Multi-view Convolution for Lipreading

Tsubasa Maeda* and Satoshi Tamura†
Gifu University, Gifu, Japan
* E-mail: maeda@asr.info.gifu-u.ac.jp
† E-mail: tamura@info.gifu-u.ac.jp

***Abstract*—This paper presents a lipreading scheme using a dynamic convolution mechanism. Lipreading is a technique to convert consecutive lip images into texts, and has been investigated particularly for this decade owing to deep learning. Most of visual speech recognition systems employ Convolutional Neural Networks (CNNs). Assuming real applications, lipreading needs to adapt any view from frontal to profile, however, existing CNN models do not take it into account. In our scheme, we choose a CNN model having a dynamic convolution mechanism, where each kernel would correspond to a particular view, which we call "Multi-view convolution". To induce view-dependent kernels, we impose regularization to dynamic convolution via view metadata. We further adapt soft attention in our scheme. We carried out evaluation experiments using a multi-view corpus OuluVS2. Only frontal and profile lip images were adopted to train our model, while recognition was conducted for every views. It is found that using a dynamic convolution with label smoothing and cross entropy worked well, and finally our model achieved better accuracy than conventional basic schemes.**

I. Introduction

Lipreading, also known as Visual Speech Recognition (VSR), is the process of recognizing speech content by observing only lip movements without having access to audio signal. Recently, Automatic Speech Recognition (ASR) has become widespread, due to the rapid progress of deep learning. However, there is still a major challenge; the accuracy of ASR decreases in noisy environments. Since visual features are unaffected by acoustic noise, lipreading has a potential to improve ASR in noisy environments. A lipreading system is also expected to be used as an application on some devices, such as a communication aid for people with speech disabilities.

In realistic scenarios, a subject does not always face to the front. For example, assume that in a web meeting a camera is placed in front of a subject; the speaker may turn right or left to look at another person, resulting a diagonal or profile face is observed. Therefore, we should consider a pose-invariant lipreading technique. Because most works have focused only on frontal faces, there has been an increasing interest in lipreading for non-frontal poses.

There are several public databases available, offering multiple synchronized views of speakers' faces; particularly, [1], [2] have been widely used for multi-view lipreading. Among them, the OuluVS2 database [2] has attracted significant interest to many research works e.g. [3], [4], [5], [6], [7], [8], [9], [10], in which five synchronous movies are available recorded at fixed angles, namely, a frontal facial view at $0°$, side views at $30°$, $45°$, and $60°$, and a profile one at $90°$, respectively. However, due to its small data size and vocabulary, it is difficult to realize a non-frontal and sentence-level lipreading system that is required in more realistic scenarios. This may because higher cost of producing a multi-view synchronized data set.

This paper proposes a new operator design for multi-view lipreading. Our motivation is twofold. First, to achieve multi-view lipreading using only a few kinds of views for training because a multi-view synchronized data set is expensive producing. Second, to design a mechanism that can be applied to existing models. At present, works on pose-invariant VSR and lipreading for large vocabularies are independent, and it is expected that the two will be integrated in the future. For this motivation, we propose "Multi-view convolution" based on dynamic convolution [11], which uses a set of multiple convolutional kernels instead of using a single convolution kernel per layer. Convolution kernels are aggregated dynamically for each individual input feature via an attention mechanism. Although dynamic convolution was proposed to increase the representation capability without noticeably increasing the computational cost, we focus on the mechanism of aggregating multiple convolution kernels. The role of each kernel in dynamic convolution is unknown. We impose regularization giving a black-boxed kernel an explicit role, e.g. for a particular view, and then aggregate them to create kernels for any unknown view. This allows us to interpolate kernels for any view that is not used for training. Since our method only replaces a traditional convolution operation in existing models, our scheme can be easily applied to any multi-view lipreading.

We conducted evaluation experiments using OuluVS2. Experimental results show that imposing view regularization to dynamic convolution with soft attention (softmax with temperature and label smoothing) enables us to accomplish robust lipreading for unknown views.

The rest of this paper is organized as follows. Section II briefly introduces related works to our approach. Our former and proposing schemes are explained in Section III. In Section IV, we describe experimental setup and results, as well as discussion. Section V finally concludes this article.

II. Related Work

In realistic scenarios, there are two problems against realizing lipreading applications: pose-invariant and open-world lipreading. This section mentions related research works regarding the problems.

A. Pose-invariant lipreading

Lipreading systems should recognize different poses which are not used for model training without accuracy decrease. There are two main approaches for such pose-invariant lipreading according to [9]. The first approach is training a recognition network model using data from all available views in order to build generalized network [10]. The second approach applies a pose-mapping from non-frontal views to a particular view, usually the frontal view [3], [12]. For example, the view2view [12] transforms a non-frontal face to a frontal one based on pix2pix [13], which is one of the generative adversarial networks.

B. Task-invariant lipreading

There are many research works investigating lipreading, however, due to the limitation of existing lipreading and audio-visual speech recognition corpora, only a few works focused on real-environment or large-vocabulary tasks [14], [15], [16]. Among such the works, for example, Lipnet [14] employed CNN, Recurrent Neural Network (RNN), and Connectionist Temporal Classification (CTC) loss. This indicates CNN is an essential technique for task-invariant lipreading.

III. Methods

A. Dynamic Convolution

First of all, we explain dynamic convolutional neural networks [11] used in this work. Fig 1(a) shows an overview. In order to increase the representation capability in deep learning, it is well known to make a network wider or deeper. However, it costs computationally a lot, and thus are not suitable for efficient networks. The dynamic convolution [11] aims at increasing the representation capability with slight additional computational cost.

Traditional convolution can be done as:

$$\boldsymbol{y} = \boldsymbol{W} * \boldsymbol{x} \tag{1}$$

where $*$ indicates a convolution operator and $\boldsymbol{W}$ is a single convolution kernel. In contrast, dynamic convolution employs a dynamic kernel $\widetilde{\boldsymbol{W}}(\boldsymbol{x})$ that aggregates a set of K convolution kernels $\{\boldsymbol{W}_k\}$ for each input $\boldsymbol{x}$:

$$\boldsymbol{y} = \widetilde{\boldsymbol{W}}(\boldsymbol{x}) * \boldsymbol{x}, \tag{2}$$

$$\widetilde{\boldsymbol{W}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}) \boldsymbol{W}_k$$

$$\text{s.t. } 0 \leq \pi_k(\boldsymbol{x}) \leq 1, \ \sum_{k=1}^{K} \pi_k(\boldsymbol{x}) = 1 \tag{3}$$

where π_k is an attention weight for a k^{th} kernel $\boldsymbol{W}_k$. Dynamic convolution applies Squeeze-and-Excitation (SE) module [17] to the attention weights π_k. The SE module employs self excitation to adaptively recalibrate channel-wise feature responses. The global spatial information is firstly squeezed by global average pooling. Then we use two Fully-Connected (FC) layers and the softmax function to generate attention weights for K convolution kernels.

B. Multi-view Convolution

In this paper, we propose "Multi-view convolution" based on dynamic convolution for lipreading. Fig 1(b) shows an overview. Although dynamic convolution sums up multiple kernels with attention weights depending on input data, model developers cannot control or determine the role of each kernel. We focus on improving this mechanism and impose regularization on SE modules in dynamic convolution. Imposing regularization is expected to give each kernel an explicit role corresponding to each view. We can then create kernels for two views, e.g. frontal and profile facial ones, and aggregate them to create kernels for any view between them. This allows us to interpolate kernels for any diagonal view, which is not used for training.

View regularization: When training, the loss function for our model is designated to combine the visual speech recognition loss $\mathcal{L}_{rec}$ and the penalty term $\mathcal{L}_{view}^{(m)}$ which is obtained from view metadata, on an m^{th} convolutional layer:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{view} \sum_{m=1}^{M} \mathcal{L}_{view}^{(m)}, \tag{4}$$

where λ_{view} is a hyperparameter that balances the contribution of the recognition loss and the regularized loss, and M is the number of multi-view convolution layers. $\mathcal{L}_{view}$ is obtained by calculating a cross-entropy between attention outputs to each kernel and the metadata.

Soft attention: To interpolate any view from the views used for training, multi-view convolution applies soft attention based on label smoothing (LS) [18], and softmax with temperature. Since a softmax mechanism often generates a vector close to a one-hot output, it is almost equivalent to selecting the best kernel among kernels. Our goal is to combine multiple kernels effectively and obtain a new kernel from them, thus we need a flatter attention.

Label Smoothing is a regularization technique that introduces noise for labels. Label smoothing replaces a one-hot encoded label vector $\boldsymbol{y}_{hot}$ into a mixture of $\boldsymbol{y}_{hot}$ and uniform distribution:

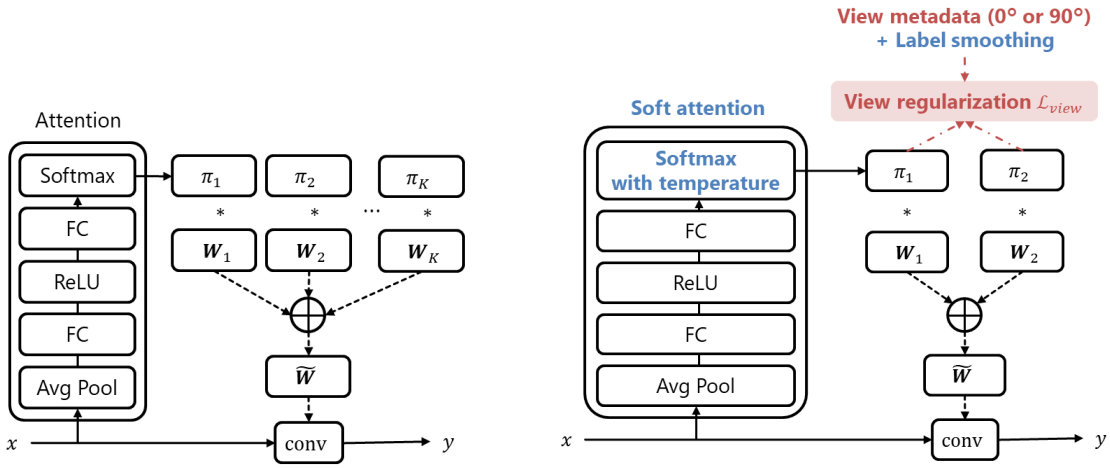
$$\boldsymbol{y}_{ls} = (1 - \alpha)\boldsymbol{y}_{hot} + \frac{\alpha}{L}, \tag{5}$$

where L is the number of label classes, and α is a hyperparameter that determines the amount of smoothing. If $\alpha = 0$, we obtain the original one-hot encoded $\boldsymbol{y}_{hot}$. If $\alpha = 1$, we get the uniform distribution.

Compared to the original softmax, softmax with temperature outputs a flattened attention as follows:

$$\pi_k = \frac{\exp(z_k/\tau)}{\sum_j \exp(z_j/\tau)}, \tag{6}$$

where z_k is an output of the second FC layer in attention, and τ is the temperature. If $\tau = 1$, the output equals to the original softmax. As τ increases, the output becomes increasingly flat. In order to improve training efficiency, dynamic convolution starts at a large temperature (e.g. $\tau = 30$) and reducing τ toward $\tau = 1$ in the first 10 epochs. We set the minimum value



(a) Dynamic convolution (b) Multi-view convolution

Fig. 1. The overview of dynamic and multi-view convolution.

TABLE I
DATA SETS IN OUR EXPERIMENTS.

Data Set	Speaker IDs	Views	Samples
Training	1,2,3,10,11,12,13,18,19,20, 21,22,23,24,25,27,33,35,36, 37,38,39,45,46,47,48,50,53	0°, 90°	1,680
Validation	4,5,7,14,16,17,28,31,32, 40,41,42	0°, 90°	720
Test	6,8,9,15,26,30,34,43,44, 49,51,52	0°, 30°, 45°, 60°, 90°	1,800

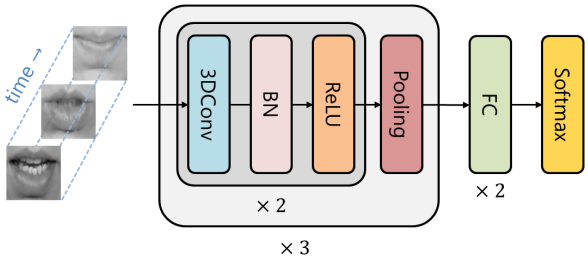


Fig. 2. The baseline architecture for our experiment.
† The 3DConv module was replaced into dynamic 3DConv or multi-view 3DConv for comparison.

of τ as a hyperparameter to get a flatter output compared to the original softmax.

IV. EXPERIMENTS

We conducted experiments to investigate the effectiveness of multi-view convolution in multi-view visual speech recognition.

A. Data

The dataset used in our work is OuluVS2 [2], which is a publicly available multi-view database with five lip views including 0°, 30°, 45°, 60°, and 90°. OuluVS2 contains video recordings from 52 speakers with five different camera views. Each subject uttered three collections of 10 continuous digit strings, 10 daily-use short English phrases, and 5 randomly selected TIMIT sentences.

OuluVS2 provides Region-Of-Interest (ROI) videos, which were preprocessed by segmenting individual utterances and cropping off ROIs, for digit strings and phrases collection. Every phrase was uttered three times in this collection, thus the total number of samples is 52 (speakers) × 5 (views) × 3 (utterances) × 10 (phrases) = 7,800. The data were divided into train, validation, and test sets with reference to Lee et al. [4] as shown in Table I. In this work, only frontal and profile lip images were adopted to train our model, while recognition was conducted for every views.

Each input video was gray-scaled and normalized, then we aligned the number of frames in each utterance to 64. We performed data augmentation with random cropping of which size was 64×64.

B. Models and experimental setup

The baseline architecture in this experiment was illustrated in Fig. 2. The models was a 3DCNN-based VGGNets [19] with the depth of 3, the kernel size of 3, batch normalization, the ReLU activation function and two FC layers. The network had 32 feature maps on the first layer, in which the next layer had doubled as the network got deeper. Since the input data were video streams, 3D convolution can deal with not only the spatial representation but also time-series information. We replaced each convolution layer into dynamic convolution or multi-view convolution to compare and evaluate the models.

We used Adam optimizer [20] for all the networks with a learning rate of 0.001 and batch size of 64 for 100 epochs.

TABLE II
TEST ACCURACY RESULTS FOR EACH CAMERA VIEW.

Method	Training Data	Accuracy of Test Data (%)					
		0°	30°	45°	60°	90°	Average
A : Baseline 3DCNN	$0^\circ + 90^\circ$	86.9	84.3	84.5	83.3	83.7	84.6
B : Dynamic 3DCNN [11]		86.4	84,4	81.4	81.4	81.4	83.0
C : B + temperature		87.0	86.0	86.4	83.7	82.2	85.1
D : B + view regularization		85.4	78.7	73.9	78.2	80.3	79.3
E : D + temperature		86.5	82.2	80.8	81.7	82.4	82.7
F : D + LS		88.5	85.8	84.3	83.3	83.0	84.9
G : D + temperature + LS (Multi-view 3DCNN)		**88.6**	**89.4**	**88.1**	**85.6**	**83.9**	**87.1**

The learning rate decayed with rates of 10% every 5 epochs.

Through fine tuning on the validation set, we found that the optimal values of hyperparameters are the loss weight $\lambda_{view} = 1$, the smoothing $\alpha = 0.2$ and the minimum value of temperature in softmax $\tau_{min} = 3$.

C. Result and discussion

Table II shows test accuracy results for each camera view. It is found that our multi-view convolution (Method G) achieved the highest accuracy for all views. In particular, the accuracy for the side views, which were not used for training, was significantly higher. It is observed that our method can perform robust recognition for all the views.

Since our multi-view convolution consists of several components, we analyzed which component contributed to the accuracy improvement much more.

- Dynamic Convolution (Method B) causes performance degradation compared to the baseline (Method A), irrelated to the number of kernels. The dataset used in this paper has only lip videos, causing lower variation. In contrast, the other datasets such as ImageNet [21] include various kinds of images. This may affect recognition performance; Method B has too many kernels to represent such the lower variation data. On the other hand, our proposed method seems to properly adjust the representation capability owing to regularization.
- Comparing Methods B and D, adding only view regularization to dynamic convolution significantly decrease the accuracy. As mentioned, the softmax output, which is the attention weight, became close to a one-hot vector for either 0° or 90°. This might cause that each kernel was built almost only for one view, and only one kernel might be selected for a view that was not used for training. For example, for a 45° test sample, a kernel for 0° was selected as a result. Consequently, only adding view regularization to dynamic convolution caused too hard attention leading accuracy decrease.
- Next, we would like to discuss the case where view regularization and soft attention are applied together to dynamic convolution. We introduced two methods, softmax with temperature and label smoothing. When only one of them was applied, the accuracy was improved. It is found that introducing soft attention enables us to combine multiple kernels for an unknown view. Furthermore, when both of them were applied, the accuracy was also improved, indicating that both could contribute to the improvement independently. These results it is clarified that multi-view convolution, which combines dynamic convolution with view regularization and soft attention (softmax with temperature and label smoothing), is robust against unknown views.

From these results, we can conclude that multi-view convolution, which combines dynamic convolution with view regularization and soft attention (softmax with temperature, label smoothing), ensures effective for robustness against unknown views.

V. CONCLUSION

In this paper, we introduce multi-view convolution based on dynamic convolution for lipreading. In order to realize multi-view lipreading using only a few kinds of views for training and to design a mechanism that can be easily applied to existing model, we impose view regularization to dynamic convolution with soft attention (softmax with temperature and label smoothing). Experimental results indicate that our multi-view convolution is robust against unknown views.

As future works, the following can be considered. In this work, we evaluated lipreading schemes with OuluVS2. Since OuluVS2 has a limited vocabulary and data size, so far we cannot evaluate our multi-view convolution in much more realistic scenarios. Therefore, we will explore a new data set to evaluate our method in real environments or in large-vocabulary tasks.

REFERENCES

[1] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," IEEE Trans. Multimedia, vol. 17, no. 5, pp. 603–615, 2015.
[2] I. Anina, Z. Zhou, G. Zhao, and M. Pietikainen, "OuluVS2:a multi-view audiovisual database for non-rigid mouth motion analysis," in Proc. FG, pp. 1–5, 2015.
[3] P. Lucey, S. Sridharan, and D. B. Dean, "Continuous pose-invariant lipreading," In Interspeech, pages 2679–2682, 2008.
[4] P. Lucey, G. Potamianos, and S. Sridharan, "An extended pose-invariant lipreading system," In International Workshop on Auditory-Visual Speech Processing, 2007.
[5] M. Zimmermann, M. Mehdipour Ghazi, H. K. Ekenel, and J.- P. Thiran, "Visual speech recognition using PCA networks and LSTMs in a tandem GMM-HMM system," in Computer Vision – ACCV 2016 Workshops, Part II, C.-S. Chen, J. Lu, and K. K. Ma, Eds. 2017, vol. LNCS 10117, pp. 264–276, Springer.
[6] D. Lee, J. Lee, and K.-E. Kim, "Multi-view automatic lipreading using neural network," in Computer Vision – ACCV 2016 Workshops, Part II, C.-S. Chen, J. Lu, and K. K. Ma, Eds. 2017, vol. LNCS 10117, pp. 290–302, Springer.

[7] T. Watanabe, K. Katsurada, and Y. Kanazawa, "Lip reading from multi view facial images using 3D-AAM," in Computer Vision – ACCV 2016 Workshops, Part II, C.-S. Chen, J. Lu, and K. K. Ma, Eds. 2017, vol. LNCS 10117, pp. 303–316, Springer.
[8] T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikainen, "Concatenated frame image based CNN for visual speech recognition," in Computer Vision – ACCV 2016 Workshops, Part II, C.-S. Chen, J. Lu, and K. K. Ma, Eds. 2017, vol. LNCS 10117, pp. 277–289, Springer.
[9] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end multiview lipreading," in BMVC, 2017.
[10] M. Zimmermann, M. Mehdipour Ghazi, H. K. Ekenel, and J.-P. Thiran, "combining multiple views for visual speech recognition," in AVSP, 2017.
[11] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu, "Dynamic convolution: Attention over convolution kernels," In CVPR, 2020.
[12] A. Koumparoulis, G. Potamianos, "Deep view2view mapping for view-invariant lipreading," IEEE SLT Workshop (2018), pp. 588-594, 2018.
[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-imagetranslation with conditional adversarial networks," In CVPR, 2017.
[14] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: Sentence-level lipreading," In ICLR, 2017.
[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," In CVPR, 2017.
[16] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," In Interspeech, 2017.
[17] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," In CVPR, 2018.
[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Zb. Wojna, "Rethinking the inception architecture for computer vision," In CoRR, abs/1512.00567, 2015.
[19] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," In ICLR, 2015.
[20] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," In ICLR, 2014.
[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," In International Journal of Computer Vision, 2015.