INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Lip Graph Assisted Audio-Visual Speech Recognition Using Bidirectional Synchronous Fusion

Hong Liu, Zhan Chen, Bing Yang

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, China
{hongliu, zhanchen_cz}@pku.edu.cn, bingyang@sz.pku.edu.cn

Abstract

Current studies have shown that extracting representative visual features and efficiently fusing audio and visual modalities are vital for audio-visual speech recognition (AVSR), but these are still challenging. To this end, we propose a lip graph assisted AVSR method with bidirectional synchronous fusion. First, a hybrid visual stream combines the image branch and graph branch to capture discriminative visual features. Specially, the lip graph exploits the natural and dynamic connections between the lip key points to model the lip shape, and the temporal evolution of the lip graph is captured by the graph convolutional networks followed by bidirectional gated recurrent units. Second, the hybrid visual stream is combined with the audio stream by an attention-based bidirectional synchronous fusion which allows bidirectional information interaction to resolve the asynchrony between the two modalities during fusion. The experimental results on LRW-BBC dataset show that our method outperforms the end-to-end AVSR baseline method in both clean and noisy conditions.

Index Terms: Audio-visual speech recognition, spatial-temporal graph convolution, deep learning

1. Introduction

Audio-visual speech recognition (AVSR) has been investigated intensively over the last few decades. It is inspired by human bimodal speech perception which leverages not only acoustic information but also visual information to reduce speech confusion [1, 2]. Although acoustic signal carries most speech information, it is not reliable in the acoustic noisy environment. In this case, introducing the visual information can help to improve the performance of the speech recognition system, since the visual information is not affected by acoustic noise [3, 4].

Many studies focus on visual feature extraction which directly affects the performance of the recognition system. In earlier studies, transforms like principal component analysis (PCA) [5] and discrete cosine transform (DCT) [6] were applied to the mouth region of interest (ROI) to extract visual feature. With the development of deep learning, the traditional transforms are replaced by deep autoencoder [7], which achieves a great improvement than the traditional transforms. The adoption of CNNs such as VGG [8], ResNet [3] and DenseNet [9] which are used to extract visual features from the raw mouth images further improves the performance of system and even outperforms the professional lip reader [8]. However, these appearance-based visual features which are used to model characteristics of the mouth region exhibit greater sensitivity to the environmental condition changes such as illumination, which limits the performance of the AVSR. The literature shows that the addition of extra visual information (*e.g.* optical flow which provides complementary temporal visual information and shape-based features which extract geometrical measurements of the lip) to appearance-based features significantly improves lipreading performance [10, 11]. Wang et al. [12] used extra 3D lip information obtained from Kinect to improve the performance of multimodal speech recognition. Tao et al. [13] combined six distance features which are used to describe the shape of mouth with the appearance-based features to improve the robust against speaker variability. Wang et al. [14] concatenated the appearance-based features with the optical flow which was used to capture lip motion information, and this method significantly improved the performance. Although, the distance features and the optical flow improve the performance of the appearance-based features, more representative visual features still need to be explored.

The fusion strategy adopted to fuse the information of audio and visual modalities is also crucial to AVSR and can be broadly categorized into two kinds, namely feature fusion [3, 15, 16] and decision fusion [12, 17, 18]. Feature fusion is a commonly used approach since the feature fusion benefits from the correlation of modalities at the feature level [19]. The feature fusion faces the problem that the audio stream and visual stream may be not synchronized [19, 20]. The asynchrony between speech and visual clues involves the anticipatory coarticulation which refers to one gesture beginning in advance and the preservatory coarticulation which refers to a gesture continues after [19]. The time-variant phase between the two modalities can be hundreds of milliseconds [20], which degrades the speech recognition performance. To solve the asynchronous problem between the two modalities. Bregler and Konig [21] assumed that lip movements preceded speech and the best synchronization was a shift of 120 milliseconds. However, the time-variant phase between two streams is obviously uncertain, and it is closely related to the words. Tao et al. [22] proposed an attention based data driven alignment neural network to generate the aligned visual feature according to audio feature, which means the alignment is highly related to the reliability of acoustic information. Sterpu et al. [23] proposed an attention-based audio-visual fusion strategy which allows the acoustic modality to learn correlation from the visual modality. Despite these methods have achieved desired results, the design of unidirectional information interaction causes the system to over-rely on one modality.

In order to deal with the above two challenges, we propose a novel lip graph assisted AVSR method which uses the bidirectional synchronous fusion strategy. The adopted baseline is the end-to-end AVSR method in [3], and we extend the baseline method with two aspects. First, we add a graph branch in the visual stream. The graph branch uses the spatial-temporal graph convolutional networks (ST-GCN) [24] followed by a 2-layer bidirectional gated recurrent unit (BGRU) to extract shape-based features in an end-to-end manner. The graph branch regards the key points of the lip as a lip graph in a non-Euclidean space with the key points as nodes and their natural connections in the mouth as edges, rather than independent feature

http://dx.doi.org/10.21437/Interspeech.2020-3146

points or distances. Thus, the graph branch can exploit the relationships among the key points, and the relationships are crucial for understanding visual speech. Second, we propose an attention based bidirectional synchronous fusion to process the asynchrony between the two modalities which is ignored in [3]. The sync block allows bidirectional information interaction between the audio and visual modalities to explore the correlation between the two modalities, which can synchronize the two modalities and balance the dependence of system on the two modalities. The experiments on LRW-BBC dataset show an absolute improvement of 0.39% and 6.49% over the end-to-end AVSR baseline method, at 20 dB and -5 dB, respectively.

2. Approach

The pipeline of the proposed AVSR method is shown in Figure 1. We use an audio stream to extract audio feature from the acoustic signal and a hybrid visual stream to extract visual feature from visual signal. A bidirectional synchronous fusion is applied to fuse the audio feature and visual feature.

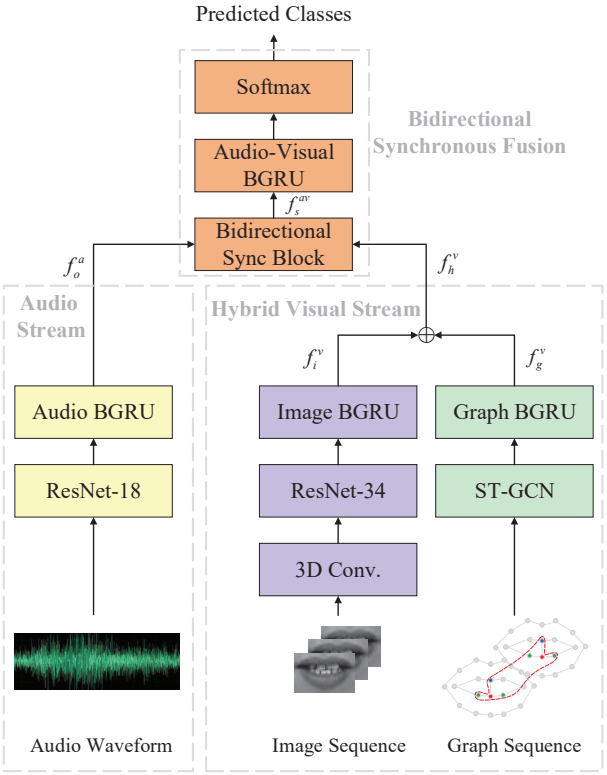

Figure 1: *Overview of the proposed audio-visual speech recognition method.*

2.1. Audio Stream

As shown in Figure 1, the audio stream uses the same architecture as [3]. Audio features are extracted from the audio waveforms by the ResNet-18, and divided into multiple frames. In order to model the temporal dynamics of the audio, these framed features are fed into a 2-layer BGRU which consists of 1024 cells in each layer. The output of the audio stream is the extracted audio feature f_o^a.

2.2. Hybrid Visual Stream

2.2.1. Image feature extraction

The image branch consists of a spatial-temporal convolution followed by a 34-layer ResNet and a 2-layer BGRU. where the spatial-temporal convolution is used to capture the short-term temporal dynamics, and the ResNet and the 2-layer BGRU are used to extract spatial features from the image sequences and model the dynamics of the mouth region, respectively. The output of the image branch is image feature f_i^v.

2.2.2. Graph feature extraction

The graph branch aims to learn shape-based features of the lips in an end-to-end manner. To the best of our knowledge, it is the first end-to-end model which models the lip and extracts shape-based features by learning both the key points and their relationships. The key points and the relationships between them constitute the lip graph with the key points as the nodes and their connections as the edges. Then we apply ST-GCN together with BGRUs to model the lip, since GCN is a general and effective framework for learning representation of graph structured data and various GCN variants have achieved the state-of-the-art results on many task [24, 25, 26].

ST-GCN consists of a series of the ST-GCN blocks. The illustration of a ST-GCN block is shown in Figure 2. Each block contains a spatial graph convolution (SGC) followed by a temporal graph convolution (TGC), which extracts spatial and temporal features alternatively.

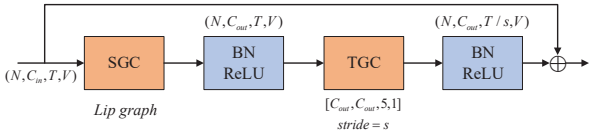

Figure 2: *Illustration of the ST-GCN block. N, C, T, V denote the mini-batch, channel, frames and nodes, separately.*

The temporal graph convolution operation is similar to the 2D convolution and performs $\mathcal{T} \times 1$ filters on the lip graph sequences to capture short-term dynamic information of the lip movement. $\mathcal{T}$ is the temporal size of the filters and is set to 5 in our work.

The spatial graph convolution operation is the key component in ST-GCN. We consider a lip graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of M nodes and $\mathcal{E}$ is the set of edges. The neighbor set of a node v_i is defined as $\mathcal{N}(v_i) = \{v_j | d(v_i, v_j) \leq D\}$, where $d(v_i, v_j)$ is the minimum path length from v_j to v_i. A graph labeling function $\mathcal{L} :\rightarrow \{1, 2, ...K\}$ is designed to assign the labels $\{1, 2, ...K\}$ to each graph node $v_i \in \mathcal{V}$, which can partition the neighbor set $\mathcal{N}(v_i)$ of node v_i into a fixed number of K subsets. Figure 3(a) shows the uniform graph which is the simplest and most straight forward graph, while K is set to 1. The graph is suboptimal on our task as the local differential properties could be lost in this operation [24]. (b) shows the shape graph which is modified by considering the shape of the lips. Since we set D to 1 in this work, there are two different weight vectors and they are capable of modeling the shape transformation of the lip, and K is set to 2. (c) shows the lip graph. The edges in lip graph represent not only the shape information but also the motion information, since the symmetric nodes of the root node can represent the mouth opening and closing and K is set to 3. Thus, in our work, we use the lip

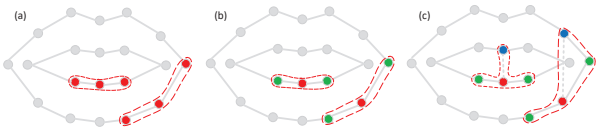

Figure 3: *Graphical spatial dependencies between different nodes. (a)* ***Uniform graph****: all nodes in a neighborhood has the same label (red). (b)* ***Shape graph****: the neighboring nodes are divided into two classes which are the root node (red) itself and the other neighboring points (green). (c)* ***Lip graph****: the neighboring nodes are divided into three classes: the root nodes(red), the shape-linked nodes (green) and the symmetric nodes (blue).*

graph, and the graph convolution can be generally computed as:

$$Y_{out}(v_i) = \sum_{v_j \in \mathcal{N}(v_i)} \frac{1}{Z_i(v_j)} X(v_j) W(\mathcal{L}(v_j)), \quad (1)$$

where $X(v_j)$ is the feature of node v_j. $W(\cdot)$ is a weight function that allocates a weight indexed by label $\mathcal{L}(v_j)$ from K weights. $Z(\cdot)$ is the number of the corresponding subset, which normalizes feature representations. $Y_{out}(v_i)$ denotes the output of graph convolution at node v_i. More specifically, with the adjacency matrix, the Eq.(1) can be represented as:

$$Y_{out}(v_i) = \sum_{k=1}^{K} \Lambda_k^{-\frac{1}{2}} A_k \Lambda_k^{\frac{1}{2}} X W_k, \quad (2)$$

where A_k is the adjacency matrix in lip-graph configuration of the label $k \in \{1, 2, ..., K\}$. $\Lambda_k^{ii} = \sum_j A_k^{ij}$ is a degree matrix.

The ST-GCN used in graph branch consists of 10 ST-GCN blocks. The first four blocks have 64 channels for output, the following three blocks have 128 channels for output, and the last three blocks have 256 channels for output. With the hierarchical structure of ST-GCN block, the ST-GCN is capable to model both shape information and motion information of the lip which is highly related to the speech. And a 2-layer BGRU is also added on the ST-GCN to capture the long-term temporal dynamics. The output of the graph branch is graph feature f_g^v.

2.2.3. Visual feature combination

With the extracted features from image and graph sequences, these two kinds of visual features are combined by a weighting scheme which can be formulated as:

$$f_h^v = \lambda \times f_g^v + (1 - \lambda) \times f_i^v, \quad (3)$$

where f_h^v denotes the hybrid visual feature. λ is a hyper-parameter, which trades off the importance between the graph feature and the image feature. In our work, λ is set to 0.3.

2.3. Bidirectional Synchronous Audio-Visual Fusion

In order to fuse the information of these two modalities, we propose an attention-based bidirectional synchronous fusion which consists of a bidirectional sync block and a 2-layer BGRU. Specifically, the bidirectional sync block is applied to synchronize the audio feature and visual feature, then the synchronized audio feature f_s^a and visual feature f_s^v are concatenated as synchronized audio-visual feature f_s^{av}. To further fuse the information of these two modalities, the synchronized audio-visual feature is fed into a 2-layer BGRU.

The bidirectional sync block is proposed to achieve bidirectional feature synchronization and represent audio-visual feature in a more meaningful way. The visual features need to learn synchronization information from the audio features to reduce the asynchrony between the two modalities. Similarly, the audio features need to learn from visual features to achieve audio-visual synchronization. We take the advantage of the attention mechanism which allows probabilistic many-to-many relations to implement the idea of synchronizing visual and audio features. Thus, the above processes can be formulated as:

$$f_s^v = Sync(f_o^a, f_h^v), \quad (4)$$
$$f_s^a = Sync(f_s^v, f_o^a), \quad (5)$$

Through these two steps, we can achieve bidirectional synchronization of audio and visual features and allow the system to learn the correlation between these two modalities.

Figure 4 shows the bidirectional sync block and the first step in detail. The queries in attention mechanisms are from audio feature, while the keys and values are generated by visual feature. To explore the synchronization between audio and visual feature, the synchronous matrix W is generated by comparing the similarity between the queries which contain the high-level semantic acoustic information and the keys which contain the high-level visual information. Finally, we can obtain the synchronized visual features by further inference about synchronous matrix and visual information. With this architecture, the cross modal information integration and synchronization are realized. The second step is similar to the first step. The parameters used in the two synchronization processes are shared.

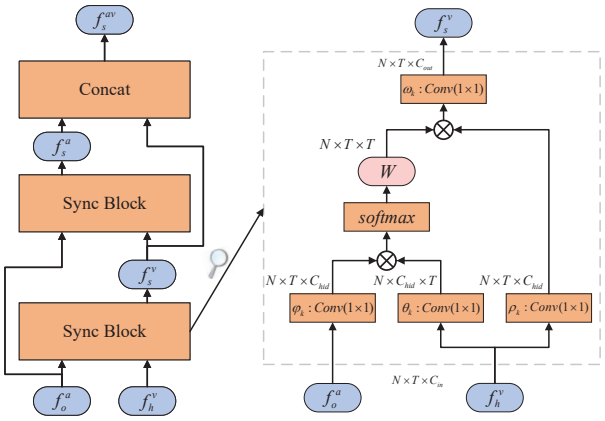

Figure 4: *Illustration of bidirectional sync block.* 1×1 *denotes the kernel size of convolution. W denotes the weight to synchronize features.* $\otimes$ *denotes the matrix multiplication.*

3. Experiments and Discussions

3.1. Dataset

To perform a fair comparison, our experiments are conducted on the same dataset with [3]: Lip Reading in the Wild (LRW) database [8]. It contains over 480, 000 video clips and the clips are performed by over 1, 000 subjects. There are 500 word classes in total and each class contains 800 to 1000 samples for training, 50 samples for validation and 50 samples for test. In our work, the mouth ROI is extracted with a fixed bounding box of 96 by 96, and then the frames are transformed to grayscale

and normalized. Since the video clips in the LRW do not contain the locations of lip key points, we use the public available Dlib [27, 28] toolbox to estimate the 2D location of 20 lip key points on each frame of the clips, named as "LRW Lip".

3.2. Training

The same data augmentation as in [3] is performed on both image sequences and audio sequences during training. Since directly end-to-end training leads to suboptimal performance, we follow the same training procedure as in [29]. First, a temporal convolution back-end combined with ResNet or ST-GCN is trained to initialize ResNet or ST-GCN. Second, the temporal convolution back-end is replaced by 2-layer BGRU, and the BGRU is initialized by training for 5 epochs, keeping the parameters of ResNet or ST-GCN fixed. Finally, the Adam optimizer with an initial learning rate of 0.0003 and a mini-batch of 36 is used to optimize each stream. After each stream has been trained, the sync block followed by another 2-layer BGRU is added on the top of two streams to fuse the audio stream and the visual stream. The parameters of sync block and audio-visual BGRU are initialized by training for 5 epochs, keeping the weights of single stream fixed. The final end-to-end training for the whole network uses the Adam optimizer with an initial learning rate of 0.0001 and a mini-batch of 18.

3.3. Experimental Results

Different visual models are compared to verify the effectiveness of our modified graph convolution for lip reading. ResNet-34 & pseudo-image regards the nodes and frames of the lip graph sequences as the height and width of image and use a ResNet-34 to model the lip, while graph branch regards the lip key points as a graph. All the models use the same estimated 2D location of 20 lip key points as input to perform fair comparison. The results are presented in Table 1. Compared with ResNet-34 & pseudo-image, the GCN-based methods outperform the ResNet-based method which shows that our method can make good use of the relationship between the lip key points to model the lip shape. Among the multiple ways of partitioning strategy, our proposed implementation (lip graph) achieves the best result which indicates that the lip graph can boost the ability of representing temporal dynamics and capture more discriminative features.

Table 1: *Classification accuracy (%) of different visual model on "LRW Lip".*

Method	Accuracy
ResNet-34 & pseudo-image	43.66
Graph branch & uniform graph	47.95
Graph branch & shape graph	48.84
Graph branch & lip graph	**49.31**

To verify the effectiveness of the hybrid visual stream, we compare it with the optical flow method which is usually used to provide complementary information. The optical flow uses the same architecture as the image branch, and uses the optical flow as input. As the results shown in Table 2, combining image branch and graph branch (lip graph) leads to 0.96% increase, better than optical flow (0.74%). These results indicate that the graph branch can provide rich complementary information and the addition of graph branch to image branch can better learn the corresponding visual information of the speech signal.

Table 2: *Classification accuracy (%) of different visual model on LRW dataset.*

Method	Accuracy
Image branch [3]	83.29
Optical flow	76.55
Image branch + Optical flow	84.03
Image branch + Graph branch (proposed)	**84.25**

In order to investigate the robustness of different audio-visual fusion approaches to acoustic noise, the comparison experiment is carried out in the environments with different acoustic noise levels. The acoustic signal is corrupted by additive babble noise from the NOISEX database [30]. The signal to noise ratio (SNR) ranges from -5dB to 20dB.

We test different strategies to achieve audio-visual synchronization. (1) the unidirectional synchronization strategy uses unidirectional sync block which allows one modality to learn from another modality. Specifically, **AV-A2V** allows the visual modality to learn from audio modality, while **AV-V2A** allows the audio modality to learn from visual modality. (2) Bidirectional synchronization strategy (**AV-Bi**) uses bidirectional sync block which allows bidirectional synchronization and information interaction. The results are presented in Table 3. AV-A2V, AV-V2A and AV-Bi outperform the baseline model, and the improvement proves that the proposed sync block can well solve the asynchronous between audio and visual streams and learn the correlation. AV-Bi achieves best performance compared with other methods. The great improvement especially in strong noise environment indicates that even if one modality is heavily corrupted, the bidirectional information interaction can still explore mutual relations between the two modalities and reduce the dependence of system on a single modality.

Table 3: *Classification accuracy (%) of the audio-only (A) and audio-visual (AV) models on LRW dataset.*

SNR(dB)	-5	0	5	10	15	20	clean
A [3]	72.39	90.28	95.26	96.92	97.47	97.68	97.78
AV [3]	86.24	95.87	97.26	97.57	98.18	98.12	98.39
AV-A2V	91.18	95.72	97.37	97.99	98.22	98.27	98.27
AV-V2A	91.75	95.86	97.27	97.86	98.05	98.17	98.19
AV-Bi	**92.73**	**96.62**	**97.77**	**98.24**	**98.41**	**98.51**	**98.49**

4. Conclusion

In this work, we propose a lip graph assisted AVSR method using bidirectional synchronous fusion. First, a graph branch is proposed to extract additional shape-based features, and then combined with the image branch to extract more discriminative visual features. Second, an attention-based bidirectional sync block is proposed to achieve more reliable audio and visual synchronization and boost the ability to explore the correlation between the two modalities. Experimental results on the largest publicly available database show that our method achieves significantly improvements compared to the baseline method.

5. Acknowledgement

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

6. References

[1] T. Saitoh, K. Morishita, and R. Konishi, "Analysis of efficient lip reading method for various languages," in *International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," 2011.

[3] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6548–6552.

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

[5] L. Cappelletta and N. Harte, "Phoneme-to-viseme mapping for visual speech recognition." in *International Conference on Pattern Recognition Applizations and Methods (ICPRAM)*, 2012, pp. 322–329.

[6] X. Hong, H. Yao, Y. Wan, and R. Chen, "A pca based visual dct feature extraction method for lip-reading," in *International Conference on Intelligent Information Hiding and Multimedia*, 2006, pp. 321–326.

[7] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with lstms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2592–2596.

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

[9] J. Wei, F. Yang, J. Zhang, R. Yu, M. Yu, and J. Wang, "Three-dimensional joint geometric-physiologic feature for lip-reading," 2018, pp. 1007–1012.

[10] 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.

[11] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[12] J. Wang, J. Zhang, K. Honda, J. Wei, and J. Dang, "Audio-visual speech recognition integrating 3d lip information obtained from the kinect," *Multimedia Systems*, vol. 22, no. 3, pp. 315–323, 2016.

[13] F. Tao and C. Busso, "Lipreading approach for isolated digits recognition under whisper and neutral speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

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

[15] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3444–3453.

[16] G. Galatas, G. Potamianos, A. Papangelis, and F. Makedon, "Audio visual speech recognition in noisy visual environments," in *International Conference on PErvasive Technologies Related to Assistive Environments*, 2011, pp. 1–4.

[17] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, 2016.

[18] R. Ding, C. Pang, and H. Liu, "Audio-visual keyword spotting based on multidimensional convolutional neural network," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4138–4142.

[19] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015.

[20] T. J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 1082–1089, 2006.

[21] C. Bregler and Y. Konig, "" eigenlips" for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 1994, pp. II–669.

[22] F. Tao and C. Busso, "Aligning audiovisual features for audiovisual speech recognition," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[23] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *ACM International Conference on Multimodal Interaction*, 2018, pp. 111–115.

[24] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI conference on artificial intelligence*, 2018.

[25] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1227–1236.

[26] Y. Cai, L. Ge, J. Liu, J. Cai, T.-J. Cham, J. Yuan, and N. M. Thalmann, "Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks," in *IEEE International Conference on Computer Vision*, 2019, pp. 2272–2281.

[27] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE conference on computer vision and pattern recognition*, 2014, pp. 1867–1874.

[28] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

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

[30] A. Varga and H. J. 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.