INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Automatic Lip-reading with Hierarchical Pyramidal Convolution and Self-Attention for Image Sequences with No Word Boundaries

Hang Chen[1], *Jun Du*[1,*], *Yu Hu*[1], *Li-Rong Dai*[1], *Bao-Cai Yin*[2], *and Chin-Hui Lee*[3]

[1] University of Science and Technology of China, Hefei 230027, China
[2] iFlytek, Hefei 230088, China
[3] Georgia Institute of Technology, Atlanta, GA. 30332-0250, USA

ch199703@mail.ustc.edu.cn, ✉jundu@ustc.edu.cn, yuhu@iflytek.com
lrdai@ustc.edu.cn, bcyin@iflytek.com, chl@ece.gatech.edu

Abstract

In this paper, we propose a novel deep learning architecture for improving word-level lip-reading. We first incorporate multi-scale processing into spatial feature extraction for lip-reading using hierarchical pyramidal convolution (HPConv) and self-attention. Specifically, HPConv is proposed to replace the conventional convolution features, leading to an improvement over the model's ability to discover fine-grained lip movements. Next to deal with fixed-length image sequences representing words in a given database, a self-attention mechanism is proposed to integrate local information in all lip frames without assuming known word boundaries, so that our deep models automatically utilize key feature in relevant frames of a given word. Experiments on the Lip Reading in the Wild corpus show that our proposed architecture achieves an accuracy of 86.83%, yielding a relative error rate reduction of about 10% from that obtained with a state-of-the-art scheme of averaging frame scores for information fusion. A detailed analysis of the experimental results also confirms that weights learned from self-attention tend to be zero at both sides of an image sequence and focus non-zero weights in the middle part of a given word.

Index Terms: visual speech recognition, lip-reading, multi-scale convolution, self-attention

1. Introduction

Automatic lip-reading, also known as visual speech recognition, aims at recognizing the speech content only based on visual information, especially the lip movements that are composed of a sequence of basic visual units also called visemes [1]. Lip-reading is a challenging task for both human and machine, due to the ambiguity introduced by the one-to-many mapping [2] between visemes and phonemes. Nonetheless a robust lip-reading system has a broad range of applications when the audio data is unavailable, such as silent speech control system [3], assisting audio-based speech recognition in noisy environments [4], and biometric authentication [5].

Conventional approaches to automatic lip-reading (e.g., [6, 7]) are usually consisted of a spatial feature extractor, such as discrete cosine transform [8, 9, 10] of the lip Regions of Interest (RoIs), and followed by a sequence modeling scheme, such as a hidden Markov model [11, 12, 13, 14], to capture the temporal dynamics of lip movements. Recently, automatic lip-reading has been significantly improved due to advances in two aspects, namely: (1) a use of deep neural network models [15, 16, 17, 18, 19, 20], and (2) an availability of a large scale data set for training [21, 22, 23, 24, 25]. Most deep-learning-based models usually consist of a frontend and a backend, which are similar to the feature extractor and the sequential models in the conventional approaches, respectively. However, using end-to-end training, the frontend module can often extract better image features than those obtained in conventional extractors, for the backend module to capture discriminative temporal information for improved lip-reading.

In this study, we focus on word-level lip-reading and adopt the Lip Reading in the Wild (LRW) [21] corpus for our experiments. LRW is the first and the largest publicly available data set with word-level labels in English. It consists of fix-length segments (1.16 seconds of 29 image frames at 25 frames per second with no specified word boundaries) extracted from BBC news and talk shows. There are more than $1,000$ speakers and 500 target words, which is much higher than existing lip-reading databases used for word recognition. A total of $538,766$ segments in this set are split into three subsets consisting of $488,766$, $25,000$ and $25,000$ samples for training, validation and testing, respectively. This task is quite challenging due to a large variation in head poses and lighting illuminations for the videos used in the LRW corpus.

Table 1: *Review of the existing models on LRW.* **Acc.**: *Word accuracy.*

Method	Frontend	Backend	Consensus	Acc.
[21]	VGG-M	-	-	61.10
[23]	VGG-M	LSTM	Average	76.20
[26]	3D Conv+ResNet-34	BLSTM	Average	83.00
[27]	3D Conv+ResNet-34	BGRU	Average	83.40
[28]	3D Conv+ResNet-18	BLSTM	Average	84.30
[29]	ResNet-34+3D DenseNet	Conv-BLSTM	Average	83.30
[30]	I3D *2	BLSTM	Average	84.07
[31]	3D Conv+P3D-ResNet-50	BLSTM	Average	84.48
[32]	3D Conv+ResNet-18	MS-TCN	Average	85.30
[33]	3D Conv+ResNet-34+ST-GCN	BGRU	Average	84.25

Since LRW was released, numerous novel models have been proposed. In addition to the frontend and backend modules mentioned above, the word-level lip-reading models usually contain an additional consensus module which merges scores at a frame level in all time steps to obtain an overall score at the sequence level to predict the recognized word. As shown in Table 1, we list existing systems with their respective frontend, backend, consensus and word accuracy when evaluated on the LRW data set. We observe that all models have kernels with a single spatial size in the frontend and average along the temporal dimension in consensus decisions.

The state-of-the-art performance on LRW [32] was achieved by a system consisting of a 3D Convolutional layer followed by a 18-layer residual network (ResNet-18) [34] as

[*] corresponding author

http://dx.doi.org/10.21437/Interspeech.2021-723

a frontend and a multi-scale temporal convolutional network (MS-TCN) backend. The final score vector at the sequence level for the 500 words to be predicted was obtained by averaging the output of the backend along the time dimension. It serves as our baseline for performance comparisons. In this paper, we improve the baseline frontend by proposing a novel hierarchical pyramidal convolution (HPConv), that is capable of processing the input with multiple spatial resolution, to replace the standard 2D-convolution in ResNet-18. Moreover, our proposed backend utilizes a self-attention alternative to the baseline averaging consensus and focuses on the frames of the relevant video segment that corresponds to the spoken word, to achieve an improved classification accuracy. To the best of our knowledge, this is the first work incorporating multi-scale processing into the frontend and adopting non-uniform self-attention weights in consensus for word-level lip-reading.

2. Our Proposed Approach

As shown in Fig. 1, our system can be divided into three main parts: frontend, backend and consensus modules. The frontend module takes a gray scale sequence of lip RoIs $X \in \mathbb{R}^{T \times H \times W}$ as input to produce a feature matrix $F_2 \in \mathbb{R}^{T \times C_1}$, where T denotes the temporal dimension and H, W represent the height and width of the gray scale lip image, respectively. Then the spatial knowledge F_2 is summarized by applying average pooling over the spatial dimensionality. The backend module is next employed to model the temporal dynamics. The output score matrix $F_3 \in \mathbb{R}^{T \times C_2}$ is then passed through the consensus module to merge temporal information. Finally, the posterior probability of each word P is predicted by the ensuing full connection and SoftMax layers.

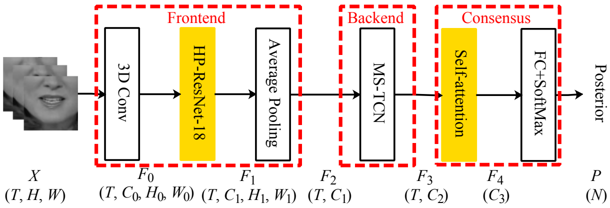

Figure 1: *Block diagram of the proposed system. The input to each module and its corresponding dimensionality are also shown below each arrow. Our contributions are the frontend and consensus modules highlighted in yellow.*

The MS-TCN module in the baseline [32] is kept as our backend and we change the frontend by replacing the standard convolution in the ResNet-18 with our proposed hierarchical pyramidal convolution and modify the consensus from averaging to our proposed self-attention based consensus.

2.1. Hierarchical Pyramidal Convolution

The ResNet-18 in the frontend of the baseline uses the standard 2D-convolution to extract spatial feature maps. It contains only a single kernel type with a single spatial size (K_1, K_1) (in the case of square kernels). Since all kernels have the same spatial resolution, the extracted feature maps only contain fixed-sized spatial context information.

We analyze some errors produced by the baseline, and find that the classification accuracy of a word is improved with an increasing number of visemes contained in the word, i.e., the model performs poorly on words with little visemic content. This is reasonable because words with fewer visemes often imply fewer lip movements in the corresponding image segments, making it challenging for the model to correctly classify these samples. Accordingly, we propose using different spatial-sized kernels to extract complementary context information, enabling the frontend to obtain discriminative feature maps. These enhanced features help boost the modeling capability for fine-grained lip movements and improve the classification accuracy for words with only a few visemes.

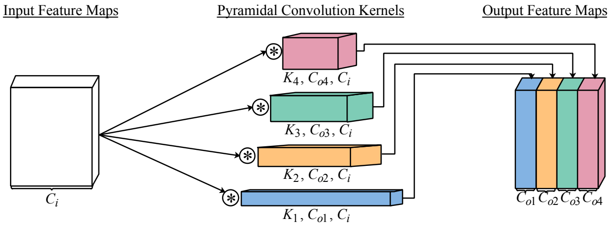

Figure 2: *Illustration of PyConv [35] with ⊛ denoting the convolution operation with hyperparameters given in kernelSize, outChannels, inChannels format.* $K_4 > K_3 > K_2 > K_1$ *and* $C_o = C_{o1} + C_{o2} + C_{o3} + C_{o4}$.

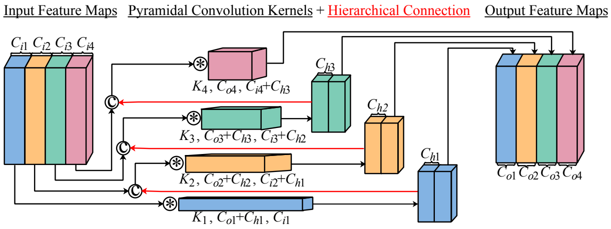
Input Feature Maps
Pyramidal Convolution Kernels + Hierarchical Connection
Output Feature Maps
C_{i1} C_{i2} C_{i3} C_{i4}
$K_4, C_{o4}, C_{i4}+C_{h3}$
$K_3, C_{o3}+C_{h3}, C_{i3}+C_{h2}$
$K_2, C_{o2}+C_{h2}, C_{i2}+C_{h1}$
$K_1, C_{o1}+C_{h1}, C_{i1}$
C_{h3}
C_{h2}
C_{h1}
C_{o1} C_{o2} C_{o3} C_{o4}

Figure 3: *Illustration of the proposed HPConv with hierarchical connections between adjacent layers of the pyramid (red lines in figure). ©denotes the concatenation over channel dimensions with* $C_i = C_{i1} + C_{i2} + C_{i3} + C_{i4}$.

To validate the effectiveness of multi-scale processing, we first incorporate pyramidal convolution (PyConv) [35] as illustrated in Fig. 2 into the frontend. It contains a pyramid with n levels of different types of kernels (we set $n = 4$ as default in our experiments, which is consistent with the figure). The kernels at each level contain an increasing spatial size from the bottom of the pyramid to the top (we set $K_{1,2,3,4} = 3, 5, 7, 9$ as default in our experiments). The kernels with a smaller spatial size can focus on extracting feature maps with local context information, while the larger-sized kernels can provide more global context information. The model can explore a good combination of different kernel types through learning. For every basic block of ResNet-18, we replace the second standard convolution layer with PyConv. We call this modification as the Pyramidal ResNet-18 (Py-ResNet-18).

Based on PyConv, we propose hierarchical pyramidal convolution (HPConv) as illustrated in Fig. 3. The novelty here is that we establish a hierarchical connection between adjacent layers of the pyramid (red lines in Fig. 3). As mentioned above, the local and global feature maps in PyConv are extracted from the input feature maps. As with the hierarchical connections, local feature maps are used as parts of the output and also as an input for global feature extraction. This bottom-up information aggregation can further improve the classification performance of the model, especially for words with only a few visemes. For every basic block of ResNet-18, we replace the second convolution layer with HPConv and call this modification Hierarchical Pyramidal ResNet-18 (HP-ResNet-18).

2.2. Self-attention Based Consensus

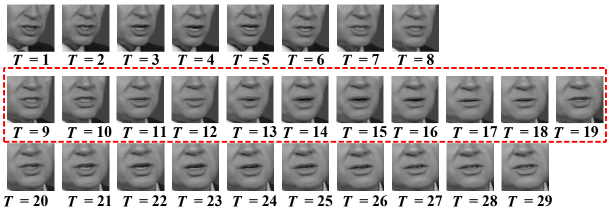

Figure 4: *An example of a video sample segment annotated as "ABOUT". Only frames at the time steps $T = 9 \sim 19$ are related to the word "ABOUT".*

The most popular consensus method currently being used is to average over scores produced at all time steps, as shown in all the systems in Table 1. Given the feature maps at the frame level, $F_3 \in \mathbb{R}^{T \times C_2}$, the final score vector at the sequence level, $F_4 \in \mathbb{R}^{C_3}$, is calculated as follows:

$$F_4 = \frac{\sum_{t=0}^{T-1} F_{3,t}}{T}. \tag{1}$$

The averaging based consensus assumes that every frame provides an equal contribution to the final decision, which is often not a good way for the LRW data being used here. As shown in Fig. 4, the video sample annotated as "ABOUT" includes 29 frames in total, but only frames at time steps $T = 9 \sim 19$, highlighted in the middle red dashed box, are related to the word "ABOUT". The accurate word boundaries of individual words are often difficult to locate in labeling. We therefore propose a self-attention [36] based consensus mechanism to ensure the model pays more attention to the frames which are more relevant to the annotated word, but less to other irrelevant frames. The proposed non-uniform self-attention based consensus can be expressed as:

$$(Q_n, K_n, V_n) = (F_3 W_n^Q, F_3 W_n^K, F_3 W_n^V) \tag{2}$$

$$head_n = A_n^{\mathsf{T}} V_n = \text{SoftMax}\left(\frac{\sum_{t=0}^{T-1} Q_{n,t} K_n^{\mathsf{T}}}{T\sqrt{d_k}}\right) V_n \tag{3}$$

$$F_4 = W^O \text{Concat}(head_0, \cdots, head_{N-1}) + \frac{\sum_{t=0}^{T-1} F_{3,t}}{T} \tag{4}$$

where for the n-th attention head, $W_n^Q \in \mathbb{R}^{C_2 \times d_k}$, $W_n^K \in \mathbb{R}^{C_2 \times d_k}$, $W_n^V \in \mathbb{R}^{C_2 \times d_v}$ and $W^O \in \mathbb{R}^{N d_v \times C_3}$ are the projection matrices, and $A_n \in \mathbb{R}^T$ is the attention weight vector. Here we employ $N = 8$ and $d_k = d_v = 64$, same as those in [36].

3. Experiments

In this section, we compare our proposed framework with the baseline system that already achieved the best word accuracy on the LRW task [32]. We pre-process each fixed-length video segment and train all models following the same procedures used in the baseline. The readers are referred to [32] for more detail. To better understand the two proposed frontend and consensus approaches as highlighted in Fig. 1, we also provide an in-depth analysis of the experimental results to illustrate the contributions of HPConv and self-attention.

Table 2: *A comparison of word accuracies (in %) of different systems. 3D Conv in the frontend is omitted for simplicity.* ***Acc.:*** *Word accuracy.*

System	Frontend	Consensus	Boundary	Acc.
Baseline	ResNet-18	Average	F	85.30
N1	ResNet-18	Self-attention	F	86.47
N2	Py-ResNet-18	Average	F	85.88
N3	HP-ResNet-18	Average	F	86.45
N4	HP-ResNet-18	Self-attention	F	**86.83**
N5	ResNet-18	Average	T	88.60
N6	ResNet-18	Self-attention	T	88.59
N7	HP-ResNet-18	Average	T	89.38
N8	HP-ResNet-18	Self-attention	T	89.38

Table 2 lists the results of all systems. Compared to the baseline model, our proposed HPConv in N3 performs better than PyConv in N2, and our proposed self-attention in N1 does better than "Average" in the baseline. The overall system (denoted as N4) achieves an accuracy of 86.83%, attaining the best performance on LRW. With known word boundaries in N5-N8, we can see N6, N7 and N8 are all better than N5.

3.1. Analysis on Hierarchical Pyramidal Convolution

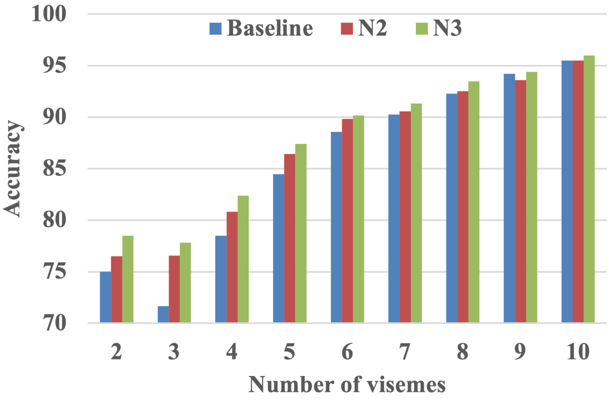

Figure 5: *A comparison of the accuracy among Baseline, N2 and N3 on different categories with the same number of visemes in the annotated word.*

To verify the effectiveness of our proposed HPConv frontend, we compare the result of the system using only HP-ResNet-18 (denoted as N3 in Table 2) with the result obtained with Py-ResNet-18 (denoted as N2), both performing multi-scale feature extraction. In comparison with the baseline using ResNet-18, applying multi-scale kernels enhances the model classification performances over the baseline without multi-scale spatial feature extraction. Moreover, our proposed HPConv (N3) benefits more from it than PyConv (N2).

We further analyze error samples obtained with different frontend features. Based on the number of visemes in the annotated words, we divide the whole test set into 9 categories and plot the corresponding accuracies in Fig. 5. We can see that both N2 and N3 perform better than the baseline in almost all cases of viseme lengths. The improvements are more significant for words with smaller number of visemss than those with more visemes. Moreover, our proposed HPConv introduces hierarchical connections from local to global information, which further improves classification accuracies over PyConv on words with few visemes.

3.2. Analysis on Self-attention Based Consensus

One of the most significant differences between our proposed framework and previous methods is the proposed self-attention based consensus. It ensures that the model pays more attention on the relevant frames during classification. Therefore in Table 2, the result of the system using only self-attention (denoted as N1) improves the classification performance over the average based consensus in Baseline. To further analyze why our proposed self-attention can outperform the conventional average based consensus, we retrain the models used in Baseline, N1, N3 and N4 using the word boundary information provided by [21] and obtain four improved systems denoted as N5, N6, N7 and N8, respectively. The major difference here is to apply average or self-attention based consensus only on the frames which are related to the annotated words. We can observe that in the situation of using manual word boundaries, the accuracies obtained are almost the same for both average and self-attention based consensus when comparing N5 versus N6 and N7 versus N8. It is conjectured that the learned attention weights act like "soft word boundaries". Although not exact, they function in a similar way to manual word boundaries.

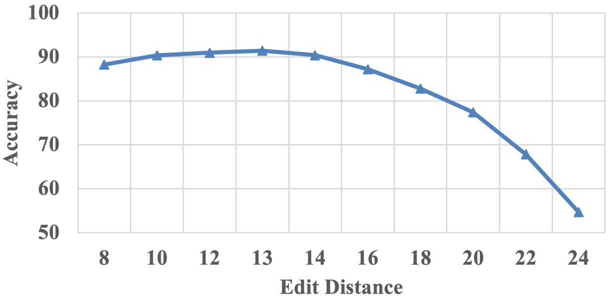

Figure 6: *Classification accuracy of the baseline model on different edit distances between the manual and the learned word boundaries.*

To verify our assumption, we categorize all test samples by the edit distance [37] between the manual word boundary vector, $B_{man} = [0, \cdots, 0, 1, \cdots, 1, 0, \cdots, 0]^\top \in \mathbb{R}^T$ and the word boundary vector with averaging, $B_{avg} = [1, \cdots, 1]^\top \in \mathbb{R}^T$, and the learned vector with self-attention, $B_{att} = \mathrm{u}(\sum_{n=0}^{N-1} A_n/N - \alpha)$, where $\mathrm{u}(\cdot)$ is the unit step function using a threshold constant $\alpha = 0.01$. First, we plot the classification accuracies of the baseline model as a function of the edit distance in Fig. 6. We observe that the accuracy tends to decline with increasing edit distances between the learned and manual word boundaries. Next in Fig. 7, we plot the number of samples obtained with the baseline and N1 systems corresponding to each edit distance. Clearly, the self-attention based consensus can learn word boundaries better and result in much smaller edit distances than average based consensus. These two observations can well explain the effectiveness of the self-attention based consensus mechanism.

The weights learned with the proposed self-attention consensus for the example in Fig. 4 are plotted in Fig. 8. We can find that even though each head has a different focus, the attention weights on all irrelevant frames are quite small (mostly equal to 0), which helps the model to ignore noisy information for better classification performance.

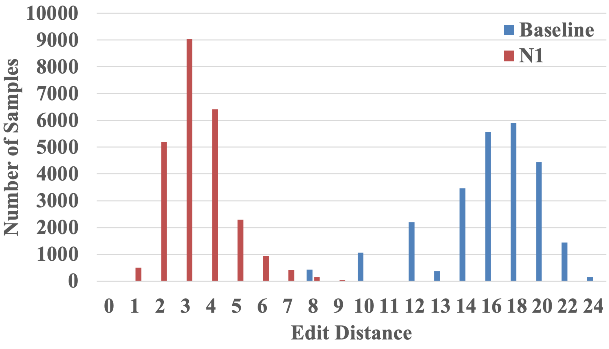

Figure 7: *A comparison of the number of samples among Baseline and N1 on different edit distances between the manual and the learned word boundaries.*

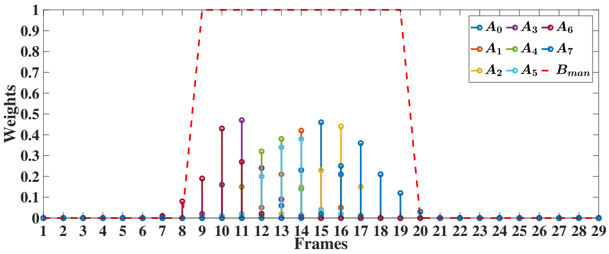

Figure 8: *The weights of each frame learned with self-attention for the example in Fig. 4. Stem-and-leaf plots of different colors show the learned weights of different attention heads for each frame. The red dashed line denotes the manual word boundary with all weights equal to 1.*

4. Conclusion

We propose hierarchical pyramidal convolution and self-attention based consensus to replace the standard convolution and the average based consensus commonly used in state-of-the-art lip-reading systems. Extensive experiments and analyses empirically validate that our proposed HPConv improves our model's utilization of slight lip movements and the self-attention based consensus ensures the model pays more attention to the relevant image frames. Together, our system achieves the best word accuracy on the LRW lip-reading task.

In the future, it would be interesting to explore more effective network structures to simultaneously utilize spatial and temporal context information with multi-scale processing. And we also further study how to improve the accuracy of learned word boundaries.

5. Acknowledgements

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDC08050200.

6. References

[1] S. L. Taylor, M. Mahlerc, B.-J. Theobald, and I. Matthews, “Dynamic units of visual speech,” in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2012, p. 275–284.

[2] H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: the good, the bad, and the ugly,” *Speech Communication*, vol. 95, p. 40–67, Dec 2017.

[3] K. Sun, C. Yu, W.-N. Shi, L. Liu, and Y.-C. 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, p. 581–593.

[4] 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*, pp. 1–1, 2018.

[5] Y. M. Assael, B. Shillingford, S. Whiteson, and N. Freitas, “Lipnet: End-to-end sentence-level lipreading,” 2016.

[6] Z.-H. Zhou, G.-Y. Zhao, X.-P. Hong, and M. Pietikäinen, “A review of recent advances in visual speech decoding,” *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.

[7] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, “Audio-visual automatic speech recognition: An overview,” *Issues in visual and audio-visual speech processing*, vol. 22, p. 23, 2004.

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

[9] 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.

[10] G. Potamianos, C. Neti, G. Iyengar, A. W. Senior, and A. Verma, “A cascade visual front end for speaker independent automatic speechreading,” *International Journal of Speech Technology*, vol. 4, no. 3-4, pp. 193–208, 2001.

[11] V. Estellers, M. Gurban, and J. Thiran, “On dynamic stream weighting for audio-visual speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1145–1157, 2012.

[12] X. Shao and J. Barker, “Stream weight estimation for multistream audio-visual speech recognition in a multispeaker environment,” *Speech Communication*, vol. 50, no. 4, p. 337–353, 2008.

[13] 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.

[14] D. Stewart, R. Seymour, A. Pass, and J. Ming, “Robust audio-visual speech recognition under noisy audio-video conditions,” *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 175–184, 2014.

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

[16] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Lipreading using convolutional neural network,” in *fifteenth annual conference of the international speech communication association*, 2014.

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

[18] 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)*, 2016, pp. 2722–2726.

[19] S. Petridis, J. Shen, D. Cetin, and M. Pantic, “Visual-only recognition of normal, whispered and silent speech,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6219–6223.

[20] S. Petridis, Y.-J. Wang, Z.-W. Li, and M. Pantic, “End-to-end multi-view lipreading,” in *British Machine Vision Conference, BMVC 2017*, 2017.

[21] J.-S. Chung and A. Zisserman, “Lip reading in the wild,” in *Asian Conference on Computer Vision*, 2016.

[22] S. Yang, Y.-H. Zhang, D.-L. Feng, M.-M. Yang, C.-H. Wang, J.-Y. Xiao, K.-Y. Long, S.-G. Shan, and X.-L. 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*, 2019, pp. 1–8.

[23] 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*, 2017.

[24] 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.

[25] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, “An audio-visual corpus for multimodal automatic speech recognition,” *Journal of Intelligent Information Systems*, vol. 49, 10 2017.

[26] T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with lstms for lipreading,” in *Proc. Interspeech 2017*, 2017, pp. 3652–3656.

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

[28] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, “Pushing the boundaries of audiovisual word recognition using residual networks and lstms,” *Computer Vision and Image Understanding*, vol. 176-177, pp. 22–32, 2018.

[29] C.-H. Wang, “Multi-grained spatio-temporal modeling for lip-reading,” in *30th British Machine Vision Conference*, 2019, p. 276.

[30] X.-S. Weng and K. Kitani, “Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading,” in *30th British Machine Vision Conference*, 2019.

[31] B. Xu, C. Lu, Y.-D. Guo, and J. Wang, “Discriminative multi-modality speech recognition,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 433–14 442.

[32] B. Martinez, P.-C. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6319–6323.

[33] H. Liu, Z. Chen, and B. Yang, “Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion,” in *Proc. Interspeech 2020*, 2020, pp. 3520–3524.

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

[35] I. C. Duta, L. Liu, F. Zhu, and L. Shao, “Pyramidal convolution: Rethinking convolutional neural networks for visual recognition,” *arXiv preprint arXiv:2006.11538*, 2020.

[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, pp. 5998–6008, 2017.

[37] R. W. Hamming, “Error detecting and error correcting codes,” *The Bell system technical journal*, vol. 29, no. 2, pp. 147–160, 1950.