Discriminative Multi-modality Speech Recognition

Bo Xu, Cheng Lu, Yandong Guo and Jacob Wang

Xpeng motors

xiaoboboer@gmail.com

Abstract

Vision is often used as a complementary modality for audio speech recognition (ASR), especially in the noisy environment where performance of solo audio modality significantly deteriorates. After combining visual modality, ASR is upgraded to the multi-modality speech recognition (MSR). In this paper, we propose a two-stage speech recognition model. In the first stage, the target voice is separated from background noises with help from the corresponding visual information of lip movements, making the model 'listen' clearly. At the second stage, the audio modality combines visual modality again to better understand the speech by a MSR sub-network, further improving the recognition rate. There are some other key contributions: we introduce a pseudo-3D residual convolution (P3D)-based visual front-end to extract more discriminative features; we upgrade the temporal convolution block from 1D ResNet with the temporal convolutional network (TCN), which is more suitable for the temporal tasks; the MSR sub-network is built on the top of Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU), which is more effective than Transformer in long sequences. We conducted extensive experiments on the LRS3-TED and the LRW datasets. Our two-stage model (audio enhanced multi-modality speech recognition, AE-MSR) consistently achieves the state-of-the-art performance by a significant margin, which demonstrates the necessity and effectiveness of AE-MSR.

1. Introduction

In the book *The Listening Eye: A Simple Introduction to the Art of Lip-reading* [17], Clegg mentions that "When you are deaf you live inside a well-corked glass bottle. You see the entrancing outside world, but it does not reach you. After learning to lip read, you are still inside the bottle, but the cork has come out and the outside world slowly but surely comes in to you." Lip reading is an approach for people with hearing impairments to communicate with the world, so that they can interpret what other say by looking at the movements of lips [7, 16, 22, 32, 46]. Lip reading is a dif-
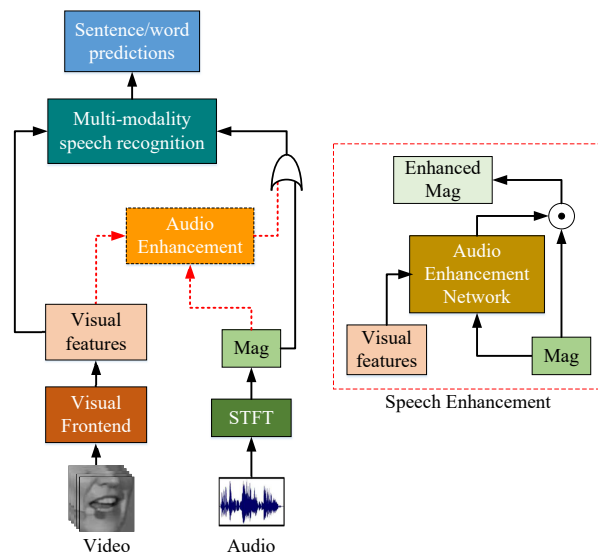

Figure 1: Overview of the audio enhanced multi-modality speech recognition network (AE-MSR). **Mag:** magnitude. Different from other MSR methods [2, 14, 47, 41, 35] with only single visual awareness, we firstly filter the voices of speakers and background noises with help from visual awareness. Then we combine visual awareness again for MSR to benefit speech recognition.

ficult skill for human to grasp and requires intensive training [20, 41]. Lip reading is also an inexact art, because different characters may exhibit the similar lip movements (*e.g.* 'b' and 'p') [2]. Consequently, several machine lip reading models are proposed to discriminate such subtle difference [18, 31, 34]. However they still suffer difficulties on extracting spatio-temporal features from the video.

Automatic lip reading becomes achievable due to rapid development of deep neural network in computer vision [30, 39, 43], and with help from large scale training datasets [14, 15, 18, 19, 37, 47]. In addition to serving as a powerful solution to hearing impairment, lip reading can also contribute to audio speech recognition (ASR) in adversary environments, such as in high noise level where hu-

man speaking is inaudible. Multi-modality (video and audio) is more effective than single modality (video or audio) in terms of both robustness and accuracy. Multi-modality (audio-visual) speech recognition (MSR) is one of the main extended applications of multi-modality. But similar to ASR, there is a significant deterioration in performance for MSR in noisy environment as well [2]. Compared to audio modality operating in a clean voice environment, the one in noisy environment shows less gain because of upgrading from ASR to MSR. [2] demonstrates that the noisy level of audio modality directly impacts the performance gain of MSR compared to single modality.

The goal of this paper is to introduce a two-stage speech recognition method with double visual-modality awareness. In the first stage, we reconstruct the audio signal which only contains the target speaker's voice with the guiding visual information (lip movements). In the second stage, the enhanced audio modality is combined with the visual modality again to yield better speech recognition. Compared to typical MSR methods with single time of visual modality awareness, our method is more advantageous in terms of robustness and accuracy.

We propose a deep neural network model named audio-enhanced multi-modality speech recognition (AE-MSR) with double visual awareness to implement the method. The AE-MSR model consists of two sub-networks, the audio enhancement (AE) and MSR. Before being fed into the MSR sub-network, audio modality is enhanced with help from the first visual awareness in the AE sub-network. After enhancement, audio stream and revisited visual stream are then fed into the MSR sub-network to make speech predictions.The techniques we incorporated into AE-MSR include pseudo 3D residual convolution (P3D), temporal convolutional network (TCN), and element-wise attention gated recurrent unit (EleAtt-GRU). Ablation study shown in the paper demonstrates the effectiveness of each of the above sub-modules and the combination of them. The MSR sub-network is also built on top of EleAtt-GRU.

The intuition of our AE-MSR is as follows. Typically, a deep learning-based MSR uses symmetric encoders for both audio and video. Though visual encoder is trained in an e2e fashion, we conduct experiments to show this is not the optimal way to leverage the visual information. The reason might be that the intrinsic architecture of the typical MSR implicitly suggests equal importance of audio and video. However we tell from various experiments that audio is still much more reliable to recognize speech, even in a noisy environment. Therefore, we re-design the architecture to embed this bias between video and audio as a prior.

Overall, the contributions of this paper are:

- We propose a two-stage double visual awareness MSR

model, where the first visual awareness is applied to remove the audio noise.

- We introduce the P3D as visual front-end to extract more discriminative visual features and EleAtt-GRU to adaptively encode the spatio-temporal information in AE and MSR sub-networks, benefiting performance of both networks.

- We upgrade the temporal convolution block of 1D ResNet to a TCN one in AE sub-network for establishing temporal connections.

- Extensive experiments demonstrate that AE-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 also 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 audio enhancement (AE) driven by visual information and multi-modality speech recognition (MSR).
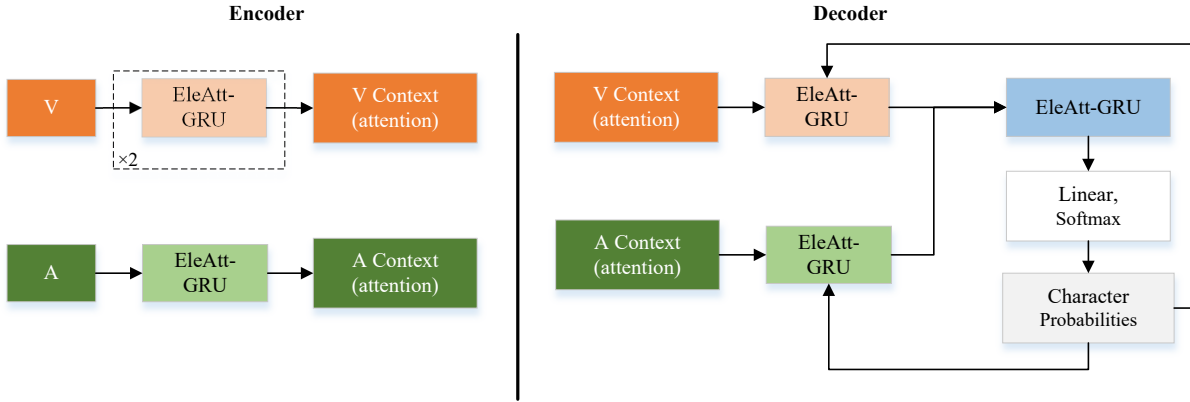
2.1. Audio enhancement

A few researchers have demonstrated that the target audio signal can be separated from other speakers' voices and background noises, *e.g.* Gabbay *et al.* [23] introduce a trained silent-video-to-speech model previously proposed by [21] to generate speech predictions as a mask on the noisy audio signal which is then discarded in the pipeline of audio enhancement. Gabbay *et al.* [24] also use the convolution neural networks (CNNs) to encode multi-modality features. The embedding vectors of audio and vision are concatenated before audio decoder and fed into transposed convolution of audio decoder to produce enhanced mel-scale spectrograms. Hou *et al.* [29] build a visual driven AE network on the top of CNNs and fully connected (FC) layers to generate enhanced speech and reconstructed lip image frames. Afouras *et al.* [3] use 1D ResNet as temporal convolution unit to process audio and visual features individually. Then the multi-modality features are concatenated and encoded into a mask by another 1D-ResNet-based encoder to remove noisy components in the audio signal. In their latest article, they propose a new approach that replaces the multi-modality feature encoder with Bi-LSTM [6].

2.2. Multi-modality speech recognition

Vision is often used as a complementary modality for audio speech recognition (ASR), especially in noisy environments. After combining visual modality, ASR is upgraded
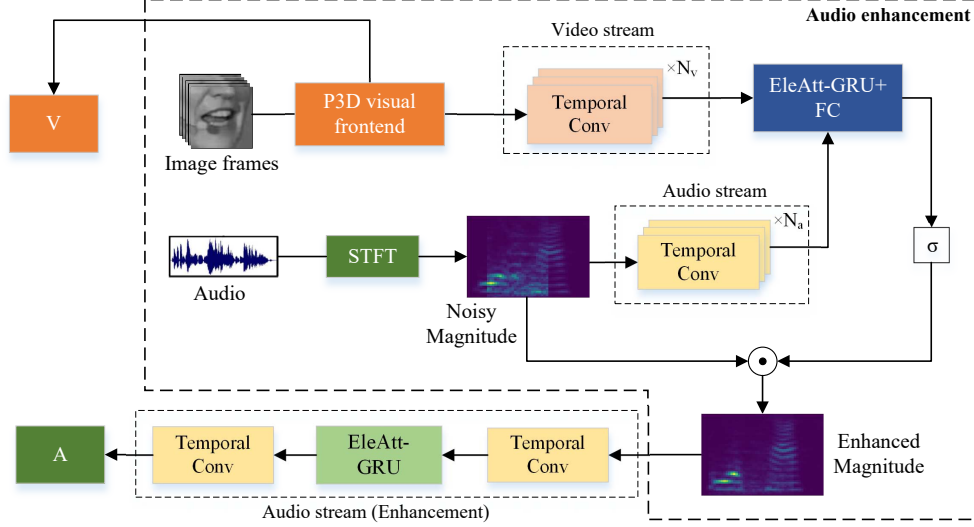
(a) Audio enhancement (AE) sub-network

(b) Multi-modality speech recognition (MSR) sub-network

Figure 2: Architecture of the multi-modality speech recognition network with double visual awareness (**AE-MSR**). The AE-MSR network consists of two sub-networks: **a)** the audio enhancement (AE) network. The network receives image frames and audio signals as inputs, outputting the enhanced magnitude spectrograms that the noisy spectrograms are filtered. **V:** visual features; **A:** enhanced audio magnitude. **b)** the multi-modality speech recognition (MSR) network.

to the multi-modality speech recognition (MSR). Reciprocally, MSR is also an upgrade to the lip reading and benefits people with hearing impairments to recognize speech by generating meaningful text.

In the field of deep learning, research on lip reading has longer history than MSR [49]. Assael *et al.* [7] propose LipNet, an end-to-end model on top of spatio-temporal convolutions, LSTM [28] and connectionist temporal classification (CTC) loss on variable-length sequence of video frames. Stafylakis *et al.* [41] introduce the state-of-the-art word-level classification lip reading network on LRW dataset [15]. The network consists of spatio-temporal convolution, residual network and Bi-LSTM.

On the basis of lip reading, MSR is developed [14,

2]. Various MSR methods often use encoder-to-decoder (enc2dec) mechanism which is inspired by machine translation [8, 10, 25, 26, 42, 45]. Chung *et al.* [14] use a dual sequence-to-sequence model with enc2dec mechanism. Visual features and audio features are encoded separately by LSTM units. Then multi-modality features are combined and decoded into characters. Afouras *et al.* [2] introduce a sequence-to-sequence model of encoder-to-decoder mechanism. The encoder and decoder of the model are built based on the transformer [45] attention architecture. In encoder stage, each modality feature is encoded with self-attention individually. After multi-head attention in decoder stage, the context vectors produced by multiple modalities are concatenated and fed to the feed forward layers to produce
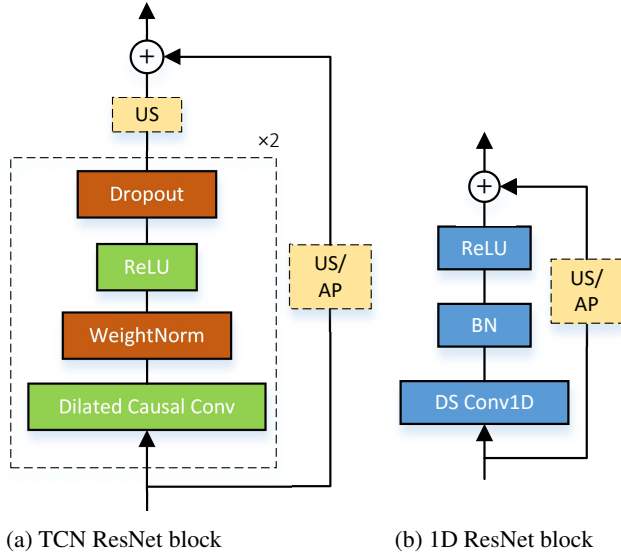
(a) TCN ResNet block	(b) 1D ResNet block

Figure 3: Temporal convolution blocks. **a) TCN ResNet block**. **US:** Up-sample; **AP:** Average Pooling [27]. **b) The 1D ResNet block**. **DS:** Depthwise separable [13]; **BN:** Batch Normalization. The non-upsample convolution layers are all depthwise separable.

probable characters. However, their state-of-the-art method suffer in noisy scenarios. In noisy environments, the performance dramatically decreases, this is the main reason why we propose the method of AE-MSR. In this paper, we qualitatively evaluate performance of the AE-MSR model for speech recognition in the noisy environments.

3. Architectures

In this section, we describe the double visual awareness multi-modality speech recognition (AE-MSR) network. It first learns to filter magnitude spectrogram from the voices of other speakers or background noises with help from the information of visual modality (*Watch once to listen clearly*). The subsequent MSR then revisits the visual modality and combines it with filtered audio magnitude spectrogram (*Watch again to understand precisely*). The model architecture is shown in detail in Figure 2.

3.1. Watch once to listen clearly

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, corresponding to 75% overlap. We multiply the resulting magnitude by a mel-spaced filter to compute the audio feature of mel-scale magnitude with mel-frequency bins of 80 between 0 to 8 kHz.

Visual features. We produce image frames by cropping original video frames to 112×112 pixel patches and choose mouth patch as region of interest (ROI). To extract video features, we build a 3D CNN (C3D) [44] -P3D [36] network to produce a more powerful visual spatio-temporal representation instead of using C3D plus 2D ResNet [27] which is mentioned in many other lip-reading papers [2, 3, 4, 6, 14, 41].

C3D is a beneficial method to capture spatio-temporal features of videos and widely adopted [41, 2, 3, 35, 6]. Multi-layer C3D can achieve even better performances in temporal tasks than a single layer one, however they are both computationally expensive and memory demanding. We use P3D to replace part of the C3D layers to alleviate this situation. The three block versions of P3D are shown in Supplementary Material, P3D ResNet is implemented by separating $N \times N \times N$ convolutions into $1 \times 3 \times 3$ convolution filters on spatial domain and $3 \times 1 \times 1$ convolution filters on temporal domain to extract spatial-temporal features. P3D ResNet achieves superior performances over 2D ResNet in different temporal tasks [36]. We implement a 50-layer P3D network by cyclically mixing the three blocks in the order of P3D-A, P3D-B, P3D-C.

The visual front-end is built on a 3D convolution layer with 64 filters of kernel size $5 \times 7 \times 7$, followed by batch normalization (BN), ReLU activation 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 maintaining 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 a 512-dimensional vector representing each video frame). The visual feature and corresponding magnitude spectrogram are then fed into audio enhancement sub-network.

Audio enhancement with the first visual awareness. Noise-free audio signal achieves satisfactory performance on audio speech recognition (ASR) and multi-modality speech recognition (MSR). However there is a significant deterioration in recognition performance in noisy environments [2, 3]. Architecture of the audio enhancement sub-network is illustrated in Figure 2a, where the visual features are fed into a temporal convolution network (video stream). The video stream consists of N_v temporal convolution blocks, outputting video feature vectors. We introduce two versions of temporal convolution blocks, one is the temporal convolutional network (TCN) proposed by [9] and the other is 1D ResNet block proposed by [6].

Architectures of temporal convolution blocks are shown in Figure 3, the residual block of TCN consists of two dilated causal convolution layers, each layer followed by a weight normalization (WN) [38] layer and a rectified linear unit (ReLU) [33] layer. A spatial dropout [40] layer

is added after ReLU layer for regularization [9]. Identity skip connection are added after the second dilated causal convolution layer. By combining causal convolution and dilated convolution, TCN guarantees no leakage from the future to the past and effectively expands the receptive field to maintain a longer memory size [9]. The 1D ResNet block is based on 1D temporal convolution layer, followed by a batch normalization (BN) layer. Residual connection is added after ReLU activation layer.

Two of the intermediate temporal convolution blocks containing transposed convolution layers up-sample the video features by 4 to match the temporal dimension of the audio feature vectors ($4T$). Similarly, the noisy magnitude spectrograms are proposed by a residual network (audio stream) which consists of N_a temporal convolution blocks, outputting audio feature vectors. Then the audio feature vectors and the video feature vectors are fused in a fusion layer by simply concatenating 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 with a 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 3.2. The enhanced magnitude is produced by multiplying the original magnitude spectrogram with the target enhancement mask element-wise. Architecture details of the audio enhancement sub-network are given in Supplementary Material.

3.2. Watch again to understand precisely

Multi-modality speech recognition with the second visual awareness. Visual information can help enhance audio modality by separating target audio signal from noisy background. After audio enhancement by visual awareness, the multi-modality speech recognition (MSR) is implemented by combining enhanced audio and the revisited visual representation to benefit the performance of speech recognition further.

We use encoder-to-decoder (enc2dec) mechanism in the MSR sub-network. Instead of using transformer [45], which demonstrates decent performance on lip reading [4] and MSR [2], our network is basically built on a RNN variant model named gated recurrent unit with element-wise-attention (EleAtt-GRU) [48]. Although transformer is a powerful model emerging in machine translation [45] and lip reading [2, 4], it builds character relationships within limited length, less effective with long sequences than RNN. EleAtt-GRU can alleviate this situation, because it is equipped with an element-wise-attention gate (EleAttG) that empowers an RNN neuron to have the attentive capacity. EleAttG is designed 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 Supplementary Material. In a GRU block/layer, all neurons share the same EleAttG, which reduces the cost of computation and number of parameters.

Architecture of the AE-MSR network is shown in Figure 2, a sequence-to-sequence MSR network is built based on EleAtt-GRU. The encoder is a two-layer EleAtt-GRU for both modalities. The enhanced audio magnitude is fed into an encoder layer between two 1D-ResNet blocks with stride 2 that down-sample the temporal dimension by 4 to match the temporal dimension of video features (T). The 1D-ResNet layer are followed by another encoder layer, outputting the audio modality encoder context. The video features extracted by C3D-P3D network are fed into the video encoder to output video encoder context. In the decoder stage, video context and audio context are decoded separately by independent decoder layer. Generated context vectors of both modalities are concatenated over the channel dimensions and propagated to another decoder layer to produce character probabilities. 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 and the whole output sequence is trained with sequence-to-sequence (seq2seq) loss [42].

4. Training

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 from British television broadcasts, including news and talk shows. LRW 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.

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 audio enhancement (AE) sub-network and the multi-modality speech recognition (MSR) sub-network on the LRS3-TED dataset.

4.2. Evaluation metric

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 sub-

stitution, 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 strategy

Visual front-end. The visual front-end of C3D-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-step training strategy. In the first step, image frames are fed into a 3D convolution, which is followed by a 50-layer 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 Bi-LSTM, followed by a linear and a SoftMax layer. We use cross entropy loss to train the word classification tasks. 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.

Noisy samples. In order to train our model so that it can be resistant to background noise or speakers, we follow the noise mix method of [2], the babble noise with SNR from -10 dB to 10 dB is added to audio stream with probability $p_n = 0.25$ and the babble noise samples are synthesized by mixing the signals of 30 different audio samples from LRS3-TED dataset.

AE and MSR sub-networks. The AE sub-network is firstly trained on multi-modality of mixed noises with temporal convolution block of TCN and 1D ResNet separately. The AE sub-network is trained by minimizing the L_1 loss between the predicted magnitude spectrogram and the ground truth. Simultaneously, the multi-modality speech recognition (MSR) sub-network is trained with video features and clean magnitude spectrogram as inputs. The MSR sub-network is also trained when only single modality (audio or visual) is available. For MSR sub-network, we use a sequence-to-sequence (seq2seq) loss [12, 42].

AE-MSR. We freeze the AE sub-network and train the AE-MSR network. To demonstrate the benefit of our model, we reproduce and evaluate the state-of-the-art multi-modality speech recognition model provided by [2] at different noise levels. The training begins with one-word samples, and then the length of the training sequence gradually grows. This is a cumulative method that not only improves the convergence rate on the training set, but also reduces overfitting significantly. Output size of decoder is set to 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 [2], in which the ground truth of the previous decoding step serves as input to the decoder.

Implementation details. The implementation of the network is based on the TensorFlow library [1] and trained

Methods	Word accuracy
Chung and Zisserman [14]	76.2%
Stafylakis and Tzimiropoulos [41]	83.0%
Petridis and Stafylakis [35]	82.0%
Ours	84.8%

Table 1: Word accuracy of different word-level classification networks on the LRW dataset.

Method	Google [11]	TM-seq2seq [2]		EG-seq2seq	
	WER %				
M / **SNR** dB	A	A	V	A	V
clean	10.4	9.0	59.9	**7.2**	**57.8**
10	-	35.9	-	**35.5**	-
5	-	49.0	-	**42.6**	-
0	70.3	60.5	-	**58.2**	-
-5	-	87.9	-	**86.1**	-
-10	-	100.0	-	100.0%	-

Table 2: Word error rates (WER) of both single modality speech recognition and multi-modality speech recognition (MSR) on the LRS3-TED dataset. **M:** modality. **A:** audio modality only; **V:** visual modality only.

on a single Tesla 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 5×10^{-6}. Training of the entire network takes approximately 15 days, including the training of the audio enhancement sub-network on both of the two temporal convolution blocks and the MSR sub-network separately and the subsequent joint training.

5. Experimental results

5.1. P3D-based visual front-end and EleAtt-GRU-based enc2dec

P3D-based visual front-end. We perform lip reading experiments on both word-level and sentence-level. In section 4.3, we introduce a word-level lip reading network on the LRW dataset to classify 500 word classes to train the visual front-end of C3D-P3D. Result of this word-level lip reading network is shown in Table 1, where we report word accuracy as evaluation metric and our result surpasses the state-of-the-art [41] on the LRW dataset. It demonstrates that visual front-end network of C3D-P3D is more advantageous in extracting video feature representations than the

SNR dB / Met	AV		VA				VAV			
	TM-s2s	EG-s2s	1D-TM-s2s	T-TM-s2s	1D-EG-s2s	T-EG-s2s	1D-TM-s2s	T-TM-s2s	1D-EG-s2s	T-EG-s2s
clean	8.0%	6.8%	-	-	-	-	-	-	-	-
10	33.4%	32.2%	25.9%	24.1%	24.2%	**23.2**%	24.5%	22.0%	21.5%	**20.7**%
5	38.1%	36.8%	34.1%	31.7%	32.7%	**30.9**%	30.2%	25.6%	26.3%	**24.3**%
0	44.3%	41.1%	37.0%	33.2%	36.6%	**32.5**%	31.6%	29.6%	28.5%	**25.5**%
-5	56.2%	52.6%	50.2%	49.5%	49.3%	**46.0**%	36.7%	35.1%	32.7%	**31.1**%
-10	60.9%	57.9%	52.5%	49.8%	50.6%	**44.5**%	42.3%	42.0%	40.2%	**38.6**%

Table 3: Word error rates (WER) of both audio speech recognition (ASR) with single visual modality awareness and multi-modality speech recognition (MSR) with double visual modality awareness on the LRS3-TED dataset. **Met:** method. **TM-s2s:** TM-seq2seq; **EG-s2s:** EG-seq2seq; **1D-TM-s2s:** an AE-MSR model, which consists of 1DRN-AE and TM-seq2seq; **T-TM-s2s:** an AE-MSR model, which consists of TCN-AE and TM-seq2seq; **1D-EG-s2s:** an AE-MSR model, which consists of 1DRN-AE and EG-seq2seq; **T-EG-s2s:** an AE-MSR model, which consists of TCN-AE and EG-seq2seq. **AV:** multi-modality with single visual modality awareness; **VA:** enhanced audio modality by single visual awareness for ASR; **VAV:** multi-modality by double visual awareness for multi-modality speech recognition (MSR).

C3D-2D-ResNet one used by [2].

EleAtt-GRU-based enc2dec. Results in both of Column V and A in Table 2 demonstrate that EleAtt-GRU-based enc2dec is more beneficial in speech recognition than the Transformer-based enc2dec. As shown in Table 2 Column V, our multi-modality speech recognition (EG-seq2seq) network (illustrated in Figure 2b) with only visual modality reduces word error rate (WER) by 2.1% compared to the previous state-of-the-art (TM-seq2seq) [2] WER of 59.9% on the LRS3 dataset without using language model in decoder. Furthermore, we also evaluate the EleAtt-GRU-based enc2dec model in ASR at different noise levels. As shown in Table 2 Column A, EG-seq2seq exceeds the state-of-the-art (TM-seq2seq) model on ASR at all noise levels (-10 dB to 10 dB) without extra language model. Table 2 Column A also shows that neither EG-seq2seq or TM-seq2seq works any more with only audio modality at -10 dB SNR.

Results in the columns under AV in Table 3 demonstrate the speech recognition accuracy improvement after adding the visual awareness once at the MSR stage, especially in noisy environments. Even when the audio is clean, visual modality can still play a helping role, for example the WER is reduced from 7.2% for audio modality only to 6.8% for multi-modality. EG-seq2seq outperforms the state-of-the-art (TM-seq2seq) model on MSR at different noise levels. It again demonstrates the superiority of EleAtt-GRU-based enc2dec in speech recognition. However, we notice that under very noisy conditions, audio modality can negatively impact the MSR because of its highly polluted input, when comparing lip reading (V in Table 2) with MSR (AV in Table 3) at -10 dB SNR.

5.2. Audio enhancement (AE) with the first visual awareness

In order to demonstrate the enhancement effectiveness of our AE models so that it can benefit not only our speech recognition models but also other speech recognition models. Compared with MSR in the Section 5.1, here we apply visual awareness at audio enhancement stage, instead of at MSR. We compare and analyze the results of following networks at different noise levels:

- **1DRN-TM-seq2seq:** an AE-MSR network, where the audio enhancement (AE) sub-network (1DRN-AE) uses 1D ResNet as temporal convolution unit and outputs enhanced audio modality. The MSR sub-network of this network is TM-seq2seq.

- **TCN-TM-seq2seq:** an AE-MSR network, where the AE sub-network (TCN-AE) uses the temporal convolutional network (TCN) as temporal convolution unit. The MSR sub-network is TM-seq2seq.

- **1DRN-EG-seq2seq:** an AE-MSR network, where the AE sub-network is 1DRN-AE and the MSR sub-network is EG-seq2seq.

- **TCN-EG-seq2seq:** an AE-MSR network, where the AE sub-network is TCN-AE and the MSR sub-network is EG-seq2seq.

In this section, all the models above use only audio modality at MSR stage. As shown in columns under VA in Table 3, our AE networks can benefit other speech recognition models, for example at SNR of -5 dB, the WER is reduced from 87.9% of TM-seq2seq to 50.2% of 1DRN-TM-seq2seq and 49.5% of TCN-TM-seq2seq. The enhancement

Audio source	Magnitude error %		
	SNR dB -5	0	5
Noisy	97.1	65.4	49.1
1DRN-AE	66.5	51.0	35.6
TCN-AE	59.5	46.3	33.1

Table 4: Energy errors between original noise-free audio magnitudes and enhanced magnitudes produced by different audio enhancement models.

gain is also clearly illustrated in Figure 4. Moreover, by comparing the result of columns under AV and VA in Table 3, with the same number of visual awareness, our audio enhancement approach shows more benefit in speech recognition than the multi-modality with single visual awareness in noisy environments.

Magnitudes produced by the two AE models are shown in Supplementary Material. We also introduce an energy error function to measure the effect of audio enhancement models as follow:

$$\Delta M = \frac{\| M - M_o \|_2}{\| M_o \|_2} \qquad (1)$$

where M is the magnitudes of noisy audio or enhanced audio, M_o is the original audio without mixing noises, ΔM is the deviation results between M and M_o. We chose 10,000 noise-free samples that are added to babble noises with SNR of -5 dB, 0 dB and 5 dB separately to compare the enhancement performance between 1DRN-AE and TCN-AE networks. We average the ΔM results among samples at each SNR-level. Results in Table 4 show the beneficial performance of TCN-AE.

In Supplementary Material, we list 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 multi-modality also fails to produce the right sentence, however the enhanced audio modality predict successfully. Experimental results of speech recognition in Table 3.2 also demonstrate that TCN-EG-seq2seq is more advantageous than 1DRN-EG-seq2seq in audio modality enhancement due to the TCN temporal convolution unit, which has a longer-term memory and larger receptive field by combining causal convolution and dilated convolution that more beneficial in temporal tasks.

5.3. Multi-modality speech recognition with the second visual awareness

After audio enhancement with the first visual awareness, we implement multi-modality speech recognition with
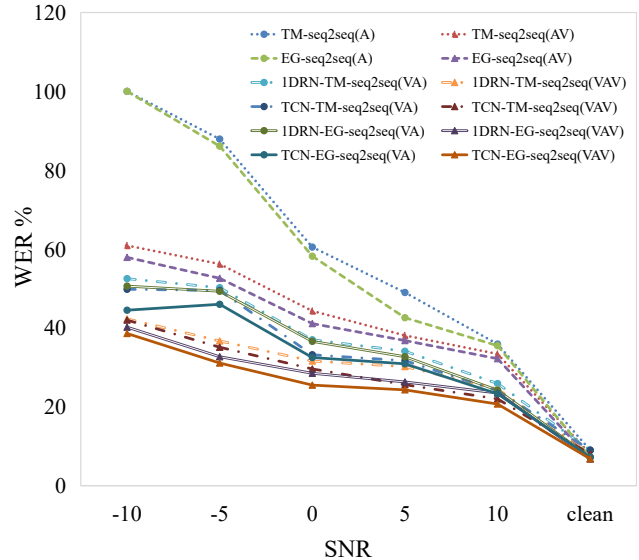

Figure 4: Word error rate (WER) on different methods. Each method in this diagram equivalent to the one with same name in Table 2.

the second visual awareness. By comparing the results in columns under VA and VAV in Table 3, MSR with double visual awareness leads to a further improvement compared to any single visual awareness method (*e.g.* AV, VA and V). For example, the WER of 1DRN-EG-seq2seq is reduced from 36.6% to 28.5% when combining the visual awareness again for speech recognition after audio enhancement, and the TCN-EG-seq2seq model reduces the WER even more. It demonstrates the performance gain because of the second visual awareness in MSR. Our AE-MSR network shows significant advantage in terms of performance after combining visual awareness twice, once for audio enhancement and the other for MSR. In Supplementary Material we list some examples that the multi-modality model (**AV**) and the AE model (**VA**) fail to predict the correct sentences, but the AE-MSR model deciphers the words successfully in some noisy environments.

6. Conclusion

In this paper, we introduce a two-stage speech recognition model named double visual awareness multi-modality speech recognition (AE-MSR) network, which consists of the audio enhancement (AE) sub-network and the multi-modality speech recognition (MSR) sub-network. By extensive experiments, we demonstrate the necessity and effectiveness of double visual awareness for MSR, and our method leads to a significant performance gain on MSR especially in noisy environments. Further, our models in this paper outperform the state-of-the-art ones on the LRS3-TED and the LRW datasets by a significant margin.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. 2016.

[2] 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.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018.

[4] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: A comparison of models and an online application. *arXiv preprint arXiv:1806.06053*, 2018.

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

[6] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. My lips are concealed: Audio-visual speech enhancement through obstructions. *arXiv preprint arXiv:1907.04975*, 2019.

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

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[9] Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

[10] William Chan, Navdeep Jaitly, Quoc Le, and Oriol 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] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. State-of-the-art speech recognition with sequence-to-sequence models. pages 4774–4778, 2018.

[12] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

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

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

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

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

[17] Dorothy G Clegg. *The listening eye: a simple introduction to the art of lip-reading*. Methuen & Company, 1953.

[18] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[19] Andrzej Czyzewski, Bozena Kostek, Piotr Bratoszewski, Jozef Kotus, and Marcin Szykulski. An audio-visual corpus for multimodal automatic speech recognition. *Journal of Intelligent Information Systems*, 49(2):167–192, 2017.

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

[21] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. *IEEE International Conference on Computer Vision*, pages 455–462, 2017.

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

[23] Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Seeing through noise: Visually driven speaker separation and enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3051–3055, 2018.

[24] Aviv Gabbay, Asaph Shamir, and Shmuel Peleg. Visual speech enhancement. *arXiv preprint arXiv:1711.08789*, 2017.

[25] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *International Conference on Machine Learning*, pages 369–376, 2006.

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

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

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

[29] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):117–128, 2018.

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

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

[32] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746, 1976.

[33] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. pages 807–814, 2010.

[34] Eric David Petajan. Automatic lipreading to enhance speech recognition (speech reading). 1985.

[35] Stavros Petridis, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552, 2018.

[36] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatiotemporal representation with pseudo-3d residual networks. *IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[37] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IEEE International Journal of Computer Vision*, 115(3):211–252, 2015.

[38] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. pages 901–909, 2016.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

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

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2015.

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. pages 4489–4497, 2015.

[45] 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*, pages 5998–6008, 2017.

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

[47] Shuang Yang, Yuanhang Zhang, Dalu Feng, Mingmin Yang, Chenhao Wang, Jingyun Xiao, Keyu Long, Shiguang Shan, and Xilin 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] Pengfei Zhang, Jianru Xue, Cuiling Lan, Wenjun Zeng, Zhanning Gao, and Nanning Zheng. Eleatt-rnn: Adding attentiveness to neurons in recurrent neural networks. *arXiv preprint arXiv:1909.01939*, 2019.

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