2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)
October 11-14, 2020. Toronto, Canada

Part-Based Lipreading for Audio-Visual Speech Recognition

Ziling Miao, Hong Liu* and Bing Yang
Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University
Shenzhen, China
zilingmiao@pku.edu.cn, hongliu@pku.edu.cn, bingyang@sz.pku.edu.cn

***Abstract*—Lipreading is an important component of audio-visual speech recognition. However, lips are usually modeled as a whole in lipreading, which ignores that each part of lip focuses on different characteristics of mouth and the overall model can not fit each part perfectly. Besides, features based on the whole lip usually vary a lot according to different speakers, which leads that the training databases usually need to contain as much speakers as possible. In this paper, A part-based lipreading (PBL) method is proposed to deal with the mismatch between an overall lip model and the separate parts of lips, also the excessive dependence of models on the speakers in training set. PBL models lips partly and predicts jointly. It employs a uniform partition strategy on convolutional features and generates several part-level sub-results for final prediction. Experiments are performed on a large publicly available dataset (LRW) and part of it (p-LRW, 65 words), in order to simulate the progressive instructions in the working scene of robots. Word accuracy of PBL reaches 82.8% on LRW and 88.9% on p-LRW. Finally, an end-to-end audio-visual speech recognition system using PBL is established and achieves 98.3% word accuracy on LRW.**

***Index Terms*—Service robots, audio-visual speech recognition, part-based lipreading**

I. INTRODUCTION

Audio-visual speech recognition is important for the human-computer interaction system, especially the service robots. It is the task of recognizing words in a video based on both audio and visual signals. The introduction of visual information can help the robots localize the speakers and understand the instructions better, which are conducive to the friendliness and effectiveness of the human-computer interaction system.

Many methods have been proposed for audio-visual speech recognition. For traditional methods, features are first extracted around the mouth region of interest (ROI) and audio waveform, and then concatenated to be matched to a normal template [1], [2]. In recent years, with the development of deep learning technology, audio-visual speech recognition has received extensive attention from researchers. Koller et al. [3] and Noda et al. [4] trained an image classifier CNN to discriminate visemes. For word recognition, in order to make full use of the deep convolutional layers to explore more highly abstract features, the deep bottleneck features (DBF) is used by Tamura et al. [5] for feature encoding. Similarity, Petridis and Pantic [6] also applied DBF on the image of every frame. Then, considering that lipreading needs both temporal and spatial information, Tran et al. [7] used 3D convolutional filters to process the image. By consulting methods in other fields, Chung et al. [8] applied an attention mechanism to both the mouth ROIs and MFCCs. In order to build a thorough end-to-end network, Petridis et al. [9] used LSTMs to extract features from the raw data. However, these methods usually take the lip as a whole, which ignores the independent functions of separate lip parts.

In this paper, PBL is proposed based on the phenomenon that different parts of lips may focus on different mouth characteristics during the recognition (smiles are judged mostly by the corners of lips). It shows that the judgement of a specific feature may not use the whole lip but only part of it. Some facts also show that not every point of the lip will be deformed during the pronunciation. For instance, the phoneme ’/ɪ/’ and ’/i:/’ are similar in central part of lips but distinct in corner, while the distinguishment of phoneme ’/ə/’ and ’/ɑ/’ is opposite. In order to focus on the deformation part of the lip and reduce the interaction between mouth characteristics, PBL models lips partly. It takes a whole lip image as the input and cuts every convolutional feature into several parts, then models each part with independent parameters. Finally, it combines the results from every part to predict. The thought of PBL is concise: the network architectures of each part keep consistent, with modifications just on the parameters of models. Based on that specific characteristics usually emerge on a certain part of the lip, PBL uses part-level models to avoid the over-fitting from using an overall model. Meanwhile, part-level models have smaller receptive fields so that they have lower sensitivity of changing lips, which means the model focuses on not which speaker the lip from but the intrinsic features it has. So PBL decreases the dependence of models on speakers in the training set. Moreover, PBL obtains fine-grained information and fits lips better by using multi-level parameters (different parts with different parameters). Besides, PBL analyzes mouth characteristics respectively and concludes jointly, which avoids the mutual influence between characteristics and ensures the integrity of information.

This paper extends the end-to-end audio-visual speech recognition network in [10] with PBL. Experiments are divided into two parts. One is video-only recognition (PBL) and the other is audio-visual recognition. For PBL, experiments are performed on the whole LRW database and part of it (p-LRW, some words in LRW are selected to be a new small database), to simulate the progressive instructions in industrial applications. PBL is

978-1-7281-8526-2/20/$31.00 ©2020 IEEE

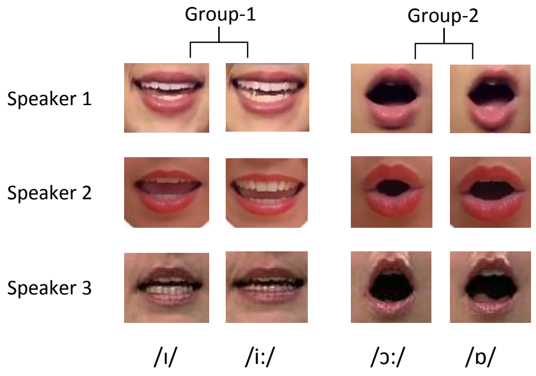

Fig. 1: Examples of lips. Each row shows one person's lips for four phonetic symbols.

0.8% higher over the whole database and 2.1% higher over the p-LRW than the baseline. For the audio-visual recognition, a 0.3% absolute promotion is accomplished over the whole LRW. The contributions of our paper are as following: 1) PBL builds models based on part-level features, which are more robust on lips from various speakers, 2) PBL fits lips in multiple levels and obtains more fine-grained information, 3) PBL is proved to be effective in both robot applications and academic tasks using lipreading.

II. Proposed Method and System Framework

A. Part-Based Lipreading

PBL aims to learn from separate lip parts to enhance the system robustness and optimize the fitting degree of models. It originates from a common sense: similar pronunciations lead to similar lip shapes. Closer observation shows that similar lip shapes may only differ obviously in parts, as shown in Fig. 1. There are two situations when we try to distinguish two similar pronunciations of the same speaker. In Group-1, different characteristics are mainly in the middle of lip, like the longitudinal open-close degree of mouth, teeth and radian of the lower lip. While in Group-2, differences are mainly from the corner part, such as the transversal open-close degree and the angle of mouth corner. For further investigations, lips under uniform partition are introduced as in Fig. 2. In Group-1, differences in the center part are more obvious than the right and left corner, the former shows more variety in mouth height, teeth and radian. Group-2 is opposite, angle of the mouth corner is distinct among comparison, and it only relates to the right and left corner of mouth. Above all, PBL learns from each part independently, in order to extract more part-level lip variety as possible. In other words, changes of some characteristics occur only in one part of lip (teeth, mouth height and lip radian in middle part, mouth angle and transversal mouth length in corner part), so modeling all kinds of variety with an overall framework may not perfectly fit each one. PBL aims to solve this problem by processing each part with different framework parameters. Besides, according the experience from other fields, part-level features are more robust on various objects. In lipreading, part-level lip features may fit the various lips well and decrease the dependence of models on speakers in the training set.

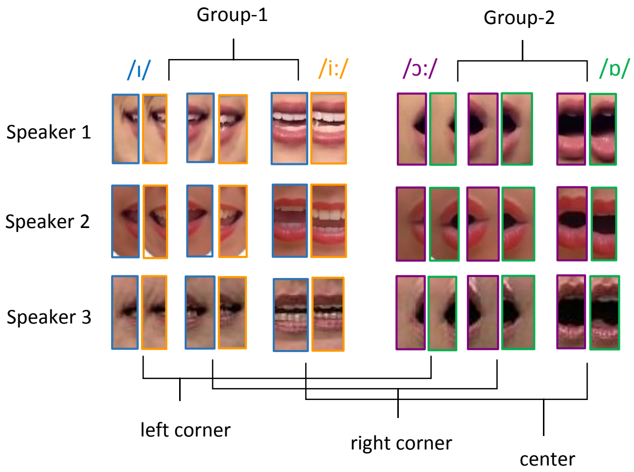

Fig. 2: Uniform divided lips. Each figure from Fig. 1 is divided into three parts: right corner, left corner and center considering the symmetrical structure of lips. Frame colors are consistent with the corresponding phonetic symbols. In each group, for each speaker, the same part from different pronunciations (different colors) are put together for comparison.

Inspired by the part-based method in person re-identification field [11], a PBL network is introduced to lipreading as shown in Fig. 3. First the lip sequences are sent into a front-end spatio-temporal convolutional unit, which consists of a 3D convolutional layer, a BN layer, a ReLU layer and a MaxPooling layer. There are 64 kernels in the 3D convolutional layer, each has 5 by 7 by 7 size. Next the outputs are sent into a 34-layer ResNet [12] for feature extraction. Then, each feature map is divided into three parts according to the real space relationship they have. For each part, two 512-cell Bidirectional Gated Recurrent Unit (BGRU) layer and a SoftMax layer are added to independently model their temporal varieties. After that, there are three intermediate predictions corresponding with three losses. We summed them up as the final loss for the ultimate prediction as follows:

$$loss = -\sum_{n=1}^{N}\sum_{d=1}^{D}[\frac{y_d}{D}log(p_d^n) + \frac{1-y_d}{D}log(1-p_d^n)], \quad (1)$$

where D stands for the output dimension (just the number of total classes), and d is the index of D. N is the number of lip parts, and n is the index of N. p_d^n is the prediction result on d^{th} frame from the n^{th} part-level network. y_d expresses right label the d^{th} frame should be.

The spatio-temporal convolutional unit is proven to be skilled in extracting the short-term dynamics of the mouth region even when there are recurrent networks deployed after [13]. In this paper, considering that ResNet is prepared for static colored images from ImageNet or CIFAR, the 34-layer ResNet in PBL is trained from zero without any pre-trained models. Moreover, the same weights are allocated to three parts during the loss

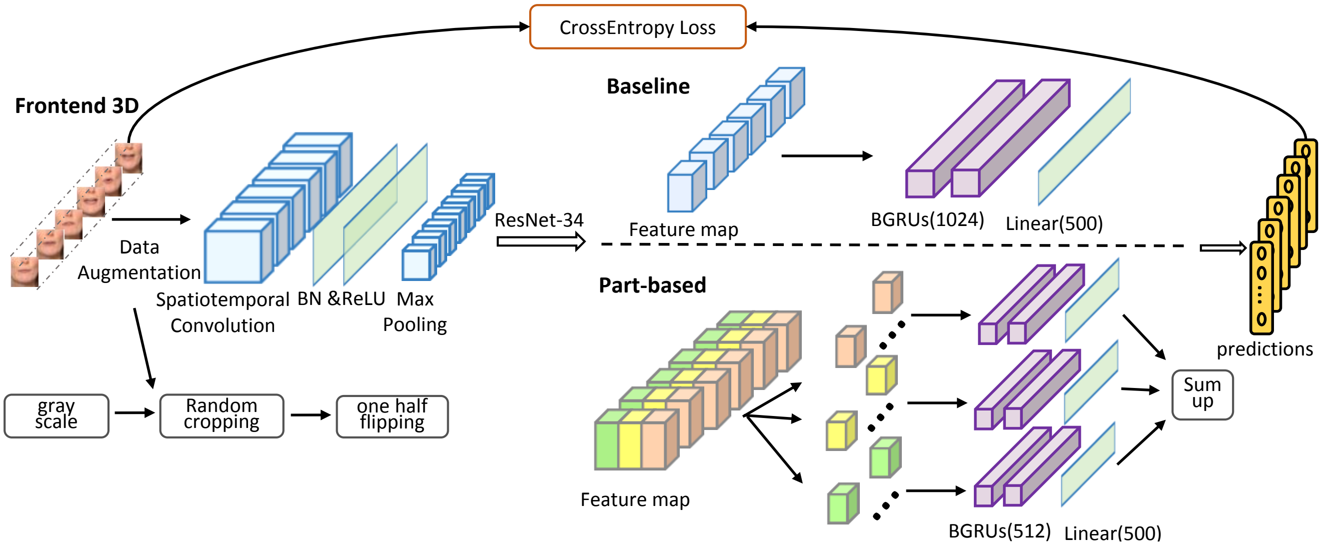

Fig. 3: Structure of PBL and the baseline. Lower left shows the data augmentation talked in III-A. Visual stream of [10] is taken as the baseline, which takes the lip feature map as a whole.

summation, because of the uncertainty when the middle part is significant, and when the corner parts are more important during lipreading.

B. Audio-Visual Recognition Framework

A brief framework of the end-to-end audio-visual speech recognition system is shown in Fig. 4, which is based on [10]. The "BGRU-V" unit contains two 1024-cell BGRU layers in baseline and six 512-cell BGRU layers in PBL. Details about the video stream are in Section II-A. In the following, we will first describe the audio stream and then the audio-visual fusion stream.

Audio stream is built with an 18-layer ResNet followed by two BGRU layers as in [10]. The raw audio signals are first fed into a front-end temporal convolutional unit. It mainly contains an 1D convolutional layer. Outputs of the unit are adjusted into 29 windows in order to keep up with the video frame rate. Then they are fed into the ResNet to extract much deeper features. Finally, the high-level features are fed into two 1024-cell BGRU layers to model the long term temporal information.

For audio-visual fusion, the outputs of two modalities are one-to-one concatenated and then fed into two 1024-cell BGRU layers. Ultimately, a SoftMax layer with 500 cells is used for classification, which is formulated by

$$s_i = \frac{e^{p_i}}{\sum_{i=1}^{n} e^{p_i}}, \tag{2}$$

where i is the index of output dimension, $i \in [1, 500]$. p_i is the initial prediction result, and s_i is the normalized prediction result. The final result is decided by the highest average probability.

III. EXPERIMENTS SETTINGS AND RESULTS

A. Database and Preprocessing

The Lip Reading in the Wild (LRW) database [14] is the largest publicly available English-based word-level lipreading dataset. It contains around 500000 video slices from BBC TV programs. Each slice has fixed resolution and the same length (1.16 seconds), which offers much convenience to the community. Moreover, there are more than 1000 speakers and 500 classes of words, much larger than other earlier word-level lipreading databases [15]–[17].

Videos of each class have been divided into three parts: training, validation and test sets. There are less than 1000 and more than 800 samples in training set, where some classes have more samples than others, and 50 samples in both validation set and test set. Moreover, in order to strengthen the robustness to noise, several levels of babble noise (SNR from -5dB to 20dB) from the NOISEX database [18] are added under uniform distribution during training.

In order to establish a single-variable comparative experiment, we frame out the mouth ROI in fixed size of 96 by 96 pixels as in [10]. Then the frames are changed into grayscale and finally normalized using unified mean and variance. During training, random cropping and horizontal flipping with probability 50% are performed for data augmentation.

Given that signals from various natural environments have different levels of loudness, each audio slice is normalized with mean zero and standard deviation one.

B. Implementation Details

The Adam training algorithm [19] is used by both the entire audio-visual network and the single streams. The size of mini-batch is 16 and the initial learning rate is set to 0.0003. BatchNorm is added after all convolutional and linear layers if they are not the preceding one of the SoftMax layer. The cross entropy loss is used per time step during training.

The whole training stage starts with the independent training of visual and audio stream, then the audio-visual network is trained end-to-end. According to [10] and [20], the training phase of the two streams are divided into 3 steps: first, a

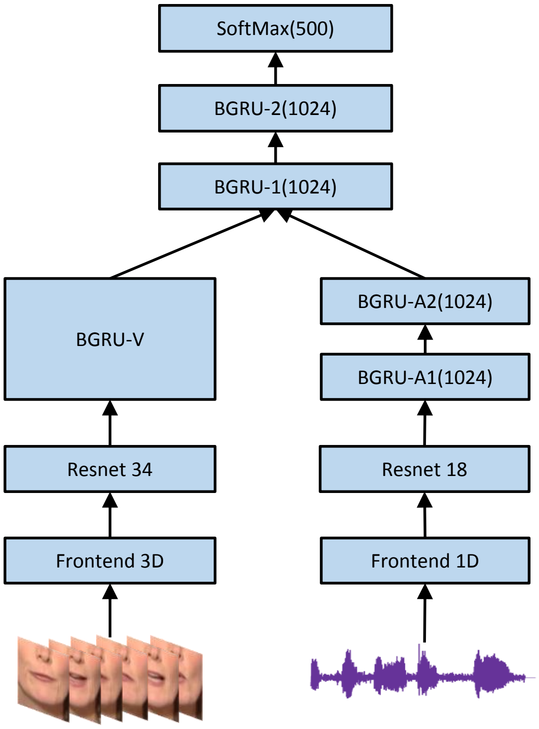

Fig. 4: Framework of the end-to-end audio-visual speech recognition system.

temporal convolutional backend is added after the ResNet to form a classification network. Then it is trained until working to its best. Second, the convolutional backend is detached and the BGRU layers are attached. Then the network is fine-tuned for 5 epochs with fixed parameters of the front-end convolutional unit and the ResNet. Last, the network is trained end-to-end. For the audio-visual training, there are two steps: first, BGRU layers are fine-tuned for 5 epochs keeping the parameters fixed in two single streams. Second, the whole network is trained end-to-end.

More specifically, the temporal convolutional backend mentioned before consists of two units. The first unit mainly contains two temporal convolutional layers and a MaxPooling layer. The second unit mainly contains a linear and a SoftMax layer.

C. Results

Video-only results: The same network is experimented twice on two databases, in order to verify if PBL works well with the progressive instructions, which can be formulated by

$$output_i = f(input_i|database_i), \tag{3}$$

where i is the index of processes. $output_i$ is the recognition result of the $input_i$ based on $database_i$. $f(\cdot)$ is the speech recognition system. Inputs here are a series of words. While $i = 1$, the first instruction $input_1$ has no prior knowledge, so the system works on the whole dataset $database_1$. Then the database is limited to a smaller scale $database_2$, according to the latest recognition result $output_1$. And the second input $input_2$ are recognized based on $database_2$. A regular is that following databases are updated by their previous outputs.

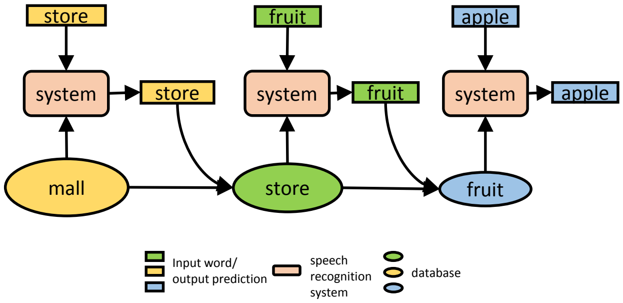

Fig. 5: Progressive instructions for shopping guide robot. After the recognition of word "mall", the database can be reduced into subset "store" from the initial "mall", the recognition of word "fruit" is performed on "store" database, and the recognition of word "apple" is performed on "fruit" database.

Table 1: Two databases for experiments. p-LRW is combined by part of words from LRW.

Database	***# of Classes***	***Instructions***
LRW	***500***	***various, copious***
p-LRW	***65***	***few species, small quantity***

An example of progressive instructions is shown in Fig. 5. Our experiments simulate a simple scene of progressive instructions: the length of input sequence is set to 2. So that two databases are needed, one is the initial whole database and the other is a smaller subset. Details of the databases are shown in Table 1. We perform PBL and the baseline models on them. During experiments, we train models only on the whole LRW and test them on both LRW and p-LRW. Like most researches on lipreading, we train and test models on the LRW dataset to verify the effectiveness of PBL academically. Apart from that, the ability of practical application is also important. So the p-LRW is introduced for another test. It is based on the progressive instructions in actural case, where models are trained once on the initial database and are not allowed to be retrained when the database update.

Results are shown in Table 2. A slightly improvement of 0.8% is shown when we test on the LRW database. A conclusion can be obtained that PBL works better than the traditional method in baseline, which takes the lip as a whole. Meanwhile, an obvious superiority of 2.1% on p-LRW database is achieved. We can see that the superiority of PBL becomes more significant with the shrinking database. It certificates that PBL owns the ability to keep efficient with a changing database, thus it has a wide industrial application prospect. Because actural cases own larger lip variety than laboratory cases and PBL provides more system robustness, so it performs better than the baseline especially in actual cases. Besides, PBL works well when the database changes from LRW to p-LRW, which provides the possibility for system transformation from complicated and universal scenes to simple and specific scenes without changing the structure of models and re-training. It

Table 2: Word accuracy (WA) for the baseline and PBL video-only stream on LRW and p-LRW. [20] is a similar end-to-end model.

Visual	**WA**	
Stream	***LRW***	***p-LRW***
baseline	***82.0%***	***86.8%***
ours	***82.8%***	***88.9%***
[20]	***83.0%***	——

Table 3: Word accuracy (WA) for the baseline and PBL based audio-visual network in the noise-free environment.

Network	***AR(LRW)***
baseline A-V	***98.0%***
PBL A-V	***98.3%***

further proves that part-level models can fit different data fields well. In addition, PBL is slightly lower than 83.0%, because a facial landmark regression algorithm [21] is used in [20]. But in this work, PBL just crops the mouth ROI by the centre of image as in baseline.

Moreover, we measure the computational cost of models by the processing FLOPs and model parameters. For an 88×88 input image, the FLOPs of the baseline [10] and PBL are 3.23G and 3.22G seperately. In addition, the model parameters of the baseline and PBL are 29.03M and 27.59M. The PBL achieves better performance with a lightweight model.

Audio-visual fusion system results: Results for the whole audio-visual fusion network are shown in Table 3, system using PBL brings 0.3% superior than the baseline. Although slight, the leading seems momentous especially when the baseline has received an accuracy up to 98%. A conclusion can be drawn from the result that PBL is effective whether it works separately or as a component of other tasks. The part-level features and models of lips have stable superiority in various tasks.

IV. CONCLUSIONS

We propose a PBL method and establish an end-to-end audio-visual fusion system using PBL. PBL is based on the phenomenon that different parts of the lip focus on different kinds of characteristics. It employs uniform partition strategy on lip feature maps and assembles part-informed classification results together. PBL enhances the system robustness on lips and fitting degree of models. Experiments on LRW and p-LRW (a database for the progressive instructions in the working scene of robots) show that PBL works better academically and practically with a lightweight framework. The end-to-end audio-visual system using PBL also outperforms the baseline model, which shows the stable superiority of part-level features and models.

ACKONWLEDGMENT

This work is supported by National Natural Science Foundation of China (No. 61673030, U1613209), National Natural Science Foundation of Shenzhen (No. JCYJ20190808182209321).

REFERENCES

[1] 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.

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

[3] O. Koller, H. Ney, R. Bowden. “Deep learning of mouth shapes for sign language,” IEEE International Conference on Computer Vision Workshops. 2015, pp. 85–91.

[4] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, T. Ogata. “Lipreading using convolutional neural network,” INTERAPEECH, 2014, pp. 1149–1153.

[5] S. Tamura, H. Ninomiya, N. Kitaoka, S. Osuga, Y. Iribe, K. Takeda, S.Hayamizu. “Audio-visual speech recognition using deep bottleneck features and high-performance lipreading,” Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015, pp. 575–582.

[6] S. Petridis, M. Pantic. “Deep complementary bottleneck features for visual speech recognition,” International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 2304–2308.

[7] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri “Learning spatiotemporal features with 3d convolutional networks,” The IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4489-4497.

[8] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3444-3450.

[9] S. Petridis, Y. Wang, Z. Li, and M. Pantic, “End-to-end audiovisual fusion with LSTMs,” arXiv preprint arXiv:1709.04343, 2017.

[10] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, M. Pantic. “End-to-End Audiovisual Speech Recognition,” International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 6548-6552.

[11] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang. “Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline),” European Conference on Computer Vision (ECCV), 2018, pp. 480-496.

[12] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[13] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas. “Lipnet: Sentence-level lipreading,” arXiv preprint arXiv:1611.01599, 2016.

[14] J. S. Chung and A. Zisserman. “Lip reading in the wild,” Asian Conference on Computer Vision (ACCV), 2016, pp. 87–103.

[15] M. Cooke, J. Barker, S. Cunningham, and X. Shao. “An audio-visual corpus for speech perception and automatic speech recognition,” The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421–2424, 2006.

[16] I. Anina, Z. Zhou, G. Zhao, and M. Pietik¨ainen. “Ouluvs2: A multi-view audiovisual database for nonrigid mouth motion analysis,” IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015, pp. 1–5.

[17] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. “Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus,” EURASIP Journal on Advances in Signal Processing, vol. 208541, no. 2002, pp. 1189–1201, 2002.

[18] A. Varga and H. Steeneken. “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” Speech Communication, vol. 12, no. 3, pp. 247–251, 1993.

[19] D. Kingma and J. Ba. “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[20] T. Stafylakis and G. Tzimiropoulos. “Combining residual networks with LSTMs for lipreading,” INTERSPEECH, 2017, pp. 3652–3656.

[21] A. Bulat and G. Tzimiropoulos. “Two-stage convolutional part heatmap regression for the 1st 3D face alignment in the wild (3DFAW) challenge,” European Conference on Computer Vision (ECCV). 2016, pp. 616–624.