Transformer-Based Video Front-Ends for Audio-Visual Speech Recognition for Single and Multi-Person Video

Dmitriy Serdyuk *Otavio Braga* *Olivier Siohan*

Google, 111 8th Ave, New York, 10011 USA

{dserdyuk,obraga,siohan}@google.com

Abstract

Audio-visual automatic speech recognition (AV-ASR) extends speech recognition by introducing the video modality as an additional source of information. In this work, the information contained in the motion of the speaker's mouth is used to augment the audio features. The video modality is traditionally processed with a 3D convolutional neural network (e.g. 3D version of VGG). Recently, image transformer networks [1] demonstrated the ability to extract rich visual features for image classification tasks. Here, we propose to replace the 3D convolution with a video transformer to extract visual features. We train our baselines and the proposed model on a large scale corpus of YouTube videos. The performance of our approach is evaluated on a labeled subset of YouTube videos as well as on the LRS3-TED public corpus. Our best video-only model obtains 31.4% WER on YTDEV18 and 17.0% on LRS3-TED, a 10% and 15% relative improvements over our convolutional baseline. We achieve the state of the art performance of the audio-visual recognition on the LRS3-TED after fine-tuning our model (1.6% WER). In addition, in a series of experiments on multi-person AV-ASR, we obtained an average relative reduction of 2% over our convolutional video frontend.

***Index Terms*—** Audio-visual speech recognition, lip reading, video transformer, deep learning.

1. Introduction

Many real-world applications of speech recognition operate on a video input (e.g. YouTube videos, webcasts, internet streams, TV broadcasts). *Audio-visual automatic speech recognition* (AV-ASR, [2, 3, 4]) adds the video modality to the traditional speech recognition. It has been shown that the video may help recognition, especially in adverse audio conditions [4]. The extreme case occurs when the audio is unavailable, a scenario known as *lip reading* [5].

A typical end-to-end system for AV-ASR requires a strong visual feature extractor – the *video front-end*. This critical component of the AV-ASR system encodes the movements of the speaker's lips movements into the features used downstream for recognition. Usually, the video front-end is a trainable 3D convolutional network (e.g. a 3D variant of VGG, [6]).

In order to improve the video front-end, we draw inspiration from the recent works in the area of NLP. A self attention-based [7] *transformer* [8] architecture was proposed for a variety of sequential tasks. The transformer was instrumental for developing strong NLP [9, 10] and ASR [11] systems. In the area of computer vision the convolutional networks are the model of choice for image processing. Recently, it has been shown that a transformer architecture (*vision transformers*, ViT, [1]) is viable for image classification. The proposed transformer is able to achieve parity or superior performance to the convolutional networks. Later, this work was extended to the video classification [12, 13].

Inspired by the success of vision transformers, we propose to use a transformer-based architecture for the video front-end of the AV-ASR system. We design a video transformer front-end which takes a sliding 3D window of the video. This window is split into 3D patches of the size 32x32x8 pixels. Then following [1, 12], we apply an affine transform to the patches followed by a transformer encoder.

This work extends our workshop paper [14]. We add the experiments with a stronger conformer [15] encoder. We conduct extra experiments on the LRS3-TED with the artificially added noise.

Compared to our previous work in [14], the contributions of this paper are:

- We test the feasibility of the transformer-based video front-end for the AV-ASR and lip reading. We design a model that uses an off-the-shelf transformer for the video encoding.
- We experimentally evaluate the proposed model. We train on a large scale dataset of YouTube videos. We experiment with the transformer encoder and the conformer audio-visual encoders. We evaluate on the YTDEV18 and LRS3-TED datasets. The experiments show that the transformer video front-end works at least as good as the convolution. Moreover, we achieve the state of the art performance on the LRS3-TED dataset after fine-tuning our model.
- We investigate the proposed model for the multi-person data. Our model outperfroms the convolution baseline.

2. Related Work

Audio-Visual Automatic Speech Recognition. Audio-visual speech recognition [2] made significant progress thanks to the introduction of the end-to-end approaches [16, 17, 4]. Using deep neural networks and end-to-end training allowed these and other works to tackle audio-visual speech recognition "in the wild", i.e. unconstrained open-world utterances.

Recently, a remarkable work [18] achieved the state of the art on the tasks of audio-visual speech recognition and lip reading by combining CTC [19] and seq2seq [7] losses with a conformer network [15].

Transformer-based Models for Video. Since the attention-based [7] transformer architecture was introduced in [8], it quickly became a model of choice for natural language processing [9, 10]. Later, it was employed for other sequence modeling tasks, such as speech recognition [11]. A highly influential paper on *visual transformers*, ViT, [1], was the first work demonstrating that the transformer architecture performs at least as well as convolutions.

3D CNN is the traditional method of using audio-visual networks.

They used YT videos and LSR3-TED to train and verify the model with better accuracy.

Transformer based architecture could help audio visual networks for speech recognition in environments with noise.

Video Transformer with certain specifics for video.

ViT, Visual Transformer is as good as standard convolutional neural network!

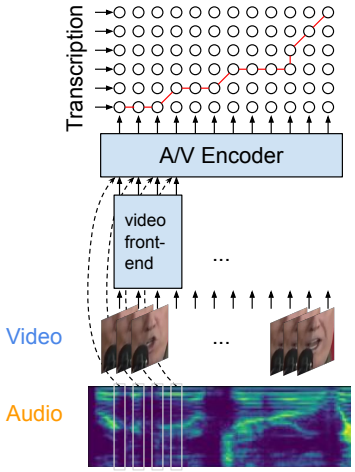

Fig. 1: *An overview of end-to-end AV-ASR and lip reading models. The video is encoded with a video front-end. The visual features and the acoustic features are concatenated and fed through the AV encoder to be used for the RNN-T loss.*

Next, the visual transformer was extended from still images to video [12, 20, 21], in particular for tasks such as video classification and action recognition. In contrast to these works, this paper focuses on a sequence-to-sequence task. While the sequence output provides a signal stronger than classification, the task is harder due to the fact that the network is required to learn the alignment.

3. Model

In this section we outline the general pipeline for the audio-visual ASR used in our experiments. Then, we describe the video front-ends, which is a focus of this paper. We introduce the baseline convolutional front-ends and the proposed transformer-based video front-end.

3.1. Common A/V ASR Model Architecture

The common AV-ASR pipeline (Fig. 1) is shared between all of our experiments.

Acoustic Features. We extract 80 log Mel filterbank features from the 16kHz input signal with a 25ms wide Hann window with steps of 10ms. Then, we fold each 3 consecutive features to produce 240-dimensional input which we denote as $\mathbf{A} \in \mathbb{R}^T \times \mathbb{R}^{D_a}$, where T is the number of time-steps, $D_a = 240$ is the dimensionality of the acoustic features. This corresponds to the acoustic features with the frequency of ≈ 33.3Hz.

Visual Features. The source videos have varying frame rates in the range from 23 to 30 frames per second. In order to synchronize the features, we re-sample the video frames at the frequency of the audio features (33.3Hz) using the nearest neighbor interpolation. Then, we crop the videos centering near the mouth region to produce frames of the size 128×128. The video is fed then into the video front-end (Section 3.2) which yields the video features $\mathbf{V} \in \mathbb{R}^T \times \mathbb{R}^{D_v}$ of dimension $D_v = 512$.

Encoder. The encoder combines two modalities and embeds them for the use in the decoder. In all our audio-visual experiments we concatenate the 240-dimensional audio features with the 512-dimensional video features to produce the fused features $\mathbf{F} = [\mathbf{A}; \mathbf{V}] \in \mathbb{R}^T \times \mathbb{R}^{D_a + D_v}$, totalling 752 input features at each time step. For the video-only (lip reading) we ignore the audio features and decrease the input dimension of the encoder (512 input features), which is $\mathbf{F} = \mathbf{V}$. We use two architectures for the encoder:

1. Transformer: a 14 layer transformer encoder [8] with 512 hidden dimensions, 8 attention heads, and the relative positional embedding. The self-attention window is limited to 100 timesteps on left and right.
2. Conformer: a conformer encoder [15] with 17 layers. The hidden dimension is again 512 and the kernel size of 32.

Decoder. The decoder is a two layer LSTM network, where each layer has 2048 units. The RNN-T [22] loss produces the character level output.

3.2. Video Front-Ends

The pre-processed video is fed into a video front-end. In this work we use two types of the video front-ends: the baseline (2+1)D convolutional network and the video transformer network.

3.2.1. (2+1)D ConvNet Baseline

Our baseline system uses a VGG 3D front-end [4] with the following change. We decompose each filter into the spatial and temporal dimensions. For example, a $[3, 3, 3]$ filter becomes two filters $[1, 3, 3]$ and $[3, 1, 1]$. This modification reduces the memory requirements and regularizes the model. In all our experiments we use a 10 layer convolutional network[1]. We refer to this baseline as the VGG (2+1)D.

3.2.2. Video Transformer Front-end

One of the goals of our design of the video transformer is to reuse the available implementations of the transformer architecture. We aim to use the minimal number of modifications for the existing models. This model is visualized in the Fig. 2.

First, we extract patches of size $P_w \times P_h$ $(= 32 \times 32)$ resulting into $N = W/P_w \times H/P_h$ $(= 4 \times 4 = 16)$ patches at each timestep of the video. Then, we fold each P_d $(= 8)$ consecutive frames which yields a series of *tubelets* $\mathbf{P}_c \in \mathbb{R}^{P_w} \times \mathbb{R}^{P_h} \times \mathbb{R}^{P_d}$ (3D version of a patch). After repeating this for each channel and each timestep, we flatten the tubelet tensor $\mathbf{P} \in \mathbb{R}^T \times \mathbb{R}^N \times \mathbb{R}^{P_w} \times \mathbb{R}^{P_h} \times \mathbb{R}^{P_d} \times \mathbb{R}^C \rightarrow \mathbf{P}^{flat} \in \mathbb{R}^T \times \mathbb{R}^N \times \mathbb{R}^{P_w \times P_h \times P_d \times C}$. Next, the flattened tensor of patches is fed into an affine transform shared in between patches and the timesteps: $W_{kl}\mathbf{P}^{flat}_{ijk} + b_k$ which yields a sequence of features $\mathbf{P}^{feats} \in \mathbb{R}^T \times \mathbb{R}^N \times \mathbb{R}^{D_v}$ $(D_v = 512)$.

The transformed tensor of patches is combined with relative positional embeddings (we found that the relative and the absolute positional embeddings have the same performance). Finally, we run a transformer encoder over the N dimension treating the T as the batch dimension. We use a 6 layer transformer encoder [8]. Each layer consists of the self-attention and the feed-forward layer, which are combined with the layer-norm and a residual connection.

Finally, we take the first output of the transformer encoder and discard the rest thus producing a set of the video features $\mathbf{V} \in \mathbb{R}^T \times \mathbb{R}^{D_v}$.

[1] Kernel sizes are: 23, 64, 230, 1

The Transformer and Conformer both combine with two LSTMs in order to make this approach work!

Transformers are as good as CNNs which means that they could be a great alternative!

They use positional embeddings, something that I explored in my Original Work projects!

Acoustic and Visual features talk about how data is being preprocessed to be used successfully for the model by focusing on an optimized size and cropped videos.

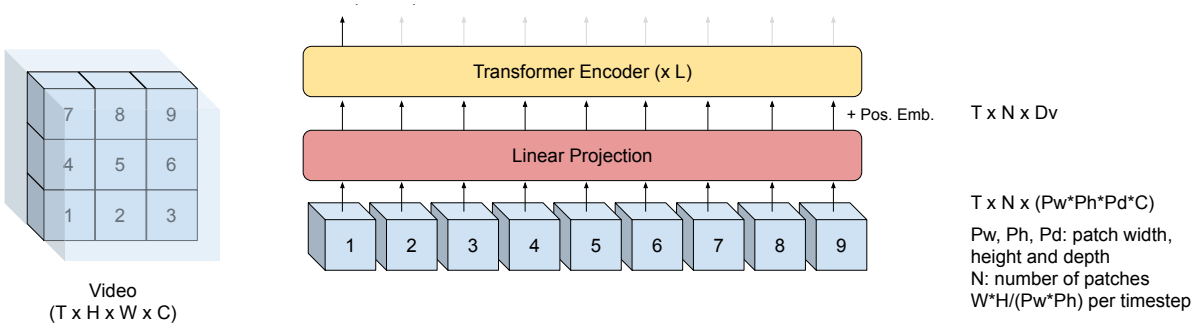

Fig. 2: *An overview of the proposed architecture for the video-encoding transformer. The input video is split into 'tubelets'. The tubelets are embedded with a linear projection and fed into a transformer.*

4. Experiments

In this section we describe our experiments and report the results. We started by training a lip reading system (Section 4.1) with a transformer and conformer encoders. In both cases the proposed video transformer front-end outperforms the convolutional baseline. Then, we train an audio-visual model using both input modalities. We find that the video transformer matches the baseline or performs slightly better than the baseline.

Datasets. We train on a dataset mined from public YouTube videos. We use a semi-supervised procedure proposed in [23] and adapted to include videos in [4, 24]. This procedure extracts short segments of the video where the force-aligned user uploaded transcript matches the production quality ASR system with high confidence. Then, only the segments are kept where the video track matches the audio with high confidence. The resulting dataset contains about 90k hours of transcribed video segments limited to 512 frames (15 seconds).

Datasets are used from Google's resources in order to create hours of video with 512 frames.

A separate set of YouTube videos is used for the development and eval sets. These videos were transcribed by professionals – the YTDEV18 set [4]. In order to compare to prior publications, we use the LRS3-TED [25] eval set.

Training. In all our experiments we use the batch size 8 and the Adam [26] optimizer ran synchronously on 128 accelerators (totalling the batch size of 1024). We use the multi-style training (MTR, [27]), which increases robustness to the noise.

The transformer models were trained with the following learning rate schedule. First, the learning rate linearly warmed up to $1e^{-4}$ for the first $30,000$ iterations. Then, it is constant until iteration $200,000$. Finally, it is annealed exponentially down to $1e^{-6}$ until iteration $300,000$.

The conformer models use the learning rate schedule which warms up linearly and then anneals exponentially. The maximum learning rate is $1.7e^{-2}$ and the number of warm up iterations is $15,000$.

4.1. Lip Reading

We summarize our findings for the lip reading models in Table 1. The proposed ViT 3D model model outperforms the VGG baseline when using the transformer encoder (4% relative improvement on YTDEV18 and 8% on LRS3-TED) and the conformer encoder (10% improvement on YTDEV18 and 9% on LRS3-TED). Furthermore, our models outperform the previous publications [4, 18, 28] with a caveat that we use a different training set.

ViT 3D model outperforms the transfer learning based model (VGG) on lip reading!

From these experiments we conclude that the ViT is able to provide strong visual features for the lip reading task.

4.2. Audio-Visual Automatic Speech Recognition

The experiments on the combined audio and video follow the same protocol as the lip reading with an exception that the encoder concatenates the audio features to the extracted video features. Then we artificially add the audio noise to the YTDEV18 (following [4]). We add the babble noise of signal-to-noise ratios 20dB, 10db, and 0dB. The noise was drawn from the NoiseX database [29]. Then, we overlap a fixed random utterance from the test set (denoted "Overlap"). The results are summarized in Table 2.

We find that our ViT front-end matches the performance of the VGG baseline. When using the transformer encoder we observe a slight increase in the performance for 0dB noise condition.

The lower part of the table refers to a stronger conformer encoder. The ViT front-end is able to match the convolutional front-end.

For both the transformer and conformer encoders we see a small dip in the performance for the overlap noise. One of possible reasons for this is that the model is trained with MTR which mitigates the babble noise but not the overlap noise. Notice that the system is still able to improve upon the video-only setup in Table 1 (29.9% vs 31.4%).

Tested with noise and the model is able to perform as good as a VGG. Babble noise vs overlap noise makes a difference for the system!

4.3. Multi-Person Audio-Visual Recognition

In this section, we briefly summarize our model applied to multi-person A/V ASR. We closely follow the setup in [30, 31], where multiple videos are encoded with the shared video front-end followed by the attention to choose the active speaker. The evaluation sets were artificially constructed by mixing the existing sets. The audio and video is taken from one utterance, then several (2, 4, or 8) video distractors are added.

We report the WER results with a varying number of video distractor speakers in the Figure 3. The solid lines are the baseline VGG video front-end, and the dash lines stand for the proposed ViT front-end. Our model outperforms the baseline in the majority of conditions.

Table 1: *Lip-reading performance, %WER. The proposed model (ViT 3D) outperforms the convolutional baseline (VGG) for both the transformer and conformer encoders.*

Model	YTDEV18	LRS3-TED
TM-seq2seq [28]	–	58.9
ResNet+Conf [18]	–	43.3
RNN-T [4]	48.5	33.6
Transformer encoder:		
VGG (2+1)D	40.5	28.2
ViT 3D	38.8	25.9
Conformer encoder:		
VGG (2+1)D	34.9	20.0
ViT 3D	**31.4**	**17.0**

Table 2: *Audio-visual ASR performance, %WER. (∞dB) is the clean subset; 20db, 10dB, 0dB – data with artificial noise added; "Overlap" – contains overlapped utterances. The proposed ViT model matches the VGG baseline.*

Model	∞dB	20dB	10dB	0dB	Overlap
Audio-only	16.5	17.0	19.8	42.9	35.0
Transformer:					
VGG (2+1)D	14.4	14.5	15.6	23.4	31.2
ViT 3D	14.4	14.6	15.6	23.1	31.9
Conformer:					
VGG (2+1)D	13.6	13.7	14.5	19.3	29.3
ViT 3D	13.4	13.5	14.3	19.3	29.9

4.4. Audio-Visual Recognition on LRS3-TED and Fine-Tuning

The LRS3-TED tends to have higher audio quality compared to the majority of the YouTube videos we use for training. In order to close this gap, we fine-tune our models on the LRS3-TED training set. More specifically, we train our models for $10,000$ steps on a 50-50 mix of the YouTube and the LRS3-TED training data. We use the maximum learning rate of $1e^{-5}$ which was warmed up linearly from 0 across the first 200 steps and then held constant.

The results for A/V are reported in Table 3. The fine-tuned transformer model matches the previous state of the art [18] for supervised models. The conformer-based models are reported in the lower section of Table 3. We found that the audio signal is strong enough to achieve the WER of 1.6%. Both the baseline AGG AV-ASR and the proposed AV ViT models match this result which demonstrates that the performance on TED is nearly saturated. Therefore, we corrupt the LRS-TED test set by adding the babble noise of 20dB, 10dB, and 0dB. The performance of the audio only model rapidly drops down to 6.1% for 0dB noise. In comparison, the AV models score 3.1% and 2.9% for the VGG and ViT front-ends.

As a side note, we were surprised that all the tested models performed so well in the presence of the 0dB babble noise (compare this to the performance drop in Table 2). The main reason for this is the very high quality of the audio.

Finally, we did not observe any benefit in fine-tuning our lip reading models on LRS3-TED. We hypothesise that the domain shift between our train data and the LRS3-TED test set is the greatest for the audio modality (clean, professionally recorded audio).

They found that audio-visual woul use a mix of both non-noise and noise data that helps them optimize speech recognition!

Table 3: *AV-ASR performance on the LRS3-TED dataset. Models denoted with * are trained on a large dataset of YouTube videos. † denotes self-supervised pre-training. All conformer models are fine-tuned for the TED data.*

Model	WER, %	20dB	10dB	0dB
TM-CTC [28]	27.7	–	–	–
EG-s2s [32]	6.8	–	–	–
RNN-T [4]*	4.5	–	–	–
ResNet+Conf [18]	2.3	–	–	–
AV-Hubert [33]†	**1.3**	–	–	–
Transformer encoder:				
VGG (2+1)D*	3.3	3.3	3.4	6.4
ViT 3D*	3.3	3.3	3.3	6.2
+ fine-tune	2.3	2.4	2.6	5.1
Conformer encoder:				
audio-only*	**1.6**	1.5	1.8	6.1
VGG (2+1)D*	**1.6**	1.7	1.8	3.1
ViT 3D*	**1.6**	1.6	1.7	**2.9**

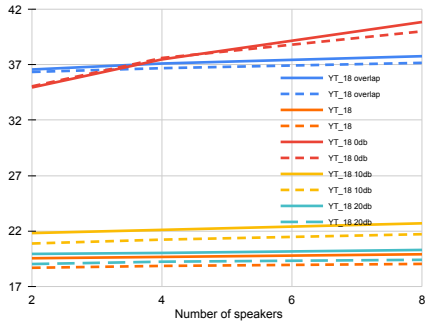

Fig. 3: *WER for multi-person recognition on YT_18 data.*

5. Conclusions

We compared a transformer-based front-end for video encoding to the convolutional front-end. We conclude that the transformer is a promising new architecture that is at least as good as the convolution. Furthermore, the ViT outperforms the convolutional baseline in certain settings, such as lip reading and the noisy LRS3-TED.

We fine-tuned our models on the public LRS3-TED dataset. This allowed the state of the art results on this set. We observed that the proposed ViT model outperforms the convolutional baseline and the audio-only recognition.

We are unable to use the LRS2-BBC dataset due to licence restrictions, which prohibits the dataset use by the private and industry researchers. Therefore, we cannot directly compare to some of the previously reported results.

ViT indeed outperforms the convolutional baseline in certain settings which might make it a better option to use for the Final Product!

6. Safe AI Principles

We are aware of the sensitive nature of the AV-ASR research and other AI technologies used in this work. Therefore, we ensure that this work abides by the Google AI Principles [34].

7. Acknowledgements

We would like to acknowledge support and advice of all our team mates, especially Hank Liao, Oscar Chang, Basi Garcia, and Kishan Sachdeva.

8. References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, Oct. 2020.

[2] Chalapathy Neti, Gerasimos Potamianos, Juergen Luettin, Iain Matthews, Herve Glotin, Dimitra Vergyri, June Sison, and Azad Mashari, "Audio-visual speech recognition," Tech. Rep., IDIAP, 2000.

[3] Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze, "Visual features for context-aware speech recognition," in *ICASSP*, Mar. 2017.

[4] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan, "Recurrent neural network transducer for audio-visual speech recognition," in *ASRU*, Nov. 2019.

[5] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "ASR is all you need: cross-modal distillation for lip reading," in *ICASSP*, Nov. 2019.

[6] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, Sept. 2014.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, June 2017.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv:1810.04805*, Oct. 2018.

[10] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," in *https://github.com/openai/gpt-2*, 2019.

[11] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and RNN-T loss," in *arXiv:2002.02562*, Feb. 2020.

[12] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid, "ViViT: A video vision transformer," in *arXiv:2103.15691*, Mar. 2021.

[13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani, "Is space-time attention all you need for video understanding?," in *arXiv:2102.05095*, Feb. 2021.

[14] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan, "Audio-visual speech recognition is worth $32\times32\times8$ voxels," 2021.

[15] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, May 2020.

[16] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas, "LIPNET: Sentence-level lipreading," in *arXiv:1611.01599*, 2016.

[17] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Lip reading sentences in the wild," in *CVPR*, Nov. 2016.

[18] Pingchuan Ma, Stavros Petridis, and Maja Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, Feb. 2021.

[19] Alex Graves, Santiago Fernandez, Faustino Gomez, and Jurgen Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, June 2006.

[20] Gilad Sharir, Asaf Noy, and Lihi Zelnik-Manor, "An image is worth 16x16 words, what is a video worth?," in *arXiv:2103.13915*, Mar. 2021.

[21] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann, "Video transformer network," in *arXiv:2102.00719*, Feb. 2021.

[22] Alex Graves, "Sequence transduction with recurrent neural networks," in *Representation Learning Workshop ICML*, Nov. 2012.

[23] Hank Liao, Erik McDermott, and Andrew Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *ASRU*, Dec. 2013, pp. 368–373.

[24] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, Marie Mulville, Misha Denil, Ben Coppin, Ben Laurie, Andrew Senior, and Nando de Freitas, "Large-scale visual speech recognition," in *Interspeech*, ISCA, Sept. 2019, ISCA.

[25] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," in *arXiv:1809.00496*, Sept. 2018.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, Dec. 2014.

[27] Xiaodong Cui, Vaibhava Goel, and Brian Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1469–1477, Sept. 2015.

[28] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, Dec. 2018.

[29] Andrew Varga and Herman JM 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*, 1993.

[30] Otavio Braga, Takaki Makino, Olivier Siohan, and Hank Liao, "End-to-end multi-person audio/visual automatic speech recognition," in *ICASSP*, May 2020.

[31] Otavio Braga and Olivier Siohan, "A closer look at audio-visual multi-person speech recognition and active speaker selection," in *ICASSP*, June 2021.

[32] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang, "Discriminative multi-modality speech recognition," in *CVPR*, May 2020.

[33] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed, "Learning Audio-Visual speech representation by masked multimodal cluster prediction," Jan. 2022.

[34] "Our principles – google AI," https://ai.google/principles/, Accessed: 2021-8-19.