This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

Importance-Aware Information Bottleneck Learning Paradigm for Lip Reading

Changchong Sheng, Li Liu*, *Senior Member, IEEE*, Wanxia Deng, Liang Bai, Zhong Liu*, Songyang Lao, Gangyao Kuang, *Senior Member, IEEE,* and Matti Pietikäinen, *Fellow IEEE*

***Abstract*—Lip reading is the task of decoding text from speakers' mouth movements. Numerous deep learning-based methods have been proposed to address this task. However, these existing deep lip reading models suffer from poor generalization due to overfitting the training data. To resolve this issue, we present a novel learning paradigm that aims to improve the interpretability and generalization of lip reading models. In specific, a Variational Temporal Mask (VTM) module is customized to automatically analyze the importance of frame-level features. Furthermore, the prediction consistency constraints of global information and local temporal important features are introduced to strengthen the model generalization. We evaluate the novel learning paradigm with multiple lip reading baseline models on the LRW and LRW-1000 datasets. Experiments show that the proposed framework significantly improves the generalization performance and interpretability of lip reading models.**

***Index Terms*—Lip Reading, Information Bottleneck, Visual Speech Recognition, Deep Learning.**

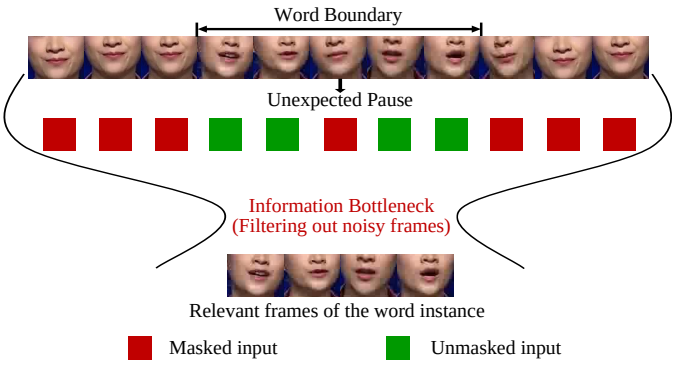


Fig. 1. An illustration of the information bottleneck-based temporal mask framework. A Boolean mask is introduced to automatically filter out noisy frames and force the network to make predictions depending on task-relevant frames.

I. Introduction

AUTOMATIC Lip Reading (ALR), also known as Visual Speech Recognition (VSR), aims to decode speech from speakers' mouth movements. As an emerging and challenging topic in the inter-field of computer vision and natural language processing, VSR has been attracting increasing attention in recent years due to its significant role in many academic and practical applications, including multimodal speech recognition and enhancement, speaker recognition and verification [1], medical assistance, security, entertainment, human-computer interaction, *etc*.

To provide some application examples, in the field of speech recognition and enhancement, visual speech can be treated as a complementary signal to increase the accuracy and robustness of current audio speech recognition and separation under a range of unfavorable acoustic conditions [2], [3], [4], [5]. In the medical domain, solving the VSR task can also help the hearing impaired [6] and people with vocal cord lesions. In public security, VSR can be applied to face forgery detection [7] and liveness detection [8]. In human-computer interaction, visual speech can serve as a new type of interactive information that improves the diversity and robustness of interactions [9], [10]. In the entertainment domain, VSR can be used to transcribe archival silent films. Recently, advances in deep learning technology, along with the availability of large-scale audio-visual speech datasets [11], [12], [13] have tremendously boosted the development of VSR.

Given a talking face video, a VSR system first crops the video and obtains the mouth-centered cropped video. Subsequently, it decodes the cropped video into a specific type of text (words, phrases or sentences). According to its recognition targets, VSR can be devided into two subtypes: word-level and sentence-level. Word-level VSR aims to classify the input video into one of a set of predefined word categories; For its part, sentence-level VSR attempts to predict consecutive sentences from the input video. Despite the significant progress made in this area, some issues remain unsolved. Among the most critical of these issues is that task-irrelevant frame-level noises are widespread in talking videos of most existing lip reading datasets [11], [12]. Due to the limited training data and the widespread frame-level noise, existing deep lip reading models have historically suffered from serious overfitting. For example, Feng *et al*. [14] proved that the presence of prediction-irrelevant information (*e.g*., video frames out of the actual word boundary) will significantly impair the models' prediction performance. To address this issue, they introduced additional word boundary annotations to improve the model training. However, these word boundary annotations are fairly costly to attain, especially for the more challenging (sentence-level) lip reading task. In addition, some noisy frames caused by unexpected pauses, stutters, repetitions, *etc*., can not be modeled by the word boundary, making the problem more

Changchong Sheng, Wanxia Deng and Gangyao Kuang are with the College of Electronic Science and Technology, National University of Defense Technology (NUDT), China. Li Liu, Liang Bai, Zhong Liu and Songyang Lao are with the College of Systems Engineering, NUDT. Matti Pietikäinen is with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland.

Li Liu (lilyliu_nudt@163.com) and Zhong Liu (liuzhong@nudt.edu.cn) are the corresponding authors.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

difficult.

In this paper, we focus on the word-level lip reading task. Different from previous works that are devoted to the design of model backbones, our goal is to suppress frame-level noises and improve both the interpretability and the generalization of baseline models from an information-theoretic perspective. As shown in Fig. 1, the core idea of this work is to force the network to automatically filter out noisy frames and make decisions that are primarily based on task-relevant frames.

To achieve this, we introduce Information Bottleneck (IB) [15], [16], [17], [18] technology to aid us in realizing noisy frame analysis. IB aims to learn an encoding that is maximally expressive about the target while maximally compressing the original data. Inspired by the idea of IB, we introduce the deep information bottleneck principles to explore the attribution importance of video frames. By applying this approach, the temporal attribution importance learning can be treated as a problem of optimizing the IB trade-off.

In general terms, a deep learning-based VSR architecture consists of two sub-networks: a visual frontend network and a sequence backend network. In this paper, a variational temporal mask (VTM) module is developed based on information theory, then inserted between the visual frontend network and the sequence backend network so that it can be trained jointly with the whole model. Notably, this architecture design can guarantee decoupling of the VTM module from the visual frontend network and the sequence backend network, making the VTM module transplantable and applicable to arbitrary baseline network structures. The VTM module attempts to restrict the unimportant or noisy visual features flowing to the sequence backend network, which forces the model to make decisions based on more important visual features. Furthermore, we formalize the specific variational objective function to optimize the VTM module. Experiments show that this simple module can improve the interpretability and prediction performance of multiple different baseline models. Because the VTM module is deployed as a plug-and-play module, it is model-agnostic and can be applied to any lip reading baseline models. The major contributions of this work can be summarized as follows:

- We propose an IB-based variational temporal mask (VTM) module, which mines the importance of frame-level features to improve both model interpretability and prediction accuracy.
- The VTM module, as a plug-and-play module, can be integrated into any lip reading baseline models to facilitate better performance.
- Experiments clearly show that the proposed learning paradigm significantly improves the baseline lip reading models on several large-scale lip reading benchmarks.

The remainder of this paper is structured as follows. The related work of VSR and deep information bottlenecks is summarized in Section II. Section III describes the overall pipeline of our model and the proposed VTM module. The ablation study, comparison with the state-of-the-art methods, and some quantitative results and discussions are presented in Section IV. Finally, Section V concludes the paper.

II. Related Work

A. Visual Speech Recognition

A talking face video contains a large amount of redundant information (such as pose, illumination, gender, *et al.*) that is unrelated to the VSR task; The information most closely related to the VSR task is lip movement. The core research point of VSR tasks, spatial-temporal feature extraction, is to effectively filter out redundant information while retaining as much lip movement information as much as possible.

Before the emergence of deep learning-based VSR methods, researchers focused primarily on traditional feature engineering driven spatial-temporal feature extraction from videos. There are two mainstream types of traditional methods: appearance-based and shape-based. The former uses the pixel value of ROI as the original feature space, then utilizes different data dimension reduction methods to obtain compact and effective feature representations. For dimension reduction methods, linear transformation methods such as Principal Component Analysis (PCA) [19], Discrete Cosine Transform (DCT) [20], Linear Discriminant Analysis (LDA) [20], and Maximum Likelihood Linear Transformation (MLLT) [21] are commonly used; Moreover, optical flow, Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [22], along with manifold learning and graph embedding methods such as Locality Discriminant Graph (LDG) [23], and Random Forest Manifold Alignment (RFMA) [24] are also used for feature extraction. Shape-based methods perform feature extraction based on the shape of the ROI (lips, chin, cheeks, *et al.*). Compared to appearance-based methods, these methods have better interpretability and generalization, although they also require more fine-grained manual annotation. The main attributes of lip contour (height, width *et al.*) or Articulatory Features (AFs) [25], [26] primarily used for small-scale recognition tasks; Furthermore, the Active Shape Model (ASM) [27] is one of the most commonly used shape-based methods that employ facial landmarks to extract spatio-temporal features. In addition, some researchers proposed a more powerful method, the Active Appearance Model (AAM) [28], which further improves the performance by combining appearance-based and shape-based methods. For classifiers, Support Vector Machine (SVM), template matching, Maximum a Posteriori (MAP), and Regularized Discriminant Analysis (RDA) are mainstream classifiers for isolated recognition tasks, while the Hidden Markov Model (HMM) is widely used for continuous recognition tasks.

As for the deep learning-based VSR methods, a deep VSR system typically contains three sub-modules. The first sub-module is visual feature extraction, which aims to extract compact and effective visual feature vectors from mouth-centered cropped videos. The second sub-module is temporal context aggregation, which aims to aggregate temporal context information for better text script decoding and recognition. The above two sub-modules are also the cores of deep learning-based VSR methods. The final sub-module is text decoding, which converts the feature representations to text. Existing works on deep VSR focus primarily on the

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

architecture design of these two sub-networks: visual frontend networks and sequence backend networks.

As for the design of visual frontend networks, many works have utilized deep CNNs to perform visual features extraction. For example, Stafylakis *et al*. [29] proposed a simple variation of ResNet (changing the first 2D convolution layer to 3D convolution layer), referred to as C3R-ResNet. This model consisted of a shallow 3D CNN and a deep 2D CNN, and achieved 83% recognition accuracy for word-level lip reading on the LRW [12] dataset. Due to the promising performance of this model, numerous lip reading models [30], [31] adopted it as the backbone network for visual features extraction. Recently, Feng *et al*. [14] improved this architecture by integrating the Squeeze-and-Extract module. Moreover, deep 3D CNNs have also been used to extract visual features. In [32], the authors successfully migrated the two-stream (the raw grayscale video stream and the dense optical flow stream) I3D model to lip reading, thereby achieving comparable performance on word-level lip reading. However, dense optical flow calculation is very time consuming, resulting in low recognition efficiency. In addition, Graph Convolution Networks (GCNs) [33], [34] and Visual Transformer (VIT) networks [35] have also been explored for VSR. For example, Sheng *et al*. [33] proposed a graph convolution network based module (ASST-GCN) to learn semantic-preserved local visual features and integrate local and global information in order to improve the discrimination of visual features. Afouras *et al*. [35] designed an end-to-end visual transformer-based pooling mechanism that learns to track and aggregate lip movement representations. The proposed visual backbone network can reduce the need for complicated preprocessing, thereby improving the robustness of visual representation.

For the design of sequence backend networks, there are three typical network architectures: RNN based, temporal convolution network (TCN) based and Transformer [36] based. As for the RNN based network, BiGRU and BiLSTM are the most commonly used. For example, Stafylakis *et al*. [29] created the baseline word-level lip reading model with a 2-layer BiLSTM backend network [29], and Feng *et al*. [14] achieved the SOTA performance (48.2%) with a 3-layer BiGRU network on the LRW-1000 dataset. Martinez *et al*. [37] improved the baseline model by replacing the BiLSTM backend with Multi-Scale TCN (MSTCN), achieving SOTA performance (87.9%) with multiple iteration self-distillation on the LRW dataset. Transformer based backend network has not been widely used in word-level lip reading tasks. Compared with RNN based and TCN based networks, these appraoches are more prone to overfitting, as existing public datasets are not large enough.

B. Deep Variational Information Bottleneck

Information Bottleneck, first proposed as a representation compression technique in [16], aims to learn a hidden encoding that is maximally expressive about the target while maximally compressing about the original input. Let X, Y, and Z denote the input, the class probability distribution, and the hidden encoding, respectively. Tishby *et al*. proposed to learn the optimal mapping of the data by solving the following Lagrangian relaxation:

$$min\,\mathcal{L}_{IB} = \min_{p(Z|X)} \beta I(Z,X) - I(Z,Y) \tag{1}$$

Here, $I(\cdot,\cdot)$ represents mutual information, while β is a positive parameter that controls the trade-off between the compression and preservation. In 2015, Tishby *et al*. further pointed out that IB can also be applied on deep neural networks [15]. Specifically, they proposed to analyze DNNs by means of Information Plane visualization, *i.e*., the plane of the Mutual Information values preserved by each layer on the input and output variables. However, they did not provide any experimental results because the iterative Blahut-Arimoto algorithm cannot feasibly be applied to DNNs.

Unfortunately, optimizing the IB principle applied to DNNs remains a difficult problem for the following reasons: 1) both the compression and encoding objectives are non-convex, meaning that there is no guarantee that a global optimum can be found; 2) the computation of mutual information is intractable. To address the drawback, some IB variations were proposed recently. For example, Strouse *et al*. [38] proposed the deterministic information bottleneck (DIB), which significantly outperforms the IB in terms of the cost function and computational efficiency by optimizing the compression term in the original IB. Kolchinsky and Tracey [39] identified the three caveats in any situation in which Y is a deterministic function of X, and proposed a simple modification of the IB functional that can recover the IB curve in all cases. Moreover, as previous IB has been mainly developed for limited cases, Kolchinsky *et al*. [40] proposed Nonlinear IB (NIB) for performing IB in much more general settings, *i.e*., both inputs and outputs can be either discrete-valued or continuous-valued, and with any desired joint distribution.

In 2016, Alemi *et al*. [17] proposed Deep Variational Information Bottleneck (Deep VIB). Deep VIB utilizes variational inference to approximate the bounds on mutual information and obtain the final variational upper bound of the IB objective function, which makes the IB objective easier to apply on DNNs. In addition to information compression, deep VIB can also be useful in many important fields of DNNs, such as graph representation learning [41], [42], disentangled representation learning [43], [44], generative adversarial learning [45], [46], network compression and quantization [47], [48], Semantic Segmentation [49], [50], and the interpretability and generalization of DNNs [18], [51], [52]. It should be noted here that Luo *et al*. [49], [50] proposed the significance-aware information bottleneck (SIB) for feature purification. This work has a similar motivation to [49], [50], but for different tasks and goals.

In this paper, we focus on Deep VIB. To give an example of the interpretability and generalization analysis of Deep VIB, Schulz *et al*. [18] proposed to adopt deep VIB for the attribution analysis of pretrained image classification DNNs. The authors proposed a novel IB-based attribution (IBA) method that estimates the amount of information (relevance score) provided by an image region to the network's final prediction.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

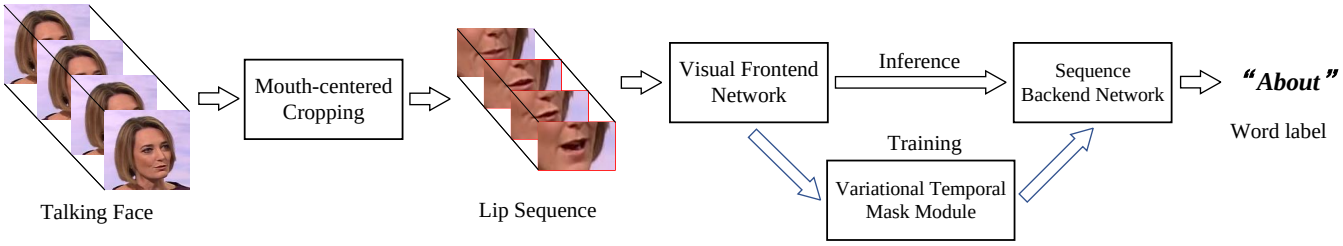


Fig. 2. The flowchart of the proposed importance-aware information bottleneck learning paradigm.

Specifically, they restricted the flow of information and quantified (in bits) how much information is provided by image regions for classification by adding noise to feature maps.

This work is motivated by [18] but goes beyond the post-hoc interpretability analysis. In more detail, we aim to improve both prediction accuracy and interpretability without resorting to additional human annotations like word boundary annotation. We will provide a more detailed description of how IB formulation in our work.

III. Proposed Method

This section introduces the proposed importance-aware information bottleneck learning paradigm for the word-level lip reading. Firstly, we introduce the general pipeline for the word-level lip reading task. And the novel learning framework based on the general pipeline is described in detail. Then, the Variational Temporal Mask (VTM) is formulated within the information bottleneck principle, which is the core idea underpinning the importance-aware information bottleneck learning paradigm.

A. The Overall Pipeline

In terms of the general training framework, when given a video of a talking face, an essential element of preprocessing for lip reading is mouth-centered region of interest (ROI) cropping. Once the aligned and mouth-centered cropped input video has been generated, the objective of the visual front-end network is to extract visual spatio-temporal features that represent visual speech patterns and dynamics. The visual front-end network has a relatively small receptive field on the temporal dimension, making it insufficient for lip reading tasks. The sequence back-end network focuses on further aggregating long-term temporal contextual information. The proposed importance-aware information bottleneck learning paradigm is illustrated in Fig. 2. Compared with the general flowchart of supervised word-level lip reading models, we introduce the novel VTM module between the visual frontend network and sequence backend network. As the proposed VTM module is only used in the training stage, there is no extra memory or time required in the inference stage.

Fig. 3 provides a more detailed pipeline of our learning paradigm. The upper half (shaded in gray) is the general pipeline of general deep lip reading models, consisting of a visual frontend network E^V, a sequence backend network E^S, a Global Average Pooling (GAP) layer and a linear classifier. The output dimension of the final linear classifier layer is equal to the total number of word classes.

To facilitate better understanding, we formalize the lip reading task here. Given a lip-centered video $I_{1:T} = [i_1, i_2, ..., i_T]$ with T frames, where $i_t \in \mathbb{R}^{W \times H \times 1}$ is the gray image. The visual frontend network E^V aims to extract frame-wise continual visual features $X = [x_1, x_2, ..., x_T] \in \mathbb{R}^{T \times C}$ from $I_{1:T}$, where C is the feature dimension. Next, the sequence backend network E^S aggregates the temporal information and outputs $H = [h_1, h_2, ..., h_{T'}] \in \mathbb{R}^{T' \times C}$. Finally, the global pooling features are sent to the linear classifier for final prediction.

Based on the general pipeline, we propose a more optimized training framework that aims to improve the interpretability and generalization of baseline models. As shown in Fig. 3, assuming that the visual frontend network E^V is properly pre-trained[1], we introduce the VTM module to automatically learn the importance of frame-level features.

Moreover, the output of the VTM module is $Z = [z_1, z_2, ..., z_T]$, *i.e.*, the binary masked version of raw frame-level features X. The VTM module attempts to filters out task-irrelevant frame-level features without reducing the prediction accuracy, thereby making the network more interpretable.

Beyond interpretability, we also expect the network to be more generalizable. To achieve this goal, a contrastive loss $\mathcal{L}_M$ is introduced that forces the network to make similar predictions under the raw features X and the masked features Z. We use Kullback-Leibler divergence to measure the prediction differences, it can be written as follows:

$$\mathcal{L}_M = \mathrm{KL}[p(Y|X), p(Y|Z)] \tag{2}$$

Here, Y is the probability distribution over word classes, while $\mathrm{KL}[\cdot, \cdot]$ denotes Kullback-Leibler divergence. The core idea of the proposed training framework is to automatically learn the importance of frame-wise features and force the network to make decisions based on important features. In the inference stage, only the global prediction is required. That is to say, the VTM module is inserted only to assist with training, and no additional memory or computation costs are incurred.

B. Variational Temporal Mask

In this section, we first introduce the background of the variational inference applied on the information bottleneck.

[1]https://github.com/Fengdalu/learn-an-effective-lip-reading-model-without-pains

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

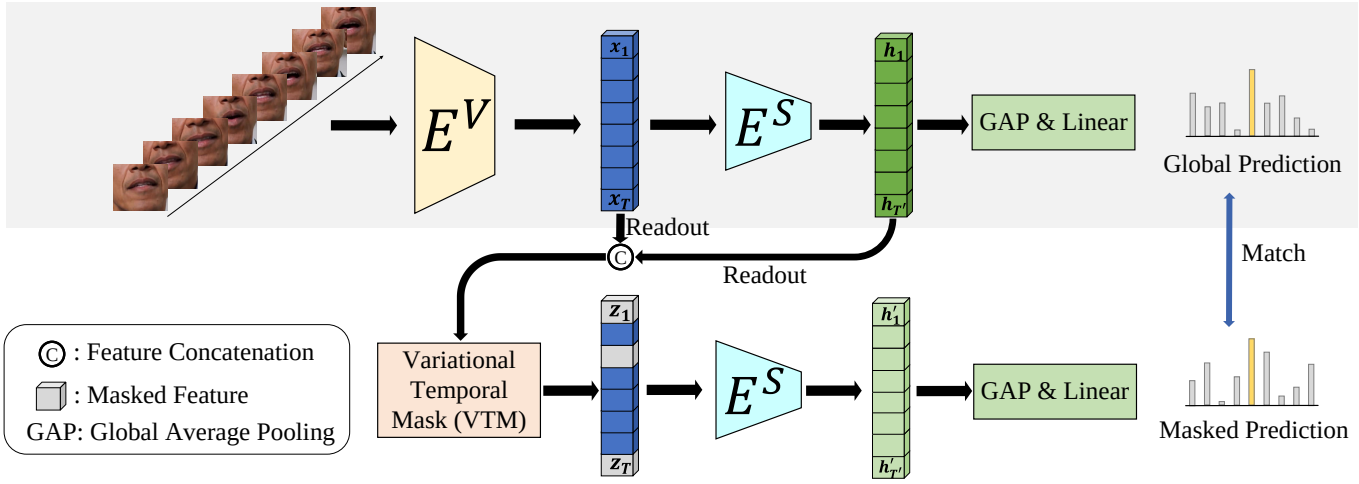


Fig. 3. The overall pipeline of the proposed importance-aware information bottleneck training framework. The upper half (shaded in gray) is the general pipeline of word-level lip reading. E^V is the visual frontend network, and E^S is the sequence backend network. Networks marked in the same color have shared parameters. The VTM module controls which features are selected based on the attribution importance analysis.

Subsequently, we will explain how the VTM module works in detail.

1) Variational Information Bottleneck: Let Y and Z denote the probability distribution over word classes and the output of a hidden layer, respectively. To ensure that Z contains enough information to predict Y while containing less redundant information from X, we follow the standard formulation in the information bottleneck theory [16] with the following objective function:

$$\mathcal{L}_{IB} = \beta I(X,Z) - I(Y,Z) \tag{3}$$

Here, $I(\cdot,\cdot)$ is the mutual information, while β is a hyperparameter that controls the trade-off between predicting Y and compressing X. Eq. 3 gives an intuitive optimization goal on Z; the main associated challenge lies on the intractable computation of $I(X,Z)$ and $I(Y,Z)$.

Rather than computing $I(X,Z)$ and $I(Y,Z)$ directly, we follow the variational information bottleneck (VIB) proposed by [17]. The mutual information between Y and Z can be described as follows:

$$I(Y,Z) = \int dydzp(y,z)\log\frac{p(y|z)}{p(y)} \tag{4}$$

Since $p(y|z)$ is intractable in this case, let $q(y|z)$ be a variational approximation of $p(y|z)$. Based on the fact that the Kullback-Leibler divergence is always positive, we have the following:

$$\begin{aligned}&\mathrm{KL}[p(Y|Z),q(Y|Z)] \geq 0 \Rightarrow \\ &\int dyp(y|z)\log p(y|z) \geq \int dyp(y|z)\log q(y|z)\end{aligned} \tag{5}$$

The variational lower bound of $I(Y,Z)$ is then constructed as follow:

$$\begin{aligned}I(Y,Z) &\geq \int dydzp(y,z)\log\frac{q(y|z)}{p(y)} \\ &= \int dydzp(y,z)\log q(y|z) + H(Y)\end{aligned} \tag{6}$$

$H(Y)$ is the entropy of the label distribution, which is independent of the optimization, and can thus be ignored. Based on the Markov assumption[2] of X, Z, Y, we have:

$$p(y,z) = \int dxp(x)p(y|x)p(z|x) \tag{7}$$

Thus, Eq. 6 can be rewritten as follows:

$$\begin{aligned}I(Y,Z) &\geq \int dxdydzp(x)p(y|x)p(z|x)\log q(y|z) + H(Y) \\ &= \mathbb{E}_{z\sim p_\theta(z|x)}[\log q_\phi(y|z)] + H(Y)\end{aligned} \tag{8}$$

Here, θ and ϕ represent the VTM network parameters and sequence backend network parameters, respectively.

Similarly, for $I(X,Z)$, we have:

$$I(X,Z) = \int dxdzp(x,z)\log\frac{p(z|x)}{p(z)} \tag{9}$$

In the above, $p(z)$ is intractable; thus, let $r(z)$ be a variational approximation of $p(z)$. Next, based on the fact that the KL divergence $\mathrm{KL}[p(z),r(z)] \geq 0$, we can construct the variational upper bound of $I(X,Z)$ as follows:

$$\begin{aligned}I(X,Z) &\leq \int dxdzp(x)p(z|x)\log\frac{p(z|x)}{r(z)} \\ &= \mathbb{E}_{x\sim p(x)}\mathrm{KL}[p(z|x),r(z)]\end{aligned} \tag{10}$$

Combining Eq. 3, 8, 10, we obtain the following variational upper bound of $\mathcal{L}_{IB}$:

$$\begin{aligned}\mathcal{L}_{IB} \leq &\mathbb{E}_{z\sim p_\theta(z|x)}[-\log q_\phi(y|z)] \\ &+ \beta\mathbb{E}_{x\sim p(x)}\mathrm{KL}[p(z|x),r(z)]\end{aligned} \tag{11}$$

[2] $Z \leftrightarrow X \leftrightarrow Y$. Y and Z are independent given X.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

In Eq. 11, the former term forces z to contain enough information to predict y. In general classification tasks, this term is equivalent to the cross-entropy loss, defined as $\mathcal{L}_{CE}$. And the latter term tries to compress the information of x as much as possible.

2) Variational Temporal Mask Module: As we have explained in Section III-A, the VTM module compresses X by selectively filtering out frame-level features by automatically learning the attribution importance score of frame-level features based on the information bottleneck principle. As shown in Fig. 3, the VTM module takes the concatenation of X and H as input, which is realized based on the readout [53] operation. In the upper forward pass, no mask is added and we collect the different feature maps. As the temporal receptive field of the feature maps differs, the concatenated feature maps contain multi-scale temporal information. As illustrated in Fig. 4, the VTM module takes the concatenated feature maps as input, then predicts the binary mask $M = [m_1, m_2, ..., m_T] \in \{0, 1\}$. Finally, Z is equal to the predicted binary mask applied on X, *i.e.*,

$$Z = X \odot M \tag{12}$$

Here, $\odot$ means element-wise multiplication.

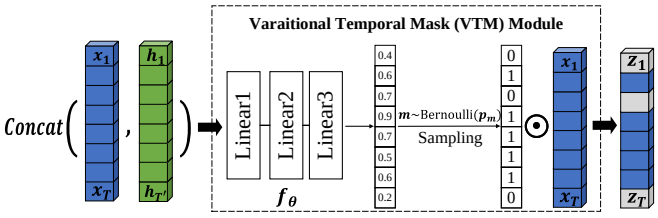


Fig. 4. The architecture of the proposed variational temporal mask module, where f_θ denotes the 3-layer linear network with a sigmoid output layer.

According to Eq. 12, the prior distribution $r(z)$ on variable Z is difficult to predefine directly. Unlike previous information bottleneck methods, which define the prior distribution on variable z, we introduce a prior distribution $r(M)$ on the boolean variable M. To make the second term $\mathrm{KL}[p_\theta(z|x), r(z)]$ in Eq. 11 tractable, we introduce the following two assumptions:

- The binary mask variable m is conditionally independent over individual frames. This assumption aims to guarantee independent analysis of frame-by-frame importance attribution.
- Adapting to the binary properties of variable m, we assume that the prior distribution $r(m)$ follows a Bernoulli distribution, *i.e.*, $r(m) = \mathrm{Bernoulli}(\pi)$, where $\pi \in (0, 1)$ is a constant.

Given the above two assumptions, the posterior distribution over z is a unit impulse mixed distribution:

$$p_\theta(z|x) = (1 - f_\theta(x)) \cdot \delta(z) + f_\theta(x) \cdot \delta(z - x) \tag{13}$$

Here, f_θ denotes the network used in the VTM module, and $\delta(\cdot)$ is a standard unit impulse function. In this case, through combination with Eq. 13, Eq. 9 can be simplified as follows:

$$\begin{aligned} I(X, Z) = \int dx p(x)[p(z = x|x) \log \frac{p(z = x|x)}{p(z = x)} \\ + p(z = 0|x) \log \frac{p(z = 0|x)}{p(z = 0)}] \end{aligned} \tag{14}$$

According to Eq. 12, Eq. 14 can be derived as follows:

$$\begin{aligned} I(X, Z) &= \int dx p(x)[p(m = 1|x) \log \frac{p(m = 1|x)}{p(m = 1)} \\ &\quad + p(m = 0|x) \log \frac{p(m = 0|x)}{p(m = 0)}] \\ &= \int dx p(x) \sum_m p(m|x) \log \frac{p(m|x)}{p(m)} \end{aligned} \tag{15}$$

Here, $p(m)$ is intractable. Based on the fact that the Kullback-Leibler divergence $\mathrm{KL}[p(m), r(m)] \geq 0$, in our case, we can construct the variational upper bound of $I(X, Z)$ as follows:

$$\begin{aligned} I(X, Z) &\leq \int p(x) dx \sum_m p(m|x) \log \frac{p(m|x)}{r(m)} \\ &= \mathbb{E}_{x \sim p(x)} \mathrm{KL}[p(m|x), r(m)] \end{aligned} \tag{16}$$

In the VTM module, $p(m|x)$ is modeled based on the network $f_\theta(x)$. Accordingly, Eq. 16 can be derived as follows:

$$\begin{aligned} \mathrm{KL}[p(m|x), r(m)] =& (1 - f_\theta(x)) \cdot \log \frac{1 - f_\theta(x)}{1 - \pi} \\ &+ f_\theta(x) \cdot \log \frac{f_\theta(x)}{\pi} \end{aligned} \tag{17}$$

Combining Eq. 8 and Eq. 16, the information bottleneck loss in this case can be written as follows:

$$\begin{aligned} \mathcal{L}_{IB} =& \mathbb{E}_{z \sim p_\theta(z|x)}[-\log q_\phi(y|z)] \\ &+ \beta \mathbb{E}_{x \sim p(x)} \mathrm{KL}[p(m|x), r(m)] \end{aligned} \tag{18}$$

Since the sampling from the Bernoulli distribution is not differentiable, we adopt the Gumbel-Softmax [54] reparameterization trick to generate a differentiable approximation m^*. In more detail, m_t has the probability $1 - f_\theta(x)$ and $f_\theta(x)$ to take 0 or 1 respectively. To achieve this, we first draw samples from a standard Gumbel(0, 1) distribution as follows:

$$g = -\log(-\log(u)), u \sim \mathrm{Uniform}(0, 1) \tag{19}$$

Next, m^* is produced via Gumbel-Softmax sampling:

$$m^* = \sigma(\frac{\log f_\theta(x) + g}{\tau}) \tag{20}$$

Here, $\sigma(\cdot)$ is the sigmoid function, while τ is the temperature hyper-parameter; we set $\tau = 1.0$ in this paper.

In summary, the final objective function of our training framework can be expressed as follows:

$$\mathcal{L}_{all} = \mathcal{L}_M + \lambda \cdot \mathcal{L}_{IB} \tag{21}$$

Here, λ is the hyper-parameter controlling the trade-off of the two loss terms.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

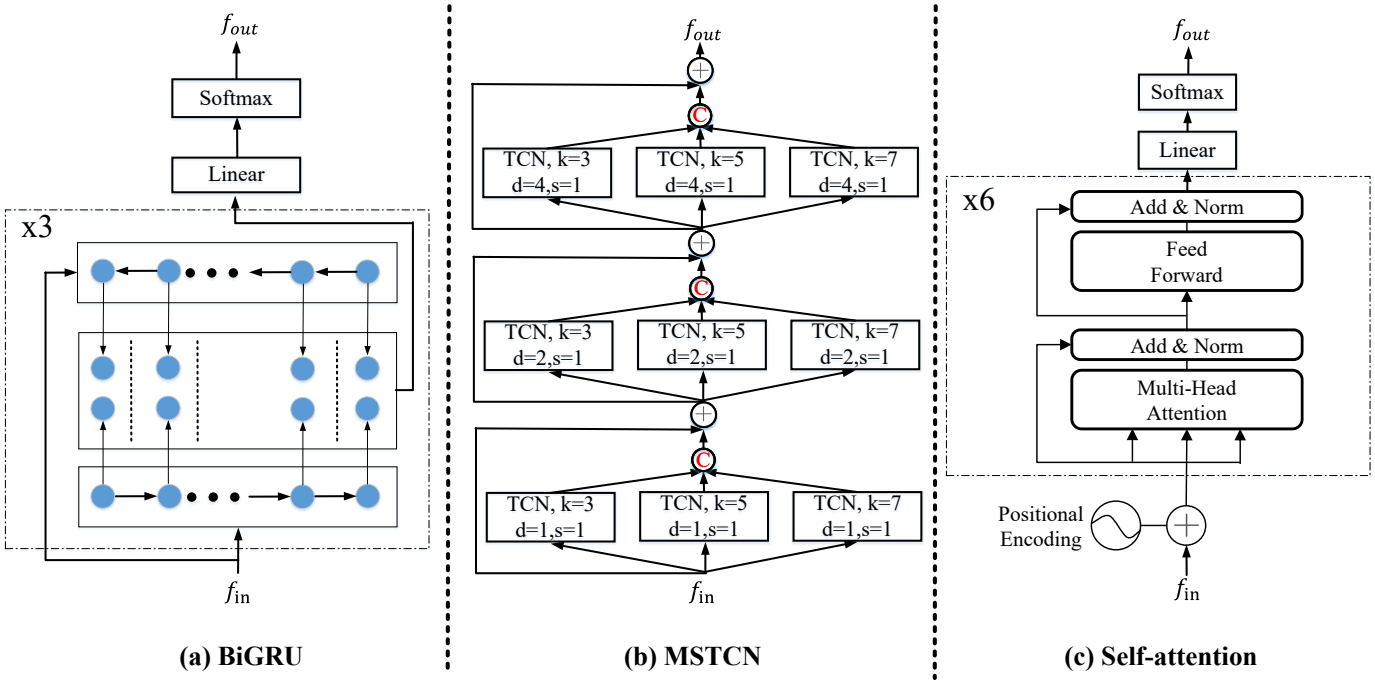


Fig. 5. The three commonly used sequence backend networks. (a) 3-layer BiGRU architecture; (b) Multi-scale TCN architecture; (c) Self-attention architecture.

IV. Experiments and Analysis

In this section, we first introduce all the network architectures used in this work in detail. We then describe the datasets used to evaluate our methods, along with some technical details regarding the training process. Finally, we elaborate the experimental results and detailed analysis based on the above datasets.

A. Network Architectures

Visual Frontend Network E^V. We adopt a simple variation of ResNet18 [55], called SE-C3D-ResNet18 in this paper, as the visual frontend network. This architecture was first used by [14] and achieve SOTA performance on word-level lip reading tasks. Building on the standard C3D-ResNet18 architecture [29], the Squeeze-and-Extract module [56] is introduced to improve the model. Full details of all layers are provided in Tab. I.

Layer Type	Filters	Output dimension
Conv3d MaxPool3d	$5 \times 7 \times 7, 64, /[1,2,2]$ $1 \times 3 \times 3/[1,2,2]$	$T \times 64 \times \frac{H}{4} \times \frac{W}{4}$
SE-Res Conv2d SE-Res Conv2d	$[3 \times 3, 64] \times 2/1$ $[3 \times 3, 64] \times 2/1$	$T \times 64 \times \frac{H}{4} \times \frac{W}{4}$
SE-Res Conv2d SE-Res Conv2d	$[3 \times 3, 128] \times 2/2$ $[3 \times 3, 128] \times 2/1$	$T \times 128 \times \frac{H}{8} \times \frac{W}{8}$
SE-Res Conv2d SE-Res Conv2d	$[3 \times 3, 256] \times 2/2$ $[3 \times 3, 256] \times 2/1$	$T \times 256 \times \frac{H}{16} \times \frac{W}{16}$
SE-Res Conv2d SE-Res Conv2d	$[3 \times 3, 512] \times 2/2$ $[3 \times 3, 512] \times 2/1$	$T \times 512 \times \frac{H}{32} \times \frac{W}{32}$
GlobalPool2d	-	$T \times 512 \times 1 \times 1$

TABLE I
Architecture details for the spatio-temporal visual front-end (SE-C3D-ResNet18).

Sequence Backend Network E^S. To verify the effectiveness of our learning paradigm, we conduct experiments on three commonly used sequence backend network architectures: RNN based network, TCN based network and self-attention based network. As shown in Fig. 5, for the RNN based network, a 3-layer BiGRU is adopted [14] that performs the best on the LRW-1000 dataset. As the SOTA model on the LRW dataset, the Multi-scale Temporal Convolution Network (MSTCN) [37] is also used to verify the effectiveness of our method. For the self-attention based network, the encoder sub-network in standard Transformer [36] is utilized.

Variational Temporal Mask f_θ. Due to the conditionally independent assumption over individual frames, f_θ does not need to capture temporal context information. Motivated by this, we adopt a simple 3-layer linear network (Input & Output feature dimension: $512 \rightarrow 128 \rightarrow 32 \rightarrow 2$) with ReLU activation to learn the importance attribution score of each frame.

B. Datasets and Technical Details

Large scale datasets play a key role in lip reading research. There are two commonly used large-scale word-level lip reading datasets, *i.e.*, LRW [12] and LRW-1000 [13]. The statistics of the two datasets are listed in Tab. II.

LRW. The LRW dataset is commonly used for the word-level lip reading classification task. It consists of up to 1000 utterances of 500 different English words, spoken by hundreds of different speakers. The length of each video is 1.16 seconds (29 frames), and the target word is uttered in the middle of the video.

LRW-1000. LRW-1000 is a naturally distributed large-scale benchmark for Chinese word-level lip reading. It contains

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

Dataset	Subset	Word instances	Hours	Vocabularies
LRW	Train	∼489k	∼157h	500
	Val	25k	∼8h	500
	Test	25k	∼8h	500
LRW-1000	Train	∼574k	∼45h	1000
	Val	∼72k	∼6h	1000
	Test	∼72k	∼6h	1000

TABLE II
THE STATISTICS OF THE LRW AND LRW-1000.

1000 classes with 718018 samples from more than 2000 individual speakers. To ensure that all videos have the same number of frames, a practical setting on LRW-1000 is that continuous 40 frames are selected for each word, and the target word is put at the center to make it similar to LRW data.

We only utilize the word labels provided by these datasets with no extra annotations (*e.g.*, word boundary annotations).

Technical Details. For all the dataset samples, lip-centered videos of size 96×96 pixels are cropped based on the detected facial landmark points. All video inputs are converted to grayscale, and all frames are normalized to [0, 1].

In all experiments, we use the same data augmentation techniques as proposed in [2] for visual input (such as horizontal flipping and random shifts). In the training phase, we randomly crop each mouth-centered video to 88×88 as the input of the network. The Adam optimizer is employed as the default optimizer. The initial learning rate is 0.0003 and the weight decay is 10^{-4}. We set the total number of epochs to 40 and reduce the learning rate to 10^{-6} based on the standard cosine scheduler. Moreover, the dropout rate is set to 0.2 for all baseline models.

Hyperparameters settings. There are three hyperparameters in our importance-aware information bottleneck learning paradigm: the Bernoulli distribution prior parameter π, the IB trade-off parameter β, and the loss term trade-off parameter λ. We use the grid search algorithm to find the optimal values of hyperparameters, and ultimately we set $\pi = 0.5$, $\beta = 0.1$ and $\lambda = 1.0$.

C. Generalization Evaluation & Analysis

Backend Network	Training Paradigm	LRW		LRW-1000	
		val set	test set	val set	test set
GRU	GRU-base	86.98	86.29	49.49	48.48
	GRU-Attn	86.31	85.92	48.94	48.03
	GRU-VTM	**87.25**	**86.92**	**50.96**	**49.74**
MSTCN	MSTCN-base	84.35	84.27	44.51	43.44
	MSTCN-Attn	84.70	84.45	45.24	44.13
	MSTCN-VTM	**85.57**	**85.31**	**47.28**	**45.91**
Transformer	Transformer-base	83.21	82.98	42.68	42.01
	Transformer-Attn	82.68	82.15	41.94	41.16
	Transformer-VTM	**85.68**	**85.18**	**43.43**	**42.93**

TABLE III
PREDICTION ACCURACY (%) OF DIFFERENT MODELS WITH DIFFERENT TRAINING FRAMEWORK ON THE LRW AND LRW-1000 DATASETS. ALL MODELS ADOPT THE SE-C3D-RESNET18 ARCHITECTURE AS THE VISUAL FRONTEND NETWORK.

Ablation study on the VTM module. We verify the proposed learning paradigm on the LRW and LRW-1000 datasets with three different types of sequence backend networks,namely 3-layer GRU, MSTCN, and 6-layer transformer encoder. Tab. III lists the prediction accuracy of different models on the validation and test sets. Here, X-base represents the baseline model with general training framework, while X-VTM means our IB training framework. Moreover, we also provide the results of baseline models equipped with the temporal attention network (X-Attn), since the motivation of the proposed VTM is somewhat similar to that of the temporal attention mechanism.

As the table shows, the proposed IB training framework outperforms the baseline framework on all sequence backend networks, especially for the transformer backend network (up to 2.47% performance improvement). Previous works have demonstrated that transformer architecture suffers from serious overfitting when the training data is insufficient. In the training stage, we observe that the prediction performance of our training framework is similar to that of the baseline framework in all of these experiments. Therefore, the performance improvements on the validation and test datasets clearly demonstrate the effectiveness of the proposed VTM module. In other words, our training framework can help to improve the generalization ability of deep lip reading models.

As for the temporal attention network, baseline models equipped with X-Attn do not significantly improve the performance, especially for the GRU and Transformer backend networks. The main problem with using the temporal attention mechanism is that it is not especially effective for solving overfitting. In contrast, the proposed VTM provides a trade-off between recognition and anti-noise, which can aid in improving generalization.

Comparison with other regularization methods. Considering that there is no extra change applied to the architecture of the visual frontend network and sequence backend network, the proposed VTM module can also be treated as a regularization method that is applied to baseline models from the perspective of network optimization. Thus, we further provide quantitative assessment with other regularization methods, such as dropout [57], weight decay [58], mix-up [59] and label smoothing [60]. We conduct comparative experiments on the LRW dataset with the 3-layer GRU backend network. It is worth noting that all the above-mentioned regularization methods, except for the VTM, have already been adopted in existing SOTA methods. The results are listed in Tab. IV. As we can see, the label smoothing technology contributes most to the improvements in prediction accuracy (83.12% $\rightarrow$ 84.18% in LRW). The main reason for this is that some word categories in the LRW dataset, like *allow & allowed* and *again & against* have high similarities; moreover, the label smoothing technology can alleviate overfitting caused by similar labels to a certain extent. As an effective data augmentation method, the mix-up technology also significantly improves the performance (83.12% $\rightarrow$ 84.07% in LRW), as the LRW and LRW-1000 datasets are insufficiently large-scale. In addition, combining all these regularization methods with VTM further improves the performance, which illustrates that the proposed VTM can be easily integrated with other regularization methods in order to further improve the generalization ability.

Another interesting finding concerns the relationship

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

Regularization Method	LRW	LRW-1000
Baseline	83.12	45.53
+Dropout (0.2)	83.73	46.56
+Weight Decay (10^{-4})	83.32	45.88
+Mix-up	84.07	47.38
+Label Smoothing (0.1)	84.18	47.02
+VTM	84.04	46.95
+All above	**86.92**	**49.74**

TABLE IV
PREDICTION ACCURACY (%) OF DIFFERENT MODELS WITH DIFFERENT REGULARIZATION METHODS.

Method	LRW	LRW-1000
Chung *et al.* [12]	61.1	/
Stafylakis *et al.* [29]	83.0	38.2
Wang *et al.* [61]	83.3	36.9
Weng *et al.* [32]	84.1	/
Zhao *et al.* [62]	84.4	38.8
Xiao *et al.* [63]	84.1	41.9
Martinez *et al.* [37]	85.3	41.4
Sheng *et al.* [33]	85.5	/
Ma *et al.* [64]	**87.9**	45.3
Feng *et al.* [14]	86.2	48.3
Ours	86.9	**49.7**

TABLE V
PREDICTION ACCURACY (%) COMPARISON WITH THE STATE-OF-THE-ART. OUR WORK HAS THE SAME NETWORK ARCHITECTURE AS THAT USED IN [*Feng et al., 2020*] WITH DIFFERENT TRAINING FRAMEWORK. NOTE THAT ALL RESULTS LISTED ABOVE DO NOT USE EXTRA ANNOTATIONS (*e.g.*, WORD BOUNDARY OR AUDIO INFORMATION).

between the VTM module and the dropout technology. More specifically, as shown in Tab. III, the performance improvement caused by the VTM module is 1.26% (48.48% $\rightarrow$ 49.74%) with the 3-layer GRU backend network on LRW-1000 test set. In Tab. IV, without other regularization methods, the performance improvement caused by the VTM module is 1.42% (45.53% $\rightarrow$ 46.95%). The small performance difference (1.26% *v.s.* 1.42%) is due to some similar effects of dropout and VTM, as VTM can be treated as a particular form of dropout that acts on the entire feature vector.

Comparison with SOTA methods. The results of the comparison between our proposed approach and the baseline methods are provided in Tab. V. As is evident, the proposed method outperforms the baseline model [14] and achieves SOTA performance on LRW-1000 dataset. Ma *et al.* [64] achieved the SOTA performance on LRW. However, this SOTA result is realized based on multi-stage knowledge distillation; considering that our approach does not increase any computational and memory load in the inference stage, the proposed framework can be concluded to be effective and efficient.

Besides, we also investigate the impact of different choices of the hyperparameters β and π on the recognition performance of GRU-based sequence networks. The results are provided in Tab. VI. As a core hyperparameter that controls the information compression and class prediction of the training set, β significantly influences the model generalization. The results show that the generalization on the test set is best when $\beta = 0.1$. The other hyperparameter, π, is an estimate of the proportion of recognition-related frames in all video frames, which is highly related to the property of the dataset. As we can see from Tab. VI, $\pi = 0.5$ is more suitable for the LRW dataset, while $\pi = 0.3$ is best for the LRW-1000 dataset.

TABLE VI
GENERALIZATION ANALYSIS ON HYPERPARAMETERS β & π

β	π	LRW	LRW-1000
$\beta = 0.01$	$\pi = 0.3$	86.12	48.83
	$\pi = 0.5$	86.40	48.54
	$\pi = 0.7$	86.17	48.26
$\beta = 0.1$	$\pi = 0.3$	86.73	**49.82**
	$\pi = 0.5$	**86.92**	49.74
	$\pi = 0.7$	86.42	49.38
$\beta = 1.0$	$\pi = 0.3$	85.33	48.24
	$\pi = 0.5$	85.51	48.17
	$\pi = 0.7$	85.56	47.52

D. Interpretability Evaluation & Analysis

The results in Section IV-C have demonstrated the generalization improvement achieved by our method. In this section, we intend to quantitatively prove the interpretability improvement of our method. More specifically, we adopt one of the most popular interpretability methods LIME [65], (*i.e.*, Local Interpretable Model-agnostic Explanations), to generate frame-wise explanations for baseline and the proposed framework. LIME explanations evaluate the interpretable coefficients [66] for each frame in a video. The interpretability coefficient metric the magnitude of the contribution made by each frame to the current prediction.

As we are the first to analyze the interpretability of deep lip reading models, there are no reference interpretability evaluation metrics available that can be directly utilized. It is therefore essential to design a proper benchmark for interpretability evaluation. Motivated by this, we propose two interpretability evaluation metrics for deep lip reading models.

Average interpretable coefficients evaluation. The first evaluation metric is inspired by intuitive knowledge obtained from a cognition perspective. As for the word-level lip reading task, it is natural to conclude that a model is interpretable when it makes decisions that depends more on the word itself than other low-relevance information. Therefore, we introduce extra word boundaries as auxiliary information to evaluate interpretability. More specifically, we argue that a lip reading model is more interpretable than other models when its predictions are more dependent on frames within the word boundary. To illustrate the difference in interpretability between our framework and the baseline framework, we compare the Average Interpretable Coefficients within the Word Boundary (abbreviated as AIC-WB). We conduct the experiments on the test set of the LRW-1000 and the LRW dataset, and list the results in Tab. VII. As we can see, our training framework is more interpretable than the baseline framework on all sequence backend networks.

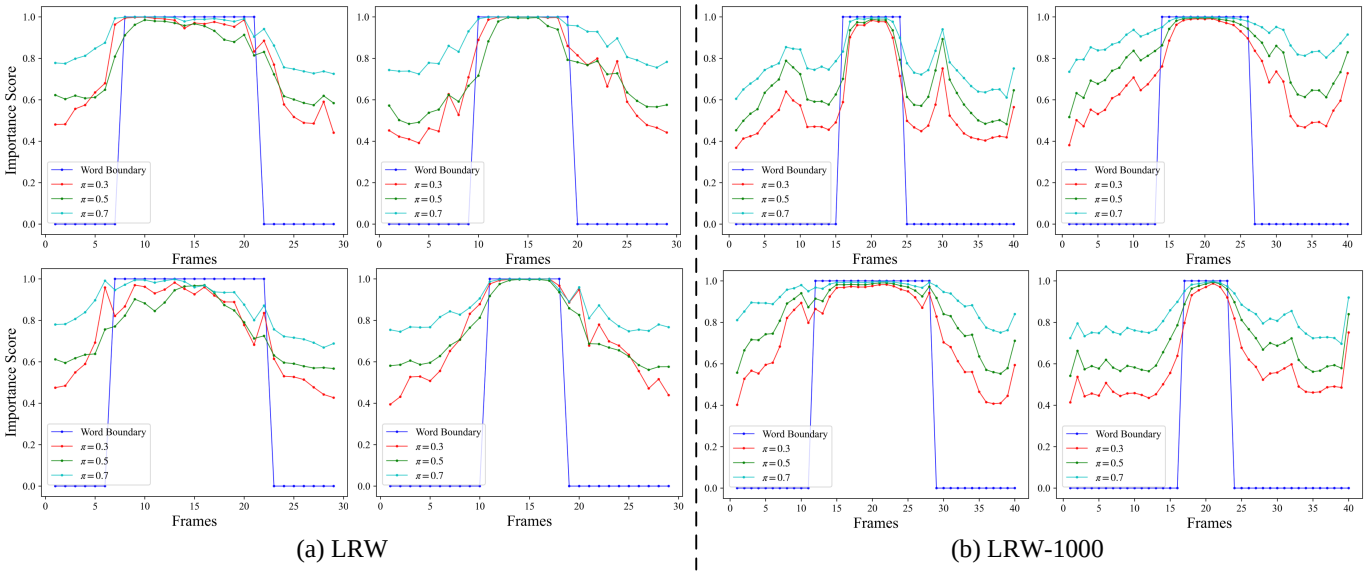
An interesting finding is that the Transformer based sequence backend network is more interpretable than the GRU and MSTCN backend networks. In contrast, the generalization of the Transformer based sequence backend network is worse than that of the GRU and MSTCN backend network in Tab. III. The reason for this phenomenon is as follows. The LIME explanation is realized based on feature

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

Backend Network	Training Framework	LRW		LRW-1000	
		AIC-WB	AOPC score	AIC-WB	AOPC score
GRU	GRU-base	0.0412	18.6%	0.0258	22.6%
	GRU-VTM	**0.0487**	**20.7%**	**0.0309**	**25.5%**
MSTCN	MSTCN-base	0.0558	20.2%	0.0248	22.7%
	MSTCN-VTM	**0.0691**	**21.9%**	**0.0281**	**23.9%**
Transformer	Transformer-base	0.0693	21.8%	0.0318	26.4%
	Transformer-VTM	**0.0718**	**22.4%**	**0.0327**	**28.8%**

TABLE VII
AVERAGE INTERPRETABLE COEFFICIENTS (INTERPRETABILITY ANALYSIS OVER GROUND-TRUTH LABELS) WITHIN WORD BOUNDARY (AIC-WB) AND AOPC SCORE ON LRW-1000 TEST SET. FOR BOTH AIC-WB AND AOPC SCORE, HIGHER IS BETTER.



(a) LRW (b) LRW-1000

Fig. 6. Visualization Effect of varying π on the VTM module ($\beta = 0.1$). The four graphs in the left half are sampled from the LRW dataset, and graphs in the right half are sampled from the LRW-1000 dataset. The blue lines indicate the actual word boundary, while the red, green and cyan lines represent the importance score of each frame under different π (0.3, 0.5 and 0.7), respectively.

importance. Meanwhile, due to their inner self-attention structures, transformers will naturally learn information about the feature importance. Thus, within the LIME framework, transformers are more interpretable than RNNs and TCNs.

AOPC evaluation. The area over the perturbation curve (AOPC) [67], [68] metric is utilized to evaluate the local fidelity of post-hoc explanations to models. By deleting the top N frames with the highest relevance (N is set to 5 in our case), it calculates the average accuracy change of prediction probability on the predicted class over all test data. AOPC can be formalized as follows:

$$\text{AOPC} = \frac{1}{N+1}\sum_{k=1}^{N}[f_\phi(\mathbf{x}) - f_\phi(\mathbf{x}_{/1\ldots k})]_{p(x)} \tag{22}$$

Here, ϕ represents the sequence backend network parameters, while $f_\phi(\mathbf{x}_{/1\ldots k})$ is the probability for the predicted class when the top-k most relevant frames are removed. $[\cdot]_{p(x)}$ denotes the expectation over the test set. Due to computation costs, we randomly select 3,000 examples for evaluation. Tab. VII shows the AOPC scores of different models on the LRW-1000 and the LRW datasets by deleting top-5 most relevant frames calculated by the LIME algorithm. The AOPC scores of the X-VTM models are higher than these obtained by X-base models, indicating that the VTM module can significantly improve the model's interpretability to post-hoc explanations.

Visualization Analysis. The AOPC evaluation metric is not sufficiently intuitive. To prove the interpretability of VTM in an easy-to-understand manner, we further conduct visualization analysis of the VTM module for qualitative interpretability evaluation. In Fig. 6, we visualize the effect of different values of π on LRW and LRW-1000 respectively. As is evident, the average importance scores of frames within the word boundary are higher than those of frames outside of the word boundary. These visualization samples provide qualitative evidence for our motivation.

V. CONCLUSION

In this paper, we propose a novel learning paradigm designed to improve the interpretability and generalization of lip reading models for word-level lip reading. In our training framework, an information bottleneck-based variational temporal mask that learns the importance of frame-level features is propsoed. Furthermore, to facilitate the generalization performance of baseline models, we further regularize the model by reducing the prediction differences between global information and important local features. Quantitative and qualitative

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

results on public lip reading datasets prove that the model trained with the proposed framework significantly outperforms the general learning paradigm for multiple different types of sequence backend networks. Meanwhile, analysis of the interpretability results also proves that our method is more trustworthy than the general paradigm. In addition, the effectiveness of the learning paradigm also opens up the possibility of breakthroughs on more challenging tasks, *i.e.*, sentence-level lip reading. Therefore, we will explore this matter further in the future.

ACKNOWLEDGMENT

This work was partially supported by the National Key R&D Program of China No.2021YFB3100800, the Academy of Finland under grant 331883 and the National Natural Science Foundation of China under Grant 61872379.

REFERENCES

[1] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018.

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

[5] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[6] N. Tye-Murray, M. S. Sommers, and B. Spehar, "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing," *Ear and hearing*, 2007.

[7] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5039–5049.

[8] Z. Akhtar, C. Micheloni, and G. L. Foresti, "Biometric liveness detection: Challenges and research opportunities," *IEEE Security & Privacy*, vol. 13, no. 5, pp. 63–72, 2015.

[9] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "Human machine interaction via visual speech spotting," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2015, pp. 566–574.

[10] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "Lip-interact: Improving mobile device interaction with silent speech commands," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 581–593.

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

[12] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian conference on computer vision*. Springer, 2016, pp. 87–103.

[13] 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 *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–8.

[14] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an effective lip reading model without pains," *arXiv preprint arXiv:2011.07557*, 2020.

[15] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *2015 ieee information theory workshop (itw)*. IEEE, 2015, pp. 1–5.

[16] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv:physics/0004057*, 2000.

[17] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *ICLR*, 2017.

[18] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," in *ICLR*, 2020.

[19] S. J. Cox, R. W. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, "The challenge of multispeaker lip-reading." in *AVSP*. Citeseer, 2008, pp. 179–184.

[20] P. Lucey, G. Potamianos, and S. Sridharan, "A unified approach to multi-pose audio-visual asr," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[21] I. Almajai, S. Cox, R. Harvey, and Y. Lan, "Improved speaker independent lip reading using speaker adaptive training and deep neural networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2722–2726.

[22] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[23] Y. Fu, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, "Lipreading by locality discriminant graph," in *2007 IEEE International Conference on Image Processing*, vol. 3. IEEE, 2007, pp. III–325.

[24] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 129–136.

[25] H. E. Cetingul, Y. Yemez, E. Erzin, and A. M. Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2879–2891, 2006.

[26] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, "Visual speech recognition with loosely synchronized feature streams," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 2. IEEE, 2005, pp. 1424–1431.

[27] J. Luettin and N. A. Thacker, "Speechreading using probabilistic models," *Computer vision and image understanding*, vol. 65, no. 2, pp. 163–178, 1997.

[28] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[29] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

[30] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6548–6552.

[31] X. Zhang, F. Cheng, and S. Wang, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 713–722.

[32] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3d cnns for lipreading," *arXiv preprint arXiv:1905.02540*, 2019.

[33] C. Sheng, X. Zhu, H. Xu, M. Pietikainen, and L. Liu, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," *IEEE Transactions on Multimedia*, 2021.

[34] H. Liu, Z. Chen, and B. Yang, "Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion." in *INTERSPEECH*, 2020, pp. 3520–3524.

[35] T. Afouras, A. Zisserman *et al.*, "Sub-word level lip reading with visual attention," *arXiv preprint arXiv:2110.07603*, 2021.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[38] D. Strouse and D. J. Schwab, "The deterministic information bottleneck," *Neural computation*, vol. 29, no. 6, pp. 1611–1630, 2017.

[39] A. Kolchinsky, B. D. Tracey, and S. Van Kuyk, "Caveats for information bottleneck in deterministic scenarios," *arXiv preprint arXiv:1808.07593*, 2018.

[40] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, "Nonlinear information bottleneck," *Entropy*, vol. 21, no. 12, p. 1181, 2019.

[41] T. Wu, H. Ren, P. Li, and J. Leskovec, "Graph information bottleneck," *Advances in Neural Information Processing Systems*, vol. 33, pp. 20 437–20 448, 2020.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

[42] J. Yu, T. Xu, Y. Rong, Y. Bian, J. Huang, and R. He, "Recognizing predictive substructures with subgraph information bottleneck," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[43] Z. Pan, L. Niu, J. Zhang, and L. Zhang, "Disentangled information bottleneck," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 9285–9293.
[44] M. Yamada, H. Kim, K. Miyoshi, T. Iwata, and H. Yamakawa, "Disentangled representations for sequence data using information bottleneck principle," in *Asian Conference on Machine Learning*. PMLR, 2020, pp. 305–320.
[45] J. Chen, Z. Zhang, X. Xie, Y. Li, T. Xu, K. Ma, and Y. Zheng, "Beyond mutual information: Generative adversarial network for domain adaptation using information bottleneck constraint," *IEEE Transactions on Medical Imaging*, 2021.
[46] I. Jeon, W. Lee, M. Pyeon, and G. Kim, "Ib-gan: Disengangled representation learning with information bottleneck generative adversarial networks," in *35th AAAI Conference on Artificial Intelligence/33rd Conference on Innovative Applications of Artificial Intelligence/11th Symposium on Educational Advances in Artificial Intelligence*. ASSOC ADVANCEMENT ARTIFICIAL INTELLIGENCE, 2021, pp. 7926–7934.
[47] B. Dai, C. Zhu, B. Guo, and D. Wipf, "Compressing neural networks using the variational information bottleneck," in *International Conference on Machine Learning*. PMLR, 2018, pp. 1135–1144.
[48] S. S. Lorenzen, C. Igel, and M. Nielsen, "Information bottleneck: Exact analysis of (quantized) neural networks," *arXiv preprint arXiv:2106.12912*, 2021.
[49] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6778–6787.
[50] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
[51] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *arXiv preprint arXiv:1703.00810*, 2017.
[52] I. Fischer, "The conditional entropy bottleneck," *Entropy*, vol. 22, no. 9, p. 999, 2020.
[53] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2014.
[54] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," *arXiv preprint arXiv:1611.01144*, 2016.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[56] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[57] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
[58] A. Krogh and J. Hertz, "A simple weight decay can improve generalization," *Advances in neural information processing systems*, vol. 4, 1991.
[59] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
[60] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[61] C. Wang, "Multi-grained spatio-temporal modeling for lip-reading," *arXiv preprint arXiv:1908.11618*, 2019.
[62] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual information maximization for effective lip reading," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 420–427.
[63] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation flow based two-stream network for lip reading," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 364–370.
[64] P. Ma, B. Martinez, S. Petridis, and M. Pantic, "Towards practical lipreading with distilled and efficient models," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7608–7612.
[65] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.
[66] D. Mardaoui and D. Garreau, "An analysis of lime for text data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 3493–3501.
[67] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.
[68] D. Nguyen, "Comparing automatic and human evaluation of local explanations for text classification," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1069–1078.

Changchong Sheng received the B.E degree in information engineering and the M.E degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2015 and 2017, respectively. He was serving as a visiting Ph.D at the Machine Vision Group at the University of Oulu, Finland, in 2020. He is currently a Ph.D candidate in NUDT, and his current research interests include computer vision, visual speech analysis.

Li Liu received her Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), China, in 2012. From 2016.12 to 2018.11, she worked as a senior researcher at the Machine Vision Group at the University of Oulu, Finland. Her current research interests include computer vision, pattern recognition and machine learning. She serves as Area Chair of ICME 2020, ICME 2021 and ACCV 2020.

Wanxia Deng received the university B.E. degree in electronic information science and technology from Xiamen University, Xiamen, China, and the M.S. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2014 and 2016, respectively. She is currently pursuing the Ph.D. degree in information and communication engineering from NUDT, Changsha, China. Her research interests include domain adaptation, transfer learning.

Liang Bai received the B.E. and B.M. degrees from Xi'an Jiaotong University, in 2002, and the M.E. and Ph.D. degrees from the National University of Defense Technology, in 2005 and 2008, respectively. He is currently a Professor with College of Systems Engineering, National University of Defense Technology. His research interests include multimedia content analysis and access, particularly for video and images, big multimedia data, complex system modeling, and network analysis.

Zhong Liu received the Ph.D. degree in management science from the National University of Defense Technology (NUDT), Changsha, China, in 2000, where he is currently a Professor. He is also the Vice-Dean of the College of Systems Engineering Laboratory, NUDT, where he is also a Senior Advisor of the Research Center for Computational Experiments and Parallel Systems. His main research interests include planning systems, computational organization, and intelligent systems.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3210761

Songyang Lao received the B.S. degree in information system engineering and the Ph.D. degree in system engineering from the National University of Defense Technology, Changsha, China, in 1990 and 1996, respectively. He joined the faculty at the National University of Defense Technology in 1996, where he is currently a Full Professor in the College of Systems Engineering. He was a Visiting Scholar with the Dublin City University, Irish, from 2004 to 2005. His current research interests include image and video analysis and human-computer interaction.

Gangyao Kuang received the B.S. and M.S. degrees in geophysics from the Central South University of Technology, Changsha, China, in 1988 and 1991, respectively, and the Ph.D. degree in communication and information from the National University of Defense Technology, Changsha, in 1995. He is currently a Professor with the College of Electronic Science and Technology, National University of Defense Technology. His current research interests include remote sensing and SAR image processing.

Matti Pietikäinen received the doctor of science degree in technology from the University of Oulu, Finland. He is an emeritus professor with the Center for Machine Vision and Signal Analysis, University of Oulu. He was president of the Pattern Recognition Society of Finland from 1989 to 1992, and was named its honorary member in 2014. He is a fellow of the IEEE for contributions to texture and facial image analysis for machine vision.

© 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Oulu University. Downloaded on October 03,2022 at 06:41:15 UTC from IEEE Xplore. Restrictions apply.