LCANet: End-to-End Lipreading with Cascaded Attention-CTC

Article · March 2018

4 authors, including:

Kai Xu
Arizona State University
13 PUBLICATIONS   **97** CITATIONS

SEE PROFILE

LCANet: End-to-End Lipreading with Cascaded Attention-CTC

Kai Xu[1], Dawei Li[2], Nick Cassimatis[2], Xiaolong Wang[2]*

[1] Arizona State University, Tempe, AZ 85281, USA

[2] Samsung Research America, Mountain View, CA 94043, USA

Machine lipreading is a special type of automatic speech recognition (ASR) which transcribes human speech by visually interpreting the movement of related face regions including lips, face, and tongue. Recently, deep neural network based lipreading methods show great potential and have exceeded the accuracy of experienced human lipreaders in some benchmark datasets. However, lipreading is still far from being solved, and existing methods tend to have high error rates on the wild data. In this paper, we propose LCANet, an end-to-end deep neural network based lipreading system. LCANet encodes input video frames using a stacked 3D convolutional neural network (CNN), highway network and bidirectional GRU network. The encoder effectively captures both short-term and long-term spatio-temporal information. More importantly, LCANet incorporates a cascaded attention-CTC decoder to generate output texts. By cascading CTC with attention, it partially eliminates the defect of the conditional independence assumption of CTC within the hidden neural layers, and this yields notably performance improvement as well as faster convergence. The experimental results show the proposed system achieves a 1.3% CER and 3.0% WER on the GRID corpus database, leading to a 12.3% improvement compared to the state-of-the-art methods.

Keywords-Lipreading, ASR, attention mechanism, CTC, cascaded attention-CTC, deep neural network, 3D CNN, highway network, Bi-GRU.
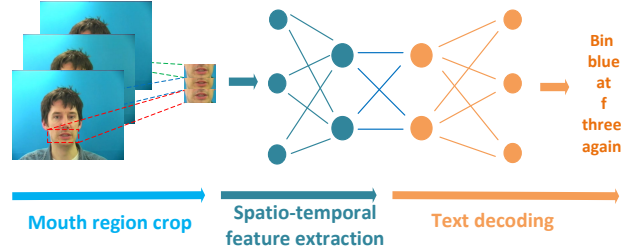
Figure 1: The proposed end-to-end lipreading system includes three major steps. (1) The mouth regions are cropped from the input video frames from aligned faces. (2) An encoder extracts spatio-temporal features from the input sequence. (3) The cascaded attention-CTC decoder generates text from encoded hidden features.

I. INTRODUCTION

Lipreading, the ability to understand what people are saying from only visual information, has long been believed to be a superpower as shown in some science fiction movies. The recent development of deep learning techniques has begun to unveil the secret behind it and demonstrated the potential to make it available for commons. Compared to the acoustic-based speech recognition technique, lipreading solves a much broader range of practical problems, such as aids for hearing-impaired persons, speech recognition in a noisy environment, silent movie analysis, etc.

Lipreading is a challenging problem because the differences among some of the phonemes or visemes are subtle (e.g., the visemes corresponding to "p" and "b") although

their corresponding utterances can be easily distinguished. Such fine distinction is the major obstacle for humans to read from lips and as reported previously only around 20% reading accuracy can be achieved [1]. On the other hand, recent work shows that machine lipreading has got remarkable progress especially with the advancement of deep learning techniques [2]–[6].

Deep learning techniques have been the cornerstone for recent success [7]–[11] of many traditionally hard machine learning tasks [12]–[17]. Generally, deep learning based lipreading methods follow the state-of-the-art sequential modeling solutions that have been widely applied to problems like acoustic based speech recognition [18]–[20] and neural machine translation [21]–[24]. Two most widely used approaches are Connectionist Temporal Classification (CTC) [25] and the attention-based sequence to sequence (seq2seq) model [21].

CTC approach works well on acoustic-based speech recognition. For lipreading problem, [3] uses the CTC loss function [25] to train an end-to-end deep neural network, named "LipNet", and the model outperforms experienced human lipreaders on the GRID benchmark dataset [26]. However, CTC loss function assumes conditional independence of separate labels (i.e., the individual character symbols), and each output unit is the probability of observing one particular label at a time. Therefore, although CTC is applied on top of Recurrent Neural Networks (RNNs), it tends to focus too much on local information (nearby frames) [5]. This may work well for acoustic phonemes, but

is not well fit for predicting visemes which require longer context information to discriminate their subtle differences.

Attention-based seq2seq models were initially introduced for neural machine translation [21], for which the input sequence and output sequence can be chronologically inconsistent. Such kind of model has also been adapted to solve the lipreading problem [5]: an encoder encodes the input frames (optionally with audios) into hidden features which are fed into an attention-based decoder to obtain the text output. By eliminating the conditional independence assumption, the attention-based method achieves better performance than CTC. However, recent work [27], [28] shows that the attention model is too flexible to predict proper alignments for speech recognition, and is hard to train from scratch due to the misalignment on longer input sequences.

In this paper, we propose a cascaded attention-CTC deep learning model, LCANet, for end-to-end lipreading. LCANet incorporates an encoder that is composed of 3D convolution, highway network and bi-directional GRU layers, which is able to effectively capture both short-term and long-term spatio-temporal information from the input video frames. To deal with the conditional independence assumption of the CTC loss, the encoded features are fed into a cascaded attention-CTC decoder. The proposed decoder explicitly absorbs information from the longer context from the hidden layers using the attention mechanism. Extensive evaluation result shows that LCANet achieves the state-of-the-art performance and faster convergence than existing methods.

The contribution of this paper is as follows:

- We propose an end-to-end lipreading deep neural network architecture, LCANet, which relies purely on the visual information.
- LCANet leverages a cascaded attention-CTC decoder to generate text from encoded spatio-temporal features. Such design partially eliminate the defect of the conditional independence assumption in the vanilla CTC-based method.
- We demonstrate the performance improvement compared with existing work. The benefits of the cascaded attention-CTC decoder are also discussed.

II. RELATED WORK

A. Deep Learning for Speech Recognition

Recent advancement of speech recognition technique is primarily due to the prosperity of deep learning research [18], [19], [29]–[33], and the reported speech recognition error rate is already below that of a human. This makes speech a convenient input method for computers as well as mobile or embedded devices. Together with the advancement of natural language understanding techniques [34]–[37] and recent software [38]–[40] and hardware acceleration [41], [42] methods, the intelligent virtual personal assistant has become a hot market (e.g., Siri of iPhone, Alexa of Amazon Echo, Google Now, Samsung Bixby, etc.).

The most successful deep speech recognition models are built with either CTC [25] or the sequence-to-sequence (seq2seq) model. The CTC based methods [18], [19] predict a label (i.e., a character label) for each audio frame and is trained to optimize the alignment between the prediction and the target label. Seq2seq models for speech recognition [32] improved significantly with attention mechanism [21] which enables more effective flow of information than a recurrent neural network. Attention-based seq2seq model has exceeded the CTC based model on several benchmark datasets. A recent work [43] shows that attention-based model may not perform well on noisy speech data, and proposes a joint decoding algorithm for end-to-end ASR with a hybrid CTC/attention architecture. The hybrid method minimizes a joint loss which is the weighted sum of both CTC loss and attention loss, and this effectively utilizes both advantages in decoding. Different from [43], LCANet uses a cascaded attention-CTC decoder. The benefit of such design includes (1) it avoids tuning the weight parameter in the loss function and (2) it generates the output directly from the softmax output instead of merging the results from both CTC and attention branches.

B. Machine Lipreading

Generally, machine lipreading contains two phases: visual feature extraction from the lip movement and text prediction based on classifiers. The traditional visual features extracted from the mouth region of interest (ROI) can be broadly classified into four categories: 1) visual feature extracted directly from images through image transformations such as DCT, DWT [44]–[47]; 2) geometry-based features, e.g., height, width, area of the lip region; 3) motion-based features, e.g., optical flow of the motion of the lip region [48]; 4) model-based features. The shape and appearance of the lip ROI are explicitly modeled by active shape models (ASMs) [49] or active appearance models (AAMs) [50], [51]. The quality of such model-based features are highly dependent on the accuracy of hand-labeled training data which requires lots of efforts to obtain. To utilize the temporal dynamics in the extracted visual features, a dynamic classifier such as hidden Markov models (HMMs) is used for classification [44], [52]. However, such classifiers are based on the conditional independent assumption and usually fails to model the long-term dependencies.

Recently, deep neural network emerges to be used for lipreading and achieves promising performance compared to the conventional methods. [6] propose to use CNN for visual feature extraction, and a Gaussian mixture observation model for an isolated word recognition task by modeling the temporal dependence of the generated phoneme label sequences. [3] proposes an end-to-end sentence-level lipreading model. A stacked 3D CNN are used for dynamic visual feature extraction. The mapping between lip movement and the corresponding text representation is modeled
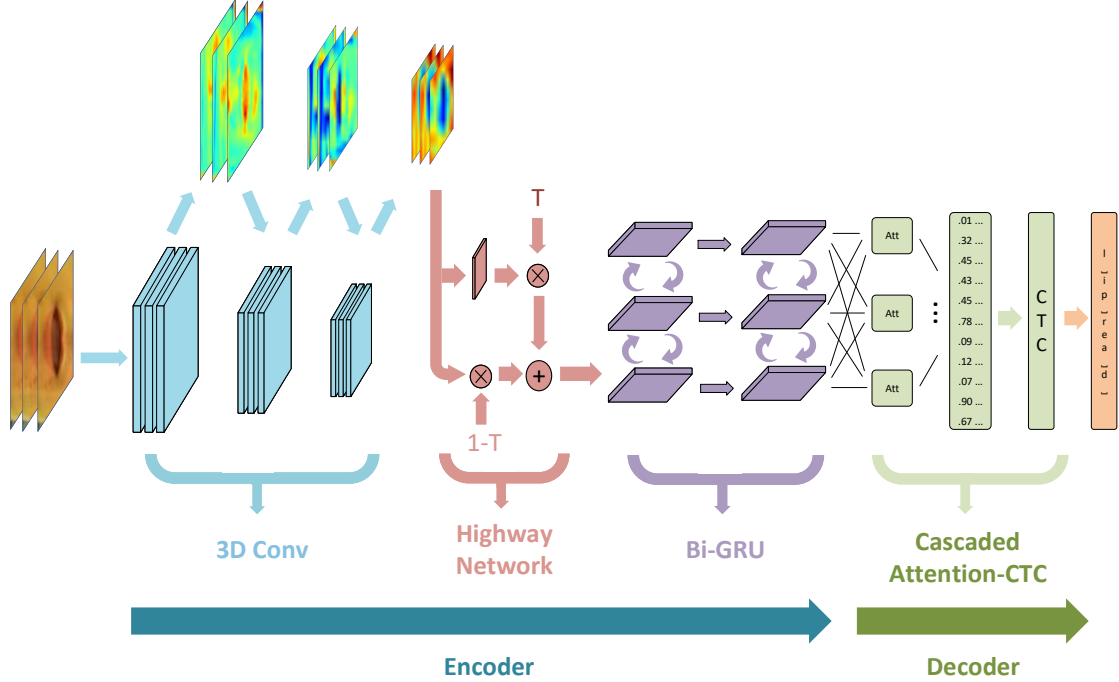
Figure 2: The architecture of LCANet. The pre-processed mouth-region frames are used as input, which are processed by the following 3D-CNN, the highway network, the Bi-GRU and the cascaded attention-CTC decoder. A softmax layer calculates the probability distribution of the output of the attention module. Each attention unit is related to one character output. The CTC decoder generates text predictions from the probability distribution.

by the Gated Recurrent Unit (GRU). CTC loss is used for calculating the difference between the predictions and the text labels which have different lengths. [2] presents an end-to-end lipreading system which admits two input streams: mouth images that encode static information and mouth differences that capture local temporal dynamics. The encoding layers are pre-trained using RBMs and the temporal dynamics are modeled by an LSTM. The fusion of the two streams is performed through a Bi-directional LSTM (BLSTM). [5], [53] propose to use combined visual speech features for multi-modal speech recognition where video and audio features are fused as input. The former method is based on the '*EESEN*' framework [19], which uses CTC loss function to solve the temporal alignment problem. The latter approach is based on the '*Watch, Listen, Attend and Spell*' network [33], which leverages the attention mechanism to model rich contextual information between the input video/audio frames and the character predictions.

III. LCANET ARCHITECTURE

In this section, we describe the neural network architecture of the proposed LCANet (Figure 2). The network is composed of a spatio-temporal video encoder which maps a sequence of video frames $\mathbf{x} = (x_1, ..., x_n)$ to a sequence of continuous latent representations $\mathbf{h} = (h_1, ..., h_n)$, and a cascaded CTC-attention decoder which generates the output

sequence of character symbols $\mathbf{y} = (y_1, ..., y_m)^1$. In the following, we first introduce the spatio-temporal video encoder, and then we present the cascaded attention-CTC decoder.

A. The Spatio-temporal Video Encoder

The video encoder of LCANet has three components: *3D-CNN*, *highway network*, and *Bidirectional GRU (Bi-GRU)*.

3D-CNN: LCANet feeds the input video frames into a 3D convolution neural network (3D-CNN) [54] to encode both visual and short temporal information. Compared to 2D-CNN which only deals with the spatial dimensions , 3D-CNN also convolves across time steps: a group of adjacent k frames $x_{t:t+k}$ are processed by a set of P temporal convolutional filters $C_p(w, h, k, c)$, where w and h are the width and height of the filter in the spatial dimension and c is the number of channels. The value of k in the filter is small (e.g., 3 or 5) to capture only short temporal information, which in lipreading can be mapped to an individual or two nearby visemes.

Highway Network: We stack two layers of highway networks [55] on top of the 3D-CNN. The highway network module (as shown in Figure 3) has a pair of *transform* gate t and *carry* gate $1-t$ that allows the deep neural network to carry some input information directly to the output. The

---

[1]For lipreading, m is normally smaller than n, meaning each character symbol corresponds to 1 or more video frames.

highway network is formulated as follows:

$$\mathbf{t} = \sigma(\mathbf{W}_T\mathbf{x} + \mathbf{b}_T), \tag{1}$$

$$\mathbf{g} = \mathbf{t} \odot \sigma(\mathbf{W}_H\mathbf{x} + \mathbf{b}_H) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{x}. \tag{2}$$

where $\mathbf{x}$ is the input, $\mathbf{g}$ is the network output, and $\odot$ denotes element-wise multiplication. A recent study on language modeling [34] shows that, by adaptively combining local features detected by the individual filters of CNN, highway network enables encoding of much richer semantic features. We show that such a design stably enhances the modeling capability of the encoder. An alternative design is to use a 3D CNN with a ResNet, which however highly increases the complexity of the encoder (e.g., a 34-layer ResNet is used in [56]).

Bi-GRU: LCANet encodes long-term temporal information with the Bi-GRU [57]. 3D-CNN only captures short viseme-level features, however for lipreading problem, there are some visemes visually very similar to each other. Such ambiguity can only be distinguished with longer temporal context which can be well modeled with the state-of-the-art recurrent neural networks such as GRU. A GRU has two gates: 1) a reset gate r which determines how much of the previous memory state should be mixed with the current input; 2) an update gate z which determines how much of the previous memory state should be kept with the current hidden state. The standard formulation is:

$$
\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_z\mathbf{g}_t + \mathbf{U}_z\mathbf{h}_{t-1} + \mathbf{C}_z\mathbf{c}_t + b_z), \\
\mathbf{r}_t &= \sigma(\mathbf{W}_r\mathbf{g}_t + \mathbf{U}_r\mathbf{h}_{t-1} + \mathbf{C}_r\mathbf{c}_t + b_r), \\
\tilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h\mathbf{g}_t + \mathbf{U}_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{C}_p\mathbf{c}_t), \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t,
\end{aligned} \tag{3}
$$

where r_t and z_t is the reset and update gate, respectively, C is the linear projection matrix for integrating the context vector, and g is the output of the highway network. To fully exploit both the past and future context, we use Bi-GRU to capture both the forward and backward information flow. Compared to the vanilla GRU, Bi-GRU can leverage long-range context to improve the motion modeling capacity.

B. Cascaded Attention-CTC Decoder

1) Connectionist temporal classification (CTC): Many state-of-the-art ASR systems train an end-to-end deep neural network using the Connectionist Temporal Classification (CTC) approach [25]. CTC is an objective loss function that eliminates the requirement of explicit alignment between the input sequence and the target sequence during the model training process. Given an input $\mathbf{x}$ of length T, the output of the neural network is a softmax layer which indicates the possibility of emitting the symbol (either a character or '$blank'$) at every time-step. For the same target sequence, there can be many alignment candidates since the blank symbol can be inserted at any position. For example, the paths (x, -, -, y, -, z), (-, x, -, y, -, z) and (-, x, -, y, z, -) all
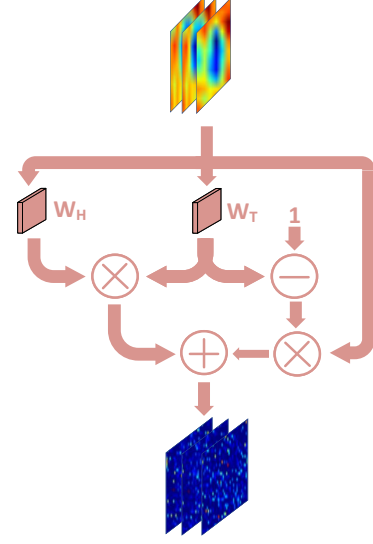

Figure 3: Information flow in the highway network. W_t is the weight of the transform gate and $1 - W_t$ is the weight of the carry gate. The highway network leverages adaptive gating units to regulate the information flow, which allows gradients to flow through many layers.

map to the same target label of (x, y, z). Now, the possibility of generating one correct path π for the given input $\mathbf{x}$ is:

$$p(\pi|\mathbf{x}) = \prod_{t=1}^{T} p(\pi_t|\mathbf{x}). \tag{4}$$

The probability of generating the target sequence $\mathbf{l}$ is the sums over all possible paths π:

$$p(\mathbf{l}|\mathbf{x}) = \sum_{i} p(\pi_i|\mathbf{x}). \tag{5}$$

A forward-backward dynamic programming algorithm is proposed to calculate equation (5) efficiently and makes the network training possible. Finally, the CTC network is trained to minimize the negative log-likelihood of the target sequence $\mathbf{l}$ as:

$$\mathcal{L}_{CTC} = -\ln p(\mathbf{l}|\mathbf{x}). \tag{6}$$

According to equation 4, the CTC approach assumes conditional independence of separate labels (i.e., the character symbols predicted by $p(\pi_t|\mathbf{x})$), and thus obstructing further performance improvement.

2) Cascaded Attention-CTC: To capture information explicitly from longer context, LCANet feeds the encoded spatiotemperal features into the cascaded attention-CTC decoder (Figure 4). As illustrated here, the decoder uses an attention mechanism to find an alignment between each element of the decoded sequence ($\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3$) and the hidden states from the encoder ($\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3$). The attention score is

calculated as:

$$e_{j,t} = \mathbf{V}_a \tanh(\mathbf{W}_a \mathbf{s}_{t-1} + \mathbf{U}_a \mathbf{h}_j), \quad (7)$$

$$\alpha_{j,t} = \frac{\exp(e_j)}{\sum_{k=1}^{T} \exp(e_k)}. \quad (8)$$

The context vector is computed as the weighted average over all the source hidden states.

$$c_t = \sum_{k=1}^{T} \alpha_{k,t} h_k \quad (9)$$

The prediction result for the frame at time step t is computed based on the context, hidden state, and the previous prediction:

$$y_t = \sigma(\mathbf{W}_o \mathbf{E} \mathbf{y}_{t-1} + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{C}_o \mathbf{c}_t). \quad (10)$$

where $\mathbf{E}$ is the character-level embedding matrix which is much simpler than the word embedding matrix used and learned in neural machine translation tasks. The output prediction is fed into the CTC loss function to guide the model training process.
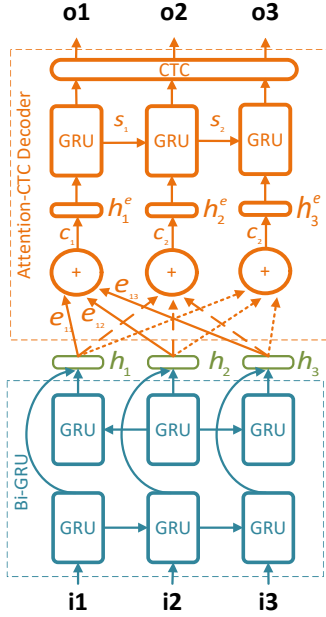

Figure 4: Diagram of the proposed cascaded attention-CTC decoder. i and o are the input and output sequence, respectively. h_i is the intermediate hidden state of the Bi-GRU. $e_{i,j}$ is the weight of the attention unit, where i, j denotes the index of the output and input node, respectively. Each attention unit generate a fixed-lenght context vector based on the input sequence.

Attention mechanism debilitates the constraint of the conditional independence assumption in CTC loss. Therefore it improves the modeling capability on the lipreading problem and can give better predictions on visually similar visemes. On the other hand, compared to the attention only architecture, the combination of attention mechanism and CTC substantially reduces the irregular alignments during the training stage, as CTC will guide the attention decoder to eliminate the unnecessary non-sequential predictions between the hypothesis and reference. Therefore, the attention module in LCANet requires fewer glimpses to find the correct focus-then-decode pattern. From the experimental result, the cascaded attention-CTC decoder outperforms existing methods.

IV. Experiments

In this section, we evaluate the performance of the proposed LCANet and compare it with other state-of-the-art lipreading methods on public benchmark datasets.

A. Datasets

GRID The GRID dataset contains 28 hours of videos which are recorded from 34 different speakers each speaks 1000 sentences in a controlled lab environment. Each sentence has a fixed 6-word structure: *command + color + preposition + letter + digit + adverb*, e.g., *"set blue with h seven again"*. There are 51 different words including four commands, four colors, four prepositions, 25 letters, ten digits and four adverbs, yielding 64,000 possible sentences. For each sentence, the six words are randomly combined. Until the paper submission, GRID is still the largest available public sentence level lip-reading dataset. We follow the evaluation protocol of [3], [5] to randomly divide the dataset into train, validation and test sets, where 255 sentences of each speaker are used for testing. All videos are recorded with a frame rate of 25fps and each lasts for 3 seconds (i.e., 75 frames per sample). Based on 68 facial landmarks extracted by Dlib [58], we crop 100x50 affine-transformed mouth-centered region as the input data. We then apply an affine transformation to extract an aligned mouth region and each has a resolution of 100×50. The obtained RGB mouth images are mean-reduced and normalized. We perform data augmentation on the training data similar to that used in [3] to avoid overfitting.

B. Evaluation Metrics

We use character error rate (CER), word error rate (WER), and BLEU [59] to benchmark the proposed framework. CER is computed with the minimum number of operations required to transform the predicted text to the ground truth (e.g., the Levenshtein distance), divided by the number of characters in the ground truth. WER is similar to CER except it operates on a word level. Smaller CER/WER means higher prediction accuracy, while a larger BLEU score is preferred.

Table I: The detailed architecture of the proposed method.

Layer name	Output size	Kernel	Stride	Pad
3d-conv1	75x50x25x32	3, 5, 5	1, 2, 2	1, 2, 2
bn/relu/drop	75x50x25x32			
pool1	75x25x12x32	1, 2, 2	1, 2, 2	
3d-conv2	75x25x12x64	3, 5, 5	1, 1, 1	1, 2, 2
bn/relu/drop	75x25x12x64			
pool2	75x12x6x64	1, 2, 2	1, 2, 2	
3d-conv3	75x12x6x96	3, 5, 5	1, 2, 2	1, 2, 2
bn/relu/drop	75x12x6x96			
pool3	75x6x3x96	1, 2, 2	1, 2, 2	
highway1	75x1728			
highway2	75x1728			
gru1	75x512			
gru2	75x512			
atten1	75x28			
CTC Loss	75			

C. Implementation Details

We implement and train LCANet using Tensorflow [60]. The detailed parameters of each layer in the architecture are summarized in Table I. ADAM optimizer [61] is used with an initial learning rate of 0.0001 and the batch size is 64. We train the propose model 50 epochs on GRID with early stopping.

The proposed approach is compared with [3], [5], [62], which is referred as 'Lipnet', 'LRSW' and 'LRW', respectively. LipNet is the first end-to-end sentence-level lipreading model. LRW and LRSW achieve state-of-the-art performance on the "Lip Reading Sentences" dataset and "Lip Reading in the wild" dataset, respectively. In addition, we compare LCANet with several of its variances to explore the impact of the highway network and the cascaded attention-CTC decoder. The structures of these models are described in Table II and the simplified diagrams are shown in Figure 5. The model architectures in subfigures 5a, 5c and 5d all use a single CTC loss but varies on whether the highway network layers or the attention mechanism is used. We also compare a multi-task learning architecture [43] (subfigure 5b) which optimizes a joint loss in two branches: CTC processes one branch, and the joint attention/cross-entropy loss handle another. The objective loss function is defined as:

$$\mathcal{L}_{total} = \lambda * \mathcal{L}_{ctc} + (1 - \lambda) * \mathcal{L}_{atten}. \quad (11)$$

The final prediction result is generated with a joint decoder as described in [43]. In the experiment, we tried different λ values (from 0.1 to 0.9) and only report the best result here.

D. Results

The performance comparison between the proposed method and the baselines on the GRID dataset is summarized in Table III. Our architecture performs character-level prediction on input frames. A word is regarded as correctly predicted when "every" character is correct. LCANet

Table II: Structure description of the proposed method and its variants. AH-CTC, H-CTC, and A-CTC use only the CTC loss function. While the AH-CTC-CE model leverages the cross-entropy loss on the attention branch and the CTC loss on another branch.

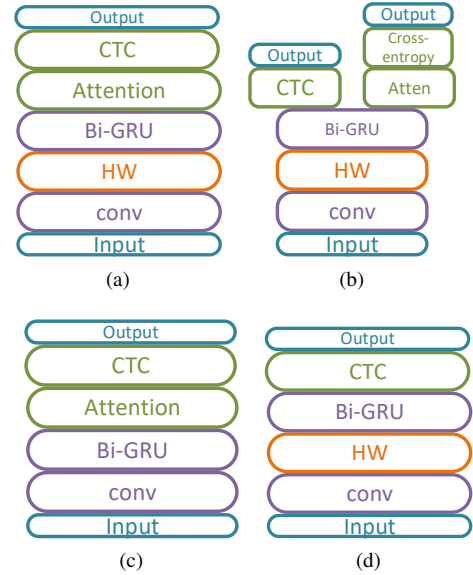| Method	Structure			
	Attention	Highway network	CTC loss	Cross-entropy loss
AH-CTC-CE	✓	✓	✓	✓
H-CTC		✓	✓	
A-CTC	✓		✓	
AH-CTC(LCANet)	✓	✓	✓	

Figure 5: Diagram of the variant models for comparison. (a) AH-CTC (LCANet), (b) AH-CTC-CE, (c) A-CTC, (d) H-CTC.

achieves 1.3% CER and 2.9% WER. Compared to Lipnet, LCANet has 31.6% and 37.5% improvement on CER and WER, receptively. The LRSW model is pre-trained on the large-scale "Lip Reading Sentences" dataset proposed in [5]. However, our method still achieves better CER and WER.

The effectiveness of the highway network layers and the cascaded attention-CTC decoder is demonstrated by the comparison among models $\{3, 4, 5, 6\}$ in Table III. First, we show that the LCANet (model $\{6\}$) is better than the multi-task or multi-branch variant (model $\{3\}$). The reason is partially because only the shared encoder is jointly updated by both the attention and CTC module in the multi-task model. Also, since the multi-task model relies on the CTC loss and the cross-entropy loss, it is difficult to coordinate the behavior of the two modules. In contrast, the proposed cascaded architecture uses a single CTC loss to update both the encoder and decoder. The leverage of the attention mechanism guarantees output sequences generated from

input sources with long-term coherence. The CTC decoder enforces monotonic alignment to guide the attention module to concentrate on the salient sequential parts of the input video frames. Second, the benefit of the highway network is illustrated by comparing LCANet with model {5}, which yields 0.4% improvement on CER and 1.1% improvement on WER. The function of the highway network is similar to the ResNet [63]. The difference is the highway network leverages adaptive gating units to regulate the information flow, which allows gradients to flow through many layers without attenuation. Finally, removing the attention mechanism (model {4}) also greatly compromises the overall performance. In addition, model {5} can be regarded as the attention-only model. The comparison between model {5} and model {6} illustrates the proposed cascaded attention-CTC model is better than its attention-only variant.

Table III: Performance on the GRID dataset.

#	Method	CER %	WER %	BLEU %
1	Lipnet [3]	1.9	4.8	96.0
2	LRSW [5]		3.0	
3	AH-CTC-CE	2.1	4.9	95.8
4	H-CTC	1.8	4.3	96.4
5	A-CTC	1.7	4.1	96.7
6	**AH-CTC(LCANet)**	**1.3**	**2.9**	**97.4**

E. Convergence speed

The training and validation loss on GRID dataset for the proposed approach and Lipnet are shown in Figure 6. In the figure, we mark the required number of epochs for the loss to reduce to 4. LCANet reaches the loss value 4 in 14 epochs for training and 20 epochs for validation. On the contrary, the converge is much slower for Lipnet, which reaches the same loss in 38 epochs for training and 40 epochs for validation. The speed-up comes from the cascaded attention-CTC module. The attention mechanism generates output sequences with long-term temporal correlation, which overcomes the disadvantage of CTC that assumes inputs are conditionally independent. Besides, the attention mechanism can be treated as a pre-alignment operation that eliminates unnecessary paths for the CTC decoder. Meanwhile, CTC guarantees a monotonic alignment between video frames and text labels, which helps to guide the attention mechanism to focus on the sequential order video-text pairs. Therefore, less irrelevant samples will be generated that leads to the observed speed-up.

F. Confusion Matrix

We use the mapping proposed by the IBM ViaVoice database [64]. 43 phonemes are grouped into 13 visemes classes (including a silence class), denoted by: lip-rounding based vowels (V), Alveolar-semivowels (A), Alveolar-fricatives (B), Alveolar (C), Palato-alveolar (D), Bilabial (E),
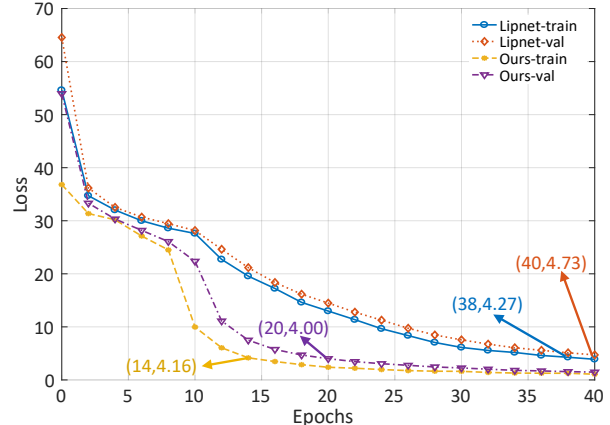

Figure 6: Training curve. The value pair (#iteration, loss) in the figure denotes the required iterations to achieve the specified loss.

Dental (F), Labio-dental (G), Velar (H). We plot the confusion matrix for the viseme class of lip-rounding based vowels, Bilabial in Figure 7 since they capture the most confusing phonemes. For example, in Figure 7a, $\{/AA/, /AY/\}$, $\{/AE/, /IH/\}$ and $\{/AO/, /UW/\}$ are frequently incorrectly classified during the text decoding process. This phenomenon is caused by the similar pronunciation between the text pair of $\{r, i\}$, $\{at, bin\}$, $\{four, two\}$. The intra-visemes categorical confusion matrix is shown in Figure 7c. The misclassification between different viseme categories is not statistically significant. There is a tiny uncertainty between viseme class V and D, which demonstrates the proposed method is effective on the lipreading task.

V. CONCLUSIONS

In this paper, we have presented LCANet, an end-to-end deep neural network architecture for machine lipreading. To accurately predict visemes, LCANet introduced a cascaded attention-CTC decoder which effectively compensates the defect of the conditional independence assumption of the CTC approach. In addition, LCANet stacks highway network layers over 3D-CNN layers to further improve performance. Extensive experiment results show that LCANet achieves the state-of-the-art accuracy as well as faster convergence.

Our future work lies in two directions. First, LCANet currently deals with the conditional independence assumption of CTC using attention mechanisms, but the loss function itself is not modified. Therefore, we will investigate solutions that can be directly applied on the loss function. Second, we would like to extend LCANet with joint audio and visual input as [5] to explore further performance improvement.

REFERENCES

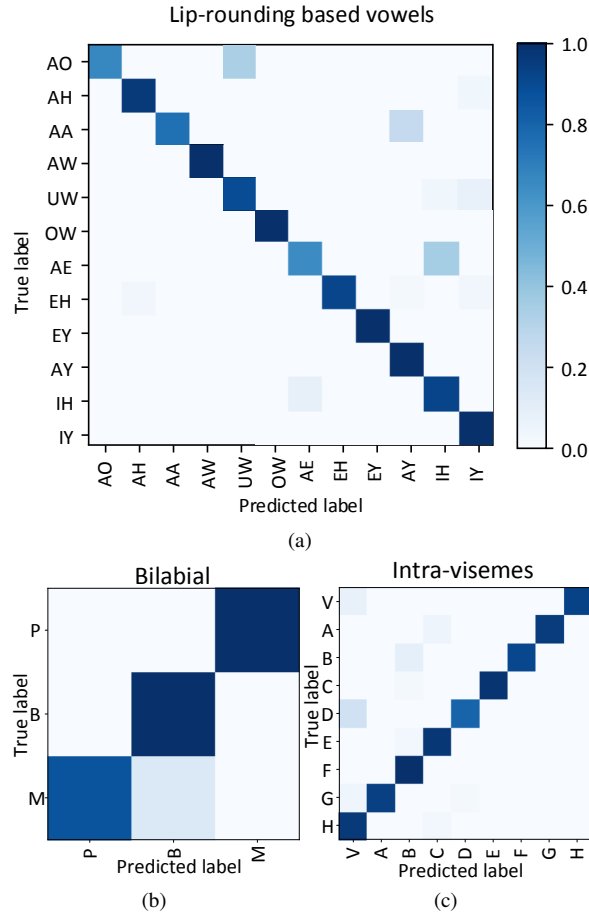[1] S. Hilder *et al.*, "Comparison of human and machine-based lip-reading," in *INTERSPEECH*, 2009.

Figure 7: (a) Lip-rounding based vowels, (b) Bilabial , (c) Intra-visemes.

[2] S. Petridis *et al.*, "End-to-end visual speech recognition with lstms," in *ICASSP*, 2017.

[3] Y. M. Assael *et al.*, "Lipnet: Sentence-level lipreading," *CoRR*, vol. abs/1611.01599, 2016.

[4] M. Wand *et al.*, "Lipreading with long short-term memory," in *ICASSP*, 2016.

[5] J. S. Chung *et al.*, "Lip reading sentences in the wild," in *CVPR*, 2017.

[6] K. Noda *et al.*, "Lipreading using convolutional neural network," in *INTERSPEECH*, 2014.

[7] Y. Geng, G. Zhang, W. Li, Y. Gu, R.-Z. Liang, G. Liang, J. Wang, Y. Wu, N. Patil, and J.-Y. Wang, "A novel image tag completion method based on convolutional neural transformation," in *International Conference on Artificial Neural Networks*. Springer, 2017, pp. 539–546.

[8] G. Zhang, G. Liang, W. Li, J. Fang, J. Wang, Y. Geng, and J.-Y. Wang, "Learning convolutional ranking-score function by query preference regularization," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2017, pp. 1–8.

[9] S. Miao, Z. J. Wang, and R. Liao, "A cnn regression approach for real-time 2d/3d registration," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1352–1363, 2016.

[10] D. Tome, C. Russell, and L. Agapito, "Lifting from the deep: Convolutional 3d pose estimation from a single image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] X. Peng, X. Yu, K. Sohn, D. Metaxas, and M. Chandraker, "Reconstruction for feature disentanglement in pose-invariant face recognition," *arXiv preprint arXiv:1702.03041*, 2017.

[12] J. Wu, E. E. Abdel-Fatah, and M. R. Mahfouz, "Fully automatic initialization of two-dimensional–three-dimensional medical image registration using hybrid classifier," *Journal of Medical Imaging*, vol. 2, no. 2, p. 024007, 2015.

[13] J. Wu and M. R. Mahfouz, "Robust x-ray image segmentation by spectral clustering and active shape model," *Journal of Medical Imaging*, vol. 3, no. 3, p. 034005, 2016.

[14] M. Ding and G. Fan, "Articulated and generalized gaussian kernel correlation for human pose estimation," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 776–789, 2016.

[15] J. Liang, C. Chen, Y. Yi, X. Xu, and M. Ding, "Bilateral two-dimensional neighborhood preserving discriminant embedding for face recognition," *IEEE Access*, vol. 5, 2017.

[16] X. Wang, L. Pollock, and K. Vijay-Shanker, "Automatic segmentation of method code into meaningful blocks: Design and evaluation," *Journal of Software: Evolution and Process*, 2013.

[17] ——, "Automatically generating natural language descriptions for object-related statement sequences," in *Proceedings of the 24th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Feb 2017, pp. 205–216.

[18] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *ICML*, 2016.

[19] Y. Miao *et al.*, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding." in *In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.

[20] Y. Zhang *et al.*, "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks," *arXiv:1701.02720*, 2017.

[21] D. Bahdanau *et al.*, "Neural Machine Translation by Jointly Learning to Align and Translate," *arXiv:1409.0473*, 2014.

[22] R. Sennrich *et al.*, "Neural machine translation of rare words with subword units," *arXiv:1508.07909*, 2015.

[23] J. Gehring *et al.*, "Convolutional sequence to sequence learning," *aarXiv:1705.03122*, 2017.

[24] A. Vaswani *et al.*, "Attention is all you need," *arXiv:1706.03762*, 2017.

[25] A. Graves *et al.*, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006.

[26] C. Chandrasekaran *et al.*, "The natural statistics of audiovisual speech," *PLoS computational biology*, vol. 5, no. 7, p. e1000436, 2009.

[27] Z. Chen *et al.*, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *arXiv:1707.07048*, 2017.

[28] H. Erdogan *et al.*, "Multi-channel speech recognition: Lstms all the way through," in *CHiME-4 workshop*, 2016.

[29] G. E. Dahl *et al.*, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 1, pp. 30–42, 2012.

[30] A. Graves *et al.*, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.

[31] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[32] D. Bahdanau *et al.*, "End-to-end attention-based large vocabulary speech recognition," in *ICASSP*, 2016.

[33] W. Chan *et al.*, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *ICASSP*, 2016.

[34] Y. Kim *et al.*, "Character-aware neural language models," in *AAAI*, 2016.

[35] A. Kumar *et al.*, "Ask me anything: Dynamic memory networks for natural language processing," in *ICML*, 2016.

[36] A. Conneau *et al.*, "Very deep convolutional networks for natural language processing," *arXiv:1606.01781*, 2016.

[37] J. Weston *et al.*, "Towards ai-complete question answering: A set of prerequisite toy tasks," *arXiv:1502.05698*, 2015.

[38] D. Li, X. Wang, and D. Kong, "Deeprebirth: Accelerating deep neural network execution on mobile devices," *CoRR*, vol. abs/1708.04728, 2017. [Online]. Available: http://arxiv.org/abs/1708.04728

[39] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding," *CoRR*, vol. abs/1510.00149, 2015. [Online]. Available: http://arxiv.org/abs/1510.00149

[40] F. N. Iandola, M. W. Moskewicz, K. Ashraf, S. Han, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size," *CoRR*, vol. abs/1602.07360, 2016. [Online]. Available: http://arxiv.org/abs/1602.07360

[41] S. Han, J. Kang, H. Mao, Y. Hu, X. Li, Y. Li, D. Xie, H. Luo, S. Yao, Y. Wang, H. Yang, and W. B. J. Dally, "Ese: Efficient speech recognition engine with sparse lstm on fpga," in *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, ser. FPGA '17. New York, NY, USA: ACM, 2017, pp. 75–84. [Online]. Available: http://doi.acm.org/10.1145/3020078.3021745

[42] Y. Mao, J. Oak, A. Pompili, D. Beer, T. Han, and P. Hu, "Draps: Dynamic and resource-aware placement scheme for docker containers in a heterogeneous cluster," in *2017 IEEE 36th International Performance Computing and Communications Conference (IPCCC)*, Dec 2017, pp. 1–8.

[43] T. Hori *et al.*, "Joint ctc/attention decoding for end-to-end speech recognition," in *Annual Meeting of the Association for Computational Linguistics*, 2017.

[44] D. Wu and Q. Ruan, "Lip reading based on cascade feature extraction and hmm," in *International Conference on Signal Processing (ICSP)*, 2014.

[45] S. S. Morade and S. Patnaik, "Lip reading using dwt and lsda," in *IEEE International Advance Computing Conference (IACC)*, 2014.

[46] X. Liu and Y. m. Cheung, "Learning multi-boosted hmms for lip-password based speaker verification," in *IEEE Transactions on Information Forensics and Security*, 2014.

[47] G. Zhao *et al.*, "Local spatiotemporal descriptors for visual recognition of spoken phrases," in *Proceedings of the International Workshop on Human-centered Multimedia*, 2007.

[48] A. A. Shaikh *et al.*, "Lip reading using optical flow and support vector machines," in *International Congress on Image and Signal Processing*, 2010.

[49] J. Luettin *et al.*, "Visual speech recognition using active shape models and hidden markov models," in *ICASSP*, 1996.

[50] K. Palecek, "Comparison of depth-based features for lipreading," in *International Conference on Telecommunications and Signal Processing (TSP)*, 2015.

[51] T. F. Cootes *et al.*, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[52] G. Potamianos *et al.*, "An image transform approach for hmm based automatic lipreading," in *ICIP*, 1998.

[53] R. Sanabria *et al.*, "Robust end-to-end deep audiovisual speech recognition," *arXiv:1611.06986*, 2016.

[54] D. Tran *et al.*, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015.

[55] R. K. Srivastava *et al.*, "Training very deep networks," in *NIPS*, 2015.

[56] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *Interspeech*, 2017.

[57] J. Chung *et al.*, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv:1412.3555*, 2014.

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

[59] K. Papineni *et al.*, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*, 2002.

[60] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016.

[61] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." in *ICLR*, 2014.

[62] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *ACCV*, 2016.

[63] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016.

[64] P. Neti *et al.*, "Audio-visual speech recognition," in *Technical report. Center for Language and Speech Processing, Johns Hopkins University, Baltimore*, 2000.