ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414567

END-TO-END AUDIO-VISUAL SPEECH RECOGNITION WITH CONFORMERS

Pingchuan Ma, Stavros Petridis, Maja Pantic

Department of Computing, Imperial College London, UK

ABSTRACT

In this work, we present a hybrid CTC/Attention model based on a ResNet-18 and Convolution-augmented transformer (Conformer), that can be trained in an end-to-end manner. In particular, the audio and visual encoders learn to extract features directly from raw pixels and audio waveforms, respectively, which are then fed to conformers and then fusion takes place via a Multi-Layer Perceptron (MLP). The model learns to recognise characters using a combination of CTC and an attention mechanism. We show that end-to-end training, instead of using pre-computed visual features which is common in the literature, the use of a conformer, instead of a recurrent network, and the use of a transformer-based language model, significantly improve the performance of our model. We present results on the largest publicly available datasets for sentence-level speech recognition, Lip Reading Sentences 2 (LRS2) and Lip Reading Sentences 3 (LRS3), respectively. The results show that our proposed models raise the state-of-the-art performance by a large margin in audio-only, visual-only, and audio-visual experiments.

Index Terms— audio-visual speech recognition, end-to-end training, convolution-augmented transformer

1. INTRODUCTION

Audio-Visual Speech Recognition (AVSR) is the task of transcribing text from audio and visual streams, which has recently attracted a lot of research attention due to its robustness against noise. Since the visual stream is not affected by the presence of noise, an audio-visual model can lead to improved performance over an audio-only model as the level of noise increases.

Traditional audio-visual speech recognition methods follow a two-step approach, feature extraction and recognition [9, 26]. Several End-to-End (E2E) approaches have been recently presented by combining feature extraction and recognition inside a deep neural network, and this has led to a significant improvement in Visual Speech Recognition (VSR) and Automatic Speech Recognition (ASR), respectively. In VSR, Assael *et al.* [4] developed the first end-to-end network based on 3D convolution with Gated Recurrent Units (GRUs) for recognising visual speech on GRID [6]. Shillingford *et al.* [27] proposed an improved version of the model called Vision to Phoneme (V2P) which predicts phoneme distributions, instead of characters, from video clips. Chung and Zisserman [5] developed an attention-based sequence-to-sequence model for VSR in-the-wild. Zhang *et al.* [36] proposed a Temporal Focal block to capture temporal dynamics locally in a convolution-based sequence-to-sequence model. In ASR, [22, 35] have been recently shown to achieve better recognition performance by replacing the hand-crafted features such as log-Mel filter-bank features with deep representations from networks.

Several audio-visual approaches have been recently presented where pre-computed visual or audio features are used [1, 19, 25, 29, 34]. Afouras *et al.* developed a transformer-based sequence-to-sequence model by using pre-computed visual features and log-Mel filter-bank features as inputs. [19, 29, 34] focus on using video clips and log-Mel filter-bank features as inputs to train an audio-visual speech recognition model in an end-to-end manner. Few audiovisual studies are truly E2E, in the sense that they are trained with raw pixels and audio waveforms [17, 24]. In particular, [24] was applied only to word classification while [17] was tested on a constrained environment.

In this work, we extend our previous audio-visual model presented in [25] to an end-to-end model, which extracts features directly from raw pixels and audio waveform, and introduce a few changes which significantly improve the performance. In particular, we integrate the feature extraction stage with the hybrid CTC/attention back-end and train the model jointly. This results in a significant improvement in performance. We also replace the recurrent networks with conformers, which further push the state-of-the-art performance. Finally, we replace the RNN-based Language Model (RNN-LM) with a transformer-based LM which enhances the performance even more. We also perform a comparison between audio-only models trained with log-Mel filter-bank features and raw waveforms. Although in clean conditions they both perform similarly, the raw audio model performs slightly better in noisy conditions. We evaluate the proposed architecture on the largest in-the-wild audio-visual speech datasets, LRS2 and LRS3. The state-of-the-art performance is raised by a large margin for audio-only, visual-only and audio-visual experiments on both datasets, even outperforming methods trained on much larger external datasets.

2. DATASETS

For the purpose of this study, we use two large-scale publicy available audio-visual datasets, LRS2 [5] and LRS3 [3]. Both datasets are very challenging as there are large variations in head pose and illumination. LRS2 [5] consists of 224.1 hours with 144 482 video clips from BBC programs. In particular, there are 96 318 utterances for pre-training (195 hours), 45 839 for training (28 hours), 1 082 for validation (0.6 hours), and 1 243 for testing (0.5 hours).

LRS3 [3] collected from TED and TEDx talks is twice as large as the LRS2 dataset. LRS3 contains 151 819 utterances (438.9 hours). Specifically, there are 118 516 utterances in the pre-training set (408 hours), 31 982 utterances in the training-validation set (30 hours) and 1 321 utterances in the test set (0.9 hours).

3. ARCHITECTURE

The proposed architecture for audio-visual speech recognition is shown in Fig. 1. The encoder of the audio-visual model is comprised of three components, the front-end, the back-end, and the fusion modules, as explained below.

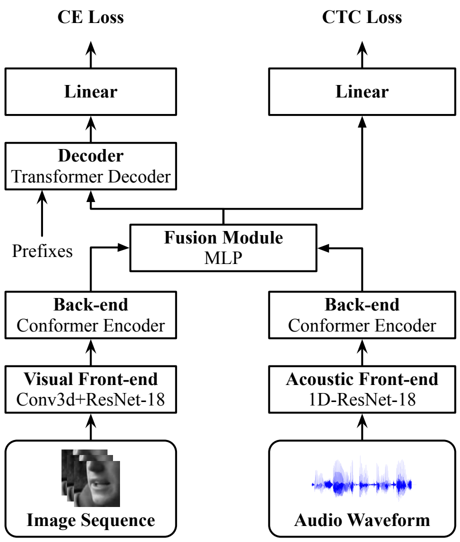

Fig. 1. End-to-end audio-visual speech recognition architecture. The inputs are pixels and raw audio waveforms.

Front-end The acoustic and visual front-ends architectures are shown in Table 1. For the visual stream, we use a modified ResNet-18 [11, 28] in which the first convolutional layer is replaced by a 3D convolutional layer with a kernel size $5 \times 7 \times 7$. The visual features at the end of the residual block are squeezed along the spatial dimension by a global average pooling layer. For the audio stream, we use a ResNet-18 based on 1D convolutional layers, where the filter size at the first convolutional layer is set to 80 (5ms). To downsample the time-scale, the stride is set to 2 at every block. The only exception is the first block, where we set the stride to 4. At the end of the front-end module, acoustic features are down-sampled to 25 frames per second so the match the frame rate of the visual features.

Back-end We use the recently proposed conformer encoder [10] as the back-end for temporal modeling. It is comprised of an embedding module, followed by a set of conformer blocks. In the embedding module, a linear layer projects the features from ResNet-18 to a d_k-dimensional space. The projected features are encoded with relative position information [7]. In each conformer block, a feed-forward module, a Multi-Head Self-Attention (MHSA) module, a convolutional module, and a feed-forward module are stacked in order. In particular, the feed-forward module is composed of a d^{ff}-dimensional linear layer, followed by Rectified Linear Units (ReLU), a dropout layer, and a second linear layer with an output size of d^k. The MHSA module receives queries Q, keys K, and values V as inputs, where $Q \in \mathbb{R}^{T\times d_k}$, $K \in \mathbb{R}^{T\times d_k}$, and $V \in \mathbb{R}^{T\times d_v}$, T denotes the sequence length and d_k and d_v are the dimensions for queries/keys and values, respectively. Suppose $Q = K = V$ in the encoder and W_i^Q, W_i^K and W_i^V are denoted as the weights of linear transformation for Q, K and V, respectively, the matrix of outputs at i-th head self-attention is computed through Scaled Dot-Product Attention [31]: $f_i(Q_i', K_i', V_i') = \text{softmax}(Q_i'K_i'^T)/d_k^{0.5}V_i'$, where $Q_i' = QW_i^Q, K_i' = KW_i^K, V_i' = VW_i^V$. The convolutional module contains a point-wise convolutional layer with an expansion factor of 2, followed by Gated Linear Units (GLU) [8], a temporal depth-wise convolutional layer, a batch normalisation layer, a swish activation layer, a point-wise convolutional layer, and a layer normalisation layer. This combination has been shown to improve ASR performance compared to the transformer architecture as it better captures temporal information locally and globally [10].

stage	Input audio waveform $(T_a \times 1)$	Input image sequence $(T_v \times W \times H)$
conv_1	$\text{conv1d}, 80, 64, \text{stride } 4$	$\text{conv3d}, 5 \times 7^2, 64, \text{stride } 1 \times 2^2$
		$\text{maxpool}, 1 \times 3^2$
res_2	$\begin{bmatrix}\text{conv1d}, 3, 64 \\ \text{conv1d}, 3, 64\end{bmatrix} \times 2$	$\begin{bmatrix}\text{conv2d}, 3^2, 64 \\ \text{conv2d}, 3^2, 64\end{bmatrix} \times 2$
res_3	$\begin{bmatrix}\text{conv1d}, 3, 128 \\ \text{conv1d}, 3, 128\end{bmatrix} \times 2$	$\begin{bmatrix}\text{conv2d}, 3^2, 128 \\ \text{conv2d}, 3^2, 128\end{bmatrix} \times 2$
res_4	$\begin{bmatrix}\text{conv1d}, 3, 256 \\ \text{conv1d}, 3, 256\end{bmatrix} \times 2$	$\begin{bmatrix}\text{conv2d}, 3^2, 256 \\ \text{conv2d}, 3^2, 256\end{bmatrix} \times 2$
res_5	$\begin{bmatrix}\text{conv1d}, 3, 512 \\ \text{conv1d}, 3, 512\end{bmatrix} \times 2$	$\begin{bmatrix}\text{conv2d}, 3^2, 512 \\ \text{conv2d}, 3^2, 512\end{bmatrix} \times 2$
pool_6	average pooling, stride 20	global average pooling

Table 1. The architecture of acoustic and visual Front-end. The dimensions of kernels are denoted by $\{\text{temporal size} \times \text{spatial size}^2, \text{channels}\}$. The acoustic model and visual backbones have 3.85 M and 11.18 M parameters, respectively. T_a and T_v denote the number of input samples and frames, respectively.

Fusion Layers The acoustic and visual features from the back-end modules are then concatenated and projected to d_k-dimensional space by an MLP. The MLP is composed of a linear layer with an output size of $4 \times d_k$ followed by a batch normalization layer, ReLU, and a final linear layer with an output dimension d_k.

Decoder We use the transformer decoder proposed in [31], which is composed of an embedding module, followed by a set of multi-head self-attention blocks. In the embedding module, a sequence of the prefixes from index 1 to $l - 1$ is projected to embedding vectors, where l is the target length index. The absolute positional encoding [31] is also added to the embedding. A self-attention block is comprised of two attention modules and a feed-forward module. Specifically, the first self-attention module uses $Q = K = V$ as input and future positions at its attention matrix are masked out. The second attention module uses the features from the previous self-attention module as Q and the representations from the encoder as K and V $(K = V)$. The component in the feed-forward module is the same as in the encoder.

Loss functions Let $\mathbf{x} = [x_1, ..., x_T]$ and $\mathbf{y} = [y_1, ..., y_L]$ be the input sequence and target symbols, respectively, with T and L representing the input and target lengths, respectively. Recent works in audio-visual speech recognition rely mostly on CTC [17] or attention-based models [1, 5] for audio-visual recognition. CTC loss assumes conditional independence between each output prediction and has a form of $p_{\text{CTC}}(\mathbf{y}|\mathbf{x}) \approx \prod_{t=1}^{T} p(y_t|\mathbf{x})$. An attention-based model gets rid of this assumption by directly estimating the posterior on the basis of the chain rule, which has a form of $p_{\text{CE}}(\mathbf{y}|\mathbf{x}) = \prod_{l=1}^{L} p(y_l|y_{<l}, \mathbf{x})$. In this work, we adopt a hybrid CTC/Attention architecture [32] to force monotonic alignments and at the same time get rid of the conditional independence assumption. The objective function is computed as follows:

$$\mathcal{L} = \alpha \log p_{\text{CTC}}(\mathbf{y}|\mathbf{x}) + (1-\alpha) \log p_{\text{CE}}(\mathbf{y}|\mathbf{x}) \tag{1}$$

where α controls the relative weight in CTC and attention mechanisms.

4. EXPERIMENTS

Pre-processing In each video, 68 facial landmarks are detected and tracked using dlib [14]. To remove differences related to rotation and scale, the faces are aligned to a neural reference frame using a similarity transformation. A bounding box of 96 × 96 is used to crop the mouth ROIs. The cropped patch is further converted to gray-scale and normalised with respect to the overall mean and variance on the training set. Each raw audio waveform is normalised by removing its mean and dividing by its standard deviation.

Data augmentation Following [20, 28], random cropping with a size of 88 × 88 and horizontal flipping with a probability of 0.5 are performed for each image sequence. For each audio waveform, additive noise, time masking, and band reject filtering are performed in the time domain. Babble noise from the NOISEX corpus [30] is added to the original audio clip with an SNR level from [-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB]. The selection of one of the noise levels or the use of a clean waveform is done using a uniform distribution. Similarly to [13], 2 sets of consecutive audio samples with a maximum length of 0.4 seconds are set to zero and 2 sets of consecutive frequency bands with a maximum width of 150 Hz are rejected. In audio-only experiments, we add speed perturbation by setting the speed between 0.9 and 1.1.

Experimental settings The network is initialised randomly, with the exception of the front-end modules in the encoder part, which in some experiments are initialised based on the publicly available pre-trained models on LRW [18] [1]. The back-end modules use a set of hyper-parameters ($e = 12$, $d^{\text{ff}} = 2048$, $d^{\text{k}} = 256$, $d^{\text{v}} = 256$), where e denotes the number of conformer blocks. The number of heads n^{head} is set to 4 in visual-only models and 8 in audio-only/audio-visual models, respectively. Kernel size is set to 31 in each depth-wise convolutional layer. The transformer decoder uses 6 self-attention blocks, where the hyper-parameters settings in feed-forward and self-attention modules are the same as in the encoder. The Adam optimizer [15] with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ is used for end-to-end training with a mini-batch size of 8. Following [31], the learning rate increases linearly with the first 25 000 steps, yielding a peak learning rate of 0.0004 and thereafter decreases proportionally to the inverse square root of the step number. The whole network is trained for 50 epochs. Note that the utterances with more than 600 frames in the pre-training set are excluded during training.

Language Model We train a transformer-based language model [12] for 10 epochs. The language model is trained by combining the training transcriptions of LibriSpeech (960 h) [21], pre-training and training sets of LRS2 [5] and LRS3 [3], with a total of 16.2 million words. The weighted prior score from the language model is incorporated through a shallow fusion, which is described in Eq. 2.

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\hat{\mathcal{Y}}}{\operatorname{argmax}} \{\lambda \log p_{\text{CTC}}(\mathbf{y}|\mathbf{x}) + (1-\lambda)\log p_{\text{CE}}(\mathbf{y}|\mathbf{x}) + \beta \log p_{\text{LM}}(\mathbf{y})\} \quad (2)$$

where $\hat{\mathcal{Y}}$ is a set of predictions of target symbols. λ is a relative CTC weight at the decoding phase, and β is the relative weight for the language model. In our work, we set λ to 0.1 and β to 0.6, respectively.

5. RESULTS

Ablation Studies In this section, we investigate the impact of each change on the baseline hybrid CTC/Attention model [25]. Results on LRS2 are shown in Table 2. We first train a model from scratch in an end-to-end manner, resulting in an absolute improvement of 12.6 % over the two stage approach, where visual features are first extracted and then fed to the back-end. We initialise the visual front-end with a model pre-trained on LRW and a further absolute improvement of 4.7 % is observed. Then, we replace the LSTM encoders and decoders with a conformer encoder and a transformer decoder, respectively, which results in an absolute improvement of 3.8 %. We also replace the RNN-based language model with a transformer-based language model and achieve a WER of 37.9 %. This leads to an absolute improvement of 4.5 %.

Method	WER
Baseline [24]	63.5
+ E2E	50.9
+ LRW pre-training	46.2
+ Conformer encoder	42.4
+ Transformer LM	37.9

Table 2. Ablation study on visual speech recognition performance on LRS2.

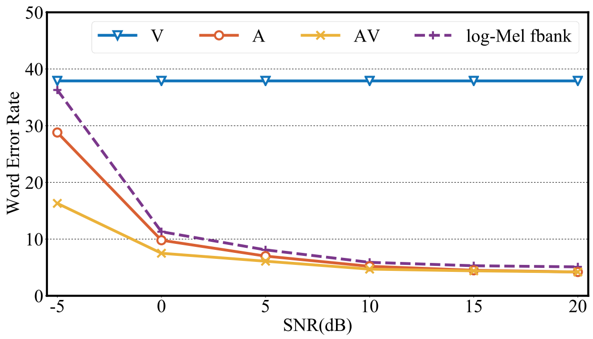

Fig. 2. Word Error Rate (WER) as a function of the noise level. A: End-to-End audio model. V: End-to-End visual model, AV: End-to-End audio-visual model. log-Mel filter-bank: A conformer model trained with log-Mel filter-bank features.

Results on LRS2 Results on LRS2 are reported in Table 3. The proposed visual-only model reduces the WER from 48.3 % to 39.1 %, while using 6 × fewer training data [1]. In case, we use the pre-trained LRW model for initialisation the WER drops further to 37.9 %. The E2E audio-only model using audio waveforms for training achieves a WER of 4.3 %, resulting in an absolute improvement of 2.4 %. over the current state-of-the-art. For comparison purposes, we also run an experiment using 80-dimension log-Mel filter-bank features following [25, 32]. Similarly to the WavAugment [13], we augment the log-Mel filter-bank features via SpecAugment [23]. By replacing the raw audio features with the log-Mel filter-bank features, we observe the same performance, WER 4.3 %, which indicates deep acoustic speech representations based on the proposed temporal network can be directly learnt from audio waveforms. To better investigate their differences, we conduct noisy experiments varying different levels of babble noise. The results are shown in Fig. 2. It is interesting to observe that the performance of the raw audio model slightly outperforms the log-Mel filter-bank based over varying levels of babble noise with a maximum absolute margin of 7.5 % at -5 dB. This indicates deep speech representations are more robust to noise than the log-Mel filter-bank features. In case, we initialise the audio encoder with a model pretrained on LRW then the WER drops to 3.9 %.

It is evident that the audio-visual model which directly learns from audio waveforms and raw pixels leads to a small improvement over the audio-only models. We also run audio-only, visual-only, and audio-visual experiments varying the SNR levels of babble noise.

[1] Pre-trained LRW models are available at `https://sites.google.com/view/audiovisual-speech-recognition`

Method	Training Data (Hours)	WER
Visual-only (↓)		
MV-WAS [5]	LRS2 (224)	70.4
LIBS [37]	MVLRS (730) + LRS2 (224)	65.3
CTC/Attention [25]	LRW (157) + LRS2 (224)	63.5
Conv-seq2seq [36]	LRW (157) + LRS2&3$^{v0.0}$ (698)	51.7
KD + CTC [2]	VC2^{clean} (334) + LRS2&3$^{v0.4}$ (632)	51.3
TDNN [34]	LRS2 (224)	48.9
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	48.3
Ours (V)	LRS2 (224)	**39.1**
Ours (V)	LRW (157) + LRS2 (224)	**37.9**
Audio-only (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	9.7
CTC/Attention [25]	LRS2 (224)	8.3
CTC/Attention [16]	LibriSpeech (960) + LRS2 (224)	8.2
TDNN [34]	LRS2 (224)	6.7
Ours (filter-bank)	LRS2 (224)	**4.3**
Ours (raw A)	LRS2 (224)	**4.3**
Ours (raw A)	LRW (157) + LRS2 (224)	**3.9**
Audio-visual (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	8.5
CTC/Attention [25]	LRW (157) + LRS2 (224)	7.0
TDNN [34]	LRS2 (224)	5.9
Ours (raw A + V)	LRS2 (224)	**4.2**
Ours (raw A + V)	LRW (157) + LRS2 (224)	**3.7**

Table 3. Word Error Rate (WER) of the audio-only, visual-only and audio-visual models on LRS2. VC2^{clean} denotes the filtered version of VoxCeleb2. LRS2&3 consists of LRS2 and LRS3. LRS3$^{v0.4}$ is the updated version of LRS3 with speaker-independent settings.

The results are shown in Fig 2. Note that both audio-only and audio-visual models are augmented with noise injection. It is clear that the audio-visual model achieves better performance than the audio-only model. The gap between raw audio-only and audio-visual models becomes larger by the presence of high level of noise. This demonstrates that the audio-visual model is particularly beneficial when the audio modality is heavily corrupted by background noise.

Results on LRS3 Results on LRS3$^{v0.4}$ are reported in Table 4. The best visual-only model has a WER of 43.3 %. We observe that our visual-only model outperforms other methods by a large margin while using fewer training data. For the audio-only and audio-visual experiments, our model pushes the state-of-the-art performance to 2.3 % and 2.3 %, respectively, outperforming [19] by 2.5 % and 2.2 %, respectively. It is worth pointing out that our model is trained on a dataset which is 52× smaller than [19], 595 vs 31000 hours.

We should note that some works use the old version of LRS3 (denoted as v0.0), where some speakers appear both in the training and test sets. For fair comparisons, we also report the performance of audio-only, visual-only, and audio-visual model on this version of LRS3 as well. Specifically, the audio-only model achieves a WER to 1.3 %. The visual-only model reduces the WER to 30.4 %. The audio-visual model reduces the WER to 1.2 % which is the new state-of-the-art performance for this set. These significant improvements over LRS3$^{v0.4}$ are mainly due to the fact that in LRS3$^{v0.0}$ overlapped identities appear in both pre-training and test sets.

Method	Training Data (Hours)	WER
Visual-only (↓)		
Conv-seq2seq [36]	LRW (157) + LRS2&3$^{v0.0}$ (698)	60.1
KD + CTC [2]	VC2^{clean} (334) + LRS3$^{v0.4}$ (438)	59.8
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	58.9
EG-seq2seq [33]	LRW (157) + LRS3$^{v0.0}$ (474)	57.8
V2P [27]	YT (3 886)	55.1
RNN-T [19]	YT (31 000)	33.6
Ours (V)	LRS3$^{v0.4}$ (438)	**46.9**
Ours (V)	LRW (157) + LRS3$^{v0.4}$ (438)	**43.3**
Ours (V)	LRW (157) + LRS3$^{v0.0}$ (474)	**30.4**
Audio-only (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	8.3
EG-seq2seq [33]	LRS3$^{v0.0}$ (474)	7.2
RNN-T [19]	YT (31 000)	4.8
Ours (filter-bank)	LRS3$^{v0.4}$ (438)	**2.3**
Ours (raw A)	LRS3$^{v0.4}$ (438)	**2.3**
Ours (raw A)	LRW (157) +LRS3$^{v0.4}$ (438)	**2.3**
Ours (raw A)	LRW (157) +LRS3$^{v0.0}$ (474)	**1.3**
Audio-visual (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	7.2
EG-seq2seq [33]	LRW (157) + LRS3$^{v0.0}$ (474)	6.8
RNN-T [19]	YT (31 000)	4.5
Ours (raw A + V)	LRW (157) + LRS3$^{v0.4}$ (438)	**2.3**
Ours (raw A + V)	LRW (157) + LRS3$^{v0.0}$ (474)	**1.2**

Table 4. Word Error Rate (WER) of the audio-only, visual-only and audio-visual models on LRS3. VC2^{clean} denotes the filtered version of VoxCeleb2. LRS2&3 consists of LRS2 and LRS3. LRS3$^{v0.4}$ is the updated version of LRS3 with speaker-independent settings.

6. CONCLUSIONS

In this work, we present an encoder-decoder attention-based architecture for audio-visual speech recognition, which can be trained in an end-to-end fashion and leads to state-of-the-art results on LRS2 and LRS3. Additionally, the audio-visual experiments show that the audio-visual model significantly outperforms the audio-only model especially at high levels of noise. It would also be interesting to investigate in future work an adaptive fusion mechanism that learns to weigh each modality based on the noise levels.

Acknowledgements. We would like to thank Dr. Jie Shen for his help with face tracking. The work of Pingchuan Ma has been partially supported by Honda and “AWS Cloud Credits for Research”.

7. REFERENCES

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” *IEEE PAMI*, 2018.

[2] T. Afouras, J. S. Chung, and A. Zisserman, “ASR is all you need: Cross-modal distillation for lip reading,” in *ICASSP*, 2020, pp. 2143–2147.

[3] T. Afouras, J. S. Chung, and A. Zisserman, “LRS3-TED: A large-scale dataset for visual speech recognition,” *arXiv preprint arXiv:1809.00496*, 2018.

[4] Y. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, “Lipnet: End-to-end sentence-level lipreading,” *arXiv preprint arXiv:1611.01599*, 2016.

[5] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *CVPR*, 2017, pp. 3444–3453.

[6] 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, pp. 2421–2424, 2006.

[7] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V Le, *et al.*, “Transformer-XL: Attentive language models beyond a fixed-length context,” in *ACL*, 2019, 2978–2988.

[8] Y. N Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *ICML*, 2017, 933–941.

[9] 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.

[10] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” in *Interspeech*, 2020, pp. 5036–5040.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.

[12] K. Irie, A. Zeyer, R. Schlüter, and H. Ney, “Language modeling with deep transformers,” in *Interspeech*, 2019, pp. 3905–3909.

[13] E. Kharitonov, M. Rivière, G. Synnaeve, L. Wolf, P. Mazaré, *et al.*, “Data augmenting contrastive learning of speech representations in the time domain,” *CoRR*, vol. abs/2007.00991, 2020.

[14] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[15] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

[16] C. Li and Y. Qian, “Listen, watch and understand at the cocktail party: Audio-visual-contextual speech separation,” in *Interspeech*, 2020, pp. 1426–1430.

[17] P. Ma, S. Petridis, and M. Pantic, “Investigating the lombard effect influence on end-to-end audio-visual speech recognition,” in *Interspeech*, 2019, pp. 4090–4094.

[18] P. Ma, B. Martínez, S. Petridis, and M. Pantic, “Towards practical lipreading with distilled and efficient models,” *CoRR*, vol. abs/2007.06504, 2020.

[19] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, *et al.*, “Recurrent neural network transducer for audio-visual speech recognition,” in *ASRU*, 2019, pp. 905–912.

[20] B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in *ICASSP*, 2020, pp. 6319–6323.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *ICASSP*, 2015, pp. 5206–5210.

[22] T. Parcollet, M. Morchid, and G. Linares, “E2e-sincnet: Toward fully end-to-end speech recognition,” in *ICASSP*, 2020, pp. 7714–7718.

[23] D. S Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, *et al.*, “SpecAugment: A simple data augmentation method for automatic speech recognition,” in *Interspeech*, 2019, pp. 2613–2617.

[24] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, *et al.*, “End-to-end audiovisual speech recognition,” in *ICASSP*, 2018, pp. 6548–6552.

[25] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, “Audio-visual speech recognition with a hybrid CTC/attention architecture,” in *SLT*, 2018, pp. 513–520.

[26] 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.

[27] B. Shillingford, Y. Assael, M. W Hoffman, T. Paine, C. Hughes, *et al.*, “Large-scale visual speech recognition,” in *Interspeech*, 2019, pp. 4135–4139.

[28] T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” in *Interspeech*, vol. 9, 2017, pp. 3652–3656.

[29] G. Sterpu, C. Saam, and N. Harte, “How to teach DNNs to pay attention to the visual modality in speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1052–1064, 2020.

[30] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, *et al.*, “Attention is all you need,” in *NIPS*, 2017, pp. 6000–6010.

[32] S. Watanabe, T. Hori, S. Kim, J. R Hershey, and T. Hayashi, “Hybrid CTC/attention architecture for end-to-end speech recognition,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[33] B. Xu, C. Lu, Y. Guo, and J. Wang, “Discriminative multimodality speech recognition,” in *CVPR*, 2020, pp. 14 433–14 442.

[34] J. Yu, S. Zhang, J. Wu, S. Ghorbani, B. Wu, *et al.*, “Audio-visual recognition of overlapped speech for the LRS2 dataset,” in *ICASSP*, 2020, pp. 6984–6988.

[35] N. Zeghidour, N. Usunier, G. Synnaeve, R. Collobert, and E. Dupoux, “End-to-end speech recognition from the raw waveform,” in *Interspeech*, 2018, pp. 781–785.

[36] X. Zhang, F. Cheng, and S. Wang, “Spatio-temporal fusion based convolutional sequence learning for lip reading,” in *ICCV*, 2019, pp. 713–722.

[37] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, *et al.*, “Hearing lips: Improving lip reading by distilling speech recognizers,” in *AAAI*, 2020, pp. 6917–6924.