

Audio-Visual Efficient Conformer for Robust Speech Recognition

Maxime Burchi, Radu Timofte

Computer Vision Lab, CAIDAS, IFI, University of Würzburg, Germany

{maxime.burchi, radu.timofte}@uni-wuerzburg.de

Abstract

End-to-end Automatic Speech Recognition (ASR) systems based on neural networks have seen large improvements in recent years. The availability of large scale hand-labeled datasets and sufficient computing resources made it possible to train powerful deep neural networks, reaching very low Word Error Rate (WER) on academic benchmarks. However, despite impressive performance on clean audio samples, a drop of performance is often observed on noisy speech. In this work, we propose to improve the noise robustness of the recently proposed Efficient Conformer Connectionist Temporal Classification (CTC)-based architecture by processing both audio and visual modalities. We improve previous lip reading methods using an Efficient Conformer back-end on top of a ResNet-18 visual front-end and by adding intermediate CTC losses between blocks. We condition intermediate block features on early predictions using Inter CTC residual modules to relax the conditional independence assumption of CTC-based models. We also replace the Efficient Conformer grouped attention by a more efficient and simpler attention mechanism that we call patch attention. We experiment with publicly available Lip Reading Sentences 2 (LRS2) and Lip Reading Sentences 3 (LRS3) datasets. Our experiments show that using audio and visual modalities allows to better recognize speech in the presence of environmental noise and significantly accelerate training, reaching lower WER with 4 times less training steps. Our Audio-Visual Efficient Conformer (AVEC) model achieves state-of-the-art performance, reaching WER of 2.3% and 1.8% on LRS2 and LRS3 test sets. Code and pretrained models are available at <https://github.com/burchim/AVEC>.

1. Introduction

End-to-end Automatic Speech Recognition based on deep neural networks has become the standard of state-of-the-art approaches in recent years [25, 47, 18, 16, 17, 31, 7]. The availability of large scale hand-labeled datasets and sufficient computing resources made it possible to train power-

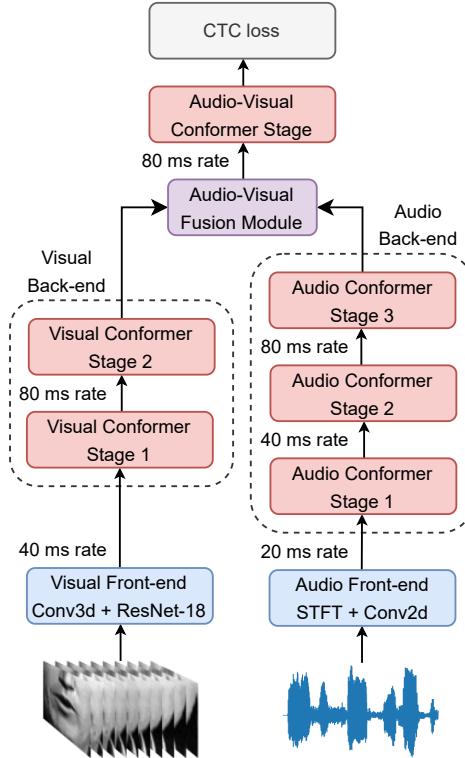


Figure 1: **Audio-Visual Efficient Conformer architecture.** The model is trained end-to-end using CTC loss and takes raw audio waveforms and lip movements from the speaker as inputs.

ful deep neural networks for ASR, reaching very low WER on academic benchmarks like LibriSpeech [34]. Neural architectures like Recurrent Neural Networks (RNN) [15, 19], Convolution Neural Networks (CNN) [10, 28] and Transformers [12, 23] have successfully been trained from raw audio waveforms and mel-spectrograms audio features to transcribe speech to text. Recently, Gulati *et al.* [16] proposed a convolution-augmented transformer architecture (Conformer) to model both local and global dependencies using convolution and attention to reach better speech recognition performance. Concurrently, Nozaki *et al.* [33]

improved CTC-based speech recognition by conditioning intermediate encoder block features on early predictions using intermediate CTC losses [14]. Burchi *et al.* [7] also proposed an Efficient Conformer architecture using grouped attention for speech recognition, lowering the amount of computation while achieving better performance. Inspired from computer vision backbones, the Efficient Conformer encoder is composed of multiple stages where each stage comprises a number of Conformer blocks to progressively downsample and project the audio sequence to wider feature dimensions.

Yet, even if these audio-only approaches are breaking the state-of-the-art, one major pitfall for using them in the real-world is the rapid deterioration of performance in the presence of ambient noise. In parallel to that, Audio Visual Speech Recognition (AVSR) has recently attracted a lot of research attention due to its ability to use image processing techniques to aid speech recognition systems. Preceding works have shown that including the visual modality of lip movements could improve the robustness of ASR systems with respect to noise while reaching better recognition performance [41, 42, 36, 1, 45, 29]. Xu *et al.* [45] proposed a two-stage approach to first separate the target voice from background noise using the speakers lip movements and then transcribe the filtered audio signal with the help of lip movements. Petridis *et al.* [36] uses a hybrid architecture, training an LSTM-based sequence-to-sequence (S2S) model with an auxiliary CTC loss using an early fusion strategy to reach better performance. Ma *et al.* [29] uses Conformer back-end networks with ResNet-18 [20] front-end networks to improve recognition performance.

Other works focus on Visual Speech Recognition (VSR), only using lip movements to transcribe spoken language into text [4, 9, 48, 3, 49, 37, 30]. An important line of research is the use of cross-modal distillation. Afouras *et al.* [3] and Zhao *et al.* [49] proposed to improve the lip reading performance by distilling from an ASR model trained on a large-scale audio-only corpus while Ma *et al.* [30] uses prediction-based auxiliary tasks. Prajwal *et al.* [37] also proposed to use sub-words units instead of characters to transcribe sequences, greatly reducing running time and memory requirements. Also providing a language prior, reducing the language modelling burden of the model.

In this work we focus on the design of a noise robust speech recognition architecture processing both audio and visual modalities. We use the recently proposed CTC-based Efficient Conformer architecture [7] and show that including the visual modality of lip movements can successfully improve noise robustness while significantly accelerating training. Our Audio-Visual Efficient Conformer (AVEC) reaches lower WER using 4 times less training steps than its audio-only counterpart. Moreover, we are the first work to apply intermediate CTC losses between

blocks [27, 33] to improve visual speech recognition performance. We show that conditioning intermediate features on early predictions using Inter CTC residual modules allows to close the gap in WER between autoregressive and non-autoregressive AVSR systems based on S2S. This also helps to counter a common failure case which is that audio-visual models tend to ignore the visual modality. In this way, we force pre-fusion layers to learn spatiotemporal features. Finally, we replace the Efficient Conformer grouped attention by a more efficient and simpler attention mechanism that we call patch attention. Patch attention reaches similar performance to grouped attention while having a lower complexity. The contributions of this work are as follows:

- We improve the noise robustness of the recently proposed Efficient Conformer architecture by processing both audio and visual modalities.
- We condition intermediate Conformer block features on early predictions using Inter CTC residual modules to relax the conditional independence assumption of CTC models. This allows us to close the gap in WER between autoregressive and non-autoregressive methods based on S2S.
- We propose to replace the Efficient Conformer grouped attention by a more efficient and simpler attention mechanism that we call patch attention. Patch attention reaches similar performance to grouped attention with a lower complexity.
- We experiment on publicly available LRS2 and LRS3 datasets and reach state-of-the-art results using audio and visual modalities.

2. Method

In this section, we describe our proposed Audio-Visual Efficient Conformer network. The model is composed of 4 main components: An audio encoder, a visual encoder, an audio-visual fusion module and an audio-visual encoder. The audio and visual encoder are separated into modality specific front-end networks to transform each input modality into temporal sequences. Efficient Conformer back-end networks are used to model local and global temporal relationships. The full model is trained end-to-end using intermediate CTC losses between Conformer blocks in addition of the output CTC layer. The complete architecture of the model is shown in Figure 1.

2.1. Model Architecture

Audio front-end. The audio front-end network first transforms raw audio wave-forms into mel-spectrograms using a short-time Fourier transform computed over windows of 20ms with a step size of 10ms. 80-dimensional

mel-scale log filter banks are applied to the resulting frequency features. The mel-spectrograms are processed by a 2D convolution stem to extract local temporal-frequency features, resulting in a 20ms frame rate signal. The audio front-end architecture is shown in Table 1.

Table 1: Audio Front-end architecture, 1.2 Millions parameters. T_a denotes the input audio sample length.

Stage	Layers	Output Shape
Fourier Transf	STFT: 400 window length 160 hop length, 512 ffts	(257, $T_a // 160 + 1$)
Mel Scale	Mel Scale: 80 mels	(80, $T_a // 160 + 1$)
Stem	Conv2d: 3^2 , 180 filters, 2^2 stride	(180, 40, $T_a // 320 + 1$)
Proj	Linear, 180 units	($T_a // 320 + 1$, 180)

Visual front-end. The visual front-end network [29] transforms input video frames into temporal sequences. A 3D convolution stem with kernel size $5 \times 7 \times 7$ is first applied to the video. Each video frame is then processed independently using a 2D ResNet-18 [20] with an output spatial average pooling. Temporal features are then projected to the back-end network input dimension using a linear layer. The visual front-end architecture is shown in Table 2.

Table 2: Visual Front-end architecture, 11.3 Millions parameters. T_v denotes the number of input video frames.

Stage	Layers	Output Shape
Stem	Conv3d: 5×7^2 , 64 filters, 1×2^2 stride MaxPoo3d: 1×3^2 , 1×2^2 stride	(64, T_v , 22, 22)
Res 1	$2 \times$ [Conv2d: 3^2 , 64 filters Conv2d: 3^2 , 64 filters]	(T_v , 64, 22, 22)
Res 2	$2 \times$ [Conv2d: 3^2 , 128 filters Conv2d: 3^2 , 128 filters]	(T_v , 128, 11, 11)
Res 3	$2 \times$ [Conv2d: 3^2 , 256 filters Conv2d: 3^2 , 256 filters]	(T_v , 256, 6, 6)
Res 4	$2 \times$ [Conv2d: 3^2 , 512 filters Conv2d: 3^2 , 512 filters]	(T_v , 512, 3, 3)
Pool	Global Average Pooling	(T_v , 512)
Proj	Linear, 256 units	(T_v , 256)

Back-end networks. The back-end networks use an Efficient Conformer architecture. The Efficient Conformer encoder was proposed in [7], it is composed of several stages where each stage comprises a number of Conformer blocks [16] using grouped attention with relative positional encodings. The temporal sequence is progressively downsampled using strided convolutions and projected to wider feature dimensions, lowering the amount of computation while achieving better performance. We use 3 stages in the audio back-end network to downsample the audio signal to a 80 milliseconds frame rate. Only 2 stages are necessary to downsample the visual signal to the same frame rate. Table 6 shows the hyper-parameter of each back-end network.

Table 3: Back-end networks hyper-parameters. *InterCTC blocks* indicates Conformer blocks having a post Inter CTC residual module.

Network	Visual Back-end	Audio Back-end	Audio-Visual Encoder
Num Params (M)	13.6	17.9	15.9
Num Stages	2	3	1
Blocks per Stage	6, 1	5, 6, 1	5
Total Num Blocks	7	12	5
Stage Feature Dim	256, 360	180, 256, 360	360
Conv Kernel Size	15	15	15
Stage Patch Size	1, 1	3, 1, 1	1
InterCTC Blocks	3, 6	8, 11	2

Audio-visual fusion module. Similar to [36, 29], we use an early fusion strategy to learn audio-visual features and reduce model complexity. The acoustic and visual features from the back-end networks are concatenated and fed into a joint feed-forward network. The concatenated features of size $2 \times d_{model}$ are first expanded using a linear layer with output size $d_{ff} = 4 \times d_{model}$, passed through a Swish activation function [38] and projected back to the original feature dimension d_{model} .

Audio-visual encoder. The audio-visual encoder is a single stage back-end network composed of 5 Conformer blocks without downsampling. The encoder outputs are then projected to a CTC layer to maximize the sum of probabilities of correct target alignments.

2.2. Patch Attention.

The Efficient Conformer [7] proposed to replace Multi-Head Self-Attention (MHSA) [44] in earlier encoder layers with grouped attention. Grouped MHSA reduce attention complexity by grouping neighbouring temporal elements along the feature dimension before applying scaled dot-product attention. Attention having a quadratic computational complexity with respect to the sequence length, this caused the network to have an asymmetric complexity with earlier attention layers requiring more flops than latter layers with shorter sequence length. In this work, we propose to replace grouped attention with a simpler and more efficient attention mechanism that we call patch attention (Figure 2). Similar to the pooling attention proposed by the Multiscale Vision Transformer (MViT) [13] for video and image recognition, the patch attention proceed to an average

Table 4: Attention variants complexities including query, key, value and output linear projections. n and d are the sequence length and feature dimension respectively.

Attention Variant	Hyper Parameter	Full Attention Complexity
Regular	-	$O(n \cdot d^2 + n^2 \cdot d)$
Grouped	Group Size (g)	$O(n \cdot d^2 + (n/g)^2 \cdot d \cdot g)$
Patch	Patch Size (k)	$O(n/k \cdot d^2 + (n/k)^2 \cdot d)$

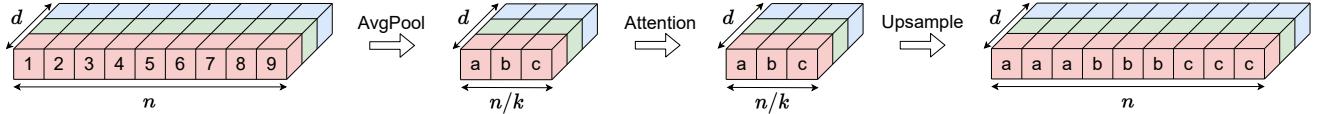


Figure 2: Patch Multi-Head Self-Attention. The input sequence is downsampled using an average pooling before applying multi-head self-attention. The output sequence is then upsampled via nearest neighbor upsampling, reducing attention complexity from $O(n^2 \cdot d)$ to $O((n/k)^2 \cdot d)$ where k defines the pooling / upsampling kernel size. Patch attention is equivalent to regular attention when $k = 1$.

pooling on the input sequence before projection the query, key and values.

$$X = \text{AvgPooling1d}(X_{in}) \quad (1)$$

$$\text{with } Q, K, V = XW^Q, XW^K, XW^V \quad (2)$$

Where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$ are query, key and value linear projections parameter matrices. MHSA with relative sinusoidal positional encoding is then performed at lower resolution as:

$$MHSA(X) = \text{Concat}(O_1, \dots, O_H) W^O \quad (3)$$

$$\text{with } O_h = \text{softmax} \left(\frac{Q_h K_h^T + S_h^{rel}}{\sqrt{d_h}} \right) V_h \quad (4)$$

Where $S^{rel} \in \mathbb{R}^{n \times n}$ is a relative position score matrix that satisfy $S^{rel}[i, j] = Q_i E_{j-i}^T$. E is the linear projection of a standard sinusoidal positional encoding matrix with positions ranging from $-(n_{max} - 1)$ to $(n_{max} - 1)$. The attention output sequence is then projected and up-sampled back to the initial resolution using nearest neighbor up-sampling.

$$X_{out} = \text{UpsampleNearest1d}(MHSA(X)) \quad (5)$$

In consequence, each temporal element of the same patch produce the same attention output. Local temporal relationships are only modeled in the convolution modules while global relationships are modeled by patch attention. We use 1-dimensional patches in this work but patch attention

could also be generalized to image and video data using 2D and 3D patches. We leave this to future works. The computational complexity of each attention variant is shown in Table 4. Path attention further reduce complexity compared to grouped attention by decreasing the amount of computation needed by Query, Key, Value and Output fully connected layers while keeping the feature dimension unchanged. Similar to previous work [7], we only use patch attention in the first audio back-end stage to reduce complexity while maintaining model recognition performance. Figure 3 shows the amount of FLOPs for each attention module variant with respect to encoded sequence length n and model feature dimension d . Using patch or grouped attention variants instead of regular MHSA greatly reduce the amount of FLOPs in the first audio back-end stage.

2.3. Intermediate CTC Predictions.

Inspired by [27] and [33] who proposed to add intermediate CTC losses between encoder blocks to improve CTC-based speech recognition performance, we add Inter CTC residual modules (Figure 4) in encoder networks. We condition intermediate block features of both audio, visual and audio-visual encoders on early predictions to relax the conditional independence assumption of CTC models. During both training and inference, each intermediate prediction is summed to the input of the next layer to help recognition. We use the same method proposed in [33] except that we do not share layer parameters between losses. The l^{th} block output X_l^{out} is passed through a feed-forward network with residual connection and a softmax activation function:

$$Z_l = \text{Softmax}(\text{Linear}(X_l^{out})) \quad (6)$$

$$X_{l+1}^{in} = X_l^{out} + \text{Linear}(Z_l) \quad (7)$$

Where $Z_l \in \mathbb{R}^{T \times V}$ is a probability distribution over the output vocabulary. The intermediate CTC loss is then computed using the target sequence y as:

$$L_l^{inter} = -\log(P(y|Z_l)) \quad (8)$$

$$\text{with } P(y|Z_l) = \sum_{\pi \in \mathcal{B}_{CTC}^{-1}(y)} \prod_{t=1}^T Z_{t, \pi_t} \quad (9)$$

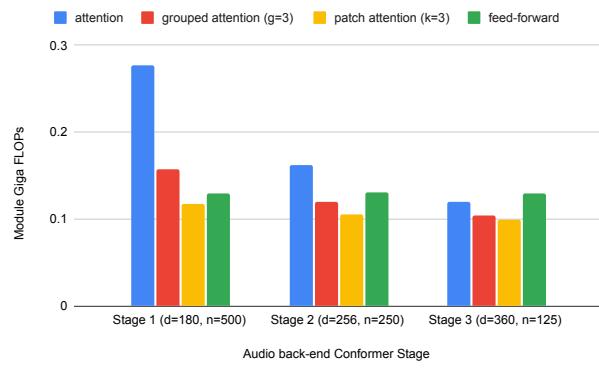


Figure 3: Audio-only back-end modules FLOPs (Billion).

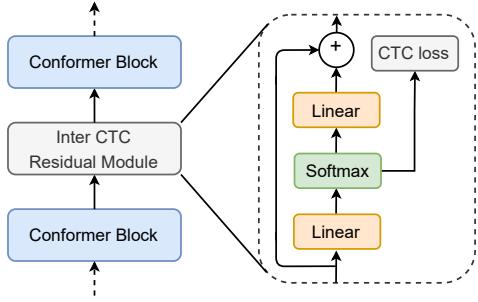


Figure 4: **Inter CTC residual module.** Intermediate predictions are summed to the input of the next Conformer block to condition the prediction of the final block on it. Intermediate CTC losses are added to the output CTC loss for the computation of the final loss.

Where $\pi \in V^T$ are paths of tokens and \mathcal{B}_{CTC} is a many-to-one map that simply removes all blanks and repeated labels from the paths. The total training objective is defined as follows:

$$L = (1 - \lambda)L^{CTC} + \lambda L^{inter} \quad (10)$$

$$\text{with } L^{inter} = \frac{1}{K} \sum_{k \in interblocks} L_k^{inter} \quad (11)$$

Where *interblocks* is the set of blocks having a post Inter CTC residual module (Figure 4). Similar to [33], we use Inter CTC residual modules every 3 Conformer blocks with λ set to 0.5 in every experiments.

3. Experiments

3.1. Datasets

We use 3 publicly available AVSR datasets in this work. The Lip Reading in the Wild (LRW) [8] dataset is used for visual pre-training and the Lip Reading Sentences 2 (LRS2) [1] and Lip Reading Sentences 3 (LRS3) [2] datasets are used for training and evaluation.

LRW dataset. LRW is an audio-visual word recognition dataset consisting of short video segments containing a single word out of a vocabulary of 500. The dataset comprise 488,766 training samples with at least 800 utterances per class and a validation and test sets of 25,000 samples containing 50 utterances per class.

LRS2 & LRS3 datasets. The LRS2 dataset is composed of 224.1 hours with 144,482 video clips from the BBC television whereas the LRS3 dataset consists of 438.9 hours with 151,819 video clips extracted from TED and TEDx talks. Both datasets include corresponding subtitles with word alignment boundaries and are composed of a pre-train split, train-val split and test split. LRS2 has 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). Whereas LRS3 has 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). All videos contain a single speaker, have a 224×224 pixels resolution and are sampled at 25 fps with 16kHz audio.

3.2. Implementation Details

Pre-processing Similar to [29], we remove differences related to rotation and scale by cropping the lip regions using bounding boxes of 96×96 pixels to facilitate recognition. The RetinaFace [11] face detector and Face Alignment Network (FAN) [6] are used to detect 68 facial landmarks. The cropped images are then converted to gray-scale and normalised between -1 and 1 . Facial landmarks of the LRW, LRS2 and LRS3 datasets are obtained from previous work [30] and reused for pre-processing to get a clean comparison of the methods. A byte-pair encoding tokenizer is built from LRS2&3 pre-train and trainval splits transcripts using sentencepiece [26]. We use a vocabulary size of 256 including the CTC blank token following preceding works on CTC-based speech recognition [31, 7].

Data augmentation Spec-Augment [35] is applied on the audio mel-spectrograms during training to prevent overfitting with two frequency masks with mask size parameter $F = 27$ and five time masks with adaptive size $pS = 0.05$. Similarly to [30], we mask videos on the time axis using one mask per second with the maximum mask duration set to 0.4 seconds. Random cropping with size 88×88 and horizontal flipping are also performed for each video during training. We also follow Prajwal *et al.* [37] using central crop with horizontal flipping at test time for visual-only experiments.

Training Setup We first pre-train the visual encoder on the LRW dataset [8] using cross-entropy loss to recognize words being spoken. The visual encoder is pre-trained for 30 epochs and front-end weights are then used as initialization for training. Audio and visual encoders are trained on the LRS2&3 datasets using a Noam schedule [44] with 10k warmup steps and a peak learning rate of 1e-3. We use the Adam optimizer [24] with $\beta_1 = 0.9$, $\beta_2 = 0.98$. L2 regularization with a 1e-6 weight is also added to all the trainable weights of the model. We train all models with a global batch size of 256 on 4 GPUs, using a batch size of 16 per GPU with 4 accumulated steps. Nvidia A100 40GB GPUs are used for visual-only and audio-visual experiments while RTX 2080 Ti are used for audio-only experiments. The audio-only models are trained for 200 epochs while visual-only and audio-visual models are trained for 100 and 70 epochs respectively. Note that we only keep videos shorter than 400 frames (16 seconds) during training. Finally, we average models weights over the last 10 epoch checkpoints using Stochastic Weight Averaging [22] before evaluation.

Table 5: Comparison of WER (%) on LRS2 / LRS3 test sets with recently published methods using publicly and non-publicly available datasets for Audio-Only (AO), Visual-Only (VO) and Audio-Visual (AV) models.

Method	Model Criterion	Training Datasets	Total Hours	test WER		
				AO	VO	AV
(↓) Using Publicly Available Datasets (↓)						
Petridis <i>et al.</i> [36]	CTC+S2S	LRW, LRS2	381	8.3 / -	63.5 / -	7.0 / -
Zhang <i>et al.</i> [48]	S2S	LRW, LRS2&3	788 / 790	-	51.7 / 60.1	-
Afouras <i>et al.</i> [3]	CTC	VoxCeleb2 ^{clean} , LRS2&3	1,032 / 808	-	51.3 / 59.8	-
Xu <i>et al.</i> [45]	S2S	LRW, LRS3	595	- / 7.2	- / 57.8	- / 6.8
Yu <i>et al.</i> [46]	LF-MMI	LRS2	224	6.7 / -	48.9 / -	5.9 / -
Ma <i>et al.</i> [29]	CTC+S2S	LRW, LRS2&3	381 / 595	3.9 / 2.3	37.9 / 43.3	3.7 / 2.3
Prajwal <i>et al.</i> [37]	S2S	LRS2&3	698	-	28.9 / 40.6	-
Ma <i>et al.</i> [30]	CTC+S2S	LRW, LRS2&3	818	-	27.3 / 34.7	-
Ours	CTC	LRW, LRS2&3	818	2.8 / 2.1	32.6 / 39.2	2.5 / 1.9
+ Neural LM	CTC	LRW, LRS2&3	818	2.4 / 2.0	29.8 / 37.5	2.3 / 1.8
(↓) Using Non-Publicly Available Datasets (↓)						
Afouras <i>et al.</i> [1]	S2S	MVLRs, LRS2&3	1,395	9.7 / 8.3	48.3 / 58.9	8.5 / 7.2
Zhao <i>et al.</i> [49]	S2S	MVLRs, LRS2	954	-	65.3 / -	-
Shillingford <i>et al.</i> [40]	CTC	LRVSR	3,886	-	- / 55.1	-
Makino <i>et al.</i> [32]	Transducer	YouTube-31k	31,000	- / 4.8	- / 33.6	- / 4.5
Serdyuk <i>et al.</i> [39]	Transducer	YouTube-90k	91,000	-	- / 25.9	- / 2.3
Prajwal <i>et al.</i> [37]	S2S	MVLRs, TEDx _{ext} , LRS2&3	2,676	-	22.6 / 30.7	-
Ma <i>et al.</i> [30]	CTC+S2S	LRW, AVSpeech, LRS2&3	1,459	-	25.5 / 31.5	-

Language Models. Similarly to [28], we experiment with a N-gram [21] statistical language model (LM) and a Transformer neural language model. A 6-gram LM is used to generate a list of hypotheses using beam search and an external Transformer LM is used to rescore the final list. The 6-gram LM is trained on the LRS2&3 pre-train and train-val transcriptions. Concerning the neural LM, we pre-train a 12 layer GPT-3 Small [5] on the LibriSpeech LM corpus for 0.5M steps using a batch size of 0.1M tokens and finetune it for 10 epochs on the LRS2&3 transcriptions.

3.3. Results

Table 5 compares WERs of our Audio-Visual Efficient Conformer with state-of-the-art methods on the LRS2 and LRS3 test sets. Our Audio-Visual Efficient Conformer achieves state-of-the-art performances with WER of 2.3%/1.8%. On the visual-only track, our CTC model competes with most recent autoregressive methods using S2S criterion. We were able to recover similar results but still lack behind *Ma et al.* [30] which uses auxiliary losses with pre-trained audio-only and visual-only networks. We found our audio-visual network to converge faster than audio-only experiments, reaching better performance using 4 times less training steps. The intermediate CTC losses of the visual encoder could reach lower levels than in visual-only experiments showing that optimizing audio-visual layers can help pre-fusion layers to learn better representations.

3.4. Ablation Studies

We propose a detailed ablation study to better understand the improvements in terms of complexity and WER brought by each architectural modification. We report the number of operations measured in FLOPs (number of multiply-and-add operations) for the network to process a ten second audio/video clip. Inverse Real Time Factor (Inv RTF) is also measured on the LRS3 test set by decoding with a batch size 1 on a single Intel Core i7-12700 CPU thread. All ablations were performed by training audio-only models for 200 epochs and visual-only / audio-visual models for 50 epochs.

Efficient Conformer Visual Back-end. We improve the recently proposed visual Conformer encoder [29] using an Efficient Conformer back-end network. The use of byte pair encodings for tokenization instead of characters allows us to further downsample temporal sequences without impacting the computation of CTC loss. Table 6 shows that using an Efficient Conformer back-end network for our visual-only model leads to better performances while reducing model complexity and training time. The number of model parameters is also slightly decreased.

Table 6: Ablation study on visual back-end network.

Visual Back-end	#Params (Million)	LRS2 test	LRS3 test	#FLOPs (Billion)	Inv RTF
Conformer	43.0	39.53	47.14	87.94	5.17
Eff Conf	40.4	37.39	44.96	84.52	5.26

Reference

the authors looked at papers written over a 10 year period and hundreds had to be thrown out

Outputs

Block 3: the otho looing pa people we over s any your per and conndries that aboutent threghow

Block 6: the autthherss looking paperss we overai year paiod and hundreds that about thrououtow

Block 9: the authors looked at papers witen over ainght year period and hundreds that to been throw out

Block 12: the authors looked at papers written over 10 year period and hundreds had to be thrown out

Figure 5: Output example of our Visual-only model using greedy search decoding on the LRS3 test set with intermediate CTC prediction every 3 blocks. The sentence is almost correctly transcribed except for the missing ‘a’ before ‘10 year’.

Inter CTC residual modules. Similar to [33], we experiment adding Inter CTC residual modules between blocks to relax the conditional independence assumption of CTC. Table 7 shows that using intermediate CTC losses every 3 Conformer blocks greatly helps to reduce WER, except for the audio-only setting where this do not improve performance. Figure 5 gives an example of intermediate block predictions decoded using greedy search without an external language model on the test set of LRS3. We can see that the output is being refined in the encoder layers by conditioning on the intermediate predictions of previous layers. Since our model refines the output over the frame-level predictions, it can correct insertion and deletion errors in addition to substitution errors. We further study the impact of Inter CTC on multi-modal learning by measuring the performance of our audio-visual model when one of the two modalities is masked. As pointed out by preceding works [8, 1, 32], networks with multi-modal inputs can often be dominated by one of the modes. In our case speech recognition is a significantly easier problem than lip reading which can cause the model to ignore visual information. Table 8 shows that Inter CTC can help to counter this problem by forcing pre-fusion layers to transcribe the input signal.

Table 7: Ablation study on Inter CTC residual modules.

Model Back-end	#Params (Million)	LRS2 test	LRS3 test	#FLOPs (Billion)	Inv RTF
<i>Audio-only (↓)</i>					
Eff Conf	31.5	2.83	2.13	7.54	51.98
+ Inter CTC	32.1	2.84	2.11	7.67	50.30
<i>Visual-only (↓)</i>					
Eff Conf	40.4	37.39	44.96	84.52	5.26
+ Inter CTC	40.9	33.82	40.63	84.60	5.26
<i>Audio-visual (↓)</i>					
Eff Conf	60.9	2.87	2.54	90.53	4.84
+ Inter CTC	61.7	2.58	1.99	90.66	4.82

Table 8: Impact of Inter CTC on audio-visual model WER (%) for LRS2 / LRS3 test sets in a masked modality setting.

Inter CTC	Audio-Visual Eval Mode		
	masked video	masked audio	no mask
No	4.48 / 3.22	52.77 / 59.10	2.87 / 2.54
Yes	3.39 / 2.38	37.62 / 46.55	2.58 / 1.99

Patch multi-head self-attention. We experiment replacing grouped attention by patch attention in the first audio encoder stage. Our objective being to increase the model efficiency and simplicity without arming performance. Grouped attention was proposed in [7] to reduce attention complexity for long sequences in the first encoder stage. Table 9 shows the impact of each attention variant on our audio-only model performance and complexity. We start with an Efficient Conformer (M) [7] and replace the attention mechanism. We find that grouped attention can be replaced by patch attention without a loss of performance using a patch size of 3 in the first back-end stage.

Table 9: Ablation study on audio back-end attention.

Attention Type	Group / Patch Size	LRS2 test	LRS3 test	#FLOPs (Billion)	Inv RTF
Regular	-	2.85	2.12	8.66	49.86
Grouped	3, 1, 1	2.82	2.13	8.06	50.27
Patch	3, 1, 1	2.83	2.13	7.54	51.98

3.5. Noise Robustness

We measure model noise robustness using various types of noise and compare our Audio-Visual Efficient Conformer with recently published methods. Figure 6 shows the WER evolution of audio-only (AO), visual-only (VO) and audio-visual (AV) models with respect to multiple Signal to Noise Ratio (SNR) using white noise and babble noise from the NoiseX corpus [43]. We find that processing both audio and visual modalities can help to significantly improve speech recognition robustness with respect to babble noise. Moreover, we also experiment adding babble noise during training as done in previous works [36, 29] and find that it can further improve noise robustness at test time.

Robustness to various types of noise. We gather various types of recorded audio noise including sounds and music. In Table 10, we observe that the Audio-Visual Efficient Conformer consistently achieves better performance than its audio-only counterpart in the presence of various noise types. This confirm our hypothesis that the audio-visual model is able to use the visual modality to aid speech recognition when audio noise is present in the input.

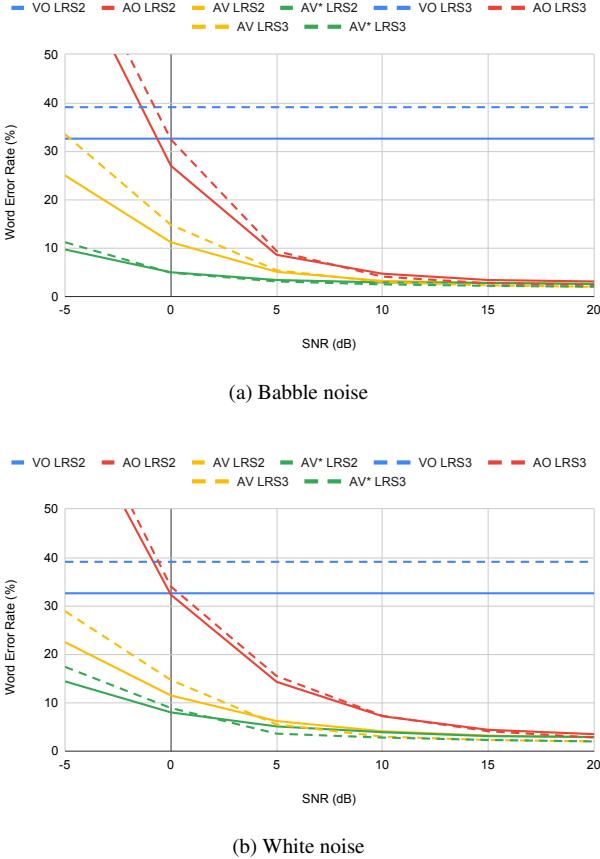


Figure 6: **LRS2 and LRS3 test WER (%) as a function of SNR (dB).** * indicates experiments being trained with babble noise. We measure noise robustness by evaluating our models in the presence of babble and white noise.

Table 10: LRS3 test WER (%) as a function of SNR (dB).

Noise	Mode	SNR (dB)				
		-5	0	5	10	15
babble	AO	75.9	32.4	9.3	4.1	2.7
	AV	33.5	14.8	5.4	3.0	2.0
	AV*	11.2	4.9	3.1	2.5	2.0
white	AO	77.6	34.0	15.5	7.3	4.1
	AV	28.9	14.7	5.5	3.0	2.0
	AV*	17.4	8.9	3.6	2.8	2.0
birds	AO	51.8	23.9	10.9	5.9	3.7
	AV	21.6	11.5	6.2	4.1	2.9
	AV*	15.9	8.3	4.9	3.4	2.4
chainsaw	AO	82.9	41.2	14.8	5.5	3.7
	AV	37.8	17.3	7.6	3.9	2.6
	AV*	25.8	10.8	5.0	3.2	2.4
jazz	AO	25.3	9.7	4.1	3.1	2.6
	AV	13.9	6.0	3.2	2.4	2.0
	AV*	10.6	4.2	2.8	2.4	2.0
street raining	AO	58.4	23.8	8.9	4.6	3.0
	AV	27.12	10.8	5.7	3.1	2.7
	AV*	15.9	6.9	3.8	2.7	2.2
washing dishes	AO	47.8	24.5	11.5	6.0	3.7
	AV	21.3	11.5	6.1	3.6	2.8
	AV*	14.2	7.3	4.3	2.2	2.6
train	AO	51.3	18.6	7.0	4.0	2.9
	AV	23.1	10.1	4.7	3.0	2.4
	AV*	14.5	6.2	3.5	2.6	2.2

Comparison with other methods. We compare our method with results provided by Ma *et al.* [29] and Petridis *et al.* [36] on the LRS2 test set. Table 11 show that our audio-visual model achieves lower WER in the presence of babble noise, reaching WER of 9.7% at -5 dB SNR against 16.3% for Ma *et al.* [29].

Table 11: Comparison with Ma *et al.* [29]. LRS2 test WER (%) as a function of SNR (dB) using babble noise.

Method	Mode	SNR (dB)				
		-5	0	5	10	15
Ma <i>et al.</i> [29]	VO	37.9	37.9	37.9	37.9	37.9
	AO*	28.8	9.8	7	5.2	4.5
	AV*	16.3	7.5	6.1	4.7	4.4
Ours	VO	32.6	32.6	32.6	32.6	32.6
	AO	70.5	27	8.6	4.7	3.4
	AV	25	11.2	5.1	3.2	2.8
	AV*	9.7	5	3.4	2.9	2.8

Table 12: Comparison with Petridis *et al.* [36]. LRS2 test WER (%) as a function of SNR (dB) using white noise.

Method	Mode	SNR (dB)				
		-5	0	5	10	15
Petridis <i>et al.</i> [36]	VO	63.5	63.5	63.5	63.5	63.5
	AO*	85.0	45.4	19.6	11.7	9.4
	AV*	55.0	26.1	13.2	9.4	8.0
Ours	VO	32.6	32.6	32.6	32.6	32.6
	AO	73.1	32.3	14.3	7.2	4.4
	AV	22.5	11.5	6.2	4.1	3.2
	AV*	14.4	8.0	5.1	3.9	3.1

4. Conclusion

In this paper, we proposed to improve the noise robustness of the recently proposed Efficient Conformer CTC-based architecture by processing both audio and visual modalities. We showed that incorporating multi-scale CTC losses between blocks could help to improve recognition performance, reaching comparable results to most recent autoregressive lip reading methods. We also proposed patch attention, a simpler and more efficient attention mechanism to replace grouped attention in the first audio encoder stage. Our Audio-Visual Efficient Conformer achieves state-of-the-art performance of 2.3% and 1.8% on the LRS2 and LRS3 test sets. In the future, we would like to explore other techniques to further improve the noise robustness of our model and close the gap between recent lip reading methods. This includes adding various audio noises during training and using cross-modal distillation with pre-trained models. We also wish to reduce the visual front-end network complexity without arming recognition performance and experiment with the RNN-Transducer learning objective for streaming applications.

Acknowledgments

This work was partly supported by The Alexander von Humboldt Foundation (AvH).

References

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

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Asr is all you need: Cross-modal distillation for lip reading. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2143–2147. IEEE, 2020.

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

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.

[7] Maxime Burchi and Valentin Vielzeuf. Efficient conformer: Progressive downsampling and grouped attention for automatic speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE, 2021.

[8] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian conference on computer vision*, pages 87–103. Springer, 2016.

[9] Joon Son Chung and AP Zisserman. Lip reading in profile. 2017.

[10] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[11] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[12] Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888. IEEE, 2018.

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[14] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, pages 369–376, 2006.

[15] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee, 2013.

[16] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[17] Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE, 2021.

[18] Wei Han, Zhengdong Zhang, Yu Zhang, Jiahui Yu, Chung-Cheng Chiu, James Qin, Anmol Gulati, Ruoming Pang, and Yonghui Wu. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. *arXiv preprint arXiv:2005.03191*, 2020.

[19] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197, 2011.

[22] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[23] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.

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

[25] Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Joelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE, 2020.

[26] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detok-

enizer for neural text processing. In *EMNLP*, pages 66–71, 2018.

[27] Jaesong Lee and Shinji Watanabe. Intermediate loss regularization for ctc-based speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE, 2021.

[28] Jason Li, Vitaly Lavrukhin, Boris Ginsburg, Ryan Leary, Oleksii Kuchaiev, Jonathan M Cohen, Huyen Nguyen, and Ravi Teja Gadde. Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*, 2019.

[29] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.

[30] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *arXiv preprint arXiv:2202.13084*, 2022.

[31] Somshubra Majumdar, Jagadeesh Balam, Oleksii Hrinchuk, Vitaly Lavrukhin, Vahid Noroozi, and Boris Ginsburg. Citrinet: Closing the gap between non-autoregressive and autoregressive end-to-end models for automatic speech recognition. *arXiv preprint arXiv:2104.01721*, 2021.

[32] 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 *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019.

[33] Jumon Nozaki and Tatsuya Komatsu. Relaxing the conditional independence assumption of ctc-based asr by conditioning on intermediate predictions. *arXiv preprint arXiv:2104.02724*, 2021.

[34] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[35] Daniel S Park, Yu Zhang, Chung-Cheng Chiu, Youzheng Chen, Bo Li, William Chan, Quoc V Le, and Yonghui Wu. Specaugment on large scale datasets. In *ICASSP*, pages 6879–6883, 2020.

[36] Stavros Petridis, Themis Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520. IEEE, 2018.

[37] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172, 2022.

[38] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[39] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth 32x32x8 voxels. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 796–802. IEEE, 2021.

[40] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cian Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrainy Bennett, et al. Large-scale visual speech recognition. *arXiv preprint arXiv:1807.05162*, 2018.

[41] Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6447–6456, 2017.

[42] George Sterpu, Christian Saam, and Naomi Harte. Attention-based audio-visual fusion for robust automatic speech recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 111–115, 2018.

[43] 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*, 12(3):247–251, 1993.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[45] Bo Xu, Cheng Lu, Yandong Guo, and Jacob Wang. Discriminative multi-modality speech recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14433–14442, 2020.

[46] Jianwei Yu, Shi-Xiong Zhang, Jian Wu, Shahram Ghorbani, Bo Wu, Shiyin Kang, Shansong Liu, Xunying Liu, Helen Meng, and Dong Yu. Audio-visual recognition of overlapped speech for the lrs2 dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6984–6988. IEEE, 2020.

[47] 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 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE, 2020.

[48] Xingxuan Zhang, Feng Cheng, and Shilin Wang. Spatio-temporal fusion based convolutional sequence learning for lip reading. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 713–722, 2019.

[49] Ya Zhao, Rui Xu, Xinchao Wang, Peng Hou, Haihong Tang, and Mingli Song. Hearing lips: Improving lip reading by distilling speech recognizers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6917–6924, 2020.