Audio visual speech recognition with multimodal recurrent neural networks

Conference Paper · May 2017

DOI: 10.1109/IJCNN.2017.7965918

5 authors, including:

Weijiang Feng
National University of Defense Technology
13 PUBLICATIONS **143** CITATIONS

SEE PROFILE

Naiyang Guan
National Innovation Institute of Defense Technology
80 PUBLICATIONS **1,997** CITATIONS

SEE PROFILE

Zhigang Luo
National University of Defense Technology
152 PUBLICATIONS **2,614** CITATIONS

SEE PROFILE

Audio Visual Speech Recognition with Multimodal Recurrent Neural Networks

Weijiang Feng[1], Naiyang Guan[2], Yuan Li[1], Xiang Zhang[1], Zhigang Luo[2]
[1]Institute of Software, College of Computer
National University of Defense Technology, Changsha, Hunan, P.R. China, 410073
[2]Science and Technology on Parallel and Distributed Laboratory
National University of Defense Technology, Changsha, Hunan, P.R. China, 410073
Email: fengweijiang14@nudt.edu.cn, ny_guan@nudt.edu.cn (corresponding author)
liyuan94@nudt.edu.cn, zhangxiang_43@aliyun.com, zgluo@nudt.edu.cn

***Abstract*—Studies on nowadays human-machine interface have demonstrated that visual information can enhance speech recognition accuracy especially in noisy environments. Deep learning has been widely used to tackle such audio visual speech recognition (AVSR) problem due to its astonishing achievements in both speech recognition and image recognition. Although existing deep learning models succeed to incorporate visual information into speech recognition, none of them simultaneously considers sequential characteristics of both audio and visual modalities. To overcome this deficiency, we proposed a multimodal recurrent neural network (multimodal RNN) model to take into account the sequential characteristics of both audio and visual modalities for AVSR. In particular, multimodal RNN includes three components, i.e., audio part, visual part, and fusion part, where the audio part and visual part capture the sequential characteristics of audio and visual modalities, respectively, and the fusion part combines the outputs of both modalities. Here we modelled the audio modality by using a LSTM RNN, and modelled the visual modality by using a convolutional neural network (CNN) plus a LSTM RNN, and combined both models by a multimodal layer in the fusion part. We validated the effectiveness of the proposed multimodal RNN model on a multi-speaker AVSR benchmark dataset termed AVletters. The experimental results show the performance improvements comparing to the known highest audio visual recognition accuracies on AVletters, and confirm the robustness of our multimodal RNN model.**

***Index Terms*—Audio visual speech recognition; deep learning; multimodal learning; recurrent neural networks; LSTM**

I. Introduction

Human beings recognize speeches of a speaker according to multimodal information. Besides speech audio, visual information such as lip and tongue movements can also provide a cue for speech understanding. Using visual information, mostly by watching lip movements, to recognize what a person is saying is often referred to as lipreading. The lipreading techniques can be utilized by hearing-impaired listeners to understand spoken speech. Even for people with normal hearing, Lipreading can help people to understand speeches, especially in noisy environments [1], [2]. The relationship between audio and visual information can be demonstrated by McGurk effect [3] where conflicting audio and visual stimuli can lead to perceptual confusion.

The great enhancement of visual information to speech recognition motivates researchers to integrate vision with hearing in a speech recognition system. Potamianos *et al.* [4] reviews the main components of audio-visual speech recognition (AVSR) system. Petajan [5] used dynamic time-warping with visual features extracted from mouth opening and showed that AVSR outperforms either vision or audio lonely. Goldschen [6] applied the hidden Markov models (HMMs) [7] to visual speech recognition and largely improved speech recognition accuracy. Ever since, many approaches have been applied to AVSR. One typical work was done by Matthews *et al*. [1], who integrated video cues and acoustic signals for speech recognition of isolated letters by linearly combining the probabilities output by each recognizer. In the situations when the SNR of acoustic signals is low, the video cues can compensate the acoustic signals, and thus their method significantly improve the recognition accuracy.

Although the aforementioned methods have achieved great successes, they can not generalize enough on unseen dataset because they employs the hand-crafted features specially designed for AVSR system. To overcome this deficiency, we propose a multimodal recurrent neural network model (multimodal RNN) to learn representations from two modalities of sequential data including visual data and audio data. In contrast to the previous hand-crafted feature extraction methods, the multimodal RNN model can automatically learn features from training data. Similar to the model proposed by Mao *et al*. [8] for image captioning, our multimodal RNN model contains an audio part, a visual part, and a fusion part. The audio part is a RNN, which learns sequential feature representation from audio data by utilizing the sequential modelling capability of this RNN. The visual part first utilizes a CNN to extract sequential features from visual data, then these extracted visual features are fed into another RNN as inputs. Thus sequential characteristics of both audio data and visual data are taken into consideration. The fusion part combines the audio part and the visual part by a multimodal layer.

In the following sections, we first describe the related works, then present our multimodal recurrent neural network model, finally report experimental results and conclude this paper.

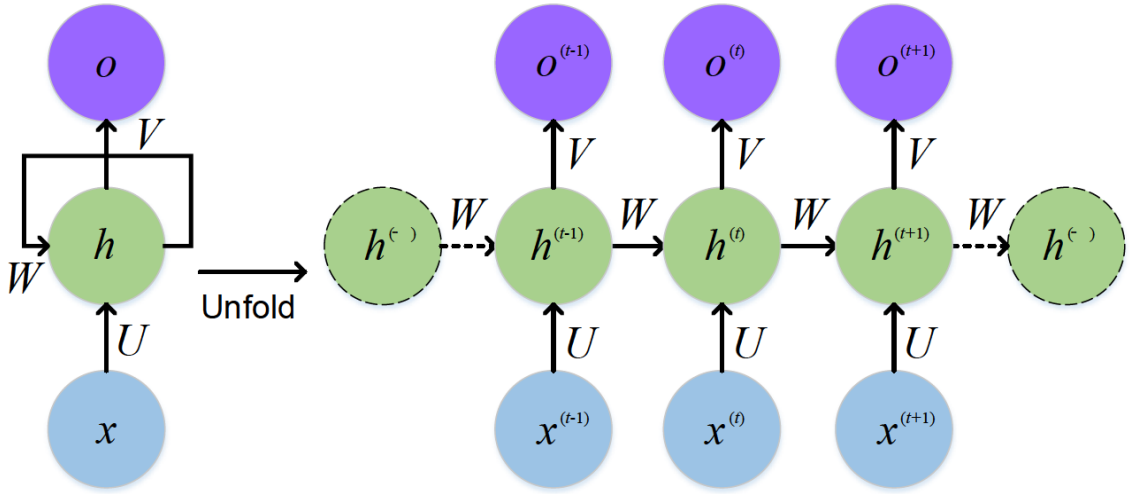

Fig. 1. The standard RNN and unfolded RNN.

II. Related Works

In this section, we briefly reviewed the related recurrent neural network (RNN) models, including: the standard RNNs, bidirectional RNNs, long short-term memory (LSTM), and also the related deep learning models for AVSR.

A. Recurrent Neural Networks

The standard recurrent neural network (RNN) model has a loop on the hidden unit as shown in Figure 1. It has three types of layers: the input layer x, the hidden layer h, and the output layer o. If we unfold this loop, the standard RNN can be considered as copying the same structure multiple times, and the state h of each copy is taken as an input to its successor. Denoting the input layer, hidden layer and output layer at time t as $x^{(t)}$, $h^{(t)}$ and $o^{(t)}$, respectively, the output $o^{(t)}$ can be calculated by:

$$\begin{aligned} a^{(t)} &= b_1 + Wh^{(t-1)} + Ux^{(t)} \\ h^{(t)} &= \sigma\left(a^{(t)}\right) \\ o^{(t)} &= b_2 + Vh^{(t)}, \end{aligned} \tag{1}$$

where b_1 and b_2 are bias vectors, U, V, and W are the weighting matrices of the input-to-hidden connection, hidden-to-output connection, and hidden-to-hidden connection, respectively, and σ is an activation function. Usually, we utilized the sigmoid function $\sigma(x) = (1+e^{-x})^{-1}$ as the activation function.

In many applications, e.g., speech recognition, handwriting recognition, the output $o^{(t)}$ depends on the whole input sequence. To meet this requirement, Schuster *et al.* [9] proposed a bidirectional recurrent neural network which combines two RNNs: one starts from the beginning of the sequence and moves forward through time, the other starts from the end of the sequence and moves backward through time. Figure 2 depicts the structure of a typical bidirectional RNN. Denoting the state of the forward-RNN that moves forward through time as $h^{(t)}$, and the state of the backward-RNN that moves backward through time as $g^{(t)}$, the output $o^{(t)}$ can be computed as:

$$\begin{aligned} a_f^{(t)} &= b_f + W_f h^{(t-1)} + U_f x^{(t)} \\ a_b^{(t)} &= b_b + W_b g^{(t-1)} + U_b x^{(t)} \\ h^{(t)} &= \sigma\left(a_f^{(t)}\right) \\ g^{(t)} &= \sigma\left(a_b^{(t)}\right) \\ o^{(t)} &= c + V_f h^{(t)} + V_b g^{(t)}, \end{aligned} \tag{2}$$

where b_f, U_f, V_f, W_f are parameters for the forward-RNN, and b_b, U_b, V_b, W_b are parameters for the backward-RNN. Due to its high effectiveness, the bidirectional RNN model has been successfully applied in speech recognition [10], [11], and handwriting recognition [12], [13].

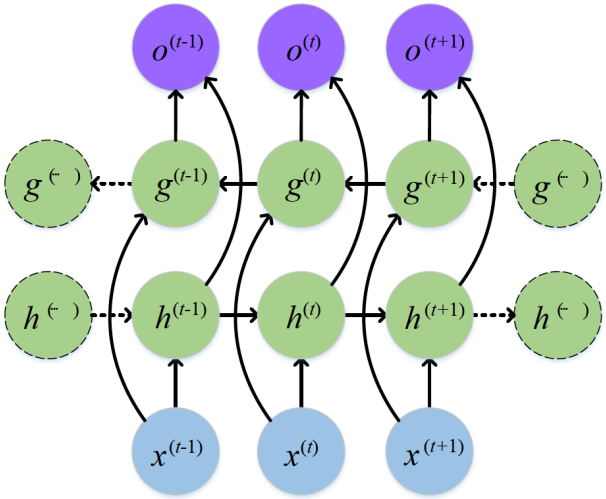

Fig. 2. Structure of a bidirectional RNN.

Training standard RNNs is a challenging due to the gradient vanishing and exploding problem [14]. To address this challenge, Hochreiter and Schmidhuber introduced a long short-term memory (LSTM) model [15]. Particularly, LSTM

replaces the hidden layer of the standard RNNs with a memory cell c, as illustrated in Figure 3, and utilizes three gates to control whether to forget the current cell (forget gate f), whether to read its input (input gate i), and whether to output the new cell value (output gate o). Each of these gates effects to one layer. If the gate is 1, LSTM keeps the value of corresponding layer, and if the gate is 0, it shrinks this value to zero. The definitions of all gates, cell update and output at time t are given as follows:

$$\begin{aligned} i^{(t)} &= \sigma\left(b_i + U_i x^{(t)} + W_i h^{(t-1)}\right) \\ f^{(t)} &= \sigma\left(b_f + U_f x^{(t)} + W_f h^{(t-1)}\right) \\ o^{(t)} &= \sigma\left(b_o + U_o x^{(t)} + W_o h^{(t-1)}\right) \\ c^{(t)} &= f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot \sigma\left(b + U x^{(t)} + W h^{(t-1)}\right) \\ h^{(t)} &= o^{(t)} \odot \tanh\left(c^{(t)}\right), \end{aligned} \tag{3}$$

where $\odot$ represents element-wise multiplication, and b_i, U_i, W_i are parameters for the input gate i, b_f, U_f, W_f are parameters for the forget gate f, b_o, U_o, W_o are parameters for the output gate o, b, U, W are parameters for the input. Because the memory cell contains a node within it, and the node has a self-looped recurrent edge of weight 1, LSTM ensures that the gradients can pass through many time steps and overcomes both vanishing and exploding deficiencies of the gradients. Due to its high efficiency, LSTM has been applied with great success to sequence generation [16], machine translation [17], [18], and image captioning [19], [20].

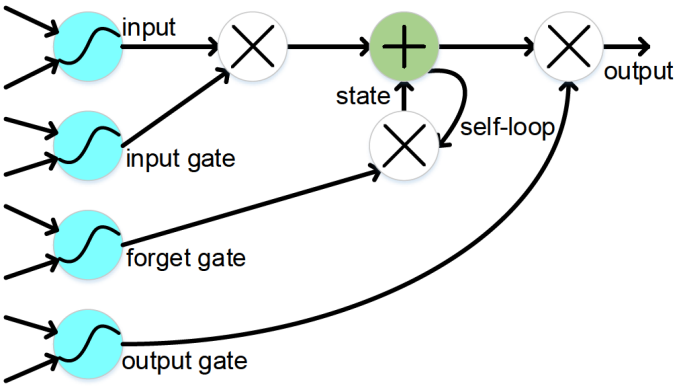

Fig. 3. LSTM cell block.

B. Deep Neural Networks Models for AVSR

Recently, the impressive successes of deep learning both in speech recognition and in image recognition tasks [11], [21]–[23] have appealed researchers to apply it to AVSR. Ngiam *et al*. [24] used deep autoencoders in cross-modality unsupervised feature learning, and obtained a significant improvement on classification accuracy on the AVLetters datasets [1]. However, their methods require too much training data.

Huang and Kingsbury [25] constructed deep belief networks (DBNs) to extract audio and visual features for noise robust speech recognition on a continuously spoken digit recognition task. They utilized a feature fusion method to combine mid-level features learned by single-modality DBNs. Their multimodal DBN reduces digit recognition error rate by 21% relative over the baseline audio-visual system. But DBNs can not model sequential information inherent in the audio and visual data.

Noda *et al*. [26] introduced a connectionist-hidden Markov model system for noise-robust AVSR. They first utilized a deep denoising autoencoder to acquire noise-robust audio features, then utilized a convolutional neural network (CNN) to extract visual features from raw mouth area images, and finally applied a multi-stream HMM to integrate the acquired audio and visual HMMs. That work is similar to our multimodal RNN model, but we use RNN to model sequential data, rather than HMM. And we do not need a deep denoising autoencoder to extract features from audio data.

Mroueh *et al*. [27] presented a deep neural network (DNN) architecture that used a bilinear softmax layer to incorporate class specific correlations between audio and video modalities, and yielded a phone recognition error rate of 34.03% on the IBM large vocabulary audio-video studio dataset. However, this architecture completely ignore the sequential characteristics of the data either.

III. Multimodal Recurrent Neural Network Model

Although the existing works have applied several deep learning models to AVSR, few of them can simultaneously consider the sequential characteristics of audio data and visual data. In this paper, we propose a multimodal RNN model to address this problem. Figure 4 illustrates the structure of our multimodal RNN model, which is composed of three components including a visual part, an audio part, and a fusion part. The visual part contains a CNN plus bidirectional LSTM layer for visual data, and the audio part contains a bidirectional LSTM layer for audio data. Both parts contain weighted state layers to generate semantically consistent outputs for fusion. Based on the weighted state layers, the multimodal RNN utilizes a multimodal layer for fusing both modalities and a softmax layer for output.

Let n_V denote the number of input video frames, and n_A denote the number of input audio frames. These video frames are first processed by an identical CNN and then transformed into n_V one-dimensional feature vectors. These feature vectors are then inputted into a bidirectional LSTM RNN layer, where the length of input sequence is n_V. Similarly, these audio frames are inputted into another bidirectional LSTM RNN layer, where the length of input sequence is n_A. Let $o_V^{(i)}$ and $o_A^{(j)}$ denote the output of video bidirectional LSTM RNN and audio bidirectional LSTM RNN corresponding to the i-th and j-th input, respectively, we can compute the weighted state layer for video data o_V and audio data o_A as follows:

$$o_V = \sum_{i=1}^{n_V} w_V^{(i)} o_V^{(i)}, \tag{4}$$

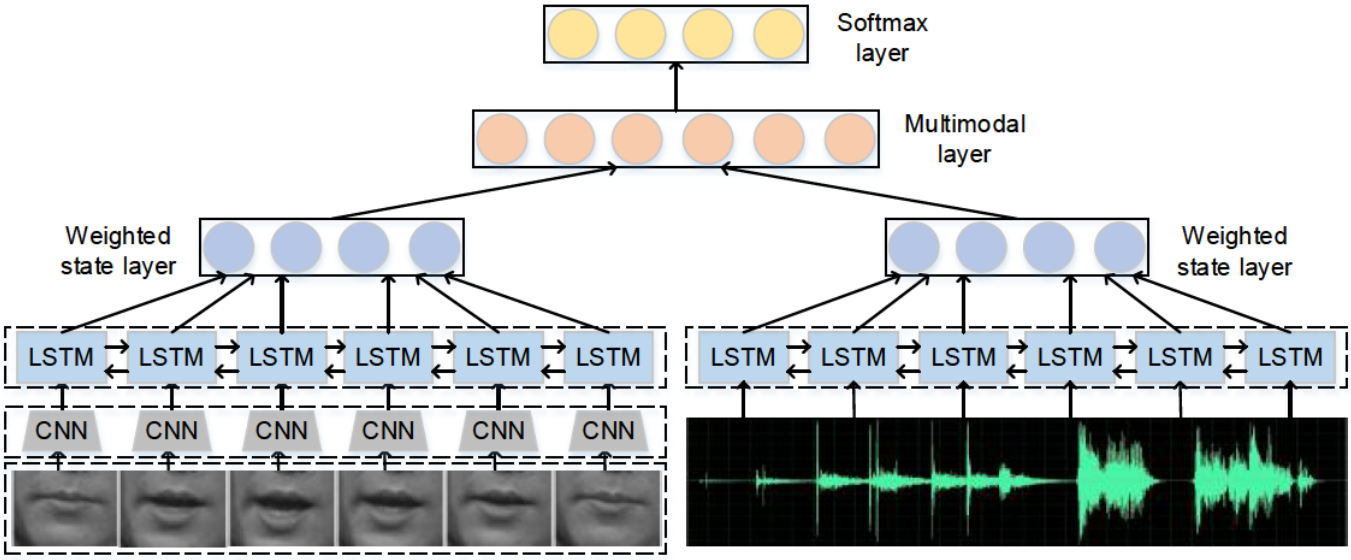

Fig. 4. Structure of our multimodal RNN model.

$$o_A = \sum_{j=1}^{n_A} w_A^{(j)} o_A^{(j)}, \tag{5}$$

where $w_V^{(i)} \in \mathbb{R}$ $(1 \le i \le n_V)$ and $w_A^{(j)} \in \mathbb{R}$ $(1 \le j \le n_A)$ are both learnable parameters.

In the fusion part, the multimodal RNN concatenates both $o_V^{(i)}$ and $o_A^{(j)}$ together and builds a multimodal layer which is fully connected to both weighted state layers. Assume $o_V \in \mathbb{R}^{d_V}$, $o_A \in \mathbb{R}^{d_A}$, and the number of neurons in the multimodal layer is d_M, then we compute output of the multimodal layer o_M as follows:

$$o_M = \sigma\left(W_V^{(M)} o_V + W_A^{(M)} o_A + b^{(M)}\right), \tag{6}$$

where $W_V^{(M)} \in \mathbb{R}^{d_M \times d_V}$, $W_A^{(M)} \in \mathbb{R}^{d_M \times d_A}$, and $b^{(M)} \in \mathbb{R}^{d_M}$ are the learnable parameters for the multimodal layer. Above the multimodal layer, the multimodal RNN puts a Softmax layer to obtain the predictions, i.e., the probability of the input example belonging to each class:

$$o = Softmax\left(W^{(S)} o_M + b^{(S)}\right), \tag{7}$$

where $W^{(S)} \in \mathbb{R}^{C \times d_M}$ and $b^{(S)} \in \mathbb{R}^C$ are the parameters for the Softmax layer with C denoting the number of classes.

A. Multimodal RNN Variants

As shown in Figure 4, the proposed multimodal RNN contains an audio part, a visual part, and a fusion part, which are constructed by bidirectional LSTM RNN, CNN plus bidirectional LSTM RNN, and multimodal layer, respectively. We can vary this configuration and obtain various multimodal RNN variants by replacing any part with the corresponding variant.

For the audio part, we can explore a simpler variant: unidirectional LSTM RNN that moves from the beginning of the audio sequence to the end of the audio sequence. Similarly, for the visual part, we can explore the CNN plus unidirectional LSTM RNN variant. To demonstrate the effectiveness of RNN on sequential data, we can also explore the CNN variant without LSTM RNN. For this CNN variant, we use an early fusion strategy: regard the input n_V video frames as n_V channels input to the first convolutional layer. This CNN structure contains two convolutional layers, each of which is followed by a pooling layer and a fully connected layer. To reduce parameters, the CNN structure in the CNN plus LSTM RNN variant is simpler, which has only one convolutional layer, one pooling layer and one fully connected layer. For the fusion part, we explore three choices: (1) taking the learned features by the visual part as the initial state for the audio part, and the output probability of the audio part is the output probability of the architecture, (2) taking convex combination of both visual logits and audio logits (parameter for this convex combination is trainable), and then taking the logits as inputs to the Softmax layer, simultaneously taking output of the Softmax layer as output probability of the architecture, and (3) concatenates the learned features by visual part and the original audio inputs as the inputs to audio part, where its output is the output probability of the architecture .

B. Training the Model

To train the proposed multimodal RNN model, we adopt the average cross-entropy cost function as the loss function and incorporate a regularization term to avoid overfitting, i.e.,

$$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} \log\left(o^{(i)}\right) + \left(1 - y^{(i)}\right) \log\left(1 - o^{(i)}\right)\right) + \lambda_\theta \cdot \|\theta\|_2^2, \tag{8}$$

where N denotes the number of samples in the training set, $y^{(i)}$ denotes the class label of the i^{th} sample, $o^{(i)}$ denotes the output probability of our model corresponding to $y^{(i)}$, and

θ represents the model parameters. The objective of training is to minimize the cost function, which is differentiable. The stochastic gradient descent algorithm is utilized to learn the model parameters in and end-to-end manner. We implemented the proposed multimodal RNN model and conducted the following experiments in TensorFlow [28] on a computer with Intel Xeon E5-2680 CPU, 256GB memory and 4 K80 GPU processors. We train each model with 30,000 iterations and save model parameters every 500 iterations. We evaluate each model on testing dataset with these saved parameters and report the highest accuracy as the result.

IV. Experiments

A. Datasets

We evaluated our multimodal RNN model on the benchmark audio-visual database termed AVletters [1]. The AVletters database contains 10 speakers including five males (two with moustaches) and five females, and each has three repetitions of the isolated letters A to Z, thus the total number of utterances is 780.

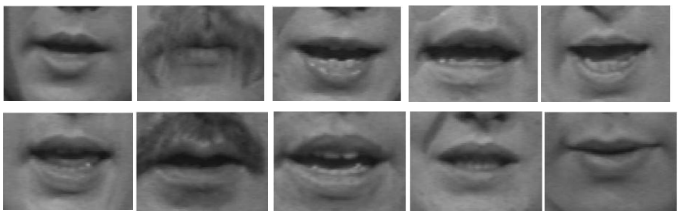

Fig. 5. Example frame for the 10 speakers in the AVletters database.

The video data captures the video of mouth region in 60×80 pixels. The number of video frames for these utterances ranges from 23 to 79. For each of the 10 speakers, Figure 5 shows one of their example frames. The audio data are captured by using 26 mel-frequency cepstral coefficients (MFCC), and the number of MFCC frames for these utterances ranges from 12 to 40. All the utterances start and end with the mouth closed, and the corresponding audio data include periods of silence-utterance-silence.

For the training/testing dataset splitting, we followed the way in [1]. All models were trained for all letters using examples from all the speakers in the database. Following [1], the training set is composed of the first two utterances of each of the letters from all speakers (520 utterances), and the test set is composed of the third utterance from all speakers (260 utterances).

For each training utterance, we extracted multiple video frame to build multiple blocks and extracted the corresponding audio frame to build multiple blocks, and set the number of frames for these two kinds of blocks to 6 and 12, respectively. By using this procedure, we obtained 9165 training examples. While for each testing utterance, we extracted central 6 frames from the video data, and central 12 frames from the audio data, and obtained 260 testing examples.

B. Visual-Only Recognition

For visual-only speech recognition task, we used only the visual part of our multimodal RNN model and the visual data for both training and testing. The visual part has three variants: "CNN only", "CNN plus LSTM RNN", and "CNN plus bidirectional LSTM RNN". The recognition accuracies of these three variants and other methods are listed in Table I. Due to the random initialization, each of these variants was tested 10 individual times, and the average accuracies and standard deviations are reported.

TABLE I
VISUAL-ONLY RECOGNITION ACCURACIES AND STANDARD DEVIATIONS FOR DIFFERENT FEATURE REPRESENTATION.

Feature Representation	Accuracy
Multiscale Spatial Analysis [1]	44.6%
Local Binary Pattern [2]	58.85%
Video-Only Deep Auto-encoder [24]	$\mathbf{64.4\% \pm 2.4\%}$
DCT+DBNF-C [29]	58.1%
RTMRBM [30]	64.63%
CNN only	$49.9\% \pm 1\%$
CNN & LSTM RNN	$\mathbf{57.7\% \pm 0.8\%}$
CNN & bidirectional LSTM RNN	$49.4\% \pm 0.9\%$

In Table I, "Multiscale Spatial Analysis" and "Local Binary Pattern" were hand-crafted features. The "Video-Only Deep Auto-encoder" first used unsupervised feature learning on a large cross-modality dataset, and all these three feature representations were specifically designed for this visual-only recognition task. "DCT+DBNF-C" combines discrete cosine transform features and deep bottleneck features. "RTMRBM" utilizes a recurrent temporal multimodal RBM and models multimodal sequences by transforming the sequence of connected multimodal RBMs into a probabilistic series model. In contrast, features learned by our three variants are not hand-crafted, and these three variants are neither pre-trained on other datasets nor specifically designed for this task, they are only the output of the visual part of our multimodal RNN model.

From Table I, we can observe that the highest accuracy of our method (shown in bold) is competitive to the best hand-crafted features, but inferior to the features learned by deep auto-encoder model. The "CNN plus LSTM RNN" variant

TABLE II
AVERAGE ACCURACIES AND STANDARD DEVIATIONS OF DIFFERENT METHODS ON CLEAN AND DEGRADED AUDIO DATA.

SNR \ Methods	Unidirectional LSTM RNN	Bidirectional LSTM RNN
Clean	$74.6\% \pm 2.4\%$	$\mathbf{75.6\% \pm 1.6\%}$
20db	$73.5\% \pm 2\%$	$\mathbf{75.0\% \pm 2\%}$
10db	$69.8\% \pm 1.7\%$	$\mathbf{70.0\% \pm 2.3\%}$
6db	$\mathbf{64.1\% \pm 2.9\%}$	$60.5\% \pm 2.8\%$
3db	$\mathbf{50.6\% \pm 2.4\%}$	$46.7\% \pm 1.7\%$
0db	$\mathbf{40.1\% \pm 1.4\%}$	$35.3\% \pm 1.7\%$

TABLE III
RECOGNITION ACCURACIES OF DIFFERENT FUSION PART VARIANTS ON CLEAN AND DEGRADED DATA.

SNR \ Methods	multimodal layer	initial state	linear combination	concatenation
Clean	$\mathbf{84.4\% \pm 1.7\%}$	$81.0\% \pm 1.7\%$	$81.4\% \pm 1.6\%$	$81.5\% \pm 1.2\%$
20db	$\mathbf{81.4\% \pm 1.8\%}$	$79.2\% \pm 1.8\%$	$79.8\% \pm 1.2\%$	$79.1\% \pm 1.4\%$
10db	$\mathbf{80.1\% \pm 1.1\%}$	$79.4\% \pm 1.8\%$	$78.0\% \pm 1.1\%$	$78.7\% \pm 1.8\%$
6db	$76.8\% \pm 1.3\%$	$76.0\% \pm 1.7\%$	$\mathbf{77.1\% \pm 1.0\%}$	$76.9\% \pm 1.6\%$
3db	$72.4\% \pm 1.5\%$	$\mathbf{72.8\% \pm 2.0\%}$	$71.3\% \pm 1.3\%$	$72.3\% \pm 1.0\%$
0db	$64.5\% \pm 0.9\%$	$\mathbf{67.1\% \pm 2.5\%}$	$66.3\% \pm 1.2\%$	$66.6\% \pm 1.0\%$

outperforms the "CNN only" variant, demonstrating the effectiveness of RNN on dealing with sequential data. However, in contrast to our expectation, the "CNN plus bidirectional LSTM RNN" variant behaves more poorly than the "CNN plus unidirectional LSTM RNN" variant. That is perhaps because that the training data is insufficient to train the "CNN plus bidirectional LSTM RNN" variant, which had the maximum number of parameters among these variants.

C. Audio-Only Recognition

To get some intuitions on how visual data improves AVSR, we also reported audio-only speech recognition results. Similar to the setting of above visual-only recognition experiments, we used only the audio part of our multimodal RNN model and the audio data for both training and testing. The audio part has two variants: "unidirectional LSTM RNN", and "bidirectional LSTM RNN". To demonstrate the robustness of the method to noise, following [1], we added different levels of white Gaussian noise to the original clean audio data, and the SNRs of these degraded audio data are: 0db, 3db, 6db, 10db, 20db. The recognition accuracies for these two variants on different audio data are listed in Table II. Each experiment was repeated 10 times, and both the average accuracies and standard deviations are reported.

From Table II, we can see that, for the relatively large SNR audio data, the bidirectional LSTM RNN outperforms unidirectional LSTM RNN, while the things turn inverse for small SNR audio data. This phenomenon may be explained as follows: for large SNR audio data, the effects of Gaussian white noise is slight, while for the smaller SNR audio data, the effects of noise change significant, and the bidirectional procedure may amplify the effects of noise.

D. Audio-Visual Recognition

For the audio visual recognition, our aim is to investigate how visual information helps to improve speech recognition. We first selected the best fusion model among the four fusion part variants, which are denoted as "multimodal layer", "initial state", "linear combination", and "concatenation", respectively. The recognition accuracies for these four fusion part variants are shown in Table III, where the visual part was "CNN plus LSTM RNN", and the audio part was "LSTM RNN". Similarly, such experiment was repeated 10 times, and the average values and standard deviations are reported. Both visual data and audio data were utilized for training and testing, while only audio data were degraded by white Gaussian noise.

From Table III, the "multimodal layer" variant achieves the best performance in most cases, especially for large SNR, thus in the following experiments, we evaluated the "multimodal layer" variant only. Figure 6 shows the recognition accuracies over a range of SNR for all models when the fusion part is "multimodal layer". "CNN & LSTM RNN" means the visual part is CNN and the audio part is "LSTM RNN", and the rest titles for each sub-figure are pronounced in the same way. All these models successfully improve the speech recognition accuracy by incorporating visual information, and the benefits of visual information became much more significant as the white Gaussian noise level increases. The performance achievements obtained by using all six configurations are consistently high, and the "CNN plus LSTM RNN" visual model and the "bidirectional LSTM RNN" audio model are slightly better. Due to the limitation of the number of training examples, the "CNN plus bidirectional LSTM RNN & bidirectional LSTM RNN" model did not perform the best. This observation inspires us to choose appropriate configuration for a specific task according to its characteristics.

TABLE IV
THE BEST RECOGNITION ACCURACIES FOR DIFFERENT METHODS ON CLEAN AND DEGRADED AUDIO DATA.

SNR \ Methods	multimodal RNN	Matthews *et al.* [1]
Clean	$\mathbf{87.7\%}$	86%
20db	$\mathbf{85.0\%}$	67%
10db	$\mathbf{83.0\%}$	53%
6db	$\mathbf{81.0\%}$	48%
3db	$\mathbf{77.0\%}$	46%
0db	$\mathbf{70.0\%}$	42%

We also compared our best results achieved by all variants with the best results of Matthews *et al*. [1] over a range of SNR, as shown in Table IV. The best results of our multimodal RNN model consistently are consistently superior to those of Matthews *et al*., especially for the smaller SNR. That is because, as the noise level increased, the performance of Matthews et al. decreased dramatically, while the decline of recognition accuracy for our method was not so remarkable.

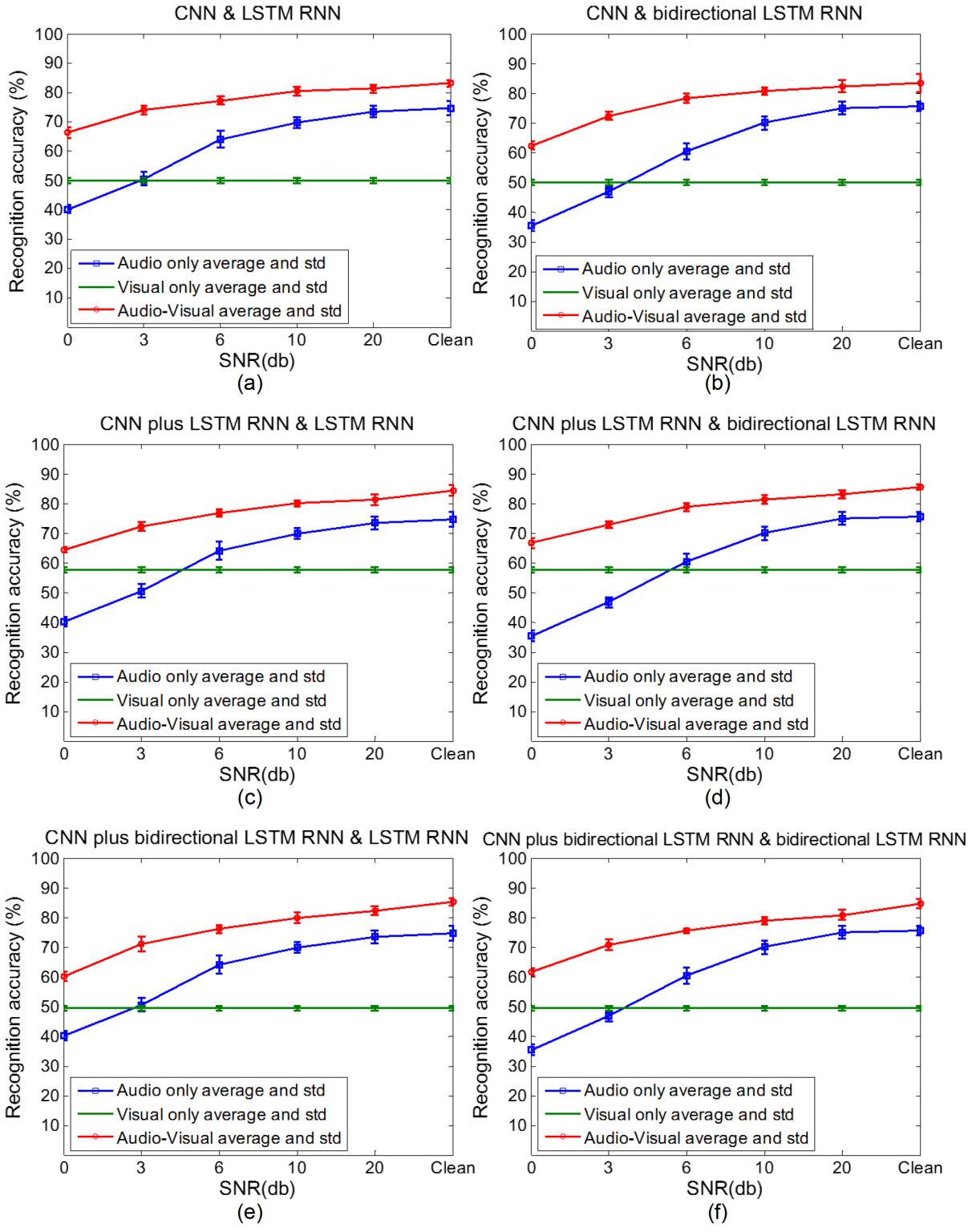

Fig. 6. Audio-Visual recognition results for all "multimodal layer" fusion part models. (a) "CNN" visual part model and "LSTM RNN" audio part model, (b) "CNN" model and "bidirectional LSTM RNN" model, (c) "CNN plus LSTM RNN" model and "LSTM RNN" model, (d) "CNN plus LSTM RNN" model and "bidirectional LSTM RNN" model, (e) "CNN plus bidirectional LSTM RNN" model and "LSTM RNN" model, (f) "CNN plus bidirectional LSTM RNN" model and "bidirectional LSTM RNN" model.

This observation demonstrates the robustness of the proposed multimodal RNN model to noise.

V. Conclusion

We propose a multimodal recurrent neural network (multimodal RNN) model for audio visual speech recognition. The

model comprises of a CNN followed by a LSTM RNN to handle visual modality data, a LSTM RNN to handle audio modality data, and a multimodal layer to fuse the outputs of both modalities. Experimental results confirm that the proposed multimodal RNN model and its variants succeed to incorporate visual information into speech recognition, and outperform the best known audio visual speech recognition results on AVletters, especially for noisy data. For future work, we consider to take deep CCA [31] as our fusion part to improve the correlation of audio part and visual part.

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (under grant NO. 61502515), Research Fund for the Doctoral Program of Higher Education of China, SRFDP (under grant No. 20134307110017) and National High Technology Research and Development Program ("863" program) of China (under grant No. 2015AA020108).

REFERENCES

[1] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 198–213, 2002.

[2] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[3] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[4] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[5] E. D. Petajan, "Automatic lipreading to enhance speech recognition (speech reading)," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1984.

[6] A. J. Goldschen, "Continuous automatic speech recognition by lipreading," 1993.

[7] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *The Bell System Technical Journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[8] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[9] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5, pp. 602–610, 2005.

[11] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[12] A. Graves, M. Liwicki, H. Bunke, J. Schmidhuber, and S. Fernández, "Unconstrained on-line handwriting recognition with recurrent neural networks," in *Advances in Neural Information Processing Systems*, 2008, pp. 577–584.

[13] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[14] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[17] 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.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[20] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *arXiv preprint arXiv:1512.00567*, 2015.

[23] A.-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[25] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7596–7599.

[26] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Audio-visual speech recognition using deep learning," *Applied Intelligence*, vol. 42, no. 4, pp. 722–737, 2015.

[27] Y. Mroueh, E. Marcheret, and V. Goel, "Deep multimodal learning for audio-visual speech recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2130–2134.

[28] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2016.

[29] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *ICASSP*, 2016.

[30] D. Hu, X. Li, and X. Lu, "Temporal multimodal learning in audiovisual speech recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3574–3582.

[31] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. 1247–1255.