mathematics

MDPI

Article

Research on Robust Audio-Visual Speech Recognition Algorithms

Wenfeng Yang *, Pengyi Li, Wei Yang, Yuxing Liu, Yulong He, Ovanes Petrosian and Aleksandr Davydenko

Faculty of Applied Mathematics and Control Processes, Saint Petersburg State University, 198504 Saint Petersburg, Russia
* Correspondence: st098651@student.spbu.ru

Abstract: Automatic speech recognition (ASR) that relies on audio input suffers from significant degradation in noisy conditions and is particularly vulnerable to speech interference. However, video recordings of speech capture both visual and audio signals, providing a potent source of information for training speech models. Audiovisual speech recognition (AVSR) systems enhance the robustness of ASR by incorporating visual information from lip movements and associated sound production in addition to the auditory input. There are many audiovisual speech recognition models and systems for speech transcription, but most of them have been tested based in a single experimental setting and with a limited dataset. However, a good model should be applicable to any scenario. Our main contributions are: (i) Reproducing the three best-performing audiovisual speech recognition models in the current AVSR research area using the most famous audiovisual databases, LSR2 (Lip Reading Sentences 2) LSR3 (Lip Reading Sentences 3), and comparing and analyzing their performances under various noise conditions. (ii) Based on our experimental and research experiences, we analyzed the problems currently encountered in the AVSR domain, which are summarized as the feature-extraction problem and the domain-generalization problem. (iii) According to the experimental results, the Moco (momentum contrast) + word2vec (word to vector) model has the best AVSR effect on the LRS datasets regardless of whether there is noise or not. Additionally, the model also produced the best experimental results in the experiments of audio recognition and video recognition. Our research lays the foundation for further improving the performance of AVSR models.

Keywords: multi-model deep learning; speech recognition; lip reading; audiovisual speech recognition; model comparison; MOCO

MSC: 68T01

Citation: Yang, W.; Li, P.; Yang, W.; Liu, Y.; He, Y.; Petrosian, O.; Davydenko, A. Research on Robust Audio-Visual Speech Recognition Algorithms. *Mathematics* **2023**, *11*, 1733. https://doi.org/10.3390/math11071733

Academic Editor: Catalin Stoean

Received: 3 March 2023
Revised: 31 March 2023
Accepted: 3 April 2023
Published: 5 April 2023

1. Introduction

With the advent of artificial intelligence and deep learning, automatic speech recognition (ASR) systems have made significant progress in tasks set in controlled environments. However, current automatic speech recognition systems have certain requirements for obtaining high recognition rates, such as test data being homogeneous with training data, a relatively quiet recording environment, a normal speech rate, and reading aloud. To compensate for the shortcomings of automatic speech recognition (ASR) in noisy situations and based on the human bimodal speech perception mechanism, some researchers have proposed a new approach to improving the robustness of ASR systems: combining visual signals with audio signals. Speech recognition based on audiovisual signals is called audio-visual speech recognition (AVSR). AVSR technique provides a good idea for the purpose of "natural language communication between human and machine" by simulating the human bimodal speech perception process based on visual information, such as lip movements. In addition, this technique can be combined with traditional robust speech-recognition techniques to further improve the performances of speech recognition systems.

Based on the multimodal complementary nature of AVSR, the AVSR model has a wide range of applications. For example, it can be applied in conference recording systems to enhance the transcription performance of conference recording systems through the recognition of lip movements by cameras; it can also be applied in command recognition in land vehicles [1], cell-phone text translation [2–4], lip reading for hearing-impaired people [5], speech recognition of individual speakers who speak at once [6], and so on.

In order to obtain a better-performing AVSR model, researchers have tried to make the model see as much as possible during the training phase. Therefore, a good dataset is very important. LRW (Lip Reading in the Wild), LRS2, and LRS3 are audio-visual speech recognition datasets collected from in-the-wild videos [7–9].

In this article, we selected the three best-performing models for reproduction based on the audio-visual speech recognition on LRS3-TED leaderboard and the audio-visual speech recognition on LRS2 leaderboard. For the recurring models, we designed a series of experiments for comparison, and analyzed the advantages and disadvantages of each model based on the comparison results, providing effective support for later work.

Our work is described in detail in the following section. Section 2 presents AVSR and Lip reading research results and related work are presented. Section 3 presents the detailed definition of the AVSR task, the dataset description, and the experimental conditions. Section 4 details the detailed rationale for the three models compared in this paper. Section 5 presents the experiments related to the experimental results obtained and analyzes the advantages and disadvantages of the models based on the results. Section 6 summarizes all the work described in this paper and presents an outlook for the future.

2. Related Work

In 2018, Stavros Petridis et al. [10] proposed an end-to-end audiovisual model based on residual networks and bi-directional gated recurrent units (BGRU). It learns to extract features from both image pixels and audio waveforms, while also being able to identify words in context using a publicly available dataset (LRW). Its accuracy reached 83.39 percent. The first application of a transformer to the AVSR domain was made by Triantafyllos Afouras et al. in 2018 [11]. Two lip reading models were compared. One using CTC (connectionist temporal classification) loss, and another using sequence-to-sequence loss, were constructed on the transformer self-attentive architecture. A new dataset, LRS2-BBC, which was created for audiovisual speech recognition and includes thousands of natural sentences from British television, was also constructed and released. Its model greatly exceeds the performance of all previous work on the lip-reading benchmark dataset. The LF-MIMI (Lip-reading for Multiple-Speaker Identification) time-delayed neural network (TDNN) word error rate (WER) absolute reduction proposed by Jianwei Yu et al. in 2020 [12] outperformed the audio-only baseline, LF-MMI DNN, by up to 29.98 percent. The absolute performance of WER reduction was improved by 4.89 percent compared to the baseline AVSR system using feature fusion [13]. A ResNet-18 and convolutional enhancement converter (conformer)-based hybrid CTC/attention model has been suggested. This model can be trained in an end-to-end manner. Audio and visual coders are capable of extracting features from original pixels and audio waveforms, respectively. These features are then fed to the conformer, which is fused using a multilayer perceptron (MLP). The model uses a combination of CTC and attention mechanisms to learn to recognize characters. The model achieved a good error rate of 3.7 percent of words. In 2022, Bowen Shi et al. [14] proposed a self-supervised AVSR framework based on audiovisual HuBERT (AV-HuBERT), a state-of-the-art model for learning audiovisual speech representations. The model achieved an error rate of 1.4 percent and ranked first in performance on the largest available AVSR benchmark dataset, LRS3. In 2022, Xichen Pan et al. [15] effectively implemented unimodal self-supervised learning to support multimodal AVSR. They trained audio and visual front-ends on a single-peaked dataset and then combined these components into a larger multimodal framework. This framework can recognize parallel audiovisual data as characters by adapting a combination of CTC and seq2seq decoding. The results indicate

that the two inherited components from unimodal self-supervised learning function well together and can produce impressive outcomes by fine-tuning the multimodal framework. The model achieved an excellent error word rate of 2.6 percent on the LRS2 dataset. By far, the model performed better than others on the LRS2 dataset.

Lip reading is one of the most popular methods for visual speech recognition. The main implementation method is to extract the lips of the speaker in the video by frame, arrange the frames in temporal order, and input them into an artificial neural network. Many researchers have been trying various methods to increase the accuracy of lip reading. In 2017, Themos Stafylakis et al. [16] created a deep learning architecture that functions end-to-end and is intended for visual speech recognition at the word level. It is a combination of spatio-temporal convolution, residuals, and bidirectional long and short-term memory networks. The system achieved an accuracy of 83 percent. In 2020, Peratham W. [17] proposed a novel deep learning architecture called SpotFast for lip reading at the word level. SpotFast is a modified version of the advanced SlowFast network designed for action recognition. It utilizes time windows as point paths and fast paths that include all frames. In combination with the memory-enhanced transversal transformer, SpotFast's accuracy was improved by 3.7 percent. The final score was 84.4 percent. In 2022, Dalu Feng et al. [18] obtained a good score of 88.4 percent using 3D-ResNet and Bi-GRU networks for modeling, along with MixUp for data enhancement. The same year, Pingchuan Ma et al. [19] stated that a sequence of research studies have proven that temporal masking (TM) is the most crucial method for enhancing data, followed by MixUp. Meanwhile, the densely connected temporal convolutional network (DC-TCN) is the most effective model for isolated word lip reading. Self distillation and word boundary indicators also contribute to an improvement in performance, albeit to a lesser extent. Using all the above-mentioned methods together resulted in a classification accuracy of 93.4 percent, improving the current state-of-the-art performance on the LRW dataset by 4.6 percent.

There are six SOTA approaches for lip-reading recognition on the LRW dataset that have been able to achieve more than 88% accuracy [20]. The first model is Vosk + MediaPipe + LS + MixUp + SA + 3DResNet-18 + BiLSTM + Cosine WR, which is able to achieve 88.7% accuracy, In 2022, Koumparoulis et al. [21] showed that 3D Conv + EfficientNetV2 + Transformer + TCN can reach 89.52% accuracy, and in the same year, Pingchuan Ma et al. [19] reported that 3D Conv + ResNet-18 + DC-TCN + KD can achieve the best current accuracy of 94.1%. One SOTA method has only two models and achieved a less than 20% error rate on the LRS2 dataset. In 2022, Haliassos et al. [22] proposed RAVEn Large WER, which had an error rate of 18.6%. In 2023, Pingchuan Ma et al. [23] proposed CTC/attention WER, which was able to reach a current-best error rate of 14.6%. Haliassos et al. [22] showed that their WER of RAVEn Large model could reach an error rate of 23.4%, and in the same year, Pingchuan Ma et al. [23] showed that the WER of CTC/attention model could reach a current-best error rate of 19.1%.

3. Audiovisual Speech Recognition

3.1. Problem Statement

AVSR technology provides a good idea for realizing "human–machine natural language communication" by simulating the human bimodal speech perception process based on visual information such as lip movement. The model captures feature information simultaneously through visual signals and audio signals, and recognizes the speaker's speech content based on this feature information. The detailed mathematical definition of the problem is given below: $D = \{v_i, y_i\}, i = 1, 2, \ldots, n$ represents a data set with n samples, where v_i represents the ith video sample and y_i represents the label corresponding to the video sample.

After the original data are preprocessed, the original video sample is divided into audio samples and image samples according to a certain strategy: $Audio_i, Image_i := Pre(v_i)$. In the fusion stage of the multimodal model, for the early fusion strategy, the audio samples and image samples are first combined and then fed to the feature extractor for feature

extraction $f_{ai} := Encoder_{ai}(Concat(Audio_i, Image_i))$. For the late fusion strategy, first, the audio samples and image samples are fed to the corresponding encoder to obtain the corresponding feature f—$f_{audio} = Encoder_{audio}(Audio_i)$, $f_{image} = Encoder_{image}(Image_i)$—and then we merge the two features: $f_{ai} = Concat(f_{audio}, f_{image})$. After feature fusion, the fused feature f_{ai} is fed into the *Predictor* to realize speech recognition: $Result := Predictor(f_{ai})$.

3.2. Description of Datasets

The audio-visual speech recognition datasets, LRW [7], LRS2 [8], and LRS3 [9], have been gathered from videos recorded in natural settings.

3.2.1. Lip Reading in the Wild (LRW)

The dataset is composed of 1000 utterances of 500 unique words, spoken by several speakers. Each video in the dataset is 29 frames long, which is equivalent to 1.16 s. The word is spoken in the middle of the video. The metadata contain the duration of the word, providing information on the start and end frames of the word. The dataset statistics are given in Table 1.

Table 1. LRW.

Set	Dates	Class	per Class
Train	01/01/2010–31/08/2015	500	800–1000
Validation	01/09/2015–24/12/2015	500	50
Test	01/01/2016–30/09/2016	500	50

3.2.2. Lip Reading Sentences 2 (LRS2)

The BBC television dataset comprises numerous spoken sentences. Each sentence has a length of up to 100 characters. The training, validation, and test sets are categorized based on the date of the broadcast. The dataset's details are outlined in the provided table.

The LRS2-BBC dataset is partitioned into two groups: the development set, which includes the train and validation sets based on the broadcast date, and the test set. This dataset also possesses a "pre-train" subset, containing sentence excerpts that may differ in length from those included in the development set. These excerpts are marked with the alignment boundaries of each word. The pre-training set comprises partial sentences and multiple sentences, and the training set only includes complete single sentences or phrases. Additionally, there is some intersection between the pre-training and training sets.

The LRS2 dataset is made up of 144,482 video clips sourced from multiple BBC programs, totaling up to 224.1 h of content. Within the dataset, there are four distinct groups of utterances. The pre-training group contains 96,318 utterances, which covers 195 h of content. The training group contains 45,839 utterances, which spans 28 h of content. The validation group comprises 1082 utterances and takes up 0.6 h of content, whereas the testing group contains 1243 utterances and spans 0.5 h of content.

It is possible that there could be some inaccuracies in the labeling of both the pre-training and training sets. However, the test set has been thoroughly inspected and confirmed to be accurate, based on our current understanding. As a result, we believe that the test set does not contain any mistakes. The dataset's statistics are given in Table 2.

Table 2. LRS2.

Set	Dates	Utterances	Word Instances	Vocab
Pre-train	11/2010–06/2016	96,318	2,064,118	41,427
Train	11/2010–06/2016	45,839	329,180	17,660
Validation	06/2016–09/2016	1082	7866	1984
Test	09/2016–03/2017	1243	6663	1698

3.2.3. Lip Reading Sentences 3 (LRS3)

This dataset comprises over 400 h of video content, taken from 5594 TED and TEDx talks in English, which were downloaded from YouTube. The dataset contains cropped face tracks in MP4 format, which have a resolution of 224 × 224 and are encoded with the h264 codec at a frame rate of 25 fps. Audio tracks are also included in a single-channel 16-bit 16 kHz format. Furthermore, the dataset includes plain text files containing the corresponding text transcripts of every word and alignment boundaries. The dataset is sorted into three categories: pre-train, train-val, and test. The first two categories share some content, and the last category is entirely separate. The dataset statistics are given in Table 3. The LRS3 dataset is twice the size of LRS2, and it contains 151,819 utterances, equaling 438.9 h of content. Precisely, the pre-training group comprises 118,516 utterances accounting for 408 h, and the training–validation group consists of 31,982 utterances for 30 h of content. Lastly, the test group contains 1321 utterances and spans 0.9 h of content.

Table 3. LRS3.

Set	Dates	Utterances	Word Instances	Vocab
Pre-train	11/2010–06/2016	96,318	2,064,118	41,427
Train	11/2010–06/2016	45,839	329,180	17,660
Validation	06/2016–09/2016	1082	7866	1984
Test	09/2016–03/2017	1243	6663	1698

4. Models Description

We selected the three best-performing models for replication from the audiovisual speech recognition ranking of LRS3-TED and the audiovisual speech recognition ranking of LRS2. They are Transformer-CTC from [11], Av-HuBERT from [14], and Moco-word2vec from [15].

4.1. Transformer-CTC

In this section, a variant of the transformer-based model is described: it consists of two parts, the transformer [24] and the CTC [25]. Its structure is outlined in Figure 1. This model receives two input parameters, one for video (V) and one for audio (A). The specific model implementation is shown in Figure 2.

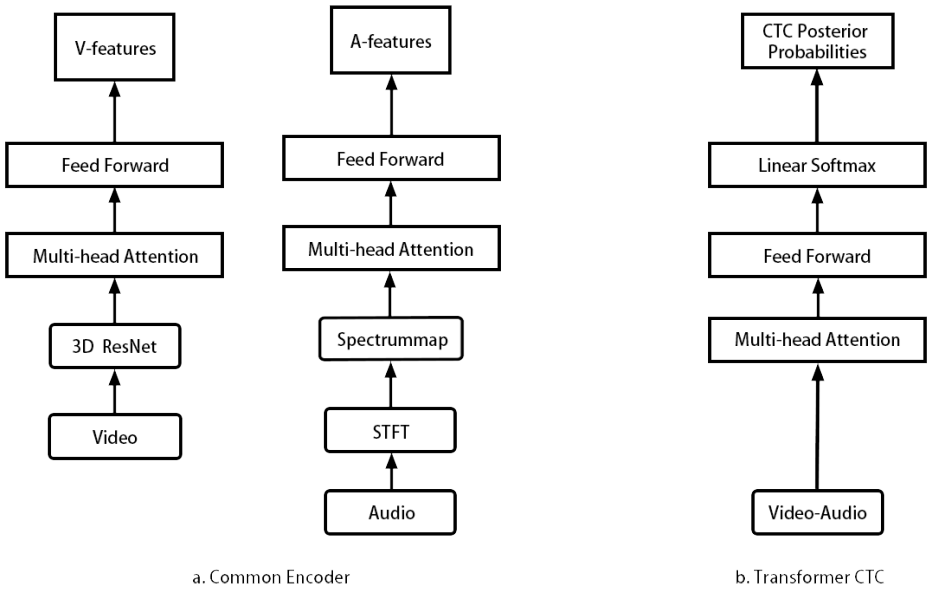

Figure 1. Audio-visual speech recognition models.

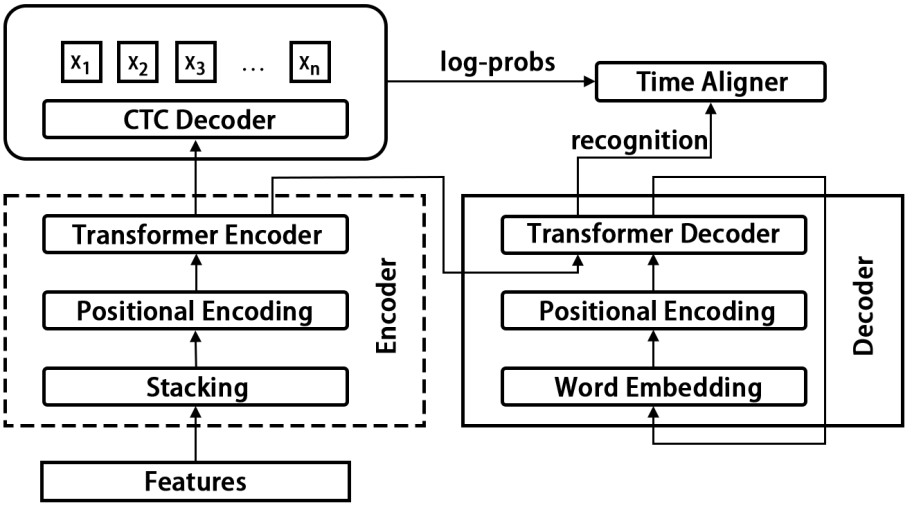

Figure 2. Audio-visual speech recognition process.

4.1.1. Model Architecture

Transformer-CTC, a model consisting of a stack of self-attentive and feedforward layers, generates CTC posterior probabilities for each input frame. CTC: uses recurrent neural networks to label unsegmented sequence data.

The TM-CTC model connects the video and audio encoding and propagates the results through a number of self-attentive/feedforward blocks, the same as those used in the encoder. The output of the network is the posterior probability of the CTC of each input frame, and the entire stack is trained using CTC losses.

4.1.2. Transformer

The paper "Attention Is All You Need" introduced the transformer architecture [25]. The definition of a transformer is given in the paper as "Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution". In the transformer architecture, the conventional CNN and RNN models are abandoned, and the network's structure is exclusively made up of an attention mechanism. To be specific, the transformer solely comprises self-attention and feed-forward neural network modules. By chance, the transformer also employs this design, which involves utilizing multiple layers of self-attention- and dot-product-based fully connected layers in both the encoder and decoder. This is illustrated in the left and right parts of Figure 3, respectively. Transformer's self-attention mechanism allows the attention of the neural network to be focused on the important parts.

1. Attention

Attention is the core part of the transformer. It takes all the global features into consideration, and the weight share of each global part in a certain part is obtained by the calculation of three matrices of queries, keys, and values in it.

As shown in Equation (1). A set of queries is loaded into matrix Q, and the keys and values are loaded into matrices K and V. d_k represents the dimensions of the keys. The effect of dividing by $\sqrt{d_k}$ is to attenuate the effect of the vector dimensionality on the resultant values.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

2. Multi-Head Attention

To make the model's attention more objective and fair, the transformer also provides a multi-headed attention mechanism. The principle of multi-headed attention mechanism

is to initialize multiple weights at the same time, calculate multiple attention results, and finally stitch them together in sequence to get the final result.

$$\begin{aligned} MultiHead(Q,K,V) &= Concat(head1,\ldots,headh)W^O \\ where\ head_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

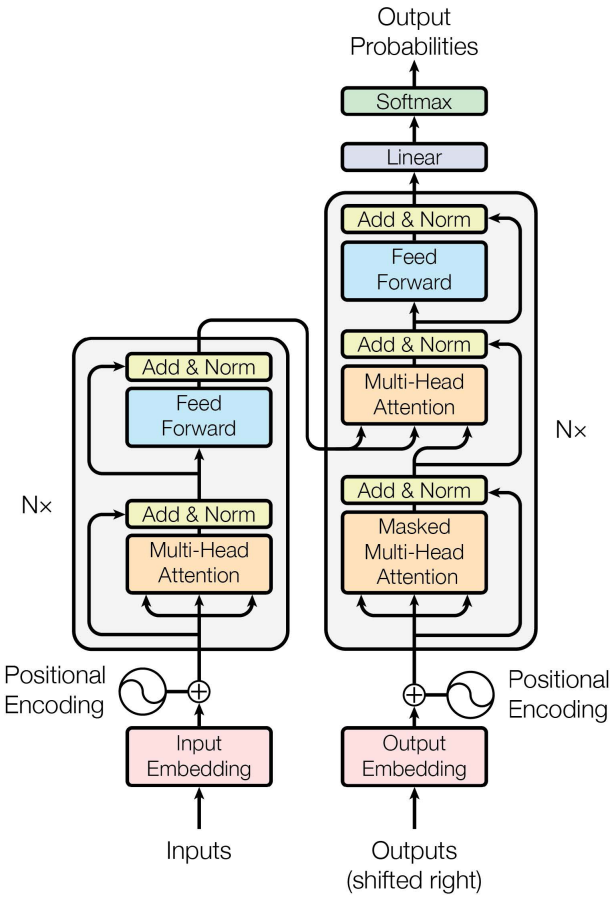

Figure 3. The transformer model's architecture.

4.1.3. Connectionist Temporal Classification

The CTC loss function will serve as the guide for the training of the neural network (CNN). We will only input the output matrix of the CNN and the corresponding ground truth (GT) text to the CTC loss function. However, one might wonder how it recognizes the location of each character. In reality, it does not have this information. Instead, it tests every possible sequence of the GT text in the image and calculates the total score for each permutation. As a result, if the sum of the alignment scores for a GT text is high, then that particular text has a high score.

4.2. Audio-Visual Hidden Unit BERT

4.2.1. Model Architecture

Audio-visual hidden unit BERT (AV-HuBERT) is a multimodal, selfsupervised speech-representation learning framework. It encodes masked audio and image sequences into audio-visual features via a hybrid ResNet-transformer architecture to make a forecast for a set of predetermined categories in a specific order. The target cluster assignments are initially generated from signal processing-based acoustic features and iteratively refined using the features learned by the audio-visual encoder via k-means clustering. AV-HuBERT captures both linguistic and phonetic information from lip movements and audio streams in unmasked regions, encoding their long-term temporal relationships to solve the masked-prediction task. Given a pre-trained AV-HuBERT model, we keep both its audio and video

fronts during fine-tuning. We use a sequence-to-sequence model for AVSR, where AV-HuBERT serves as the encoder module. In contrast to pretraining, we do not apply input masking or modality dropout during finetuning. Additionally, we froze the pretrained AV-HuBERT encoder for a certain number of training steps, after which we updated all model weights.

4.2.2. BERT

Definition

Google AI Research introduced BERT (bidirectional encoder representation from transformers), a pretrained model in October 2018, which performed amazingly well on the top-level machine reading comprehension test SQuAD V1.1. (This is a dataset that has been created for the purpose of question answering and reading comprehension. The dataset is composed of a collection of Wikipedia articles.) It outperformed humans across the board by both measures and set SOTA performance for 11 different NLP tests, including improving the GLUE benchmark to 80.4% (an absolute improvement of 7.6%) and achieving 86.7% (an absolute improvement of 5.6%) in MultiNLI precision, making it a landmark model in the history of NLP development.

BERT Framework

The network architecture of BERT uses the multilayer transformer structure proposed in Attention is All You Need, as shown in Figure 4. The most prominent characteristic of the transformer's structure is abandoning the conventional RNN and CNN models and utilizing the attention mechanism to transform the distance between two words at any position into a value of one. This approach efficiently resolves the complicated long-term dependency issue in NLP, and as a result, the transformer structure has been widely used in the field of NLP. The BERT framework can be broken down into two stages: pre-training and fine-tuning. In the pre-training phase, models are trained on unlabeled labeled data. In the fine-tuning phase, BERT models are first initialized by the pre-trained model's parameters, and then all parameters are trained with downstream labeled data.

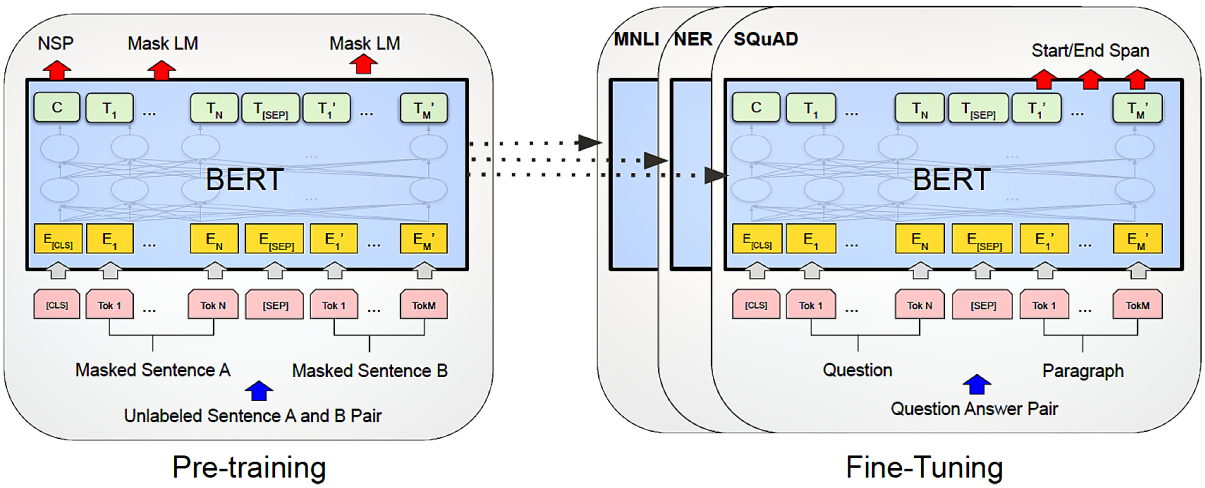

Figure 4. BERT structure.

4.2.3. HuBERT

Definition

HuBERT uses clustering to provide labels for the bosses used in BERT, and then a BERT-like mask-style boss allows the model to train the acoustic and linguistic models with the continuous speech data. The structure is shown in Figure 5.

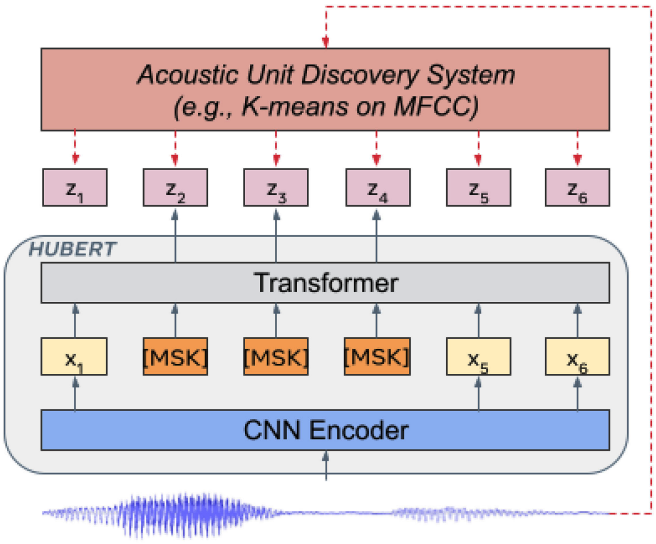

Figure 5. HuBERT structure.

4.2.4. Main Improvements

The architecture of the audio-visual HuBERT model is show as in Figure 6. It undergoes iterative training by alternating between predicting and feature clustering, in a manner similar to Visual HuBERT. However, it has four notable enhancements:

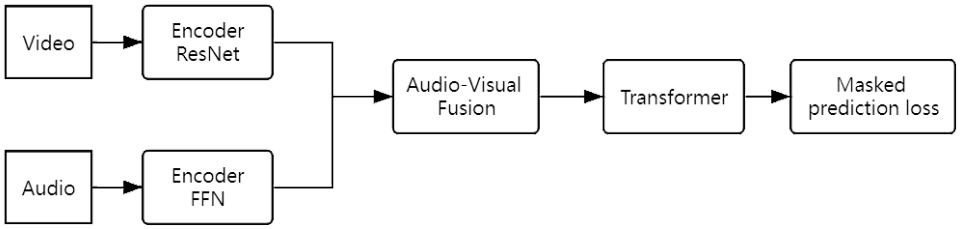

Figure 6. Illustration of AV-HuBERT.

1. Audio-visual input

The AV-HuBERT model is designed to incorporate both acoustic and image frames during the masked-prediction training stage. This approach facilitates better modeling and enables the distillation of correlations between the two modalities. Specifically, the AV-HuBERT model employs lightweight encoders for processing image sequences and acoustic features to generate intermediate features. These intermediate features are then merged and passed through a shared backbone transformer encoder, which predicts masked cluster assignments. The model generates targets by clustering audio features or previously extracted features from an earlier iteration of the AV-HuBERT model. When the AV-HuBERT model is utilized for lip-reading, we omit the audio input to focus solely on the visual input. Modality dropout is utilized to address any differences between inputs.

2. Modality dropout

When it comes to relating audio input to lexical output, audio-visual speech recognition models tend to be more effective than visual input streams. However, this often leads to the audio modality dominating model decisions. The same problem exists in this model. To prevent the model from becoming overly dependent on the audio stream in our joint model, only the acoustic input is encoded with a linear layer to force the audio encoder to learn simple features.

3. Audio-visual clustering

Pre-training on both modalities provides an advantage in that it allows for the generation of multimodal cluster assignments, which can serve as target labels for the masked

prediction task of the next iteration. This differs from the approach used in cross-modal visual HuBERT, where targets are generated from audio-based features or a prior audio HuBERT model. In the case of AV-HuBERT, targets are naturally multimodal after the first iteration due to the complementary information provided by lip movement sequences to the audio stream.

4. Masking by substitution

We introduce a novel masking strategy. Specifically, segments in the visual stream are masked by substituting them with random segments from the same video.

4.3. Moco and wav2vec

4.3.1. Model Architecture

Visual Front-End

One approach to adapt the MoCo v2 model, which is pre-trained on ImageNet, for use in the AV-HuBERT framework, is to truncate the first convolutional layer and replace it with a layer of 3D convolution. Additionally, converting the RGB input image to grayscale before feeding it into the model can be beneficial, as it prevents the model from learning chromatic aberration information.

Audio Front-End

We used the wav2vec 2.0 model pre-trained on Libri-Light for our audio front-end. Specifically, we transfered both the 1D convolutional layers and the stacked transformer encoder layers of the pre-trained wav2vec 2.0 model into our audio front-end. The audio front-end takes as input raw audio waves sampled at 16 kHz and produces one vector representation every 20 ms.

Visual Back-End

The feature dimensions of the MoCo v2 output that enter the visual back-end number 2048, and this occurs at a rate of 25 vectors per second. In the visual back-end, we maintain this frequency but decrease the feature size to 512.

Audio Back-End

The wav2vec 2.0 outputs that are received by the audio back-end have a feature size of 1024 and occur at a rate of 50 vectors per second. To reduce the frequency, we set the stride of the 1D convolutional layer to two.

Fusion Module

In this part, the features from the audio and visual modes are combined to create a 1024-dimensional vector representation at a slower rate of 25 Hz. Before being concatenated on the feature dimension, we apply LayerNorm to each modality separately. This is necessary because it prevents any one modality from dominating the entire representation due to a larger variance.

Loss Function

In this work, we used a so-called hybrid CTC/attention loss for our training process. The form of the CTC loss assumes that there is conditional independence between each prediction of output:

$$p_{CTC}(\mathbf{y} \mid \mathbf{x}) \approx \prod_{t=1}^{T} p(y_t \mid \mathbf{x})$$

Conversely, an autoregressive decoder eliminates this assumption by estimating the posterior through the chain rule directly, which has a form of:

$$p_{\mathrm{CE}}(\mathbf{y} \mid \mathbf{x}) = \prod_{l=1}^{L} p(y_l \mid y_{<l}, \mathbf{x})$$

The overall objective function is computed as follows:

$$\mathcal{L} = \lambda \log p_{\mathrm{TC}}(\mathbf{y} \mid \mathbf{x}) + (1-\lambda) \log p_{\mathrm{CE}}(\mathbf{y} \mid \mathbf{x})$$

The weight factor in the hybrid CTC/attention mechanisms controls the balance between the CTC and seq2seq losses. This weight is essential not only when combining the two losses into a single training loss, but also when integrating the two predictions during decoding, as we will discuss in subsequent subsections.

5. Experiment

In this section, we design a series of experiments that reproduce each of the four excellent models mentioned in Section 3, and we experimented with them in a comprehensive manner.

5.1. Hardware Resources

The composition of the computing resources used in the experiment was as follows: Intel(R) Xeon(R) Gold 6134 CPU @ 3.20 GHz, GPU-Tesla V100 16 GB*2, SSD—5TB, OS—CentOS 6.0.

5.2. Input Features

- Audio Features

For audio input, we used a spectral size of 321 dimensions, computed using a 40 ms window with a 10 ms jump length, and a 16 kHz sample rate. Since the sampling rate of the video is 25 frames (40 ms per frame), each video input frame corresponds to 4 acoustic feature frames. We concatenated audio features into groups of 4 to reduce the length of the input sequence while achieving a common time scale for both modalities [8].

- Vision Module (VM)

For every video, we utilized dlib to identify and monitor 68 specific features on the face. The input image was 224 × 224 pixels, sampled at 25 frames/s, and contained the speaker's face. We cropped a 112 × 112 patch covering the area around the mouth, as shown in Figure 7.

Figure 7. Preprocessed images.

5.3. Model Setup

5.3.1. Transformer-CTC

The PyTorch library was used to implement the system, and it was trained on a Tesla V100 GPU with 16 GB of memory. The Adam optimizer was used with default parameters and an initial learning rate of 10^{-4}, which was decreased by a factor of 2 when the validation error stopped improving, eventually reaching a final learning rate of 10^{-6}. For all models, we used a dropout with $p = 0.1$ and label smoothing.

When dealing with a large number of timesteps, sequence-to-sequence learning can be slow to converge because the decoder initially struggles to extract relevant information from all the input steps. Despite not having any recurrent modules in our models, we found that implementing a curriculum instead of training on full sentences immediately was helpful.

Our approach involves starting with single-word examples and gradually increasing the sequence length as the network trains. These shorter sequences are parts of the longer

sentences in the dataset. We noticed that the rate of convergence on the training set was much faster, and the curriculum reduced overfitting by augmenting the data in a natural way.

We first trained the networks on the frozen features of the pre-trained sets from MV-LRS, LRS2-BBC, and LRS3-TED. To handle the variation in utterance lengths, we padded the sequences with zeros to a maximum length, which we steadily increased. We then fine-tuned the model end-to-end on the train-val set of either LRS2-BBC or LRS3-TED, depending on which set we were evaluating.

5.3.2. AV-HuBERT

The model takes in lip regions of interest (ROIs) for visual data and log filterbank energy features for audio data. The image encoder is based on a modified version of ResNet-18, and the audio encoder is a simple linear projection layer. There are two different model configurations: BASE which has 12 transformer blocks, and LARGE which has 24 transformer blocks. For BASE, each transformer block has an embedding dimension/feedforward dimension/attention head of 768/3072/12. For LARGE, these values are 1024/4096/16. The numbers of parameters in BASE and LARGE are 103 M and 325 M, respectively.

To improve the robustness of the AV-HuBERT model, noise enhancement was performed on the audio information. By training the model on noisy data, the ability of the model to resist noise was significantly improved.

To improve the input audio quality, we used discourse mixing, which selects random speech samples from the same small batch. To enhance the noise in pre-training, we used a wide range of sources, including non-speech noise, such as ambient and babble noise. To ensure the primary discourse is identified correctly, the intersection between the secondary and primary discourse should be less than 50 percent in WavLM, which focuses on pure audio self-supervised learning. Our approach is flexible and not restricted in terms of noise mixing, as the accompanying visual stream helps differentiate the primary and secondary discourse.

Our approach is unconstrained and more flexible in terms of mixing noise, since the accompanying visual stream disambiguates the primary and secondary discourse.

5.3.3. Moco

The final AVSR model was achieved through a series of training stages.

Firstly, the audio front-end was pre-trained through self-supervised learning using wav2vec 2.0 for the audio modality. Then, the audio front- and back-end were trained with dedicated decoders through the audio-only (AO) setting.

For the visual modality, the visual front-end was pre-trained through self-supervised learning and then modified to be trained through sequence classification at the word level for video clips in LRW data. The visual front-end was then used in the visual-only (VO) model with the visual back-end and dedicated decoders.

The final AVSR model was trained after the audio-only and visual-only models converged. Due to computational constraints, the audio and visual back-end outputs were pre-computed, and only the parameters in the fusion module and decoders were learned in the final stage.

5.4. Experiment Results

To evaluate the performances of the models more objectively, we used the best-known audiovisual databases LSR2 and LSR3 to evaluate the performance of each model in the presence and absence of noise in each of the three datasets.

5.4.1. ASR (Audio Speech Recognition)

The results for pure speech recognition are shown in Table 4. It can be seen that in the experiments, the Transformer-CTC model performed best in automatic speech recognition

with a WER of 0.123 on the LRS2 dataset and 0.244 on the LRS3 dataset. Some examples of successful model predictions are shown in Table 5.

Table 4. ASR results.

Model	TM-CTC	AV-Hubert	Moco
LRS2 without noisy	0.1230	0.1705	0.0513
LRS2 with noisy	0.5860	0.6031	0.3501
LRS3 without noisy	0.2440	0.0235	0.0501
LRS3 with noisy	0.7330	0.4031	0.3370

Table 5. ASR examples.

REF	HYP
i do not know	i do not know
that is pretty cool	that is pretty cool
raise your hands	raise your hands
thank you very much	thank you very much
what about polio	what about polio
did you find the gold	did you find the goal
life is good	life is good
how does it happen	how does it happen

5.4.2. Lips Only (Visual Speech Recognition)

The results for pure visual speech recognition are shown in Table 6. It can be seen that in the experiments, the ResNet-18 and convolution-augmented transformer (conformer) model performed best for lip reading (pure visual speech recognition), which achieved a WER of 0.261 on the LRS2 dataset and 0.323 on the LRS3 dataset. Some examples of successful model predictions are shown in Table 7.

Table 6. VSR results.

Model	TM-CTC	AV-Hubert	Moco
LRS2 without noisy	0.565	0.5785	0.4550
LRS3 without noisy	0.836	0.4231	0.4017

Table 7. VSR examples.

REF	HYP
i do not know	and so
that is pretty cool	spoto
the board of ed	I're moving it
I were wrong	I were wrong
this is great	this is great
thank you very much	thank you very much
you're in there	or in
lots of money	lots of money

5.4.3. Audio-Visual Speech Recognition

The results for audio-visual speech recognition are shown in Table 8. It can be seen that in the experiments, the AV-HuBERT model performed best in audiovisual speech recognition, with a WER of 0.137 on the LRS2 dataset and a WER of 0.0176 on the LRS3 dataset. Although the Transformer-CTC-based audio-visual speech recognition model performed well on the LRS2 dataset with a WER of 0.076, it did not perform as well on the LRS3 dataset with a WER of 0.164. An example of using the model to predict LRS3 is shown in the following Table 9. We can see that most of the sentences in the LRS3 dataset that

were predicted to have errors are relatively long, and the language model in the TM-CTC decoder was trained on the LRS2 dataset, which has relatively shorter sentences compared to LRS3. Therefore, in the future, we could try to use a mixed dataset of LRS2 and LRS3 and more audiovisual datasets as the training set of the language model, which is expected to achieve better training results.

Table 8. AVSR results.

Model	TM-CTC	AV-Hubert	Moco
LRS2 without noisy	0.0760	0.1370	0.0509
LRS2 with noisy	0.2310	0.3142	0.2452
LRS3 without noisy	0.1640	0.0176	0.0489
LRS3 with noisy	0.4360	0.1920	0.2130

Table 9. AVSR examples.

REF	HYP
i do not know	i do not know
that is pretty cool	that is pretty cool
raise your hands	raise your hands
thank you very much	thank you very much
what about polio	what about polio
did you find the gold	did you find the gold
life is good	life is good
how does it happen	how does it happen

5.5. Analysis

Through a series of experiments, we found that the performance of Moco+word2vec was the best among the three replicate models in both ASR and AVSR. By combining a CNN and a transformer, the model has both the excellent feature extraction ability of a CNN and the excellent receptive field of a transformer, so the feature extraction ability of the model is greatly improved, especially in lip reading. Since the main purpose of AVSR is to improve the accuracy of speech recognition through lip reading, the accuracy of the lip reading part has a very important impact on the overall performance of AVSR. Xichen Pan et al. designed a pre-training front-end model using LRW (word-level dataset) to train an excellent word-level-lip-reading front-end model and used this model as a pre-training model to greatly improve the performance of the lip reading part. The performance of the lip-reading part was greatly improved, and thus, the overall AVSR model performed much better than other models. Moreover, the method uses a hybrid of Conv and ransformer to extract features, which was found to be effective in extracting more valuable information.

However, the current study still has some shortcomings: Not only the three models we reproduced, but also almost all AVSR models in existence have been trained with a single dataset. Take the TM-CTC model as an example. Since it was trained on the training set of LRS2, its performance on the LRS3 test set is much worse than that on the LRS2 test set. In fact, LRS2 and LRS3 are similar data sets, just like their names imply—similar but different. This situation often occurred in our practical applications of the models as well. No matter which dataset the model is trained on, the training set must not contain the actual scenarios of our future applications in real life. Such an ability to adapt the model to all domains that are similar to but different from the training set is called the domain-adaptation capability.

Therefore, the research in the field of AVSR is mainly limited by feature-extraction ability and domain-generalization ability, and we can continue to go deeper in this direction in future research to further improve the performances of AVSR models.

6. Conclusions

In this paper, we presented the principles associated with several state-of-the-art visual speech recognition models and compared the performances of these models on the LRS2 and LRS3 datasets by replication. We used WER as an evaluation criterion. By observing the experimental results, we found that Moco built the network framework by using a combination of a convolutional network and transformer, and that the model outperforms the other two models in terms of AO, VO, and AVSR. While comparing their performances, we analyzed the problems of the current AVSR model, mainly in feature extraction and domain generalization, based on the experience accumulated during the research. Our research lays the foundation for further improving the performance of the AVSR model.

Although many important results have been achieved in audiovisual speech recognition, there are still many issues that need further investigation and research—for example, improving the recognition accuracy in noisy environments and enhancing its anti-interference capability. In addition, it is important to optimize the algorithm, reduce the use of computational resources, and improve its real-time performance and efficiency. Moreover, adding more diverse corpora to the database is also an important way to improve recognition performance. With the continuous development of technology and the expansion of application scenarios, there is still a lot of room for development in the field of audiovisual speech recognition processing. It is expected that the future research directions will focus more on practical applications, such as those in smart homes, smart customer service, smart driving, and other fields. At the same time, with the popularization of 5G and other new generation communication technologies, it will provide more efficient technical support for the field of audio-visual speech-recognition processing.

Author Contributions: Conceptualization, W.Y. (Wenfeng Yang); Methodology, W.Y. (Wenfeng Yang); Software, W.Y. (Wenfeng Yang) and P.L.; Validation, P.L., W.Y. (Wei Yang), Y.L., Y.H. and A.D.; Writing—original draft, W.Y. (Wenfeng Yang); Supervision, O.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Saint Petersburg State University (project ID: 94062114).

Data Availability Statement: LRW, LRS2, LRS3. LRW: https://www.robots.ox.ac.uk/vgg/data/lip_reading/lrw1.html; LRS2: https://www.robots.ox.ac.uk/vgg/data/lip_reading/lrs2.html; LRS3: https://www.robots.ox.ac.uk/vgg/data/lip_reading/lrs3-lang.html.

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

References

1. Biswas, A.; Sahu, P.K.; Chandra, M. Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *Int. J. Speech Technol.* **2016**, *19*, 159–171. [CrossRef]
2. Koguchi, Y.; Oharada, K.; Takagi, Y.; Sawada, Y.; Shizuki, B.; Takahashi, S. A mobile command input through vowel lip shape recognition. In *Human–Computer Interaction. Interaction Technologies: 20th International Conference, HCI International 2018*; Springer: Cham, Switzerland, 2018; pp. 297–305.
3. Jang, S.B.; Kim, Y.G.; Ko, Y.W. Mobile video communication based on augmented reality. *Multimed. Tools Appl.* **2017**, *76*, 16893–16909. [CrossRef]
4. Sun, K.; Yu, C.; Shi, W.; Liu, L.; Shi, Y. Lip-interact: Improving mobile device interaction with silent speech commands. In Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, Berlin, Germany, 14 October 2018; pp. 581–593.
5. Mohammed, A.; Mansour, A.; Ghulam, M.; Mohammed, Z.; Mesallam, T.; Malki, K.; Mohamed, F.; Mekhtiche, M.; Mohamed, B. Automatic speech recognition of pathological voice. *Indian J. Sci. Technol.* **2015**, *8*, 1–6. [CrossRef]
6. Afouras, T.; Chung, J.S.; Zisserman, A. The conversation: Deep audio-visual speech enhancement. *arXiv* **2018**, arXiv:1804.04121.
7. Chung, J.S.; Zisserman, A. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision*; Springer: Cham, Switzerland, 2017; pp. 87–103.
8. Son Chung, J.; Senior, A.; Vinyals, O.; Zisserman, A. Lip reading sentences in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6447–6456.
9. Chung, J.S.; Zisserman, A. Lip reading in profile. In Proceedings of the Ritish Machine Vision Conference (BMVC 2017), London, UK, 4–7 September 2017.

10. Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; Pantic, M. End-to-end audiovisual speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 6548–6552.
11. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *44*, 8717–8727. [CrossRef] [PubMed]
12. Yu, J.; Zhang, S.X.; Wu, J.; Ghorbani, S.; Wu, B.; Kang, S.; Liu, S.; Liu, X.; Meng, H.; Yu, D. Audio-visual recognition of overlapped speech for the lrs2 dataset. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6984–6988.
13. Ma, P.; Petridis, S.; Pantic, M. End-to-end audio-visual speech recognition with conformers. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7613–7617.
14. Shi, B.; Hsu, W.N.; Mohamed, A. Robust self-supervised audio-visual speech recognition. *arXiv* **2022**, arXiv:2201.01763.
15. Pan, X.; Chen, P.; Gong, Y.; Zhou, H.; Wang, X.; Lin, Z. Leveraging uni-modal self-supervised learning for multimodal audio-visual speech recognition. *arXiv* **2022**, arXiv:2203.07996.
16. Stafylakis, T.; Tzimiropoulos, G. Combining residual networks with LSTMs for lipreading. *arXiv* **2017**, arXiv:1703.04105.
17. Wiriyathammabhum, P. SpotFast networks with memory augmented lateral transformers for lipreading. In *Neural Information Processing: 27th International Conference, ICONIP 2020*; Springer: Cham, Switzerland, 2020; pp. 554–561.
18. Feng, D.; Yang, S.; Shan, S.; Chen, X. Learn an effective lip reading model without pains. *arXiv* **2020**, arXiv:2011.07557.
19. Ma, P.; Wang, Y.; Petridis, S.; Shen, J.; Pantic, M. Training strategies for improved lip-reading. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8472–8476.
20. Ivanko, D.; Ryumin, D.; Kashevnik, A.; Axyonov, A.; Karnov, A. Visual Speech Recognition in a Driver Assistance System. In Proceedings of the 2022 30th European Signal Processing Conference (EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 1131–1135.
21. Koumparoulis, A.; Potamianos, G. Accurate and Resource-Efficient Lipreading with Efficientnetv2 and Transformers. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 8467–8471.
22. Haliassos, A.; Ma, P.; Mira, R.; Petridis, S.; Pantic, M. Jointly Learning Visual and Auditory Speech Representations from Raw Data. *arXiv* **2022**, arXiv:2212.06246.
23. Ma, P.; Haliassos, A.; Fernandez-Lopez, A.; Chen, H.; Petridis, S.; Pantic, M. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. *arXiv* **2023**, arXiv:2303.14307.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
25. Graves, A.; Fernández, S.; Gomez, F.; Schmidhuber, J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.