Deep Audio-visual System for Closed-set Word-level Speech Recognition

Yougen Yuan*
Wei Tang*
Minhao Fan*
Yue Cao*
School of Computer Science, Northwestern Polytechnical University, Xi'an, China

Peng Zhang†
Lei Xie†
zh0036ng@nwpu.edu.cn
lxie@nwpu-aslp.org
School of Computer Science, Northwestern Polytechnical University, Xi'an, China

ABSTRACT

Audio-visual understanding is usually challenged by the complementary gap between audio and visual informative bridging. Motivated by the recent audio-visual studies, a closed-set word-level speech recognition scheme is proposed for the Mandarin Audio-Visual Speech Recognition (MAVSR) Challenge in this study. To achieve respective audio and visual encoder initialization more effectively, a 3-dimensional convolutional neural network (CNN) and an attention-based bi-directional long short-term memory (Bi-LSTM) network are trained. With two fully connected layers in addition to the concatenated encoder outputs for the audio-visual joint training, the proposed scheme won the first place with a relative word accuracy improvement of 7.9% over the solitary audio system. Experiments on LRW-1000 dataset have substantially demonstrated that the proposed joint training scheme by audio-visual incorporation is capable of enhancing the recognition performance of relatively short duration samples, unveiling the multi-modal complementarity.

CCS CONCEPTS

• **Information systems** → *Multimedia and multimodal retrieval*; • **Human-centered computing** → *Natural language interfaces*; *HCI theory, concepts and models*; • **Computing methodologies** → *Speech recognition*; *Computer vision tasks*; *Neural networks*.

KEYWORDS

Audio-visual, convolutional neural network, long short-term memory, multi-model

ACM Reference Format:
Yougen Yuan, Wei Tang, Minhao Fan, Yue Cao, Peng Zhang, and Lei Xie. 2019. Deep Audio-visual System for Closed-set Word-level Speech Recognition. In *2019 International Conference on Multimodal Interaction (ICMI '19), October 14–18, 2019, Suzhou, China*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3340555.3356102

1 INTRODUCTION

Audio-visual speech recognition [4, 23] has received increasing attention in recent years, aiming at exploring the complementarity between visual and acoustic information. Beyond solitary acoustic modeling, recurrent neural networks [5, 6] are more capable of capturing temporal dependency for variable-length audio sequence data. They have been successfully applied in speech applications [13, 24, 30] and also naturally introduced into audio-visual speech recognition [18, 20]. Other studies of visual and audio systems often apply joint training via both visual and audio data [18, 20, 21]. In [18, 20], fully connected layers together with bi-directional long short-term memory (bi-LSTM) or bi-directional GRU (bi-GRU) were used for classification. In a similar way [21] , a connectionist temporal classification (CTC) and attention based joint training approach was followed for decoding.

More commonly, the strategy of joint-classification has been introduced to visual speech recognition (also known as lipreading) via feature extraction from both visual and audio data [1, 14, 19, 22]. Deep neural networks (DNNs) [2, 16, 26, 28] or end-to-end architectures [9, 11, 15, 25] were used for such purpose. With their success in action recognition [27], 3-dimensional convolutional neural network (3D-CNN) based models (like LipNet [3]

*The first four authors contributed equally to this research.
†Corresponding author.

ICMI '19, October 14–18, 2019, Suzhou, China

ACM ISBN 978-1-4503-6860-5/19/10…$15.00
https://doi.org/10.1145/3340555.3356102

and DenseNet in 3D version [29]) were widely used in lipreading. This has become a motivation for us to train a 3D-CNN by transforming image sequence into spatial-temporal features, followed by gated recurrent units (GRUs) for classification. Unfortunately, although decent performance has been achieved by these methods, it is still unclear to what degree the visual information can complement to the counterpart audio information. This has become an unexplained complementary gap of informative bridging due to training on different datasets.

To address this issue more clearly, a two-step joint training strategy is introduced in this study. More specifically, we train an audio system to recognize the words via acoustic data and a visual system to capture speech-associated mouth movements for visual speech recognition, separately. Then we set both trained systems as the initialized model in another network for joint training. With the LRW-1000 dataset and a benchmark for real-word speech recognition provided by the Mandarin Audio-Visual Speech Recognition (MAVSR) Challenge [1], the proposed audio-visual system takes the trained 3D-CNN visual model and the attention-based bi-LSTM acoustic model for joint training. To the best of our knowledge, this is the first audio-visual system in a naturally-distributed large-scale dataset for closed-set word-level speech recognition. Convincing experiments on the LRW-1000 dataset have shown superior performance of our proposed audio-visual system, and the audio-visual complementarity is studied across different sample duration.

2 METHODS

Visual system

The task of lipreading is to transfer the mouth movement of image sequences into words or sentences through learning linguistic information. In recent years, 3D-CNN models have been successfully applied in lipreading [3] with superior performance. It also becomes the foundation to build our visual system.

Figure 1 shows the diagram of the training visual system for closed-set word-level speech recognition. The model of the visual system consists of one spatiotemporal convolution block, followed by several dense blocks, transition blocks and bi-GRU layers. The spatiotemporal convolution block contains a convolutional layer with 3-dimensional (3D) kernels of $5\times7\times7$ (time/width/height), a batch normalization (BN) with rectified linear units (ReLU) layer, and a 3D spatiotemporal max-pooling layer with the size of $1 \times 3 \times 3$. For a batch of image frame sequences with $B/Tv/C/W/H$ size (batch/image frame/channel/width/height), the spatiotemporal convolution block is able to capture the short-term dynamics of the mouth region.

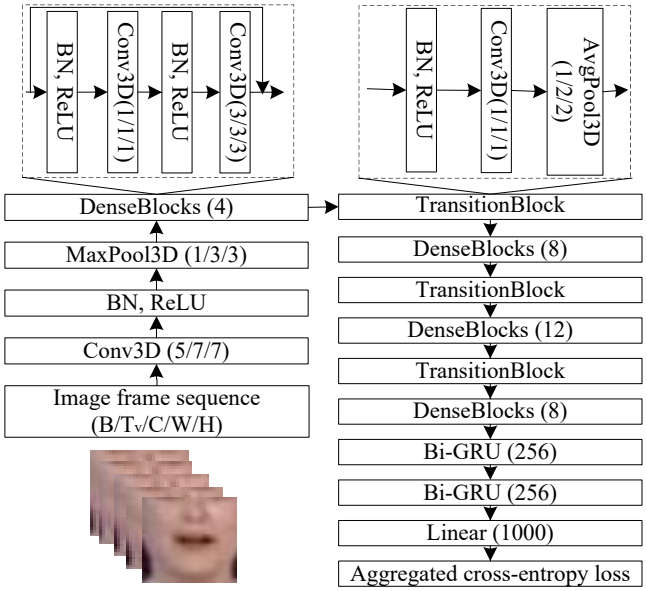

Figure 1: The diagram of training visual system for closed-set word-level speech recognition.

Each dense block contains a BN+ReLU layer, a 3D convolutional layer with $1 \times 1 \times 1$ size, followed by a BN+ReLU layer, a 3D convolutional layer with $3 \times 3 \times 3$ size and a skip connection. We use $[4, 8, 12, 8]$ dense blocks, and a transition block is followed in the first three dense blocks. Each transition block contains a BN+ReLU layer, a 3D convolutional layer with $1 \times 1 \times 1$ size and a 3D average pooling layer with the size of $1 \times 2 \times 2$. The connections of dense blocks and transition blocks are similar to the ResNet [7], which aim to drop progressively the spatial dimensionality and capture rich motion information in compressed outputs.

As the bi-GRU network is more capable of capturing temporal dependency with variable-length sequential data in a fixed-dimensional space, we add two bi-GRU layers on the compressed outputs to learn the image sequence of motion information. Each bi-GRU layer consists of 256 hidden units per direction. Given an image sequence $[x_1, ..., x_t, ..., x_{T_v}]$, a fully connected layer is added to produce 1000-dimensional outputs $[f(x_1), ..., f(x_t), ..., f(x_{T_v})]$.

In training visual system, the word label y is repeated at every time step. Hence, an aggregated cross-entropy loss is performed on all time-steps to summarize the negative logarithm of word posteriors. It can be calculated

[1] http://vipl.ict.ac.cn/homepage/mavsr/index.html

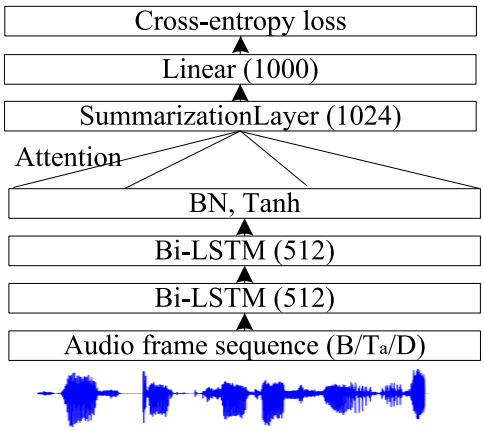

Figure 2: The diagram of training audio system for closed-set word-level speech recognition.

as follows:

$$L_{ace}(x, y) = -\sum_{t=1}^{t=T_v}\sum_{n=1}^{n=N}(f(x_t)_n * \log y_{tn}), \quad (1)$$

where N denotes number of classes in corresponding closed-set word-level speech recognition.

Audio system

DNN-based acoustic models have been extensively used to transcribe audio sequences into words or sentences [8]. Here we also build an audio system to predict words from audio sequences. Figure 2 shows the diagram of training the audio system based on bi-LSTM networks with attention mechanism.

In Figure 2, the audio system consists of two bi-LSTM layers, a BN with hyperbolic tangent (Tanh) layer, an attention layer, and a fully connected layer. Unlike the end-to-end visual system, acoustic feature extraction is performed before training the audio model. For a batch of audio frame sequences, mel-frequency cepstral coefficients (MFCCs) are extracted as neural input with $B/T_a/D$ size (batch/audio frame/dimension). In addition, since the MFCCs of one audio frame sequence is a 2D matrix, it is not necessary to use 3D-CNN models. Recent studies on speech recognition have also indicated that bi-LSTM network can bring superior performance. Hence we adopt a bi-LSTM network to capture temporal dependency from variable-length speech data.

Each bi-LSTM layer consists of 512 hidden units per direction. Current acoustic models often generate the outputs from the last time step in both directions, but this encoding method cannot reflect the attention degree to different phonemes in a speech sequence [5, 6]. With

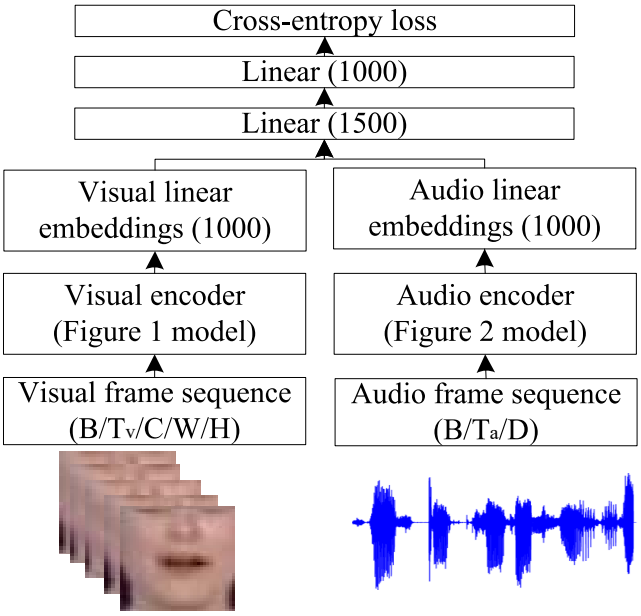

Figure 3: The proposed audio-visual system for closed-set word-level speech recognition.

the success of attention mechanism in sentence embedding [12], an attention layer is added to learn a sequence of weights for the whole bi-LSTM outputs. Given an audio sequence $[x_1, ..., x_t, ..., x_{T_a}]$, a BN with Tanh layer is added on two bi-LSTM layers to produce the outputs $[h(x_1), ..., h(x_t), ..., h(x_{T_a})]$, and the weight coefficients a for all the time steps can be calculated by:

$$a(x) = softmax(W_{s2}\tanh(W_{s1}h^T)), \quad (2)$$

where W_{s2} and W_{s1} are matrices with the size of $[d_a, 2 \times 512]$ and $[1, d_a]$, respectively. d_a represents the dimension of attention weights. With the weight coefficients a, the summarizing vector S is calculated by:

$$S(x) = \sum_{t=1}^{t=T_a} a_t h(x_t), \quad (3)$$

Then, a fully connected layer is added to produce 1000-dimensional outputs $f(S(x))$.

In training audio system, the word label y is a one-hot vector from 1000 classes in the closed-set word-level speech recognition, and thus a cross-entropy loss is performed to guide error back-propagation as follows:

$$L_{ce}(x, y) = -\sum_{n=1}^{n=N}(f(S(x))_n * \log y_n). \quad (4)$$

Audio-visual system

Figure 3 shows the proposed audio-visual system for the closed-set word-level speech recognition task. This

system aims to make use of visual and audio complementarity via joint training both the visual and audio models. For this purpose, first, the encoders of visual and audio are applied for extracting embeddings from the visual and audio frame sequence separately. Instead of randomly initializing network parameters, the pre-trained visual system (see Figure 1) and audio system (see Figure 2) are set to respectively initialize its visual and audio encoder. The outputs of both encoders are concatenated to be a 2000-dimensional embedding vector. In order to combine the information from both visual and audio systems, one fully connected layer with 1500 hidden units is added on top of the concatenated vector, and another fully connected layer is then used to output 1000-dimensional predictions.

For the whole system, the network parameters are fine-tuned throughout the audio-visual joint training phase. In addition, the same word label y and the cross-entropy loss in Equation 4 as in the audio system are used to guide the joint training.

3 EXPERIMENTS

Experimental setup

To evaluate the effectiveness of the proposed audio-visual system, experiments are conducted on the closed-set word-level speech recognition task from the MAVSR Challenge [2]. This task aims to use the audio and the corresponding mouth movement images to recognize a closed set of words by classification. The training data of the task takes the LRW-1000 dataset [29] that is a naturally-distributed large-scale benchmark for lipreading in the wild. The LRW-1000 dataset contains 1,000 classes with 718,018 samples from about 840 raw videos (57 hours). Each class corresponds to a Mandarin word composed of one or multiple Chinese characters. The minimum/average/maximum duration of all samples are about 0.01/0.3/2.25 seconds, respectively. For each sample, the corresponding images and audio files are provided. Each image sequence is sampled at 25 frames per second (fps) from the corresponding video and it has already been cropped to contain only the lip region with the size of 112×112. Each audio sequence has the sample rate of 16000 Hz with 16 bits single channel and it is also cropped with the actual word in the center with 200 milliseconds temporal context information. Another 125,884/68,896 samples are used in the task for validation/test, respectively.

The training of the proposed audio-visual system is implemented via two stages: firstly, the 3D-CNN visual system and the attention-based bi-LSTM audio system are trained independently, and then both trained systems are set as the initialized models, followed by two fully connected layers for the audio-visual joint training. In training the visual system, the max length of visual frames T_v is set to 30, and all the frame sequences are padded with zeros to that length. Data augmentation is performed on the visual frame sequence by applying random cropping and horizontal flips with a probability of 50%. An Adam optimizer [10] is applied for updating the weights with a mini-batch size of 32 and an initial learning rate of 0.0001. The 3D-CNN which gives the best results on the validation set within 5 epochs is set as the pre-trained visual model. In training the audio system, the max length of audio frames T_a is set to 170, and all the audio frame sequences are padded with zeros to that length. The dimensions D of MFCCs and the dimensions d_a of attention weights are set to 43 and 256, respectively. A mini-batch size of 1000 and a learning rate of 0.001 are used. The attention-based bi-LSTM network which gives the best results on the validation set within 30 epochs is used as the pre-trained audio model. During the joint training of audio-visual system, a mini-batch size of 32 and an initial learning rate of 0.0001 are also applied for fine-tuning the visual and audio sequential data. An early stop strategy with 50000 batches is used to obtain the best audio-visual model. The implementation of all systems is based on Pytorch [17]

Table 1: Word accuracy of different systems on closed-set word-level speech recognition in the MAVSR Challenge

System	input sequence	Word accuracy(%) validation	test
System 1	Visual	32.80	34.59
Visual	Visual	37.13	37.05
System 2	Visual	37.18	37.51
Audio	Audio	74.86	76.72
Audio-visual	Audio+visual	**80.49**	**82.78**

The performance of the closed-set word-level speech recognition task is evaluated by word accuracy, which corresponds to the percentage of correct word predictions in all samples. Higher word accuracy represents better performance.

Results

Table 1 summarizes the word accuracy of different systems on closed-set word-level speech recognition in the MAVSR Challenge. There are five systems trained using the LRW-1000 dataset, including two systems (System 1

[2] http://vipl.ict.ac.cn/homepage/mavsr/index.html#tasks

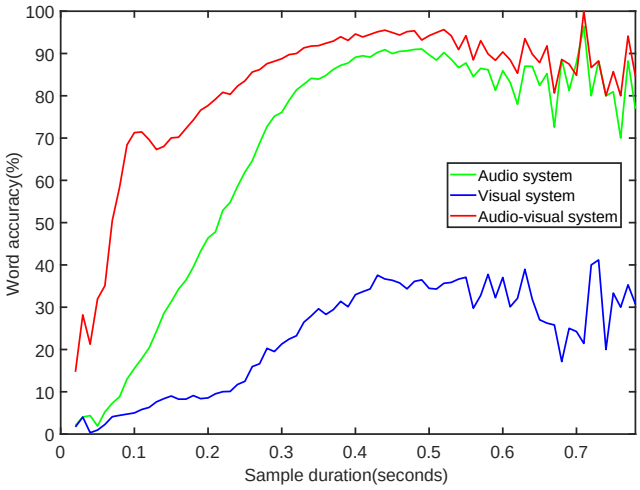

Figure 4: Word accuracy of our different systems across sample duration.

and System 2) submitted from other participants, our visual system, our audio system and our audio-visual system. Not suprisingly, we can find that the audio system has a much better performance than the visual systems. It improves the word accuracy from 37.51% (the best visual system) to 76.72% on the test set. The result indicates that audio speech can provide more discriminative information than visual speech. Hence the audio system has more capability to predict the linguistic information than the visual system as many visually confusing pronunciation units are acoustically separable.

More importantly, when joint training is applied to the audio and visual systems, the joint system holds the best performance over all the other submitted systems. In comparison to the audio system, the word accuracy of the audio-visual system on the test set is relatively improved by 7.9%. It also indicates that systems built with both audio and visual data are better than those built with solitary audio/visual data. This result is in line with previous works [1, 20].

Audio-visual complementarity

In order to further study the audio-visual complementarity, the word accuracy is further compared across sample duration. Short samples are more difficult to recognize by both humans and machines as the discriminative information is quiet limited. We expect the use of audio-visual information can help to mitigate the problem. In our investigation, we select the samples with less than 0.8 second from the validation set. As shown in Figure 4, all systems obtain roughly better performance when the sample duration is increased. This result demonstrates that the samples with long duration are more accurate than that with short duration as they capture richer information for prediction. The audio system performs much better than the visual system, but it still performs poorly in the very short duration samples. The audio-visual system consistantly achieves superior performance over the audio- and visual-only systems. More importantly, it achieves a large improvement on those short duration samples, which convincingly verifies the importance of audio-visual integration. In summary, joint training on both audio and visual systems is critical for improving the word recognition of relatively short duration samples.

4 CONCLUSIONS

We have proposed an audio-visual system for closed-set word-level speech recognition. The audio-visual system sets the pre-trained 3D-CNN model and the attention-based Bi-LSTM model as the initial visual and audio encoders, respectively. Then we add another two fully connected layers on the concatenated encoder outputs for joint training the audio-visual system. Experiments on the LRW-1000 dataset show that compared to the audio system, the word accuracy of the proposed audio-visual system on the test set is relatively improved by 7.9%, and joint training on both audio and visual systems is critical for improving the word recognition of relatively short duration samples.

ACKNOWLEDGMENTS

This research has been supported by the National Natural Science Foundation of China (No. 61571363 and No. 61571362), the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JM6015), and the Fundamental Research Funds for the Central Universities (No. 3102019ZY1004).

REFERENCES

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[2] Ibrahim Almajai, Stephen Cox, Richard Harvey, and Yuxuan Lan. 2016. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In *Proc. ICASSP*. 2722–2726.

[3] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. 2016. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599* (2016).

[4] Stéphane Dupont and Juergen Luettin. 2000. Audio-visual speech modeling for continuous speech recognition. *IEEE transactions on multimedia* 2, 3 (2000), 141–151.

[5] Alex Graves and Navdeep Jaitly. 2014. Towards end-to-end speech recognition with recurrent neural networks. In *Proc.*

ICML. 1764–1772.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proc. ICASSP*. 6645–6649.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proc. CVPR*. 770–778.

[8] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine* 29 (2012).

[9] Jing Huang and Brian Kingsbury. 2013. Audio-visual deep learning for noise robust speech recognition. In *Proc. ICASSP*. 7596–7599.

[10] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[11] Yiting Li, Yuki Takashima, Tetsuya Takiguchi, and Yasuo Ariki. 2016. Lip reading using a dynamic feature of lip images and convolutional neural networks. In *Proc. ICIS*. 1–6.

[12] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[13] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Proc. INTERSPEECH*.

[14] Hiroshi Ninomiya, Norihide Kitaoka, Satoshi Tamura, Yurie Iribe, and Kazuya Takeda. 2015. Integration of deep bottleneck features for audio-visual speech recognition. In *Proc. INTERSPEECH*.

[15] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. 2014. Lipreading using convolutional neural network. In *Proc. INTERSPEECH*.

[16] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. 2015. Audio-visual speech recognition using deep learning. *Applied Intelligence* 42, 4 (2015), 722–737.

[17] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *Proc. NIPS-W*.

[18] Stavros Petridis, Zuwei Li, and Maja Pantic. 2017. End-to-end visual speech recognition with LSTMs. In *Proc. ICASSP*. 2592–2596.

[19] Stavros Petridis and Maja Pantic. 2015. Prediction-based audiovisual fusion for classification of non-linguistic vocalisations. *IEEE Transactions on Affective Computing* 7, 1 (2015), 45–58.

[20] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *Proc. ICASSP*. 6548–6552.

[21] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. 2018. Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture. In *Proc. SLT*. 513–520.

[22] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* 91, 9 (2003), 1306–1326.

[23] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. 2004. Audio-visual automatic speech recognition: An overview. *Issues in visual and audio-visual speech processing* 22 (2004), 23.

[24] Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Proc. INTERSPEECH*.

[25] Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining residual networks with LSTMs for lipreading. *arXiv preprint arXiv:1703.04105* (2017).

[26] Kwanchiva Thangthai, Richard W Harvey, Stephen J Cox, and Barry-John Theobald. 2015. Improving lip-reading performance for robust audiovisual speech recognition using DNNs.. In *Proc. AVSP*. 127–131.

[27] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proc. ICCV*. 4489–4497.

[28] Michael Wand and Jürgen Schmidhuber. 2017. Improving speaker-independent lipreading with domain-adversarial training. *arXiv preprint arXiv:1708.01565* (2017).

[29] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin Chen. 2018. LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild. *arXiv preprint arXiv:1810.06990* (2018).

[30] Yougen Yuan, Cheung-Chi Leung, Lei Xie, Hongjie Chen, and Bin Ma. 2019. Query-by-Example Speech Search using Recurrent Neural Acoustic Word Embeddings with Temporal Context. *IEEE Access* 7, 1 (2019), 67656–67665.