Learning from the Master: Distilling Cross-modal Advanced Knowledge for Lip Reading

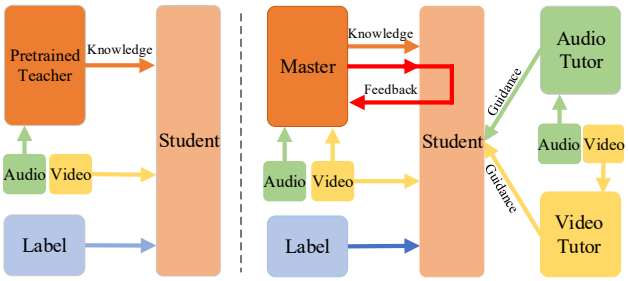
Sucheng Ren[1], Yong Du[2,1]*, Jianming Lv[1], Guoqiang Han[1], Shengfeng He[1]†
[1] School of Computer Science and Engineering, South China University of Technology
[2] Department of Computer Science and Technology, Ocean University of China

Abstract

Lip reading aims to predict the spoken sentences from silent lip videos. Due to the fact that such a vision task usually performs worse than its counterpart speech recognition, one potential scheme is to distill knowledge from a teacher pretrained by audio signals. However, the latent domain gap between the cross-modal data could lead to a learning ambiguity and thus limits the performance of lip reading. In this paper, we propose a novel collaborative framework for lip reading, and two aspects of issues are considered: 1) the teacher should understand bi-modal knowledge to possibly bridge the inherent cross-modal gap; 2) the teacher should adjust teaching contents adaptively with the evolution of the student. To these ends, we introduce a trainable "master" network which ingests both audio signals and silent lip videos instead of a pretrained teacher. The master produces logits from three modalities of features: audio modality, video modality, and their combination. To further provide an interactive strategy to fuse these knowledge organically, we regularize the master with the task-specific feedback from the student, in which the requirement of the student is implicitly embedded. Meanwhile, we involve a couple of "tutor" networks into our system as guidance for emphasizing the fruitful knowledge flexibly. In addition, we incorporate a curriculum learning design to ensure a better convergence. Extensive experiments demonstrate that the proposed network outperforms the state-of-the-art methods on several benchmarks, including in both word-level and sentence-level scenarios.

1. Introduction

Lip reading, which is also referred to as visual speech recognition, aims at predicting words or sentences being spoken from muted lip videos. This vision task enables to switch speech to text without relying on hearing, and therefore, it could apply to many practical scenarios, such as dubbling for silent films, creating a voice for aphonia patients, and serving for security systems. To tackle this problem, early researches usually adopt HMM with designed hand-crafted features [6, 10, 19], whereas recent works exploit deep neural networks [18, 29, 23] for lip reading.



(a) Traditional KD (b) Ours

Figure 1: (a) Traditional knowledge distillation from pretrained audio teacher to a video student. (b) Our method distills advanced knowledge from the master that trained not only with both audio and video data, but also the feedback from the student, leading to a more compatible knowledge transfer. Furthermore, the introduced audio and video tutors provide additional cues to the student for further bringing the cross-modal gap.

Notwithstanding the tremendous success of deep learning and large benchmark construction [8, 1, 2], the admitted most critical challenge is the inherent limitation of visual information, which severely impedes a good performance of lip reading models. For example, different characters "p" and "b" share a similar lip shape so that they are hard to be distinguished in video clips. In contrast, such an uncertainty can be uniquely identified by audio information, and audio based speech recognition would not be affected by ambiguities caused by the intrinsic limitation of visual information. Specifically, the counterpart of lip reading task, *i.e.*, speech recognition, could achieve a much more accurate translation of speech to text. The performance gap can even reach 40% [1] on the metric of word error rate (WER).

Based on this fact, one potential solution is to transfer knowledge from audio data to video data via knowledge distillation (KD) [15] (Fig. 1a). Several previous approaches [29, 31] attempt to build KD-based models, which consist of a teacher network pretrained by audio signals and a student network utilized for lip reading. The audio information is introduced in this way and supposed to be

*The first two authors contribute equally.
†Corresponding author (hesfe@scut.edu.cn).

Table 1: The WER (%) of applying knowledge distillation between audio and video model on LRS2-BBC dataset.

Task	Distilled?	Teacher	WER
Lip reading	✗	-	57.5
Speech recognition	✗	-	**15.7**
Lip reading (KD)	✓	Video	**53.4**
Lip reading (KD)	✓	Audio	54.2

a complementary clue for facilitating the performance of the student. Due to the existed heterogeneity between two modalities, however, such a general audio teacher may only provide limited hidden knowledge to the student for promotion. This observation is examined in Tab. 1, in which we conduct an experiment on LRS2-BBC [1] dataset to test the impact of the teachers which share a same structure but pretrained on different modalities. Note that except for the teachers, the structures of both student networks are also the same. Several interesting facts can be observed: (i) The performance gap (about 40% in WER) between lip reading and speech recognition tasks is similar as reported in previous research [1]. (ii) When distilling knowledge from an audio teacher, the performance of the student is even worse than that of the student with distilled knowledge from the video teacher. Combining (i) and (ii), we can conclude that: a "master" that merely has an advanced accuracy (audio teacher in this case) does not act as a good teacher to the video student; while the video teacher which shares the same data domain with the student, can supervise the student for learning more distilled knowledge representations. Obviously, cross-modal gap is the reason of why this phenomenon occurs. Then we spontaneously raise a question: In lip reading, how can the visual student learn more comprehensible knowledge from the audio "master"?

We consider the above problem from two aspects. First, the teacher should understand bi-modal knowledge to possibly bridge the inherent cross-modal gap. Second, the teacher should dynamically adjust teaching contents in consistent with the evolution of the -student. In this way, the changing requirement of the student would help the teacher to regulate and emphasize the important knowledge from different modalities. Therefore, the cross-modal gap could be fused with a clear aim. In this paper, we propose an innovative deep lip reading model (see in Fig. 1b). Instead of using a pretrained teacher network, we design a trainable and much more powerful network named "*master*". To produce a more compatible knowledge with regard to visual student, the *master* takes not only speech audio data but also lip video data as inputs, and produces three types of probabilities respectively based on audio modality, video modality, and their combination. Then to fuse these knowledge adaptively, we incorporate a couple of "*tutor*" networks into our framework as knowledge fusion guidance, which are pretrained from audio and video data respectively. Based on the interactions among the "*master*", the "*tutors*", and the requirement (feedback) of the student, we design a *dynamic fusion loss* to balance yet fuse different types of knowledge. We also propose a curriculum learning strategy to mitigate the learning ambiguity during training, ensuring a better convergence.

In summary, our key contributions are as follows:

- We propose a collaborative learning based framework for lip-reading. Unlike most other existing methods directly using a pretrained teacher to distill knowledge, we embed an advanced trainable master network into our system. The master could be adjusted according to the feedback of the student, and thus provides bimodal knowledge dynamically for the student to learn in a better way.
- We incorporate a couple of tutor networks into our system, which are respectively pretrained by audio and video data. To get the master, the tutors, and the student to cooperate, we particularly tailor a dynamic fusion loss to guide the student to learn audio-visual probabilities, which alleviates ambiguities caused by the cross-domain gap.
- We present a curriculum learning strategy for lip-reading. By measuring and sorting the difficulty of samples, it could enhance the effectiveness of model training as well as ensuring a better convergence.
- We outperform state-of-the-art lip-reading methods on three benchmarks, indicating the effectiveness of the proposed method.

2. Related work

Lip reading. Lip reading have drawn increasing attention in recent years. Most deep learning based lip reading methods include a pair of frontend and backend, and focus on the corresponding architecture design. The frontend is used to extract visual features, and usually is ResNet [14] or VGG [21] with some modifications. In particular, Chung *et al*. [8] evaluate the performance of different frontends with multiple transform. While the backend is used to map visual features to natural language, mainly including temporal convolution [23], recurrent neural network (RNN) [23], and transformer [1]. Specifically, Stafylakis *et al*. [23] propose to combine residual network as frontend with long short term memory network (LSTM) [16] as backend. Afouras *et al*. [1] introduce transformer to replace RNN as backend. Zhang *et al*. [27] focus on short-range temporal information with temporal focal block and local self-attention.

To deal with the problem that one output word is corresponded to several frames in the lip-reading task, two architectures are designed for the alignment purpose. The first

one is sequence to sequence (Seq2Seq) [24] model, which reads all visual features before prediction. The other one is connectionist temporal classification (CTC) [11], which is an emission model that predicts the results for each frame and searches the optimal alignment for final prediction. LipNet [3] is based on this architecture.

As multi-modal learning develops, lip reading methods attempt to extract information from audio data as a complement for a better performance. Afouras *et al*. [31] propose to distill knowledge from a pretrained audio teacher to guide lip reading. Zhao *et al*. [29] propose frame-, sequence- and context-level distillation methods into the prediction. However, the cross-modal gap between the two modalities is completely ignored, which seriously limits the lip reading performance.

Knowledge distillation. Knowledge Distillation [15] (KD) aims at transferring knowledge from teachers to students. There are two main factors that may affect the performance of KD. The first is the type of knowledge, and lots of previous works propose to discover more useful knowledge as guidance. For example, FitNets [20] utilizes hidden layer features rather the logits for distillation. Zagoruyko *et al*. [26] introduce an attention mechanism to selectively transfer knowledge. Yang *et al*. [25] propose a more tolerant teacher with softer logits, which is easy for the student to mimic. Furlanello *et al*. [9] integrate multiple teachers to boost the utility of knowledge. However, these methods are designed from a perspective of model capacity rather than data capacity, and thus limited in distilling knowledge from a different data modality due to the cross-modal gap.

The other factor is the strategies of distillation. Online distillation, also known as collaborative distillation, is of great interest recently. It aims to alleviate the model capacity gap between the student and the teacher. By treating all the students as teacher, Zhang *et al*. [28] propose DML to make the students learn from each other. ONE [30] constructs the teacher using a multiple branch design. CLNN [22] establishes multiple head branches in the teacher with a corresponding scaling gate for the branch diversity. KDCL [12] ensembles the predictions of students, generating soft targets as supervision. Note that collaborative distillation methods transfer knowledge between students without considering the feedback of each student. It means that when different students learn for different tasks, the potential task gap may distract the learning objective.

3. Approach

3.1. Overview

Given a sequence of lip frames $\{X_V^n \in R^{H\times W\times 3} | n = 1, 2, ..., N_V\}$, our system aims at predicting the spoken word $U \in R^K$ or sentence $\{Z_q \in R^K | q = 1, 2, ..., Q\}$ from lip movements, where Q indicates the length of the sentence in the sentence-level scenarios and K represents the size of vocabulary. It means that a sentence is predicted character by character, while a word is directly recognized in the word-level scenarios.

Fig.2 illustrates the pipeline of the proposed method. Our network includes four modules: *master*, $tutor_A$, $tutor_V$ and *student*, here the subscripts A and V respectively denotes the audio modality and the video modality. The master $f_m(\cdot)$ takes audio signals X_A and video clips X_V as inputs, and provides three types of knowledge: $f_m(X_A; \theta_A)$ generated from audio stream, $f_m(X_V; \theta_V)$ generated from visual stream, and $f_m(X_A, X_V; \theta_A, \theta_V)$ generated from audio-visual combination. The student takes X_V as input, and outputs probabilites $f_s(X_V; \theta_s)$. The pretrained $tutor_A$ outputs logits $f_{t_A}(X_A)$ in regard to speech recognition, whereas the pretrained $tutor_V$ produces logits $f_{t_V}(X_V)$ for lip reading. Note that the pretrained tutors in our model are to provide the information for balancing knowledges from different modalities as well as evaluating the training data by difficulty, not for directly transferring knowledge to the student like pretrained teachers in most existing approaches.

We alternatively train the master and the student, and thus the whole training procedure includes two stages. (i) In the student training stage, the student is optimized by both the distilled cross-modal knowledge from the master and the supervision with lip-reading labels. (ii) In the master training stage, the master is optimized not only by the supervision of labels, but more importantly, the feedback from the temporarily updated student.

3.2. Network Architecture

Let us start with the definitions of audio stream, visual stream, and audio-visual combination. As the names suggest, (i) audio stream is a stream that produces the logits based on audio data, (ii) visual stream is a stream that produces the logits based on video data, and (iii) audio-visual combination aims to combine audio and video features for the logits prediction. Each stream is comprised of a feature extraction frontend, a feature mapping backend and a final classifier. The $tutor_A$ is an audio stream based model, the $tutor_V$ and the student are visual stream based models, and the master consists of both audio stream and visual stream. Next, we will introduce the architectures of these modules in detail.

Audio stream. Audio stream adopts a ResNet-18 [14] as a frontend. Regarding the deployed backend, we have two model variants: (a) Temporal convolution (TC) backend [23] is utilized in word-level lip reading, and (b) transformer sequence to sequence (TM-Seq2Seq) backend [1] is used in both word-level and sentence-level scenarios.

Different from original ResNet, we replace all 2D convolution kernels in the frontend of audio stream with 1D

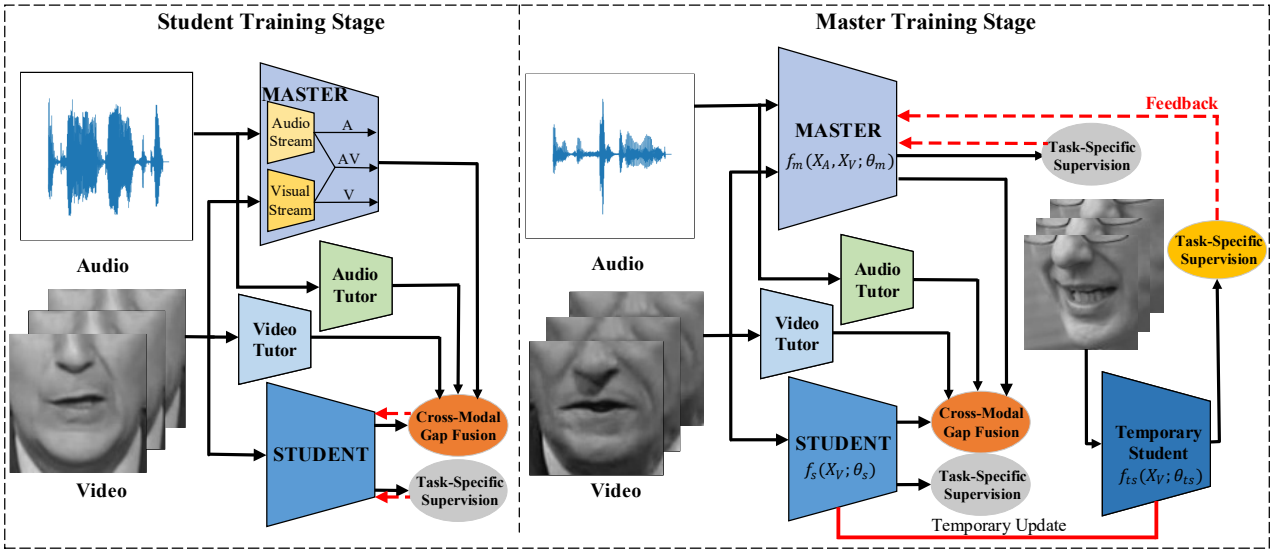


Figure 2: Illustration of our method. The whole training process is split into the student training stage and the master training stage. During the student training stage, the student absorbs task-specific (lip reading) supervision and knowledge distilled from the master with cross-modal guidance provided by video and audio tutors. During the master training stage, a temporary student is introduced as an auxiliary network, feeding task-specific supervision into the master for its update.

filters since a waveform lies in 1D space, and the filter size of the first convolution layer is changed to 5ms with a stride of 0.25ms. As for the TM-Seq2Seq backend, we follow the same setting as that of Afouras *et al.* [1], to include the multi-head attention and the feedforward block.

Visual stream. The structure of visual stream is almost the same as audio stream. The only difference is that, we replace the first 2D convolution layer in the frontend with kernel size of 7×7 by a 3D convolution layer with kernel size of $5 \times 7 \times 7$, because of an extra time dimension.

Audio-visual combination. Audio-visual combination is only set up in the master for generating logits of merged features, which are derived from both audio stream and visual stream. Specifically, in the case of model (a), we directly concatenate the output feature vectors, which are respectively produced by the backends of audio and visual streams, into a new vector. In the case of model (b), the audio and video feature vectors are separately attended by the context vectors and then concatenated into a new vector before feeding in the classifier.

3.3. Learning From the Master

Existing teacher networks are usually pretrained, or lacking of considering to be trained based on that how well the students perform on their tasks. The teachers thus tend to be inflexible in adjusting their teaching contents due to the neglect of the needs of the students. On the other hand, the students require more comprehensive knowledge distilled from the teachers to achieve a better performance. Our method is implemented to address these problems, of which the training procedure is divided into two stages: the student training stage and the master training stage. Next, we will take word-level recognition as an example to explain the proposed method.

Student Training Stage. During the student training stage, we only update the student parameters θ_s. The training objective contains two terms: cross entropy loss $\mathcal{L}_{CE}$ and dynamic fusion loss $\mathcal{L}_{DF}$. $\mathcal{L}_{CE}$ controls the classification accuracy, and $\mathcal{L}_{DF}$ is used to match the logits between the teacher and the student.

Given a one-hot label $y = [y_1, y_2, ..., y_K]$, here K indicates the size of vocabulary as defined in Sec 3.1, the cross-entropy loss $\mathcal{L}_{CE}$ between the prediction $\hat{y} = [\hat{y}_1, \hat{y}_2, ..., \hat{y}_K]$ and label y is defined as follows:

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{k=1}^{K} y_k \log \hat{y}_k. \tag{1}$$

Note that in sentence-level prediction, Eq. (1) is corresponding to the loss generated by a character, and the same is true for subsequent other losses.

Regarding the dynamic fusion loss $\mathcal{L}_{DF}$ which is different from traditional distillation loss [15], extra cross-modal guidance is affiliated within its design. We will concretely discuss it soon.

Therefore, the overall objective function $\mathcal{L}_s$ is formulated as follows:

$$\mathcal{L}_s = \mathcal{L}_{CE}\big(y, f_s(X_V; \theta_s)\big) + \lambda_s \mathcal{L}_{DF}, \tag{2}$$

where λ_s indicates a balance factor for regularization, and the optimized parameters θ_s^* is then calculated by

$$\theta_s^* = \arg\min_{\theta_s} \mathcal{L}_s. \tag{3}$$

Cross-Modal Gap Fusion. Distilling knowledge from audio modality to video modality for lip reading is arguably necessary, since the ambiguities caused by different phonemes with almost identical visemes could be avoided. However, a derived trouble is the existed cross-modal gap, which is verified in Table 1. On the other hand, we observe from Table 1 that a video teacher can also promote the performance of the video student. This phenomenon may be attributed to a transferring of more distilled knowledge.

Based on these facts, we drive the teacher outputs different types of knowledge, *i.e.*, audio knowledge, video knowledge, and audio-visual knowledge, for further distillation. This design makes it possible for the student to make its own trade-offs on which modality to learn more from. Then the problem turns into that, how to fuse the cross-modal knowledge.

We tackle this issue by introducing two pretrained audio and video tutors, *i.e.*, $tutor_A$ and $tutor_V$. We exploit the fixed features generated by the tutors as guidance, and encode them into weighting factors w_A and w_V for measuring the preference of different types of knowledge from audio or video modality. Note that we use the output feature vectors $\{H_A, H_V\}$ of the backends of the tutors here for richer representations, instead of the logits produced by the classifiers. We further present a *dynamic fusion loss* $\mathcal{L}_{DF}$ as a regularizer during the student training stage, that is

$$\begin{aligned}\mathcal{L}_{DF} \;=\; & \mathcal{L}_F\big(f_s(X_V;\theta_s), f_m(X_A,X_V;\theta_A,\theta_V)\big) \\ & + w_A\mathcal{L}_F\big(f_s(X_V;\theta_s), f_m(X_A;\theta_A)\big) \\ & + w_V\mathcal{L}_F\big(f_s(X_V;\theta_s), f_m(X_V;\theta_V)\big).\end{aligned} \tag{4}$$

Here we use the focal loss $\mathcal{L}_F(\cdot)$ [17] to ease the difficulty-imbalance and class-imbalance (in the sentence-level scenario) problems. We calculate two weighting factors $\{w_A, w_V\}$ by

$$\begin{aligned}& H'_A = \text{FC}(H_A;\theta_{FA}), H'_V = \text{FC}(H_V;\theta_{FV}), \\ & w = \phi\big(\text{FC}(H'_A \oplus H'_V;\theta_{FAV})\big), \\ & w_A = w,\ w_V = 1-w,\end{aligned} \tag{5}$$

where $\text{FC}(\cdot;\theta_*)$ denotes a fully connected layer with parameters θ_*, $\oplus$ denotes the concatenation operation, and $\phi(\cdot)$ indicates the sigmoid function. All the parameters of the FCs are trained as part of the master.

Master Training Stage. During the master training stage, the values of the parameters θ_s are fixed. Nevertheless, we introduce a temporarily updated student network $f_{ts}(\cdot;\theta_{ts})$ as an auxiliary module. The rationale behind this setting is that, the master needs to receive a task-specific (*i.e.*, lip reading task) feedback for adaptively adjusting its teaching contents, meanwhile the feedback from the student should be renewed with the update of the master. In other words, $f_{ts}(\cdot;\theta_{ts})$ and $f_m(\cdot;\theta_m)$ are alternatively updated during this stage, where $\theta_m = \{\theta_A, \theta_V, \theta_{FA}, \theta_{FV}, \theta_{FAV}\}$.

Algorithm 1 Learning from the master

Input: Training audio, video, label pairs $\{X_A^n, X_V^n, y^n\}$; The initialization of parameters θ_m, θ_s, α, λ_s, λ_m, G_0, P, ξ; The pretained $tutor_A$ and $tutor_V$.
Output: The master $f(\cdot;\theta_m^*)$; The student $f(\cdot;\theta_s^*)$.

```
repeat
    while training the student do
        Sample training audio, video, label pairs {X_A, X_V, y};
        Calculate the logits of the student f_s(X_V; θ_s);
        Calculate the logits of the master f_m(X_A; θ_A),
        f_m(X_V; θ_V), f_m(X_A, X_V; θ_A, θ_V);
        Update the parameters θ_s by Eq. (3).
    end while
    while training the master do
        Sample training audio, video, label pairs {X'_A, X'_V, y'};
        Update the parameters θ_ts by Eq. (6);
        Resample training audio, video, label pairs {X''_A, X''_V, y''};
        Update the parameters θ_m by Eq. (7).
    end while
until converged
```

Specifically, the temporary student first learns from the master, and is governed by the same loss $\mathcal{L}_s$. The updated parameters θ_{ts} is given by

$$\theta_{ts} = \theta_s - \alpha\frac{\mathrm{d}\mathcal{L}_s}{\mathrm{d}\theta_s}, \tag{6}$$

where α represents the learning rate.

Then the master is trained according to both data labels and the requirement of the updated student. We define the optimization of the master as follows:

$$\begin{aligned}\min_{\theta_m}\; & \mathcal{L}_{CE}\big(y, f_{ts}(X_V;\theta_{ts})\big) \\ & + \lambda_m\Big(\mathcal{L}_{CE}\big(y, f_m(X_A,X_V;\theta_A,\theta_V)\big) \\ & + \mathcal{L}_{CE}\big(y, f_m(X_A;\theta_A)\big) \\ & + \mathcal{L}_{CE}\big(y, f_m(X_V;\theta_V)\big)\Big),\end{aligned} \tag{7}$$

where λ_m denotes a balancing factor. Note that the parameters of the FCs are updated by the first term, and here we omit the derivation based on the chain rule.

In practice, the updates of the temporary student and the master adopt different training samples. Because when updating the master, the student should utilize a validation set instead of a training set. Finally, the algorithm of "Learning from the Master" is summarized in Algorithm 1.

3.4. Curriculum Learning

Previous lip reading methods [1, 7, 27] are proposed to randomly sample training data first and then feed them into

the network. However, the order of training samples is ignored, which may influence the effectiveness of training process. A more desired scheme is to make the framework learn from an easy start and gradually increase the difficulty to facilitate a better convergence. Therefore, we incorporate a curriculum learning [4, 13] based strategy into the training stage of our approach. It is implemented by two steps.

First, we build a *rating function* $\mathcal{R}(\cdot)$ to evaluate the difficulty of each sample pair X_A^n, X_V^n (with the same label), that is

$$\mathcal{R}(X_A^n, X_V^n) = \text{sort}\Big(C\big(f_{t_A}(X_A^n)\big) + C\big(f_{t_V}(X_V^n)\big)\Big), \quad (8)$$

where $C(\cdot)$ denotes the confidence, sort$(\cdot)$ indicates a sorting operation. It is easy to understand that the higher the function value, the easier the sample pair. Note that when several sample pairs have the same rating value, we resort them according to the video modality based confidence $C\big(f_{t_V}(X_V^n)\big)$.

Given a series of training data sorted by their difficulties, we secondly introduce a fixed exponential *pacing function* [13] to establish the increment of samples during training, which is given by

$$G_i = \min(G_0 \times P^{\lfloor \frac{i}{\xi} \rfloor}, 1), i > 0, \quad (9)$$

where G_i indicates the input percentage of training data in the ith iteration (G_0 is the percentage at the start), P is an exponential factor, and ξ is the number of iterations in each step.

Based on the rating function and the pacing function, the order and the increment of the training samples are determined more sensibly. Such a strategy could reduce the learning ambiguities at the beginning of training and also enable a better convergence. More discussions will be presented in Sec. 4.4.

4. Experiment

4.1. Experimental Datasets and Preprocessing

To evaluate the proposed method, we utilize three benchmark datasets, *i.e.*, one word-level dataset LRW [8] and two sentence-level datasets LRS2-BBC [1], LRS3-TED [2].

LRW. Lip Reading in the Wild dataset is a large word-level dataset with 500 words and 450000 utterances. Each video is 1.16 second long with 29 frames.

LRS2-BBC. The Lip Reading Sentence dataset comes from BBC talks including 143 kilo utterances and 2.3 million words. The dataset is split into pretrain/train-val/test data, while train-val and pretrain data comes from the same date and pretrain data contain a word-level boundary alignment.

LRS3-TED. LRS3-TED dataset comes from TED talks with 150 kilo utterances and over 4.2 million words. This is the most difficult audio-video lip reading dataset. The dataset construction is also divided into the same setting as LRS2-BBC.

Preprocessing. To crop aligned center mouth areas of the video data, we first use dlib [5] to detect the facial landmarks. Then we randomly crop and interpolate the results to obtain 112×112 lip-centered images. Lastly, we transform all faces for removing rotations and different scalings.

4.2. Implementation Details and Metric

Vocabulary size. In the word-level prediction, the size of the vocabulary is set to 500, which is in line with the number of words in LRW. Regarding the sentence-level scenarios, *i.e.*, LRS2-BBC and LRS3-TED, we set the size of vocabulary as 40, including 26 letters, 10 digits, and 4 special tokens ([SPACE], [PAD], [EOS] and punctuation mark).

Training protocol. The student and the master are trained alternatively using SGD optimizer with a momentum of 0.9 and a weight decay of 1e-4. In audio stream, we take the raw wave as input. While in visual stream, the input videos are sampled at 25fps. We follow the same data augmentation and audio-subtitle forced alignment strategies as in Afouras *et al*. [1], and training together with the proposed curriculum learning described in Sec.3.4.

The whole training process includes both pretraining and fine-tuning steps. Specifically, we first pretrain our model with a backend of TC at word level, using LRW and the pretrain sets of LRS2-BBS and LRS3-TED. In regard to the word-level prediction, we then fine-tune the pretrained model with LRW. While for the sentence-level prediction, we replace TC with TM-Seq2Seq as the backend in the pretrained model, and continue training with the pretrain set of LRS2-BBS or LRS3-TED. The new pretrained model is then fine-tuned with the related train-val set.

The learning rate α is set to 10^{-3} during pretraining. During fine-tuning, α is initialized to 10^{-4} and decreased by half every time the validation error flats, down to a final learning rate 10^{-6}. The other hyperparameters in Algorithm 1 are set as follows: $\lambda_s = 10, \lambda_m = 10, G_0 = 0.25, P = 1.75, \xi = 10^7$.

Evaluation metric. For all the experiments, we adopt the word error rate (WER) as the measurement, which is defined as WER $= \frac{S+D+I}{NUM}$, where S, D, I is the number of substituted, deleted, and inserted words respectively to get from the reference to the inference, and NUM is the total number of words in the reference.

4.3. Comparisons with state-of-the-arts

We compare our method with several state-of-the-art methods, including MT [8], Temporal Convolution [23], WAS [7], Bi-LSTM [23], TM-CTC [1], TM-Seq2Seq [1], Conv-Seq2Seq [27], LIBS [29], and TM-CTC-KD [31].

Table 2: Quantitative evaluation (WER) with the state-of-the-art methods on LRW, LRS2-BBC, and LRS3-TED. Note that Ours-TC and Ours-TM respectively indicate model (a) and model (b) described in Sec 3.2.

Method	WER(%)		
	LRW	LRS2	LRS3
MT [8]	38.9	-	-
Tem. Conv. [23]	25.4	-	-
WAS [7]	23.8	70.4	
Bi-LSTM [23]	17.0	-	-
TM-CTC [1]	-	65.0	74.7
TM-Seq2Seq [1]	-	49.8	59.9
Conv-Seq2. [27]	16.3	51.7	60.1
LIBS [29]	-	65.3	-
TM-CTC-KD[31]	-	51.3	59.8
(a) Ours-TC	18.7	-	-
(b) Ours-TM	**14.3**	**49.2**	**59.0**

Unless stated otherwise, all the reported results are directly copied from the original papers.

Word-Level Lip Reading. Table 2 reports the quantitative comparisons with existing approaches from LRW dataset, regarding word-level lip reading. It can be seen that, Ours-TC significantly outperforms its baseline Temporal Convolution which is without the master module, with a WER improvement of 6.7%. Besides, Ours-TM achieves the best performance compared to all the other methods. In particular, it acquires an increase of 2% in WER compared to that of the second best method Conv-Seq2Seq.

Sentence-Level Lip Reading. We also performed experiments on sentence-level lip reading, and the results are listed in the last two columns of Table 2. It can be observed that Our-TM performs the best on both LRS2-BBC and LRS3-TED in comparison with the other methods. More importantly, compared to TM-Seq2Seq, which incorporates a same backend as Ours-TM and is trained on an extra non-public dataset MV-LRS, our method achieves a better performance with WER improvements 0.6% on LRS2-BBC and 0.9% on LRS3-TED in the case of using much less training data. Furthermore, compared to Conv-Seq2Seq, which uses a more advanced structure than that of our task-specific network (the student), Ours-TM still achieves better performance, with WER improvements 2.5% on LRS2-BBC, 1.1% on LRS3-TED.

Tendency of misrecognition. We further investigate the top-4 cases on LRW with the highest error rate and list the comparison results of Ours-TC w/o KD and Ours-TC in Table 3. It can be observed that when multiple phonemes are mapped to one viseme, *e.g.*, phonemes TH and DH vs. viseme /t/, our method can reach an average improvement of nearly 6% in accuracy with the comparison.

Combining all the above observations, it implies that: (i) Introducing the proposed master can effectively promote the performance of our task-specific network. (ii) Though our model mainly focuses on the benefits over standard distillation methods, it has the potential to achieve better performance when replacing the architecture of the task-specific network with a more advanced one.

Table 3: The error rate (%) of misrecognition on LRW.

Method	Top-4 cases			
	'THESE'	'THERE'	'THING'	'UNDER'
Ours-TC w/o KD	74%	70%	70%	66%
Ours-TC	70%	59%	68%	60%

Table 4: WER(%) results for the teachers, the masters and the students learned from different pretrained teachers or co-evolved masters.

Method	Distilled from	LRS2-BBC
Audio Teacher	✗	17.2
Student1	Audio Teacher	54.2
Video Teacher	✗	57.5
Student2	Video Teacher	53.4
Audio-Visual Teacher	✗	15.6
Student3	Audio-Visual Teacher	54.1
Audio Master	✗	19.1
Student4	Audio Master	52.1
Video Master	✗	59.1
Student5	Video Master	53.0
Audio-Visual Master	✗	16.9
Student6	Audio-Visual Master	**51.5**

4.4. Ablation Study

In this section, we investigate the effectiveness of our proposed modules, including the master network, the cross-modal gap fusion, and the curriculum learning strategy. We use a single-modal lip-reading network as baseline, *i.e.*, Video Teacher in Table 4.

Effectiveness of the Master. To explore the efficacy of the master, we examined 6 different pairs of teacher-student or master-student designs based on different modalities, and tested their performances on LRS2-BBC. The results are summarized in Table 4. Note that the reported performance of the Audio-Visual Master comes from its AV branch, and the architectures of the pretrained teachers are totally the same as those of their counterpart trainable masters. Besides, the curriculum learning strategy is not used here. Then we have the following observations and analyses. (i) In the case of single models without KD, no matter whether the models are trainable (*i.e.*, the masters) or not (*i.e.*, the teachers), the descending order of their performance in different modalities is always {audio-visual modality (AV), audio modality (A) and video modality (V)}. This verifies the importance of learning from cross-modal data instead of

Table 5: WER (%) results with different weight settings.

Method	Weight w.r.t. modality AV : A : V	LRS2-BBC
Baseline	0 : 0 : 0	57.5
T1	1 : 0 : 0	54.1
T2	0 : 1 : 1	53.1
T3	1 : 1 : 1	53.2
M1	1 : 0 : 0	51.5
M2	0 : 1 : 1	51.9
M3	1 : 1 : 1	50.4
M4	1 : w_a: w_v	**49.6**

single-modal data. (ii) In the cases of distilling knowledge from the teachers and the masters, the descending orders of the performance of the students in different modalities are always {V, AV, A} and {AV, A, V}, respectively. The first sorting order implies that, compared to audio modality, audio-visual modality could provide additional information and thus help to alleviate the ambiguities caused by the cross-modal gap, but using a simple fusion strategy (concatenation) is limited. Whereas the other sorting order reveals the effectiveness of the master that could reduce the cross-modal gap to some extent, due to its dynamic regulation based on the task-specific feedback of the student. (iii) No matter which modality is used, the teachers always perform better than their counterpart masters. While no matter which modality is used, the students learned from the masters perform better than those learned from the teachers all the time. These facts demonstrate that the proposed master is more effective than a pretrained teacher because of its adaptability to the student, although a sacrifice in its own performance is obtained.

Effectiveness of Cross-Modal Gap Fusion. A simple fusion strategy by concatenating the features from cross modalities shows a limitted improvement, which is verified in Table 4. This motivates us to come up with the design of cross-modal gap fusion. Further, to evaluate the effectiveness of this module, we examined 7 different and representative weight settings and reported the results in Table 5. Note that "T" and "M" respectively indicate that the student is distilled from a teacher or a master, and the curriculum learning strategy is also not used here. We have the following observations and analyses: (i) T1 performs worse than T2, and T2 performs similarly to T3. It seems that introducing AV modality is not a better choice compared to incorporate {A,V}; (ii) However, both M1 and M3 perform better than M2. This reveals the importance of the interaction between audio stream and visual stream in the master, which confirms the necessity of AV. Furthermore, A and V are also useful but should be treated dynamically rather than equally. As a result, M4 achieves the best performance

Table 6: Quantitative evaluation (WER) with more advanced backends on LRW. B.-2 and B.-8 are respectively short for baselines with 2 and 8 TC layers. Training time is reported in parentheses.

Method	B.-2	Ours-TC	B.-8	Ours-TC	[18]	Ours-TC
Distilled?	✗	✔	✗	✔	✗	✔
w/o CL	-	19.8(70h)	-	14.0(77h)	-	11.9(81h)
w/ CL	24.3	18.7(64h)	17.0	13.5(71h)	14.7	11.8(75h)

compared to other methods, proving the effectiveness of the proposed cross-modal gap fusion module. Note that we fix the weight of AV to avoid increasing the system complexity.

Effectiveness of Curriculum Learning on various backends. The proposed method has a specific focus on distillation performance and thus a lightweight structure (2 TC layers in the backend) is used. However, our work can be used as a plug-in for any state-of-the-art method, like [18] which uses a more advanced backend. To demonstrate the effectiveness of our curriculum learning (CL) strategy, as well as the superiority of our framework, Table 6 reports the corresponding comparisons that replacing our backend from 2 TC layers to 8 TC layers and the one used in [18]. Note that here we only use LRW as training data, which is the same training setting as in [18]. It can be seen that, when using a lightweight model, CL would be helpful to improve the performance of our method as well as acceralating convergence. Nonetheless, when replacing with an advanced model, the improvement of CL degrades, and it mainly helps to accelerate the training. Specifically, our method can achieve better performance against the comparisons, and still outperforms [18] when without CL. This reveals the power of our CL strategy and tailored distillation approach.

5. Conclusion

In this paper, we propose a novel framework for lip reading based on knowledge distillation. Instead of distilling knowledge from a pretrained audio teacher, we particularly introduce a trainable master into our model. With the aid of the task-specific feedback from the student, the cross-modal gap fusion module, and the curriculum learning strategy, the master could adjust its teaching contents to the student adaptively. Extensive experiments demonstrate the validity of the proposed method.

Acknowledgements: This work is supported by NSF Grant 61972162, 61702194; the China Postdoctoral Science Foundation under Grant No. 2020M682240; the Fundamental Research Funds for the Central Universities under Grant No. 202113035, and the CCF-Tencent Open Research fund (CCF-Tencent RAGR20190112).

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE TPAMI*, 2018. 1, 2, 3, 4, 5, 6, 7

[2] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv:1809.00496*, 2018. 1, 6

[3] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv:1611.01599*, 2016. 3

[4] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *ICML*, pages 41–48, 2009. 6

[5] Donatella Castelli and Pasquale Pagano. Opendlib: A digital library service system. In *ICTPDL*, pages 292–308, 2002. 6

[6] Greg I Chiou and Jenq-Neng Hwang. Lipreading from color video. *IEEE TIP*, 6(8):1192–1195, 1997. 1

[7] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017. 5, 6, 7

[8] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, pages 87–103, 2016. 1, 2, 6, 7

[9] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pages 1607–1616, 2018. 3

[10] Alan J Goldschen, Oscar N Garcia, and Eric D Petajan. Continuous automatic speech recognition by lipreading. pages 321–343, 1997. 1

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006. 3

[12] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *CVPR*, pages 11020–11029, 2020. 3

[13] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *ICML*, pages 2535–2544, 2019. 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 2, 3

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv:1503.02531*, 2015. 1, 3, 4

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[17] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. 5

[18] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Lipreading using temporal convolutional networks. In *ICASSP*, pages 6319–6323, 2020. 1, 8

[19] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition. *IEEE TASLP*, 17(3):423–435, 2009. 1

[20] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *ICLR*, 2015. 3

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 2

[22] Guocong Song and Wei Chai. Collaborative learning for deep neural networks. In *NeurIPS*, pages 1832–1841, 2018. 3

[23] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *Interspeech*, pages 3652–3656, 2017. 1, 2, 3, 6, 7

[24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NeurIPS*, pages 3104–3112, 2014. 3

[25] Chenglin Yang, Lingxi Xie, Siyuan Qiao, and Alan L Yuille. Training deep neural networks in generations: A more tolerant teacher educates better students. In *AAAI*, volume 33, pages 5628–5635, 2019. 3

[26] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv:1612.03928*, 2016. 3

[27] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *ICCV*, pages 713–722, 2019. 2, 5, 6, 7

[28] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 3

[29] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *AAAI*, pages 6917–6924, 2020. 1, 3, 6, 7

[30] Xiatian Zhu, Shaogang Gong, et al. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, pages 7517–7527, 2018. 3

[31] AP Zisserman, T Afouras, and JS Chung. Asr is all you need: cross-modal distillation for lip reading. In *ICASSP*, 2020. 1, 3, 6, 7