Attention-based Audio-Visual Fusion for Robust Automatic Speech Recognition

George Sterpu*
ADAPT Centre, Sigmedia Lab, EE
Engineering, Trinity College Dublin
Dublin, Ireland
sterpug@tcd.ie

Christian Saam*
ADAPT Centre, Sigmedia Lab,
Trinity College Dublin
Dublin, Ireland
saamc@scss.tcd.ie

Naomi Harte
ADAPT Centre, Sigmedia Lab, EE
Engineering, Trinity College Dublin
Dublin, Ireland
nharte@tcd.ie

ABSTRACT

Automatic speech recognition can potentially benefit from the lip motion patterns, complementing acoustic speech to improve the overall recognition performance, particularly in noise. In this paper we propose an audio-visual fusion strategy that goes beyond simple feature concatenation and learns to automatically align the two modalities, leading to enhanced representations which increase the recognition accuracy in both clean and noisy conditions. We test our strategy on the TCD-TIMIT and LRS2 datasets, designed for large vocabulary continuous speech recognition, applying three types of noise at different power ratios. We also exploit state of the art Sequence-to-Sequence architectures, showing that our method can be easily integrated. Results show relative improvements from 7% up to 30% on TCD-TIMIT over the acoustic modality alone, depending on the acoustic noise level. We anticipate that the fusion strategy can easily generalise to many other multimodal tasks which involve correlated modalities.

KEYWORDS

Lipreading; Audio-Visual Speech Recognition; Multimodal Fusion; Multimodal Interfaces

1 INTRODUCTION

Human speech interaction is inherently multimodal in nature: we both watch and listen when communicating with other people. Under clean acoustic conditions, the auditory modality carries most of the useful information, and recent state of art systems [5] are capable of automatically transcribing spoken utterances with an accuracy above 95%. The visual modality becomes most effective

---

when the audio channel is corrupted by noise, as it allows us to recover some of the suppressed linguistic features.

Exploiting both modalities in the context of Automatic Audio-Visual Speech Recognition (AVSR) has been a challenge. One reason is the inconclusive research on what are good visual features for Large Vocabulary Continuous Speech Recognition (LVCSR) [14] that match the well established Mel-frequency cepstral coefficients for acoustic speech. Another reason is the need for a fusion strategy of two correlated data streams running at different frame rates [11].

Our work addresses both these challenges and we make a number of contributions. We introduce an audio-visual fusion strategy that learns to align the two modalities in the feature space. We embed this into a state of the art attention-based Sequence-to-Sequence (Seq2seq) Automatic Speech Recogition (ASR) system consisting of Recurrent Neural Network (RNN) encoders and decoders. We evaluate the strategy on a LVCSR task involving the two largest publicly available audio-visual datasets, TCD-TIMIT and LRS2, which contain complex sentences of both read speech and in-the-wild recordings. Using both of these datasets offers repeatability and allows other researchers to compare their systems directly to ours. We demonstrate improved performance over an acoustic-only ASR system for clean speech and also for three types of noise: white noise, cafe noise and street noise.

The paper is organised as follows: In section 2 we review related work in the field of AVSR. In section 3 we describe our framework and introduce our audio-visual fusion strategy. Section 4 presents our experimental results, and we discuss our findings in Section 5.

2 RELATED WORK

ASR. The work of [5, 15] shows that attention-based Seq2seq models [1] can surpass the performance of traditional ASR systems on challenging speech transcription tasks such as dictation and voice search. These models are characterised by a high degree of simplicity and generality, requiring minimal domain knowledge and making no strong assumptions about the modelled sequences. These properties make them good candidates for the AVSR task, where the potential benefit of the visual modality has been limited by an inability to exploit human domain knowledge for modelling. In our work, we use attention-based Seq2seq models to encode representations of both modalities.

Lip-reading. Transcribing speech from the visual modality alone, i.e. lip-reading, has improved significantly in recent years thanks to the advancements in neural networks. While there have been multiple contributions to the word classification task [7, 13, 20, 22], the work of [19] transcribes large vocabulary speech at the character level, and this is the work most comparable to ours. The visual

processing front-end consists of a Convolutional Neural Network (CNN) which learns a hierarchy of features from the training data. We adopt the CNN variant with residual connections [10], as it attains competitive results in many computer vision tasks.

Audio-Visual fusion. Some of the earliest audio-visual fusion strategies are reviewed in [11, 14], broadly classified into *feature fusion* and *decision fusion* methods. In AVSR, feature fusion is the dominant approach [13, 19]. Yet, as pointed out by [14], concatenating features does not explicitly model the stream reliabilities. Furthermore, [19] observe that the system becomes over-reliant on the audio modality and apply a regularisation technique where one of the streams is randomly dropped out during training [12]. Our method differs by always learning from the two streams at the same time. In addition, while [19] concatenates the sentence summaries, we learn to correlate the acoustic with the visual features at every time step, an opportunity to make corrections at much finer detail. Several other approaches from related areas focus on modelling cross-view or cross-modality temporal dynamics and alignment with modifications of gated memory cells [16, 18, 23, 24]. While these RNN architectures were introduced to address difficulties in learning long-term dependencies e.g. Long Short-Term Memory (LSTM) cells [8] or Gated Recurrent Units [6], such dependencies were still found problematic in Seq2seq models based on gated RNN cells and lead to the introduction of temporal attention mechanisms [1]. Thus it seems reasonable to explore such attention mechanisms for potentially long-range cross-stream interactions rather than putting additional memorisation burdens on the basic building blocks. A gating unit is introduced in [21], designed to filter out unreliable features. However, it still takes concatenated audio-visual inputs, and the synchrony of the streams is only ensured through identical sampling rate, limiting the temporal context.

3 METHOD

We begin by reviewing attention-based Seq2seq networks in order to introduce our fusion strategy more clearly in Section 3.5.

3.1 Background

The network consists of a sequence encoder, a sequence decoder and an attention mechanism. The encoder is based on an RNN which consumes a sequence of feature vectors, generating intermediate latent representations (coined as *memory*) and a final state representing the sequence latent summary. The decoder is also an RNN, initialised from the sequence summary, which predicts the language units of interest (e.g. characters). Because the encoding process tends to be lossy on long input sequences, an attention mechanism was introduced to soft-select from the encoder memory a context vector to assist each decoding step.

Attention mechanisms. Attention typically consists in computing a context vector as a combination of state vectors (*values*) taken from a memory, weighted by their correlation score with a target state (or *query*).

$$\alpha_{ij} = score\,(value_j,\ query_i) \tag{1}$$

$$c_i = \sum_j \alpha_{ij}\,query_i \tag{2}$$

Table 1: CNN Architecture. All convolutions use 3x3 kernels, except the final one. The Residual Block is taken from [10] in its *full preactivation* variant.

layer	operation	output shape
0	Rescale [-1 ... +1]	36x36x**3**
1	Conv	36x36x**8**
2-3	Res block	36x36x**8**
4-5	Res block	18x18x**16**
6-7	Res block	9x9x**24**
8-9	Res block	5x5x**32**
10	Conv 5x5	1x1x**128**

The context vector c is then mixed with the *query* to form a context-aware state vector. So far, this has been used successfully in Seq2seq decoders, where the *query* is the current decoder state and the *values* represent the encoder memory.

3.2 Inputs

Our system takes auditory and visual input concurrently. The **audio** is the raw waveform signal of an entire sentence. The **visual** input consists of video frame sequences, centred on the speaker's face, which correspond to the audio track. We use the OpenFace toolkit [3] to detect and align the faces, then we crop around the lip region.

3.3 Input pre-processing

Audio input. The audio waveforms are re-sampled at 22,050 Hz, with the option at this step to add several types of acoustic noise at different Signal to Noise Ratios (SNR). Similar to [5], we compute the log magnitude spectrogram of the input, choosing a window length of 25ms with 10ms stride and 1024 frequency bins for the Short-time Fourier Transform, and a frequency range from 80Hz to 11,025Hz with 30 bins for the mel scale warp. Finally, we append the first and second order derivatives of the log mel features, ending up with a feature of size 90 computed every 10 ms.

Visual input. The lip regions are 3-channel RGB images downsampled to 36x36 pixels. A ResNet CNN [10] processes the images to produce a feature vector of **128 units** per frame. The details of the architecture are presented in Table 1.

3.4 Sequence encoders

Audio and visual feature sequences differ in length, being sampled at 100 and 30 Frames per Second (FPS) respectively. Across training examples, the sequences also have variable lengths. We process them using two LSTM [8] RNNs. The architecture of the LSTMs consists of **3 layers** of **256 units**. We collect the top-layer output sequences of both LSTMs and also their final states, referring to them as encoding memories and sequence summaries respectively.

3.5 Audio-Visual fusion strategy

This subsection describes the key architectural contribution. Our premise is that conventional dual-attention mechanisms, such as the one in [19], overburden the decoder in Seq2seq architectures. In the uni-modal case, a typical decoder has to perform both language modelling and acoustic decoding. Adding another attention
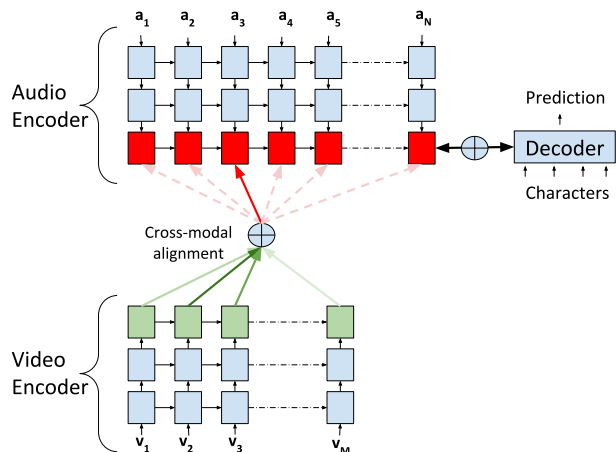
Figure 1: Proposed Audio-Visual Fusion Strategy. The top layer cells of the Audio Encoder (red) attend to the top layer outputs of the Video Encoder (green). The Decoder receives only the Audio Encoder outputs after fusion with the Video outputs. For clarity, the Decoder is not fully shown.

mechanism that attends to a second modality requires the decoder to also learn correlations among the input modalities.

We aim to make the modelling of the audio-visual correlation more explicit, while completely separating it from the decoder. Thus, we move this task to the encoder side. Our strategy is to decouple one modality from the decoder and to introduce a supplementary attention mechanism to the top layer of the coupled modality that attends to the encoding memory of the decoupled modality. The decoder only receives the final state and memory of the coupled encoder's top layer like a standard uni-modal attention decoder. A diagram of the fusion strategy is shown in Figure 1.

The *queries* from Eq. (1, 2) come from the states of the top audio encoder layer, while the *values* represent the visual encoder memory. The acoustic encoder's top layer can no longer be considered to hold acoustic-only representations. They are fused audio-visual representations based on corresponding high level features from the two modalities matched via attention. This layer could also be viewed separately as a higher level encoder (red layer in Figure 1) operating on acoustic and visual hidden representations.

The following intuition motivates our choice. The top layers of stacked RNNs encode higher order features, which are easier to correlate than lower levels. They provide speech related abstractions from visual and acoustic features. In addition, any time one feature stream is corrupted by noise its encodings may be automatically corrected by correlated encodings of the other stream.

The audio and visual modalities can have two-way interactions. However, by design only the acoustic modality learns from the visual. This is because in clean speech, the acoustic modality is dominant and sufficient for recognition, while the visual one presents intrinsic ambiguities: the same mouth shape can explain multiple sounds. The design assumes that acoustic encodings can be partially corrected or even reconstructed from visual encodings. One disadvantage is that an alignment score has to be computed for each timestep of the much longer audio sequence.

We also note that this fusion strategy explicitly models the alignment between each acoustic and all visual encodings. This elegantly addresses the problem of different frame rates, that traditionally required slower modalities to be interpolated in order to match the frame rate of the fastest modality.

3.6 Decoding

Our decoder is a single layer LSTM of **256 units**. As in [5], we use four attention heads to improve the overall performance, while still attending to a single enhanced memory. The decoder predicts characters, word level results are inferred by splitting at blanks.

4 EXPERIMENTS AND RESULTS

We apply our audio-visual fusion method in the context of LVCSR. This is a well suited task for measuring the fusion performance, as correlations between modalities can be captured at the sub-word level. We also consider that the visual modality has a lot more to benefit from full sentences instead of isolated words, as longer contexts are critical to disambiguate the spoken message, thus supplying more informative features to the auditory modality.

4.1 Datasets

TCD-TIMIT [9] and LRS2 [4] are the largest publicly available audio-visual datasets suitable for continuous speech recognition. We report our results on both datasets.

TCD-TIMIT consists of 62 speakers and more than 6,000 examples of read speech, typically 4-5 seconds long, from both phonetically balanced and natural sentences. The video is laboratory recorded at 1080p resolution from two fixed viewpoints. We follow the speaker-dependent train/test protocol of [9].

LRS2 contains more than 45,000 spoken sentences from BBC television. Unlike TCD-TIMIT it contains more challenging head poses, but a much lower image resolution of 160x160 pixels.

Our results on LRS2 are not directly comparable with the ones on LRS [19], which is completely different from LRS2 and was never publicly released. In addition, LRS2 has a lot more diversity in content and head poses than LRS. To the best of our knowledge, this is the first audio-visual baseline on LRS2.

We used the suggested evaluation protocol on LRS2 with one difference. The face detection [3] failed on several challenging videos, so we kept only those where the detection was successful for at least 90% of the frames. With this rule, we kept approximately 87% of the training videos and 91% of the test videos. To foster reproducibility, we make available the list of files used in our subset.

4.2 Training and evaluation procedures

We train several uni- and bimodal Seq2seq systems. The unimodal networks process only the acoustic input, either in clean form or additively corrupted by three types of noise: *a) White Gaussian Noise (WGN)*, *b) Cafeteria Noise*, and *c) Street Noise*. The bimodal networks take both audio and visual inputs at the same time. We compare our method (**AV Align**) to an audio-only system (**A**) and dual attention feature concatenation [19] (**AV Cat**).

In training, we directly optimise the cross entropy loss between ground-truth character transcription and predicted character sequence via the AMSGrad optimiser [17]. In evaluation, we measure
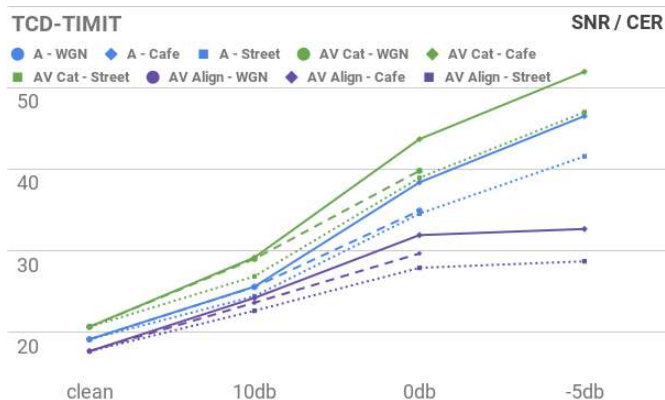
Figure 2: Character Error Rate (CER) [%] on TCD-TIMIT

Figure 3: Character Error Rate (CER) [%] on LRS2

the Levenshtein edit distance between them, normalised by the ground truth length. Our results report this Character Error Rate (CER) metric in percents. Our results are plotted in Figures 2-3, while complete numerical values are listed in the appendix.

5 DISCUSSION

We begin by analysing our results on TCD-TIMIT. We first notice relative improvements starting at 7% on clean speech (17.7% CER down from 19.16%), up to 30% at -5db SNR (32.68% CER down from 46.52%) using our fusion strategy (**AV Align**) over audio only (**A**). The feature concatenation approach (**AV Cat**) performed between 8% and 14% worse than **A**, suggesting that the **AV Cat** strategy is far from optimal and may be difficult to train. Indeed, [19] reported training for 500,000 iterations. They also used a stream dropout regularisation technique which failed to improve convergence in our case after training for longer than **AV Align** took to converge.

Despite again clearly outperforming **AV Cat**, we see no benefit of **AV Align** over audio-only recognition (**A**) on LRS2. Our visual front-end, while effective on the limited variability conditions of TCD-TIMIT, may not be large enough to cope with the much more demanding conditions, varying poses and larger number of faces. Although not directly comparable, [19] used a more powerful front-end architecture with additional pre-training. It is plausible that an under-performing video feature extraction cannot provide strong enough support to the audio stream.

Further, we observe in the network outputs a progression through several stages of learning. At first, the decoder forms a strong language model learning correct words and phrases. Later, the influence of the acoustic decoding increases and the network learns letter to sound rules, over-generalising like a child and unlearning some of the correct spellings for words. The comparatively much larger size of LRS2 allows re-learning the multitude of letter to sound rules much more reliably, such that it may become the factor dominating the error rate and driving the training. We may then expect that the network needs to train longer and past this stage to start to fully exploit the visual information on LRS2.

With **AV Align** the network appears to learn how to cope with noise by leveraging the visual modality. Furthermore, few studies prior to [19] show improvements in the clean condition, prompting
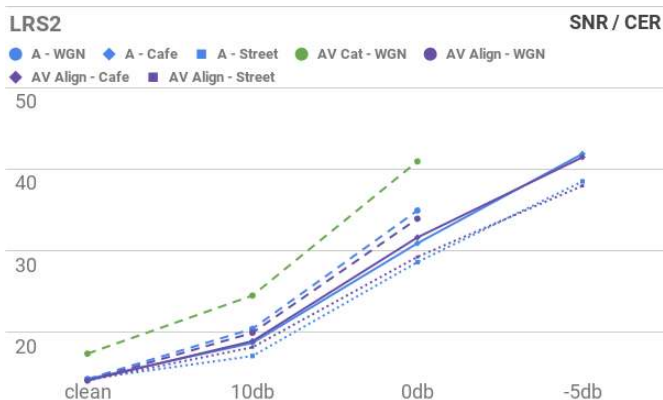
[2] to call the interactions supplementary rather than complementary. That our model is able to consistently outperform audio only recognition on TCD-TIMIT even under clean conditions points to the inherent acoustic confusability of speech and the potentially sub-optimal speech recognition process for which the visual signal may offer some complementary information. We conclude that the **AV Align** strategy effectively separates the stream alignment task from the speech sound classification and language modelling tasks in the decoder, as opposed to the joint modelling in [19].

One shortcoming of our method is the lack of explicit modelling of the visual stream's reliability. However, the result on LRS2 suggests that the network may have learnt to discard noisy or uninformative visual features. An open question is whether the attention-based fusion strategy is powerful enough to cope with visual noise, rendering unnecessary a more complex structure.

6 CONCLUSIONS

In this work, we introduce an audio-visual fusion strategy for speech recognition. The method uses an attention mechanism to automatically learn an alignment between acoustic and visual modalities, leading to an enhanced representation of speech. We demonstrate its effectiveness on the TCD-TIMIT and LRS2 datasets, observing relative improvements up to 30% over an audio-only system on high quality images, and no significant degradation when the visual information becomes harder to exploit. Since our strategy was able to discover structure in an audio-visual speech recognition task, we expect it to generalise to others tasks where the input modalities are semantically correlated.

Further to its performance, what makes our fusion strategy attractive is its straightforward formulation: it can be applied to attention-based Seq2seq approaches by reusing existing attention code. Not only may it effectively generalise to many other multimodal applications, but it also allows researchers to easily integrate our method into their solutions.

ACKNOWLEDGMENTS

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2018. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*. http://arxiv.org/abs/1409.0473

[2] T. Baltrušaitis, C. Ahuja, and L. P. Morency. 2018. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018), 1–1.

[3] T. Baltrušaitis, P. Robinson, and L. P. Morency. 2016. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–10. https://doi.org/10.1109/WACV.2016.7477553

[4] BBC and Oxford University. 2017. The BBC-Oxford Multi-View Lip Reading Sentences 2 (LRS2) Dataset. http://www.robots.ox.ac.uk/~vgg/data/lip_reading_sentences/. (2017). Online, Accessed: 11 August 2018.

[5] Chung-Cheng Chiu, Tara Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J. Weiss, Kanishka Rao, Katya Gonina, Navdeep Jaitly, Bo Li, Jan Chorowski, and Michiel Bacchiani. 2018. State-of-the-art Speech Recognition With Sequence-to-Sequence Models. In *ICASSP*. https://arxiv.org/pdf/1712.01769.pdf

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1724–1734. https://doi.org/10.3115/v1/D14-1179

[7] J. S. Chung and A. Zisserman. 2016. Lip Reading in the Wild. In *Asian Conference on Computer Vision*.

[8] F.A. Gers, J. Schmidhuber, and F. Cummins. 1999. Learning to forget: continual prediction with LSTM. *IET Conference Proceedings* (January 1999), 850–855(5). http://digital-library.theiet.org/content/conferences/10.1049/cp_19991218

[9] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia* 17, 5 (May 2015), 603–615.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 630–645.

[11] A. K. Katsaggelos, S. Bahaadini, and R. Molina. 2015. Audiovisual Fusion: Challenges and New Approaches. *Proc. IEEE* 103, 9 (Sept 2015), 1635–1653. https://doi.org/10.1109/JPROC.2015.2459017

[12] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*. 689–696.

[13] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. 2018. End-to-end Audiovisual Speech Recognition. In *ICASSP*. http://arxiv.org/abs/1802.06424

[14] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE* 91, 9 (Sept 2003), 1306–1326. https://doi.org/10.1109/JPROC.2003.817150

[15] Rohit Prabhavalkar, Kanishka Rao, Tara N. Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. 2017. A Comparison of Sequence-to-Sequence Models for Speech Recognition. In *Proc. Interspeech 2017*. 939–943. https://doi.org/10.21437/Interspeech.2017-233

[16] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Roland Goecke. 2016. Extending Long Short-Term Memory for Multi-View Structured Learning. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*. Springer, Cham, 338–353. https://doi.org/10.1007/978-3-319-46478-7_21

[17] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the Convergence of Adam and Beyond. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ryQu7f-RZ

[18] Jimmy Ren, Yongtao Hu, Yu-Wing Tai, Chuan Wang, Li Xu, Wenxiu Sun, and Qiong Yan. 2016. Look, Listen and Learn – A Multimodal LSTM for Speaker Identification. In *Thirtieth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12386 00014.

[19] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2017. Lip Reading Sentences in the Wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining Residual Networks with LSTMs for Lipreading. In *Proc. Interspeech 2017*. 3652–3656. https://doi.org/10.21437/Interspeech.2017-85

[21] F. Tao and C. Busso. 2018. Gating Neural Network for Large Vocabulary Audiovisual Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 7 (July 2018), 1286–1298. https://doi.org/10.1109/TASLP.2018.2815268

[22] Michael Wand and Jürgen Schmidhuber. 2017. Improving Speaker-Independent Lipreading with Domain-Adversarial Training. In *Proc. Interspeech 2017*. 3662–3666. https://doi.org/10.21437/Interspeech.2017-421

[23] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *AAAI Conference on Artificial Intelligence*. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17341

[24] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention Recurrent Network for Human Communication Comprehension. In *AAAI Conference on Artificial Intelligence*. https://aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17390