Understanding Pictograph with Facial Features: End-to-End Sentence-Level Lip Reading of Chinese

Xiaobing Zhang,[1,2] **Haigang Gong,**[1] **Xili Dai,**[1] **Fan Yang,**[1] **Nianbo Liu,**[1] **Ming Liu**[1]
[1]University of Electronic Science and Technology of China, UESTC
[2]CETC Big Data Research Institute Co.,Ltd, Guizhou, China
{zhangxiaobing, daixili, yangfan}@std.uestc.edu.cn {hggong, liunb, csmliu}@uestc.edu.cn

Abstract

With the breakthrough of deep learning, lip reading technologies are under extraordinarily rapid progress. It is well-known that Chinese is the most widely spoken language in the world. Unlike alphabetic languages, it involves more than 1,000 pronunciations as Pinyin, and nearly 90,000 pictographic characters as Hanzi, which makes lip reading of Chinese very challenging. In this paper, we implement visual-only Chinese lip reading of unconstrained sentences in a two-step end-to-end architecture (LipCH-Net), in which two deep neural network models are employed to perform the recognition of Picture-to-Pinyin (mouth motion pictures to pronunciations) and the recognition of Pinyin-to-Hanzi (pronunciations to texts) respectively, before having a jointly optimization to improve the overall performance. In addition, two modules in the Pinyin-to-Hanzi model are pre-trained separately with large auxiliary data in advance of sequence-to-sequence training to make the best of long sequence matches for avoiding ambiguity. We collect 6-month daily news broadcasts from China Central Television (CCTV) website, and semi-automatically label them into a 20.95 GB dataset with 20,495 natural Chinese sentences. When trained on the CCTV dataset, the LipCH-Net model outperforms the performance of all state-of-the-art lip reading frameworks. According to the results, our scheme not only accelerates training and reduces overfitting, but also overcomes syntactic ambiguity of Chinese which provides a baseline for future relevant work.

Introduction

Lip reading is the task of decoding what is being said from the movement of a speaker's mouth, which plays an important role in speech comprehension and human communication. It is a proverbially difficult skill, and very challenging for humans, especially applied to the Chinese language. Chinese lip reading actuations, besides the face, lip, tongue and teeth, are actually latent and hard to disambiguate. For example, it is intrinsically ambiguous at the Pinyin letter level - different letters share exactly the same lip shape (e.g. 'm' and 'f', 'p' and 'b'). However, these ambiguities can be largely solved when using a language model to capture the context and inherent relationship in a sentence.

Automatic lip reading has enormous practical applications: speech recognition in noisy environment, silent dictation in public places, improved hearing, silent-movie processing, and so on. Now, such automation is promising due to the progress across computer vision tasks: the application of deep neural network models and the use of large datasets for training. Recently, proposed deep learning architectures surpass professional lip readers by a large margin, at least for the constrained vocabulary defined by the database (Chung et al. 2017) (Assael et al. 2016).

Existing lip reading approaches can be classified into two categories according to their modeling units: (1) words-based (Wand, Koutnik, and Schmidhuber 2016) and (2) phonemes or visemes-based (Assael et al. 2016). A phoneme is the smallest distinguishable unit of sound which continuously form a spoken word and a viseme is its corresponding visual equivalent. The former approach is considered more related to tasks of single word detection, recognition and classification, while the latter to large vocabulary continuous speech recognition and sentence-level classification. However, the majority of related works focus on English or other languages, as opposed recognising to Chinese sentence-level sequences.

In this paper, we present LipCH-Net, which is to the best of our knowledge, the first unconstrained sentence-level Chinese lip reading model. Linguistically, Chinese is not English. In English, all words are made up of 26 letters, and the exact word can be roughly determined according to its pronunciation. On the contrary, the elements of Chinese pronunciation are composed of 23 initials, 24 vowels and 4 kinds of intonation marks. On average, each syllable corresponds to almost 3-90 different Hanzi. Moreover, there exist homophones and polyphones for most Hanzi. According to a survey (Tian 2012), Chinese is the language with the largest information entropy, which means each basic unit in Chinese carries much more information than any other languages. Thus, extracting discriminative features from this highly ambiguous language is a critical but challenging task for Chinese lip reading.

Based on the distinct characteristics of Chinese, we implement Chinese lip reading in a two-step architecture, in which two different neural models are employed to solve the recognition of Picture-to-Pinyin and Pinyin-to-Hanzi respectively before being jointly optimized to improve the overall performance finally. During experiments, we tried many tricks (such as scheduled sampling, attention mechanism, and so

on) to enhance the robustness and overall performance of LipCH-Net. The empirical results show that our strategies can accelerate training and greatly overcome intrinsic ambiguity of Chinese.

We totally spent 15 days to collect and label 6-month news with 20,495 Chinese sentences from CCTV website (the most famous daily news TV program in China). During experiments, LipCH-Net can achieve 50.2% and 58.7% accuracies in sentence-level and Pinyin-level task respectively, which surpass the results of all state-of-the-art lip reading frameworks trained on CCTV dataset.

Our main contributions are as: (1) We are the first to solve Chinese lip reading which is harder than that of English or other languages. Our experimental results provide a baseline for future related Chinese lip reading work. (2) The proposed LipCH-Net model takes lip regions alone as input and transcribes them to Hanzi sequence, without audio auxiliary information. (3) We collect a Chinese dataset for lip reading including over 20,000 natural sentences from CCTV website, which has already been a useful resource for this community. In addition, we are constantly enlarging and completing it. The data will be released as a resource for evaluation.

Related work

Audio-Visual Speech Recognition (AVSR): The fields of lip reading and AVSR are closely associated. (Mroueh, Marcheret, and Goel 2015) applies a deep neural network to classify phoneme using an audio-visual dataset. (Noda et al. 2015) applies a pre-train CNN to extract visual features and mHMMS to do fusion and classification. As with lip reading, a number of efforts have been made to develop AVSR to word recognition application.

Lip Reading: Before the advent of deep learning, most lip reading works are based on hand-extracted features, which are usually modeled by HMM-based pipeline (Potamianos et al. 2003). Optical flow, SVM classifiers and active appearance models have also been proposed. The traditional lip reading literatures are too vast to adequately cover, so we refer readers to (Zhou et al. 2014) for an analytic review.

More recent deep learning methods either for engineering 'deep' features (Thangthai et al. 2015) or for making end-to-end frameworks. In (Huang and Kingsbury 2013), Deep Belief Networks are applied for audio-visual recognition, which generate 21% relative improvement over a baseline audio-visual GMM/HMM system. In (Petridis and Pantic 2017), deep bottleneck features (DBF), which are extracted from deep autoencoder, are concatenated with discrete cosine transform (DCT) features, and the overall architecture is trained jointly with a Long Short-Term Memory (LSTM) classifier. (Wand, Koutnik, and Schmidhuber 2016) proposes a fully LSTM framework with HOG input features to recognise short phrases, which achieves superior results compared to traditional methods on the GRID dataset. (Chung and Zisserman 2016) trains a two-stream ConvNet architecture to learn mouth features, which are taken as inputs to LSTM. Their work is to make an audio-visual max-margin matching for word classification, which is far from a lip reading task. (Assael et al. 2016) applies a spatiotemporal CNN with Bi-LSTM and Connectionist Temporal Classification (CTC) (Graves and Gomez 2006). Their prediction is speaker-independent performance and limited on the constrained grammar of 51 word vocabulary, which is not entirely natural sentence-level work. (Chung et al. 2017) proposes WLAS network, which uses audio signal to augment English lip reading results. Their best character and word recognition results are 60.5% and 49.8% when only using lip pictures for input.

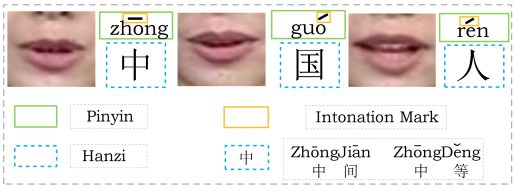

Figure 1: Examples of Chinese phrase, Pinyin and Hanzi. Here, we use different colored boxes to represent them: Yellow box: intonation mark; Green box: Pinyin sequence; Blue dotted box: single Hanzi.

Characteristics of Chinese

Overview. In this section, we give our deep consideration of particular characteristics of the Chinese language before modeling Chinese lip reading.

The Chinese character, usually called Hanzi, is a visual symbol of the unity of "shape, sound and meaning". A single Hanzi possesses abundant information and signifies a certain meaning only when grouped with others to form a phrase. For example, the phrase in Fig.1, composed of three separate Hanzi in blue dotted box, indicates the specific meaning of "Chinese people" in English. The same Hanzi can combine with different others to express different meanings. "zhōng" can join with "Jiān" or "Děng" together to form new phrases for expressing "middle" and "medium" respectively.

Pinyin: Chinese Pinyin (phonetic symbol of Chinese character) has been used as a tool to assist the pronunciation of Hanzi since 1955. It is similar to the phonetic symbol in English, but more different from them. Pinyin is consisted of three basic components: initial, vowel and intonation mark. Removing irregular pronunciations and adding 4 kinds of intonation marks, the total number of Pinyin is nearly 1,000. Nevertheless, there are more than 90,000 Hanzi in Chinese, and 3,000 of them are most used.

Syllable: Syllable is the basic unit of language that can be distinguished through hearing. In Chinese, each individual Hanzi is an independent syllable symbol. As shown in Fig.1, three different Hanzi in blue dashed box represent the syllables of "zhōng", "guó", "rén" respectively. Whereas, one identical syllable (i.e. pronunciation) corresponds to at least 3 but at most 120 Hanzi. In addition, homophones account for more than 85% among all Hanzi (91,251).

With so many pronunciations and Pinyin possibilities, as well as polyphones and homophones, Chinese is an ambiguous language with its unique characteristics. To overcome

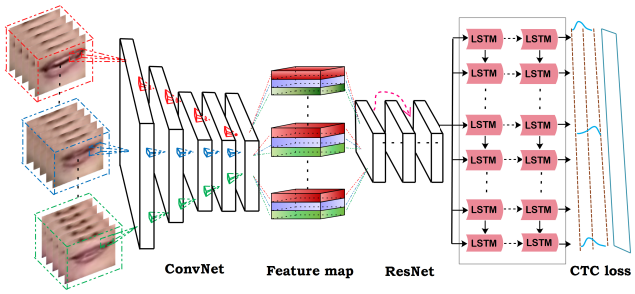

Figure 2: The Picture-to-Pinyin model. The input is continuous lip-pictures which will be transformed into gray-scale images at first. There are 5 convolutional layers (96, 256, 512, 512, 512) with 3x3 kernels and batch normalization are utilized in each ConvNet layer.

the ambiguity in translating sentence, we split Pinyin-to-Hanzi model into two modules and use large additional data to pre-train them with tricks separately. Instead of improving the recognition of single Hanzi, this strategy can make the best of long sequence matches with natural broken sentences and increase the robustness of the model.

LipCH-Net

Picture-to-Pinyin model

Different from other image recognition tasks, lip reading should be able to capture the slight changes in mouth pictures. Fig. 2 displays the configuration of our Picture-to-Pinyin model. It is a combination of convolutional, residual and LSTMs networks with CTC loss.

- ConvNet: First set of the model utilizes convolutional layers to the preprocessed lip frames. Convolutional layers can capture short-dynamics of the lip region. It is based on the VGG-M model (Chatfield et al. 2014), which has a fairly good classification performance on ImageNet and is fast to train during experiments.
- ResNet: The building blocks of residual network (He et al. 2016) are composed of two convolutional layers with BN and ReLU. Its skip connections facilitate information propagation. The output of ResNet (modified to 14-layer) is a single dimensional tensor each timestep. We did not employ pre-trained models, because they are optimized for specific diverse tasks.
- LSTM: We stack 2-layer LSTMs to absorb the features generated from ResNet. For its optimization criterion, three methods are tried during experiments. The first method is to add a softmax layer after the LSTMs output, when the overall sequence is encoded by the LSTMs. It is feasible to propagate the errors back to the first timestep of the sequence during backpropagation. The second method is to make the criterion for each timestep, which is similar to the application of LSTMs in speech recognition. Instead of viseme labels, the Pinyin label is repeated at every 3 timestep. The last approach is to apply CTC loss. After experimentation with three approaches, our conclusion is that the last method leads to much higher Pinyin accuracy.
- CTC: According to (Maas et al. 2015), CTC loss can eliminate the need for aligning input training data to target outputs. It is utilized in Picture-to-Pinyin to automatically segment pictures and Pinyin sequences, which has a better performance even than traditional hidden markov model. Given a discrete distribution over Pinyin sequence which is supplemented with "blank" label, CTC defines the probability of a sequence by counting over all possible sequences that are regarded as equivalent to this sequence. This deals with variable-length sequences and removes the need for the alignment at the same time. Let L represents the Pinyin letter labels, and $\bar{L} = L \cup \smile$ represents the blank-supplemented ones where $\smile$ is the blank token. Define the mapping Γ: $\bar{L}^{\star} \rightarrow L^{\star}$, which removes blank tokens and deletes adjacent duplicate Pinyin letters when given a string over $\bar{L}$. For a sequence $y \in L^{\star}$, CTC computes $P(y|\mathrm{x}) = \sum P(v_1, \ldots, v_K|\mathrm{x})(s.t. v \in \Gamma^{-1}(y), |v| = K)$, where K is the number of timesteps in the sequence model. For example, if K = 3, CTC computes the probability of "ni" as $p(nni) + p(nii) + p(\smile ni) + p(n \smile i) + p(ni \smile)$, which is calculated efficiently through dynamic programming.

As shown in Fig.2, the lip pictures are transformed into gray-scale ones before feeding to the feature extractor. ConvNet takes every 5 lip pictures as a input, moving 2 frames at each timestep. If the number of lip pictures is n, the length of input sequence will be $\lfloor (n-4)/2 \rfloor$. Based on pixels inside lip pictures $X = x_1, x_2, x_3, \ldots, x_t$ (x_i involves 5 gray lip pictures), residual network computes 256-dimensional image features for every input timestep as follows:

$$v_i = K[\mathbf{Res}(\mathbf{CNN}_{\theta_c}(x_i))] + b \tag{1}$$

where $\mathrm{CNN}(x_i)$ transforms picture x_i into 512-dimensional vector before passing to the residual part. The matrix K has the dimension $h \times 256$, where h is the size of the embedding space of LSTMs. Then, the LSTMs absorb the v_i and transform it into a h'-dimension vector v'_i. Simultaneously, it generates a state vector s at the last t timestep. The LSTMs network can be denoted as:

$$\begin{aligned} (v'_i, h_i) &= \mathbf{LSTMs}(v_i, h_{i-1}) \\ s &= h_t \end{aligned} \tag{2}$$

We should emphasize that there is a full connection layer between LSTMs and CTC loss to transform feature v'_i into 26-dimensional vector.

Pinyin-to-Hanzi model

Picture-to-Hanzi model is based on RNN transducer with the attention mechanism (Bahdanau, Cho, and Bengio), which has given state-of-the-art results in a variety of sequence processing tasks recently.

According to (Gehring et al. 2017), it is hard to train a sequence-to-sequence model when the number of timesteps is too large. Therefore, we separate the Pinyin-to-Hanzi model into two language modules, encoder and decoder (Fig. 3), to overcome ambiguity when translating Pinyin

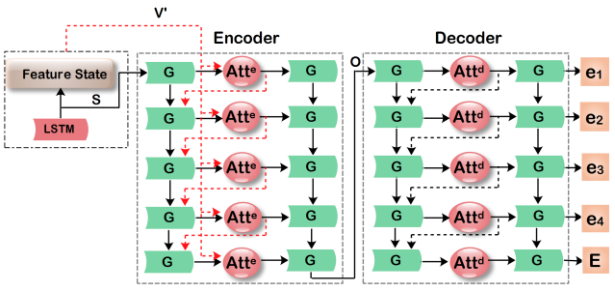

Figure 3: The Pinyin-to-Hanzi model. The red dashed line works when Picture-to-Pinyin and Pinyin-to-Hanzi model are jointly optimized to improve the overall LipCH-Net performance.

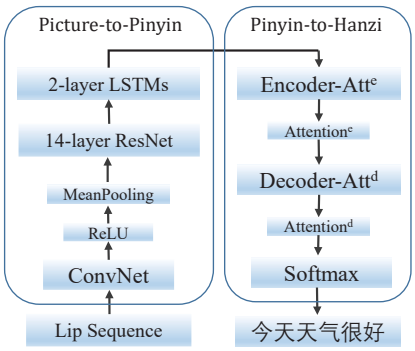

Figure 4: The block-diagram of LipCH-Net architecture.

sequence into Hanzi sentence. We apply two dependent attention mechanisms $Attention^e$ and $Attention^d$ in the encoder and decoder module. Note that $Attention^e$ only works during the overall training of LipCH-Net when given the feature vectors v' and state vector s. Before having a jointly sequence-to-sequence optimization, encoder and decoder are trained respectively with supplementary data. Experimental results show that Gated Recurrent Unit (GRU) performs better with lower perplexity value and converges faster than LSTMs in Pinyin-to-Hanzi model.

During the training process of encoder, the input and labels are both Pinyin sequence. We convert the Pinyin sequence $C = c_1, c_2, c_3, \ldots, c_l$ into embedding space before passing them through weighted connections to compute the output vector sequence $O = o_1, o_2, o_3, \ldots, o_l$. Each output vector o_i is used to parameterize a predictive distribution $\Pr(c_{i+1}|o_i)$ over the possible next input c_{i+1}. The encoder modeling objective is to maximize the total log probability of the training sequence $\sum_{i=0}^{l-1} \log \Pr(c_{i+1}|c_{\leq i})$.

During the training of decoder, the input and labels are both Hanzi sequence. The attention vectors are mixed with the output states to generate the Hanzi vectors r_k which compress the information necessary to generate the next timestep output. The probability distribution of the output Hanzi is computed by a full connection layer with softmax as follows:

$$\begin{aligned} t_k, g_k &= \mathbf{GRU}(g_{k-1}, e_{k-1}, r_{k-1}) \\ r_k &= t \cdot \mathbf{Attention}^d(t, g_k) \\ P(e_i|e_{<i}) &= \mathbf{softmax}(f(t_k, r_k)) \end{aligned} \tag{3}$$

At every step k, t_k is output vectors, g_k is state vectors, r_k is context vectors, and e_k is the output. In the end, we employ the parameters preserved during the separate training procedure of the encoder and decoder to initialize our sequence-to-sequence Pinyin-to-Hanzi model.

During experiments, we have found that the attention mechanism is indispensable for the sequence-to-sequence system to work. When lacking of it, the model seems to forget previous useful information and generates the output sentence which is very little relevant to the input sequence. The benefits are clearly shown in the experimental results section.

LipCH-Net Architecture

Our whole architecture, summarised in Fig. 4, combines the two key components: "Picture-to-Pinyin" and "Pinyin-to-Hanzi" to make a joint optimization. During the training time of LipCH-Net, CTC loss in the Picture-to-Pinyin part is removed. The output of each LSTMs unit (i.e. feature vector) in Picture-to-Pinyin is fed into the corresponding GRU unit of encoder in Pinyin-to-Hanzi one by one. Meanwhile, the state vector s and feature vectors v' are put into the GRU and $Attention^e$ unit directly. Here, $Attention^e$ has the similar processing procedure to $Attention^d$ as Equations 3.

Dataset

Lip reading datasets are plentiful (AVLetters(Matthews et al. 2002), AVLetters2(Cox et al. 2008), OuluVS1(Zhao, Barnard, and Pietikainen 2009), OuluVS2(Anina et al. 2015), GRID(Wand, Koutnik, and Schmidhuber 2016), BBCTV(Chung et al. 2017)). However, there is no public Chinese lip reading data. In this section, we show the multi-stage pipeline to semi-automatically collect a large-scale Chinese dataset.

We adopt a variety of CCTV new broadcast recorded between April and October in 2016. The use of CCTV programs are based on the following facts: (1) CCTV programs involve many anchors and abundant contents; (2) all speaking anchors are sitting at the center on screen during the programs and shot changes are less frequent; (3) all anchors speak standard Mandarin at an appropriate speed. Here, we give a brief of some key procedures.

- **Video Clip:** Individual statement clips in programs are detected by comparing color histograms across consecutive pictures (Ope 2016). Then, the *CORELX9* (COR 2015) is used to convert each video clip into frame pictures, following the criteria of 25 pictures per second.
- **Text Processing:** According to the talking content in each clip, we download the corresponding manuscripts from the website and save them as a separate TXT file. The content in TXT file is divided into lines based on the standard that sentence in each line has a clear meaning with no more than 25 Hanzi. Files are additionally checked by two persons to ensure their credibility and accuracy.

- **Timestamp Tag:** We use *OksrtClient* (Oks 2016), which supports the automatic alignment between video and Chinese text with more than 95% accuracy, to align video clips and TXT file. *OksrtClient* can generate the start and end timestamps of each sentence and save them simultaneously.
- **Alignment:** According to the time interval between the start and end time of each sentence, as well as the sample rate of 25fps, we automatically locate and map the corresponding lip pictures of each sentence.
- **Lip Detect:** According to the facial landmarks, extracted by face recognition API of *OpenCV* (Ope 2016), lip region is cropped with size of 120*120. Some tricks such as magnifying the pictures and constraining the regions are used to improve the accuracy up to 100%.

Using this pipeline, we have collected a labeled 20.95GB Chinese dataset, including 20,495 natural Chinese sentences along with the corresponding lip pictures. The training, validation, and test data are divided according to the proportion of 7:1:2. Each sub-data contains all different speakers which makes the model to be speaker agnostic.

Experiment and Evaluation

In this section, we evaluate our LipCH-Net and training strategies. We also compare our model to the state-of-the-art lip reading architectures on CCTV dataset.

Data Cleaning We remove special symbols (such as @, !, $*$, ?, $<$) in labels and delete the sentence with less than 5 Hanzi. Moreover, sentences are grouped into subsets according to the length of 5-10, 11-15 and 16-25 Hanzi, including 9,452, 7,619 and 2,868 sentences respectively.

Evaluation protocol LipCH-Net is assessed on the independent test data. In experiments, we compute Pinyin-level Accuracy Rate (PAR), Hanzi Accuracy Rate (HAR) and Perplexity value. PAR and HAR are defined as $1 - \mathbf{errorrate} = 1 - (\mathbf{d}+\mathbf{i}+\mathbf{r})/\mathbf{m}$, where $\mathbf{d}$, $\mathbf{i}$, $\mathbf{r}$ is the amount of $\mathbf{deletions}$, $\mathbf{insertions}$ and $\mathbf{replacements}$ to obtain from the reference to the hypothesis, respectively. m is the amount of Pinyin letters (PAR) or Hanzi (HAR) in the reference. $Perplexity$ is a measurement of probability distribution, and a low perplexity value indicates the distribution is good at predicting the output.

Experiments of Picture-to-Pinyin

Preprocessing: Before training the Picture-to-Pinyin model, each Chinese sentence label is transformed into its corresponding Pinyin sequence. For example, the Chinese sentence in Fig.4 is translated into "jin tian tian qi hen hao". Because the mechanism of CTC requires input sequences are longer than the output labels, we remove the blank spaces as "jintiantianqihenhao". In addition, the sample whose input sequence is shorter than its label is also eliminated.

Training strategy: Convergence ability of ConvNet and LSTMs is inconsistent and different. In the training time, ConvNet is provided with the Batch Normalization (BN), while LSTMs not. Thus, we apply $different\ learning\ rate$ to promote two different networks converge simultaneously. Moreover, LSTMs are trained with $gradient\ clipping$ which drop large gradient so as to greatly avoid the bad convergence problem. Experimental results indicate the initial learning rate of 0.1 in ConvNet and 0.001 in LSTMs can make the Picture-to-Pinyin model fully converge.

Table 1: Pinyin accuracies (PAR) using temporal convolution (TemConv).

	Network	PAR
Nt1	3D + ResNet + TemConv + CTC	35.1%
Nt2	AlexNet + ResNet+ TemConv + CTC	32.6%
Nt3	IncepV2 + ResNet + TemConv + CTC	37.4%
Nt4	VGG-M + ResNet + TemConv + CTC	**41.2%**
Nt5	VGG-M + DNN + TemConv + CTC	36.7%

Table 2: Pinyin accuracies (PAR) using different LSTMs/GRUs.

	Netwotk	PAR
Nt6	VGG-M + ResNet + 1-LSTMs + CTC	40.1%
Nt7	VGG-M + ResNet+ 2-LSTMs + CTC	**44.2%**
Nt8	VGG-M + ResNet + Bi-LSTMs + CTC	41.4%
Nt9	VGG-M + ResNet + 2-GRUs + CTC	34.8%

Evaluation:

To evaluate the contribution of each individual module in the model, we start by applying a simpler one than the proposed LipCH-Net. In Table.1, Nt1 utilize 3D convolution instead of 2D, which is followed by the ResNet. We replace LSTMs with two temporal convolutional layers, each of which is followed by BN, ReLU and max-pooling with a factor of 2. The results of the same configuration, but with AlexNet/IncepV2/VGG-M are also listed in Table.1 (Nt2/3/4). Then, we use 6 fully connected hidden layers with BN and ReLU (Nt5) instead of the ResNet part to evaluate its effectiveness. The DNN progressively changes the feature size as 512-256-128-64-128-256.

In order to verify the performance of LSTMs, we use them to replace temporal convolutions. In Table.2, Nt6 utilize 1-layer LSTMs, while Nt7 utilize 2-layer LSTMs. Bidirectional LSTMs have been applied in some sequence learning tasks (Chorowski et al. 2015) due to its ability to generate output based on future content as well as the past content. In experiments, we also tried bidirectional LSTMs instead of unidirectional ones (Nt8). Experimental results show that it took longer training time, while offering no distinct performance improvement. This is probably because the Pinyin-to-Hanzi model with attention mechanism is also based on the whole input sequence and provide extra local focus. The results of the same configuration, but with GRUs are represented as Nt9. While training the LSTMs, the VGG-M and ResNet remain fixed. Thus, these four networks are not trained in end-to-end fashion.

In Table.3, we evaluate different loss methods as mentioned in Picture-to-Pinyin model part. Nt10 and Nt11 have the same configuration of Nt7, but with a softmax layer and making the criterion for each timestep to calculate the loss. Finally, we apply end-to-end training method for the overall network. The last Nt12* is the same as Nt7, but trained end-

Table 3: Pinyin accuracies (PAR) using different loss or training methods.

	Network	PAR
Nt10	VGG-M + ResNet + 2-LSTMs + Softmax	29.7%
Nt11	VGG-M + ResNet + 2-LSTMs + CET	35.4%
Nt12*	VGG-M + ResNet + 2-LSTMs + CTC	**46.3%**

to-end, using the weight of Nt7 as initialization parameters.

We calculate the average confusion values of initials in the output Pinyin sequence, and visualize them as shown in Fig.5. It is quite clear that each initial basically has at least three highly similar confusing pronunciations, which can combine with different vowels to indicate various syllables. Meanwhile, it shows the ambiguity and high challenge of Chinese lip reading recognition.

Discussion and error analysis: Some summaries can be made from the results listed above. Firstly, the VGG-M makes absolute improvement over other models. In addition, CNN exhibits strong ability of feature extraction and the use of 3D instead of 2D does not work much better especially in Chinese lip reading task. By applying ResNet instead of 6-layer DNN, a further 4.5% absolute improvement is attained. Moreover, we find the LSTMs offer 3.0% better accuracy compared to temporal convolutional network, which demonstrates the expressive power of LSTMs in model temporal sequences again. By training Nt7 with an end-to-end fashion (Nt12*), we get a 2.1% improvement, which indicates the importance of end-to-end training method towards achieving higher Pinyin recognition accuracy.

Experiments of Pinyin-to-Hanzi

Training strategy: We download extra CCTV manuscripts from January 1st, 2015 to June 15th, 2017 to serve as *auxiliary data*. In the same way, we sort and group the sentences into three subsets for pre-training according to their length. The total number of Hanzi and sentence in the *auxiliary data* is 20,972,355 and 1,215,697. In order to improve model robustness, Pinyin sequences in *auxiliary data* are replaced or deleted randomly to simulate the generated ones from Picture-to-Pinyin model which are not absolutely correct. The final random error rate is chosen from 0 to 0.25.

In the training of recurrent neural network, the ground truth of previous timestep is treated as the next timestep input. However, during the inference, the ground truth is absent which leads to worse performance since the model was not trained with poor predictions at some timesteps. To eliminate this discrepancy between training and inference, we apply *Scheduled sampling* approach proposed by (Bengio et al. 2015). At training time, the sampling rate from the previous output is selected from 0 to 0.20. We find that the model would become unstable once the rate greater than 0.2.

Evaluation: Parameter initialization range is selected from -0.02 to +0.02 and the initial learning rate is 0.001. During pre-training, the translation model yields nearly 95 percent accuracy with abundant auxiliary data. We test the performance of combining different RNNs with various cell

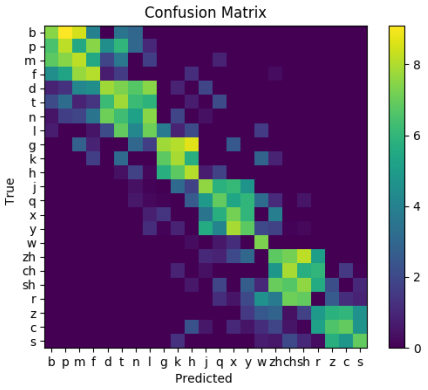

Figure 5: The confusion matrix of initials.

Table 4: Performance of LipCH-Net on CCTV news with different training strategies. LR: learning rate; CL: curriculum learning; AD: auxiliary data; SS: scheduled sampling; HFW: high frequency words; BS: beam search; PAR: Pinyin accuracy; HAR: Hanzi accuracy.

Models	Strategy	PAR	HAR	Perplexity
Picture-to-Pinyin	-	46.3%	-	3.29
	Different LR	48.7%	-	3.01
	Different LR+CL	**52.5%**	-	**2.87**
Pinyin-to-Hanzi	-	-	35.4%	7.54
	AD	-	39.8%	3.71
	AD+SS	-	40.2%	3.09
	AD+SS+Att^d	-	**45.7%**	**2.94**
LipCH-Net	-	43.7%	34.5%	12.06
	HFW	49.8%	40.8%	7.90
	HFW+BS	50.4%	42.7%	5.82
	HFW+BS+CL	52.6%	43.1%	3.32
	HFW+BS+CL+Att^e	**58.7%**	**50.2%**	**2.46**

units and layers. The best perplexity value in the experiment can decline to 2.69 when using 2-layer 1024-GRU unit. Due to the memory of GTX 1080 GPU used for training is only 8G, 2-layer 512-GRU which can produce the perplexity value with 2.94 is applied in Pinyin-to-Hanzi model finally.

Experiments of Lip Reading Model

Training strategy: We calculate high frequency words and phrases in the dataset and split out the corresponding lip pictures. The LipCH-Net is pre-trained with these high frequency words in the beginning to learn priori knowledge. Similar to (Chan et al. 2016), decoder in the Pinyin-to-Hanzi part is executed with *beam search* of width 6. In Fig.6, [C] shows there are no obvious benefits when increasing the width more than 6 in Chinese lip reading task.

Inspired by the strategy of *curriculum learning* in (Chung et al. 2017), we start training on short sentences with 5-10 Hanzi and then make the sequence length grow as the network trains. We observe that the convergence rate on the training set is a few times faster, and it also largely reduces overfitting.

Evaluation: Our implementation is based on the Tensorflow library. The network is trained using stochastic gradient descent on 4 Nvidia GTX 1080 GPU with 8GB memory

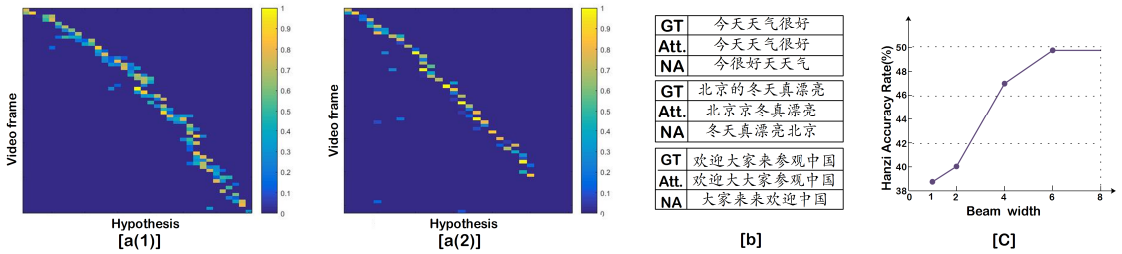

Figure 6: Alignment between the video frames and the Pinyin sequence ([a(1)])/ Hanzi sequence ([a(2)]). [b] is example of some LipCH-Net results. GT: ground truth; Att./NA: prediction with/without attention mechanism. [C] is the effect of beam width on Hanzi Accuracy Rate.

Table 5: The comparison results of LipCH-Net with other existing frameworks on CCTV news as well as their respective details.

Lipreading Model	Architecture	Data & Language	AiC(%)	AiS(%)	AiP(%)
WLAS(Chung et al. 2017)	CNN+LSTM+Attention	BBCTV & English	49.8%	**36.7%**	**51.0%**
(Assael et al. 2016)	STCNN+BiLSTM+CTC.	GRID & English	93.4%	28.9%	41.6%
(Wand, Koutnik, and Schmidhuber 2016)	NN+LSTM	GRID & English	79.6%	16.7%	30.5%
(Noda et al. 2014)	CNN+GMM+HMM.	JAVD & Japanese	37%	18.6%	35.7%
(Garg, Noyola, and Bagadia 2016)	CNN+LSTM	MIRACL-VC & English	76%	29.15%	47.3%
LipCH-Net	CNN+LSTM+GRU+CTC	CCTV& Chinese	–	**50.2%**	**58.7%**

and Intel Xeon processor E5-2620 with 32GB memory. The initial learning rate of 0.001 was applied and decreased by shrinking three times, if the training error did not increase for 10,000 iterations. Training was stopped when the validation error did not increase for 2,000 iterations. The total end-to-end model was trained for around 4 days.

All the training strategies discussed previous can make contributions to the performance. The detailed results are given in Table.4. We observe that *curriculum learning* and *attention* mechanism offer absolute improvement in PAR and HAR. In Fig.6, [a(1)] visualises the alignment of the Pinyin letter "Huan ying da jia lai can guan zhong guo" with the corresponding video frames, [a(2)] is the alignment of its equivalent Chinese sentence (i.e. the seventh one in Fig.6[b]) with the video frames. It demonstrates that attention mechanism makes straightforward alignment between the input video frames and hypothesis output. [C] shows some examples where the LipCH-Net model successfully deciphers the sentences when applied attention mechanism.

Comparison

Table. 5 shows comparative paradigms including their proposed recognition languages ("Data&Language") and model components ("Architecture"). The accuracies listed in column "AiC" are mentioned in their papers evaluated on their own dataset. The results listed in column "AiS"(sentence-level) and "AiP"(Pinyin-level) are the best accuracies they can generate on Chinese lip reading dataset during the whole retraining procedure. In the five comparative models, only WLAS is about sentence level prediction. When retrained on CCTV news, WLAS makes worse performance because there is no audio part in our data which can help to improve the model performance demonstrated in (Chung et al. 2017). The other four models all work on word classification. Thus, we employ the high frequency words and phrases in CCTV news as their training data. When retrained on CCTV news, each model retains its original structure and the labels of output units are transformed to Pinyin letter sequence to get the average Pinyin-level (AiP) accuracy. Then, the generated Pinyin sequences are translated into Chinese sentences or phrases to get the average sentence-level (AiS) accuracy via our Pinyin-to-Hanzi model. As the results show, LipCH-Net can achieve 50.2%(sentence-level) and 58.7%(Pinyin-level) accuracies which are better than Google's work (36.7% and 51.0% correspondingly). The comparison results show that the design of the model should consider language characteristics, especially intrinsic ambiguity of Chinese.

Conclusion and Discussion

In this paper, we introduce LipCH-Net, the first end-to-end model that can transcribe lip picture sequence to Chinese sentence. In the training time, two different neural network models are applied to solve the recognition of Picture-to-Pinyin and Pinyin-to-Hanzi respectively before jointly optimizing. The end-to-end model removes the need to segment video into phrases when predicting a sentence. LipCH-Net surpasses the performance of all previous lip reading networks on Chinese lip reading CCTV dataset. In future work, there are several extensions to this work which we hope to investigate: (1) we aim to apply LipCH-Net to a composite audio-visual model, in which audio information can assist with recognition accuracy and visual information can help improve model robustness in noisy environment. (2) we will employ LipCH-Net to larger datasets to demonstrate the suggestion in (Amodei et al. 2015) that recognition performance can be improved with more data; (3) we will explore the application of LipCH-Net in different dialects, such as Cantonese, rather than only in Mandarin.

Acknowledgments

This work is supported in part by National Science Foundation of China under Grant No.61572113, and the Fundamental Research Funds for the Central Universities under Grants No.ZYGX2015J155, ZYGX2016J084, ZYGX2016J195.

References

Amodei, D.; Anubhai, R.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Chen, J.; Chrzanowski, M.; Coates, A.; and Diamos, G. 2015. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR* abs/1512.02595.

Anina, I.; Zhou, Z.; Zhao, G.; and Pietikainen, M. 2015. Ouluvs2: A multi-view audiovisual database for non-rigid mouth motion analysis. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 1–5.

Assael, Y. M.; Shillingford, B.; Whiteson, S.; and De Freitas, N. 2016. Lipnet: End-to-end sentence-level lipreading. *CoRR* abs/1611.01599.

Bahdanau, D.; Cho, K.; and Bengio, Y. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *International Conference on Neural Information Processing Systems*, 1171–1179.

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

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. *CoRR* abs/1405.3531.

Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; and Bengio, Y. 2015. Attention-based models for speech recognition. *Computer Science* 10(4):429–439.

Chung, J. S., and Zisserman, A. 2016. Out of time: Automated lip sync in the wild. *ACCV 2016 Workshops. Lecture Notes in Computer Science, vol 10117*.

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

2015. Corelx9. http://www.huishenghuiying.com.cn/.

Cox, S.; Harvey, R.; Lan, Y.; Newman, J.; and john Theobald, B. 2008. The challenge of multispeaker lip-reading. In *International Conference on AuditoryVisual Speech Processing*.

Garg, A.; Noyola, J.; and Bagadia, S. 2016. Lip reading using cnn and lstm. *Technical report, Stanford University, CS231n project report*.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and N. Dauphin, Y. 2017. Convolutional sequence to sequence learning. *CoRR* abs/1705.03122.

Graves, A., and Gomez, F. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, 369–376.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

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

Maas, A.; Xie, Z.; Dan, J.; and Ng, A. 2015. Lexicon-free conversational speech recognition with neural networks. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 345–354.

Matthews, I.; Cootes, T. F.; Bangham, J. A.; Cox, S.; and Harvey, R. 2002. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 24(2):198–213.

Mroueh, Y.; Marcheret, E.; and Goel, V. 2015. Deep multimodal learning for audio-visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2130–2134.

Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H. G.; and Ogata, T. 2014. Lipreading using convolutional neural network. 1149–1153.

Noda, K.; Yamaguchi, Y.; Nakadai, K.; Okuno, H. G.; and Ogata, T. 2015. Audio-visual speech recognition using deep learning. *Applied Intelligence* 42(4):722–737.

2016. Oksrtclient. http://www.zimudashi.cn/.

2016. Opencv api reference. http://docs.opencv.org/2.4.13/modules/refman.html.

Petridis, S., and Pantic, M. 2017. Deep complementary bottleneck features for visual speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2304–2308.

Potamianos, G. N.; Gravier, C.; Garg, G.; Senior, A.; and W, A. 2003. Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE* 91(9):1306–1326.

Thangthai, K.; Harvey, R.; Cox, S.; and Theobald, B. J. 2015. Improving lip-reading performance for robust audiovisual speech recognition using dnns. In *Faavsp - the Joint Conference on Facial Analysis, Animation and Auditory-Visual Speech Processing*.

Tian, Q. 2012. Comparing information entropy between chinese and english language. In *IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems*, 748–751.

Wand, M.; Koutnik, J.; and Schmidhuber, J. 2016. Lipreading with long short-term memory. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6115–6119.

Zhao, G.; Barnard, M.; and Pietikainen, M. 2009. Lipreading with local spatiotemporal descriptors. *IEEE Transactions on Multimedia* 11(7):1254–1265.

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