algorithms

MDPI

Article

Efficient DNN Model for Word Lip-Reading

Taiki Arakane and Takeshi Saitoh *

Department of Artificial Intelligence, Kyushu Institute of Technology, Fukuoka 820-8502, Japan
* Correspondence: saitoh@ai.kyutech.ac.jp

Abstract: This paper studies various deep learning models for word-level lip-reading technology, one of the tasks in the supervised learning of video classification. Several public datasets have been published in the lip-reading research field. However, few studies have investigated lip-reading techniques using multiple datasets. This paper evaluates deep learning models using four publicly available datasets, namely Lip Reading in the Wild (LRW), OuluVS, CUAVE, and Speech Scene by Smart Device (SSSD), which are representative datasets in this field. LRW is one of the large-scale public datasets and targets 500 English words released in 2016. Initially, the recognition accuracy of LRW was 66.1%, but many research groups have been working on it. The current the state of the art (SOTA) has achieved 94.1% by 3D-Conv + ResNet18 + {DC-TCN, MS-TCN, BGRU} + knowledge distillation + word boundary. Regarding the SOTA model, in this paper, we combine existing models such as ResNet, WideResNet, WideResNet, EfficientNet, MS-TCN, Transformer, ViT, and ViViT, and investigate the effective models for word lip-reading tasks using six deep learning models with modified feature extractors and classifiers. Through recognition experiments, we show that similar model structures of 3D-Conv + ResNet18 for feature extraction and MS-TCN model for inference are valid for four datasets with different scales.

Keywords: lip-reading; word recognition; deep neural network; LRW; OuluVS; CUAVE; SSSD; 3D convolutional layer; ResNet; WideResNet; EfficientNet; transformer; ViT; ViViT; MS-TCN

Citation: Arakane, T.; Saitoh, T. Efficient DNN Model for Word Lip-Reading. *Algorithms* **2023**, *16*, 269. https://doi.org/10.3390/a16060269

Academic Editors: Chih-Lung Lin, Bor-Jiunn Hwang, Shaou-Gang Miaou and Yuan-Kai Wang

Received: 28 April 2023
Revised: 24 May 2023
Accepted: 26 May 2023
Published: 27 May 2023

1. Introduction

This paper focuses on word lip-reading technology that estimates the utterance content from visual information only without audio information. This paper uses "words," but more precisely, it includes both words and short phrases. This technology is expected to be used in the following cases where it is difficult to use audio-based speech recognition: it is used in noisy environments where it is difficult to obtain speech, in public places where it is difficult to speak, and used by people with disabilities who cannot speak due to laryngectomy. Since various problems can be solved using lip-reading technology, it is expected to be one of the next-generation communication tools.

As an academic framework, lip-reading technology is classified as supervised learning for video data. There are several topics for research on lip-reading; shooting directions such as frontal and side [1], recognition targets such as single sound [2,3], word [4–9], and sentence [10–12]. Word recognition is an active research topic, and various algorithms have been proposed.

Word lip-reading has been studied since the early days of lip-reading technology. However, research has become active with the release of datasets such as OuluVS [13], CUAVE [14], SSSD [15], LRW [4], CAS-VSR-W1k (LRW-1000) [16], and RUSAVIC [17], and the introduction of deep learning. In particular, research groups using LRW, one of the large-scale datasets, have been competing for several percent accuracies in recent years (https://paperswithcode.com/sota/lipreading-on-lip-reading-in-the-wild, accessed on 26 April 2023).

This paper explores various deep learning models and their effectiveness in word lip-reading. While many papers use one or two datasets, this paper conducts experiments on four publicly available datasets; LRW, OuluVS, CUAVE, and SSSD.

This paper is organized as follows. Section 2 describes the related research. Section 3 summarizes the basic model of the deep learning model considered in this paper. Section 4 introduces the deep learning model investigated in this paper. Section 5 shows recognition experiments on four datasets, and Section 6 concludes this paper.

2. Related Research

There are many studies on word lip-reading techniques. Here, this paper focuses on research targeting LRW.

LRW is a dataset published by Chung et al. in 2016 [4] (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html, accessed on 26 April 2023) and is a large-scale lip-reading research dataset containing 500 English words. LRW is used as a benchmark in the lip-reading field. LRW contains utterance scenes clipped from news and discussion programs broadcast from 2010 to 2016 by the British Broadcasting Corporation (BBC), which collectively manages radio and television in the United Kingdom. While most datasets are utterance scenes in which the speaker was recording in a roughly specified posture, LRW has the feature of recording utterance scenes in a natural posture even though it is a TV program. The number of speakers is more than 1000. LRW consists of three types of data: training data, validation data, and test data, and provides video data containing 488,766 scenes, 25,000 scenes, and 25,000 scenes, respectively. The train data contain 800–1000 scenes per word, and the validation and test data contain 50 scenes per word. All scenes are extracted face images with 256 × 256 pixels, the frame rate is 25 fps, the scene length is 1.16 s, and the number of frames is 29. The total recording time of the three types is about 173 h.

CAS-VSR-W1k [16], which contains 1000 Chinese words, is a large-scale open dataset containing word utterance scenes. The state-of-the-art (SOTA) recognition accuracies of LRW and CAS-VSR-W1k are 94.1% [18] and 55.7% [19], respectively. Academically, it is desirable to target CAS-VSR-W1k, which has a high degree of task difficulty. Since we could not obtain CAS-VSR-W1k, we target LRW.

Chung et al., who published LRW, proposed four convolutional neural network (CNN) models based on VGG-M [4]. Among the four models, the multiple towers model obtained the highest recognition rate of 66.1%. This model has a structure in which the convolutional layers for all frames are provided in the first layer. The outputs of all subsequent convolutional layers are connected to one and input to the second convolutional layer. The following year, Chung et al. proposed a new network with a watch, listen, attend, and spell (WLAS) structure, obtaining 76.2% [10]. In WLAS, the encoder consists of a CNN model that is an improved version of VGG-M, which extracts features from each input frame image, and a long short-term memory (LSTM) that summarizes the output of the features from the CNN model. The decoder consists of LSTM, attention, and softmax.

According to the paper with the code site (https://paperswithcode.com/sota/lipreading-on-lip-reading-in-the-wild, accessed on 26 April 2023), SOTA in LRW is currently the result of Ma et al. [18]. The model architecture consists of 3D-Conv + ResNet18 in the front stage and a Temporal model in the backstage with a mouth region of interest (ROI) as input. The temporal model is an ensemble of three different models; densely connected temporal convolutional networks (DC-TCNs), multi-scale temporal convolutional networks (MS-TCNs), and bidirectional gated recurrent units (BGRUs). In addition to the model architecture, they applied data augmentation, self-distillation, and word boundary indicators to improve the recognition accuracy. Many other papers have recently discussed the model training strategy, but the model architecture is often 3D-Conv + ResNet18 + MS-TCN [6,8,20]. The second highest accuracy in LRW is 3D Conv + EfficientNetV2 + Transformer + TCN structure, which obtained a recognition rate of 89.5%, proposed by Koumparoulis et al. [21]. The structure with the third highest accuracy 88.7% is Vosk + MediaPipe + LS + MixUp + SA + 3DResNet-18 + BiLSTM +

CosineWR [22], where Vosk is a voice activity detection model, which can detect speech regions even in heavy acoustically noisy conditions. MediaPipe is a machine-learning library provided by Google (https://developers.google.com/mediapipe, accessed on 26 April 2023). LS means label smoothing [23], SA means a squeeze-and-attention (SA) module, and CosineWR means cosine annealing warm restarts. This study uses two datasets, LRW and RUSAVIC [17], for evaluation. The Russian Audio-Visual Speech in Cars (RUSAVIC) is a multi-speaker and multi-modal corpus. The number of speakers is 20, and the number of phrases is 68.

3. Basic Model

In this section, we explain preprocessing and summarize the basic models of deep learning, data augmentation, distance learning, and fine-tuning discussed in this paper.

3.1. Preprocessing

To begin the process, we follow the same preprocessing steps as the existing method by our previous research [15]. The datasets we are working with, namely OuluVS and CUAVE, contain not only the speaker's face but also their upper body and the background. Hence, we first extract the face rectangle from the input image using face detection processing. Several face detectors have been proposed, including non-deep learning approaches that use Haar-like features and histograms of oriented gradients (HOG) [24,25] and deep learning approaches such as RetinaFace [26]. This paper uses the face detector implemented in the dlib library (http://dlib.net/, accessed on 26 April 2023).

Facial landmark detection helps determine the location of facial parts such as eyes, eyebrows, nose, and lips. This is an important process for stable ROI extraction. In this paper, we utilize the method proposed by Kazami and Sullivan [27], which is a typical facial landmark detection method implemented in the dlib library. A total of 68 facial landmarks are detected.

The size and rotation normalization process is applied based on the detected facial landmarks. At first, two variables of d_{eye}, the distance between two eyes, and θ, the angle between two eyes, are calculated. Then, an affine transformation is applied using d_{eye} and θ. Specifically, the scale is changed so that d_{eye} becomes 200 pixels, and the image is rotated so that θ becomes 0 degrees.

Then, the following equation extracts the upper left coordinate (L, T), lower right coordinate (R, B), and the size of $S \times S$ pixels of the lipROI.

$$\begin{aligned} L &= (x_{llip} + x_{rlip})/2 - S/2, \\ T &= (y_{llip} + y_{rlip})/2 - S/3, \\ R &= L + S, \\ B &= T + S. \end{aligned}$$

Here, the two points (x_{llip}, y_{llip}) and (x_{rlip}, y_{rlip}) are the landmark coordinates of the left and right corners of the mouth, respectively. The extracted lipROI is fed to the deep learning model.

3.2. Three-Dimensional Convolution

Many deep learning models investigated in this paper will be described later in Section 4, which extract features using ResNet. The input data are time-series image data (lipROIs). For this reason, we first apply a 3D convolution. Specifically, the structure shown in Figure 1 is used.

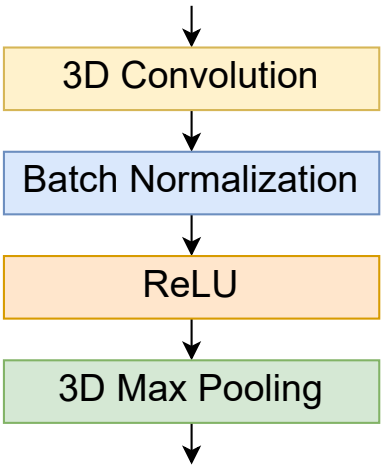

Figure 1. Structure of a 3D convolutional layer.

3.3. Resnet

Residual network (ResNet) [28] is a well-known model that introduces a residual block and a shortcut connection into an existing CNN model. The convolutional layer of CNN extracts features by combining them with the pooling layer, and it is thought that advanced and complex features can be extracted by stacking layers. However, when deep structures are used, there is a problem that training does not progress due to the gradient vanishing or exploding gradients. Therefore, ResNet solves the problem by training the residual function referenced from the layer's input instead of training the optimal output found in a layer. The residual block combines a convolutional layer and a skip connection, summing the outputs of the two passes. One of the residual blocks combines convolutional layers, and the other is the identity function. If this structure does not require transformations in additional layers, it can be handled by setting the weight to zero.

3.4. WideResNet

WideResNet [29] is an improved model of ResNet. The purpose of ResNet is to deepen the layers, but there is a problem stemming from the fact that deeper layers implies a lower computational efficiency in terms of performance. This is believed because many layer weights become meaningless, which is called the reduced feature reuse problem. WideResNet was proposed as a solution to this problem. WideResNet improves the computational efficiency and performance by increasing the number of channels for convolution in the residual block and introducing the dropout.

3.5. EfficientNet

In order to improve the accuracy of the image classification model, various measures, such as increasing the number of layers, widening the width (channel) of the model, and increasing the resolution of the input image, were implemented independently. On the other hand, EfficientNet [30] is a model that introduces a compound coefficient that simultaneously performs three changes in a well-balanced manner. EfficientNet proposed eight models, namely EfficientNet-B0–EfficientNet-B7, which are automatically designed using neural architecture search (NAS) [31]. NAS is a method that automatically optimizes a dedicated network structure to scale the composite coefficients in a balanced manner.

3.6. Transformer

Transformer is a model that uses only attention without recurrent neural network (RNN) or CNN [32]. Transformer is based on the encoder–decoder model and incorporates self-attention and a position-wise feed-forward network.

Self-attention calculates the similarity and importance among its own data. The input of the transformer is divided into Query Q, Key K, and Value V. Here, Q is the input data

and represents what to search in the input data. K is used to measure how similar the object to be searched and the Q are. V is an element that outputs an appropriate V based on K. These features are transformed in the fully connected layer, and the inner product of K and V is taken. The inner product is then normalized by softmax so that the sum of the weights for a single Q is 1.0. Finally, the output is obtained by multiplying the obtained weight by V. A position-wise feed-forward network is an independent neural network for each data position that consists of two fully connected layers. Located after the attention layer, it linearly transforms the output of the attention layer.

3.7. ViT

Vision transformer (ViT) [33] is a model that divides an image into patches and treats each patch image as a word. Specifically, the image is divided into N patches x_N and passes through the transformer E to obtain one-dimensional E_{x_i}. Let x_{cls} be the classification token and $E_p \in R^{(N+1)\times D}$ be the location information, z_0, which is the input of the transformer, which is prepared from N E_{x_i}, where i is $i = 1, \ldots,$ and N and D is the number of dimensions of the latent vector. This feeds the input z into the transformer. The transformer consists of multi-head self-attention (MSA), layer normalization (LN), and multilayer perceptron (MLP).

3.8. ViViT

Since the video vision transformer (ViViT) [34] constructs the input token z from a video that does not handle 3D data, the method of obtaining the patch x is different from ViT. In ViT, an image is divided into patches and input tokens are obtained, whereas in ViViT, tablets are obtained by collecting patches on the spatio-temporal axis.

In addition, there are encoders with two different roles as a device to capture the time-series information. One is an encoder for capturing the spatial information. It extracts the tokens from the same time frames, interacts, and creates an average classified token x_{cls} through a transformer. The concatenated x_{cls} representing each time is input to the second time-series encoder. Classification is realized using x_{cls} output from the second encoder as a classifier.

3.9. MS-TCN

TCN [35] is a network that uses CNN for series data. It achieves higher accuracy than RNN, such as LSTM, in tasks for time-series data such as natural language and music. TCN consists of a combination of 1D fully convolutional networks and casual convolutions. Furthermore, Martinez et al. proposed a model using MS-TCN [6]. MS-TCN incorporates multiple timescales into the network to mix short-term and long-term information during feature coding.

3.10. Data Augmentation

In the research field of image recognition, data augmentation (DA) is widely used to increase the number of image data by applying operations such as slightly rotating the image or flipping it horizontally. This paper applies RandAugment (RA) [36], which randomly selects the DA method. Various transformations include identity, autocontrast adjustment, histogram equalization, rotation, solarization, color adjustment, posterization, contrast, brightness, sharpness, horizontal shearing, vertical shearing, horizontal translation, vertical translation, and generate N_{RA} images. This paper applies the MixUp [37] with a weight of 0.4. MixUp is a data augmentation technique that generates a weighted combination of random image pairs from the training data.

3.11. Distance Learning

Distance learning is a method of learning a function that maps data to a feature space so that similar data are brought closer to each other and dissimilar data are separated from

each other. This paper applies ArcFace [38], which proposes distance learning using angles, one of the methods with high accuracy in face recognition.

In ArcFace, class classification can solve distance learning by replacing the softmax loss of class classification with angular margin Loss. Softmax Loss has the property of increasing the similarity between samples of the same class but not forcing the similarity of other classes to be low. In ArcFace, the input weights W and feature values x_i are normalized, and the bias b is set to 0. This gives $W_j^T x_i = \|W_j\| \|x_i\| \cos\theta_{j,i} = \cos\theta_{j,i}$. $\theta_{j,i}$ is the angular distance between the feature $x_{j,i}$ and the center position W_j of the j class. Therefore, $\cos\theta_{j,i}$ represents the cosine similarity between the feature x_i and the j class. In addition, the convergence of learning is stabilized by setting the scaling parameter s as a hyperparameter. Furthermore, a linear separation space is secured by directly adding the margin to the angle space.

3.12. Fine-Tuning

Fine-tuning (FT) is used to re-train the weights of the entire model using the weights of the trained network as the initial values to construct a highly accurate model. This paper uses four public datasets: LRW, OuluVS, CUAVE, and SSSD. Among them, LRW is larger than the other datasets. Therefore, FT using the model learned by LRW is applied to the three datasets excluding LRW.

4. Target Models

Referring to the SOTA model [18], this paper investigates six deep learning models shown in Figure 2. The numerical values in the figure indicate the layers that make up the model as one block and indicate the output size of each block. N_F is the number of input sequential image frames, and N_C is the number of classes, which is the number of units in the output layer.

4.1. 3D-Conv + ResNet18 + MS-TCN

An overview of the model diagram is shown in Figure 2a. Extract 512-dimensional features from input images using 3D-Conv + ResNet18. After that, it trains the temporal changes of the features obtained by MS-TCN. MS-TCN has three convolutional layers with kernel sizes of 3, 5, and 7, and obtains short-term and long-term information.

4.2. 3D-Conv + ResNet18 + ViT

As shown in Figure 2b, this model extracts 512-dimensional features from input images using 3D-Conv + ResNet18 and then trains temporal changes using ViT. Normally, the input of ViT is an image, but in this paper, the extracted features are regarded as image patches and input to ViT for training.

4.3. 3D-Conv + WiderResNet18 + MS-TCN

As shown in Figure 3a, this paper uses a model with permuted layers of ResNet. The activation function is changed from ReLU to swish (SiLU). Training is performed in the same way as ResNet18 + MS-TCN. In WideResNet, it is desirable to expand the number of dimensions of features to be extracted. However, this paper uses the same dimensions as ResNet18 due to resource constraints, as shown in Figure 3b.

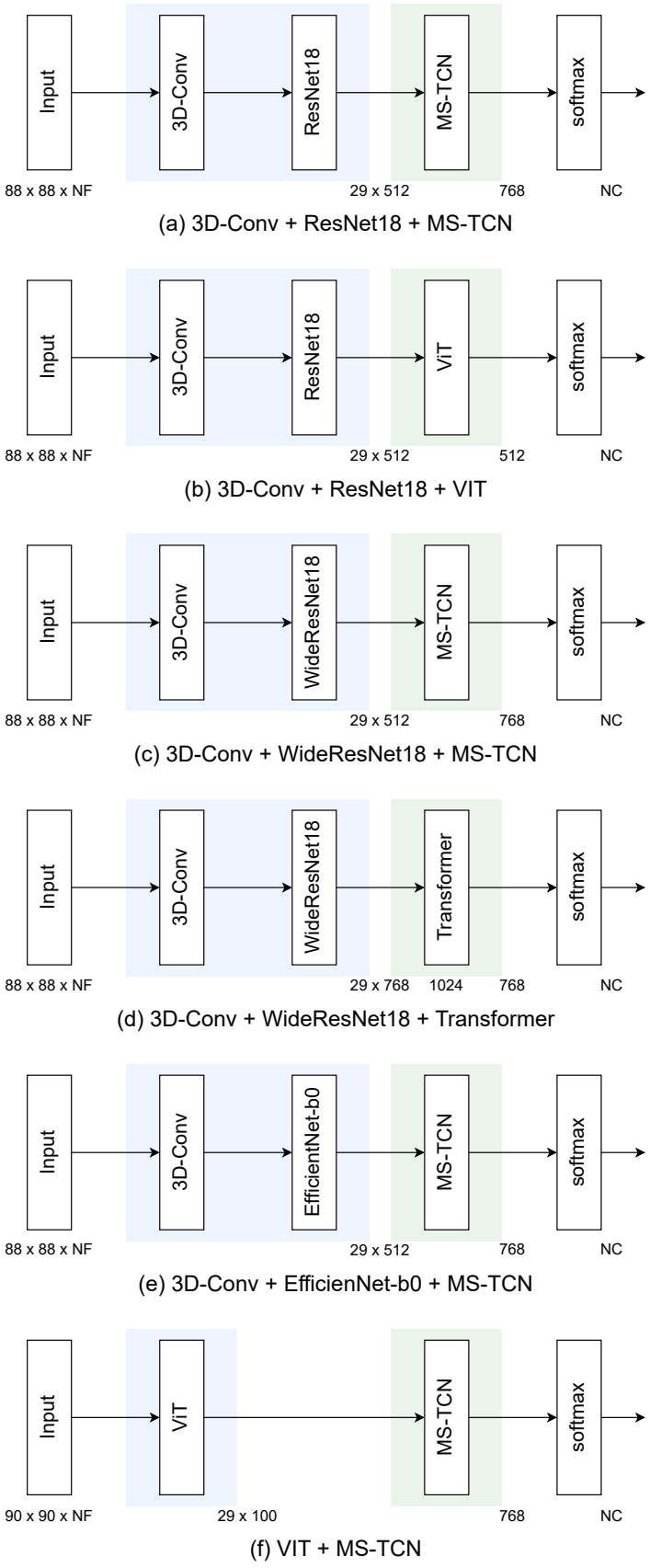

Figure 2. Target models.

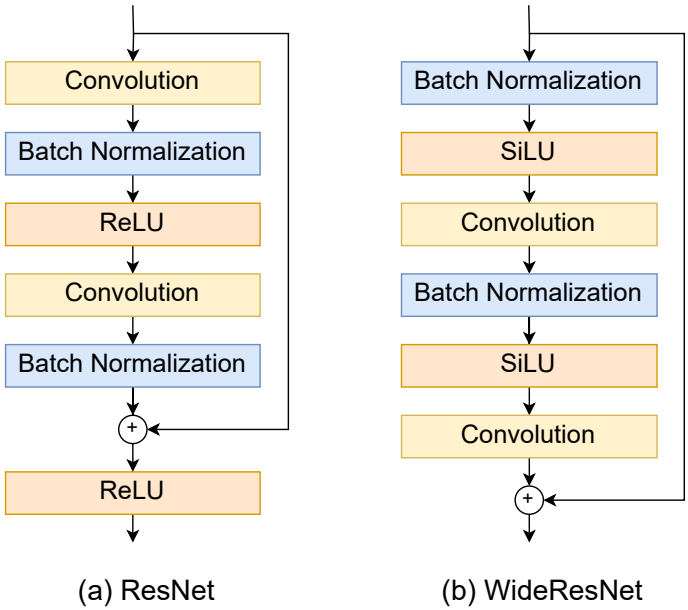

Figure 3. Structures of ResNet and WideResNet.

4.4. 3D-Conv + WiderResNet18 + Transformer

3D-Conv + WideResNet18 extracts 768-dimensional features from input images and then trains temporal changes using transformer. The transformer is originally multi-head self-attention, but single attention is used in this experiment.

4.5. 3D-Conv + EfficientNet-b0 + MS-TCN

3D-Conv + EfficientNet-b0 extracts 512-dimensional features from input images and MS-TCN trains temporal changes.

4.6. ViT + MS-TCN

It is possible to recognize by ViT alone, but in this paper, ViT is used as a feature extractor, as shown in Figure 2f. ViT extracts 100-dimensional features from each frame image, and MS-TCN trains temporal changes. In ViT, a feature vector called a class token inserted at the beginning of each frame image is extracted as a feature value of each frame image.

5. Evaluation Experiment

Several datasets have been published in the lip-reading field. Four public datasets shown in Table 1 are used in this experiment.

Table 1. Overview of the four datasets used in our experiments.

Name	Year	Language	# of Speakers	Content
LRW [4]	2016	English	1000+	500 words
OuluVS [13]	2009	English	20	10 greeting phrases
CUAVE [14]	2002	English	36	10 digits
SSSD [15]	2018	Japanese	72	25 words

We applied the preprocessing described in Section 3.1 to extract the grayscale lipROIs. The image size $S \times S$ of the lipROIs of LRW, OuluVS, CUAVE, and SSSD are 96×96 pixels, 64×64 pixels, 64×64 pixels, and 64×64 pixels, respectively. Figure 4 shows the lipROIs of OuluVS, CUAVE, and SSSD. Inputs to 3D-Conv, ViT, and ViViT are image data randomly extracted from 88×88 pixels, 90×90 pixels, and 87×87 pixels, respectively.

Figure 4. Extracted lipROIs (**left**: OuluVS, **center**: CUAVE, **right**: SSSD).

We used PyTorch to implement each model. To train our network, we utilized Adam with the decoupled weight decay (AdamW) [39] along with certain parameters such as $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and a weight decay of 0.01. The training was conducted for 80 epochs, using an initial learning rate of 0.0003 and a mini-batch size 32 for models using ResNet18 and WideResNet18 and 16 for other models. As for the transformer, single-head attention was used due to insufficient resources. We employed a cosine annealing schedule to decay the learning rate without warm-up steps. We gave $N_{RA} = 2$ as a parameter in RA.

5.1. LRW

Table 2 shows the top-1 accuracy and the number of parameters for each model described in Section 4. In the table, the upper row is the accuracy of other representative papers, and the lower row is the accuracy of this paper. "—" means that the reference does not show the number of parameters in this model. Although the recognition accuracy of this research did not reach the recognition accuracy of SOTA, the following things were clarified by examining various models.

Table 2. Recognition results (LRW).

Model	**Top-1 Acc. (%)**	**Params $\times 10^6$**
Multi-Tower 3D-CNN [4]	61.1	—
WLAS [10]	76.2	—
3D-Conv + ResNet34 + Bi-LSTM [40]	83.0	—
3D-Conv + ResNet34 + Bi-GRU [41]	83.39	—
3D-Conv + ResNet18 + MS-TCN [6]	85.3	—
3D-Conv + ResNet18 + MS-TCN + MVM [20]	88.5	—
3D-Conv + ResNet18 + MS-TCN + KD [8]	88.5	36.4
Alternating ALSOS + ResNet18 + MS-TCN [42]	87.0	41.2
Vosk + MediaPipe + LS + MixUp + SA + 3D-Conv + ResNet-18 + BiLSTM + Cosine WR [22]	88.7	—
3D-Conv + EfficientNetV2 + Transformer + TCN [21]	89.5	—
3D-Conv + ResNet18 + {DC-TCN, MS-TCN, BGRU} (ensemble) + KD + Word Boundary [18]	94.1	—
3D-Conv + ResNet18 + MS-TCN (ours)	87.4	36.0
3D-Conv + ResNet18 + MS-TCN + RA (ours)	85.3	36.0
3D-Conv + ResNet18 + MS-TCN + ArcFace (ours)	86.7	36.0
3D-Conv + ResNet18 + ViT (ours)	83.8	30.1
3D-Conv + WideResNet18 + MS-TCN (ours)	86.8	36.0
3D-Conv + WideResNet18 + Transformer (ours)	79.2	11.2
3D-Conv + EfficientNet-b0 + MS-TCN (ours)	80.6	32.3
ViT + MS-TCN (ours)	79.9	24.0
ViViT (ours)	72.4	3.9
ViViT + RA (ours)	75.6	3.9

As a feature extractor, it can be confirmed that ResNet18 is superior to other models. The analysis of mouth movements, which is the target of our experiment, has a smaller

difference in movements than in other video classification tasks, such as human action recognition. In addition, other feature extractors did not obtain the expected accuracy because the image's resolution was small. The model used in this experiment is a model that achieves high accuracy in image recognition, but these are usually input with high-resolution images such as 224 × 224 pixels. This research is video recognition, and it is necessary to put many images in the memory at once, so the experiment was performed at a low resolution due to memory constraints. Therefore, effective feature values could not be obtained in deep layers.

It was found that MS-TCN tends to obtain recognition accuracy in the inference part of the latter stage. As for MS-TCN, we conducted experiments using RA and ArcFace, but the recognition accuracy decreased. We suspect that this is because the application of the RA-generated image data is unsuitable for the task or the model's generalization performance deteriorated due to the excessive application to the training data.

We consider the details of the recognition results for 3D-Conv + ResNet18 + MS-TCN, which had the highest accuracy among the models investigated in this paper. Among 500 words, 15 words with low recognition rates are listed in Table 3. In the table, the 4th column shows words with many mistakes, and the numbers in parentheses are the misrecognition rate. From the table, it can be confirmed that words with a low recognition rate are misrecognized as similar words. Among the 500 words, 252 had a recognition rate of 90.0% or higher.

Table 3. Fifteen words with low recognition rate (LRW).

Order	Word	Top-1 Acc. (%)	Most Misrecognized Word
486	ABOUT	64	AMONG (10)
	BECAUSE	64	ABUSE (10)
488	ACTUALLY	62	ACTION (6)
	COULD	62	EUROPEAN, SHOULD (4)
	MATTER	62	AMONG (6)
	NEEDS	62	YEARS (8)
	THINGS	62	YEARS (6)
493	THEIR	60	THERE (20)
	UNDER	60	DURING, LONDON (4)
	UNTIL	60	STILL (8)
496	SPEND	58	SPENT (18)
	THESE	58	THINGS (8)
498	THING	56	BEING, NOTHING, THESE (4)
499	THINK	50	THING (16)
500	THERE	44	THEIR (12)

5.2. OuluVS

OuluVS [13] contains ten sentences spoken by 20 speakers, comprising 17 males and 3 females. The contents of the 10 sentences are (1) "excuse me", (2) "good bye", (3) "have a good time", (4) "hello", (5) "how are you", (6) "I am sorry", (7) "nice to meet you", (8) "see you", (9) "thank you", and (10) "you are welcome". For each speaker, five utterance scenes are recorded for each sentence. The image size is 720 × 576 pixels, the frame rate is 25 fps, and the speaker speaks in front of a white background.

The leave-one-person-out cross-validation method was applied to 20 speakers in the evaluation experiment, and the average recognition rate was obtained. Here, the training and test data per speaker are 19 speakers × 10 sentences × 5 scenes = 950 scenes and 1 speaker × 10 sentences × 5 scenes = 50 scenes, respectively. Table 4 shows the recognition rate of the four training conditions and other methods. Here, $N_{RA} = 3$ was set for RA, and the training data were padded. AE, MF, and AU in [5,7,43] stand for feature names based on the auto-encoder, motion feature, and action unit, respectively. FOMM in [7] means a first-order motion model and generates utterance scenes. From the table, 3D-Conv + ResNet18 + MS-TCN + RA + FT obtained the highest recognition accuracy

of 97.2%. It can be confirmed that the recognition accuracy is improved by fine-tuning with LRW.

Table 4. Recognition results (OuluVS).

Model	Top-1 Acc. (%)
Multi-Tower 3D-CNN [4]	91.4
AE + GRU [43]	81.2
FOMM → AE + GRU [7]	86.5
{MF + AE + AU} + GRU [5]	86.6
3D-Conv + ResNet18 + MS-TCN (ours)	90.1
3D-Conv + ResNet18 + MS-TCN + RA (ours)	93.1
3D-Conv + ResNet18 + MS-TCN + FT (ours)	95.1
3D-Conv + ResNet18 + MS-TCN + RA + FT (ours)	97.2

5.3. CUAVE

In CUAVE [14], utterance scenes are taken from 36 speakers, comprising 19 males and 17 females. The utterance contents are "zero", "one", "two", "three", "four", "five", "six", "seven", "eight", and "nine". A feature of CUAVE is that it includes frontal-face speech scenes and side-face speech scenes. Furthermore, not only utterances in which the speaker is standing still but scenes in which the speaker speaks while moving are recorded. It also includes scenes in which two speakers speak at the same time. The image size is 720 × 480 pixels, the frame rate is 29.97 fps, and the speaker speaks in front of a green background. In this experiment, we use a scene where the speaker stands still and speaks five samples per word.

The same leave-one-person-out cross-validation method as OuluVS was applied in the experiment, and the average recognition rate was obtained. The training and test data per speaker are 35 speakers × 10 sentences × 5 scenes = 1750 scenes and 1 speaker × 10 sentences × 5 scenes = 50 scenes, respectively. Table 5 shows the recognition rate of experimental conditions and other methods. However, $N_{RA} = 3$. From the table, 3D-Conv + ResNet18 + MS-TCN + RA + FT obtained the highest recognition accuracy as well as OuluVS.

Table 5. Recognition results (CUAVE).

Model	Top-1 Acc. (%)
AE + GRU [43]	72.8
FOMM → AE + GRU [7]	79.8
{MF + AE + AU} + GRU [5]	83.4
3D-CNN (ours)	84.4
3D-Conv + ResNet18 + MS-TCN (ours)	87.6
3D-Conv + ResNet18 + MS-TCN + RA (ours)	90.0
3D-Conv + ResNet18 + MS-TCN + FT (ours)	93.7
3D-Conv + ResNet18 + MS-TCN + RA + FT (ours)	94.1

5.4. SSSD

SSSD [15] consists of 25 utterances, comprising 10 Japanese numeric words and 15 greetings (https://www.saitoh-lab.com/SSSD/index_en.html, accessed on 26 April 2023). The 25 words are (1) /ze-ro/ (zero), (2) /i-chi/ (one), (3) /ni/ (two), (4) /sa-N/ (three), (5) /yo-N/ (four), (6) /go/ (five), (7) /ro-ku/ (six), (8) /na-na/ (seven), (9) /ha-chi/ (eight), (10) /kyu/ (nine), (11) /a-ri-ga-to-u/ (thank you), (12) /i-i-e/ (no), (13) /o-ha-yo-u/ (good morning), (14) /o-me-de-to-u/ (congratulation), (15) /o-ya-su-mi/ (good night), (16) /go-me-N-na-sa-i/ (I am sorry), (17) /ko-N-ni-chi-wa/ (good afternoon), (18) /ko-N-ba-N-wa/ (good evening), (19) /sa-yo-u-na-ra/ (goodbye), (20) /su-mi-ma-se-N/ (excuse me), (21) /do-u-i-ta-shi-ma-shi-te/ (you are welcome), (22) /ha-i/ (yes), (23) /ha-ji-me-ma-shi-te/ (nice to meet you), (24) /ma-ta-ne/ (see you), and (25) /mo-shi-mo-shi/ (hello).

Unlike OuluVS, CUAVE, and LRW, SSSD is filmed using a smart device. An image of the lower half of a face of 300 × 300 pixels extracted after normalization processing is applied for scale, and rotation is provided. The frame rate is 30 fps. The number of provided scenes is 72 speakers × 25 words × 10 samples = 18,000 scenes. As a competition using SSSD, the second machine lip-reading challenge was held in 2019, and 5000 scenes of test data were released.

For the accuracy evaluation, we used 18,000 scenes as training data and an extra 5000 scenes as test data, using the same task as the second machine lip-reading challenge to obtain recognition accuracy. The results are shown in Table 6. N_{RA} gave 3 like OuluVS and CUAVE. From the table, 3D-Conv + ResNet18 + MS-TCN + FT obtained the highest recognition accuracy of 95.14%. While OuluVS and CUAVE obtained high recognition accuracy when RA was applied, SSSD obtained the highest recognition rate when RA was not applied. We presume that SSSD has more training data than OuluVS and CUAVE and can train sufficiently without applying RA.

Table 6. Recognition results (SSSD).

Model	Top-1 Acc. (%)
LipNet	90.66
3D-Conv + ResNet18 + MS-TCN (ours)	93.08
3D-Conv + ResNet18 + MS-TCN + RA (ours)	93.68
3D-Conv + ResNet18 + MS-TCN + FT (ours)	95.14
3D-Conv + ResNet18 + MS-TCN + RA + FT (ours)	94.86

6. Conclusions

We conducted a study on word lip-reading using deep-learning models. Our goal was to find an effective model for this task. We explored different combinations of models such as ResNet, WideResNet, EfficientNet, Transformer, and ViT, referring to the SOTA model. While many papers use one or two datasets, recognition experiments were conducted using four public datasets, namely LRW, OuluVS, CUAVE, and SSSD, with different sizes and languages. As a result, we found that 3D-Conv + ResNet18 is a good model for feature extraction, and MS-TCN is a good model for inference. Although we did not propose a model that surpasses SOTA, our study confirmed the effectiveness of these models.

This paper investigates an effective word lip-reading model on four public datasets. There are other lip-reading datasets not used in this paper. In the future, we will work on experiments including other datasets. Since it has been clarified that the model structure is effective for lip-reading, we will also verify the training method of the model in the future. The recognition target of this paper is words, but sentence lip-reading has also been actively researched in recent years. Sentence lip-reading is also a target task for the future.

Author Contributions: Conceptualization, T.A. and T.S.; methodology, T.A. and T.S.; software, T.A.; validation, T.A.; formal analysis, T.A. and T.S.; investigation, T.A. and T.S.; resources, T.S.; data curation, T.A.; writing—original draft preparation, T.S.; writing—review and editing, T.S.; visualization, T.S.; supervision, T.S.; project administration, T.S.; funding acquisition, T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the JSPS KAKENHI Grant No. 19KT0029.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this article come from [4,13–15].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AE	Auto-Encoder
AU	Action Unit
BBC	British Broadcasting Corporation
BGRU	Bidirectional Gated Recurrent Unit
CNN	Convolutional Neural Network
DA	Data Augmentation
DC-TCN	Densely Connected Temporal Convolutional Network
FOMM	First-Order Motion Model
FT	Fine-Tuning
HOG	Histograms of Oriented Gradient
LN	Layer Normalization
LRW	Lip Reading in the Wild
LS	Label Smoothing
LSTM	Long Short-Term Memory
MF	Motion Feature
MLP	Multilayer Perceptron
MSA	Multi-head Self-Attention
MS-TCN	Multi-Scale Temporal Convolutional Network
NAS	Neural Architecture Search
RA	RandAugment
ResNet	Residual Network
RNN	Recurrent Neural Network
ROI	Region of Interest
RUSAVIC	Russian Audio-Visual Speech in Cars
SA	Squeeze-and-Attention
SiLU	Swish
SSSD	Speech Scene by Smart Device
SOTA	State of the Art
ViT	Vision Transformer
ViViT	Video Vision Transformer
WLAS	Watch, Listen, Attend, and Spell

References

1. Saitoh, T.; Konishi, R. Profile Lip Reading for Vowel and Word Recognition. In Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, 23–26 August 2010. [CrossRef]
2. Nakamura, Y.; Saitoh, T.; Itoh, K. 3D CNN-based mouth shape recognition for patient with intractable neurological diseases. In Proceedings of the 13th International Conference on Graphics and Image Processing (ICGIP 2021), Kunming, China, 18–20 August 2022; Volume 12083, pp. 775–782. [CrossRef]
3. Kanamaru, T.; Arakane, T.; Saitoh, T. Isolated single sound lip-reading using a frame-based camera and event-based camera. *Front. Artif. Intell.* **2023**, *5*, 298. [CrossRef] [PubMed]
4. Chung, J.S.; Zisserman, A. Lip Reading in the Wild. In Proceedings of the Asian Conference on Computer Vision (ACCV), Taipei, Taiwan, 20–24 November 2016.
5. Shirakata, T.; Saitoh, T. Lip Reading using Facial Expression Features. *Int. J. Comput. Vis. Signal Process.* **2020**, *10*, 9–15.
6. Martinez, B.; Ma, P.; Petridis, S.; Pantic, M. Lipreading using temporal convolutional networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6319–6323. [CrossRef]
7. Kodama, M.; Saitoh, T. Replacing speaker-independent recognition task with speaker-dependent task for lip-reading using First Order Motion Model. In Proceedings of the 13th International Conference on Graphics and Image Processing (ICGIP 2021), Kunming, China, 18–20 August 2022; Volume 12083, pp. 652–659. [CrossRef]
8. Ma, P.; Martinez, B.; Petridis, S.; Pantic, M. Towards Practical Lipreading with Distilled and Efficient Models. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7608–7612. [CrossRef]
9. Fu, Y.; Lu, Y.; Ni, R. Chinese Lip-Reading Research Based on ShuffleNet and CBAM. *Appl. Sci.* **2023**, *13*, 1106. [CrossRef]
10. Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Lip Reading Sentences in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017, pp. 6447–6456. [CrossRef]

11. Arakane, T.; Saitoh, T.; Chiba, R.; Morise, M.; Oda, Y. Conformer-Based Lip-Reading for Japanese Sentence. In Proceedings of the 37th International Conference on Image and Vision Computing, Auckland, New Zealand, 24–25 November 2023; pp. 474–485. [CrossRef]
12. Jeon, S.; Elsharkawy, A.; Kim, M.S. Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition. *Sensors* **2022**, *22*, 72. [CrossRef]
13. Zhao, G.; Barnard, M.; Pietikainen, M. Lipreading with local spatiotemporal descriptors. *IEEE Trans. Multimed.* **2009**, *11*, 1254–1265. [CrossRef]
14. Patterson, E.K.; Gurbuz, S.; Tufekci, Z.; Gowdy, J.N. Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus. *EURASIP J. Appl. Signal Process.* **2002**, *2002*, 1189–1201. [CrossRef]
15. Saitoh, T.; Kubokawa, M. SSSD: Speech Scene Database by Smart Device for Visual Speech Recognition. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR2018), Beijing, China, 20–24 August 2018; pp. 3228–3232. [CrossRef]
16. Yang, S.; Zhang, Y.; Feng, D.; Yang, M.; Wang, C.; Xiao, J.; Long, K.; Shan, S.; Chen, X. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG2019), Lille, France, 14–18 May 2019. [CrossRef]
17. Ivanko, D.; Ryumin, D.; Axyonov, A.; Kashevnik, A.; Karpov, A. Multi-Speaker Audio-Visual Corpus RUSAVIC: Russian Audio-Visual Speech in Cars. In Proceedings of the 13th Conference on Language Resources and Evaluation (LREC2022), Marseille, France, 21–23 June 2022; pp. 1555–1559.
18. Ma, P.; Wang, Y.; Petridis, S.; Shen, J.; Pantic, M. Training Strategies for Improved Lip-reading. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022. [CrossRef]
19. Feng, D.; Yang, S.; Shan, S.; Chen, X. Learn an Effective Lip Reading Model without Pains. *arXiv* **2020**, arXiv:2011.07557. [CrossRef]
20. Kim, M.; Yeo, J.H.; Ro, Y.M. Distinguishing Homophenes using Multi-head Visual-audio Memory for Lip Reading. In Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI), Virtual, 22 February–1 March 2022.
21. Koumparoulis, A.; Potamianos, G. Accurate and Resource-Efficient Lipreading with Efficientnetv2 and Transformers. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022. [CrossRef]
22. Ivanko, D.; Ryumin, D.; Kashevnik, A.; Axyonov, A.; Karnov, A. Visual Speech Recognition in a Driver Assistance System. In Proceedings of the 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022. [CrossRef]
23. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
24. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Kauai, HI, USA, 8–14 December 2001; Volume 1. [CrossRef]
25. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893. [CrossRef]
26. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 5203–5212.
27. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1867–1874. [CrossRef]
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778. [CrossRef]
29. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016; pp. 87.1–87.12. [CrossRef]
30. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning (ICMR), Long Beach, CA, USA, 9–15 June 2019.
31. Zoph, B.; Le, Q.V. Neural architecture search with reinforcement learning. In Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017. [CrossRef]
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS2017), Long Beach, CA, USA, 4–9 December 2017.
33. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
34. Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; Schmid, C. ViViT: A Video Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 6816–6826. [CrossRef]

35. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271. [CrossRef]
36. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 14–19 June 2020; Volume 33, pp. 18613–18624.
37. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018. [CrossRef]
38. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 4685–4694. [CrossRef]
39. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019. [CrossRef]
40. Stafylakis, T.; Tzimiropoulos, G. Combining Residual Networks with LSTMs for Lipreading. In Proceedings of the Conference of the International Speech Communication Association (Interspeech 2017), Stockholm, Sweden, 20–24 August 2018; pp. 3652–3656.
41. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end Audiovisual Speech Recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6548–6552.
42. Tsourounis, D.; Kastaniotis, D.; Fotopoulos, S. Lip Reading by Alternating between Spatiotemporal and Spatial Convolutions. *J. Imaging* **2021**, *7*, 91. [CrossRef] [PubMed]
43. Iwasaki, M.; Kubokawa, M.; Saitoh, T. Two Features Combination with Gated Recurrent Unit for Visual Speech Recognition. In Proceedings of the 14th IAPR Conference on Machine Vision Applications (MVA), Nagoya, Japan, 8–12 May 2017; pp. 300–303. [CrossRef]