with a 3-dimensional convolutional neural network (3D-CNN) and tested on a dataset of self-recorded videos containing Indonesian phrases. These studies highlight the advancements in lip-reading for classifying short speech segments. However, there is still a need to extend the scope of lip-reading to accurately recognize and classify words with different structures, such as verbs, colors, and prepositions. To address these challenges, we investigated computing accurate measurements for different word structures, including verbs, colors, prepositions, letters, and numbers present within the dataset.

Fenghour et al. [4] presented a viseme prediction model using 3D-CNN with residual networks (ResNets) architecture. This is subsequently followed by sentence prediction using generative pre-trained transformers (GPT). Even though the model achieved high accuracy in classifying visemes, there was a significant decrease in the accuracy of word classification following the conversion. During the perplexity calculation stage, misclassifications have frequently occurred due to the existence of local optima in the implementation process of local beam search. These local optima pose a challenge at each iteration of the viseme sequence, leading to incorrect classifications. Defining a viseme is challenging due to the shorter duration of visemes; there is not enough temporal information to distinguish between the various classes [4] and requires more background information to detect small variations [18].

The word-level lip-reading method had been experimented on the LRW dataset using the two-stream network, which is 3D-CNN and bidirectional long short-term memory (Bi-LSTM). Optical flow and grayscale video as inputs can further improve performance. The results demonstrated the two-stream network's effectiveness in lip-reading [19]. Another study in the same dataset [20], consisting of a 3D-CNN ResNet-18 followed by a temporal convolutional network (TCN). The efficacy of the initial TCN designs is improved through the utilization of densely connected TCN (DC-TCN) [21]. The squeeze-and-excitation block was employed by the model. This technique is employed to capture more comprehensive attributes at higher temporal resolutions. Wang et al [12] utilized a 3D convolutional vision transformer (3DCvT). The method combines the strengths of both vision transformers and 3D convolutions to extract spatiotemporal features from continuous images. By leveraging the properties of convolutions and transformers, it can effectively capture local and global information from these images. The extracted features are subsequently fed into a bidirectional gated recurrent unit (Bi-GRU) for sequence modeling to improve the capture of overall correlation among feature sequences and accurately identify crucial information.

The sentence prediction lip-reading model received more attention from researchers. At the sentence level, a hybrid lip-reading network (HLR-Net) was developed [13]. The model consisted of two distinct phases, namely an encoder and a decoder. The encoder is constructed using the three inception layers that structure the spatiotemporal CNN (StCNN), gradient, and Bi-GRU layers. The decoder uses the connectionist temporal classification (CTC) loss function and is built using the attention layer, fully connected, and activation functions. Jeon et al. [22] developed a lip-reading model at sentence level with multiple visual feature extraction methods. It was accomplished by employing a combination of 3D-CNN, 3D DenseNet, and multi-layer feature fusion (MLFF) 3D-CNN. However, performing automatic speech recognition solely based on visual speech recognition is still a challenging task due to the reliance on both acoustic and visual cues in spoken language. Visual recognition of mouth movements is a crucial aspect to consider in the creation of a lip-reading model. Furthermore, the inclusion of a temporal sequence modeling component is necessary. This component often involves training a language model capable of disambiguating distinct lip forms. One notable limitation of lip-reading models in practical settings is their considerable dimensions and insufficient computational efficacy.

Prior research has primarily focused solely on a singular variant of ResNets. The primary objective of this work was to perform a comparative examination of various ResNet architectures. This research utilizes different iterations of the

ResNets to assess the efficacy of the model, encompassing variations in the number of trained parameters and layers. The suggested methodology has multiple stages, including pre-processing, which incorporates the tasks of face detection and mouth localization, feature extraction, and classification. The feature extraction architecture consists of a 3D-CNN combined with ResNets. The classification phase involves the control of temporal sequences, which is achieved by employing the Bi-GRU model. Furthermore, this work computes metrics for several word structures present in the GRID dataset, which consist of verbs, colors, prepositions, letters, and numerals. This component has not been subject to prior investigation in previous studies. The findings obtained from this study can offer valuable insights into the assessment of the precision attributed to individual word structures.

2. Materials and Methods

2.1. Data

The dataset used was GRID [17], an audio-visual sentence corpus for research purposes, that consists of color videos and audio in English recorded from 33 speakers (18 men and 15 women) with a resolution of 360 x 288. Each speaker utters 1,000 short sentences with a six-word sequence. All videos are 75 frames (about 3 seconds) in length. The spoken sentence structure is command + color + preposition + letter + number + adverb. Each sentence was chosen randomly from the combination of words listed in Table 1.

Table 1. Word structure in GRID dataset

Command	Color	Preposition	Letter	Number	Adverb
Place	Red	With	A to Z (Not include W)	0 to 9	Soon
Bin	Green	In			Again
Set	Blue	At			Now
Lay	White	By			Please

The GRID dataset contains 51 unique words, including 25 letters (excluding the letter 'W' since it is classified as multi-syllable), 10 numbers (from zero to nine), and 4 words for each command, color, preposition, and adverb. A combination of color, letter, and number were the keywords. All speakers utter all these keywords. An example of a spoken sentence is "Bin Blue By Z Eight Now". The dataset includes metadata files that list the frame time for each word. An example of a metadata file can be seen in Table 2. The word “sil” at the beginning and end is silent time.

Table 2. Metadata file in GRID dataset

Start Frame	End Frame	Word
0	15500	sil
15500	20500	bin
20500	25500	blue
25500	30000	by
30000	37000	z
37000	42500	eight
42500	49250	now
49250	74500	sil

The training and validation datasets are divided according to speaker numbers. The total video used is 32747 of 33000, and 253 videos were corrupted. The evaluation data consisted of 2 males and 2 females from speaker numbers 1, 2, 20, and 22. The rest of the video was used for training.

2.2. Proposed Model

Lip-reading automation modeling is divided into several processes. Firstly, during data pre-processing, the input video was converted into images, and the image's mouth region was cropped and saved into image files (*.png). In the training phase, images will be labeled based on dataset metadata, followed by data augmentation and normalization. Next is the feature extraction process, which reduces data dimensions by identifying key features that are more informative and non-redundant. The extracted data were used as input for the classification process. The end of the process is sentence prediction. This section proposes an overall process for lip-reading, illustrated in Figure 1.

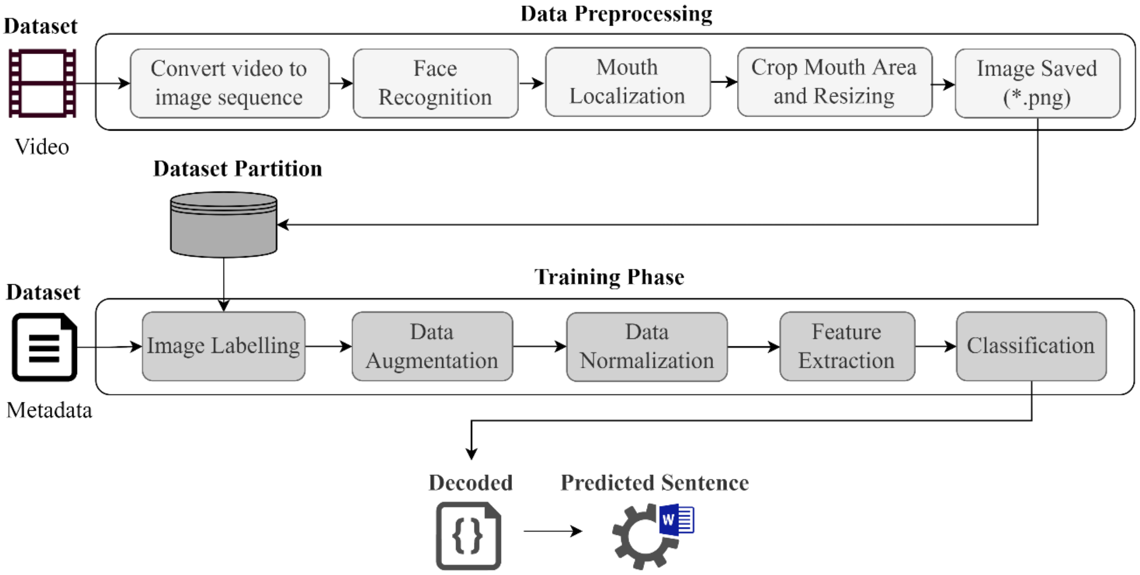

Figure 1. Overall lip-reading process

The Dlib and OpenCV libraries were used to perform mouth extraction. A shape predictor identifies features in an input image, such as the mouth, nose, and eyes. Facial landmark detection first locates the face and then detects the mouth within that region. The frames were cropped around the mouth area and downsized to 100 x 50. An illustration of mouth area localization is shown in Figure 2.

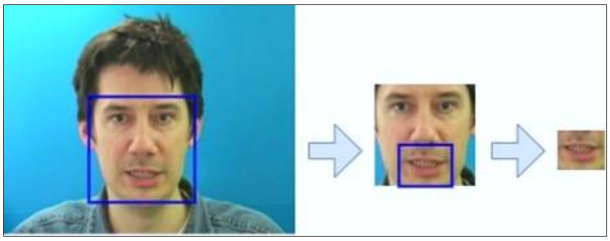

Figure 2. Mouth area localization

This model includes techniques for data augmentation; random horizontal flipping is applied to each frame with a probability of 0.5. Additionally, frames are duplicated or deleted with a probability of 0.05 per frame. Next is the data normalization process. This step ensured that each input had a similar data distribution to make convergence faster during training. The data is normalized to intervals of 0 and 1, dividing the pixel value by 255.

Figure 3 shows the proposed architecture of the lip-reading model. It uses a 3D-CNN with ResNets to extract spatial and temporal features to get fixed-length feature vectors. After the convolution and pooling layers, this architecture has a Bi-GRU layer, which is followed by a fully connected layer that uses softmax activation and CTC loss functions to classify. At the end of the sequence, it generates classes of probabilities to predict future input characters based on previous ones.

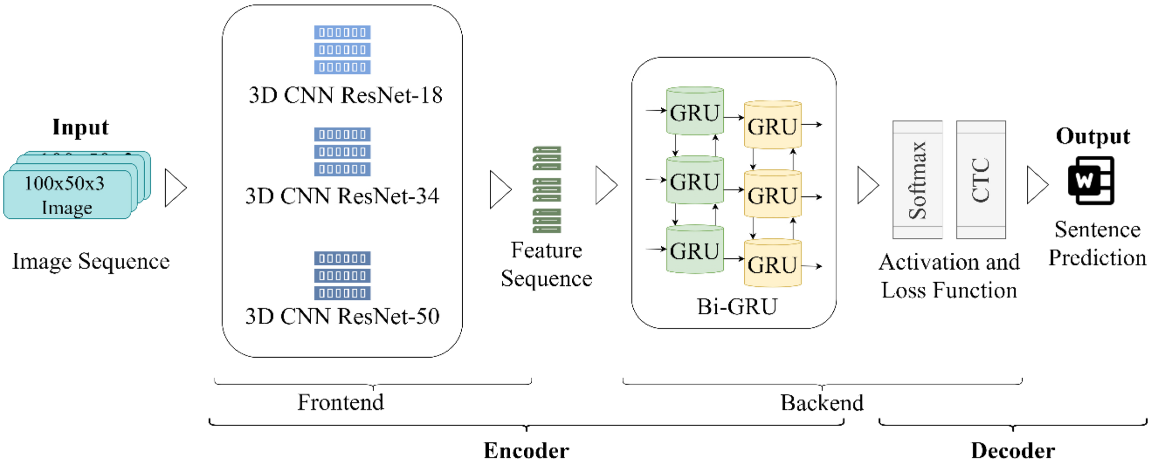

Figure 3. Proposed lip-reading model architecture with 3 compared ResNets variant

2.3. 3-Dimensional Convolutional Neural Network (3D-CNN)

Several convolutional network models have been developed to extract image features, but their applicability in video analysis is limited owing to the absence of motion modeling. The current study utilized a feature extraction model known as 3D-CNN with ResNets. 3D-CNN showed superior performance results in video analysis. It has an additional dimension that can retain temporal information from input to produce output volume. 3D-CNN efficiently summarizes object, scene, and action-related data within videos, thus offering versatility without necessitating model fine-tuning for each task [23].

2.4. Residual Neural Networks (ResNets)

ResNets is highly well-liked for image recognition and classification because it can overcome degradation by skipping layers (skipping connections). This skip connection can solve the problem of vanishing gradients by allowing gradients to carry out the alternative shortcut to skip unnecessary paths. ResNets are easier to train due to their simple topology and short interconnections among layers. Additionally, they exhibit proficiency in detecting features within lower-dimensional data representations, enabling them to acquire knowledge from small datasets [24]. ResNets consists of several variants, the distinction between them is the number of layers it forms, such as 18, 34, 50, 101, or 152 layers. The architecture of variant ResNets [25] is shown in Figure 4. For instance, Res-Net18 is built from 18 layers of a neural network. The first layer is a 7x7 kernel, followed by max pooling (stride 2), and four identical convolution layers. Each layer consists of two residual blocks and two layers with a skip connection. The size of the kernel in the convolution layer is 3x3 except for the first layer (7x7), and the number of parameters in each layer is 64, 128, 256, and 512. The maximum pooling layer uses stride 2.

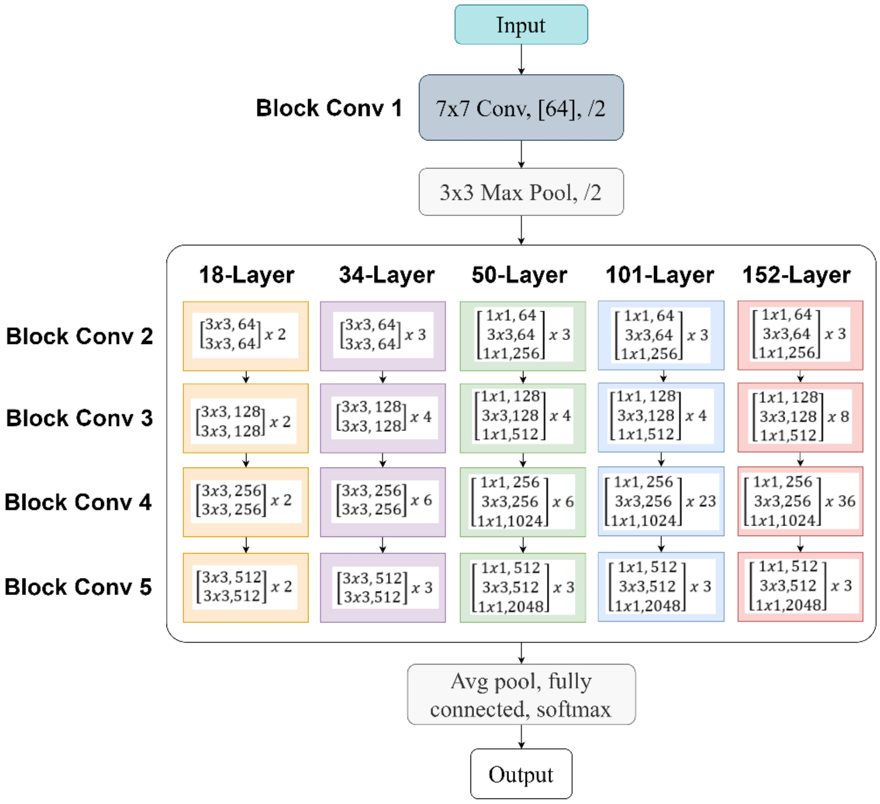
Input
Block Conv 1
7x7 Conv, [64], /2
3x3 Max Pool, /2
18-Layer
34-Layer
50-Layer
101-Layer
152-Layer
Block Conv 2
Block Conv 3
Block Conv 4
Block Conv 5
[3x3, 64; 3x3, 64] x 2
[3x3, 64; 3x3, 64] x 3
[1x1, 64; 3x3,64; 1x1,256] x 3
[1x1, 64; 3x3,64; 1x1,256] x 3
[1x1, 64; 3x3,64; 1x1,256] x 3
[3x3, 128; 3x3, 128] x 2
[3x3, 128; 3x3, 128] x 4
[1x1, 128; 3x3,128; 1x1,512] x 4
[1x1, 128; 3x3,128; 1x1,512] x 4
[1x1, 128; 3x3,128; 1x1,512] x 8
[3x3, 256; 3x3, 256] x 2
[3x3, 256; 3x3, 256] x 6
[1x1, 256; 3x3,256; 1x1,1024] x 6
[1x1, 256; 3x3,256; 1x1,1024] x 23
[1x1, 256; 3x3,256; 1x1,1024] x 36
[3x3, 512; 3x3,512] x 2
[3x3, 512; 3x3,512] x 3
[1x1, 512; 3x3,512; 1x1,2048] x 3
[1x1, 512; 3x3,512; 1x1,2048] x 3
[1x1, 512; 3x3,512; 1x1,2048] x 3
Avg pool, fully connected, softmax
Output

Figure 4. ResNets Architecture

2.5. Bidirectional Gated Recurrent Units (Bi-GRU)

The Recurrent Neural Network (RNN) is a neural network architecture that is frequently employed in several domains, including speech recognition, language modeling, and translation. Its main functionality lies in predicting the next word or character within a sequence of words. To address some limitations faced by RNNs, like vanishing gradients during training, an improved version called GRU was introduced by Cho et al. [26]. The goal behind developing GRU was to enhance the information flow throughout the network. GRUs are often considered like LSTM networks due to their comparable design and similarly promising results. However, one notable distinction lies in how they handle gating mechanisms. While LSTM employs forget gates and input gates separately for controlling memory cells' access at each time step independently, GRU combines these into an update gate. This enables efficient determination of relevant information for propagation towards output predictions.

In the backend process, we used a two-layer Bi-GRU. The Bi-GRU had the best prediction accuracy and the quickest learning convergence time compared to the unidirectional models, GRU and LSTM [18]. Bi-GRU provides information to two independent neural network topologies connected to the same output layer in both forward and reverse flows. Both networks receive complete input information, unlike the standard GRU deployment. The output of the 3D-CNN ResNets is delivered successively to the Bi-GRU layers, which produce characters as output.

2.6. CTC Loss Function

The CTC loss function obviates the necessity of pre-alignment between the sequence of input and output. It enables the independent prediction of labels for each time step. The vocabulary comprises tokens, including a 'blank' character representing '-', aiding in encoding repetitive characters. For instance, in the CTC configuration, 'Hel-lo' is the correct representation of 'Hello,' where 'l' is duplicated. The CTC loss function accepts a model output matrix consisting of scores assigned to each token at every time step alongside the actual truth sequence [27].

During training, the objective is either to optimize all possible routes leading up to the fundamental truth label or minimize negative log probability sums. Throughout the process of evaluation, a selection is made at each stage using either beam search or greedy methods to identify characters. The final recognition output sequence is then generated by removing redundant and null characters. The CTC loss function can be implemented on various levels, such as phonemes, visemes, or individual characters.

2.7. Performance Measurement

Model performance measurement uses a standard evaluation metric in automatic speech recognition at the sentence level, character error rate (CER) and word error rate (WER). CER measures how close the predicted character order is to the target character order, while WER is for words. The lower the CER or WER, the better the prediction accuracy. All models were evaluated to compare computational performance and efficiency. The CER and WER equations are determined in Equations 1 and 2, respectively. From Equations 1 and 2, N represents the entire character count in the fundamental truth, while S signifies substitution for incorrect classification. D denotes deletions of non-decoding characters, and I indicates the insertion of decoded characters not chosen.

$$CER(\%) = \left(\frac{C_S + C_D + C_I}{C_N}\right) \times 100 \tag{1}$$

$$WER(\%) = \left(\frac{W_S + W_D + W_I}{W_N}\right) \times 100 \tag{2}$$

The proposed model used phoneme-to-viseme mapping [2] to visualize the results. Table 3 shows the mapping. This mapping was considered the best match of the 15 mappings evaluated [28]. Viseme and phoneme mapping in English consists of 39 phonemes and 14 viseme classes (6 consonants, 7 vowels, and 1 silence viseme).

Table 3. Mapping of phoneme and viseme

Viseme Type	Phoneme	Viseme
Vowel	/ah/	ah
	/er/	er
	/ih/, /iy/	iy
	/uw/, /uh/	uh
	/ey/, /ae/, /eh/	eh
	/ay/, /aa/, /aw/	aa
	/ow/, /oy/, /ao/	ao
Consonant	/v/, /f/	f
	/w/, /r/	w
	/m/, /p/, /b/	p
	/jh/, /ch/, /zh/, /sh/	ch
	/z/, /s/, /t/, /d/, /dh/, /th/	t
	/n/, /k/, /g/, /l/, /y/, /ng/, /hh/	k
Silent Character	h#	#

3. Results and Discussion

The proposed models were evaluated using The Google Colab Pro+ version on GPU T4 with an allocation of 15 GB of GPU RAM and 51 GB of system RAM. A TensorFlow-CTC decoder was used to calculate the error rate scores for all experimental models, which were all developed using Keras with a TensorFlow backend. The ResNets have five variations: 18, 34, 50, 101, and 152 layers. We experimented with three variations, namely 18, 34, and 50. We employed the Adam optimizer [29] to train all our models. The learning rate used was set at 10^{-4} for a total of 250 epochs, and the batch size was 16.

3.1. CER and WER

To assess the computational effectiveness of the models, we examined the error rate in relation to the trained parameters and the duration of training. The evaluation was conducted with unseen speakers. The results are summarized in Table 4. The 3D CNN ResNet-34 and Bi-GRU combination gave the best result in terms of CER of 14.09% and WER of 28.51%. However, 3D CNN ResNet-18 and Bi-GRU have shorter training times of 205.45 hours (about 8.5 days). The 3D-CNN ResNet-50 and Bi-GRU models have the highest CER and WER values and the longest training time. The number of trainable parameters in the 3D-CNN ResNet-34 and Bi-GRU models is approximately 65.1 million, which contributes to their superior performance in sentence prediction compared to the other two models. We could not train models for layers 101 and 152 due to memory constraints in our test environment.

Table 4. Performance proposed models

Model	Trainable	Training Time	Unseen Speaker	
	Parameters	(hour)	CER (%)	WER (%)
3D CNN ResNet-18 + Bi-GRU	34.8 M	**205.45**	15.11	29.51
3D CNN ResNet-34 + Bi-GRU	65.1 M	242.82	**14.09**	**28.51**
3D CNN ResNet-50 + Bi-GRU	46.5 M	322.90	18.53	36.50
3D CNN ResNet-101 + Bi-GRU	86.9 M	-	-	-
3D CNN ResNet-152 + Bi-GRU	119.1 M	-	-	-

In Figure 5, we can observe the comparisons of CER for each model, while in Figure 6, we can observe the WER comparisons. During the initial 100 epochs, it is evident that the model utilizing a 3D CNN ResNet-50 exhibits a larger disparity in CER values compared to the other two models. However, as training progresses, it only demonstrates a slight discrepancy. On the other hand, there is only a slight difference when comparing WER between 3D CNN ResNet-18 and ResNet-34 models. The deeper architecture of 3D CNN ResNet-50 displays more significant variations than its counterparts. These observations imply that having a deeper network does not necessarily guarantee improved lip-reading performance; hence, further investigation is needed to understand this.

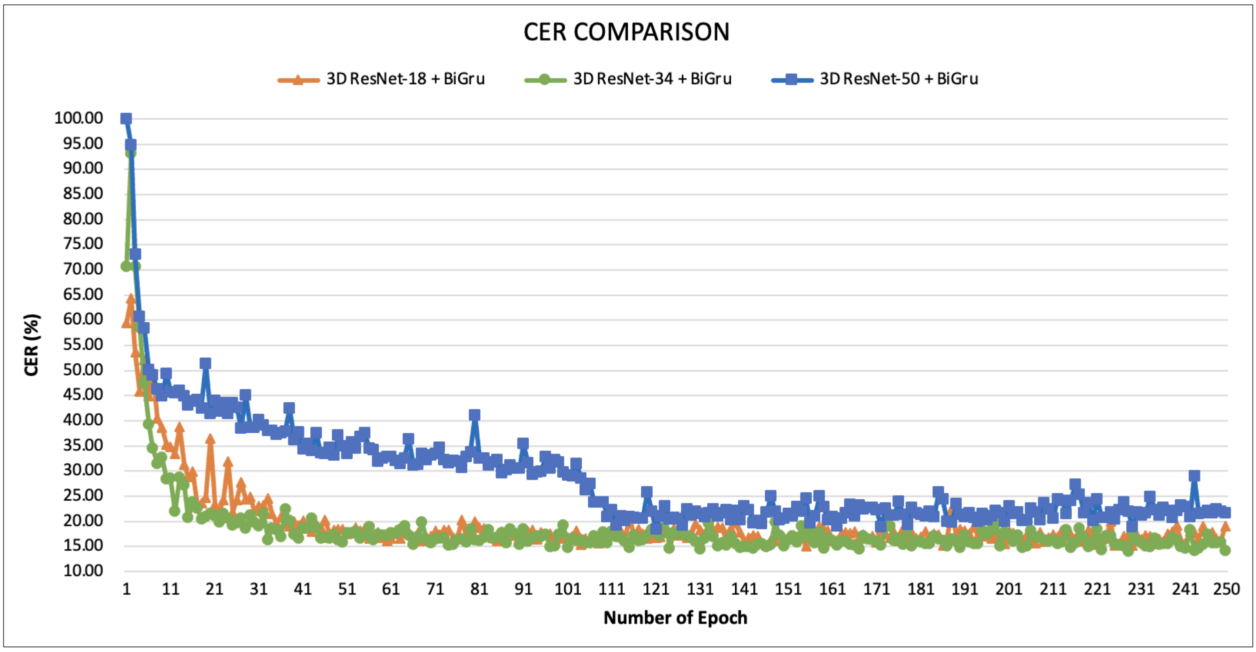

Figure 5. CER Comparison for each model

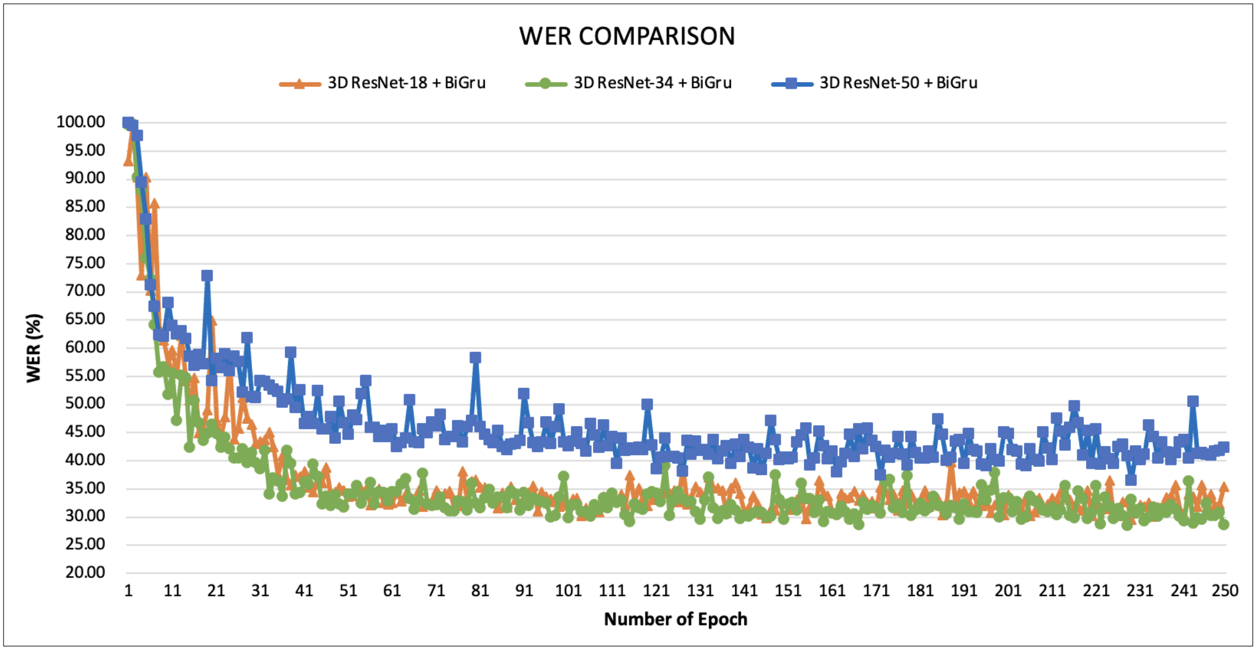

Figure 6. WER Comparison for each model

The more layers in ResNets, the longer the training time required. Increasing the number of trained parameters may be beneficial in ResNets models. However, this approach may not be suitable for other neural networks. This highlights the significance of selecting relevant features rather than relying solely on parameter quantity for enhancing model quality and performance. Training a neural network with numerous parameters poses a considerable computational challenge. The complexity increases when implementing the ResNets model due to the extensive memory requirements for storing and maintaining parameters and weight values, resulting in time-consuming training processes. Consequently, given the limitations of memory capacity in our testing environment, it is not feasible to train layers like ResNet-101 and ResNet-152. These findings also demonstrate that deeper networks such as ResNets are computationally expensive without necessarily leading to enhanced lip-reading performance.

Despite the challenges and limitations associated with training deep neural networks like ResNets for lip-reading, there are still promising opportunities for improving accuracy and performance in this field. One potential avenue for improvement is the exploration of ensemble learning methods in lip-reading models. By combining multiple models, more accurate predictions can be made. Another approach to addressing the computational challenges of training deep neural networks is through model compression. Model compression techniques, such as sparsity via regularization, weight quantization, and network pruning, have shown promise in reducing the memory usage and computation requirements of deep networks. For instance, weight quantization replaces trained network weights with lower precision or utilizes bit-wise operations. These techniques can be applied to lip-reading models based on deep convolutional neural networks such as MobileNet, VGG16, and AlexNet [30]. However, the challenges of training deep neural networks, particularly models like ResNets with their memory requirements, highlight the need for alternative approaches to improve computational efficiency and performance.

3.2. Confusion Matrix

To visualize the results, the proposed approach utilized phoneme-to-viseme mapping [2]. Figure 7 presents the confusion matrix for viseme prediction. The results revealed that while the proposed model successfully differentiated most visemes, but there were several misclassifications observed. Viseme "ch", which mapped the phoneme groups {/jh/, /ch/, /zh/, /sh/} had a high frequency of incorrect classification. Examples of data included in viseme "ch" are the letters g, h, and j. These findings are consistent with previous research [4] that has also highlighted the challenges associated with mapping multiple phonemes to a single viseme.

We selected 10 sample sentences to illustrate the results of our predictions. Table 5 displays example sentences from the GRID dataset, along with their corresponding predicted sentences. Any inaccuracies in the predictive sentences are marked using bold and underlined formatting. Some complete sentences may align perfectly with the predictions, while others might contain one or more incorrect words. This is to be expected since lip-reading is dependent primarily on the visible articulators, which include the lips, the tongue, and the teeth to some extent.

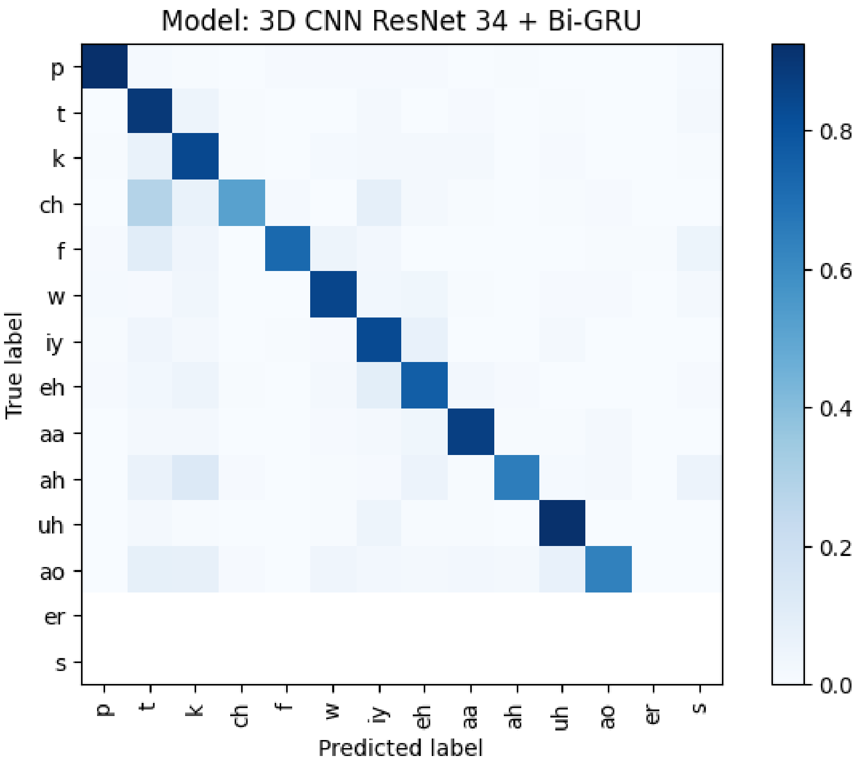

Figure 7. Confusion matrix for viseme prediction

Table 5. Example of proposed model results

Target Sentences	Predicted Sentences
lay blue with c five please	lay blue with c five please
set red by s seven now	set red by s seven now
place red with u eight again	place red with u eight again
set white by i six please	set **hited** by i six please
set white in f six soon	set white **it** f six soon
bin white in d nine now	bin white in **c** nine now
place white in n seven please	place white in **a** seven please
set white with a five please	set white with a **fire** please
place green with p six again	place green with **v** six **gain**
lay green with y five now	**let** **reed** with **b** five now

3.3. Prediction Sentence Structure

Table 6 displays the level of accuracy attained in successfully predicting different word structures. The predictive accuracy rate for letter prediction was observed to be 35.79%, indicating a considerably lower performance compared to other components of the test. On the other hand, it was observed that command words exhibited the highest rate of prediction, with an accuracy of 88.65%. This outcome was attained through a confluence of various causes. The observed discrepancy can potentially be accounted for by the temporal duration of letter sounds being shorter than 0.3 seconds. Furthermore, distinguishing visually indistinguishable visemes, such as the phonemes "p" and "b" or "f" and "v," might be a significant challenge due to their closely related visual characteristics. For instance, specific letters such as b, d, c, and e necessitate similar oral articulatory gestures for accurate pronunciation. This phenomenon impacts word prediction and poses challenges for the visual system's acquisition of novel information. This has led to the revelation of an additional technological limitation associated with visual speech recognition systems.

Table 6. Accuracy for prediction sentence structure on GRID test data

	Command	Color	Preposition	Letter	Number	Adverb
Accuracy (%)	88.65	85.85	68.99	35.79	67.18	85.40

The accuracy rates of human interpreters were used as a baseline model in an audio speech recognition study [31]. To evaluate the effectiveness of our proposed model, we conducted a comparative analysis between its performance in lip-reading tasks and that of humans in lip-reading tasks. The findings of our investigation, as provided in Table 7, demonstrate that the proposed model outperformed humans across all sentence structures. This remarkable achievement can be attributed to significant advancements in deep learning methodologies. The application of these techniques has yielded notable advancements in the field of lip-reading, enabling successful utilization in many real-life scenarios. As a result, this noteworthy accomplishment was rendered feasible.

Table 7. Accuracy for prediction sentence structure compared with human interpreters

Method	Command	Color	Preposition	Letter	Number	Adverb
Human Interpreters [31]	57.30	75.00	43.80	17.70	41.40	78.10
Proposed Model	**88.65**	**85.85**	**68.99**	**35.79**	**67.18**	**85.40**

4. Conclusion

This paper proposes a deep learning-based lip-reading system using 3D CNN ResNets and Bi-GRU. Different ResNet architectures were compared to see which one was best at predicting full sentences from a series of images of the lip area. The most accurate model was achieved by combining 3D CNN ResNet-34 with Bi-GRU, which obtained a CER of 14.09% and a WER of 28.51% on unseen speakers in experiments conducted on the GRID dataset. We demonstrated that our model outperformed humans across all sentence structures. This remarkable achievement showcases its effectiveness in real-world scenarios and highlights the transformative impact of recent developments in deep learning.

Increasing the number of layers in ResNets results in longer training durations. Although introducing more trained parameters may have its advantages, this approach may not be compatible with all neural network architectures. Instead of solely relying on parameter count to enhance model quality and performance, it is crucial to focus on selecting relevant features. Training a neural network that has many parameters poses computational challenges due to the extensive memory resources required to store and update all the weights and values. Consequently, these training processes become time-consuming. These findings also suggest that while deeper networks such as ResNets are computationally demanding, they do not necessarily translate into improved lip-reading capabilities. Despite the obstacles and constraints, there are favorable prospects for enhancing accuracy and performance. A potential approach to improving this is by investigating ensemble learning methods in lip-reading models. Ensemble learning entails the fusion of multiple models to generate more precise predictions, thereby enabling the utilization of diverse viewpoints and amplifying overall performance.

5. Declarations

5.1. Author Contributions

Conceptualization, A.A. and S.S.; methodology, A.A. and S.S.; software, A.A.; validation, A.A. and S.S.; formal analysis, A.A. and S.S.; writing—original draft preparation, A.A.; writing—review and editing, S.S.; visualization, A.A.; supervision, S.S. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.3625687 [32].

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Acknowledgements

We would like to thank Bina Nusantara University, especially the Binus Graduate Program, Master of Computer Science, for giving us the chance to conduct the research.

5.5. Institutional Review Board Statement

Not applicable.

5.6. Informed Consent Statement

Not applicable.

5.7. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6. References

[1] Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Kong, L., & Li, M. (2019). Lip Reading-Based User Authentication through Acoustic Sensing on Smartphones. IEEE/ACM Transactions on Networking, 27(1), 447–460. doi:10.1109/TNET.2019.2891733.

[2] Lee, S., & Yook, D. (2002). Audio-to-Visual Conversion Using Hidden Markov Models. PRICAI 2002: Trends in Artificial Intelligence, 563–570, Springer. doi:10.1007/3-540-45683-x_60.

[3] Bagherzadeh, S. Z., & Toosizadeh, S. (2022). Eye tracking algorithm based on multi model Kalman filter. HighTech and Innovation Journal, 3(1), 15-27. doi:10.28991/HIJ-2022-03-01-02.

[4] Fenghour, S., Chen, D., Guo, K., & Xiao, P. (2020). Lip Reading Sentences Using Deep Learning with only Visual Cues. IEEE Access, 8, 215516–215530. doi:10.1109/ACCESS.2020.3040906.

[5] El-Bialy, R., Chen, D., Fenghour, S., Hussein, W., Xiao, P., Karam, O. H., & Li, B. (2023). Developing phoneme-based lip-reading sentences system for silent speech recognition. CAAI Transactions on Intelligence Technology, 8(1), 129–138. doi:10.1049/cit2.12131.

[6] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2022). Deep Audio-Visual Speech Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12), 8717–8727. doi:10.1109/TPAMI.2018.2889052.

[7] 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:10.1109/ACCESS.2021.3107946.

[8] Luo, M., Yang, S., Chen, X., Liu, Z., & Shan, S. (2020). Synchronous Bidirectional Learning for Multilingual Lip Reading (ArXiv Preprint). doi:10.48550/arXiv.2005.03846.

[9] Ali, N. H., Abdulmunim, M. E., & Ali, A. E. (2021). Constructed model for micro-content recognition in lip reading based deep learning. Bulletin of Electrical Engineering and Informatics, 10(5), 2557–2565. doi:10.11591/eei.v10i5.2927.

[10] Thammarak, K., Sirisathitkul, Y., Kongkla, P., & Intakosum, S. (2022). Automated Data Digitization System for Vehicle Registration Certificates Using Google Cloud Vision API. Civil Engineering Journal, 8(7), 1447-1458. doi:10.28991/CEJ-2022-08-07-09.

[11] Kurniawan, A., & Suyanto, S. (2020). Syllable-Based Indonesian Lip Reading Model. 2020 8th International Conference on Information and Communication Technology (ICoICT). doi:10.1109/icoict49345.2020.9166217.

[12] Nurhidayat, I., Pimpunchat, B., Noeiaghdam, S., & Fernández-Gámiz, U. (2022). Comparisons of SVM kernels for insurance data clustering. Emerging Science Journal, 6(4), 866-880. doi:10.28991/ESJ-2022-06-04-014.

[13] Sarhan, A. M., Elshennawy, N. M., & Ibrahim, D. M. (2021). HLR-net: a hybrid lip-reading model based on deep convolutional neural networks. Computers, Materials & Continua, 68(2), 1531-1549. doi:10.32604/cmc.2021.016509.

[14] Chung, J. S., & Zisserman, A. (2017). Lip Reading in the Wild. Lecture Notes in Computer Science, 87–103, Springer. doi:10.1007/978-3-319-54184-6_6.

[15] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip Reading Sentences in the Wild. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2017.367.

[16] Anina, I., Ziheng Zhou, Guoying Zhao, & Pietikainen, M. (2015). OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia. doi:10.1109/fg.2015.7163155.

[17] Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), 2421–2424. doi:10.1121/1.2229005.

[18] Jeon, S., & Kim, M. S. (2022). End-to-End Sentence-Level Multi-View Lipreading Architecture with Spatial Attention Module Integrated Multiple CNNs and Cascaded Local Self-Attention-CTC. Sensors, 22(9), 3597. doi:10.3390/s22093597.

[19] Weng, X., & Kitani, K. (2019). Learning spatio-temporal features with two-stream deep 3d CNNs for lipreading. arXiv preprint arXiv:1905.02540. doi:10.48550/arXiv.1905.02540.

[20] Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading Using Temporal Convolutional Networks. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). do:10.1109/icassp40776.2020.9053841.

[21] Ma, P., Wang, Y., Shen, J., Petridis, S., & Pantic, M. (2021). Lip-reading with Densely Connected Temporal Convolutional Networks. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). doi:10.1109/wacv48630.2021.00290.

[22] Jeon, S., Elsharkawy, A., & Kim, M. S. (2022). Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition. Sensors, 22(1), 72. doi:10.3390/s22010072.

[23] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV). doi:10.1109/iccv.2015.510.

[24] Haque, I., Alim, M., Alam, M., Nawshin, S., Noori, S. R. H., & Habib, M. T. (2022). Analysis of recognition performance of plant leaf diseases based on machine vision techniques. Journal of Human, Earth, and Future, 3(1), 129-137. doi:10.28991/HEF-2022-03-01-09.

[25] 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). doi:10.1109/cvpr.2016.90.

[26] Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. doi:10.3115/v1/w14-4012.

[27] Oghbaie, M., Sabaghi, A., Hashemifard, K., & Akbari, M. (2021). Advances and Challenges in Deep Lip Reading. arXiv preprint arXiv:2110.07879. doi:10.48550/arXiv.2110.07879.

[28] Bear, H. L., Harvey, R. W., Theobald, B.-J., & Lan, Y. (2014). Which Phoneme-to-Viseme Maps Best Improve Visual-Only Computer Lip-Reading? Lecture Notes in Computer Science, 230–239, Springer. doi:10.1007/978-3-319-14364-4_22.

[29] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. doi:10.48550/arXiv.1412.6980.

[30] Lu, Y., Xiao, Q., & Jiang, H. (2021). A Chinese Lip-Reading System Based on Convolutional Block Attention Module. Mathematical Problems in Engineering, 2021, 1–12. doi:10.1155/2021/6250879.

[31] Le Cornu, T., & Milner, B. (2017). Generating Intelligible Audio Speech from Visual Speech. IEEE/ACM Transactions on Audio Speech and Language Processing, 25(9), 1751–1761. doi:10.1109/TASLP.2017.2716178.

[32] Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). The Grid Audio-Visual Speech Corpus. Zenodo, Open Science. doi:10.5281/zenodo.3625687.