Visual speech recognition for multiple languages in the wild

Pingchuan Ma1,2, Stavros Petridis1,2 and Maja Pantic1,2

Visual speech recognition (VSR), also known as lipreading, is the task of automatically recognizing speech from video based only on lip movements. In the past, this field has attracted a lot of research attention within the speech recognition community1,2, but it has failed to meet the initial high expectations. There are two main reasons why the first generation of VSR models fell short: (1) the lack of large transcribed audio-visual datasets resulted in models that could only recognize a limited vocabulary and work only in a laboratory environment and (2) the use of handcrafted visual features, which might not have been optimal for VSR applications, prevented the development of high-accuracy models. Recently, large audio-visual transcribed datasets, like LRS23 and LRS34, have become available, and these have allowed the development of a large vocabulary and robust models. In addition, advances in deep learning have made possible the use of end-to-end models, which learn to extract VSR-related features directly from raw images. These developments have led to a new generation of deep-learning-based VSR models that achieve much higher accuracy than older models and also work in unseen real-life situations.

The recent advances in VSR models are mainly fuelled by using increasingly larger transcribed datasets and the development of models that work well when trained with huge amounts of data. Some recent works5,6 use tens of thousands of hours of non-publicly available training data to achieve state-of-the-art performance on standard benchmarks. In contrast to this recent trend, we demonstrate that carefully designing a model is equally as important as using larger training sets. Our approach revolves around (1) addition of prediction-based auxiliary tasks to a VSR model, (2) appropriate data augmentations and (3) hyperparameter optimization of an existing architecture. This leads to a great reduction in word error rate (WER) and results in state-of-the-art performance on almost all benchmarks. This is achieved by using only publicly available datasets, which are two orders of magnitude smaller than those used in previous works. We also show that combining multiple datasets further improves the performance (which is in line with the results reported in the literature). Hence, we argue that further progress in the field can be achieved not only by increasing the size of the training data but also by careful model design and optimization.

The vast majority of existing works focus on improving the performance of English-only VSR models. There are also a few works that design models tailored to a specific language, like Mandarin7,9. In contrast to previous works, our approach is evaluated not only on English but also on Mandarin and Spanish (the two other widely spoken

Received: 22 February 2022
Accepted: 13 September 2022
Published online: 24 October 2022

1Imperial College London, London, UK. 2Meta AI, London, UK. e-mail: pingchuan.ma16@imperial.ac.uk
languages), Italian, French and Portuguese. State-of-the-art performance is achieved in all languages.

Specifically, in this Article, we make the following contributions:

- We propose a novel method for VSR that outperforms state-of-the-art methods trained on publicly available data by a large margin.
- We do so with a VSR model with auxiliary tasks that jointly performs VSR and prediction of audio and visual representations.
- We demonstrate that the proposed VSR model performs well, not only in English, but also in other languages, such as Spanish, Mandarin, Italian, French and Portuguese.
- We show that enlarging the training sets, even by including unlabelled data with automatically generated transcriptions or videos in other languages, results in improved performance. This provides further evidence for the hypothesis that the recent improvements presented in the literature are probably the result of larger training sets and not necessarily of better models.
- We discuss challenges for VSR systems that need to be solved and ethical considerations that must be taken into account before this technology can be widely applied.

**Baseline VSR model**

The baseline VSR model that we extend in this work is based on ref. 10. The model consists of a three-dimensional (3D) convolutional layer with a receptive field of five frames, followed by a 2D ResNet-18 (Fig. 1e), a 12-layer Conformer model11 and a transformer decoder as shown in Fig. 1b. The model is trained end to end using a combination of the connectionist temporal classification (CTC) loss with an attention mechanism. Data augmentation is also used during training in the form of random cropping and image flipping (applied to all frames in the same sequence). This model achieves state-of-the-art VSR performance on the LRS2 and LRS3 datasets, when only publicly available data are used for training.

**Baseline ASR model**

The baseline Automatic Speech Recognition (ASR) model that we use is based on ref. 10. The model consists of an 1D ResNet-18 (Fig. 1d), a 12-layer Conformer model and a transformer decoder as shown in Fig. 1a. This model also follows the hybrid CTC/attention architecture and is trained end to end. Time-masking is also used as data augmentation during training. At the moment, this is the state-of-the-art ASR model on the LRS2 and LRS3 datasets.

**Our approach**

In contrast to previous works, which improve the VSR performance by using increasingly larger training sets, we focus on improving the performance by carefully designing a model without relying on additional data. This is achieved by revising the training strategy and architecture of the state-of-the-art model proposed in ref. 10. First, we optimize hyperparameters and improve the language model (LM) with the aim of squeezing extra performance out of the model. Second, we introduce time-masking, which is a temporal augmentation method that is commonly used in ASR models. It substantially improves the VSR performance by forcing the model to rely more on contextual information and, as a consequence, it can better disambiguate similar lip movements that correspond to different phonemes. Finally, we use a VSR model with auxiliary tasks where the model jointly performs VSR and prediction of audio and visual representations extracted from pre-trained VSR and ASR models. This prediction task provides an additional supervisory signal and forces the model to learn better visual representations. A diagram of the architecture of our model is shown in Fig. 1c.

The performance of our model is presented in Tables 1–4. Owing to the random nature of training, we train ten models for each experiment and report the mean and standard deviation of the WER over the ten runs. This is in contrast to previous works, which report just a single value (most probably the best WER) and no standard deviation, and it provides a more robust estimate of the actual performance. However, to facilitate a fair comparison with other works, we also report the best WER of the ten runs.

**Results on LRS2**

The results on LRS2—an English audio-visual dataset—are reported in Table 1. Our model outperforms all existing works by a large margin, even when it is trained on smaller amounts of training data. In particular, it outperforms the previous state of the art—ref. 10—in terms of the best WER achieved, by 5%. This is despite the fact that, in ref. 10, training is carried out on a larger training set. When we use the same training set size as in ref. 10, our model results in a 22.8% absolute improvement. When we use additional training data, an even larger improvement of 12.4% is observed. Similarly, our approach results in a 22.8% absolute improvement in the best WER over ref. 1, which uses a training set with similar size to ours and also includes non-publicly available data.
The results on LRS3—another English audio-visual dataset—are presented in Table 2. In this case too, our proposed approach substantially outperforms all existing works that are trained using publicly available datasets. In particular, our method leads to an 8.2% absolute improvement, in terms of the best WER, over the state of the art—ref. 10—when the same training data are used. As expected, a smaller absolute improvement of 5.4% is reported when a smaller training set is used. In the case of additional training data being available, a larger absolute improvement of 11.8% is achieved.

There are also some works that rely on very large non-publicly available datasets for training. As a consequence, it is not clear whether the reported improvement in WER is due to a better model or simply to the large amount of training data. Our approach outperforms all existing works, as expected, a smaller absolute improvement of 5.4% is reported when a smaller training set is used. In the case of additional training data being available, a larger absolute improvement of 11.8% is achieved.

Results on CMLR

The results on the CMLR dataset—a Mandarin audio-visual dataset—are shown in Table 3. We report performance in terms of character error rate (CER) instead of WER, because Chinese characters are not separated by spaces. Our approach results in a substantial reduction in the CER over all existing works. We achieve an absolute improvement of 12.9% over the state of the art, ref. 9. The WER can be further reduced by 1.1% by first pre-training our model on English and then fine-tuning it on the CMLR training set.

Results on CMU-MOSEAS-Spanish

The results on the CMU-MOSEAS-Spanish dataset—an audio-visual Spanish dataset—are shown in Table 4. Given that this is a small dataset, it is not possible to train an accurate model without using additional data. For this purpose, we first pre-trained the model on English datasets and then fine-tuned it on the training sets of CMU-MOSEAS and TEDx datasets using the Spanish videos only. Because this is a new dataset and there are no results from previous works, we trained the end-to-end model presented in ref. 10 to serve as the baseline.

Table 1 | Results on the LRS2 dataset

| Method                  | Pre-training set | Training set | Training sets total size (h) | Mean ± s.d. | Best  |
|-------------------------|------------------|--------------|------------------------------|-------------|-------|
| Using publicly available datasets |
| MV-WAS S                   | –                | LRS2         | 223                          | –           | 70.4  |
| CTC/Att. S                 | LRW              | LRS2         | 380                          | –           | 63.5  |
| KD+CTC S                  | VoxCeleb2clean+LRS3 | LRS2       | 995                          | –           | 51.3  |
| KD-seq2seq S              | LRW+LRS3         | LRS2         | 818                          | –           | 49.2  |
| TDNN S                    | –                | LRS2         | 223                          | –           | 48.9  |
| CM-seq2seq S            | LRW              | LRS2         | 380                          | –           | 37.9  |
| Ours                     | –                | LRS2         | 223                          | 33.8 ± 0.5  | 32.9  |
| Ours                     | LRS2              | LRS2         | 380                          | 29.5 ± 0.4  | 28.7  |
| Ours                     | LRW+LRS3         | LRS2         | 818                          | 27.6 ± 0.2  | 27.3  |
| Ours                     | LRW+LRS3+AVSpeech | LRS2       | 1,459                        | 25.8 ± 0.4  | 25.5  |

Using non-publicly available datasets

| Method                  | Pre-training set | Training set | Training sets total size (h) | Mean ± s.d. | Best  |
|-------------------------|------------------|--------------|------------------------------|-------------|-------|
| TM-seq2seq S            | MVLRS+LRS3       | LRS2         | 1,391                        | –           | 48.3  |

'Mean ± s.d.' refers to the mean WER over ten runs and the corresponding standard deviation, and 'Best' denotes the best (lowest) WER.

Table 2 | Results on the LRS3 dataset

| Method                  | Pre-training set | Training set | Training sets total size (h) | Mean ± s.d. | Best  |
|-------------------------|------------------|--------------|------------------------------|-------------|-------|
| Using publicly available datasets |
| KD+CTC S                | VoxCeleb2clean | LRS3         | 772                          | –           | 59.8  |
| KD-seq2seq S           | LRW+LRS2         | LRS3         | 818                          | –           | 59.0  |
| CM-seq2seq S           | LRW              | LRS3         | 595                          | –           | 43.3  |
| Ours                    | –                | LRS3         | 438                          | 38.6 ± 0.4  | 37.9  |
| Ours                    | LRS3              | LRS3         | 595                          | 35.8 ± 0.5  | 35.1  |
| Ours                    | LRS3+LRS2        | LRS3         | 818                          | 34.9 ± 0.2  | 34.7  |
| Ours                    | LRS3+AVSpeech    | LRS3         | 1,459                        | 32.1 ± 0.3  | 31.5  |

Using non-publicly available datasets

| Method                  | Pre-training set | Training set | Training sets total size (h) | Mean ± s.d. | Best  |
|-------------------------|------------------|--------------|------------------------------|-------------|-------|
| TM-seq2seq S            | MVLRS+LRS2       | LRS3         | 1,391                        | –           | 58.9  |
| V2P S                   | –                | LSVSR        | 3,886                        | –           | 55.1  |
| RNN-T S                 | –                | YT-31k       | 31,000                       | –           | 33.6  |
| VI3D-TM S               | –                | YT-90k       | 90,000                       | –           | 25.9  |
| VI3D-CM S               | –                | YT-90k       | 90,000                       | –           | 17.0  |

'Mean ± s.d.' refers to the mean WER over ten runs and the corresponding standard deviation, and 'Best' denotes the best (lowest) WER.

Results on LRS3

The results on LRS3—an another English audio-visual dataset—are presented in Table 2. In this case too, our proposed approach substantially outperforms all existing works that are trained using publicly available datasets. In particular, our method leads to an 8.2% absolute improvement, in terms of the best WER, over the state of the art—ref. 10—when the same training data are used. As expected, a smaller absolute improvement of 5.4% is reported when a smaller training set is used. In the case of additional training data being available, a larger absolute improvement of 11.8% is achieved.
We observe that our proposed approach results in a 7.7% absolute reduction in the WER. A further reduction of 6.5% can be achieved by using additional training data.

**Comparison between mean and best WER/CER**
In all results shown in Tables 1–4 we report both the mean and the best performance over ten runs. We observe that the mean WER, which is more representative of the actual performance, is up to 0.8% worse than the best WER. The only exception is for the CMLR dataset (Table 3), where the mean and best CER are practically the same, mainly as a result of the large size of the test set. This difference between the mean and best WER is something that should be taken into account when comparing different models, especially when the models are tested on relatively small test sets and the results are too close.

**Applications**
Speech is the most commonly used human communication method and consists of an audio signal and the corresponding mouth movements. Speech perception is also bimodal, as demonstrated by the McGurk effect\(^2\), where the perception of a sound may change depending on the lip movements shown to the observers. In addition, it has been shown that the addition of visual speech information to a word recognition task performed by normal hearing adults is equivalent to increasing the signal-to-noise ratio (SNR) by 15 dB compared to audio-only recognition\(^3\). Hence, one of the main applications of VSR is to enhance the performance of ASR models in noisy environments. VSR models are not substantially affected by acoustic noise and can be integrated into an audio-visual speech recognition (AVSR) model to compensate for the performance drop of ASR models. Several AVSR architectures have been proposed\(^4,5,12,15-18\); these show that the improvement over ASR models is greater as the noise level increases, that is, the SNR is lower. The same VSR architectures can also be used to improve the performance of audio-based models in a variety of applications like speech enhancement\(^19\), speech separation\(^20\), voice activity detection\(^21\), active speaker detection\(^22\) and speaker diarization\(^23\).

Table 4 | Results on the CMU-MOSEAS-Spanish (CM\(_{es}\)) dataset

| Method               | Pre-training set | Training set | Training sets total size (h) | Mean ± s.d. | Best |
|----------------------|------------------|--------------|------------------------------|-------------|------|
| CM-seq2seq\(^2\)     | LRW              | CM\(_{es}\) + MT\(_{es}\) | 244                          | 58.9 ± 0.8  | 58.1 |
| Ours                 | LRW              | CM\(_{es}\) + MT\(_{es}\) | 244                          | 51.5 ± 0.8  | 50.4 |
| Ours                 | LRW + LRS2 + LRS3 | CM\(_{es}\) + MT\(_{es}\) | 905                          | 47.4 ± 0.2  | 47.2 |
| Ours                 | LRW + LRS2 + LRS3 + AVSpeech | CM\(_{es}\) + MT\(_{es}\) | 1,546                        | 44.6 ± 0.6  | 43.9 |

\(^1\)Mean ± s.d.: refers to the mean WER over ten runs and the corresponding standard deviation, and ‘Best’ denotes the best (lowest) WER.
motion blur and compression. Reduced and/or mismatched resolution and frame rate between training and test conditions can also affect performance. There is some evidence that VSR systems are robust to small or moderate amounts of noise and less robust to reduced resolution but further studies are needed to establish the impact of each noise type.

Another challenge is that a VSR model should be person-independent and pose-invariant. However, it is well known that deep networks rely heavily on texture. This can potentially degrade the performance, because unknown test subjects and head pose can substantially affect the appearance of the mouth. This is typically addressed by training the VSR models on a large number of subjects with varying poses. Some preliminary works on pose-invariant VSR have shown that this can be addressed in a more principled way, and this is another area that deserves further attention. Similarly, multi-view VSR can be beneficial, but it is not yet clear which lip views are optimal and how they should be combined. The availability of multiple cameras in meeting rooms, cars and in modern smartphones opens up a new opportunity for improving VSR systems.

The vast majority of VSR systems have focused on plain English speech. However, it is known that lip movements are affected by the context where speech is produced and the type of speech. There is evidence that lip movements tend to increase in silent speech and also when speech is produced in noise. Despite studies that show a performance drop when VSR models are tested on such conditions, this area remains unexplored. Finally, the development of non-English VSR systems that take into account the unique characteristics and accents of each language also remains an open challenge.

Ethical considerations

It is important to note that VSR is a dual-use technology, which means it can have a positive impact on society as well as a negative one. Although our objective is to build VSR systems that will be beneficial for society, like the applications mentioned above, this technology can also be misused. One example is that it can be deployed for surveillance via CCTV or even via smartphone cameras, which raises privacy concerns. A potential side effect of this is that it might discourage people from speaking in public if they believe that their conversation can be intercepted by anyone carrying a camera. Sophisticated surveillance using VSR technology might not be possible at the moment, especially via CCTV due to the low quality of CCTV camera images, compared to the high-quality data used during training, but it should not be ignored. Cameras and VSR systems are getting better, so it might become a serious privacy concern rather soon unless automatic blurring of all faces of people who did not provide an explicit consent becomes a standard.

Commercial applications of VSR technology are still at a very early stage. One of the very few examples is a smartphone application that aims to help speech-impaired individuals communicate and is currently being trialled in UK NHS hospitals. This is being developed by Liopa, which also works on keyword spotting from CCTV footage. We thus argue that appropriate government regulations for VSR systems, which address privacy concerns and potential misuse, are necessary at this early stage before the technology is fully commercialized. This will allow the proper auditing of every new application before it reaches the market, so that the risks and merits can be properly communicated to users and the public. Otherwise, VSR systems may have the same fate as face recognition technology, which was commercialized without proper regulation being in place. As a consequence, a ban on using face recognition was introduced in several cities and some companies either stopped offering such services or put restrictions on their use when the ethical concerns became widely known.

It should also be pointed out that VSR technology might be biased against specific age groups, genders, cultural backgrounds or non-native speakers. Most of the publicly available datasets have been collected from TV programmes, TED talks or YouTube videos. Hence, it is very likely that some groups are underrepresented, for example, younger people when data are collected from TV programmes or older people when data are collected from YouTube. Similarly, it is likely that people from specific cultural backgrounds or non-native speakers are also underrepresented. This will lead to VSR models that are less accurate for all these groups. Because demographic information is not available for any publicly available dataset used for training VSR models, it is not easy to verify whether such biases exist. VSR models need to be trained on demographically diverse data, including non-native speakers, to ensure similar performance across different user groups. This will lead to VSR systems whose accuracy is not lower for some users because their age, gender, cultural background or accent is underrepresented in the training data.

Methods

Our method outperforms state-of-the-art methods by a large margin for VSR in multiple languages. In what follows we explain the details of our approach and the changes that we have made to the training strategy and architecture that led to this highly improved performance.

Datasets

LRS2. Ref. describes a large-scale audio-visual English dataset collected from BBC programmes. It consists of 144,482 video clips with a total duration of 224.5 h. The videos are divided into a pre-training set with 96,318 utterances (195 h), a training set with 45,839 utterances (28 h), a validation set with 1,082 utterances (0.6 h) and a test set with 1,243 utterances (0.5 h).

LRS3. Ref. describes the largest publicly audio-visual English dataset collected from TED talks. It contains 438.9 h with 151,819 utterances. Specifically, there are 118,516 utterances in the 'pre-train' set (408 h), 31,982 utterances in the 'train-val' set (30 h) and 1,321 utterances in the 'test' set (0.9 h).

CMLR. Ref. describes a large-scale audio-visual Mandarin dataset collected from a Chinese national news programme. It contains 102,072 clips with transcriptions. The training, validation and test sets contain 71,448 (60.6 h), 10,206 (8.6 h) and 20,418 (17.3 h) clips, respectively. To the best of our knowledge, CMLR is the largest publicly available dataset in Mandarin.

CMU-MOSEAS. Ref. describes a large-scale dataset that contains multiple languages and was collected from YouTube videos. It consists of 40,000 transcribed sentences and includes Spanish, Portuguese, German and French. We consider the Spanish videos (CMU-MOSEAS) with a total duration of 16.3 h. We divided the data into training and test sets, which contain 8,253 videos (15.7 h) and 329 videos (0.6 h), respectively.

Multilingual TEDx. Ref. describes a multilingual corpus collected from TEDx talks. It covers eight languages with manual transcriptions and has a total duration of 765 h. For the purposes of this study, we consider the Spanish videos (CMU-MOSEAS) with a total duration of 16.3 h. We divided the data into training and test sets, which contain 8,253 videos (15.7 h) and 329 videos (0.6 h), respectively.

AVSpeech. Ref. is a large-scale audio-visual dataset consisting of 4,700 h of video in multiple languages. A pre-trained language recognition model, VoxLingual10, was first used to identify the English speaking videos. Two pre-trained ASR models, Wav2vec2-Base-960h (https://huggingface.co/facebook/wav2vec2-base-960h) and Wav2vec2-large-xlsr-53-english (https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english), were then used to obtain machine-generated transcriptions for these videos. We only kept the
 videos where the WER between the two generated transcriptions was below 60%, resulting in 350,991 videos with a total duration of 641 h. The transcriptions generated by the Wav2Vec2-Base-960h model were used for these videos.

Performance metrics
WER is the most common metric used in speech recognition. This measures how close the predicted word sequence is to the target word sequence. Assuming $S$ is the number of substitutions, $D$ is the number of deletions, $I$ is the number of insertions needed to get from the predicted to the target sequence and $N$ is the number of words in the target sequence, then the metric can be defined as

$$\text{WER} = \frac{S + D + I}{N}. \quad (1)$$

Similarly to WER, we can define the CER, which measures how close the predicted and target character sequences are. In this case, $S$, $D$ and $I$ are computed at the character level and $N$ is the total number of characters.

Pre-processing
We used the RetinaFace\cite{zhan2018retinaface} face detector and the Face Alignment Network (FAN)\cite{zhu2016face} to detect 68 facial landmarks. The faces were then registered to a neutral reference frame using a similarity transformation to remove translation and scaling variations. A bounding box of $96 \times 96$, centred on the mouth centre, was used to crop the mouth region of interest. The cropped patch was further converted to grey-scale and normalized with respect to the overall mean and variance of the training set.

Hyperparameter optimization
Hyperparameter optimization aims to improve the performance of a model by fine-tuning the values of the parameters that are used to control the training process or the model architecture. Some of the most common hyperparameters that are usually optimized are the following: initial learning rate, learning rate decay parameters, number of layers, size of layers, dropout rate and the loss function weights, which are used to combine the different loss terms. Additional hyperparameters related to conformers are the number and size of the self-attention heads. We performed hyperparameter optimization on the LRS2 dataset by attempting to reduce the WER on the validation set. Our conclusion was that the parameters used in the baseline model\cite{yatkin2021learning} were already optimal, so no further improvement was observed.

The next step was to optimize other hyperparameters that might not have been exhaustively optimized, like batch-size-related parameters. Again, the parameters were chosen based on the validation set performance. Further details and results are provided in Supplementary Section 4 and Supplementary Table 8, respectively. The results on the LRS2 and LRS3 test sets are shown in Supplementary Table 9. Each hyperparameter was optimized independently based on the WER on the validation set of LRS2. We used the same hyperparameters for all experiments. It is clear that hyperparameter optimization results in a substantial reduction in the WER for both datasets.

Improving LMs
A LM determines the probability of a given sequence of characters. It is used during decoding and favours sequences that are more likely to occur. To increase the capacity of the LM we use multiple text corpora for training. We also increase the number of sequences considered during decoding (beam size is set to 40). The impact of these changes is demonstrated in Supplementary Table 9, where the WER is reduced for both English datasets.

The score from the LM ($S_{\text{lm}}$) is incorporated in decoding as follows:

$$S = \lambda_{\text{CTC}} S_{\text{CTC}} + \left(1 - \lambda\right) S_{\text{att}} + \beta S_{\text{LM}}, \quad (2)$$

where $S_{\text{CTC}}$ and $S_{\text{att}}$ are the scores of the CTC and decoder branch, respectively, and $\lambda$ and $\beta$ correspond to the CTC and LM score weights. Additional details about the corpora used for training the LM in each language, as well as training details, are presented in Supplementary Section 5.

Time-masking
Data augmentation works by synthesizing additional distorted training data with the goal of reducing over-fitting. In VSR, most existing works make use of image transformations such as random cropping and horizontal flipping\cite{karim2018learning,li2020understanding}. These spatial augmentations are helpful, but they do not take into account the temporal nature of visual speech. Only a few works exist that apply temporal augmentations like deleting or duplicating frames\cite{fu2019vocal} or variable length augmentation\cite{cheng2020temporal}.

In this Article we propose the use of time-masking, which is commonly used in training ASR models\cite{sun2016temporal}. It works by randomly masking $n$ consecutive frames by replacing them with the mean sequence frame. This allows the model to more effectively use contextual information and can better disambiguate similar lip movements that correspond to different phonemes. It also makes the model more robust to short missing segments. Given that there is large variance in the video lengths, especially on the LRS2 and LRS3 datasets, the number of masks used is proportional to the length of the training sequence. Specifically, we use one mask per second and, for each mask, we randomly mask up to 40% of frames, with the masked segments chosen using a uniform distribution. Additional details about this augmentation are provided in Supplementary Section 6.

The impact of time-masking is shown in the ablation study on the LRS2 and LRS3 datasets shown in Table 5. Training a model without time-masking results in a substantial increase in the mean WER when compared to the full model.

Prediction-based auxiliary tasks
The standard approach to VSR relies on end-to-end training, which allows the entire model to be optimized towards the desired target. This is an attractive property and has led to impressive results, but also results in substantial challenges in training such a large model. One solution that has recently been proposed is the use of auxiliary tasks in the form of additional losses applied to intermediate layers of the model\cite{li2020learning,barlow2022unsupervised}. This acts as regularization, which helps the model learn better representations and leads to better generalization on test data.

Based on this observation, we propose as an auxiliary task the prediction from intermediate layers of audio and visual representations learned by pre-trained ASR and VSR models (Fig. 1c). This is inspired by the recent success of prediction tasks in self-supervised learning. In particular, good audio representations can be learned by predicting handcrafted audio features\cite{hershey2017cnn} or by using joint audio and
visual supervision\(^\text{4}\). Similarly, visual speech representations can be learned by predicting audio features\(^\text{4}\). Hence, the proposed auxiliary task provides additional supervision to the intermediate layers of the model, which in turns results in better visual representations and improved performance. Mathematically, this is formulated as a regression problem where the goal is to minimize the L1 distance between the predicted and pre-trained visual and audio features. This results in the following loss term added to the loss function:

\[
\mathcal{L}_\text{aux} = \beta_1 ||h_\alpha(f(x)), g_a(x)||_1 + \beta_2 ||h_\beta(f(x)), g_v(x)||_1
\]

where \(x\) and \(x\) are the visual and audio input sequences, respectively, \(f\) and \(g\) are the pre-trained visual and audio encoders, respectively, \(f\) is the subnetwork up to layer \(f\) whose intermediate representation is used as input to the audio and visual predictors \(h\) and \(a\), respectively, \(\beta_1\) and \(\beta_2\) are the coefficients for each loss term, and \(||\cdot||_1\) is the \(\ell^1\)-norm.

The model performs VSR and at the same time attempts to predict audio and visual representations from intermediate layers. Hence, the final loss is simply the addition of the main VSR loss and the auxiliary loss:

\[
\mathcal{L} = \mathcal{L}_\text{VSR} + \mathcal{L}_\text{aux}
\]

\[
\mathcal{L}_\text{VSR} = \alpha \mathcal{L}_\text{CTC} + (1 - \alpha) \mathcal{L}_\text{att}
\]

where \(\mathcal{L}_\text{VSR}\) is the loss of the hybrid CTC/attention architecture used, \(\mathcal{L}_\text{CTC}\) is the CTC loss, \(\mathcal{L}_\text{att}\) is the loss of the attention mechanism, and \(\alpha\) controls the relative weight of each loss term. Further details about the losses are provided in Supplementary Section 7. We emphasize that the proposed method is not architecture-dependent and can also be used with other more advanced visual front ends\(^\text{6}\).

The substantial impact of the auxiliary losses on performance can be observed from Table 5. Removing either loss, that is, either the first or second term from equation (3), leads to an increase in the mean WER for both datasets. In the case where both losses are removed, that is, no auxiliary loss is used, then the increase in the mean WER is even greater. Finally, the removal of the two losses and time-masking results in a substantial decrease in performance.

An ablation study on the effect of layer \(f\) where the auxiliary loss (equation (3)) is attached is shown in Supplementary Fig. 1. Layer 6 was found to be the optimal level based on the performance on the validation set. All results reported in all the tables are based on this configuration. Further details are presented in Supplementary Section 9.1.

Using additional training data

Using larger and larger training sets with a view to reducing the WER is a recent trend in the literature. To investigate the impact of the amount of training data, we trained models on varying amounts of data. We started by training models using only the training set of each database (seventh row of Table 1 and fourth row of Table 2). It is not possible to train a model from scratch on the LRS2 and LRS3 datasets, so we used curriculum learning. This means that we first used only short utterances and as training progresses we kept adding longer ones. Further details on curriculum learning are provided in Supplementary Section 8. We used a model trained for recognizing 500 English words\(^\text{7}\) on the LRW dataset for initialization, then fine-tuned it on the corresponding training sets of the LRS2 or LRS3 datasets (eighth row of Table 1 and fifth row of Table 2). Finally, we used the models trained on LRW + LRS3 and LRW + LRS3 as initialization and fine-tuned them further on LRS2 and LRS3, respectively (ninth row of Table 1 and sixth row of Table 2). It is clear that, as we use more datasets for training, the performance keeps improving. This is also the case for Spanish and Mandarin (sixth row of Table 3 and third row of Table 4), even when models trained on English are used for initialization. However, the reduction in WER is smaller than in English, probably due to language mismatch.

Finally, we used a subset of the AVSpeech dataset as additional training data together with the automatically generated English transcriptions. Again, the WER is reduced in all languages (tenth row of Table 1, seventh row of Table 2, last row of Tables 3 and 4), despite using transcriptions that contain errors, with the smallest reduction observed in Mandarin. This is not surprising, because Mandarin is much less similar to English than Spanish. These results are in line with the hypothesis that the reduction in the WER reported in recent works is mainly due to the larger datasets used for training.

Implementation

Our experiments were implemented using an open-source toolkit, ESPNet\(^\text{4}\). We trained the models with the Adam optimizer\(^\text{4}\) with \(\beta_1 = 0.9, \beta_2 = 0.98\) and \(\epsilon = 10^{-9}\). The learning rate increases linearly in the first 25,000 steps, yielding a peak learning rate of 0.0004 and thereafter decreasing in proportional to the inverse square root of the step number. The network was trained for 50 epochs with a batch size of 16. We used the model averaged over the last ten checkpoints for evaluation. Details regarding the network architecture are provided in Supplementary Section 2.

Conclusions

In this Article we have presented our approach for VSR and demonstrated that state-of-the-art performance can be achieved not only by using larger datasets, which is the current trend in the literature, but also by carefully designing a model. We have highlighted the importance of hyperparameter optimization, which can further improve the performance of existing architectures. We have then shown the importance of time-masking, which forces the network to focus more on the context. We have also proposed a new architecture based on auxiliary tasks where the VSR model also predicts audio-visual representations learned by pre-trained ASR and VSR models. Finally, we have provided evidence that using larger datasets improves the performance, which is in line with recent works in this field. Our approach outperforms all existing VSR works trained on publicly available datasets in English, Spanish and Mandarin, by a large margin.

Data availability

The datasets used in the current study are available from the original authors on the LRS2 (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs2.html), LRS3 (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrs3.html), CMLR (https://www.vipazoo.cn/CMLR.html), Multilingual (http://openslr.org/100) and CMU-MOSEAS (http://immortal.multicomps.cmu.edu/cache/multilingual) repositories. Qualitative results and the list of cleaned videos for the training and test sets of CMU-MOSEAS and Multilingual TEDx are available on the authors' GitHub repository (https://mpc001.github.io/lipreader.html).

Code availability

Pre-trained networks and testing code are available on a Github repository (https://mpc001.github.io/lipreader.html) or at Zenodo\(^\text{8}\) under an Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence.

References

1. Potamianos, G., Neti, C., Gravier, G., Garg, A. & Senior, A. W. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE 91, 1306–1326 (2003).
2. Dupont, S. & Luettin, J. Audio-visual speech modeling for continuous speech recognition. IEEE Trans. Multimedia 2, 141–151 (2000).
3. Chung, J. S., Senior, A., Vinyals, O. & Zisserman, A. Lip reading sentences in the wild. In Proc. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition 3444–3453 (IEEE, 2017).

4. Afouras, T., Chung, J. S., Senior, A., Vinyals, O. & Zisserman, A. Deep audio-visual speech recognition. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 1 (IEEE, 2018); https://doi.org/10.1109/TPAMI.2018.2889052.

5. Shillingford, B. et al. Large-scale visual speech recognition. In Proc. 20th Annual Conference of International Speech Communication Association 4135–4139 (ISCA, 2019).

6. Serdyuk, D., Braga, O. & Siohan, O. Audio-visual speech recognition is worth 32×32×8 voxels. In IEEE Automatic Speech Recognition and Understanding Workshop 796–802 (IEEE, 2021).

7. Zhang, X. et al. Understanding pictograph with facial features: end-to-end sentence-level lip reading of Chinese. In Proc. 33rd AAAI Conference on Artificial Intelligence 9211–9218 (AAAI, 2019).

8. Zhao, Y., Xu, R. & Song, M. A cascade sequence-to-sequence model for Chinese Mandarin lip reading. In Proc. 1st ACM International Conference on Multimedia in Asia 1–6 (ACM, 2019).

9. Ma, S., Wang, S. & Lin, X. A transformer-based model for sentence-level Chinese Mandarin lipreading. In Proc. 5th IEEE International Conference on Data Science in Cyberspace 78–81 (IEEE, 2020).

10. Ma, P., Petridis, S. & Pantic, M. End-to-end audio-visual speech recognition with conformers. In Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing 7613–7617 (IEEE, 2021).

11. Gulati, A. et al. Conformer: convolution-augmented transformer for speech recognition. In Proc. 21st Annual Conference of International Speech Communication Association 5036–5040 (ISCA, 2020).

12. Makino, T. et al. Recurrent neural network transducer for audio-visual speech recognition. In Proc. IEEE Automatic Speech Recognition and Understanding Workshop 905–912 (IEEE, 2019).

13. McGurk, H. & MacDonald, J. Hearing lips and seeing voices. Nature 264, 746–748 (1976).

14. Sun, T. W. H. & Pollock, J. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215 (1954).

15. Petridis, S., Stafylakis, T., Ma, P., Tzimiropoulos, G. & Pantic, M. Audio-visual speech recognition with a hybrid CTC/attention architecture. In Proc. IEEE Spoken Language Technology Workshop 513–520 (IEEE, 2018).

16. Yu, J. et al. Audio-visual recognition of overlapped speech for the LRS2 dataset. In Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing 6984–6988 (IEEE, 2020).

17. Yu, W., Zeiler, S. & Kolossa, D. Fusing information streams in end-to-end audio-visual speech recognition. In Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing 3430–3434 (IEEE, 2021).

18. 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 Language Process. 28, 1052–1064 (2020).

19. Afouras, T., Chung, J. S. & Zisserman, A. The conversation: deep audio-visual speech enhancement. In Proc. 19th Annual Conference of International Speech Communication Association 3244–3248 (ISCA, 2018).

20. Ephrat, A. et al. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Trans. Graph. 37, 112:1–112:11 (2018).

21. Yoshimura, T., Hayashi, T., Takeda, K. & Watanabe, S. End-to-end automatic speech recognition integrated with CTC-based voice activity detection. In Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing 6999–7003 (IEEE, 2020).

22. Kim, Y. J. et al. Look who’s talking: active speaker detection in the wild. In Proc. 22nd Annual Conference of International Speech Communication Association 3675–3679 (ISCA, 2021).

23. Chung, J. S., Huh, J., Nagrani, A., Afouras, T. & Zisserman, A. Spot the conversation: speaker diarisation in the wild. In Proc. 21st Annual Conference of International Speech Communication Association 299–303 (ISCA, 2020).

24. Denby, B. et al. Silent speech interfaces. Speech Commun. 52, 270–287 (2010).

25. Dungan, L., Karaali, A. & Harte, N. The impact of reduced video quality on visual speech recognition. In Proc. 25th IEEE International Conference on Image Processing 2560–2564 (IEEE, 2018).

26. Bear, H. L., Harvey, R., Theobald, B.-J. & Lan, Y. Resolution limits on visual speech recognition. In Proc. 21st IEEE International Conference on Image Processing 1371–1375 (IEEE, 2014).

27. Geirhos, R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Proc. 7th International Conference on Learning Representations (OpenReview, 2019).

28. Cheng, S. et al. Towards pose-invariant lip-reading. In Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing 4357–4361 (IEEE, 2020).

29. Wand, M. & Schmidhuber, J. Improving speaker-independent lipreading with domain-adversarial training. In Proc. 18th Annual Conference of International Speech Communication Association 3652–3666 (ISCA, 2017).

30. Petridis, S., Wang, Y., Li, Z. & Pantic, M. End-to-end multi-view lipreading. In Proc. 28th British Machine Vision Conference (BMVA, 2017); https://doi.org/10.5244/C.31.161.

31. Bicevski, K. et al. Effects of mouthing and interlocutor presence on movements of visible vs. non-visible articulators. Can. Acoust. 44, 17–24 (2016).

32. Śimko, J., Beňuš, Š. & Vainio, M. Hyperarticulation in Lombard speech: global coordination of the jaw, lips and the tongue. J. Acoust. Soc. Am. 139, 151–162 (2016).

33. Ma, P., Petridis, S. & Pantic, M. Investigating the Lombard effect influence on end-to-end audio-visual speech recognition. In Proc. 20th Annual Conference of International Speech Communication Association 4090–4094 (ISCA, 2019).

34. Petridis, S., Shen, J., Cetin, D. & Pantic, M. Visual-only recognition of normal, whispered and silent speech. In Proc. 43rd IEEE International Conference on Acoustics, Speech and Signal Processing 6219–6223 (IEEE, 2018).

35. Heracleous, P., Ishi, C. T., Sato, M., Ishiguro, H. & Hagita, N. Analysis of the visual Lombard effect and automatic recognition experiments. Comput. Speech Language 27, 288–300 (2013).

36. Efforts to acknowledge the risks of new A.I. technology. New York Times (22 October 2018); https://www.nytimes.com/2018/10/22/business/efforts-to-acknowledge-the-risks-of-new-ai-technology.html
40. Feathers, T. Tech Companies Are Training AI to Read Your Lips https://www.vice.com/en/article/bzvzdw/tech-companies-are-training-ai-to-read-your-lips (2021).
41. Liopa. https://liopa.ai. Accessed 24 November 2021.
42. Crawford, S. Facial recognition laws are (literally) all over the map. Wired (16 December 2019); https://www.wired.com/story/facial-recognition-laws-are-literally-all-over-the-map/
43. Flynn, S. 13 cities where police are banned from using facial recognition tech. Innovation & Tech Today (18 November 2020); https://innotechtoday.com/13-cities-where-police-a-re-banned-from-using-facial-recognition-tech/
44. An update on our use of face recognition. FaceBook (2 November 2021); https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/
45. Metz, R. Amazon will block police indefinitely from using its facial-recognition software. CNN (18 May 2021); https://edition.cnn.com/2021/05/18/tech/amazon-police-facial-recognition-ban/index.html
46. Greene, J. Microsoft won’t sell police its facial-recognition technology, following similar moves by Amazon and IBM. Washington Post (11 June 2020) https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognitiontechnology
47. Afouras, T., Chung, J. S. & Zisserman, A. LRS3-TED: a large-scale dataset for visual speech recognition. Preprint at https://arxiv.org/abs/1809.00496 (2018).
48. Zadeh, A. B. et al. CMU-MOSEAS: a multimodal language dataset for Spanish, Portuguese, German and French. In Proc. 2020 Conference on Empirical Methods in Natural Language Processing 1801–1812 (ACL, 2020).
49. Salesky, E. et al. The multilingual TEDx corpus for speech recognition and translation. In Proc. 22nd Annual Conference of International Speech Communication Association 3655–3659 (ISCA, 2021).
50. Valk, J. & Alumae, T. VoxLingua107: a dataset for spoken language recognition. In Proc. IEEE Spoken Language Technology Workshop 652–658 (IEEE, 2021).
51. Deng, J. et al. RetinaFace: single-stage dense face localisation in the wild. In Proc. 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition 5203–5212 (IEEE, 2020).
52. 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 Proc. 16th IEEE/CVF International Conference on Computer Vision 1021–1030 (IEEE, 2017).
53. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proc. 3rd International Conference on Learning Representations (OpenReview, 2015).
54. Assael, Y., Shillingford, B., Whiteson, S. & De Freitas, N. LipNet: end-to-end sentence-level lipreading. Preprint at https://arxiv.org/abs/1611.01599 (2016).
55. Ma, P., Martinez, B., Petridis, S. & Pantic, M. Towards practical lipreading with distilled and efficient models. In Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing 7608–7612 (IEEE, 2021).
56. Park, D. S. et al. SpecAugment: a simple data augmentation method for automatic speech recognition. In Proc. 20th Annual Conference of International Speech Communication Association 2613–2617 (ISCA, 2019).
57. Liu, C. et al. Improving RNN transducer based ASR with auxiliary tasks. In Proc. IEEE Spoken Language Technology Workshop 172–179 (IEEE, 2021).
58. Toshiwal, S., Tang, H., Lu, L. & Livescu, K. Multitask learning with low-level auxiliary tasks for encoder-decoder based speech recognition. In Proc. 18th Annual Conference of International Speech Communication Association 3532–3536 (ISCA, 2017).
59. Lee, J. & Watanabe, S. Intermediate loss regularization for CTC-based speech recognition. In Proc. 46th IEEE International Conference on Acoustics, Speech and Signal Processing 6224–6228 (IEEE, 2021).
60. Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A. & Bengio, Y. Learning problem-agnostic speech representations from multiple self-supervised tasks. In Proc. 20th Annual Conference of International Speech Communication Association 161–165 (ISCA, 2019).
61. Shukla, A., Petridis, S. & Pantic, M. Learning speech representations from raw audio by joint audiovisual self-supervision. In Proc. 37th International Conference on Machine Learning Workshop (PMLR, 2020).
62. Ma, P., Mira, R., Petridis, S., Schuller, B. W. & Pantic, M. LiRA: learning visual speech representations from audio through self-supervision. In Proc. 22nd Annual Conference of International Speech Communication Association 3011–3015 (ISCA, 2021).
63. Serdyuk, D., Braga, O. & Siohan, O. Transformer-Based Video Front-Ends for Audio-Visual Speech Recognition for Single and Multi-Person Video. In Proc. 23rd Annual Conference of International Speech Communication Association 2833–2837 (ISCA, 2022).
64. Watanabe, S. et al. ESPnet: End-to-end speech processing toolkit. In Proc. 19th Annual Conference of International Speech Communication Association 2207–2211 (ISCA, 2018).
65. Kingma, D. & Ba, J. Adam: a method for stochastic optimization. In Proc. 2nd International Conference on Learning Representations (OpenReview, 2014).
66. Ma, P., Petridis, S. & Pantic, M. 2022. mpc001/Visual_Speech_Recognition_for_Multiple_Languages: visual speech recognition for multiple languages. Zenodo https://doi.org/10.5281/zendao.7065080
67. Afouras, T., Chung, J. S. & Zisserman, A. ASR is all you need: cross-modal distillation for lip reading. In Proc. 45th IEEE International Conference on Acoustics, Speech and Signal Processing 2143–2147 (IEEE, 2020).
68. Ren, S., Du, Y., Lv, J., Han, G. & He, S. Learning from the master: distilling cross-modal advanced knowledge for lip reading. In Proc. 34th IEEE/CVF Conference on Computer Vision and Pattern Recognition 13325–13333 (IEEE, 2021).
69. Zhao, Y. et al. Hearing lips: improving lip reading by distilling speech recognizers. In Proc. 34th AAAI Conference on Artificial Intelligence 6917–6924 (AAAI, 2020).

Acknowledgements
All training, testing and ablation studies were conducted at Imperial College London.

Author contributions
The code was written by P.M., and the experiments were conducted by P.M. and S.P. The manuscript was written by P.M., S.P. and M.P. M.P. supervised the entire project.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-022-00550-z.
Correspondence and requests for materials should be addressed to Pingchuan Ma.
