Watch to Listen Clearly: Visual Speech Enhancement Driven Multi-modality Speech Recognition

Bo Xu, Jacob Wang, Cheng Lu and Yandong Guo
Xpeng motors
xiaoboboer@gmail.com

Abstract

Multi-modality (talking face video and audio) information helps improve speech recognition performance compared to the single modality. In noisy environments, the effect of audio modality is weakened, which further affects the performance of multi-modality speech recognition (MSR). Most of the MSR methods use noisy audio signal as input of the audio modality without any enhancement (filtering the noisy components in the audio signal). In this paper, we propose an audio-enhanced multi-modality speech recognition model. In particular, the proposed model consists of two sub-networks, one is the visual speech enhancement (VE) sub-network and the other is the multi-modality speech recognition (MSR) sub-network. The VE sub-network is able to separate a speaker's voice from background noises when given the corresponding talking face to enhance audio modality. Then the audio modality together with video modality are fed into the MSR sub-network to produce characters. We introduce a pseudo-3D residual network (P3D)-based visual front-end to extract more advantageous visual features. The MSR sub-network is built on top of the Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU) architecture which is more effective than Transformer in long sequences. We demonstrate the effectiveness of audio enhancement for MSR by extensive experiments. The proposed method surpasses the state-of-the-art MSR models on the LRS3-TED dataset and the LRW dataset.

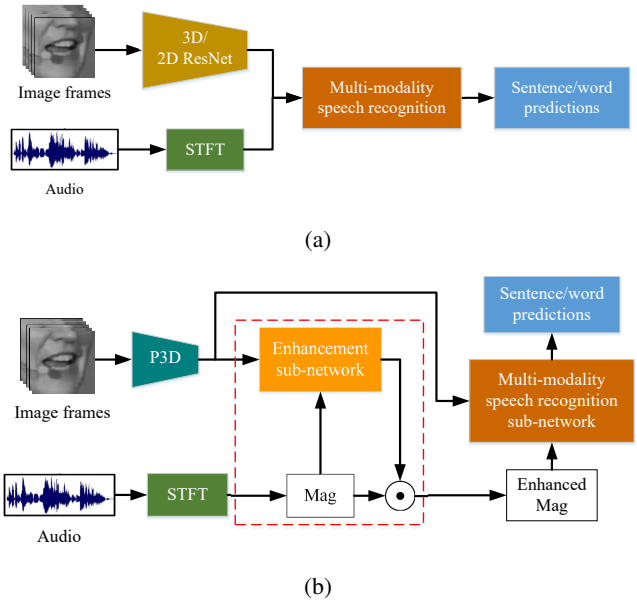


Figure 1: **a)** Outline of multi-modality speech recognition (MSR) pipeline used in [2, 14]. **b)** Outline of the visual speech enhancement driven multi-modality speech recognition (VE-MSR) pipeline. We add VE sub-network (inside the red box) which can separate target audio signal from background noises to enhance audio modality. Enhanced audio modality can also improve the performance of MSR. We use pseudo-3d residual network (P3D) [36] as visual front-end instead of 3D CNN and 2D ResNet. The P3D network has more advantages on extracting spatio-temporal features from videos.

1. Introduction

Lip reading is an approach to interpret what people say by looking at the movements of lips when they are talking [21, 33, 46]. Some people with hearing impairments use this technique to communicate with others [7, 16]. Lip reading is a difficult skill for human to grasp and requires numerous practice [19, 41]. The concept of visual speech recognition based on lip reading is proposed by Sumby in 1954 [42], where visual observation of the speaker's lip movements is examined as a indicator to test oral speech intelligibility. Subsequently, several machine lip reading models are proposed [17, 31, 34], but they suffer difficulties on extracting spatio-temporal features from the video.

In recent years, lip reading automation becomes possible to achieve due to rapid development of deep neural network, especially in computer vision [30, 40, 44]. Large scale train-

ing datasets also accelerate the development of lip reading in deep learning [14, 15, 17, 18, 38, 47]. In addition to serving as a power solution to hearing impairment, lip reading can also contribute to audio speech recognition (ASR) in audio adversary environments, such as high noise level where human speaking is inaudible. Multi-modality (video and audio) is more effective than single modality (video or audio) in terms of robustness and performance. Visual speech enhancement (VE) and multi-modality speech recognition (MSR) are two main extended applications of multi-modality. In [2] there is a significant deterioration in performance for MSR in noisy environments. Compared to audio modality in a clean voice environment, the one in noisy environment improves less on the performance of MSR. For example in [2], the word error rate (WER) is reduced from 59.9% for only video modality to 8.0% for multi-modality, equivalent to more than 50% reduction. In noisy environment of 0 dB SNR, the WER is 44.3% for multi-modality, only reduced by 15.6%. The results demonstrate that the noisy level of audio modality directly affects the performance improvement of MSR.

VE is able to strengthen the audio signal which only contains the target speaker's voice with the contribution of lip reading information (video modality). Then the enhanced audio single is used to recognize speech [3]. Our proposed method is to separate target audio signal from background noises before MSR. By combining VE and MSR, our method is more advantageous in terms of robustness and performance than any other MSR method currently proposed.

In this paper we propose a deep neural network model named visual speech enhancement driven multi-modality speech recognition (VE-MSR) for MSR, combining VE sub-network and MSR sub-network. Before being fed into the MSR sub-network, audio modality is enhanced by separating voice of target speaker from background noises through the VE sub-network. The architecture of our VE network is proposed based on [6]. Instead of 3D CNN and 2D ResNet, we use the pseudo-3D residual network (P3D) [36] as visual front-end which can extract more effective visual feature representations. We replace the encoder layer of Bi-LSTM [39] with an Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU) layer [48], which can adaptively modulate the input as a fine granularity, paying different of attention to different elements, resulting higher performance in spatio-temporal tasks than the recurrent neural networks (RNN) [37] and its variants [27, 39, 12]. The MSR sub-network is also built on top of EleAtt-GRU. The enhanced audio stream and its corresponding video stream are then fed into the MSR sub-network, outputting speech predictions.

Overall, the contributions of this paper are:

- We propose a new multi-modality speech recognition model, which reconstructs cleaner audio modality by incorporating visual features to improve the performance of MSR.
- We introduce the P3D as visual front-end to extract more advantageous visual feature representations and EleAtt-GRU to adaptively encode the spatio-temporal information in VE and MSR sub-networks, benefiting performance of VE and MSR.
- By extensive experiments, we demonstrate that VE-MSR surpasses state-of-the-art MSR model [2] both in audio clean and noisy environments on the Lip Reading Sentences 3 (LRS3-TED) dataset [5]. The word classification model we build based on P3D outperforms the word-level state-of-the-art [41] on the Lip Reading in the Wild (LRW) dataset [15].

2. Related works

In this section, we introduce some related works about visual speech enhancement (VE) and multi-modality speech recognition (MSR).

2.1. Visual speech enhancement

Various works have proposed to enhance audio speech with the help of visual information. Gabbay *et al*. use a trained silent-video-to-speech model [20] to generate speech predictions as a mask on the noisy input audio [22]. The noisy audio is not used in the pipeline of speech enhancement. [28] proposed a VE network based on convolution neural networks (CNNs) and fully connected (FC) layers to generate enhanced speech and reconstruct lip image frames. Aviv *et al*. also use CNNs to encode multi-modality features, which unify the embedding between audio and video before audio decoder [23]. They use transposed convolution in audio decoder to enhance mel-scale spectrograms. Afouras *et al*. use 1D ResNet temporal convolution networks to encode audio and video data individually, then concatenate multi-modality features [3]. They also use temporal convolution networks to decode multi-modality features into a mask to remove noisy components in the audio. Subsequently they proposed a new approach that replaces the multi-modality feature decoder with Bi-LSTM [6].

2.2. Multi-modality speech recognition

MSR is the integration of lip reading and audio speech recognition (ASR). Lip reading can contribute to ASR results, especially in noisy environments. Reciprocally, ASR can strengthen the lip reading and benefit people with hearing impairments.

Various methods in deep learning have been proposed for lip reading [49]. [7] proposed LipNet, an end-to-end model making use of spatio-temporal convolutions,

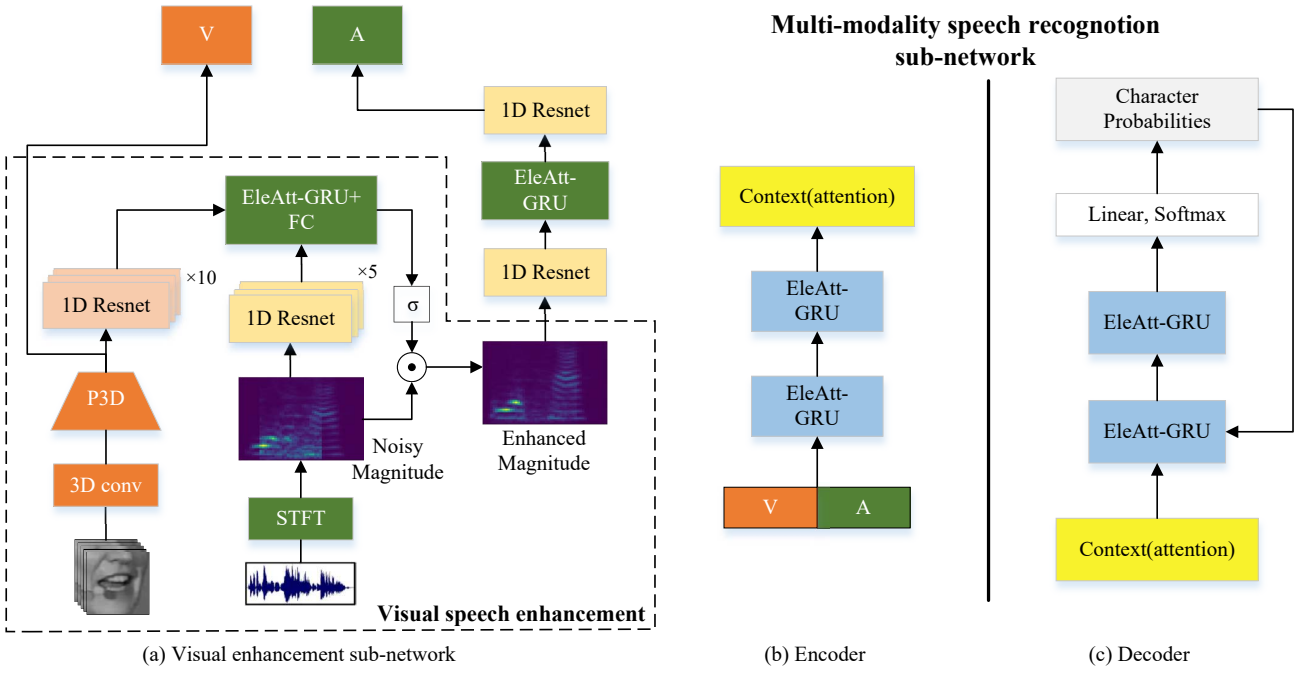


Figure 2: Architecture of the VE-MSR network. The network consists of two sub-networks, VE and MSR. **a) VE sub-network:** the mouth ROI image sequence is fed into a spatio-temporal network to extract the visual features and the audio features are extracted though STFT to audio signal. We use video stream and audio stream to produce visual and audio feature vectors, and the multi-modality vectors are combined and fed into the EleAttGRU layer which outputs a multiplicative mask that filters the noisy spectrograms. **V**– video feature vectors, **A**– enhanced audio feature vectors. **b) Encoder:** the enhanced magnitude spectrogram is fed into an audio stream to extract enhanced audio features. The enhanced audio features and the visual features extracted by the visual spatio-temporal network are combined and encoded by the EleAttGRU encoder of the MSR sub-network. **c) Decoder:** the decoder of MSR sub-network consists of two EleAttGRU layers and produces character probabilities. The output of previous decoding step is fed back into decoder to process the next decoding step.

LSTM [27] and connectionist temporal classification (CTC) loss on variable-length sequence of video frames. LipNet [7] achieves 95.2% accuracy in sentence-level on the GRID corpus [32]. [41] proposes an architecture by combining spatio-temporal convolution, residual and Bi-LSTM, they achieved 83.0% accuracy in word-level on LRW dataset [15].

Encoder-to-decoder (enc2dec) architecture has been developed for speech recognition and machine translation [9, 10, 24, 25, 43, 45]. Chung *et al*. use a dual sequence-to-sequence model with enc2dec (attention) mechanism, one for the video frame features and the other for the audio features [14]. Combining audio and visual information, they achieve 97.0% accuracy on GRID and 76.2% accuracy on LRW [15]. Afouras *et al*. propose a sequence-to-sequence model based on transformer self-attention architecture [2]. They also use encoder-to-decoder mechanism and concatenate the context vectors produced by multiple modalities after multi-head attention in decoder stage. The model transcribing spoken sentences to characters, by using sequence-to-sequence loss achieves 7.2% word error rates (WER) on LRS3-TED dataset and 8.5% WER on LRS2 dataset. With noisy audio signal it achieves 42.5% WER on LRS3-TED dataset and 34.2% WER on LRS2 dataset. Obviously, without prior speech enhancement, the results of MSR are not ideal, this is the main reason why we propose the method of VE-MSR. In this paper, we qualitatively evaluate performance of the VE-MSR model for speech recognition in the noisy environments.

3. Architectures

The visual speech enhancement driven multi-modality speech recognition (VE-MSR) network architecture consists of visual speech enhancement (VE) sub-network and multi-modality speech recognition (MSR) sub-network. The model architecture is shown in detail in Figure 2. We propose a VE sub-network for separating the speech of target speaker from background noises. The network contains

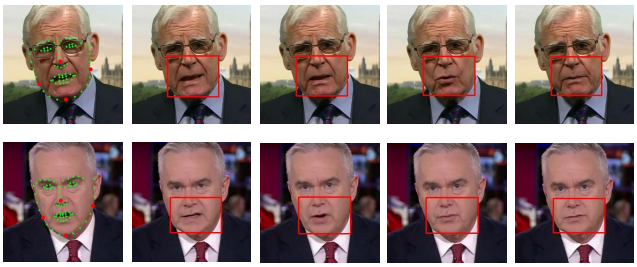


Figure 3: Examples of mouth crop. Face Landmarks are extracted by the *Dlib* toolkit and the mouth region of interest (ROI) inside the red squares are achieved by four specified reference landmarks.

two inputs (video frames and original audio spectrogram) and outputs a spectrogram of the enhanced speech. Video feature and enhanced speech spectrogram are then fed into the multi-modality speech recognition (MSR) sub-network to recognize to full words or sentences.

3.1. Video features

As shown in Figure 3, we use 4 (red points) out of 68 facial landmarks (red and green points) extracted by *Dlib* [29] and crop image frames to 112×112 pixels pitch including the mouth as region of interest (ROI). Instead of using ResNet [26]-based video feature extraction networks mentioned in many other lip-reading papers [2, 3, 4, 6, 14, 41], we use a pseudo-3D (P3D) [36] network to produce a more powerful visual spatio-temporal representation.

A few studies [41, 2, 3, 35, 6] have shown that performing 3D convolutions is a beneficial method to capture both temporal and spatial dimensions in videos. However, the computation cost and memory demand of a deep 3D CNN is very expensive. P3D alleviates this situation by simulating $N \times N \times N$ convolutions with $1 \times 3 \times 3$ convolutions filters on spatial domain (like 2D CNN) plus $3 \times 1 \times 1$ convolutions to extract temporal information on adjacent features along time. [36] devises three variants of bottleneck building blocks in a residual framework and achieves superior performances over several state-of-the-art techniques in several different tasks. The three block version is illustrated in Figure 4. We implement a 50-layer P3D network by mixing the three units, as illustrated in Figure 4d.

The spatio-temporal convolution network consists of a 3D convolution layer with 64 filters of kernel size $5 \times 7 \times 7$, followed by batch normalization, ReLU units and max-pooling layers. And then the max-pooling is followed by a 50-layer P3D ResNet that gradually decreases the spatial dimensions with depth while remaining the temporal dimension. For an input of $T \times H \times W$ frames, the output of the sub-network is a $T \times 512$ tensor (in the final stage, the feature is average-pooled in spatial dimension and processed as

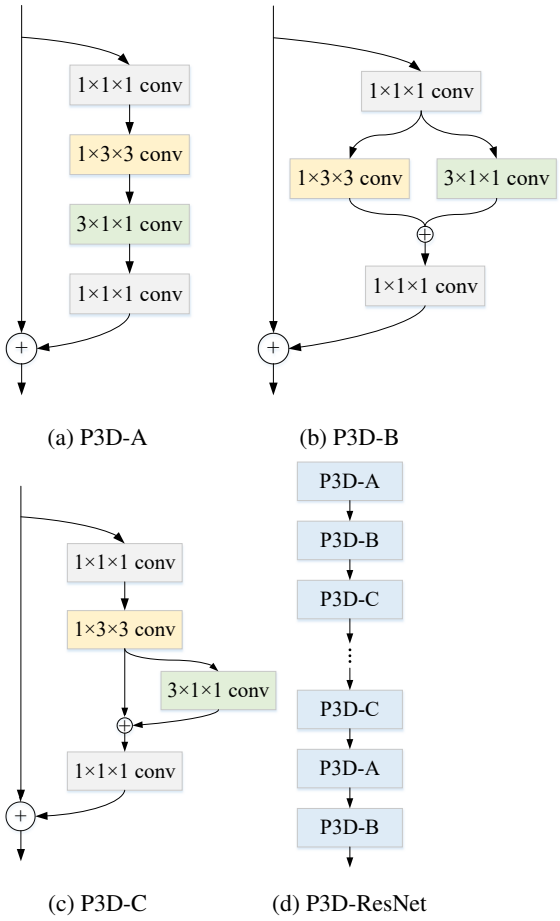


Figure 4: Bottleneck building blocks of the Pseudo-3D (P3D) and P3D ResNet network. P3D ResNet is produced by interleaving P3D-A, P3D-B and P3D-C.

a 512-dimensional vector representing each video frame).

3.2. Audio features

We use Short Time Fourier Transform (STFT) to extract magnitude spectrogram from the waveform signal at a sample rate of 16kHz. To align with the video frame rate at 25fps, we set the STFT window length to 40ms and hop length to 10ms when sample rate is 16kHz. We convert the resulting magnitude with frequency bins of 321 (representing frequencies ranging from 0 to 8 kHz) into mel-scale magnitude spectrogram with mel-frequency bins of 80. The magnitude spectrogram and corresponding video feature are then fed into VE sub-network.

3.3. Visual speech enhancement network

Previous studies have demonstrated the effect of audio signals with different noise levels on MSR. VE can improve not only the performance of ASR, but also the performance

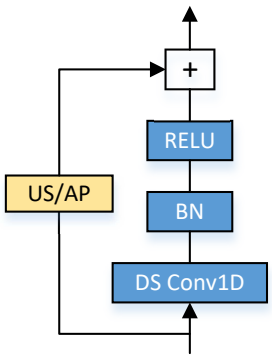


Figure 5: The 1D ResNet block. The 1D ResNet is built based on the temporal convolution block. **DS:** Depthwise separable [13]; **BN:** Batch Normalization [8]; **US:** Upsample; **AP:** Average Pooling [26]. The non-upsample convolution layers are all depthwise separable. BN, ReLU activation and identity skip connection are added after every convolution layer.

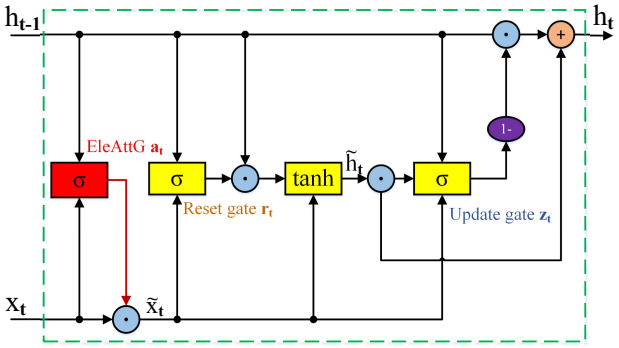


Figure 6: Illustration of Element-wise-Attention Gate (EleAttG) for a GRU block. A GRU block consists of a group of ($e.g.$, N) GRU neurons. In the diagram, each black line carries a vector. **Yellow** boxes – the units of the original GRU with the output dimension of N. **Blue** circle – element-wise operation and the brown circle denotes vector addition operation. **Red** box – EleAttG with an output dimension of D, which is the same as the dimension of the input x_t.

of MSR. As shown in Figure 2a, the video feature vectors are processed by a 1D temporal ResNet network (video stream), which consists of 10 convolution blocks of the 1D ResNet. The 1D ResNet block is proposed by [6] and its architecture is shown in Figure 5. Two of the intermediate blocks containing transposed convolution layers upsample the video features by 4 to match the temporal dimension of the audio feature vectors ($4T$). Similarly, the noisy magnitude spectrograms are fed into a residual network (audio stream) which consists of 5 convolution blocks of the 1D ResNet and outputs audio feature vectors. Then the audio feature vectors and the video feature vectors are fused in a fusion layer by simply concatenated over the channel dimension. The fused multi-modality vector is then fed into a one-layer EleAtt-GRU encoder followed by 2 fully connected layers. We use sigmoid as activation to produce a target enhancement mask (values range from 0 to 1). EleAtt-GRU is demonstrated more effective than other RNN variants in spatio-temporal tasks and its detail is introduced in section 3.4. The enhanced magnitude is produced by multiplying the original magnitude spectrogram with the target enhancement mask element-wise:

$$\hat{M} = \sigma\left(W_m^T EleAtt(f_{av})\right) \odot M_n \qquad (1)$$

where σ denotes the activation function of Sigmoid. f_{av} is the fused multi-modality vector output by the fusion layer. $EleAtt$ denotes the one-layer EleAtt-GRU encoder and W_m^T denotes the weight matrices of two FC layers followed the encoder. M_n denotes the original magnitude spectrogram and $\hat{M}$ denotes the enhanced magnitude spectrogram.

An illustration of the VE sub-network is shown in Supplementary Material. Enhancement of audio modality can improve both the performance of audio speech recognition (ASR) and the performance of MSR.

3.4. Multi-modality speech recognition network

MSR sub-network uses encoder-to-decoder (enc2dec) mechanism. As shown in Figure 2, the enhanced magnitude spectrogram is extracted from the 'noise-free' audio stream (details shown in Supplementary Material) to produce enhanced audio features. These two 1D-ResNet blocks of the new audio stream with stride 2 down-sample the temporal dimension of the audio features by 4 to match the temporal dimension of video features (T).

Transformer [45] is a self-attention model recently used in the area of lip reading [4] and MSR [2]. It has a nice performance on LRS2-BBC [2] and LRS3-TED [5] datasets. Although transformer is a powerful model emerging in lip reading [2, 4], it builds character relationships within limited length. RNN is more effective with long sequences than Transformer. Considering this situation, we introduce a recent proposed RNN variant model which named *Gated recurrent unit with element-wise-attention* (EleAtt-GRU) [48]. Compared with an RNN block (GRU, LSTM and other RNN variants in a layer), EleAtt-GRU is equipped with an element-wise-attention gate (EleAttG) that empowers an RNN neuron to have the attentive capacity. In this way, an EleAttG has the ability to modulate the input adaptively by assigning different importance levels, i.e., attention, to each element or dimension of the input. Illustration of EleAttG for a GRU block is shown in Figure 6 and cor-

responding computations are as follows:

$$\begin{aligned} \tilde{x_t} &= a_t \odot x_t \\ r_t &= \sigma\left(W_{xr}\tilde{x_t} + W_{hr}h_{t-1} + b_r\right) \\ z_t &= \sigma\left(W_{xz}\tilde{x_t} + W_{hz}h_{t-1} + b_z\right) \\ h_t{}' &= \tanh\left(W_{xh}\tilde{x_t} + W_{hh}\left(r_t \odot h_{t-1}\right) + b_h\right) \\ h_t &= z_t \odot h_{t-1} + (1 - z_t) \odot h_t{}' \end{aligned} \tag{2}$$

where σ denotes the activation function of Sigmoid. The attention response of an EleAttG is a vector a_t with the same dimension as the input x_t of GRU. a_t modulates x_t to the updated input $\tilde{x_t}$. r_t and z_t denote the reset gate and update gate of GRU. h_t and h_{t-1} are the output vectors of the hidden state and the previous hidden state. $W_{\alpha\beta}$ denotes the weight matrix related with α and β, where $\alpha \in \{x, h\}$ and $\beta \in \{r, z, h\}$. b_γ is the bias vector, where $\gamma \in \{r, z, h\}$. In a GRU block/layer, all neurons share the same EleAttG, which reduces the cost of computation complexity and number of parameters.

As shown in Figure 2, we build an sequence-to-sequence EleAtt-GRU network (EA-GRU-seq2seq) for MSR, using two EleAtt-GRU (EA-GRU) layers as the kernel of encoder and the other two EleAtt-GRU layers as the kernel of decoder. The video features extracted by P3D network are fed into the encoder along with the audio features. The number of unit of EleAtt-GRU in both encoder and decoder is 128. The decoder outputs character probabilities which are directly matched to the ground truth labels and trained with a cross-entropy loss. The MSR sub-network can also be used when only single modality (audio or visual) is available.

3.5. Loss function

The VE sub-network is trained by minimizing the L_1 loss between the predicted magnitude spectrogram and the ground truth. For sequence prediction task, we use a sequence-to-sequence (seq2seq) loss [12, 43]. For single label inference task, we use cross entropy loss.

4. Experiment

4.1. Datasets

The proposed network is trained and evaluated on LRW [15] and LRS3-TED [5] datasets. LRW is a very large-scale lip reading dataset in the wild, it consists of short clips (29 frames) from British television broadcasts, including news and talk shows. The dataset consists of up to 1000 utterances of 500 different words, spoken by more than 1000 speakers. We use LRW dataset to pre-train the P3D spatio-temporal front-end based on a word-level classification network of lip reading, similar to [41].

LRS3-TED is the largest available dataset in the field of lip reading (visual speech recognition). It consists of face tracks from over 400 hours of TED and TEDx videos, and organized into three sets: pre-train, train-val and test. We train the VE sub-network and the MSR sub-network on the LRS3-TED dataset.

	Transcription	WER %
GT	We can prevent the worst case scenario	
V	We can put and worst case scenario	34
A	We can prevent the worst case tcario	8
AV	We can prevent the worst case scenario	0
	Noisy	
GT	what would that change about how we live	
V	wouldn't at chance whole a life	53
A	that would I try all we live	50
AV	that would I chance all how we live	25
E-A	what would that change about how we live	0
	Noisy	
GT	human relationships are not efficient	
V	you man relation share are now efficient	38
A	man went left now fit	73
AV	you man today are now efficient	43
E-A	human relations are now efficient	14
E-AV	human relationships are not efficient	0

Table 1: Examples of MSR and VE-MSR results. **GT:** Ground truth; **V:** video modality only; **A:** audio modality only; **AV:** multi-modality (audio-visual); **E-A:** enhanced audio; **E-AV:** enhanced audio and visual.

4.2. Evaluation protocol

For the word-level lip reading experiment, the train, validation and test sets are provided with the LRW dataset. We report word accuracy for classification in 500 word classes of LRW. For sentence-level recognition experiments, we report the Word Error Rate (WER). WER is defined as $WER = (S + D + I)/N$, where S is the number of substitution, D is the number of deletions, I is the number of insertions to get from the reference to the hypothesis, and N is the number of words in the reference [14].

4.3. Training

The spatio-temporal visual front-end of P3D is pre-trained on a word-level classification network of lip reading with LRW dataset for 500 word classes and we adopt a two-steps training strategy. In the first step, video frames are fed into a 3D convolution, which is followed by a P3D, and the back-end is based on one dense layer. In the second step, to improve the effectiveness of the model we replace the dense layer with two layers of EleAtt-GRU, followed by a linear and a SoftMax layer. With the visual front-end frozen, we extract and save video features, as well as magnitude spectrograms for both original audio and the mix-noise one.

To demonstrate the gain of our model, we follow the noise mix method of [2], the babble noise with 0dB SNR is added to audio stream with probability p_n = 0.25 and the

Methods	Word accuracy
Chung and Zisserman [14]	76.2%
Stafylakis and Tzimiropoulos [41]	83.0%
Ours	84.8%

Table 2: Word accuracy of different word-level classification networks on the LRW dataset.

Method \ SNR		clean	0	5	10
	M	**WER**			
Google S2T API [11]	A	10.4%	70.3%	-	-
TM-seq2seq [2]	A	9.0%	60.5%	-	-
TM-seq2seq	V	59.9%	-	-	-
TM-seq2seq	AV	8.0%	44.3%	-	-
EA-GRU-seq2seq	A	7.2%	58.2%	42.6%	35.5%
EA-GRU-seq2seq	V	57.8%	-	-	-
EA-GRU-seq2seq	AV	6.8%	41.1%	36.8%	32.2%
EA-GRU-seq2seq*	A	-	36.6%	32.7%	24.2%
EA-GRU-seq2seq*	AV	-	28.5%	26.3%	23.5%

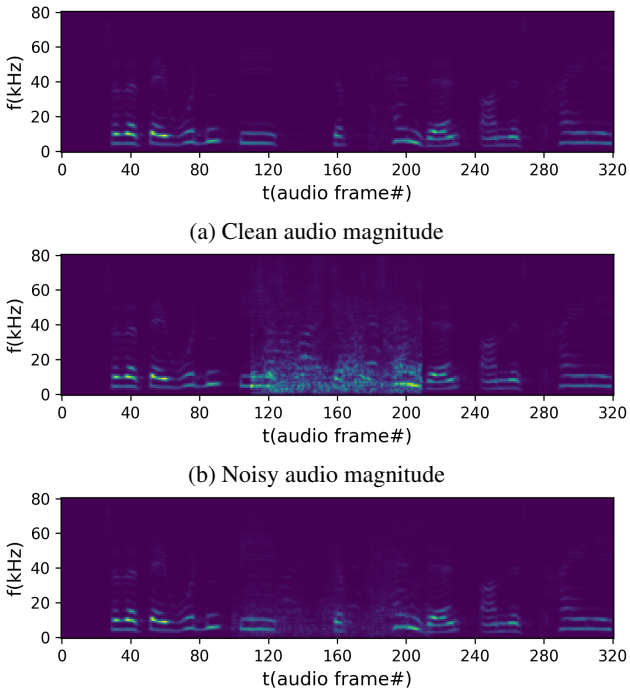
Table 3: Word error rates (WER) on the LRS3-TED dataset. V, A and AV denote video modality only, audio modality only and multi-modality (audio and video) inputs respectively. * means the VE driven ASR and VE-MSR models.

babble noise samples are synthesized by mixing the signals of 30 different audio samples from LRS3-TED dataset.

We first train the visual speech enhancement (VE) sub-network. Simultaneously, the multi-modality speech recognition (MSR) sub-network is trained with video features and clean (original audio) magnitude spectrogram as inputs. Then we freeze the enhancement sub-network and train the MSR sub-network with video features and enhanced magnitude output by the VE sub-network. Instead of immediately training on full sentences, we follow a curriculum learning training method [2]. The training starts with single word samples, and then the length of the training sequence gradually grows. This curriculum method not only improves the rate of convergence on the training set, but also reduces overfitting significantly.

The output size of decoder is 41, accounting for the 26 characters in the alphabet, the 10 digits, and tokens for [PAD], [EOS], [BOS] and [SPACE]. We also use teacher forcing method, in which the ground truth of the previous decoding step serves as input to the decoder.

The implementation of the network is based on the TensorFlow library [1] and trained on a single P100 GPU with 16GB memory. We use the ADAM optimiser to train the network with dropout and label smoothing. An initial learning rate is set to 10^{-4}, and decreased by a factor of 2 every time if the training error did not improve, the final learning rate is $5{\times}10^{-6}$. Training of the entire network takes approximately 10 days, including the training of two sub-networks separately and the subsequent joint training.



(a) Clean audio magnitude

(b) Noisy audio magnitude

(c) Enhanced audio magnitude

Figure 7: Comparison of magnitude spectrograms: **a)** the clean (original) utterance; **b)** we obtain this noisy utterance by adding babble noise to the 100 central audio frames; **c)** the enhanced audio utterance, which shows the effect of VE when compared to **b)**.

4.4. Results

Video modality only. Lip reading (video modality only) experiments are performed both on word-level and sentence-level. We train the word-level lip reading network on the LRW dataset to classify 500 word classes. We report word accuracy as evaluation metric. As shown in Table 2, our result exceeds the current state-of-the-art [41] on LRW. It demonstrates the superiority of our spatio-temporal visual front-end network in extracting video features compared to the one used in [2]. Our MSR (EA-GRU-seq2seq) network with only video modality achieves a WER of 57.8%, where it improves by 2.1% compared to the previous 59.9% of state-of-the-art [2] without language model in decoder.

MSR. We perform audio modality only and MSR experiments with both clean and noisy audio signal in the MSR sub-network. Noisy utterances are obtained by adding babble noise to the clean audio frames. The results in Table 3 demonstrate that MSR outperforms one single modality (audio or video) both in audio clean and noisy environ-

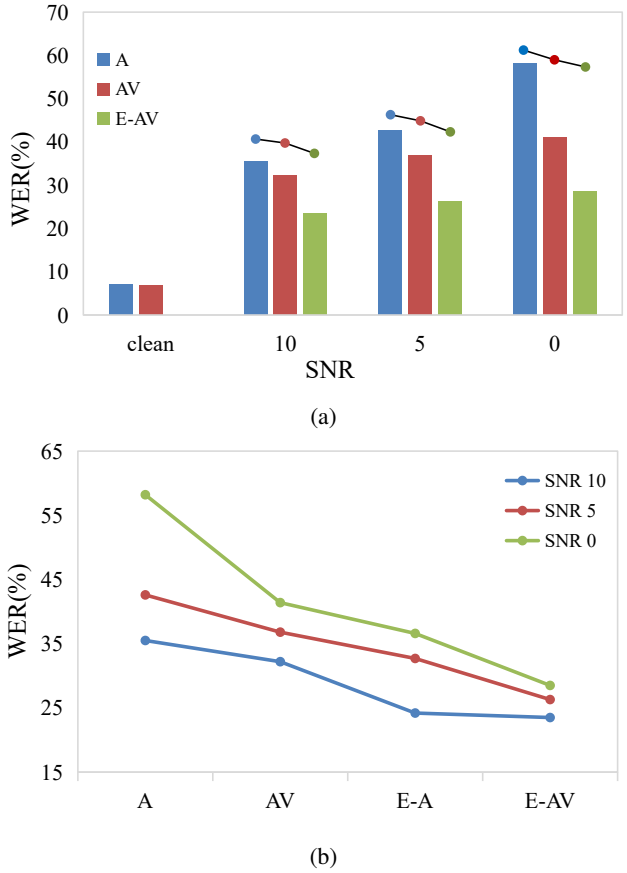


Figure 8: **a)** Word error rate (WER) for the audio modality speech recognition (**A**), MSR (**AV**) and VE-MSR (**E-AV**) networks in different SNR noisy environments. As show in the figure, multi-modality information can improve the performance of speech recognition and enhanced audio modality makes further improvements on MSR. **b)** Line chart of WER for the four kinds of speech recognition methods with different SNR audio inputs. **E-A** denotes the VE driven ASR. Shown in the diagram, with higher SNR audio inputs, video modality information can improve more both in speech enhancement and speech recognition performance.

ments. The results also demonstrate that video modality can improve ASR performance, particularly in high SNR noisy environment. For example, the WER is reduced from 7.2% for single modality of audio to 6.8% with clean audio signal; the WER is reduced from 35.5% for single modality of audio to 32.2% with 10dB audio signal and the WER is reduced from 58.2% for single modality of audio to 41.1% with 0dB SNR audio signal. As shown in Figure 8(b), with higher SNR audio inputs, video modality information can help more in speech recognition task. Table 3 shows quantitative gain of our MSR network performance compared to the previous MSR state-of-the-art [2] without extra language model. Table 1 shows some of the many examples where the single modality (video or audio alone) fails to predict the correct sentences, but these sentences are correctly deciphered by applying both modalities. It also shows that, in some noisy environment the MSR model also fails to produce the right sentence.

VE-MSR. Enhancement of one modality can contribute to the improvement of MSR performance. ASR in noisy environment is extremely challenging, so enhancement of audio can separate target voice from background noise and improve performance of ASR. As shown in Figure 7, we add babble noise to the 100 central audio frames to obtain noisy utterance. Comparing Figure 7b and c, it apparently shows the effectiveness of VE sub-network. For example, after being enhanced by our VE sub-network, the WER is reduced from 58.2% to 36.6% for ASR with 0dB SNR audio signal, improving speech recognition performance more than multi-modality, which achieves a WER of 41.1%. The results in Table 3 show that our enhancement of audio modality can further benefit the performance of MSR. For example, the WER is reduced from 41.1% to 28.5% for MSR with 0dB SNR audio signal. Our VE-MSR network shows significant advantage in terms of performance when compared to the state of the art [2], the WER is reduced by 15.8% with 0dB SNR audio signal. Table 1 shows some of the many examples where the multi-modality model fails to predict the correct sentences, but the VE driven ASR model and multi-modality model decipher the words successfully in some noisy environments.

5. Conclusion

In this paper, we introduce the visual speech enhancement driven multi-modality speech recognition (VE-MSR) network, which reconstructs cleaner audio modality by incorporating visual features to benefit the performance of MSR. In VE-MSR network, the VE sub-network is proposed to separate magnitude spectrogram of target speaker from noisy background or other speakers' voices by leveraging visual information of target speaker's lips. The enhanced audio modality together with video modality are then fed into the MSR sub-network to yield characters. We build the visual front-end of VE-MSR based on a pseudo-3D residual network (P3D) to extract more effective visual features. We introduce EleAtt-GRU to adaptively encode the spatio-temporal information in VE and MSR sub-networks, benefiting performance of VE and MSR. By extensive experiments, we demonstrate multi-modality information that helps improve speech recognition performance compared to only single modality. In noisy environment, the enhanced audio further leads to a significant performance gain on MSR.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, and X. Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016.

[2] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[3] T. Afouras, J. S. Chung, and A. Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018.

[4] T. Afouras, J. S. Chung, and A. Zisserman. Deep lip reading: A comparison of models and an online application. *arXiv preprint arXiv:1806.06053*, 2018.

[5] T. Afouras, J. S. Chung, and A. Zisserman. Lrs3-ted: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[6] T. Afouras, J. S. Chung, and A. Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019.

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

[8] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[9] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[10] W. Chan, N. Jaitly, Q. Le, and O. Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964, 2016.

[11] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, E. Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. pages 4774–4778, 2018.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[13] F. Chollet. Xception: Deep learning with depthwise separable convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017.

[14] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453, 2017.

[15] J. S. Chung and A. Zisserman. Lip reading in the wild. *Asian Conference on Computer Vision*, pages 87–103, 2016.

[16] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. *Asian Conference on Computer Vision*, pages 251–263, 2016.

[17] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[18] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, 49(2):167–192, 2017.

[19] R. D. Easton and M. Basala. Perceptual dominance during lipreading. *Perception & Psychophysics*, 32(6):562–570, 1982.

[20] A. Ephrat, T. Halperin, and S. Peleg. Improved speech reconstruction from silent video. *IEEE International Conference on Computer Vision*, pages 455–462, 2017.

[21] C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11(4):796–804, 1968.

[22] A. Gabbay, A. Ephrat, T. Halperin, and S. Peleg. Seeing through noise: Visually driven speaker separation and enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3051–3055, 2018.

[23] A. Gabbay, A. Shamir, and S. Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017.

[24] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *International Conference on Machine Learning*, pages 369–376, 2006.

[25] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. *International Conference on Machine Learning*, pages 1764–1772, 2014.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[27] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[28] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

[29] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[31] K. Kumar, T. Chen, and R. M. Stern. Profile view lip reading. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4:IV–429, 2007.

[32] C. Martin, B. Jon, C. Stuart, and S. Xu. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(1):2421–2424, 2006.

[33] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[34] E. D. Petajan. Automatic lipreading to enhance speech recognition (speech reading). 1985.

[35] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition.

IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6548–6552, 2018.

[36] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. *IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[37] D. E. Rumelhart, G. E. Hinton, R. J. Williams, et al. Learning representations by back-propagating errors. *Nature*, 5(3):1, 1988.

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IEEE International Journal of Computer Vision*, 115(3):211–252, 2015.

[39] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[40] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[41] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*, 2017.

[42] W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215, 1954.

[43] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[44] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[46] M. F. Woodward and C. G. Barber. Phoneme perception in lipreading. *Journal of Speech and Hearing Research*, 3(3):212–222, 1960.

[47] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. *IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8, 2019.

[48] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng. Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks. *arXiv preprint arXiv:1909.01939*, 2019.

[49] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen. A review of recent advances in visual speech decoding. *Image and Vision Computing*, 32(9):590–605, 2014.