TRAINING STRATEGIES FOR IMPROVED LIP-READING

Pingchuan Ma[1], Yujiang Wang[1*†], Stavros Petridis[1,2], Jie Shen[1,2], Maja Pantic[1,2]*

[1] Department of Computing, Imperial College London, UK
[2] Meta AI, UK

ABSTRACT

Several training strategies and temporal models have been recently proposed for isolated word lip-reading in a series of independent works. However, the potential of combining the best strategies and investigating the impact of each of them has not been explored. In this paper, we systematically investigate the performance of state-of-the-art data augmentation approaches, temporal models and other training strategies, like self-distillation and using word boundaries indicators. Our results show that time masking is the most important augmentation followed by mixup and Densely-Connected Temporal Convolutional Networks (DC-TCN) are the best temporal model for lip-reading of isolated words. Using self-distillation and word boundaries is also beneficial but to a lesser extent. A combination of all the above methods results in a classification accuracy of 93.4%, which is an absolute improvement of 4.6% over the current state-of-the-art performance on the LRW dataset. The performance can be further improved to 94.1% by pre-training on additional datasets. An error analysis of the various training strategies reveals that the performance improves by increasing the classification accuracy of hard-to-recognise words.

Index Terms— Visual Speech Recognition, Lip-reading, Temporal Convolutional Network, Self-Distillation

1. INTRODUCTION

Lip-reading of isolated words has received a lot of attention recently thanks to the availability of large publicly available datasets like LRW [1]. The majority of works follow the same lip-reading pipeline consisting of a visual encoder, followed by a temporal model and a softmax classification layer. The visual encoder proposed by [2] has been widely adopted in most works, hence, most recent efforts aim at improving the temporal model or the training strategy. Bidirectional Gated Recurrent Units (BGRUs) and Multi-Scale Temporal Convolutional Networks (MS-TCNs) have been the most popular temporal models in the literature and conflicting conclusions about their performance have been reported. For example,

MS-TCNs outperformed BGRUs in [3] but not in [4]. Similarly, different data augmentations have been presented like mixup [4, 5], variable length augmentation [3] and cutout [6]. Other improvements which have been proposed in the literature include the addition of word boundary indicators [7], which define the start and end frame of a word in a video, and self distillation [5] which results in a series of networks with the same architecture trained via distillation. All these improvements have been proposed separately in the literature and a study combining all of them and investigating the impact of each of them is missing.

In this work, we present a model trained with some of the most promising recent ideas and evaluate the contribution of each of them via an ablation study. This is a useful study since we can quantify the effect of each method when combined with other augmentation methods or temporal models. We also provide an error analysis demonstrating how each method improves the lip-reading accuracy. To the best of our knowledge, the only similar study that exists is [4] but despite using some of the latest methods it was only able to match the current state-of-the-art performance.

Our results demonstrate that: 1) We can achieve a new state-of-the-art performance on the LRW dataset by combining all the latest data augmentation methods, using the recently proposed DC-TCN, word boundary indicators and self-distillation. The accuracy achieved is 92.8% for a single model and 93.4% for an ensemble. The performance can be slightly improved to 93.5% and 94.1%, respectively, by pre-training on additional datasets. 2) Time masking is the most effective augmentation method followed by mixup. The use of DC-TCN significantly outperforms the MS-TCN which in turn outperforms the BGRU model. The use of word boundaries and self-distillation is also beneficial with the former resulting in greater improvement. 3) The error analysis suggests that all these methods improve performance by significantly increasing the classification accuracy of difficult words.

2. TRAINING STRATEGIES

Architecture The first building block of the model (Fig. 1c) is the most commonly used mouth region-of-interest (ROI) encoder consisting of a 3D convolutional layer, which takes as
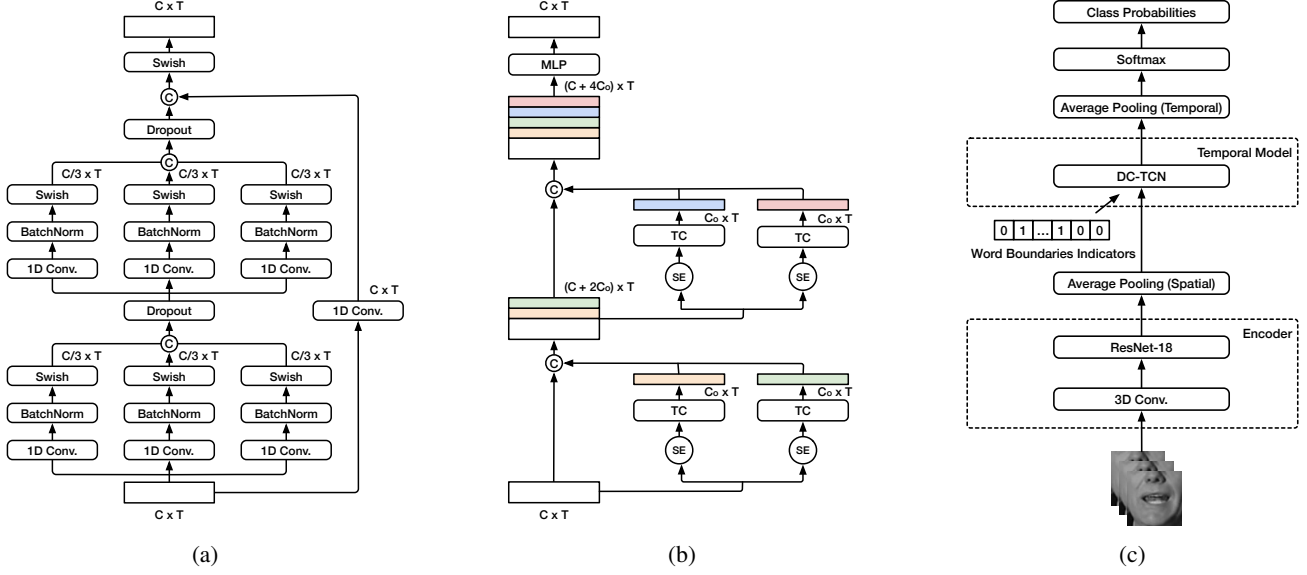

Fig. 1: (a): MS-TCN architecture. "C" and "T" refer to the channel number and sequence length, respectively. (b): DC-TCN architecture. (SE) and (C) denote the the operations of Squeeze-and-Excitation (SE) [8] and channel-wise concatenation, respectively. "TC" represents a Temporal Convolutional block, while the growth rate is denoted as "C_o". (c): Lip-reading model with a modified ResNet-18 as encoder and DC-TCN as a temporal model. The word boundary indicators are concatenated with the output features of the encoder.

input 5 consecutive frames, followed by a 2D ResNet-18 [2]. The frame-wise features from the output of the encoder are then fed to a temporal model to capture the temporal dependencies. This is followed by a softmax layer which outputs the class probabilities over the words to be classified.

In this work, we investigate the impact of three different temporal models for the recognition of isolated words, BGRUs [9], MS-TCNs [3] and DC-TCNs [10]. TCNs consist of a stack of temporal convolutional (TC) blocks , where each block consists of a few layers of dilated convolutions with kernel size k. A MS-TCN (Fig. 1a) extends the vanilla TCN by adding multiple branches each with different kernel sizes, and the features from the output of each branch are concatenated to mix information at several temporal scales. A DC-TCN (Fig. 1b) extends the vanilla TCN by adding dense connection at each TC block and using a Squeeze-and-Excitation (SE) attention mechanism.

Data Augmentation *Random Cropping:* We randomly crop an 88×88 patch from the mouth ROI during training. At test time, we simply crop the central patch. This is a commonly used augmentation method that has been used successfully in several lip-reading works [3, 9]. *Flipping:* We randomly flip all the frames in a video with a probability of 0.5. This augmentation is commonly used in combination with random cropping [3, 9]. *Mixup:* We create new augmented training examples by linearly combining two input video sequences and their corresponding targets. We set the linear combination weight λ to be 0.4 similarly to [5]. *Time Masking:* We mask N consecutive frames for each training sequence where N is sampled between 0 and $N_{\max}$ using a

uniform distribution. Each masked frame is replaced with the mean frame of the sequence it belongs to. This augmentation is based on SpecAugment [11], which has been proposed for ASR applications, and aims at making the model more robust to small segments with missing frames.

Word Boundaries Following [4, 7], we add word boundaries indicators as extra input to the temporal model. The indicators are basically binary vectors with the same length as the number of frames in the input video. All vector entries which correspond to frames where the target word is present are set to 1 and the rest are set to 0. The vector for the word boundaries indicator is concatenated with the frame-wise visual features from the encoder and the new vector is fed into the temporal model.

Self-Distillation Self-distillation [13] is based on the idea of training a series of models with the same architecture using distillation and has been recently applied to lip-reading [5]. Specifically, we first train a network that acts as a teacher for training a student model with the same architecture. The student network becomes the teacher network in the next iteration and we keep training models until no improvement is observed. The insight behind this is that the teacher network provides extra supervisory signal with inter-class similarity information. The overall loss $\mathcal{L}$ to be optimized is the weighted combination of Cross-Entropy loss $\mathcal{L}_{\mathrm{CE}}$ for hard targets and Kullback-Leibler (KL) divergence loss $\mathcal{L}_{\mathrm{KD}}$ for soft targets.

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}(y, \delta(z_s; \theta_s)) + \alpha\mathcal{L}_{\mathrm{KD}}(\delta(z_s; \theta_s), \delta(z_t; \theta_t)) \quad (1)$$

where z_s and z_t represent the embedded representations from

Temporal Model	Data Augmentation				Word Boundary	Pre-training Strategies			Top-1 Acc. (%)
	Crop	Flip	Mixup	TM		Scratch	LiRA(LRS3)	LRS2&3+AVS	
	✓	✓	✓	✓	✓	-	-	✓	**92.9**
	✓	✓	✓	✓	✓	-	✓	-	92.3
	✓	✓	✓	✓	✓	✓	-	-	92.1
	-	✓	✓	✓	✓	✓	-	-	91.8
DC-TCN [10]	✓	-	✓	✓	✓	✓	-	-	91.7
	✓	✓	-	✓	✓	✓	-	-	91.0
	✓	✓	✓	-	✓	✓	-	-	89.7
	✓	✓	✓	✓	-	✓	-	-	90.4
MS-TCN [3]	✓	✓	✓	✓	✓	✓	-	-	90.0
BGRU [9]	✓	✓	✓	✓	✓	✓	-	-	89.7

Table 1: Ablation studies of three temporal models on LRW dataset. Starting from the best-performing DC-TCN model, we remove each data augmentation and the word boundaries indicators to examine their effectiveness. Then we replace the DC-TCN with MS-TCN and BGRU. "Scratch" denotes a model trained from scratch without using external data. "LiRA(LRS3)" indicates a self-supervised pre-trained model using LiRA [12] on the LRS3 dataset, and "LRS2&3+AVS" indicates a fully supervised pre-trained model on LRS2, LRS3 and AVSpeech.

student and teacher networks, respectively, θ_s and θ_t denote learnable parameters of student and teacher models, y is the target label, $\delta(\cdot)$ stands for the softmax function, and α is the balancing weight between the two terms.

3. EXPERIMENTAL SETUP

Databases In our experiments, we employ LRW [1], which is the largest publicly available dataset for lip-reading of isolated words. The dataset is collected in a form of short clips from more than 1 000 speakers in BBC programs and contains 500 isolated words. Each clip has a duration of 29 frames (1.16 seconds). The isolated word is centred within the clip. The dataset is composed of 488 766, 25 000, and 25 000 short clips in the training, validation and test sets.

Pre-Processing We used the RetinaFace [14] tracker to detect the faces and the Face Alignment Network (FAN) [15] for landmark detection. The size and rotation differences are removed through registering the faces to the mean face in the training set. A bounding box of 96×96 is used to crop the mouth ROIs. Each frame is normalised by subtracting the mean and dividing by the standard deviation of the training set.

Training Details The model is trained for 80 epochs with a mini-batch size of 32. We use the AdamW optimizer [16] with an initial learning rate of 3e-4. The learning rate is decayed using a cosine annealing strategy without a warm-up phase. We also use the variable-length augmentation [3] for all experiments. The value for $N_{\max}$ used in time-masking (see section 2) is set to 15 frames (0.6 seconds) and was optimised in the LRW validation set.

Temporal Models *MS-TCN:* We adopt the same MS-TCN architecture as in [3], that is, each block consists of 3 branches with 3/5/7 kernel sizes, respectively, and we stack 4 such

Self-distillation Models	Top-1 Acc. (%)		
	Scratch	LiRA(LRS3)	LRS2&3+AVS
Teacher	92.1	92.3	92.9
Student 1	92.5	92.8	93.5
Student 2	92.8	92.9	93.5
Student 3	92.5	93.0	93.5
Student 4	-	92.9	93.3
Ensemble	**93.4**	**93.6**	**94.1**

Table 2: Performance of self-distillation models (Teacher = ResNet-18 + DC-TCN). The best-performing models from Table 1 are serving as teachers in first row. For each student model, the model from the line above is used as its teacher, and "Student i" stands for the model after the i-th self-distillation iteration.

blocks to formulate the MS-TCN network. *DC-TCN:* The DC-TCN used in this paper generally follows the structures in [10]. In particular, we opt for the Partially Dense (PD) architecture in each TC block, while each block consists of 9 densely-connected temporal convolutions with kernel sizes of $\{3, 5, 7\}$ and dilation rates of $\{1, 2, 5\}$. *BGRU:* A four-layer BGRU with a dropout rate of 0.2 is used with 1024 hidden units.

Initialisation To investigate the impact of initialisation we consider three cases: 1) we train the model from scratch using only the LRW training set, 2) we pre-train the encoder from Fig. 1 on the LRS3 dataset [17] using the LiRA [12] self-supervised approach and fine-tune it on the LRW training set. 3) we pre-train the encoder on LRS2, LRS3 and AVspeech datasets as described in [18].

Method	Word Boundary	Top-1 Acc. (%)
3D-CNN [1]		61.1
ResNet-34 + BLSTM [2]		83.0
2*3D-CNN + BLSTM [19]		84.1
ResNet-18 + BLSTM [7]		84.3
ResNet-18 + BGRU + Cutout [6]		85.0
ResNet-18 + BGRU [4]	✗	85.0
ResNet-18 + MS-TCN [3]		85.3
ResNet-18 + MS-TCN + S.D. [5]		88.5
ResNet-18 + DC-TCN [10]		88.4
Ours (w/o S.D., Scratch)		**90.4**
Ours (w/o S.D., LRS2&3+AVS)		**91.1**
Ours (Ensemble, Scratch)		**91.6**
Ours (Ensemble, LRS2&3+AVS)		**92.1**
ResNet-18 + BGRU [4]		88.4
ResNet-18 + BLSTM [7]		88.8
Ours (w/o S.D., Scratch)		**92.1**
Ours (w/o S.D., LiRA(LRS3))	✓	**92.3**
Ours (w/o S.D., LRS2&3+AVS)		**92.9**
Ours (Ensemble, Scratch)		**93.4**
Ours (Ensemble, LiRA(LRS3))		**93.6**
Ours (Ensemble, LRS2&3+AVS)		**94.1**

Table 3: Comparison with state-of-the-art methods on the LRW dataset in terms of classification accuracy. Experiments are divided into two groups, with and without utilising word boundaries indicators, respectively. "S.D.": self-distillation. "Scratch", "LiRA(LRS3)" and "LRS2&3+AVS" correspond to the three pre-training strategies in Table 1.

4. RESULTS

Ablation Study   Results for the ablation study are shown in Table 1. By removing one augmentation at a time we can estimate its contribution to the final model. We see the time masking is the most important augmentation, resulting in an absolute drop of 2.4% followed by mixup with a drop of 1.1%. By replacing DC-TCN with MS-TCN, we observe that the performance drops by 2.1 %, which demonstrates the importance of dense connections and the SE attention mechanism in DC-TCN. The performance drops by 2.4% by replacing DC-TCN with BGRU. Additionally, the removal of word boundary indicators drops the performance by 1.7 %, which demonstrates the benefits of including auxiliary boundary indicators. Finally, we pre-train the encoder in a self-supervised/supervised manner on the LRS3 / LRS2, LRS3 and AVspeech datasets and then fine-tune the model on the LRW training set, and this slightly increases the performance to 92.3 % / 92.9 %. It is clear from Table 3 that the proposed models significantly outperform the current state-of-the-art.

Self-Distillation   Results for self-distillation experiments are presented in Table 2. We use the best two models from
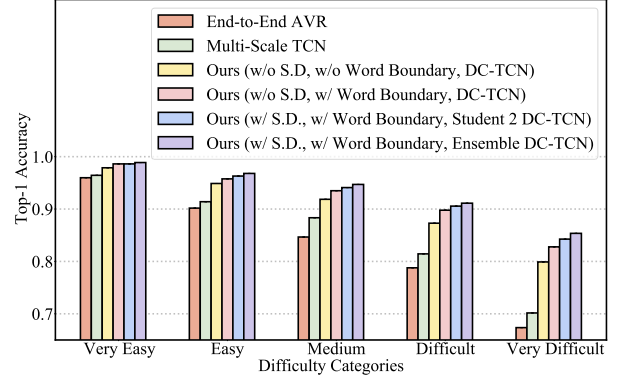

Fig. 2: A comparison of our method and two baseline methods (End-to-End AVR [9] and Multi-Scale TCN [3]) on the five difficulty groups of the LRW test set.

Table 1 as teachers in the first round. It is clear that self-distillation results in a 0.6 % to 0.7 % absolute improvement in all cases. In addition, an ensemble of all models (all students + teacher) leads to a further absolute improvement of 0.6 %. These results suggest that self-distillation is beneficial for lip-reading. However, we should point out that the improvement is smaller compared to [5], probably due to the much better teacher model which makes further improvement harder.

Error Analysis   In order to better understand how the presented models improve the word classification accuracy, we perform some error analysis. We divide the test samples in the LRW dataset into five groups [10]. Each group contains 100 distinct isolated words and it is created based on the word accuracy of the model in [9]. The 100 words with the highest classification accuracy are grouped in the "Very Easy" group, the next 100 words in the "Easy" group and so on. The average classification accuracy in each group is shown in Fig. 2. For comparison purposes, we also include the performance of [3] and [9]. We can see that our models outperform the two baselines across all groups and the improvement is more pronounced in the "Difficult" and "Very Difficult" groups.

5. CONCLUSION

In this work, we present a detailed study on the LRW dataset in terms of data augmentation and temporal models and demonstrate how state-of-the-art performance can be achieved by combining the best augmentations and training strategies. We show that time masking is the most important data augmentation method followed by mixup. We also show that DC-TCNs result in better performance than MS-TCNs or BGRUs. The use of self-distillation and word boundary indicators further improves the classification accuracy whereas the use of pre-training leads to a slight improvement. Finally, an error analysis reveals that the presented models significantly improve the classification accuracy of hard-to-recognise words.

6. REFERENCES

[1] J. S. Chung and A. Zisserman, "Lip Reading in the Wild," in *Proceedings of the 13th Asian Conference on Computer Vision (ACCV)*, 2016, vol. 10112, pp. 87–103.

[2] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," in *Proceedings of the 18th Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017, vol. 9, pp. 3652–3656.

[3] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Lipreading Using Temporal Convolutional Networks," in *Proceedings of the 45th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6319–6323.

[4] Dalu Feng, Shuang Yang, Shiguang Shan, and Xilin Chen, "Learn an Effective Lip Reading Model without Pains," *CoRR*, vol. abs/2011.07557, 2020.

[5] Pingchuan Ma, Brais Martinez, Stavros Petridis, and Maja Pantic, "Towards Practical Lipreading with Distilled and Efficient Models," in *Proceedings of the 46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 7608–7612.

[6] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition," in *Proceedings of the 15th IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2020.

[7] Themos Stafylakis, Muhammad Haris Khan, and Georgios Tzimiropoulos, "Pushing the Boundaries of Audiovisual Word Recognition using Residual Networks and LSTMs," *Computer Vision and Image Understanding*, vol. 176–177, pp. 22–32, 2018.

[8] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-Excitation Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[9] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *Proceedings of the 43rd IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6548–6552.

[10] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic, "Lip-reading with Densely Connected Temporal Convolutional Networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2857–2866.

[11] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proceedings of the Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019, pp. 2613–2617.

[12] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic, "LiRA: Learning Visual Speech Representations from Audio Through Self-Supervision," in *Proceedings of the 22nd Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2021, pp. 3011–3015.

[13] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar, "Born-again neural networks," in *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018, pp. 1602–1611.

[14] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212.

[15] Adrian Bulat and Georgios Tzimiropoulos, "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 16th IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[16] Ilya Loshchilov and Frank Hutter, "Decoupled Weight Decay Regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[17] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *CoRR*, vol. abs/1809.00496, 2018.

[18] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "Visual Speech Recognition for Multiple Languages in the Wild," *arXiv Preprint*, 2022.

[19] Xinshuo Weng and Kris Kitani, "Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2019, p. 269.