LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild

Shuang Yang[*1], Yuanhang Zhang[*2], Dalu Feng[*1,2], Mingmin Yang[*4], Chenhao Wang[2], Jingyun Xiao[2], Keyu Long[2], Shiguang Shan[1,2,3], Xilin Chen[1,2]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

[2] University of Chinese Academy of Sciences, Beijing 100049, China

[3] CAS Center for Excellence in Brain Science and Intelligence Technology

[4] Huazhong University of Science and Technology

***Abstract*— Large-scale datasets have successively proven their fundamental importance in several research fields, especially for early progress in some emerging topics. In this paper, we focus on the problem of visual speech recognition, also known as lip-reading, which has received increasing interest in recent years. We present a naturally-distributed large-scale benchmark for lip-reading in the wild, named *LRW-1000*, which contains $1,000$ classes with $718,018$ samples from more than $2,000$ individual speakers. Each class corresponds to the syllables of a Mandarin word composed of one or several Chinese characters. To the best of our knowledge, it is currently the largest word-level lip-reading dataset and also the only public large-scale Mandarin lip-reading dataset. This dataset aims at covering a "natural" variability over different speech modes and imaging conditions to incorporate challenges encountered in practical applications. It has shown a large variation in this benchmark in several aspects, including the number of samples in each class, video resolution, lighting conditions, and speakers' attributes such as pose, age, gender, and make-up. Besides providing a detailed description of the dataset and its collection pipeline, we evaluate several typical popular lip-reading methods and perform a thorough analysis of the results from several aspects. The results demonstrate the consistency and challenges of our dataset, which may open up some new promising directions for future work.**

I. INTRODUCTION

Visual speech recognition, also known as lip-reading, is a task of recognizing the speech content in a video only based on visual information. It has been demonstrated that incorporating visual information in audio-based speech recognition systems can bring obvious performance improvements, especially in cases where multiple speakers are present or the acoustic signal is noisy [10], [15].

The common procedure of lip-reading involves two steps: analyzing motion information in the given image sequence and transforming this information into words or sentences. This procedure links lip-reading to two closely related fields: audio-based speech recognition and action recognition, both of which relies on a similar analyzation to an input sequence to obtain predicted results. However, currently there exists a large performance gap between lip-reading and these two closely related tasks. One main reason is that there were few large-scale lip-reading datasets in the past, which was likely a major obstacle to the progress in lip-reading.

Fortunately, with the development of deep learning technologies, some researchers have begun to collect large-scale data for lip-reading in recent years using deep learning tools. Existing public datasets can be divided into two categories: word-level dataset and sentence-level dataset. We focus on word-level lip-reading in this paper. One outstanding benchmark is the *LRW* [6] dataset proposed in 2016, which has 500 classes and displays substantial diversity in speech conditions. In addition, all the videos in this dataset are of fixed size and length, which provides much convenience to the community. The best performance on this dataset in terms of Top-1 classification accuracy has reached as high as 83% in merely two years. Some other popular word-level lip-reading datasets include *OuluVS* [22] and *OuluVS2* [1], which were released in 2009 and 2015 respectively. There are 10 classes in both datasets and the state-of-the-art model has achieved an accuracy of more than 90% on both datasets. These exciting and encouraging results mark a significant and praiseworthy improvement in lip-reading. However, lip-reading in natural or "in-the-wild" settings remains challenging due to the large variations in the practical real-world environment. Meanwhile, these appealing results also call for more challenging datasets to trigger new progresses and inspire novel ideas for lip-reading.

To this end, we collect a naturally-distributed large-scale dataset for lip-reading in the wild. The contributions in this paper are summarized as follows.

Firstly, we present a challenging $1,000$-class lip-reading dataset. Each class corresponds to the syllables of a Mandarin word which is composed of one or several Chinese characters. The labels are also provided in the format of English letters and so anyone who knows English could understand and use the data. In total, there are $718,018$ samples from more than 2000 speakers, with over 1 million Chinese character instances covering 286 Chinese syllables. To the best of our knowledge, this database is currently the largest word-level lip-reading dataset and also the only one large-scale Mandarin lip-reading dataset.

Secondly, our benchmark aims to provide naturally-distributed data to the community, highlighted by the follow-

[*] Authors make equal contributions to this work.

978-1-7281-0089-0/19/$31.00

Fig. 1. Examples of speakers in our dataset, which show a large variation of speech conditions, including lighting conditions, resolution, speaker's age, pose, gender, and make-up etc.

ing properties: (a) it contains large variations in speech conditions, including lighting conditions, resolution of videos, and speaker's attribute variations in pose, speech rate, age, gender, make-up and so on, as shown in Fig. 1; (b) some classes are allowed to contain more samples than some others, which is consistent with the actual case that some words indeed occur more frequently than others; (c) samples of the same word are not limited to a previously specified length range to allow different speech rates. These three properties make this dataset very consistent with practical settings.

Thirdly, we provide a comprehensive comparison of the current popular lip-reading methods and perform a detailed analysis of their performance in several different settings to analyze the effect of different factors on lip-reading, including the performance with respect to image scales, word's length, speaker's pose and the model capacity on naturally-distributed data. The results demonstrate the consistency and the challenges of our benchmark, which may lead to some new inspirations to the related research communities.

II. Related Work

In this section, we provide an overview of current word-level lip-reading datasets, followed by a survey of state-of-the-art methods targeting at lip-reading.

A. Word-level Lip-reading Datasets

Some well-known word-level lip-reading datasets are summarized in Table I. All these datasets have contributed greatly to the progress of lip-reading. In this part, we will give a brief review of these well-known datasets shown in the table.

AVICAR [13] and *AVLetters* [16] were proposed in 2004 and 2002 respectively and were widely used at an early period. The words in these two datasets are composed by 10 digits and 26 letters from 100 speakers and 10 speakers respectively. These two datasets has provided an important and initial impetus for early progress in automatic lip-reading.

OuluVS [22], released in 2009, consists of 10 phrases spoken by 20 subjects with 817 sequences in total. This dataset provides cropped mouth region sequences, which brings much convenience to related researchers. However, the average number of samples in each class is merely 81.7, which is not enough to cover the various conditions in practical applications.

OuluVS2 [1], released in 2015, extends the number of subjects in *OuluVS* to 53. The speakers are recorded from five fixed different views: frontal, profile, 30°, 45° and 60°. One major difference compared with *AVLetters* and *OuluVS* is that *OuluVS2* contains several different viewpoints, which makes it more difficult than the above three datasets and is therefore widely used in previous lip-reading studies. However, the viewpoints are all fixed in this dataset, and also, there are few variations beyond the view conditions.

LRW [6], an appealing large-scale lip-reading dataset released in 2016, contains 500 classes with more than a thousand speakers. The videos are no longer posed videos recorded in controlled lab environments as above, but are extracted from TV shows and thus cover a large variation of speech conditions. This remains a challenging dataset until now and has been widely used by most existing lip-reading methods. However, one pre-defining setting of this dataset is that all the words are ensured to have a roughly equal duration and each class is specified to contain roughly the same number of samples. This setting leads to a gap between the data and practical applications because word frequencies and speech rates are actually not uniform in the real world. We believe that if a model learned from data which has a natural diversity over these two points can still achieve good performance, it should also perform well when applied to practical applications.

Although there have been many English lip-reading datasets as listed above, there are very few Mandarin lip-reading datasets available up to now. With the rapid development of scientific technologies, automatic lip-reading of any language would definitely catch more and more researchers' attention over time. Therefore, we hope *LRW-1000* could fill a part of the gap for automatic lip-reading of Mandarin.

TABLE I

A SUMMARY OF EXISTING WELL-KNOWN WORD-LEVEL LIP-READING DATASETS

Datasets	# of Classes	# of Speakers	Resolution	Pose	Envir.	Color/Gray	Year	Best Acc.
AVICAR [13]	10	100	-	Controlled	In-car	Gray	2004	37.9% [11]
AVLetters [16]	26	10	Fixed 80×60	Controlled	Lab	Gray	2002	65.13% [3]
OuluVS [22]	10	20	Fixed 80×60	Controlled	Lab	Color	2009	91.4% [6]
OuluVS2 [1]	10	53	Fixed (6 different sizes)	Controlled	Lab	Color	2015	95.6% [18]
LRW [6]	500	> 1000	Fixed 256×256	Natural	TV	Color	2016	83.0% [20]
LRW-1000	1000	> 2000	Naturally distributed	Natural	TV	Color	2018	38.19%

B. Lip reading Methods

Automated lip-reading has been studied in the computer vision fields for decades. Most early methods focus on designing appropriate hand-engineering features to obtain good representations. Some well-known features include the Discrete Cosine Transform (DCT), active appearance model (AAM), motion history image (MHI), Local Binary Pattern (LBP) and optical flow , to name a few. With the rapid development of deep learning technologies, more and more work began to perform end-to-end recognition with the help of deep neural networks (DNN). According to the types of the front-end network, modern lip-reading methods can be roughly divided into the following three categories.

(1) *Fully 2D CNN based:* Two-dimensional convolution has been proved successful in extracting representative features in image-based recognition tasks. With this inspiration, some early lip-reading work [14], [17], [9] try to obtain a discriminative representation of each frame individually with some pretrained 2D CNN models, such as theVGGNet [4] and residual networks [12]. One representative work is the multi-tower structure proposed by Chung and Zisserman in [9], where each tower takes a single frame or a T-channel image as input with each channel corresponding to a single frame in grayscale. The activations from all the towers are concatenated to produce the final representation of the whole sequence. This multi-tower structure has been proved effective by appealing results on the current challenging dataset *LRW*.

(2) *Fully 3D CNN based:* One direct reason for the wide use of 3D convolutional layers in lip-reading has much to do with the success of 3D CNN in action recognition [21]. One popular method whose front-end network is completely based on 3D convolution is the LipNet model [2]. It contains three 3D convolutional layers which transform the raw input video into spatial-temporal features and feed them to the following gated recurrent units (GRUs) to generate the final transcription. The effectiveness has been proved by its remarkable performance on the public dataset which has surpassed professional human lip-readers by a large margin.

(3) *Mixture of 2D and 3D convolution:* The regular 2D spatial convolutional layers have been proved to be effective in extracting discriminative features in the spatial domain, while spatio-temporal convolutional layers are believed to be able to better capture the temporal dynamics in a sequence. For this reason, some researchers have begun to combine the advantages of the two types to generate even stronger features. In [20], Stafylakis and Tzimiropoulos proposed to combine a spatio-temporal convolutional layer with a 2D residual network to produce the final representation of the sequence. It has achieved the state-of-the-art performance on *LRW* with an accuracy of 83%.

In this paper, we evaluate each of the above state-of-the-art approaches on our proposed benchmark and present a detailed analysis of the results which may provide some inspirations for future research.

III. DATA CONSTRUCTION

In this section, we describe the pipeline for collecting and processing the *LRW-1000* benchmark, as shown in Fig. 2. We first present the choice of television programs from which the dataset was created and then provide details of the data preprocessing procedures, which interleave automatic process with manual annotation and extra filtering efforts to make the data consistent for research.

A. Program Selection and Data Collection

In our benchmark, all collected programs are either broadcast news or conversational programs with a focus on news and current events. To encourage the diversity of speakers and speech content, we select programs from both regional and national TV stations, covering a wide range of male and female TV presenters, guests, reporters and interviewees who speak Mandarin or dialectal Chinese. The final program list is composed by 26 broadcast sources with 51 programs and yields more than 500 hours of raw videos over the two-month data collection period. This large range endows the data with a nearly full coverage of commonly used words and a natural diversity in several aspects as in practical applications.

The broadcast collection described above is retrieved daily through an IPTV streaming service in China, hosted by Northeastern University. It produces 25 fps recordings in H.264 encoding, with 1.5 to 7.5Mbps video bitrates and 128 to 160Kbps audio bitrates. The video resolution is 1920×1080 for high-definition channels and 1024×576 for standard-definition channels. This makes our data cover a wide range of scales. Since the source videos were recorded through cable TV and re-encoded in real-time, they may contain temporal discontinuities which appear as frozen frames or artifacts. We clip each video up to the first occurrence of such abnormality and feed the obtained segment to subsequent procedures.

B. Shot Boundary Detection

We firstly employ a shot boundary detector by comparing the color histograms of adjacent frames. Within each detected shot, we choose three evenly spaced frames and perform face detection with a multi-view face detector in

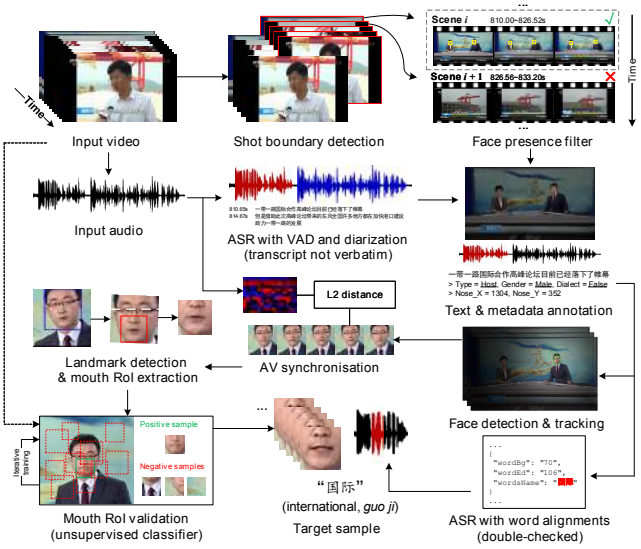


Fig. 2. Pipeline to generate samples in our dataset.

the SeetaFaceEngine2 toolkit [19]. If none of them contains a face larger than 20×20 pixels, we dismiss the shot as not containing any potential speakers. What is worth noting is that although we deliberately set a low minimum size of the candidate faces to closely mimic the in-the-wild setting, statistics still show that there are very few samples with lip resolution below 20×20, as shown in Fig. 3.

C. Annotations, Face Detection, and Face Tracking

Most Chinese TV programs have no verbatim subtitles, so we create rough transcripts of the videos with the commercial iFLYREC speech recognition service, time-aligned at the sentence level. This process automatically detects voiced segments in the audio track and diarizes it by different speakers. We then isolate sentences which are within shots retained in the previous stage, and manually annotate each video clip with the active speaker's position, gender, exact endpoints of the speech, and also the speech content. Finally, to further refine the manually-checked text annotations, a more robust ASR tool by iFLYTEK is used to produce very faithful transcripts of the utterance which are compared again with the manually checked transcripts. After several rounds of interleaved mannual and automatic check, the final annotation is believed to be accurate enough for the final use.

To associate each utterance with the corresponding speaker's face, we use the landmark detector in SeetaFaceEngine2 on the first frame and check by comparing the coordinates of each detected face with the manual annotation. Then, a kernelized correlation filter (KCF) tracker is utilized to the selected face in the given duration to obtain the whole speaking sequence. During the tracking process, we perform automatic validation of the tracking quality every 15 frames with the CNN-based face detector in SeetaFaceEngine2.

D. Audio-to-Video Synchronization

After the above process, we check for the synchronization issues and find that similar to [8] [5], the audio and video streams in a few collected videos may be out of sync, with the largest offset being less than one second. To tackle this problem, we introduce the SyncNet model in [7], which extracts visual features from 5 frames of cropped faces using a 3D VGG-M network and computes their distance to the MFCC-based audio features. The model searches for offsets within ± 15 frames, attempting to minimize the distance between the two features so that the two modalities are synchronized. We run the model over all the extracted utterances from each video and average the distances across these samples. If the determined offset is greater than ± 7 frames in any clip or the samples do not reach a consent, we will perform shifting of the video stream manually to obtain the final synchronization.

E. Facial Landmark Detection and Mouth Region Extraction

At this stage, we have obtained face tracks of individuals speaking, as well as synchronized audio with corresponding transcripts. The next step is to extract the mouth regions. We first detect facial landmarks with the SeetaFaceEngine2 toolkit. Using these landmarks, the detected faces are first rotated so that the eyes are barely on a horizontal line. Then, a square mouth-centered RoI is extracted for each frame. To account for the yaw variations, the size of the RoI is set to the horizontal distance between the two mouth corners extended by an empirically determined factor of 0.12, or twice the distance between the nose tip and the center of the mouth (d_{MN}), whichever is larger. However, this crop sometimes extends beyond the desired region for extremely small faces, so we restrict the size of the region to be no more than $3.2d_{MN}$.

In other words, the size of a RoI bounding box is determined by

$$w = \min\{3.2d_{MN}, \max\{2d_{MN}, 1.12x_r - 0.88x_l\}\},$$

where x_l and x_r are the x coordinates of the left and right mouth corners. Finally, to smooth the resulting boxes, we apply a first-order Savitzky-Golay filter with window length 3 to the estimated face rotations, the coordinates of the x and y centers, and the size of the ROIs.

F. Validating the Extracted RoIs

On some extremely challenging videos where the yaw and pitch angles are large, the landmark predictor fails and the extracted RoIs are inaccurate or even wrong. We train a binary CNN classifier to remove these non-lip images from the dataset. We begin the training process by using the initial unfiltered crops as positive samples and generate negative samples by shifting the crop region randomly in the original frame. After convergence, we filter the dataset using the trained model and fine-tune on the resulting subset. The trained model has a high recall (e.g. it easily picks up glasses at the top corner, and sometimes fails on profile views and low-resolution images, which are scarcer in the dataset), so we ask a human annotator to revise the inference results and remove false alarms.

IV. Dataset Statistics

LRW-1000 is challenging due to its large variations in scale, resolution, background clutter, and speaker's attributes including pose, age, gender, make-up and so on. They are all important factors to consider when building a robust and practical lip-reading system. To facilitate the study of lip-reading, we provide cropped lip images and so users don't have to struggle with the trivial and cumbersome details of preprocessing. To quantify the properties of the datasets, we perform a comprehensive analysis based on the statistics of several aspects.

A. Source Videos

We select 51 television programs with 840 videos in total, where each raw video has a variable duration of 20 minutes to 2 hours. All the programs fall within the class of news and current affairs. Because different programs always have different broadcasters, we split all the videos of a single program into only one subset of the train, test, and validation set to ensure that there are no or few overlapped speakers among train, test and validation set. In summary, there are 840 videos with about 508 hours' duration in total. The principle of splitting train/test/validation follows two points: (a) there are no or few overlapped speakers in these three sets; (b) the total duration of these three sets follows a ratio of about $8:1:1$, which means the number of samples in these three sets follows a similar ratio round $8:1:1$.

Considering the above two points, we finally select 634 videos of 44 programs with more than 415 hours for training, 84 videos of 4 programs with 43.4 hours for test, and 122 videos of 3 programs with 48.95 hours for validation.

B. Word Samples

In this subsection, we present the statistics about the word samples in this benchmark.

The final extracted samples in our benchmark have a duration of about 57 hours with $718,018$ video clips in total, which are selected from the above 840 raw videos. On the average, each class has about 718 samples which makes it adequate to learn deep models. The minimum and maximum length of the data are about 0.01 seconds and 2.25 seconds respectively, with an average of about 0.3 seconds for each sample. This is reasonable as some words are indeed relatively short in our practical speaking process, especially the wake-up words in many speech-assistant systems. On the other hand, the abundance of such instances in real-world settings also suggests that they should not be overlooked in our research.

C. Lip Region Resolution

Considering the diversity of scales of input videos in practical applications, we do not delete those sequences with small or large sizes. Instead, we collect these words according to their intrinsic natural distribution and found that there are indeed only few low or large sizes. Most of them contribute to some moderate value, as shown in Fig. 3. The two peaks in the figure are resulted by the two different types of the source videos: standard definition (SD) and high definition (HD). The existence of this case brings our benchmark much closer to practical applications which have a large resolution coverage.

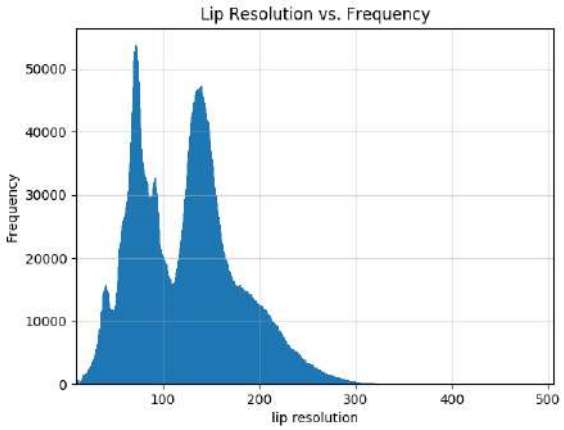


Fig. 3. Scale distribution of the data, measured by the pixel-level width of the lip region.

D. Speakers

There are more than $2,000$ speakers in the 840 videos used to construct our benchmark. The speakers are mostly interviewers, broadcasters, program guests, and so on. The large number and diversity in their identity equip the data with a broad coverage of age, pose, gender, accent, and personal speaking habits. These factors make the data very challenging for most existing lip-reading methods. We would evaluate the state-of-the-art word-level lip-reading models on our benchmark and the results should be very meaningful for designing practical lip-reading models. Among the multiple characteristics of speakers, we select pose as a statistical object because it is believed to be especially critical for the lip-reading task compared with other characteristics. We present the distribution of data in the pitch, yaw, and roll rotations respectively in Fig. 4. We can see that although we do not perform deliberate filtering, the data is still mainly comprised of frontal views.

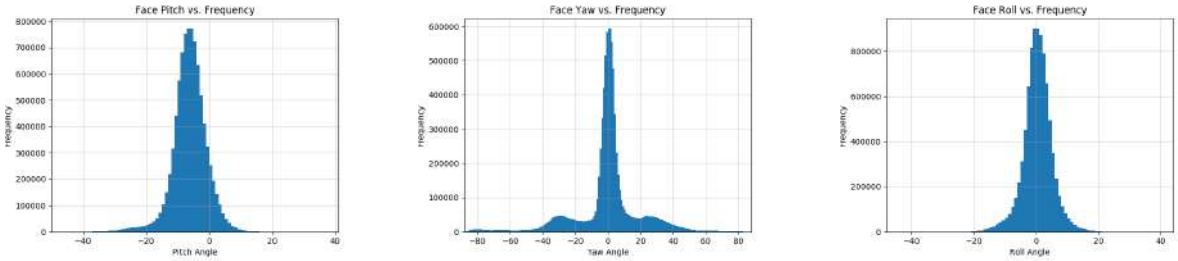
V. Experiments

In this section, we present the evaluation results of popular lip-reading methods and give a detailed analysis to illustrate the characteristics and challenges of the proposed benchmark.

A. Baseline Methods

We cast the word-level lip-reading task on our benchmark as a multi-class recognition problem and evaluate three popular methods on this dataset. Specifically, we evaluate three types of models with different types of front-end network: a fully 2D CNN based front-end, a fully 3D CNN based front-end and a front-end mixing 2D and 3D convolutional layers. Based on these three types of models, we hope to provide a relatively complete analysis and comparison of the currently popular methods.

The first network architecture in our experiments is the *LSTM-5* network based on the multi-tower structure proposed



(a) Pitch Rotation Distribution (b) Yaw Rotation Distribution (c) Roll Rotation Distribution

Fig. 4. Pose distribution of the data in our benchmark, measured by angle degrees. Note that roll has been removed after preprocessing.

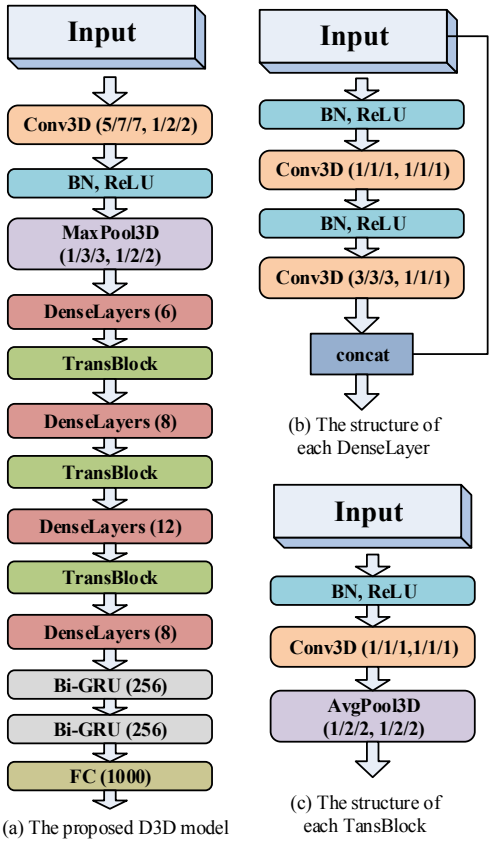
in [9], which is completely composed of 2D convolutional layers. This structure has achieved an appealing performance on the public word-level dataset *LRW* [6]. The second network is based on LipNet [2] which contains only three spatio-temporal convolutional layers as the front-end. The third network is the model proposed in [20] which contains a 3D convolutional layer cascaded with a residual network as the front-end network. During the experiments, the original LipNet consistently failed to converge. We believe that this is because this dataset is too complex to be learned by only three spatio-temporal layers. Therefore, we propose to transform the 2D DenseNet into a 3D counterpart and apply it as the fully 3D convolutional front-end. We named this model as D3D (DenseNet in 3D version), whose structure is shown in Fig. 5. These three models are abbreviated as "LSTM-5", "3D+2D" and "D3D" respectively in the experiments.

To perform a fair comparison, all the three models are combined with a back-end network of the same structure which contains a two-layer bi-directional RNN to perform the final recognition. The recurrent units used in our experiments are bidirectional Gated Recurrent Units. In the remainder of this section, we compare these three models side by side and also find some interesting observations.

B. Experimental Settings

1) *Data Preprocessing:* In our experiments, all the images are converted to grayscale and normalized with respect to the overall mean and variance. When fed into the models, the frames in each sequence are cropped in the same random position for training and centrally cropped for validation and test. All the images are resized to a fixed size of 122×122 and then cropped to a size of 112×112. As an effective data augmentation step, we also randomly flip all the frames in the same sequence horizontally. To accelerate the training process, we divide the training process into two stages. In the first stage, we choose shorter sequences with a length below 30, allowing a larger batch size for training. Then we add the remaining sequences to the training set when the models exhibit a tendency of convergence. We also randomly repeat some samples in the the training process to further accelerate the convergence.

2) *Parameters Settings:* Our implementation is based on PyTorch and the models are trained on servers with four NVIDIA Titan X GPUs, with 12GB memory of each one.



(a) The proposed D3D model (b) The structure of each DenseLayer (c) The structure of each TansBlock

Fig. 5. The proposed D3D network (DenseNet in 3D version).

We use the Adam optimizer with an initial learning rate of 0.001, with $\beta = (0.9, 0.99)$. All the networks are pretrained on *LRW*. During the training process, we apply dropout with probability 0.5 to the last layer of each model to prevent the model from being trapped in some local optima for the *LRW* dataset.

3) *Evaluation Protocols:* We provide two evaluation metrics in our experiments. The *recognition accuracy* over all $1,000$ classes is naturally considered as the base metric, since this is a classification task. Meanwhile, motivated by the large diversity of the data shown in many aspects, such as the number of samples in each class, we also provide the *Kappa coefficient* as a second evaluation metric.

C. Recognition Results

To evaluate the effects of different factors on lip-reading, we split the data into different difficulty levels according to the input scales (resolution), speaker's pose (degree), and

the sample length (number of frames), as shown in Table II. We now present a thorough comparison of the models on all three levels to obtain a complete and comprehensive analysis of the results.

TABLE II

PARTITION OF DIFFERENT DIFFICULTY LEVELS ON *LRW-1000*

Criterion	Easy	Medium	Hard
Input Scale	$\leq$150	$\leq$100	$\leq$50
Pose	$\geq$20	$\geq$40	$\geq$60
Sample Length	$\leq$30	$\leq$15	$\leq$5

(1) *General Performance:* We show the results on *LRW* and *LRW-1000* in Table III and Table IV respectively. We can see that there is a similar trend of these three models in both *LRW* and *LRW-1000*. The method combining 3D convolution together with 2D convolution performs best on both datasets. The *LSTM-5* architecture, which relies only on the 2D convolutional layers performs worse compared to the other two models. This is reasonable because 3D convolution has an advantage for capturing short-term motion information, which has been proved important in lip-reading. However, the network with a fully 3D front-end cannot surpass the model combining 2D and 3D convolutional layers. This result proves the necessity of 2D convolutional layers for extracting fine-grained features in the spatial domain, which is quite useful for discriminating words with similar lip movements. In addition, the performance gap between these three models on *LRW-1000* is not too wide and the Top-1 accuracy ranges from 25.76% to 38.19% among the $1,000$ classes, which confirms both the challenges and the consistency of our data.

TABLE III

RECOGNITION RESULTS ON *LRW*

Method	Accuracy
LSTM-5	66.0%
D3D	78.0%
3D+2D	**83.0%**

TABLE IV

RECOGNITION RESULTS ON *LRW-1000*

Method	Top-1	Top-5	Top-10	Kappa (Top-1)
LSTM-5	25.76%	48.74%	59.73%	0.24
D3D	34.76%	59.80%	69.81%	0.33
3D+2D	**38.19%**	**63.50%**	**73.30%**	**0.37**

(2) *Performance vs. Word Length:* There is a small amount of samples with a relatively short duration in our benchmark, which can be used to roughly evaluate the performance of lip-reading models in extreme cases. The length is measured by the number of frames in our experiments. As shown in Fig. 6 and Table V, all models perform similarly when the word has a relatively short duration. As the length of the word gradually increases, the performance of all three models becomes better and more stable, likely because the context included in a sample increases simultaneously with the word's length. One other possible reason is that the number of samples within classes of longer durations is larger than those of shorter durations.

(3) *Performance vs. Input Scales:* We evaluate the models on the three levels which are divided by resolution, as shown in Table II. Data with a resolution smaller than 50×50 falls in the hard level. Similarly, data with a resolution smaller than 100×100 and 150×150 fall in the medium level and the easy level, respectively. We can see that the performance of the models do tend to increase as we make a transition from the hard level to the medium level and from the medium level to the easy level. As shown in Table II and Fig. 7, the results show that higher input resolution does indeed help improve the lip-reading performance, but the performance would stabilize when the input scale is above some value. On the other hand, the performance gap between these three settings are not wide and the accuracy is still close to 30% for the $1,000$ classes even in the hard-level, where all the test sequences have a resolution below 50×50. This result again demonstrates the consistency of our data which covers a large variation over input scales.

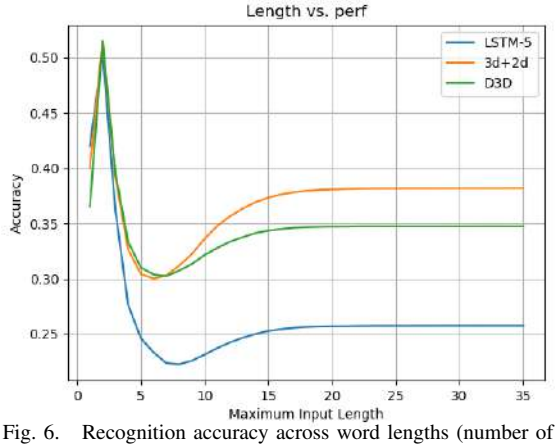


Fig. 6. Recognition accuracy across word lengths (number of frames).

TABLE V

PERFORMANCE W.R.T. WORD LENGTHS (# OF FRAMES) ON *LRW-1000*

Method	Easy	Medium	Hard	All
LSTM-5	25.76%	25.27%	24.63%	25.76%
D3D	34.75%	34.36%	**31.01%**	34.76%
3D+2D	**38.75%**	**37.34%**	30.44%	**38.19%**

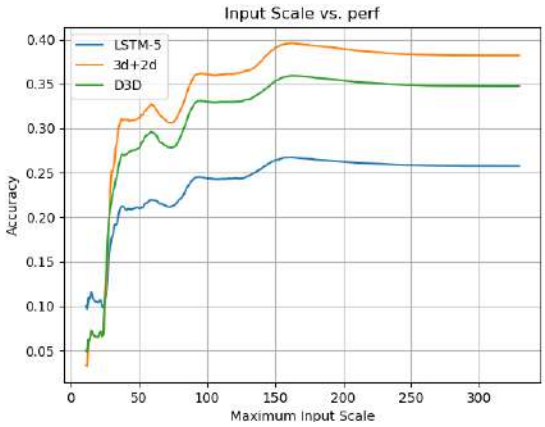


Fig. 7. Recognition accuracy across input scales.

(4) *Performance vs. speaker pose:* In this section, we evaluate the models under different poses measured by the yaw rotation. As shown in Fig. 8 and Table VII, the performance of all three models drops greatly as the yaw angle increases. This may pose a serious challenge to most

TABLE VI

PERFORMANCE W.R.T INPUT SCALES ON *LRW-1000*

Method	Easy	Medium	Hard	All
LSTM-5	26.41%	24.38%	21.02%	25.76%
D3D	35.31%	32.98%	27.75%	34.76%
3D+2D	**39.08%**	**36.07%**	**31.18%**	**38.19%**

current lip-reading models in real-world scenarios. When speakers are viewed from a large angle, there is too much occlusion in the lip region, making it hard to learn the patterns from the data. This significant drop of performance when the camera viewpoints shift from frontal to profile may point out a challenging direction worthy of deeper study.

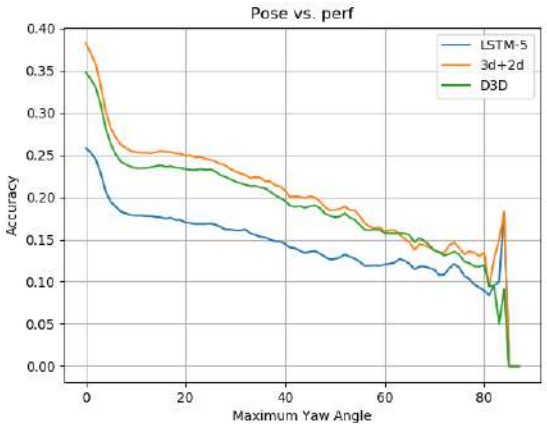


Fig. 8. Recognition accuracy across poses (measured by yaw rotation).

TABLE VII

PERFORMANCE W.R.T POSE ON *LRW-1000*

Method	Easy	Medium	Hard	All
LSTM-5	17.03%	14.51%	11.6%	25.76%
D3D	23.31%	19.95%	15.78%	34.76%
3D+2D	**24.89%**	**20.76%**	**15.9%**	**38.19%**

VI. CONCLUSIONS

In this paper, we have proposed a large-scale naturally-distributed word-level benchmark, named *LRW-1000*, for lip-reading in the wild. We have evaluated representative lip-reading methods on our dataset to compare the effects of different factors on lip-reading. With this new dataset, we wish to present the community with some challenges of the lip-reading task – scale, pose and word duration variations. These factors are ubiquitous in many real-world applications and very challenging for current lip-reading models. We look forward to new exciting research results inspired by the benchmark and the corresponding results provided in this paper.

VII. ACKNOWLEDGMENTS

This research was supported in part by the National Key R&D Program of China (grant 2017YFA0700800), Natural Science Foundation of China (grants 61702486, 61876171). Shiguang Shan and Xilin Chen are corresponding co-authors of this paper.

REFERENCES

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

[2] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. Lipnet: Sentence-level lipreading. *arXiv preprint*, abs/1611.01599:1–12, 2016.

[3] A. Bakry and A. Elgammal. Mkpls: Manifold kernel partial least squares for lipreading and speaker identification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 684–691, 2013.

[4] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

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

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

[7] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016.

[8] Joon Son Chung and Andrew Zisserman. Lip reading in profile. In *British Machine Vision Conference*, pages 1–11, 2017.

[9] Joon Son Chung and Andrew Zisserman. Learning to lip read words by watching videos. *Computer Vision and Image Understanding*, pages 1–10, 2018.

[10] S. Dupont and J. Luettin. Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151, 2000.

[11] Yun Fu, Shuicheng Yan, and T. S. Huang. Classification and feature extraction by simplexization. *Trans. Info. For. Sec.*, 3(1):91–100, March 2008.

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

[13] Bowon Lee, Mark Hasegawa-Johnson, Camille Goudeseune, Suketu Kamdar, Sarah Borys, Ming Liu, and Thomas Huang. Avicar: Audio-visual speech corpus in a car environment. In *International Conference on Spoken Language Processing*, pages 2489–2492, 1 2004.

[14] Y. Li, Y. Takashima, T. Takiguchi, and Y. Ariki. Lip reading using a dynamic feature of lip images and convolutional neural networks. In *International Conference on Computer and Information Science*, pages 1–6, 2016.

[15] 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*, 24(2):198–213, 2002.

[16] 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*, 24(2):198–213, 2002.

[17] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata. Lipreading using convolutional neural network. In *Fifteenth Annual Conference of the International Speech Communication Association*, pages 1149–1153, 2014.

[18] Stavros Petridis, Yujiang Wang, Zuwei Li, and Maja Pantic. End-to-end multi-view lipreading. *British Machine Vision Conference*, 2017.

[19] SeetaFace. Seetafaceengine2. `https://github.com/seetaface`.

[20] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with lstms for lipreading. *Interspeech*, pages 20–24, 2017.

[21] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.

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