applied sciences

MDPI

Article

# An Investigation into Audio–Visual Speech Recognition under a Realistic Home–TV Scenario

Bing Yin [1,2], Shutong Niu [1], Haitao Tang [2], Lei Sun [2], Jun Du [1], Zhenhua Ling [1] and Cong Liu [2,*]

[1] Department of Electronics and Information, University of Science and Technology of China, Hefei 230026, China
[2] IFLYTEK Research, Hefei 230088, China
* Correspondence: congliu2@iflytek.com

**Abstract:** Robust speech recognition in real world situations is still an important problem, especially when it is affected by environmental interference factors and conversational multi-speaker interactions. Supplementing audio information with other modalities, such as audio–visual speech recognition (AVSR), is a promising direction for improving speech recognition. The end-to-end (E2E) framework can learn information between multiple modalities well; however, the model is not easy to train, especially when the amount of data is relatively small. In this paper, we focus on building an encoder–decoder-based end-to-end audio–visual speech recognition system for use under realistic scenarios. First, we discuss different pre-training methods which provide various kinds of initialization for the AVSR framework. Second, we explore different model architectures and audio–visual fusion methods. Finally, we evaluate the performance on the corpus from the first Multi-modal Information based Speech Processing (MISP) challenge, which is recorded in a real home television (TV) room. By system fusion, our final system achieves a 23.98% character error rate (CER), which is better than the champion system of the first MISP challenge (CER = 25.07%).

**Keywords:** audio–visual speech recognition; pre-training; encoder–decoder; E2E

check for updates

**Citation:** Yin, B.; Niu, S.; Tang, H.; Sun, L.; Du, J.; Ling, Z.; Liu, C. An Investigation into Audio–Visual Speech Recognition under a Realistic Home–TV Scenario. *Appl. Sci.* **2023**, *13*, 4100. https://doi.org/10.3390/app13074100

Academic Editor: Douglas O'Shaughnessy

Received: 27 February 2023
Revised: 18 March 2023
Accepted: 20 March 2023
Published: 23 March 2023

## 1. Introduction

Automatic speech recognition (ASR) is an important direction in the field of speech signal processing. In the past decades, numerous research efforts have been made to boost the accuracy of automatic speech recognition tasks [1–4]. Through introducing deep learning methods, especially, ASR performance has made great advances [5–12]. In the early stages, these deep learning-based ASR systems were typically based on conventional statistic models such as hidden Markov models (HMMs), which are also called hybrid ASR systems [5,6]. Recently, benefiting from the development of deep learning technology and hardware computing capacity, a new structure for speech recognition has attracted more and more attention, which transits the ASR system from hybrid modeling to E2E modeling [7–12]. The E2E models can directly optimize the whole network through the objective functions, which have achieved potential results.

However, due to the wide applications of speech technologies, speech recognition systems are now facing more challenging scenarios (e.g., at home and in meetings [13,14]). In these scenarios, some adverse environmental factors, such as channel distortion, ambient noise and reversion, will degrade the ASR performance. In addition, the common overlapping regions in multi-speaker recordings also have negative impacts on the speech recognition system. In this case, researchers have begun to pay attention to visual information, and have found that visual information can play a positive role in the process of speech recognition [15–17]. To this end, AVSR has been developed. Modern AVSR systems based on deep learning technologies can be divided into two categories. The first type is still based on the deep neural network hidden Markov model (DNN-HMM) to build the AVSR system, such as the AVSR system proposed by Tao et al. [18]. The second one performs AVSR through end-to-end modeling. For example, Chung et al. [19] propose a 'Watch, Listen, Attend and Spell' (WLAS) network to recognize phrases and sentences through a talking face with or without the audio. Petridis et al. [20] use the hybrid connectionist temporal classification (CTC)/attention architecture for AVSR in the wild. Ref. [21] utilizes the Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU) network to build a two-stage speech recognition model. References [22,23] present the audio–visual speech recognition system based on a recurrent neural network transducer (RNN-T) architecture. A conformer has also been used in the AVSR system [24]. In addition, in order to better integrate audio and video features, the AV align framework is explored [25,26].

In the AVSR task, there are quite a number of audio–visual speech corpora for training and evaluation, such as GRID [27], OuluVS [28], OuluVS2 [29], TCD-TIMIT [30], LRW [31] and AVSpeech [32]. However, most of them are collected in a controlled environment, which makes it difficult to measure the performance of AVSR models in real scenarios. In addition, audio–visual data are more difficult to collect than pure audio data, which means that most of the above audio–visual datasets are greatly limited in terms of time duration and vocabulary size. Moreover, all the above datasets are in English, and few of the released audio–visual datasets are in Chinese [33,34]. Therefore, the MISP2021 challenge [35] was held, which contains the Chinese audio–visual speech corpus recorded in real-world application scenarios. The first MISP challenge was aimed at the home TV scenario, containing multiple speakers who were chatting in Chinese while watching TV or interacting with a smart speaker/TV [35]. Specifically, the MISP datasets were collected in real conversation, which makes the corpus contain high overlaps ratios and real domestic noise backgrounds, making it hard to obtain high recognition accuracy.

As mentioned above, audio–visual pairs are more difficult to collect than pure audio data, and the storage cost of audio–visual data is very high. Therefore, the sizes of audio–visual datasets in the real scenario are usually not very large (e.g., the duration of the MISP dataset is about 100 h), which means that methods (e.g., the E2E-based methods) with a large number of parameters and calculated quantities tend to be overfitting, as shown in [36].

In order to alleviate this problem, we explored from the perspective of pre-training to improve the accuracy and generalization ability of the AVSR model. We first investigated the impacts of different pre-training strategies on AVSR performance. On this basis, we also explored other parts of the AVSR system, including model architectures, audio and video fusion methods, system fusion methods, and so on, which may further improve the performance of the AVSR system.

We evaluated the use of different methods on the released MISP 2021 dataset [35], and the experimental results show that the pre-training methods can help the AVSR system achieve significant improvement. Specifically, after combining effective fusion strategies, the performance of our AVSR system is better than that which achieved first place in the MISP 2021 challenge [37]. The contributions of this paper can be summarized as follows:

(1) We explore different pre-training strategies in the AVSR tasks, and this can provide some useful experience for deploying AVSR systems in real scenarios where there is only a limited size of audio–visual data;
(2) Based on (1), we explore the impacts of different model architectures for audio–visual embedding extractors on the AVSR performance;
(3) We explore the performance of different audio–visual fusion methods.

Our final system achieved 23.98% CER on the first MISP 2021 AVSR corpus, which is better than the champion system [37] of the first MISP challenge (CER = 25.07%).

The rest of the paper is organized as below: In Section 2, the employed end-to-end AVSR framework is introduced. In Section 3, the explored methods are explained in detail. In Section 4, we present and analyze the experimental results. Finally, we conclude in Section 5.

## 2. End-to-End AVSR Framework

In this paper, we employ an advanced encoder–decoder based end-to-end AVSR framework as the baseline system, which is presented in Figure 1. In the encoder part, we utilize two extractors for audio and video modalities, respectively, and then splice the extracted features as the input of the conformer [38] to model the context dependencies within the inputs. In the decoder part, we employ the transformer [39] to generate the posterior probabilities of the character sequence. 'Fusion-cat' means directly splicing the features of audio and video.

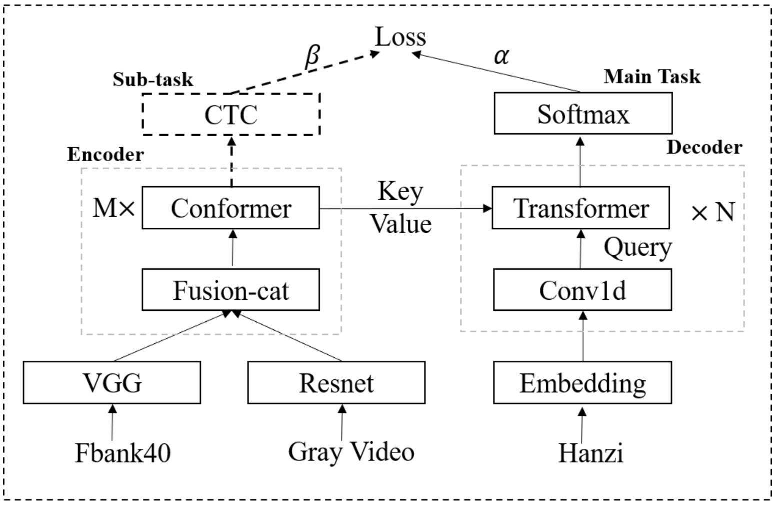

**Figure 1.** The illustration of the end-to-end AVSR baseline system.

Moreover, we utilize the multi-task learning framework to improve the robustness of the system, as in [40]. The overall multi-task learning (MTL) loss function is defined by the following equations:

$$\mathcal{L}_{\mathrm{CTC}} = -\ln p_e(z^* \mid \boldsymbol{X}) \tag{1}$$

$$\mathcal{L}_{\text{Attention}} = -\ln p_d(z^* \mid \boldsymbol{X}) = -\sum_u \ln p_d\left(z_u^* \mid \boldsymbol{X}, z_{1:u-1}^*\right) \tag{2}$$

$$\mathcal{L}_{\mathrm{MTL}} = \alpha \times \mathcal{L}_{\text{Attention}} + \beta \times \mathcal{L}_{\mathrm{CTC}} \tag{3}$$

where $z^*$ denotes the ground truth character sequence. $\boldsymbol{X}$ means the input features. $p_e\left(z^* \mid \boldsymbol{X}\right)$ and $p_d\left(z_u^* \mid \boldsymbol{X}, z_{1:u-1}^*\right)$ are the probability distributions estimated from the shared encoder and attention decoder, respectively. $u$ is the index of ground truth characters. $z_{1:u-1}^*$ is the label of the previous $u-1$ characters.

A compact E2E audio–visual framework can efficiently learn multi-modal information. However, direct training procedures do not guarantee good results, especially when the amount of training data is small. Various problems will occur, such as overfitting, non-convergence, and even training failure. Hence, pre-training becomes the most popular method to alleviate the training difficulty.

## 3. Methods

### *3.1. Different Pre-Training Strategies*

**Pre-training with the Hybrid System (HS).** First, we adopted a strategy to initialize the audio–visual encoder part of the baseline system, which is shown in Figure 2. We first employed the conventional hybrid ASR system [5,6] to generate the corresponding force alignment (FA), then applied them to pre-train the audio–visual encoder, as presented on the left side of Figure 2. After pre-training, we initialize the encoder part of the AVSR system, as shown on the right side of Figure 2, and the gray blocks represent the randomly initialized parts.

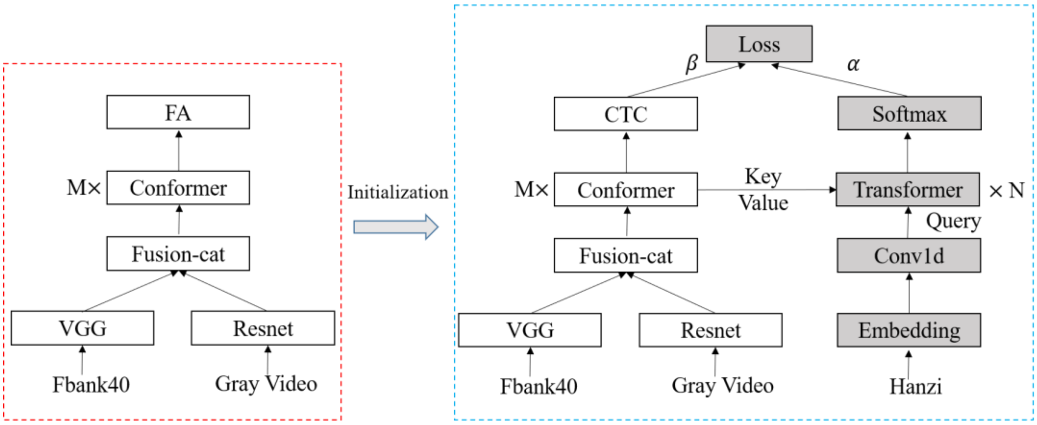

**Figure 2.** Pre-training with the hybrid system.

**Pre-training with the End-to-end System (ES).** As illustrated in Figure 3, we also used the E2E ASR system to initialize the AVSR system. Note that the key difference is whether the visual input is included in the training. Most modules in the AVSR system can directly use the parameters in the pre-training process, and the rest use random initialization (i.e., the parts marked with gray blocks). Through fine-tuning, the AVSR model can gradually absorb visual modal information and avoid training instability.

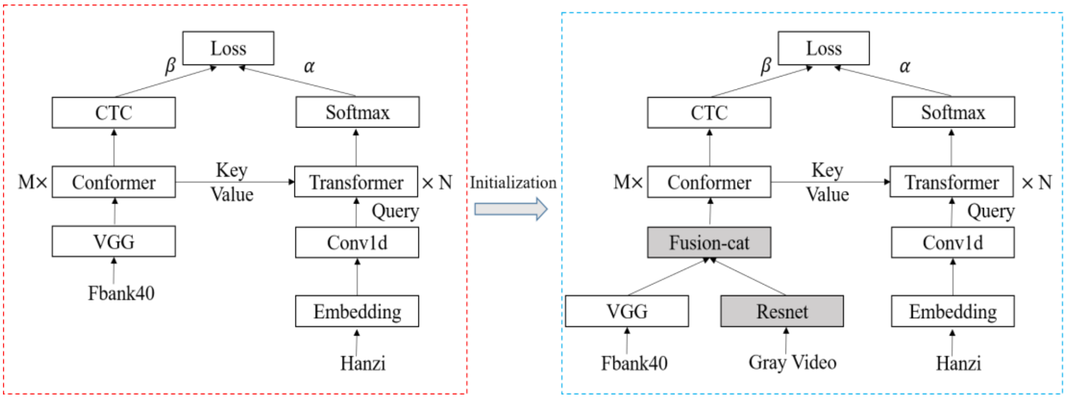

**Figure 3.** Pre-training with the end-to-end ASR system.

**Using Pre-trained Audio-Visual Representation.** Recently, self-supervised speech representation learning has attracted lots of attention. We used the Audio-Visual Hidden Unit BERT (AV-HuBERT) [41] model, which has shown great recognition performance under adverse noise conditions. Specifically, in our previous work, we found that adopting the entire face can achieve better performance on the lipreading task compared with the baseline method using lip as visual input [42]. Therefore, we used it as the feature extractor, as shown in Figure 4.

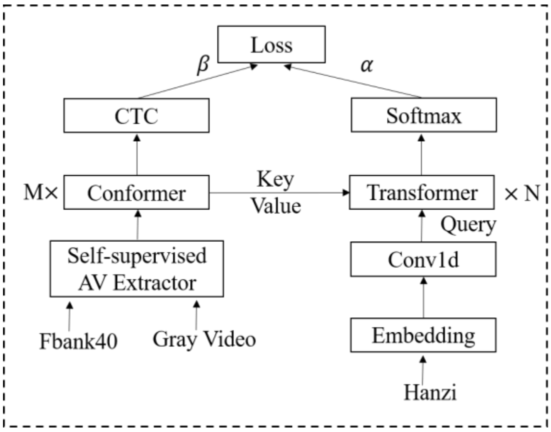

**Figure 4.** The AVSR system with pre-trained audio–visual representation.

### 3.2. Model Architectures

In this paper, we also explored the impacts of the model architectures for audio and visual embedding extractors. As shown in the blue box of Figure 5, we separately replaced the audio and visual extractors with other architectures. For the audio embedding extractor, we replaced the original VGG with gate CNN [43]. For the video embedding extractor, we replaced the original Resnet with Swin Transformer [44].

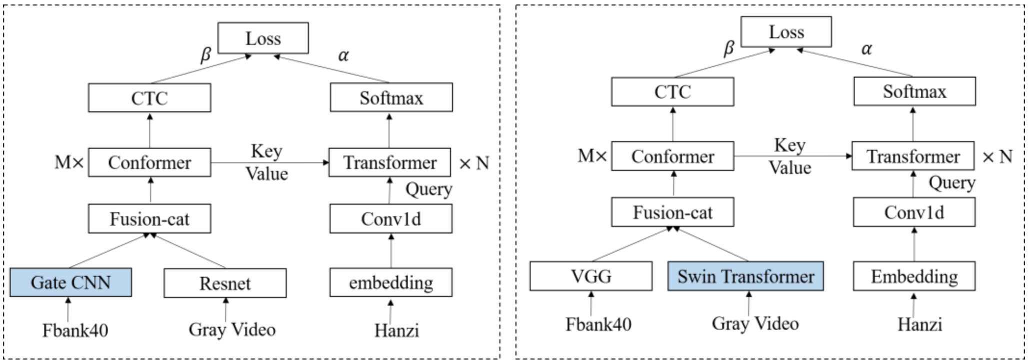

**Figure 5.** Exploration of model architectures of audio and visual embedding extractor.

### 3.3. Audio–Visual Fusion

The fusion of audio and visual information also plays an important role in AVSR. The widely used approach is to directly splice the two modal features, as described in our previous system. We also explore the different audio–visual fusion methods to replace the simple embedding concatenation from the encoder and decoder parts, respectively.

**Fusion in the Encoder Part.** For the encoder part, we employ the self-attention (SA) operation, which can be expressed as:

$$\text{Multihead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = [\mathbf{H}_1,\mathbf{H}_2,\ldots,\mathbf{H}_h]\mathbf{W}^O \tag{4}$$

$$\mathbf{H}_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q,\mathbf{K}\mathbf{W}_i^K,\mathbf{V}\mathbf{W}_i^V\right) \tag{5}$$

$$\text{Attention}(\mathbf{q},\mathbf{k},\mathbf{v}) = \text{softmax}\left(\frac{\mathbf{q}\mathbf{k}^{\mathrm{T}}}{\sqrt{d_k}}\right)\mathbf{v} \tag{6}$$

where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ denote the video, audio and audio features, respectively. $[\cdot]$ represents the concatenation operation. $h$ is the number of attention heads. $\mathbf{W}^O \in \mathbb{R}^{(h\times d_v)\times d}$ is the final linear parameter matrix. $\mathbf{W}_i^Q,\mathbf{W}_i^K \in \mathbb{R}^{d\times d_k}$ and $\mathbf{W}_i^V \in \mathbb{R}^{d\times d_v}$ are the learnable parameter for the $i$-th head. $d$, $d_k$ and $d_v$ denote the dimension of input, key and value, respectively.

**Fusion in the Decoder Part.** For the decoder part, we performed the audio–visual fusion based on the transformer [39], as shown in Figure 6. We first used the self-attention for audio and video encoder outputs, and then we fused the two attention outputs as:

$$\text{out} = [\text{Multihead}_{audio}(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1), \text{Multihead}_{video}(\mathbf{Q}_2, \mathbf{K}_2, \mathbf{V}_2)] \tag{7}$$

where $\text{Multihead}_{audio}(\bullet)$ and $\text{Multihead}_{video}(\bullet)$ are the self-attentions based on the audio and video encoder outputs, respectively. $\mathbf{Q}_1, \mathbf{K}_1$ and $\mathbf{V}_1$ are the vectors of text embedding, audio encoder output and audio encoder output, respectively. $\mathbf{Q}_2, \mathbf{K}_2$ and $\mathbf{V}_2$ are the vectors of text embedding, video encoder output and video encoder output, respectively.

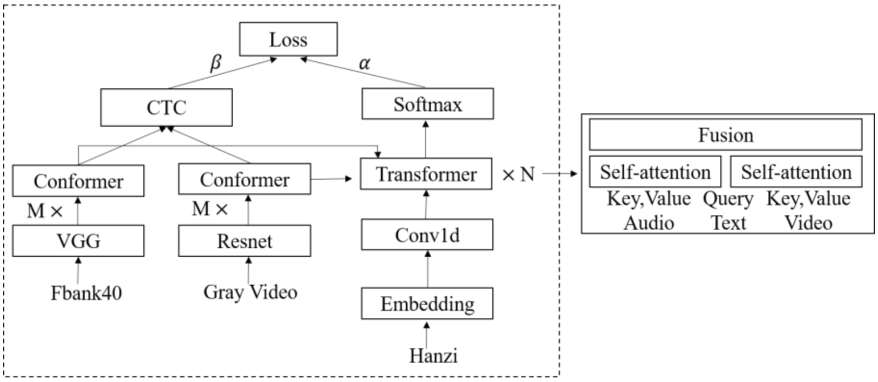

**Figure 6.** Audio–visual fusion in the decoder part.

Finally, in order to further improve the overall performance, we also performed the system fusion by averaging the posterior probabilities of various models.

## 4. Experiments and Analysis

### *4.1. Experimental Setups*

We evaluated the use of different methods on the MISP2021-AVSR corpus [35], which contained about 122 h of audio–visual data collected from 30 real-life living rooms. The dataset includes 376 sessions, with each session consisting of an approximately 20 min discussion under the real home TV scenario. There were 95 male speakers and 153 female speakers. The collected data were partitioned into three distinct subsets for training (about 101 h), development (about 9.83 h), and evaluation (about 9.94 h). Each subset was comprised of unique sets of speakers and recording rooms that do not intersect with any other subset. In order to be close to the realistic home TV scenario, we only used far-field data for training (about 100 h), which included videos from the wide-angle cameras and six-channel audios from the linear microphone arrays. The evaluation corpus was collected by far-field devices. Because low-resource training data (about 100 h) contain simple context and spectral information, catting continuous short segments (CAT) [45] was used to augment the original audio–visual data. We obtained a new 300 h CAT dataset for training. Furthermore, it is difficult for speed perturbation to align audio and video time series. We only employed speed perturbation (SP) for audio with ratios 0.8, 1.0 and 1.2. For the hybrid ASR/AVSR system, we utilized the baseline system in the MISP2021 challenge [35]. For the E2E-AVSR framework, we utilized combining the main task of attention and the sub-task of CTC as the acoustic model [24]. The two tasks shared the same encoder, and the $\alpha$ and $\beta$ in Equation (3) were set to 0.8 and 0.2, respectively. In the E2E-AVSR framework, we used two layers of VGG as the audio extractor, like in [45], and the Resnet18 as the video extractor. After VGG and Resnet, audio and video hidden representations were concatenated along time series as inputs of the conformer [38], where the number of layers $M$ was set to 12. In the decoder part, we utilized a six-layer transformer [39] (i.e., $N$ = 6). The number of attention heads was eight. All networks were implemented using Fairseq [46]. We applied SpecAugment [47] to improve the robustness. In addition, gradient clipping was adopted with a value of 10 to avoid an exploding gradient. All models were trained with a batch size for 8 k frames and a transformer [39] learning-rate schedule with 3 k warm-up steps. The max learning rate was set to 0.0007. Character error rate (CER) was employed as the evaluation metric as in [35], which is defined as follows:

$$\text{CER} = \frac{S + D + I}{N} \times 100 \tag{8}$$

where $S$, $D$, $I$ and $N$ represent the number of substitutions, deletions, insertions and characters in the ground truth, respectively.

### *4.2. Experimental Results*

#### 4.2.1. Results of the Pre-Training Strategies

**The results of hybrid system (HS) pre-training.** The results of pre-training with force alignment (FA) from the hybrid ASR system are listed in Table 1. From this table, we can make several observations: (1) the splicing of continuous segments (CAT) can effectively expand the audio data and improve the performance of ASR, as shown in the first and second rows of Table 1; (2) when the training data size is not large enough (i.e., the 100 h original far-field audio–visual training data), the AVSR model using the E2E framework (CER = 52.26%) is worse than the conventional hybrid AVSR model (CER = 44.12%), which indicates that the AVSR system based on the E2E architecture overfits when there are not enough training data. This is also our motivation to explore the pre-training strategies; (3) through utilizing the force alignment (FA) from the hybrid ASR system for pre-training, the AVSR system has been significantly improved and under the same architectures the CER decreased by 18.84% (from 52.26% to 33.78%); (4) after using the HS pre-training strategies, the recognition performance of end-to-end AVSR (CER = 33.78%) was significantly better than that of hybrid AVSR (CER = 44.12%), which also shows the effectiveness of our pre-training strategies.

**Table 1.** The results (CER %) of hybrid system pre-training. HS means the hybrid system. CAT means splicing the continuous segments.

| Task | Model | Pre-Training | Data Processing | Eval |
|---|---|---|---|---|
| ASR | Hybrid | N | N | 45.92 |
| | | | CAT | 41.69 |
| AVSR | Hybrid | N | N | 44.12 |
| | E2E [36] | N | N | 52.26 |
| | E2E | HS (CAT) | N | 35.93 |
| | E2E | HS (CAT) | CAT | 33.78 |

**The results of end-to-end system (ES) pre-training. Table** 2 presents the results of the end-to-end system (ES) pre-training. Note that, for the ASR-E2E system, we also use the hybrid system pre-training methods to perform the system initialization. From this table, we can draw the following conclusions: (1) Through utilizing the HS pre-training methods, the ASR-E2E system (CER = 36.81%) can achieve better performance than the ASR-Hybrid system (CER = 41.69%) in Table 1. (2) The recognition performance of the ASR-E2E system will be further improved by applying both SP and CAT methods at the same time. When we only use the CAT method, the CER is 36.81%. When we only use the SP method, the CER is 37.44%. When we use both SP and CAT methods simultaneously, the CER is 35.90%, indicating better performance. (3) The ES pre-training (CER = 32.21%) can achieve better results than HS pre-training (CER = 33.78%) in Table 1, which may have two reasons. First, the ASR-E2E model achieved better performance than the ASR-Hybrid model. Second, the ES pre-training initialized more modules in the AVSR system, as shown in Figures 2 and 3.

Similarly, combining CAT and SP techniques for data augmentation still yields the best results in AVSR systems.

**Table 2.** The results (CER %) of the ES pre-training. HS and ES mean the hybrid system and end-to-end system, respectively. CAT and SP mean splicing the continuous segments and speed perturbation, respectively.

| Task | Model | Pre-Training | Data Processing | Eval |
|---|---|---|---|---|
| ASR | E2E | HS (CAT) | CAT | 36.81 |
| | | HS (SP) | SP | 37.44 |
| | | HS (SP + CAT) | SP + CAT | 35.90 |
| AVSR | E2E | ES (CAT) | CAT | 33.16 |
| | | ES (SP) | CAT | 34.00 |
| | | ES (SP + CAT) | CAT | 32.21 |

**The results of pre-trained audio–visual representation.** Table 3 presents the results of utilizing the pre-trained audio–visual representation. Due to the system with pre-trained audio–visual representation being unable to converge on the model with the original configuration (M = 12, N = 6), as shown in the last row in Table 3, we also used other configurations (M = 6, N = 3 and M = 10, N = 4). As can be seen, the introduction of pre-trained audio–visual representation brought performance degradation to the AVSR system (the CER increased from 38.27% to 42.78%). Furthermore, increasing the number of model parameters (i.e., M = 10, N = 4) only leads to marginal improvement (the CER decreased from 42.78% to 41.85%). The reason for this phenomenon may be the mismatch in the data between the AV-HuBERT-V1 [42] training and the MISP2021 AVSR training. Taking the pre-trained representations as auxiliary features and splicing them with the original features may be helpful to the performance.

**Table 3.** The results (CER %) of pre-trained audio–visual representation on the AVSR task. HS means the hybrid system. CAT means splicing the continuous segments. M and N denote the number of blocks in the encoder and decoder modules, respectively. AV-HuBERT-V1 is the utilized modified AV-HuBERT model.

| Model | Pre-Training | Extractor | Data Processing | Eval |
|---|---|---|---|---|
| E2E (M = 6, N = 3) | HS (CAT) | VGG/Resnet | CAT | 38.27 |
| E2E (M = 6, N = 3) | | AV-HuBERT-V1 | | 42.78 |
| E2E (M = 10, N = 4) | | | | 41.85 |
| E2E (M = 12, N = 6) | | | | 96.98 |

### 4.2.2. Results of Different Model Architectures

The results of replacing the audio and visual extractors with better model architectures are shown in Table 4. It can be seen from the table that the performance of AVSR will be improved by adopting a better audio extractor. On the contrary, if a better video extractor is adopted, the performance of AVSR will degrade. This may be caused by the difference in the amount of information brought by the audio and video data. In order to better analyze these two modalities, we compared the differences between AVSR and ASR in the original VGG/Resnet architecture under different environments in Table 5. As can be seen, when there is considerable TV background noise (i.e., C1, C2, C5), AVSR has a significant improvement over ASR (about 5% absolute values). On the contrary, when there is no background noise (i.e., C7, C8), AVSR does not bring much improvement compared with ASR. In particular, under the C6 conditions, the performance of AVSR is worse than that of

ASR, which may be because the lights and TV interfere with the video data, leading to the introduction of wrong information. In general, the video modality only brings an absolute improvement of 2.53% for recognition performance. Therefore, compared with the audio modality, the video modality contains less effective information, which may be the reason why the recognition performances become worse when a better video extractor is used.

**Table 4.** The results (CER %) of the different model architectures. ES means the end-to-end system. SP means speed perturbation. CAT means splicing the continuous segments.

| Task | Model | Pre-Training | A/V Extractor | Eval |
|---|---|---|---|---|
| AVSR | E2E | ES<br>(SP + CAT) | VGG/Resnet | 32.21 |
| | | | Gate CNN/Resnet | 29.88 |
| | | | VGG/Swin Transformer | 36.13 |

**Table 5.** The comparison of ASR-E2E and AVSR-E2E under the same pre-training strategy and training data. For the pre-training strategy, we use the 'HS (CAT)' method. For the training data, we used the 'CAT' method.

| Task | Model | ALL | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 |
|---|---|---|---|---|---|---|---|---|---|---|
| ASR | E2E | 41.15 | 46.99 | 45.45 | 48.02 | 37.84 | 41.80 | 47.43 | 32.23 | 33.14 |
| AVSR | | 38.62 | 42.44 | 39.80 | 45.61 | 35.13 | 37.42 | 49.60 | 30.71 | 31.98 |

### 4.2.3. Results of Audio–Visual Fusion

Table 6 shows the performance of the AVSR system after modifying the fusion methods. Because changing the fusion method in the decoder part does not cause convergence, we only show the recognition results for changing the fusion method in the encoder part. Note that, when using the self-attention to fuse the audio and video information in the encoder part, we do not need to strictly align the two modalities in time series, so we can apply the speed perturbation to the audio–visual data, namely 'CAT + SP' in the 'Data Processing' column. As shown in Table 6, our modified fusion methods do not bring improvement. We suspect that this is because the difference in the amount of information between these two modalities is too large, which leads to the result that fusion methods that need to learn parameters cannot extract useful information effectively.

**Table 6.** The results (CER %) of audio–visual fusion (self-attention in the encoder part) methods. HS means the hybrid system. CAT and SP mean splicing the continuous segments and speed perturbation, respectively. 'Splicing' means directly splicing the two modal features in audio–visual fusion. SA means self-attention.

| Task | Model | Pre-Training | Fusion | Data Processing | Eval |
|---|---|---|---|---|---|
| AVSR | E2E | HS (CAT) | Splicing | CAT | 33.78 |
| | | | SA | CAT | 38.16 |
| | | | SA | CAT + SP | 37.43 |

### 4.2.4. Results of System Fusion

Finally, we performed the system fusion on the posterior probabilities, and the results are shown in Table 7. We can see that the system fusion has made significant improvements, and the performance of our final system is better than that of the champion system in the MISP2021 [37] challenge. Note that, in order to improve the diversity of the systems, we also use the Transformer [39] as the encoder in the AVSR-E2E system, as shown in 'ID 3' of Table 7. For better comparison with other methods, we also list the MISP baseline results [35] and the SJTU system results in the first MISP challenge [48] at the bottom of Table 7, which uses the same dataset as ours.

**Table 7.** The results (CER %) of system fusion. HS and ES mean the hybrid system and end-to-end system, respectively. M and N denote the number of blocks in the encoder and decoder modules, respectively. CAT means splicing the continuous segments. SP means speed perturbation.

| ID | Model | Pre-Training | Eval |
|---|---|---|---|
| 1 | E2E | HS (CAT) | 33.78 |
| 2 | E2E | ES (CAT + SP) | 32.21 |
| 3 | E2E<br>Transformer Encoder | ES (CAT + SP) | 32.54 |
| 4 | E2E<br>Gate CNN for audio | ES (CAT + SP) | 29.88 |
| 5 | 1 + 2 + 3 + 4<br>Fusion | - | 23.98 |
| 6 | NIO [37] | - | 25.07 |
| 7 | MISP Baseline [35] | - | 62.74 |
| 8 | SJTU [48] | - | 34.02 |

## 5. Conclusions

In this paper, we analyzed and investigated the AVSR system under the realistic Chinese home TV scenario. Aiming at the overfitting problem which is often encountered in AVSR-E2E systems, we proposed different pre-training strategies. Experimental results showed that effective pre-training strategies can significantly improve recognition performance. At the same time, we also explored other aspects of the AVSR system, including model architecture, audio–visual fusion, and system fusion. Some of them can further improve the model's performance. By integrating the effective methods we found, our final system was superior to the champion system from the MISP 2021 challenge. Compared to the state-of-the-art methods, the improvement of our approach is primarily attributable to the exploration of various initialization methods and the fusion of their results. In the future, we will explore how to better utilize audio–visual fusion modules and pre-training features to improve recognition performance in the AVSR task.

**Author Contributions:** B.Y. contributed the central idea, designed the entire project and wrote the manuscript. S.N., H.T. and L.S. performed the experiment. J.D. and Z.L. helped perform the data analysis with constructive discussions. C.L. provided helpful guidance at every stage of the whole work. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Key R&D Program of China (2020AAA0103600). This work was partially supported by the National Natural Science Found of China (Grant Nos. 61727809, 62076209).

**Data Availability Statement:** The datasets in our work are from the Multimodal Information Based Speech Processing (MISP) Challenge 2021. The link is https://mispchallenge.github.io/ (accessed on 18 March 2023).

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

## References

1. Baker, J.M.; Deng, L.; Glass, J.; Khudanpur, S.; Lee, C.H.; Morgan, N.; O'Shaughnessy, D. Developments and directions in speech recognition and understanding, Part 1 [DSP Education]. *IEEE Signal Process. Mag.* **2009**, *26*, 75–80. [CrossRef]
2. Deng, L.; Hinton, G.; Kingsbury, B. New types of deep neural network learning for speech recognition and related applications: An overview. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 8599–8603.

3. Deng, L.; Li, X. Machine learning paradigms for speech recognition: An overview. *IEEE Trans. Audio Speech Lang. Process.* **2013**, *21*, 1060–1089. [CrossRef]
4. Li, J. Recent advances in end-to-end automatic speech recognition. *APSIPA Trans. Signal Inf. Process.* **2022**, *11*, e8. [CrossRef]
5. Hinton, G.; Deng, L.; Yu, D.; Dahl, G.E.; Mohamed, A.R.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T.N.; et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.* **2012**, *29*, 82–97. [CrossRef]
6. Yu, D.; Deng, L. *Automatic Speech Recognition*; Springer: Berlin/Heidelberg, Germany, 2016; Volume 1.
7. Graves, A.; Jaitly, N. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*; PMLR: Bejing, China, 2014; pp. 1764–1772.
8. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv* **2014**, arXiv:1412.5567 2014.
9. Chorowski, J.; Bahdanau, D.; Cho, K.; Bengio, Y. End-to-end continuous speech recognition using attention-based recurrent NN: First results. *arXiv* **2014**, arXiv:1412.1602 2014.
10. Miao, Y.; Gowayyed, M.; Metze, F. EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 167–174.
11. Bahdanau, D.; Chorowski, J.; Serdyuk, D.; Brakel, P.; Bengio, Y. End-to-end attention-based large vocabulary speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 25 March 2016; pp. 4945–4949.
12. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 25 March 2016; pp. 4960–4964.
13. Watanabe, S.; Mandel, M.; Barker, J.; Vincent, E.; Arora, A.; Chang, X.; Khudanpur, S.; Manohar, V.; Povey, D.; Raj, D.; et al. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. *arXiv* **2020**, arXiv:2004.09249 2020.
14. Yu, F.; Zhang, S.; Fu, Y.; Xie, L.; Zheng, S.; Du, Z.; Huang, W.; Guo, P.; Yan, Z.; Ma, B.; et al. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 6167–6171.
15. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef]
16. Rosenblum, L.D. Speech perception as a multimodal phenomenon. *Curr. Dir. Psychol. Sci.* **2008**, *17*, 405–409. [CrossRef]
17. Massaro, D.W.; Simpson, J.A. *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*; Psychology Press: London, UK, 2014.
18. Tao, F.; Busso, C. Gating neural network for large vocabulary audiovisual speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1290–1302. [CrossRef]
19. 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.
20. Petridis, S.; Stafylakis, T.; Ma, P.; Tzimiropoulos, G.; Pantic, M. Audio-Visual Speech Recognition with a Hybrid Ctc/Attention Architecture. In Proceedings of the 2018 IEEE Spoken Language TechnologyWorkshop (SLT), Athens, Greece, 18–21 December 2018; pp. 513–520.
21. Xu, B.; Lu, C.; Guo, Y.; Wang, J. Discriminative multi-modality speech recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14433–14442.
22. Makino, T.; Liao, H.; Assael, Y.; Shillingford, B.; Garcia, B.; Braga, O.; Siohan, O. Recurrent neural network transducer for audio-visual speech recognition. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 905–912.
23. Braga, O.; Makino, T.; Siohan, O.; Liao, H. End-to-End Multi-Person Audio/Visual Automatic Speech Recognition. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6994–6998.
24. 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.
25. Sterpu, G.; Saam, C.; Harte, N. Attention-based audio-visual fusion for robust automatic speech recognition. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 19–20 October 2018; pp. 111–115.
26. Sterpu, G.; Saam, C.; Harte, N. How to teach DNNs to pay attention to the visual modality in speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1052–1064. [CrossRef]
27. Cooke, M.; Barker, J.; Cunningham, S.; Shao, X. An audio-visual corpus for speech perception and automatic speech recognition. *J. Acoust. Soc. Am.* **2006**, *120*, 2421–2424. [CrossRef]
28. Zhao, G.; Barnard, M.; Pietikainen, M. Lipreading with local spatiotemporal descriptors. *IEEE Trans. Multimed.* **2009**, *11*, 1254–1265. [CrossRef]

29. Anina, I.; Zhou, Z.; Zhao, G.; Pietikäinen, M. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Volume 1, pp. 1–5.
30. Harte, N.; Gillen, E. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Trans. Multimed.* **2015**, *17*, 603–615. [CrossRef]
31. Chung, J.S.; Zisserman, A. Lip reading in the wild. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016, Revised Selected Papers, Part II 13*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 87–103.
32. Ephrat, A.; Mosseri, I.; Lang, O.; Dekel, T.; Wilson, K.; Hassidim, A.; Freeman, W.T.; Rubinstein, M. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv* **2018**, arXiv:1804.03619 2018. [CrossRef]
33. Yu, J.; Su, R.; Wang, L.; Zhou, W. A multi-channel/multi-speaker interactive 3D audio-visual speech corpus in Mandarin. In Proceedings of the 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), Tianjin, China, 17–20 October 2016; pp. 1–5.
34. Liu, H.; Chen, Z.; Shi, W. Robust Audio-Visual Mandarin Speech Recognition Based on Adaptive Decision Fusion and Tone Features. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 25–28 October 2020; pp. 1381–1385.
35. Chen, H.; Zhou, H.; Du, J.; Lee, C.H.; Chen, J.; Watanabe, S.; Siniscalchi, S.M.; Scharenborg, O.; Liu, D.Y.; Yin, B.C.; et al. The First Multimodal Information Based Speech Processing (Misp) Challenge: Data, tasks, baselines and results. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 9266–9270.
36. Chen, H.; Du, J.; Dai, Y.; Lee, C.H.; Siniscalchi, S.M.; Watanabe, S.; Scharenborg, O.; Chen, J.; Yin, B.C.; Pan, J. Audio-Visual Speech Recognition in MISP 2021 Challenge: Dataset Release and Deep Analysis. In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Incheon, Korea, 18–22 September 2022; Volume 2022, pp. 1766–1770.
37. Xu, G.; Yang, S.; Li, W.; Wang, S.; Wei, G.; Yuan, J.; Gao, J. Channel-Wise AV-Fusion Attention for Multi-Channel Audio-Visual Speech Recognition. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 9251–9255.
38. Gulati, A.; Qin, J.; Chiu, C.C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100 2020.
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
40. Watanabe, S.; Hori, T.; Kim, S.; Hershey, J.R.; Hayashi, T. Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1240–1253. [CrossRef]
41. Shi, B.; Hsu, W.N.; Lakhotia, K.; Mohamed, A. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv* **2022**, arXiv:2201.02184 2022.
42. Zhang, J.X.; Wan, G.; Pan, J. Is Lip Region-of-Interest Sufficient for Lipreading? In Proceedings of the 2022 International Conference on Multimodal Interaction, Bengaluru, India, 21–22 January 2022; pp. 368–372.
43. Yuan, J.; Xiong, H.C.; Xiao, Y.; Guan, W.; Wang, M.; Hong, R.; Li, Z.Y. Gated CNN: Integrating multi-scale feature layers for object detection. *Pattern Recognit.* **2020**, *105*, 107131. [CrossRef]
44. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
45. Zhang, W.; Ye, Z.; Tang, H.; Li, X.; Zhou, X.; Yang, J.; Cui, J.; Deng, P.; Shi, M.; Song, Y.; et al. The USTC-NELSLIP Offline Speech Translation Systems for IWSLT 2022. In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), Dublin, Ireland, 26–27 May 2022; pp. 198–207.
46. Ott, M.; Edunov, S.; Baevski, A.; Fan, A.; Gross, S.; Ng, N.; Grangier, D.; Auli, M. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv* **2019**, arXiv:1904.01038 2019.
47. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779 2019.
48. Wang, W.; Gong, X.; Wu, Y.; Zhou, Z.; Li, C.; Zhang, W.; Han, B.; Qian, Y. The Sjtu System for Multimodal Information Based Speech Processing Challenge 2021. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 9261–9265.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.