applied sciences

MDPI

Article

# Comparing Speaker Adaptation Methods for Visual Speech Recognition for Continuous Spanish †

**David Gimeno-Gómez *,‡ and Carlos-D. Martínez-Hinarejos *,‡**

Pattern Recognition and Human Language Technologies Research Center, Universitat Politècnica de València, Camino de Vera, s/n, 46022 València, Spain

* Correspondence: dagigo1@dsic.upv.es (D.G.-G.); cmartine@dsic.upv.es (C.-D.M.-H.)

† This paper is an extended version of our paper published in Gimeno-Gómez, D.; Martínez-Hinarejos, C.D. Speaker-Adapted End-to-End Visual Speech Recognition for Continuous Spanish. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 41–45.

‡ These authors contributed equally to this work.

**Abstract:** Visual speech recognition (VSR) is a challenging task that aims to interpret speech based solely on lip movements. However, although remarkable results have recently been reached in the field, this task remains an open research problem due to different challenges, such as visual ambiguities, the intra-personal variability among speakers, and the complex modeling of silence. Nonetheless, these challenges can be alleviated when the task is approached from a speaker-dependent perspective. Our work focuses on the adaptation of end-to-end VSR systems to a specific speaker. Hence, we propose two different adaptation methods based on the conventional fine-tuning technique, the so-called Adapters. We conduct a comparative study in terms of performance while considering different deployment aspects such as training time and storage cost. Results on the Spanish LIP-RTVE database show that both methods are able to obtain recognition rates comparable to the state of the art, even when only a limited amount of training data is available. Although it incurs a deterioration in performance, the Adapters-based method presents a more scalable and efficient solution, significantly reducing the training time and storage cost by up to 80%.

**Keywords:** visual speech recognition; speaker adaptation; fine-tuning; Adapters; Spanish language; end-to-end architectures

check for updates

**Citation:** Gimeno-Gómez, D.; Martínez-Hinarejos, C.-D. Comparing Speaker Adaptation Methods for Visual Speech Recognition for Continuous Spanish. *Appl. Sci.* **2023**, *13*, 6521. https://doi.org/10.3390/app13116521

Academic Editor: José Luis Pérez Córdoba

Received: 8 May 2023
Revised: 24 May 2023
Accepted: 25 May 2023
Published: 26 May 2023

## 1. Introduction

Originally, Automatic Speech Recognition (ASR) was focused solely on acoustic cues [1,2]. Although today these auditory-based ASR systems are capable of understanding spoken language with outstanding quality [3,4], their performance deteriorates in adverse scenarios such as noisy environments [5–7]. Hence, influenced by different studies that have shown the relevance of visual cues during speech perception [8,9], the robustness of ASR systems has been enhanced by the design of audiovisual approaches [6,7,10,11]. In addition, these studies have encouraged the development of systems capable of interpreting speech by reading only the lips of the speaker. This challenging task, known as Visual Speech Recognition (VSR), has been a focus of interest during the last few decades [12]. Moreover, recognising speech without the need for acoustic stream data offers a wide range of applications, such as silent speech passwords [13], visual keyword spotting [14], or the development of silent speech interfaces that would be able to improve the lives of people who experience difficulties in producing speech [15–17].

Unprecedented advances have recently been achieved in the field of VSR thanks to the availability of large-scale audiovisual databases [18–20] and the design of end-to-end architectures [6,7,11,21,22]. Specifically, by combining the Connectionist Temporal Classification (CTC) [23] with Attention [24] paradigms, the so-called hybrid CTC/Attention architecture [25] stands as the current state-of-the-art in the field, reaching performance of around 25–30% Word Error Rate (WER) [22]. However, VSR remains an open research problem; by dispensing with the auditory sense, different challenges must be considered, e.g., visual ambiguities [26,27], the complex modeling of silence [28], the intra-personal variability among speakers [29], and different lighting conditions, as well as more technical aspects such as frame rate and image resolution [30–32]. Duchnowski et al. [33] argued that only 30% of speech information is visible, which highlights the relevance of modeling contextual relationships when addressing VSR [12].

These challenges can be alleviated when the VSR task is approached from a speaker-dependent perspective [34,35]. In fact, it has been proven that each person produces speech in a unique way [36], a finding that supports the idea that visual speech features are highly sensitive to the identity of the speaker [29]. However, although a wide range of works have studied the speaker adaptation of end-to-end systems in the field of ASR [37–40], only a few works in this regard have addressed VSR [41,42]. Although this speaker-dependent approach makes for a less demanding task, it should not be forgotten that speaker-adapted VSR systems could be helpful in a non-invasive and inconspicuous way for people who suffer from communication difficulties [15–17].

Another important aspect to highlight is that the most predominant approach when adapting a model to new languages or domains is to re-estimate all the parameters that compose it [22,43]. However, this technique, known as fine-tuning, requires estimating and maintaining a separate model for each task for which it is to be adapted, incurring additional time and storage costs. For this reason, Bapna and Firat [44] proposed the so-called Adapters as a scalable and parameter-efficient adaptation method. Although different works in ASR have been based on this method to address language- or domain-adaptation tasks [45,46], to the best of our knowledge it has not yet been explored for the field of VSR. Details on these adaptation methods can be found in Section 5.

In addition, languages other than English have recently received increasing interest in the field of VSR [20,22,47,48]. In these cases, the lack of available audiovisual resources [12] can be considered an additional challenge.

**Contributions:** Our research focuses on the adaptation of end-to-end VSR systems to a specific speaker. This study was motivated by the development of applications such as the one reflected in the scheme depicted in Figure 1. In this scheme, after the speaker has been automatically identified, a corresponding speaker-adapted VSR system is used to transcribe what the speaker is saying without the need for acoustic information. Experiments were conducted on the challenging Spanish LIP-RTVE database [49]. All the defined speaker-adapted VSR systems were based on a hybrid CTC/Attention architecture pretrained with hundreds of hours of data [22]. Thus, our key contributions are: (i) a comparative study on different speaker adaptation methods for end-to-end VSR systems; (ii) the proposal of three different adaptation strategies based on the fine-tuning technique; (iii) the use of the so-called *Adapters* [44] that, to the best of our knowledge, has not yet been explored in the field of VSR; (iv) an analysis of how these different adaptation approaches behave based on the amount of data available for estimation; and (v) a discussion of the deployment aspects, such as training time, storage cost, and real-time factors, of the best proposed methods.

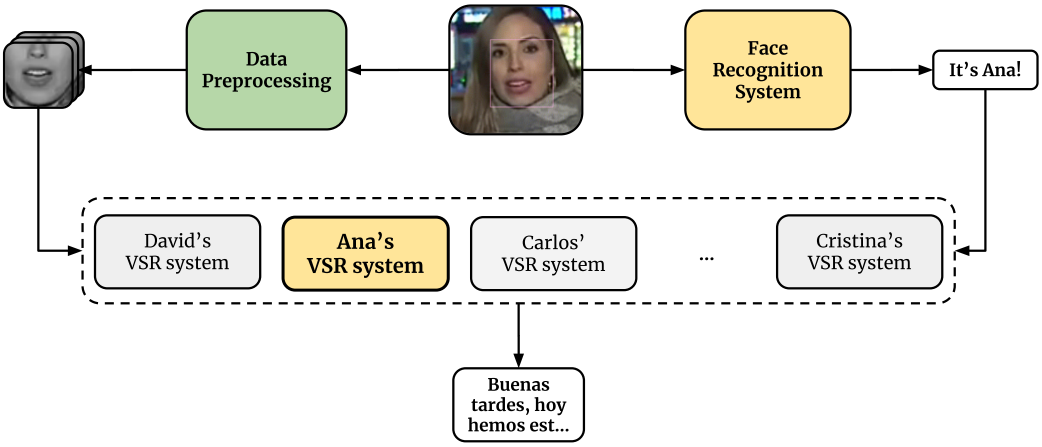

**Figure 1.** Scheme of a real application based on speaker-adapted visual speech recognition.

## 2. Related Work

**Visual Speech Recognition.** Influenced by the evolution of systems focused on auditory-based ASR, different approaches have been considered in the field [12]. Today, the current state of the art in VSR [11,21,22] has shown remarkable advances, achieving a WER of around 25–30% on the challenging LRS3-TED database [19]. As introduced above, this has been possible not only thanks to the availability of large-scale databases [18–20], but to the design of powerful end-to-end architectures [11,21,22].

**Speaker Adaptation Methods.** Although speaker adaptation has been widely studied in traditional ASR [2,50,51], in this section, due to the nature of our VSR system, we only consider those works that have addressed the problem using end-to-end architectures. A simple retraining-based adaptation was adopted in [37] to fine-tune an Attention-based system. In addition, influenced by research on conventional ASR systems [51], a hybrid CTC/Attention model was adapted by incorporating speaker identity vectors [38]. On the other hand, different works [39,40] have proposed the use of more sophisticated techniques, such as the Kullback–Leibler divergence and Linear Hidden Networks.

In our work, we decided to use a retraining-based method, also known as fine-tuning. However, as previously introduced, this approach implies estimating and maintaining a separate model for each speaker to whom it needs to be adapted. Therefore, similar to other works [44–46], we contrasted this fine-tuning technique to a more recent approach based on the so-called Adapters [44]. Although it was originally proposed in the field of neural machine translation, this parameter-efficient method has recently been applied in ASR. Thomas et al. [46] used Adapters to address language- and domain-adaptation tasks, demonstrating that, although there was a slight deterioration in terms of performance, this lightweight method presented several advantages for deployment purposes. Reporting similar conclusions, Tomanek et al. [45] focused on adaptation to atypical or accented speech while considering hundreds of different speakers. The authors found Adapters to be a feasible and scalable solution for personalized speaker-dependent models, as well as for domain-specific or dialect/accent-focused models.

However, most of these studies were conducted in the auditory-based ASR domain. Although this research describes approaches that could be adopted to any speech modality, it is noteworthy that few works have explicitly focused on speaker adaptation for VSR systems. Kandala et al. [41] defined an architecture based on the CTC paradigm [23] where, after computing visual speech features, a speaker-specific identity vector was integrated as an additional input to the decoder. Fernandez-Lopez et al. [42] approached the problem indirectly, studying how to adapt the visual front end of an audiovisual recognition system. Specifically, the authors proposed an unsupervised method that allowed an audiovisual system to be adapted when only visual data were available. However, unlike our research, these works did not address natural continuous VSR, as their experiments were evaluated on databases recorded in controlled settings. Furthermore, to the best of our knowledge, Adapters have not yet been explored in VSR.

**Spanish Visual Speech Recognition.** Although they tend to lack audiovisual resources [52], languages other than English are beginning to be considered in the field of VSR [22]. This is the case for the Spanish language, which has been the object of study on multiple occasions despite the fact that an evaluation benchmark has not yet been established. Fernandez-Lopez and Sukno [53] explored diverse approaches over the VLRF corpus [47], achieving around 70% WER in the best setting of their experiments. Ma et al. [22] designed a hybrid CTC/Attention end-to-end architecture that was fine-tuned to the Spanish language after being pre-trained with large-scale English corpora, achieving around 45% WER on the CMU-MOSEAS database [52]. As the present work focuses on the Spanish language, our VSR systems are based on this CTC/Attention architecture, for which the details can be found in Section 4.

Regarding our previous work, we first focused our research on studying different visual speech representations [48]. Subsequently, we collected the challenging LIP-RTVE database [49], an audiovisual corpus primarily conceived to deal with the Spanish VSR task and the details of which are found in Section 3. Currently, our best results with this database have been obtained with the pretrained CTC/Attention architecture proposed by Ma et al. [22], with a WER of around 40% in the speaker-dependent partition. Considering methods based on the fine-tuning technique, we then studied the development of speaker-adapted VSR systems [54], a work that is the basis of this conference paper extension.

## 3. The LIP-RTVE Database

One of the main reasons why we chose the LIP-RTVE database [49] was because it offers a suitable support to estimate VSR systems for the Spanish language against realistic scenarios. In addition, it defines a partition for both speaker-dependent and speaker-independent scenarios, each with respective training, development, and test datasets. This challenging database was compiled from TV newscast programmes recorded at 25 fps with a resolution of 480 × 270 pixels. No restrictions were considered in data collection in terms of ability to find so-called spontaneous speech phenomena, head movements, or different lighting conditions. The corpus comprises 323 speakers, providing 13 h of data with a vocabulary size of 9308 words.

It should be noted that due to the nature of our proposed study only the speaker-dependent partition of the LIP-RTVE database was considered in our experiments. More specific details on how the database was processed to carry out our research experiments according to the proposed study are described in Section 6.1.

## 4. Model Architecture

### *4.1. Visual Speech Recogniser*

The VSR system employed in our research is based on the state-of-the-art CTC/Attention architecture proposed by Ma et al. [22], which comprises about 52 million parameters. As Figure 2 reflects, this end-to-end system is composed of different modules:

- **Visual Front-end:** Consists of a 2D ResNet-18 [55] in which, in order to deal with temporal relationships in the data, the first layer has been replaced by a 3D convolutional layer with a kernel size of 7 × 7 pixels and a receptive field of five frames.
- **Conformer Encoder:** Defined as a 12-layer encoder based on Conformers [56], an approach explicitly designed to capture both global and local speech interactions from the previous latent representation provided by the visual front-end.
- **Hybrid CTC/Attention Decoder:** Composed of a six-layer Transformer decoder [24] based on the Cross-Entropy (CE) loss and a linear layer as the CTC-based decoding branch [23]. Both decoders are followed by a softmax activation layer. Details on how these loss functions are combined can be found in Section 4.3.

It should be noted that in all our experiments our VSR system was initialised using the weights publicly released by Ma et al. [22] for the Spanish language. As mentioned in Section 2, this model was able to reach around 45% WER on the Spanish partition of the CMU-MOSEAS database [52]. This performance was possible thanks to a two-step training process. First, more than 1500 h of data from different English corpora were used in a pretraining stage. Then, the model was fine-tuned using the Spanish partition of the Multilingual-TEDx [57] and CMU-MOSEAS databases [52].

### *4.2. Language Model*

As Figure 2 suggests, a character-level language model (LM) composed of six Transformer decoder layers [24] was integrated during inference in a shallow fusion manner. It comprises about 50 million parameters.

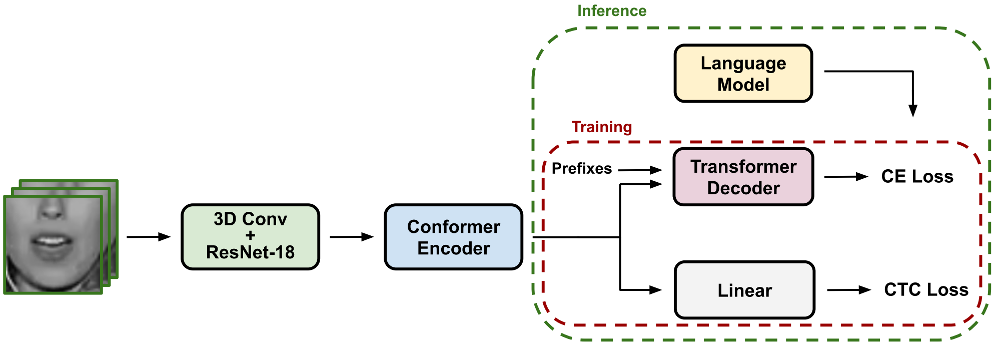

**Figure 2.** Architecture of the end-to-end VSR model based on the CTC/Attention paradigm. CE and CTC refer to Cross-Entropy and Connectionist Temporal Classification, respectively.

Similar to the VSR system, the LM used throughout all our experiments corresponds to that publicly released by Ma et al. [22] for the Spanish language. Concretely, this LM was estimated over a total of 192 million characters collected from the Spanish Multilingual TEDx [57], Common Voice [58], and Multilingual LibriSpeech [59] corpora. It should be noted that the LM was not re-estimated in any of our experiments; although experiments were carried out to adapt the LM to the newscasts domain of the LIP-RTVE database, no significant differences were observed with respect to the pretrained LM provided by Ma et al. [22]).

### *4.3. Loss Function*

The hybrid CTC/Attention architecture is an approach that has led to advances in speech processing [22,25]. By combining both paradigms, the model is able to adopt both the Markov assumptions of CTC (an aspect that is in harmony with the nature of speech) and the flexibility of the non-sequential alignments provided by the Attention-based decoder. Thus, the loss function is computed as follows:

$$\mathcal{L} = \alpha \log p_{ctc}(\mathbf{y}|\mathbf{x}) + (1-\alpha) \log p_{attn}(\mathbf{y}|\mathbf{x}) \quad (1)$$

where $p_{ctc}$ and $p_{attn}$ denote the CTC and the Attention posteriors, respectively. In both terms, $\mathbf{x}$ and $\mathbf{y}$ refer to the input visual stream and its corresponding character-level target, respectively. The $\alpha$ weight is introduced to balance the relative influence of each decoder. In our work, following the indications stated in [22], $\alpha$ was set to 0.1 in all our experiments.

### *4.4. Inference*

During inference, the VSR system and the Transformer-based LM were integrated in a shallow fusion manner, as reflected by

$$S = \lambda S_{ctc} + (1-\lambda) S_{attn} + \beta S_{lm} \quad (2)$$

where $S_{ctc}$ and $S_{attn}$ are the scores of the CTC and the Attention decoder, respectively, $\lambda$ is their corresponding relative weight, and $\beta$ and $S_{lm}$ refer to the LM decoding influence weight and the LM score, respectively. In our experiments, $\lambda$ and $\beta$ were set to 0.1 and 0.4, respectively. Then, a beam search algorithm was applied with a beam size of 10. All these hyperparameters were set according to [22] except for the beam size, which had to be reduced from 40 to 10 due to memory constraints (see Section 6.4).

## 5. Speaker Adaptation Methods

Our research aimed to study the feasibility of developing speaker-adapted systems for the VSR task. By considering the speaker-dependent scenario defined for the LIP-RTVE database, we analysed how estimating specialised end-to-end VSR systems for a specific person affected the quality of speech recognition. Specifically, two different adaptation methods are explored in this work: the conventional fine-tuning technique (Section 5.1), and the use of the so-called *Adapters* (Section 5.2). Details on the training settings used when applying these adaptation methods are found in Section 6.4.

### *5.1. Fine-Tuning*

When adapting a model (usually trained for a general task using large-scale databases) to new languages or domains, the most predominant approach is to re-estimate all the parameters that compose the model. This method, known as fine-tuning, has been widely explored in the literature [22,43]. In our work, we propose three different fine-tuning strategies:

- **Multi-Speaker Training (MST).** The VSR system is re-estimated using the training data of the entire speaker-dependent partition of the database. This strategy, as discussed throughout the paper, can be considered as a task adaptation.
- **Speaker-Adapted Training (SAT).** In this case, only data corresponding to a specific speaker is considered when fine-tuning the VSR system.
- **Two-Step Speaker-Adapted Training (TS-SAT).** As its name suggests, this strategy consists of two fine-tuning steps. First, following the MST strategy, the entire partition is used to re-estimate the VSR system and achieve task adaptation. Afterwards, the system is fine-tuned to a specific speaker using the corresponding data.

In this way, by comparing these proposed strategies, we were able to study how a VSR system can generalize common patterns from different speakers or, on the contrary, evaluate to what extent it is capable of adapting to a specific speaker.

### *5.2. Use of Adapters*

The fine-tuning technique implies estimating and maintaining a separate model for each speaker for whom the recognition system needs to be adapted. In a real application, such as the one suggested in Figure 1, this could be unfeasible in terms of training time and storage. For this reason, Bapna and Firat [44] proposed the so-called Adapters as a scalable and parameter-efficient adaptation method. As mentioned in Section 2, different works have used Adapters in the field of ASR to address language and domain adaptation tasks [45,46]. However, to the best of our knowledge, this approach has not yet been explored in VSR.

As depicted in Figure 3, this method consists of injecting an Adapter module at the end of every layer that composes both the encoder and decoder blocks. Each Adapter module is composed of two linear layers in a bottleneck manner, for which the inner dimension is the only tunable hyperparameter. The input of the Adapter is normalised, and a nonlinear ReLU activation layer follows the resulting inner projection. In addition, a residual connection is incorporated. Thus, by only estimating these injected Adapter modules (i.e., the rest of the model is kept frozen), the general VSR system can be adapted to a specific speaker. It should be noted that no parameter-sharing techniques are applied, i.e., each Adapter module injected in the architecture is independent.

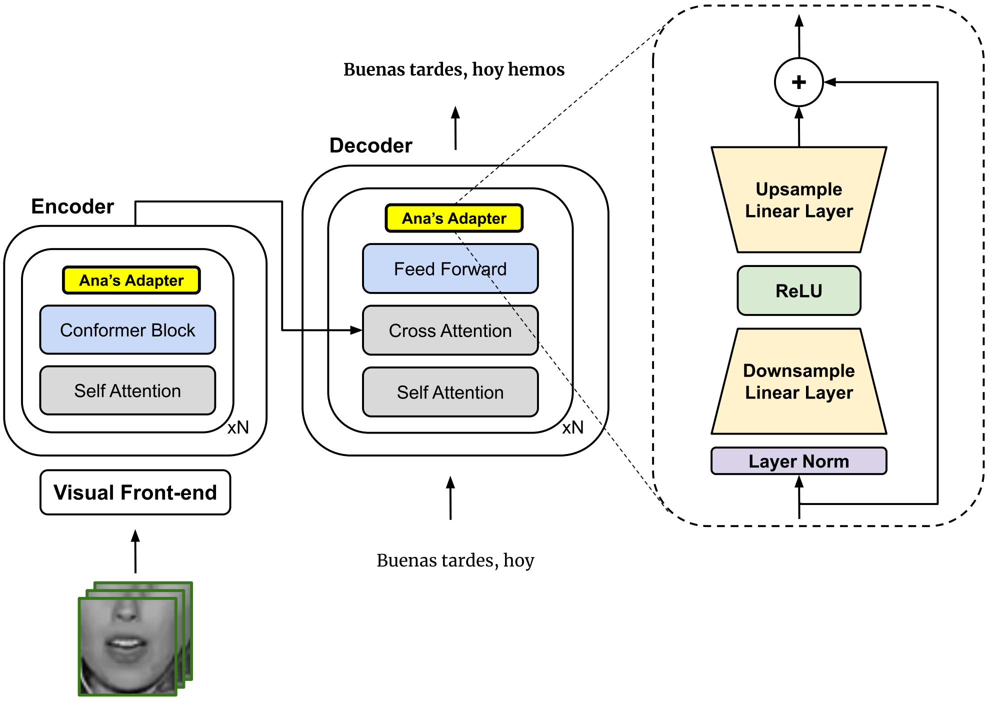

**Figure 3.** Scheme of how adapters are injected to adapt an end-to-end VSR system to a specific speaker. It should be noted that only the Adapter modules are estimated during the adapting process, while the rest of the model is kept frozen.

In our experiments, as described in Section 6.2, we explored different Adapter sizes, i.e., the dimension of their inner projections. Depending on the size of this projection, the number of learnable parameters to be estimated varies. Table 1 shows the percentage of learnable parameters with respect to the entire end-to-end system (52 million parameters; see Section 4.1) for each adapter size considered in our experiments. As can be observed, in the worst case it is only necessary to estimate about 4% of the parameters that compose the entire end-to-end system, making for a significant reduction in training time and storage cost, as discussed in Section 7.3.

**Table 1.** Percentage (%) of learnable parameters with respect to the entire end-to-end system (see Section 4.1) for each adapter size considered in our experiments. Adapter sizes are represented by the hidden units that compose its inner projection. Percentages were computed considering all the injected Adapter modules, i.e., one Adapter for each layer of the encoder and decoder (see Figure 3).

| **Adapter Size** | 8 | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| **Learnable Parameters** † | 0.15% | 0.27% | 0.52% | 1.02% | 2.02% | 4.02% |

† Percentages w.r.t. the entire end-to-end system, which comprises around 52.5 million parameters.

It should be noted that our Adapters-based VSR systems were initialised using the weights of the MST-based model in all our experiments. Although different experiments were conducted using the entire speaker-dependent partition training set, no acceptable recognition rates were obtained when adapting the Spanish VSR model provided by Ma et al. [22] to the LIP-RTVE task using Adapters.

## 6. Experimental Setup

### *6.1. Datasets*

As introduced in Section 3, due to the nature of our proposed study we only considered the speaker-dependent partition of the LIP-RTVE database in our experiments. However, in order to estimate our speaker-adapted VSR systems and adequately interpret the obtained results, the twenty most talkative speakers were selected. Therefore, depending on the adaptation method appl (see Section 6.3), two different partitions used to estimate our VSR models in our work can be distinguished:

- **The entire speaker-dependent partition**. This corresponds to the official speaker-dependent partition of the LIP-RTVE database. The training and development sets offer around 9 h and 2 h of data, respectively. As suggested in Section 5, using one of these data sets to estimate the model can be seen as a task adaptation. Concretely, this partition is used when applying the fine-tuning MST strategy, providing a model used as initialisation for the TS-SAT strategy and the Adapters-based adaptation method.
- **The speaker-specific partition.** Only the data corresponding to the aforementioned twenty speakers are considered in this case. On average, the training and development sets offer around 15 min and 4 min of data per speaker, respectively. As suggested in Section 5, when seeking to estimate a VSR system adapted to a specific speaker, only data associated with that speaker is used. Concretely, this partition is considered in all of our proposed methods with the exception of the fine-tuning MST strategy.

It should be noted that, regardless of whether a dataset from one partition or the other is used to estimate the VSR models, the speaker-specific test set (composed of around 4 min of data per speaker) was used to assess the effectiveness of all the proposed adaptation methods. In addition, as described in Section 6.2, the development set of each partition was used to analyse how the different proposed adaptation methods behaved when only a significantly reduced amount of data was available for their estimation.

### 6.2. Experiments

Different experiments regarding the speaker adaptation of end-to-end VSR systems were carried out as part of this work. The results are reported and discussed in Section 7.

**Face Recognition.** First, experiments were conducted on the automatic face recognition system that a real application would need, as depicted in Figure 1. Specifically, we explored a traditional approach based on Principal Component Analysis for feature extraction (representation known in the literature as *eigenfaces* [60]) and Support Vector Machine for classification. Results of around 99% accuracy were obtained using this approach, possibly due to the limited number of speakers considered in our case study (twenty speakers). Henceforth, we assume a perfect face recognition system throughout the rest of experiments presented in this paper. In addition, it should be noted that LIP-RTVE includes different head poses and occluded faces, as well as profile views, due to its in-the-wild nature. These type of situations were not controlled, i.e., any restrictions were considered during estimation using the face recogniser.

**Independent Speaker-Adaptation Studies.** First, each adaptation method proposed in this work was independently analysed. Section 7.1 focuses on comparing the three different fine-tuning strategies described in Section 5.1. Section 7.2 discusses the Adapters-based method's effectiveness, comparing different Adapter sizes in terms of performance. In both cases, it is studied how all of these methods behave depending on the amount of data available for their estimation, using either the training or development set of the corresponding data partition. Using the development set, we study the robustness of each proposed adaptation method against data-scarcity scenarios, as dealing with this dataset means, in average terms, using an amount of data 4.5 times less than when using the training set [49].

**Overall Analysis.** In this case, the best approaches of each proposed adaptation method are compared only in terms of performance and in terms of different deployment aspects, such as training time and storage costs.

### 6.3. Methodology

Before reporting and discussing our results (see Section 7), different aspects of how we carried out our experiments must be clarified:

- Both the MST- and SAT-based models were initially pretrained using the weights publicly released by Ma et al. [22] for the Spanish language (see Section 4.1).
- The MST-based system was fine-tuned using the entire speaker-dependent partition of the LIP-RTVE database, which is considered our task-adapted VSR system.
- An SAT-based system using the corresponding speaker-specific training data was independently estimated for each speaker considered in our study, i.e., twenty SAT-based systems were defined.
- When applying the TS-SAT strategy, the previously estimated MST-based system was used for initialisation, which can be considered an adaptation of the model to the task. Then, we followed the same fine-tuning scheme described for the SAT strategy to obtain a TS-SAT-based system for each speaker.
- Regarding the Adapters-based method, different experiments were conducted using the entire speaker-dependent partition training set. However, no acceptable recognition rates were obtained when adapting the Spanish VSR model provided by Ma et al. [22] to the LIP-RTVE task using Adapters. Therefore, we decided to use the MST-based model for initialisation. Similar to the TS-SAT strategy, this can be seen as starting from a task-adapted model. Then, speaker-specific adapter modules were injected and estimated for each speaker considered in our study while keeping the original VSR system backbone frozen, as described in Section 5.2.
- Experiments were conducted using either the training or development set for adapting. However, it should be noted that TS-SAT-based systems were always based on the MST-based system estimated with the training set, while in the second step the training or development set was used depending on the experiment. Conversely, when applying the Adapters-based method, depending on whether we used the speaker-specific training or development set, for initialisation we used the MST-based model estimated with the corresponding dataset from the entire speaker-dependent partition.
- All these VSR systems, as suggested in Section 6.1, were evaluated on the test set corresponding to each of the speakers selected in our study. The MST-based system was the same regardless of the evaluated speaker. Conversely, for the rest of the methods, the corresponding speaker-adapted system was used in each case.
- The LM used in all the tests was the one described in Section 4.2. Unlike the VSR system, it was not re-estimated in any of our experiments.

### *6.4. Implementation Details*

All of our VSR systems and the training and inference processes were implemented using the open-source ESPNet toolkit [61]. Experiments were conducted on a GeForce RTX 2080 GPU with 8GB memory.

**Data Pre-processing.** Similar to [22], using the state-of-the-art RetinaFace face detector [62] and Face Alignment Network [63], grayscale bounding boxes centered on the speaker's mouth were cropped to form images of 96 × 96 pixels. These Regions Of Interest (ROIs) covered the mouth as well as the complete jaw and cheeks of the speaker, a wider area that has shown benefits when addresing the VSR task [64].

**Data Augmentation.** This data augmentation process was influenced by [22]. First, after the ROIs were normalised with respect to the overall mean and variance of the training set, a random cropping of 88 × 88 pixels was applied. Then, additional techniques were considered, such as horizontal flipping and time masking [22], a method inspired by the work carried out in the field of auditory-based ASR [65].

**Pre-Training.** As mentioned in Section 4, both the LM and all the VSR systems considered in our experiments were pretrained or initialised using the weights of the models publicly released by Ma et al. [22] for the Spanish language. It should be noted that the LM was not re-estimated in any of our experiments.

**Training Setup**. The pretrained VSR systems were re-estimated during five epochs using the AdamW optimiser [66] and a linear one-cycle scheduler [67]. In all the cases, the learning rate was set to 5 × $10^{-4}$. This optimum was found after carrying out preliminary experiments, as described in Section 7. Regarding the CTC/Attention loss, the $\alpha$ weight specified in Equation (1) was set to 0.1. For those experiments where the proposed fine-tuning strategies were applied, the batch size was set to only 1 sample due to our GPU memory constraints. It should be noted here that we explored the accumulating gradient strategy [68]; however, no significant differences were found. We argue that despite applying this technique, the normalisation layers were affected by the actual reduced batch size. When using the light-weight Adapters-based method, the significantly reduced number of learnable parameters allowed us to use a batch size of four samples.

**Inference Setup.** As Equation (2) reflects, different weights were used to model the influence of each component during the inference process. In all cases, the $\beta$ weight was set to 0.4. The word insertion penalty was set to 0.0, while due to memory limitations the beam size was set to 10. Regarding the CTC/Attention balance, the $\lambda$ weight was set to 0.1. As described in Section 4.4, all of these hyperparameters were set according to [22].

**Evaluation Metric.** All the results reported in our experiments were evaluated using the well-known Word Error Rate (WER) with 95% confidence intervals obtained by the bootstrap method, as described in [69].

## 7. Results & Discussion

Our first experiments were focused on the training setup. Therefore, several learning rates were explored until the optimal value specified in Section 6.4 was reached. This optimum was the same for all the proposed adaptation methods. Moreover, we studied dispensing with the scheduler, concluding that its absence slowed down the learning process and worsened the VSR system's performance. Then, after the best training settings had been determined, the experiments described in Section 6.2 were carried out. A discussion of the results obtained with the fine-tuning adaptation method can be found in Section 7.1. Similarly, Section 7.2 discusses the results obtained when using the Adapters-based adaptation method. Finally, an overall comparison of both adaptation methods is presented in Section 7.3.

### *7.1. Fine-Tuning*

As described in Section 5.1, three different fine-tuning strategies are proposed in our work. Table 2 reflects a comparison between these strategies in general terms. In addition, as described in Section 6.2, we explored the use of different datasets when estimating these fine-tuning strategies. Hence, we studied how robust each proposed strategy is against data scarcity scenarios. Taking into account all these aspects and the results reported in Table 2, we following conclusions can be inferred:

- Irrespective of the dataset used for estimation, it can be observe that the MST method is significantly outperformed by the rest of the proposed strategies, a fact that supports the effectiveness of our fine-tuning speaker adaptation approaches.
- When the training set was used, the MST-based model provided a considerable quality of speech recognition. This result could mean that the end-to-end architecture employed in our experiments was able to generalise common patterns across speakers when addressing VSR.
- Regarding the amount of data used during the fine-tuning process, the results reflect a drastic deterioration of system performance when the development set was used. However, this deterioration was noticeably lower when the TS-SAT strategy was applied, showing that this approach could be more robust against situations where a given speaker presents data scarcity.
- The TS-SAT strategy stands as the best option when addressing speaker adaptation. This fact supports the idea that a two-step fine-tuning process in which the model is first adapted to the general task could benefit the final adaptation of the VSR system to a specific speaker.

- Results comparable to the current state of the art were obtained. Moreover, our findings suggest that the fine-tuning method employed in our experiments is capable of adapting VSR end-to-end architectures in a small number of epochs, even when only a limited amount of data is available.

In addition, the average performance of the VSR system without fine-tuning (i.e., using the weights published by Ma et al. [22] for the Spanish language) provided around 78% WER on the twenty speakers considered in our experiments. For reference, it should be noted that when using the entire nine-hour training set from the LIP-RTVE database, about 60% WER was obtained for the speaker-independent scenario. However, as mentioned in Section 2, the current state of the art in the field is around 25–30% WER for the English LRS3-TED database [19], a challenging speaker-independent database widely studied in the field [11,21,22]. Therefore, although we were able to obtain recognition rates of considerable quality, all these aspects suggest that further research should be considered.

**Table 2.** System performance (WER test set) in average terms depending on the dataset used to estimate the VSR system (see Section 6.1) for each proposed adaptation strategy based on fine-tuning. DEV and TRAIN refer to the development and training datasets, respectively.

| Fine-Tuning Strategy | Data Set | |
|---|---|---|
| | DEV | TRAIN |
| **MST** | 59.6 ± 1.3 | 36.4 ± 1.3 |
| **SAT** | 52.2 ± 1.4 | 29.1 ± 1.5 |
| **TS-SAT** | 32.8 ± 1.3 | 24.9 ± 1.4 |

In addition, considering only those experiments where the training set was used, we evaluated each strategy for each of the speakers selected in our study, as shown in Figure 4. Similar conclusions to those mentioned above can be inferred from these results. Nonetheless, it is noteworthy that regardless of the strategy being applied, the VSR system provides remarkably different recognition rates depending on the speaker evaluated. Hence, a study was conducted to find reasons that could explain this behaviour. Several statistics were computed for each of these speakers, such as the number of words per utterance, the perplexity of the LM in the test samples, and the number of training seconds. Then, we analysed how each of the statistics varied as the word error rate increased. However, we were not able to identify any trends or patterns from these data. Therefore, we can say that these experiments suggest that the reason why VSR systems behaved in this way could be related to aspects that are difficult to model, such as better vocalizations or certain oral physiognomies that reflect more adequate speech articulations.

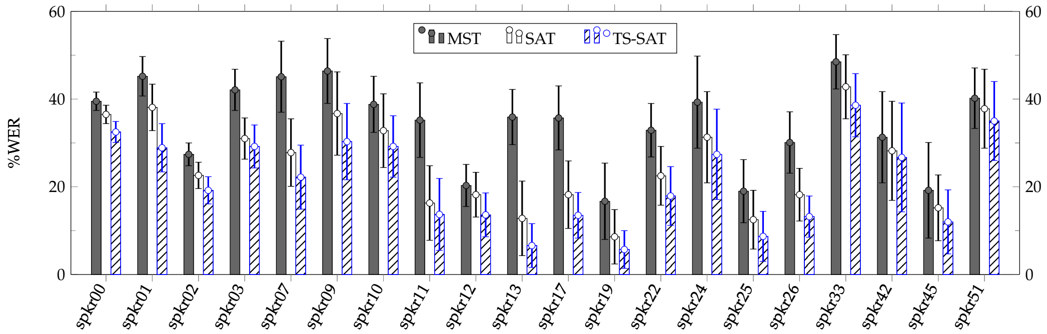

**Figure 4.** Comparison of the proposed fine-tuning adaptation strategies. System performance (WER) with 95% confidence intervals is reported for each speaker considered in the study. Only experiments that used the training dataset to estimate the VSR systems are considered.

### *7.2. Use of Adapters*

Using Adapters means estimating a set of speaker-specific modules previously injected in multiple parts of the model architecture while keeping the original model backbone frozen [44]. As suggested in Section 5.2, one of our purposes was to study different Adapter sizes while considering the percentage of learnable parameters that each one implied (see Table 1). Figure 5 depicts this analysis when using the training set and the development set. According to these results, the following conclusions can be inferred:

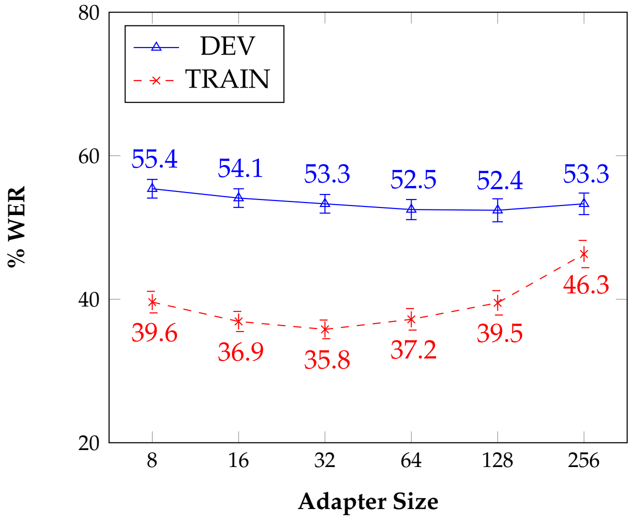

**Figure 5.** Performance comparison (WER test set) of different adapter's sizes with respect to our proposed adaptation strategy based on the use of adapters and depending on the dataset used to estimate the VSR system. The percentage of learnable parameters offered by each adapter size is specified in Table 1. DEV and TRAIN refer to the development and training datasets, respectively.

- In data scarcity scenarios, i.e., when using the development set for estimation, the Adapters-based adaptation method was able to significantly improve the MST-based model performance reported in Table 2. However, no significant differences were observed when using different Adapter sizes.
- When the training set was used, more remarkable differences could be observed regarding the Adapter size, leading to the conclusion that 32 hidden units (estimating around 0.52% of parameters with respect to the entire end-to-end model; see Table 1) provided the best recognition rate. However, the Adapters-based method did not outperform the MST-based model in this scenario, as Table 2 confirms.

Nonetheless, it should be noted that by considering the best Adapter size for each speaker we were able to obtain recognition rates of around 50.8 ± 1.4 and 32.9 ± 1.3% WER when using the development and training datasets, respectively. These results present significant differences with respect to the MST-based system (see Table 2), which shows the effectiveness of the Adapters-based method when adapting VSR systems to a specific speaker. In addition, the fact that each speaker required a tailored Adapter size could be related to the behaviour observed in Figure 4, where all the fine-tuning strategies obtained different recognition rates depending on the speaker evaluated. Furthermore, this finding highlights the flexibility that Adapters-based adaptation methods can provide.

### *7.3. Overall Analysis*

In this section, we compare the best approach of each adaptation method proposed in our work in terms of performance as well as in terms of different deployment aspects such as training time and storage costs, as reflected in Table 3. We consider the TS-SAT strategy for the fine-tuning method, as it provided the best recognition rates (see Table 2). Regarding the Adapters-based method, for each speaker we consider the adapter size that provided the best recognition rate, an approach we refer as *Best Adapter*. In both cases, we only consider experiments where the training dataset was used. Hence, we are able to identify why or in which situations one adaptation method might be more suitable.

**Table 3.** Overall comparison of the best approach for each proposed adaptation method. Only experiments where the training dataset was used are considered. *Best Adapter* refers to the approach for each considered speaker where the adapter size with the best recognition rate was used. Training times are reported in minutes (min). Storage costs are reported in Gigabytes (GB). All reported details were computed considering the estimation of twenty speaker-adapted VSR systems.

| Method | % WER | Training Time † | Storage |
|---|---|---|---|
| **TS-SAT** | 24.9 ± 1.4 | 78.1 | 4.2 |
| **Best Adapter** | 32.9 ± 1.3 | 13.6 | 0.611 |

† It should be noted that both methods are based on the MST model (see Section 5.1), which implies a training process of approximately 2 h.

First, it should be noted that regardless of the adaptation method, the end-to-end architecture employed in our experiments presented a real-time factor of around 0.75. Regarding the performance of the VSR system, it is true that the Adapter-based method implies a significant deterioration in the quality of recognition rates. However, similar to the findings discussed in [45,46], this method shows a dramatic training time reduction of about 80% due to the reduced number of learnable parameters. In addition, this method offers valuable properties in terms of storage cost. For our case study, instead of maintaining twenty full speaker-adapted systems (around 4.2 GB), it is only necessary to maintain the original end-to-end architecture adapted to the task and the Adapters modules estimated for each speaker under consideration. All these aspects support that Adapters can be considered a scalable and efficient solution for deployment purposes when adapting end-to-end VSR systems to specific speakers.

## 8. Conclusions and Future Work

In this work, we have addressed the adaptation of end-to-end VSR systems to specific speakers. We propose two adaptation methods, one based on the use of the so-called Adapters approach and the other based on the conventional fine-tuning technique, with three different strategies defined in total. Then, we conducted a thorough comparative study, reaching the conclusion that, while both approaches provide recognition rates comparable to the state of the art, the Adapters-based method presents a more scalable and efficient solution for deployment purposes. Although using this method implies a deterioration in terms of performance, in our case study it significantly reduced the training time and storage cost by up to 80% with respect to the full-model fine-tuning technique.

For future work, as suggested in [45], we are considering further research regarding the Adapters-based method and ways to improve its performance. For instance, by exploiting the flexibility that this method offers, it might be feasible to study the use of Adapter modules with different inner projection sizes depending on the layer where they were injected. In addition, taking into account the deployment of these VSR systems for real applications, we intend to consider focusing our research on reducing the real-time factor when decoding speech, as well as on exploring streaming speech recognition. The former could be addressed by the design of non-autoregressive architectures [23,70], while for the latter, a consideration could be the use of Recurrent Neural Network Transducers [71,72].

**Author Contributions:** Conceptualization, D.G.-G. and C.-D.M.-H.; methodology, D.G.-G. and C.-D.M.-H.; software, D.G.-G.; validation, D.G.-G. and C.-D.M.-H.; formal analysis, D.G.-G. and C.-D.M.-H.; investigation, D.G.-G. and C.-D.M.-H.; resources, C.-D.M.-H.; data curation, D.G.-G. and C.-D.M.-H.; writing—original draft preparation, D.G.-G. and C.-D.M.-H.; writing—review and editing, D.G.-G. and C.-D.M.-H.; visualization, D.G.-G. and C.-D.M.-H.; supervision, C.-D.M.-H.;

project administration, C.-D.M.-H.; funding acquisition, C.-D.M.-H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was partially supported by the Grant CIACIF/2021/295 funded by Generalitat Valenciana and by the Grant PID2021-124719OB-I00 under the LLEER (PID2021-124719OB-100) project funded by MCIN/AEI/10.13039/501100011033/ and by ERDF EU, “A way of making Europe”.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

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

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| ASR | Automatic Speech Recognition |
| CE | Cross-Entropy |
| CTC | Connectionist Temporal Classification |
| LM | Language Model |
| MST | Multi-Speaker Training |
| ROI | Region Of Interest |
| SAT | Speaker-Adapted Training |
| TS-SAT | Two-Step Speaker-Adapted Training |
| VSR | Visual Speech Recognition |
| WER | Word Error Rate |

## References

1. Juang, B.H.; Rabiner, L.R. Hidden Markov models for speech recognition. *Technometrics* **1991**, *33*, 251–272. [CrossRef]
2. Gales, M.; Young, S. *The Application of Hidden Markov Models in Speech Recognition*; Now Publishers Inc.: Delft, The Netherlands: 2008.
3. 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 ICASSP, Shanghai, China, 20–25 March 2016; pp. 4960–4964.
4. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust speech recognition via large-scale weak supervision. *arXiv* **2022**, arXiv:2212.04356.
5. Juang, B. Speech recognition in adverse environments. *Comput. Speech Lang.* **1991**, *5*, 275–294. [CrossRef]
6. Afouras, T.; Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Deep audio-visual speech recognition. *Trans. PAMI* **2018**, *44*, 8717–8727. [CrossRef] [PubMed]
7. Ma, P.; Petridis, S.; Pantic, M. End-To-End Audio-Visual Speech Recognition with Conformers. In Proceedings of the ICASSP, Toronto, ON, Canada, 6–11 June 2021; pp. 7613–7617.
8. McGurk, H.; MacDonald, J. Hearing lips and seeing voices. *Nature* **1976**, *264*, 746–748. [CrossRef] [PubMed]
9. Besle, J.; Fort, A.; Delpuech, C.; Giard, M.H. Bimodal speech: Early suppressive visual effects in human auditory cortex. *Eur. J. Neurosci.* **2004**, *20*, 2225–2234. [CrossRef] [PubMed]
10. Potamianos, G.; Neti, C.; Gravier, G.; Garg, A.; Senior, A. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* **2003**, *91*, 1306–1326. [CrossRef]
11. Shi, B.; Hsu, W.N.; Lakhotia, K.; Mohamed, A. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv* **2022**, arXiv:2201.02184.
12. Fernandez-Lopez, A.; Sukno, F. Survey on automatic lip-reading in the era of deep learning. *Image Vis. Comput.* **2018**, *78*, 53–72. [CrossRef]
13. Ezz, M.; Mostafa, A.M.; Nasr, A.A. A Silent Password Recognition Framework Based on Lip Analysis. *IEEE Access* **2020**, *8*, 55354–55371. [CrossRef]
14. Stafylakis, T.; Tzimiropoulos, G. Zero-shot keyword spotting for visual speech recognition in-the-wild. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018; pp. 513–529.
15. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. [CrossRef]
16. González-López, J.A.; Gómez-Alanís, A.; Martín Doñas, J.M.; Pérez-Córdoba, J.L.; Gómez, A.M. Silent Speech Interfaces for Speech Restoration: A Review. *IEEE Access* **2020**, *8*, 177995–178021. [CrossRef]
17. Matsui, K.; Fukuyama, K.; Nakatoh, Y.; Kato, Y. Speech Enhancement System Using Lip-reading. In Proceedings of the IICAIET, Kota Kinabalu, Malaysia, 26–27 September 2020; pp. 1–5.

18. Chung, J.; Senior, A.; Vinyals, O.; Zisserman, A. Lip reading sentences in the wild. In IEEE Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 3444–3453.
19. Afouras, T.; Chung, J.; Zisserman, A. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv* **2018**, arXiv:1809.00496.
20. Zhao, Y.; Xu, R.; Song, M. A cascade sequence-to-sequence model for chinese mandarin lip reading. In Proceedings of the ACM Multimedia Asia, Beijing, China, 16–18 December 2019; pp. 1–6.
21. Prajwal, K.; Afouras, T.; Zisserman, A. Sub-word level lip reading with visual attention. In Proceedings of the CVPR, New Orleans, LA, USA, 19–24 June 2022; pp. 5162–5172.
22. Ma, P.; Petridis, S.; Pantic, M. Visual Speech Recognition for Multiple Languages in the Wild. *Nat. Mach. Intell.* **2022**, *4*, 930–939. [CrossRef]
23. Graves, A.; Fernández, S.; Gómez, F.; Schmidhuber, J. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In Proceedings of the 23rd ICML, Pittsburgh, PA, USA, 25–29 June 2006; pp. 369–376.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gómez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. 2017. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 24 May 2023).
25. Watanabe, S.; Hori, T.; Kim, S.; Hershey, J.; Hayashi, T. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1240–1253. [CrossRef]
26. Bear, H.; Harvey, R.; Theobald, B.; Lan, Y. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In Proceedings of the International Symposium on Visual Computing, Las Vegas, NV, USA, 8–10 December 2014; pp. 230–239.
27. Fernández-López, A.; Sukno, F. Optimizing Phoneme-to-Viseme Mapping for Continuous Lip-Reading in Spanish. In Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics, Porto, Portugal, 27 February–1 March 2017; pp. 305–328.
28. Thangthai, K. Computer Lipreading via Hybrid Deep Neural Network Hidden Markov Models. Ph.D. Thesis, University of East Anglia, Norwich, UK, 2018.
29. Cox, S.J.; Harvey, R.W.; Lan, Y.; Newman, J.L.; Theobald, B.J. The challenge of multispeaker lip-reading. In Proceedings of the AVSP, Queensland, Australia, 26–29 September 2008; pp. 179–184.
30. Bear, H.; Harvey, R. Decoding visemes: Improving machine lip-reading. In Proceedings of the ICASSP, Shanghai, China, 20–25 March 2016; pp. 2009–2013.
31. Bear, H.; Harvey, R.; Theobald, B.; Lan, Y. Resolution limits on visual speech recognition. In Proceedings of the ICIP. IEEE, Paris, France, 27–30 October 2014; pp. 1371–1375.
32. Dungan, L.; Karaali, A.; Harte, N. The Impact of Reduced Video Quality on Visual Speech Recognition. In Proceedings of the ICIP, Athens, Greece, 7–10 October 2018; pp. 2560–2564.
33. Duchnowski, P.; Lum, D.; Krause, J.; Sexton, M.; Bratakos, M.; Braida, L. Development of speechreading supplements based on automatic speech recognition. *IEEE Trans. Biomed. Eng.* **2000**, *47*, 487–496. [CrossRef] [PubMed]
34. Thangthai, K.; Harvey, R. Improving Computer Lipreading via DNN Sequence Discriminative Training Techniques. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 3657–3661.
35. Assael, Y.; Shillingford, B.; Whiteson, S.; Freitas, N. LipNet: Sentence-level Lipreading. *arXiv* **2016**, arXiv:1611.01599.
36. Leung, K.Y.; Mak, M.W.; Kung, S.Y. Articulatory feature-based conditional pronunciation modeling for speaker verification. In Proceedings of the Interspeech, Jeju Island, Republic of Korea, 4–8 October 2004; pp. 2597–2600.
37. Ochiai, T.; Watanabe, S.; Katagiri, S.; Hori, T.; Hershey, J. Speaker Adaptation for Multichannel End-to-End Speech Recognition. In Proceedings of the ICASSP, Calgary, AB, Canada, 15–20 April 2018; pp. 6707–6711.
38. Delcroix, M.; Watanabe, S.; Ogawa, A.; Karita, S.; Nakatani, T. Auxiliary Feature Based Adaptation of End-to-end ASR Systems. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2444–2448.
39. Weninger, F.; Andrés-Ferrer, J.; Li, X.; Zhan, P. Listen, Attend, Spell and Adapt: Speaker Adapted Sequence-to-Sequence ASR. In Proceedings of the Interspeech 2019, Graz, Austria, 15–19 September 2019; pp. 3805–3809.
40. Li, K.; Li, J.; Zhao, Y.; Kumar, K.; Gong, Y. Speaker Adaptation for End-to-End CTC Models. In Proceedings of the IEEE SLT, Athens, Greece, 18–21 December 2018; pp. 542–549.
41. Kandala, P.; Thanda, A.; Margam, D.; Aralikatti, R.; Sharma, T.; Roy, S.; Venkatesan, S. Speaker Adaptation for Lip-Reading Using Visual Identity Vectors. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2758–2762.
42. Fernández-López, A.; Karaali, A.; Harte, N.; Sukno, F. Cogans For Unsupervised Visual Speech Adaptation To New Speakers. In Proceedings of the ICASSP 2020, Barcelona, Spain, 4–8 May 2020; pp. 6294–6298.
43. Yang, H.; Zhang, M.; Tao, S.; Ma, M.; Qin, Y. Chinese ASR and NER Improvement Based on Whisper Fine-Tuning. In Proceedings of the ICACT, Pyeongchang, Republic of Korea, 19–22 February 2023; pp. 213–217.
44. Bapna, A.; Firat, O. Simple, Scalable Adaptation for Neural Machine Translation. In Proceedings of the EMNLP-IJCNLP. ACL, Hong Kong, China, 3–7 November 2019; pp. 1538–1548.
45. Tomanek, K.; Zayats, V.; Padfield, D.; Vaillancourt, K.; Biadsy, F. Residual Adapters for Parameter-Efficient ASR Adaptation to Atypcal and Accented Speech. In Proceedings of the EMNLP, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 6751–6760.
46. Thomas, B.; Kessler, S.; Karout, S. Efficient Adapter Transfer of Self-Supervised Speech Models for Automatic Speech Recognition. In Proceedings of the ICASSP, Singapore, 22–27 May 2022; pp. 7102–7106.

47. Fernández-López, A.; Martínez, O.; Sukno, F. Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In Proceedings of the 12th FG, Washington, DC, USA, 30 May–3 June 2017; pp. 208–215.
48. Gimeno-Gómez, D.; Martínez-Hinarejos, C.D. Analysis of Visual Features for Continuous Lipreading in Spanish. In Proceedings of the IberSPEECH, Valladolid, Spain, 24–25 March 2021; pp. 220–224.
49. Gimeno-Gómez, D.; Martínez-Hinarejos, C.D. LIP-RTVE: An Audiovisual Database for Continuous Spanish in the Wild. In Proceedings of the LREC, Marseille, France, 20–25 June 2022; pp. 2750–2758.
50. Neto, J.; Almeida, L.; Hochberg, M.; Martins, C.; Nunes, L.; Renals, S.; Robinson, T. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. In Proceedings of the EUROSPEECH, Madrid, Spain, 18–21 September 1995; pp. 2171–2174.
51. Saon, G.; Soltau, H.; Nahamoo, D.; Picheny, M. Speaker adaptation of neural network acoustic models using i-vectors. In Proceedings of the IEEE ASRU, Olomouc, Czech Republic, 8–12 December 2013; pp. 55–59.
52. Zadeh, A.B.; Cao, Y.; Hessner, S.; Liang, P.P.; Poria, S.; Morency, L.P. CMU-MOSEAS: A Multimodal Language Dataset for Spanish, Portuguese, German and French. In Proceedings of the EMNLP, Online, 16–20 November 2020; pp. 1801–1812.
53. Fernández-López, A.; Sukno, F. End-to-End Lip-Reading Without Large-Scale Data. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 2076–2090. [CrossRef]
54. Gimeno-Gómez, D.; Martínez-Hinarejos, C.D. Speaker-Adapted End-to-End Visual Speech Recognition for Continuous Spanish. In Proceedings of the IberSPEECH, Granada, Spain, 14–16 November 2022; pp. 41–45.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the CVPR, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
56. 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. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 5036–5040.
57. Salesky, E.; Wiesner, M.; Bremerman, J.; Cattoni, R.; Negri, M.; Turchi, M.; Oard, D.; Post, M. The Multilingual TEDx Corpus for Speech Recognition and Translation. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 3655–3659.
58. Ardila, R.; Branson, M.; Davis, K.; Kohler, M.; Meyer, J.; Henretty, M.; Morais, R.; Saunders, L.; Tyers, F.; Weber, G. Common Voice: A Massively-Multilingual Speech Corpus. In Proceedings of the LREC, Online, 20–25 June 2020; pp. 4218–4222.
59. Pratap, V.; Xu, Q.; Sriram, A.; Synnaeve, G.; Collobert, R. MLS: A Large-Scale Multilingual Dataset for Speech Research. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 2757–2761.
60. Turk, M.; Pentland, A. Face recognition using eigenfaces. In Proceedings of the CVPR, Maui, HI, USA, 3–6 June 1991; pp. 586–587.
61. Watanabe, S.; Hori, T.; Karita, S.; Hayashi, T.; Nishitoba, J.; Unno, Y.; Soplin, N.; Heymann, J.; Wiesner, M.; Chen, N.; et al. ESPnet: End-to-End Speech Processing Toolkit. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 2207–2211.
62. Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; Zafeiriou, S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In Proceedings of the CVPR, Online, 14–19 June 2020; pp. 5202–5211.
63. Bulat, A.; Tzimiropoulos, G. How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017; pp. 1021–1030.
64. Zhang, Y.; Yang, S.; Xiao, J.; Shan, S.; Chen, X. Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition. In Proceedings of the 15th IEEE FG, Buenos Aires, Argentina, 16–20 November 2020; pp. 356–363.
65. Park, D.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.; Le, Q. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2613–2617.
66. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the ICLR, New Orleans, LA, USA, 6–9 May 2019.
67. Smith, L.; Topin, N. Super-convergence: Very fast training of neural networks using large learning rates. In Proceedings of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, San Jose, CA, USA, 24–28 February 2019; Volume 11006, pp. 36–386.
68. Ott, M.; Edunov, S.; Grangier, D.; Auli, M. Scaling Neural Machine Translation. In Proceedings of the 3rd Conference on Machine Translation, Brussels, Belgium, 31 October–1 November 2018; pp. 1–9.
69. Bisani, M.; Ney, H. Bootstrap estimates for confidence intervals in ASR performance evaluation. In Proceedings of the ICASSP, Seoul, Republic of Korea, 14–19 April 2004; Volume 1, pp. 409–412.
70. Higuchi, Y.; Watanabe, S.; Chen, N.; Ogawa, T.; Kobayashi, T. Mask CTC: Non-Autoregressive End-to-End ASR with CTC and Mask Predict. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 3655–3659.
71. Graves, A. Sequence transduction with recurrent neural networks. *arXiv* **2012**, arXiv:1211.3711.
72. He, Y.; Sainath, T.; Prabhavalkar, R.; McGraw, I.; Álvarez, R.; Zhao, D.; Rybach, D.; Kannan, A.; Wu, Y.; Pang, R.; et al. Streaming End-to-end Speech Recognition for Mobile Devices. In Proceedings of the ICASSP, Brighton, UK, 12–17 May 2019; pp. 6381–6385.

**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.