Towards practical lipreading with distilled and efficient models

Pingchuan Ma[†,1]*, Brais Martinez*[†,2]*, Stavros Petridis*[1,2]*, Maja Pantic*[1]

[1]Computing Department, Imperial College London, UK
[2]Samsung AI Research Center, Cambridge, UK

pingchuan.ma16@imperial.ac.uk

Abstract

Lipreading has witnessed a lot of progress due to the resurgence of neural networks. Recent work has placed emphasis on aspects such as improving performance by finding the optimal architecture or improving generalization. However, there is still a significant gap between the current methodologies and the requirements for an effective deployment of lipreading in practical scenarios. In this work, we propose a series of innovations that significantly bridge that gap: first, we raise the state-of-the-art performance by a wide margin on LRW and LRW-1000 to 88.6% and 46.6%, respectively, through careful optimization. Secondly, we propose a series of architectural changes, including a novel depthwise-separable TCN head, that slashes the computational cost to a fraction of the (already quite efficient) original model. Thirdly, we show that knowledge distillation is a very effective tool for recovering performance of the lightweight models. This results in a range of models with different accuracy-efficiency trade-offs. However, our most promising lightweight models are on par with the current state-of-the-art while showing a reduction of 8 and 4$\times$ in terms of computational cost and number of parameters, respectively, which we hope will enable the deployment of lipreading models in practical applications.

Index Terms: Visual Speech Recognition, Lip-reading, Knowledge Distillation

1. Introduction

Visual speech recognition (VSR) or lipreading is the task of recognising speech based on the visual stream only. Lipreading has attracted a lot of attention recently mainly due to its robustness in noisy environments where the audio signal might be heavily corrupted.

The traditional lipreading approach was based on the Discrete Cosine Transform and Hidden Markov Models (HMMs) [1, 2, 3]. Recently, the focus has shifted to deep models due to their superior performance. Such models consist of fully connected [4, 5, 6, 7, 8] or convolutional layers [9, 10, 11, 12] which extract features from the mouth region of interest, followed by recurrent layers or attention [12, 13] / self-attention architectures [11]. Few works have also focused on the computational complexity of visual speech recognition [14, 15], and in any case efficient methods have trailed massively behind full-fledged ones in terms of accuracy.

The state-of-the-art approach for recognition of isolated words is the one proposed in [16]. It consists of a 3D convolutional layer followed by an 18-layer Residual Network (ResNet) [17], a Temporal Convolutional Network (TCN) network and a softmax layer. It achieves the state-of-the-art performance on the LRW [12] and LRW1000 [18] datasets, which are the largest publicly available datasets for isolated word recognition.

In this work we focus on improving the performance of the state-of-the-art model and training lightweight models without considerable decrease in performance. Lipreading is a challenging task due to the nature of the signal, where a model is tasked with distinguishing between e.g. *million* and *millions* solely based on visual information. We resort to Knowledge Distillation (KD) [19] since it provides an extra supervisory signal with inter-class similarity information. For example, if two classes are very similar as in the case above, the KD loss will penalize less when the algorithm confuses them. We leverage this insight to produce a sequence of teacher-student classifiers in the same manner as [20, 21], by which student and teacher have the same architecture, and the student will become the teacher in the next generation until no improvement observed (see Fig. 1).

Our second contribution is the proposal of a novel lightweight architecture. The ResNet-18 backbone can be readily exchanged for an efficient one, such as a version of the MobileNet [22] or ShuffleNet [23] families. Furthermore, both architectures have a parameter controlling the width of the networks, effectively controlling the computational complexity of the backbone. However, there is no such equivalent for the head classifier. The key to designing the efficient backbones is the use of depthwise separable convolutions (a depthwise convolution followed by a pointwise convolution) [24] to replace the standard convolution. This operation dramatically reduces the amount of parameters and the number of FLOPs. Thus, we devise a novel variant of the Temporal Convolution Networks that relies on depthwise separable convolutions instead. The resulting efficient lipreading architecture is shown in Fig. 2. We conduct experiments on replacing only the backbone, only the back-end, and both of them, and on giving each component different capacity. The result is a full range of lightweight models with varied efficiency-accuracy trade-offs.

Our third contribution is to use the KD framework to recover some of the performance of these efficient networks. Unlike the full-fledged case, it is now possible to use a higher-capacity network to drive the optimization. However, we find that just using the best-performing model as the teacher, which is the standard practice in the literature, yields sub-optimal performance. Instead, we use intermediate networks whose architecture is in-between the full-fledged and the efficient one. Thus, similar to [25], we generate a sequence of teacher-student pairs that progressively bridges the architectural gap.

We provide experimental evidence showing that a) we achieve new state-of-the-art performance on LRW [12] and LRW-1000 [18] by a wide margin and without any increase of computational performance[1] and b) our lightweight models can

† The first two authors contributed equally.

[1] The models and code are available at https://sites.google.com/view/audiovisual-speech-recognition

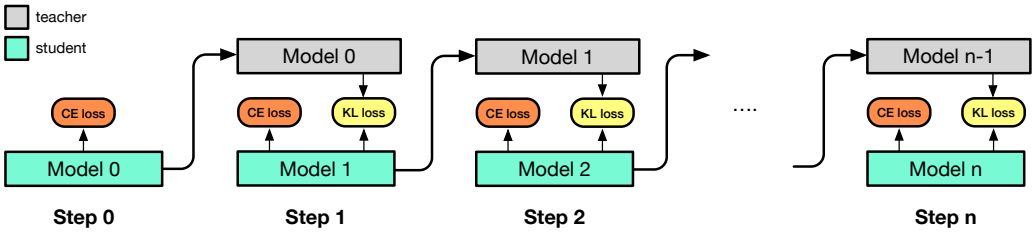

Figure 1: *The pipeline of knowledge distillation in generations*

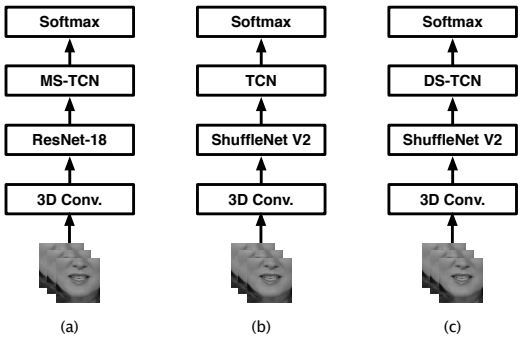

Figure 2: *(a): base architecture with ResNet18 and multi-scale TCN, (b): lipreading model with ShuffleNet v2 backbone and TCN back-end. (c): lipreading model with ShuffleNet v2 backbone and depthwise separable TCN back-end.*

achieve competitive performance. For example, we match the current state-of-the-art on LRW [16] using 8x fewer FLOPs and 4x fewer parameters.

2. Background

Knowledge Distillation: Knowledge Distillation (KD) [19] was initially proposed to transfer knowledge for model compression, so that the student capacity would be much smaller than the teacher one. Recent studies [20, 26, 21] have experimentally shown that the student can still benefit when the teacher and student network have identical architectures. This naturally gave rise to the idea of training in generations. In particular, the student of one generation is used as the teacher of the subsequent generation. This process is iterated until no further improvement is observed. Finally, an ensemble can be optionally used so as to combine the predictions from multiple generations [20]. The training pipeline is detailed in Fig. 1.

Depthwise Seperable Convolution: Standard convolutions rely on a kernel with dimensionality $k \times k \times C_{in} \times C_{out}$, where k is the spatial extent of the kernel and C_{in} and C_{out} are the number of channels of the input tensor and output tensor. Thus, the cost of the convolution is proportional to $k^2 C_{in} C_{out}$. Instead, a depthwise separable convolution separates the spatial and the channel-wise operations to reduce the computational cost. That is to say, it first applies a convolution with kernel size $k \times k \times C_{in} \times 1$ to perform the spatial convolution, where channel interactions are directly ignored, and then a convolution with kernel size $1 \times 1 \times C_{in} \times C_{out}$ to capture the correlations between channels. Thus, cost is proportional to $k^2 C_{in} + C_{in} C_{out}$, where typically the second term dominates. Depthwise Seperable Convolutions are widely used in the existing efficient network architectures including Xception [24], ShuffleNet [27, 23], and MobileNet [22, 28].

3. Towards Practical Lipreading

3.1. Architecture

Base architecture: We use the visual speech recognition architecture proposed in [16] as our base architecture. The details are shown in Fig. 2. It consists of a modified ResNet-18 backbone in which the first convolution has been substituted by a 3D convolution of kernel size $7 \times 7 \times 5$. The rest of the network follows a standard design up to the global average pooling layer. A multi-scale temporal convolutional network (MS-TCN) follows to model the short term and long term temporal information simultaneously.

Efficient backbone: The efficient spatial backbone is produced by replacing the ResNet-18 with an efficient network based on depthwise separable convolutions. For the purpose of this study, We use ShuffleNet v2 ($\beta\times$) as the backbone, where β is the width multiplier [23]. In addition to the use of depthwise in this lightweight architecture, the channel shuffle operation is designed to enable information communication between different groups of channels. Specifically, ShuffleNet v2 ($0.5\times$) has $8.5\times$ fewer parameters and $36\times$ fewer FLOPs than ResNet-18. ShuffleNet v2 ($1.0\times$) has $5.13\times$ fewer parameters and $12.1\times$ fewer FLOPs.

Depthwise Separable TCN: We note that the cost of the temporal convolutional network in MS-TCN is non-negligible. To build an efficient architecture, in addition to replacing standard convolutions with depthwise separable convolutions, we reduce the amount of heads in MS-TCN and leave only one branch with a kernel size of 3. The architecture is denoted as depthwise separable temporal convolutional network (DS-TCN). The detailed architecture is shown in Fig. 2.

3.2. Distillation Loss

This work aims to minimise the combination of cross-entropy loss ($\mathcal{L}_{CE}$) for hard targets and KL divergence loss ($\mathcal{L}_{KD}$) for soft targets. Let us denote the labels as y, the parameters of student and teacher models as θ_s and θ_t, respectively, and the predictions from the student and teacher models as z_s and z_t, respectively. $\delta(\cdot)$ denotes the softmax function and α is a hyper-parameter to balance the loss terms. The overall loss function is calculated as follows:

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

Note that we have omitted the temperature term, which is commonly used to soften the logits of the $\mathcal{L}_{KD}$ term, since we found that the proposed approach works well even without it.

4. Experimental Setup

Datasets: Lip Reading in the Wild (LRW) [12] and LRW-1000 [18] are the largest publicly available lipreading datasets. Both datasets are very challenging as they contain a large number of

Method	Top-1 Acc. (%)	
3D-CNN [29]	61.1	
Seq-to-Seq [12]	76.2	
ResNet34 + BLSTM [9]	83.0	
ResNet34 + DenseNet52 + ConvLSTM [30]	83.3	
ResNet34 + BGRU [31]	83.4	
2-stream 3D-CNN + BLSTM [32]	84.1	
ResNet18 + BLSTM [33]	84.3	
ResNet18 + BGRU + Cutout [34]	85.0	
Resnet18 + MS-TCN [16]	85.3	
Initial Model Trained With →	Adam	AdamW
Student Models Trained With →	AdamW	AdamW
ResNet18 + MS-TCN - Teacher	85.6	87.0
ResNet18 + MS-TCN - Student 1	87.4	87.8
ResNet18 + MS-TCN - Student 2	87.8	87.5
ResNet18 + MS-TCN - Student 3	**87.9**	-
ResNet18 + MS-TCN - Student 4	87.7	-
Ensemble	88.5	**88.6**

Table 1: *Comparison with state-of-the-art methods on the LRW dataset in terms of classification accuracy. Each student is trained using the model from the line above as a teacher.*

speakers, have large variations in head poses, illumination and background noise. *LRW* is based on a collection of over 1000 speakers from BBC programs. There are 488763, 25000, 25000 utterances of 500 target words on training, validation, and test sets, respectively. Each utterance is composed of 29 frames (1.16 seconds), where the target word is surrounded by other context words. *LRW-1000* is the largest Mandarin lipreading dataset collected from more than 2000 speakers with a duration of approximately 57 hours. It has 1000 Mandarin syllable-based classes with a total of 718018 utterances. It contains utterances of varying length from 0.01 up to 2.25 seconds.

Pre-processing: For the video sequences in LRW dataset, 68 facial landmarks are detected and tracked using dlib [35]. The faces are aligned to a neural reference frame to remove differences related to rotation and scale using a similarity transformation. A bounding box of 96×96 is used to crop the mouth region of interest once the centre of the mouth is located. It should be noted that the video sequences in LRW1000 dataset are already cropped so there is no need for pre-processing.

Training: The lipreading model is trained in an end-to-end manner. We train the model for 80 epochs using an initial learning rate of 3e-4, a weight decay of 1e-4 and a mini-batch of 32. We decay the learning rate using a cosine annealing schedule [36]. We should note that all models are trained from random initialisation, without using any external datasets.

Data Augmentation: During training, each sequence is flipped horizontally with a probability of 0.5, randomly cropped to a size of 88×88 and mixup [37] is used with a weight of 0.4. During testing, we use the 88×88 center patch of the image sequence. To improve robustness, we train all models with variable-length augmentation similarly to [16]. where each sequence is segmented temporally at a random point prior and after the boundary of the target word.

Method	Top-1 Acc. (%)
ResNet34 + DenseNet52 + ConvLSTM [30]	36.9
ResNet34 + BLSTM [9]	38.2
ResNet18 + BGRU [34]	38.6
Resnet18 + MS-TCN [16]	41.4
ResNet18 + BGRU + Cutout [34]	45.2 †
Initial Model Trained With →	Adam
Student Models Trained With →	AdamW
ResNet18 + MS-TCN - Teacher	43.2
ResNet18 + MS-TCN - Student 1	**45.3**
ResNet18 + MS-TCN - Student 2	44.7
Ensemble	**46.6**

Table 2: *Comparison with state-of-the-art methods on the LRW-1000 dataset in terms of classification accuracy using the publicly available version of the database (which provides the cropped mouth regions). Each student is trained using the model from the line above as a teacher.* † *This approach uses the full face version of the database, which is not publicly available, in combination with cutout augmentation.*

5. Results

5.1. Born-Again Distillation

In this set of experiments, we apply born-again distillation [20], so that student and teacher have identical architectures. An ensemble of student models is also created as suggested by [20]. Results on the LRW dataset are shown in Table. 1. We notice that when we train a model without distillation, using the AdamW optimiser [38] leads to a significant increase in performance when compared to the Adam optimiser. However, we find the best results after self-distillation are similar and only have 0.1% difference no matter what the accuracy of the initial teacher model is. This leads to a new state of the art performance on LRW by 2.6% margin over the previous one without an increase of computational cost. Furthermore, an ensemble of the models reaches an accuracy of 88.6%, which further pushes the state-of-the-art performance on LRW.

Results on the LRW-1000 dataset are shown in Table 2. In this case, our proposed best single-model accuracy results in an absolute improvement of 3.9% compared to the previous state-of-the-art accuracy on LRW-1000 among works only using the publicly available data. Furthermore, the ensemble model yields a further 1.3%, resulting in a 5.2% overall improvement. These results confirm that indeed inter-class similarity information is crucial for lipreading.

5.2. Efficient Lipreading

One of the major limitations of current lipreading models barring their use in practical applications is that of their computational cost. Many speech recognition applications rely on on-device computing, where the computational capacity is limited, and memory footprint and battery consumption are also important factors. We aim to bridge this gap by constructing very efficient models that can perform on par with competing lipreading methods. To this end, we explore replacing the ResNet-18 backbone and the TCN-based classifier head with efficient alternatives. We chose to replace the ResNet-18 with a ShuffleNet v2 architecture [23] as preliminary experiments showed superior

Student Backbone (Width mult.)	Student Back-end (Width mult.)	Distillation	Top-1 Acc.	Params $\times 10^6$	FLOPs $\times 10^9$
ResNet-18 [16]	MS-TCN (3×)	-	85.3	36.4	10.31
ResNet-34 [31]	BGRU (512)	-	83.4	29.7	18.71
MobiVSR-1 [15]	TCN	-	72.2	4.5	10.75
ShuffleNet v2 (1×)	MS-TCN (3×)	✗	84.4	28.8	2.23
	MS-TCN (3×)	✓	85.5	28.8	2.23
ShuffleNet v2 (0.5×)	MS-TCN (3×)	✗	83.1	27.9	1.69
	MS-TCN (3×)	✓	83.5	27.9	1.69
ShuffleNet v2 (1×)	TCN (2×)	✗	82.7	9.1	1.31
	TCN (2×)	✓	84.6	9.1	1.31
ShuffleNet v2 (1×)	DS-MS-TCN (3×)	✗	84.5	9.3	1.26
	DS-MS-TCN (3×)	✓	85.3	9.3	1.26
ShuffleNet v2 (1×)	TCN (1×)	✗	81.0	3.8	1.12
	TCN (1×)	✓	82.7	3.8	1.12
ShuffleNet v2 (0.5×)	TCN (2×)	✗	81.6	8.2	0.77
	TCN (2×)	✓	82.5	8.2	0.77
ShuffleNet v2 (0.5×)	TCN (1×)	✗	78.1	2.9	0.58
	TCN (1×)	✓	79.9	2.9	0.58
ShuffleNet v2 (0.5×)	DS-TCN (2×)	✗	76.2	3.5	0.58
	DS-TCN (2×)	✓	77.9	3.5	0.58

Table 3: *Performance of different efficient models, ordered in descending computational complexity, and their comparison to the state-of-the-art on the LRW dataset. We use a sequence of 29-frames with a size of 88 by 88 pixels to compute the multiply-add operations (FLOPs). The number of channels is scaled for different capacities, marked as* $0.5\times$, $1\times$, *and* $2\times$. *Channel widths are the standard ones for ShuffleNet V2, while base channel width for TCN is 256 channels.*

performance over MobileNetV2 [28] and EfficientNet-B0 [39] alternatives. In order to control the backbone complexity, we further consider a channel width multiplier of 0.5 and of 1.

The TCN-based head classifier has the following variants: TCN and MS-TCN. TCN indicates the vanilla TCN [40] with kernel of size 3. We chose that kernel size as it yields comparable performance to larger kernels yet has lower computational cost, and a smaller kernel results in large accuracy drops. MS-TCN indicates the multi-scale variant presented in [16]. Finally, for the purpose of model efficiency, we introduce novel depthwise-separable variants of these models, noted as DS-TCN and DS-MS-TCN respectively. We can similarly add capacity to the different TCN variants with a width multiplier respect to the base size of 256 channels.

We explored multiple options as teacher networks. Since now there are higher capacity models that can be used as teachers, we do not need to resort to self distillation. We explore the following options: Use the best-performing network as teacher, use an intermediate capacity network as teacher with either the same backbone or head (e.g., ShuffleNet + MS-TCN as teacher, ShuffleNet + DS-MS-TCN as student), and combining each option with further training in generations. Since each strategy works to varying degrees for the different architectures, we use performance on the validation set to choose the best-performing model and report the accuracy on the test partition. However, we found that using an intermediate-capacity network as teacher, and afterwards using distillation in generations is often the best option. Since the intermediate architecture itself is also trained through distillation, we end up with a progressive teacher-student sequence as in [25].

The results on LRW dataset are shown in Table 3. Remarkably, replacing the state-of-the-art ResNet18-MS-TCN with ShuffleNet-DS-MS-TCN provides the same accuracy than the previous state-of-the-art MS-TCN of [16], while requiring 8.2× fewer FLOPs and 3.9× fewer parameters. This is particularly

Student Backbone (Width mult.)	Student Back-end (Width mult.)	Distillation	Top-1 Acc.	Params $\times 10^6$	FLOPs $\times 10^9$
ResNet18 [16]	MS-TCN(3×)	-	41.4	36.7	15.78
3D DenseNet [18]	BGRU (256)	-	34.8	15.0	30.32
ShuffleNet v2 (1×)	TCN (1×)	✗	40.7	3.9	1.73
	TCN (1×)	✓	41.4	3.9	1.73
ShuffleNet v2 (1×)	DS-TCN (1×)	✗	39.1	2.5	1.68
	DS-TCN (1×)	✓	40.4	2.5	1.68
ShuffleNet v2 (0.5×)	TCN (1×)	✗	40.5	3.0	0.89
	TCN (1×)	✓	41.1	3.0	0.89
ShuffleNet v2 (0.5×)	DS-TCN (1×)	✗	39.1	1.6	0.84
	DS-TCN (1×)	✓	40.2	1.6	0.84

Table 4: *Performance of different efficient models on the LRW-1000 dataset. We use a sequence of 29-frame with a size of 112 by 112 to report multiply-add operations (FLOPs). The number of channels is scaled for different capacities, marked as* $0.5\times$ *and* $1\times$.*Channel widths are the standard ones for ShuffleNet v2, while base channel width for TCN is 256 channels.*

remarkable since the MS-TCN is already quite efficient, having slightly lower computational cost than the lightweight architecture of MobiVSR-1 [15]. Another remarkable combination is the ShuffleNet v2 (0.5×) + TCN model, which achieves 79.9% accuracy on LRW with as little as 0.58 GFLOPs and 2.9M parameters, a reduction of 17.8× and 12.5× respectively when compared to the ResNet18-MS-TCN model of [16].

The same pattern is also observed on the LRW1000 dataset, which is shown in Table 4. ShuffleNet v2 (0.5×) - DS-TCN (1×) provides a higher performance (4.3% absolute improvement) while requiring 9.4× fewer parameters and 36.1× fewer FLOPs than DenseNet [18]. An additional absolute improvement of 1.1% is achieved in the model ShuffleNet v2 (0.5×) - DS-TCN (1×) by using DS-TCN (1×) as the teacher model. A visual depiction of number of parameters vs. accuracy is given in Fig. 3.

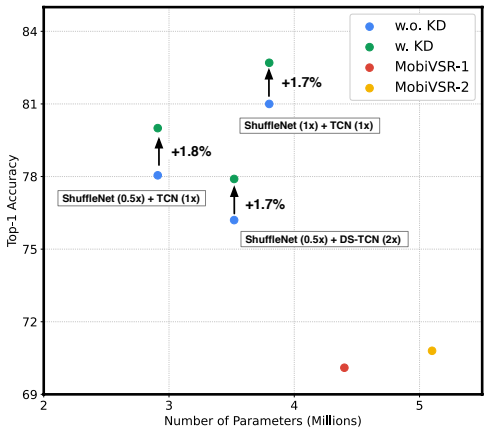

Figure 3: *Model size-vs-Accuracy on the LRW dataset. Our efficient networks significantly reduces the model size and outperforms other lightweight networks.*

6. Conclusions

In this work, we present state-of-the-art results on isolated word recognition by knowledge distillation. We also investigate efficient models for visual speech recognition and we achieve results similar to the current state-of-the-art while reducing the computational cost by 8 times. It would be interesting to investigate in future work how cross-modal distillation affects the performance of audiovisual speech recognition models.

7. References

[1] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[3] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.

[4] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *ICASSP*, 2017.

[5] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *ICASSP*, 2018, pp. 6219–6223.

[6] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multi-view lipreading," in *BMVC*, 2017.

[7] M. Wand, J. Koutnik, and J. Schmidhuber, "Lipreading with long short-term memory," in *ICASSP*, 2016.

[8] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end audiovisual fusion with LSTMs," in *AVSP*, 2017.

[9] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *INTERSPEECH*, 2017.

[10] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett *et al.*, "Large-scale visual speech recognition," in *INTERSPEECH*, 2019.

[11] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *TPAMI*, 2018.

[12] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *CVPR*, 2016.

[13] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/attention architecture," in *IEEE SLT*, 2018.

[14] A. Koumparoulis and G. Potamianos, "Mobilipnet: Resource-efficient deep learning based lipreading," in *INTERSPEECH*, 2019.

[15] N. Shrivastava, A. Saxena, Y. Kumar, R. R. Shah, D. Mahata, and A. Stent, "Mobivsr: A visual speech recognition solution for mobile devices," in *INTERSPEECH*, 2019.

[16] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP*, 2020.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[18] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *FG*, 2019.

[19] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning Workshop*, 2014.

[20] T. Furlanello, Z. C. Lipton, M. Tschannen, L. Itti, and A. Anandkumar, "Born again neural networks," in *ICML*, 2018.

[21] C. Yang, L. Xie, S. Qiao, and A. Yuille, "Knowledge distillation in generations: More tolerant teachers educate better students," in *AAAI*, 2018.

[22] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[23] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *ECCV*, 2018.

[24] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *CVPR*, 2017.

[25] B. Martinez, J. Yang, A. Bulat, and G. Tzimiropoulos, "Training binary neural networks with real-to-binary convolutions," in *ICLR*, 2020.

[26] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label refinery: Improving ImageNet classification through label progression," *arXiv:1805.02641*, 2018.

[27] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018.

[29] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.

[30] C. Wang, "Multi-grained spatio-temporal modeling for lip-reading," in *BMVC*, 2019.

[31] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *ICASSP*, 2018.

[32] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading," in *BMVC*, 2019.

[33] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and lstms," *Computer Vision and Image Understanding*, vol. 176, pp. 22–32, 2018.

[34] 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 *FG*, 2020.

[35] D. E. King, "Dlib-ml: A machine learning toolkit," *JMLR*, 2009.

[36] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[37] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2018.

[38] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[39] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019.

[40] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv:1803.01271*, 2018.