Lip Reading in Cantonese

YEWEI XIAO, LIANWEI TENG, AOSU ZHU, XUANMING LIU, AND PICHENG TIAN

School of Information Engineering, Xiangtan University, Xiangtan 411105, China
Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education, Xiangtan University, Xiangtan 411105, China
Corresponding author: Lianwei Teng (manticteng@gmail.com)

This work was supported in part by the Joint Fund for Regional Innovation and Development of NSFC under Grant U19A2083, and in part by the Science and Technology Research and Major Achievements Transformation Project of Strategic Emerging Industries in Hunan Province under Grant 2019 GK4007.

ABSTRACT Lip reading aims at recognizing texts from a talking face without audio information. Due to the rapid development of deep learning techniques, researchers have made giant breakthroughs for both word-level and sentence-level English lip reading in recent years. Unlike English, it is difficult for Chinese to distinguish the lexical meanings, because Chinese is a tonal language. In addition, most of the existing Chinese lip reading datasets are designed for Mandarin, there are few for Cantonese. In this paper, we propose a word-level Cantonese lip reading dataset called CLRW which contains 800-word classes with 400,000 samples. For better practical applications, we do not limit gender, age, postures, light conditions, and speech speed to make CLRW closer to the real scene distribution. At first, we give a detailed description of the data collection process. Next, a novel two-branch network is proposed by us, named TBGL, which consists of a global branch and a local branch. The global branch models the whole lip and the local branch divides the feature into three parts to focus on subtle local lip motion. We jointly train these two branches and achieve comparable performance on LRW, CAS-VSR-W1K, and CLRW, respectively. Finally, we benchmark our dataset and perform a comprehensively analyze of the results, which demonstrate that CLRW is full of challenge, and it will bring a positive impact on further Cantonese lip reading tasks.

INDEX TERMS Lip reading, deep learning techniques, Cantonese, CLRW, TBGL.

I. INTRODUCTION

Lip reading, also known as visual speech recognition, aims to predict the sentence being spoken with only a silent video of a talking face. Lip reading is a developing topic that has attracted increasing attention in recent years and has broad applications in practice, such as improving audio-based speech recognition in noisy environments, aiding hearing-impaired people, analysis of silent movies, biometric authentication in video authentication systems [1], and so on.

Due to the rapid development of deep learning techniques in computer vision domains, English-based lip reading methods have made a huge breakthrough in both word-level and sentence-level English lip reading in recent years. Unlike English, Chinese is a tonal language and the tone is used to distinguish the lexical and grammatical meanings [2]. In addition, there are too many homophones in Chinese, and the lip shapes between homophones are similar, which significantly increase the ambiguity of lip reading. As the most spoken language in the world, there are very few jobs in Chinese lip reading. Yang et al. [3] present the only publicly available large-scale benchmark for word-level Chinese Mandarin lip reading, named CAS-VSR-W1K, which contains 1000 classes with 718,018 samples from more than 2,000 individual speakers on CCTV programs.

Meanwhile, we should not only focus on Chinese Mandarin lip reading, but also Chinese dialects are of great significance. As the only Chinese dialect that has been publicly studied at home and abroad with a complete writing system, Cantonese can be expressed entirely in Chinese characters. It is widely used in Guangdong province, Guangxi province, Hong Kong, Macao, and overseas Chinese community, with a population of 120 million all over the world, which makes the research of Cantonese lip reading be of great significance [4]. As a branch of Chinese, Cantonese is quite different from Chinese Mandarin in terms of grammar, phonetics,
Our main contributions are as follows:

1. We present a large scale Cantonese lip reading dataset, named CLRW with 800 word classes and each class consists of one or several Chinese characters. There are a total of 400,000 samples, with an average of 500 samples per class.

2. We have established a pipeline to automatically collect lip reading datasets. To be closer to the real scene distribution, CLRW contains huge variations in diversity of speakers, background clutters, sample lengths, etc.

3. In the next step, we propose a two-branch model, named TBGL, which contains a global branch and a local branch. In this paper, we introduce a bidirectional knowledge distillation loss for jointly training the two branches. Finally, we evaluate our methods on LRW, CAS-VSR-W1K, and CLRW respectively, and demonstrate the effectiveness of our methods.

### II. GUIDELINES FOR MANUSCRIPT PREPARATION

In this section, we will firstly introduce the pronunciation rules of Cantonese and then give an overview of current mainstream lip reading datasets and some popular lip reading methods.

#### A. THE PRONUNCIATION RULES OF CANTONESE PINYIN

Cantonese is widely used in Guangdong province, Guangxi province, Hong Kong, Macao and overseas Chinese community, with a population of 120 million all over the world, which make the research of Cantonese lip reading be of great significance.

Pinyin is a tool used to assist the pronunciation of Chinese characters including initials, final and tones. Initials are called consonants and are used before finals. Different initials and finals are combined to form different syllables. The main feature of consonants is that the airflow in the mouth will be hindered in various ways during pronunciation, so the degree of facial muscle force and the shape of the mouth are different shape when different initials are pronounced. Therefore, we can obtain the content of speech based on visual information such as mouth shape changes, facial muscle changes, and jaw movements.

Chinese is a monomorphic language, and words do not have strict morphological changes. The syllables of Chinese are composed of initials, finals and tones according to the rules of pinyin. At the same time, Cantonese has a relatively

---

**TABLE 1. The differences between Cantonese Pinyin and Chinese Pinyin.**

|                | initials | Finals | tones |
|----------------|----------|--------|-------|
| Chinese Pinyin | 23       | 24     | 4     |
| Cantonese Pinyin | 19     | 58     | 9     |

---

**FIGURE 1. Examples of speakers in our dataset.**
independent pronunciation system and grammatical norms, which is quite different from Chinese Pinyin. There are three popular Cantonese Pinyin schemes at present: Cantonese Pinyin Scheme, Hong Kong Education Association Scheme, and Hong Kong Language Association Scheme [7]. In this paper, we adopt the Hong Kong Language Association Scheme issued by the Hong Kong Language Association in 1993.

The difference between Cantonese Pinyin and Chinese Pinyin are shown in Tab.1, which are as followed:

1. Chinese Pinyin has a total of 23 initials, 24 finals and 4 tones. Cantonese Pinyin has a total of 19 initials, 58 finals and 9 tones. In Cantonese, most initials are the same as in Mandarin. However, the most obvious difference is that there is no blade-palatal z, ch, sh in Cantonese, which are generally replaced by supradental z, c, and x.

2. In most cases, the same words in Cantonese and Mandarin have different pronunciations. For example, ‘βj’ - ‘reng 2’ (in Mandarin) - ‘jing 4’ (in Cantonese) and ‘βk’ - ‘song 4’ (in Mandarin) - ‘sung 3’ (in Cantonese). The same word in Mandarin and Cantonese often has different pinyin, so the pronunciation of the same word in Mandarin and Cantonese is different.

3. Tones are used to distinguish lexical or grammatical meaning that look similar on the face when pronounced. For example, even though “爱” (which means falling in love) and “哀伤” (which means sadness) have different meanings, but they have the same mouth movements and the same spelling, which adds to the ambiguity when lip reading. Therefore, the tone is an important factor for Chinese lip reading task. There are nine Cantonese tones, but only six of them are different. Tone 7 is the same as Tone 1. Tone 8 is the same as Tone 3 and Tone 9 is the same as Tone 6. But tones 7, 8, and 9 are relatively short-pronounced compared to tones 1, 3, and 6, which makes huge difficulties for Tone recognition and directly makes it difficult for native speakers of Mandarin to understand Cantonese.

### B. LIP READING DATASETS

As a data-driven process, lip reading systems are inevitably influenced by the available data, which contain rich vocabulary, multiple perspectives and complex backgrounds to more closely approximate the real scene distribution. Tab.2 lists some of the major lip reading datasets that have contributed significantly to lip reading.

Early lip reading datasets are designed for classifying isolated speech segments in the form of digits and letters. The AVICAR [8] was recorded in a moving car and proposed in 2004. There were 100 speakers and they were all asked to speak 10 digits in four different views. The AV Letters [9] was released in 2002 and consisted of 5male and 5female, each of whom was asked to utter isolated letters from A to Z three times. These two corpora have made significant progress in the development of lip reading.

However, the main reason for the early research focusing on digit and letter-level datasets is not the ease of relevant feature extraction, but the simplicity of such data collection. With the rapid development of deep learning technology, researchers have made a breakthrough in the field of face detection. We can extract all the faces in the video within a few seconds, which greatly improves the efficiency of lip recognition data collection. Researchers start to pay attention to constructing large scale datasets, and some well-known large scale lip reading datasets are summarized below.

The Oulu VS1 [10] dataset was released in 2009 with 10 phrases spoken by 17 males and 3 females. Each utterance is repeated 9 times. It has been widely used in previous works. However, the average number of samples in each class is merely 81.7, which is not enough to cover the various conditions in a realistic scene. The Oulu VS2 [11] is an extension of the Oulu VS1. It also contains 10 phrases, but the number of speakers is increased to 52. The major highlight of Oulu VS2 is that the Oulu VS2 contains 5 different viewpoints: the frontal, profile, 30°, 45°, and 60°, which make Oulu VS2 a challenging dataset.

LRW [12] is possibly one of the largest English word-level lip reading datasets released in 2016, and it is still available to us today. The dataset has extracted 1000 utterances of

| Dataset            | Language | Year | segment | Speaker | Class | Pose       | Environment |
|--------------------|----------|------|---------|---------|-------|-----------|-------------|
| AVLETTERS [9]      | English  | 1998 | Alphabet| 10      | 26    | Frontal   | Lab         |
| AVICAR [7]         | English  | 2004 | Sentence| 100     | 20    | 4 Views   | In-car      |
| Oulu VS [10]       | English  | 2009 | Sentence| 20      | 10    | Frontal   | Lab         |
| Oulu VS2 [11]      | English  | 2015 | Sentence| 53      | 10    | 5 Views   | Lab         |
| LRW Error! Reference source not found. | English | 2016 | Words   | >1000   | 500   | -30° ~ 30° | TV          |
500 different words, with over 1M word instances, and over 1000 different speakers from British television programs. Unlike traditional datasets, it is no longer limited to a fixed viewpoints and covers a wide variety of speech environments. However, we believe that it should not guarantee that all words have the same duration, which leads to a gap between the data and practical applications, because word frequencies and speech rates between different people and different words in the real-world scene are not uniform.

We have made breakthrough progress in English lip reading. At the same time, researchers have begun to focus on lip reading tasks in their native language. For different languages, there are different vocabulary and grammar. Researchers build AVS for Arabic [13], NDUTA VSC for German, and so on. For Chinese lip reading, the CA VSR1.0 researchers build AVS for Arabic [13], NDUTA VSC for different languages, there are different vocabulary and grammar. Researchers have begun to focus on lip reading tasks in their native language. For different languages, there are different vocabulary and grammar. Researchers build AVS for Arabic [13], NDUTA VSC for German, and so on. For Chinese lip reading, the CA VSR1.0 researchers build AVS for Arabic [13], NDUTA VSC for German, and so on. For Chinese lip reading, the CA VSR1.0 [3] presents the only publicly available large scale benchmark for word-level Chinese Mandarin lip reading, named CAS-VSR-W1K (the original LRW-1000), which contains 1000 classes with 718,018 samples from more than 2,000 individual speakers on CCTV programs.

It is unique in that it consists of different videos with different resolutions, which makes it useful for the natural variation of people speaking in real time, where you can have people speaking at different distances from a video camera. Furthermore, there are no particular restrictions on the character length, phoneme length, or frequency of the involved words due to the wide range of speaker pose, age, gender, speaking speed. These appealing characteristics make CAS-VSR-W1K a challenging benchmark, covering the natural variation in practice across different speech modalities and imaging conditions.

C. LIP READING METHODS

Research on lip reading has a long history. In the past, most methods focus on the hand-engineered features. Some well-known features include the Active Appearance Model(AAM), Local Binary Pattern(LBP), Discrete Cosine Transform(DCT), and Locally Discriminant Graphs(LDG) have been used in lip reading and we can find more details in Zhou’s work [18]. With the rapid development of deep learning and the release of some well-known large-scale lip-reading databases, researchers start to pay more attention to applying deep learning methods to lip-reading tasks.

Noda et al. [19] were the first to use CNNs to extract features for lip reading and they have achieved a better performance than traditional methods. However, it should be noted that the use of 2D CNNs for feature extraction when dealing with sequential inputs is limited even if dynamic frames were to be used as opposed to static features. To solve this problem, Stayfylakis et al. [20] combined 2D CNNs and 3D CNNs by changing the first 2D convolution layer of the ResNet-18 to a 3D Convolution layer to extract robust features with a Bi-LSTM based back-end network to explore the temporal information. The combined use of 3D-ResNet-18 and Bi-LSTM are preferred to be used as the backbone network due to the considerable performance of the model. Martinez et al. [21] improve lip reading accuracy of isolated words by replacing the Bi-LSTM back-end with Multi-Scale TCN. As we all known, TCNs have advantages over LSTM in that they can perform parallel computation with less training time and they are more flexible in changing receptive field size. In paper [22], Ma et al. propose a modification to the system of Martinez et al. by using a Densely Connect TCN instead of MS-TCN, and they achieved state-of-the-art on LRW with an accuracy of 88.4%.

Some impressive lip reading methods start to design specific modules to address some shortcomings of the existing networks for efficient lip reading. For example, Sheng et al. [5] used 38 lip-reading related points as the center of the patch sequence to model the lip contour deformation for capturing the motion of the mouth contour. They use an embedding layer to model the semantic information which contains the local motion information and coordinates information, but they also lose spatial information of parts of the lip inevitably. Hao et al. [23] were the first to use the Temporal Shift Module in the residual branch of each residual block, which can effectively extract the temporal information between adjacent frames without reducing the spatial feature extraction. Xiao et al. [24] proposed a Deformation Flow Network that generates deformation flow to capture the motion information of faces. To provide complementary cues for lip reading, they use a bidirectional knowledge distillation loss to help the two branches learn from each other. In paper, Li et al. [25] proposed a new dual-stream lip reading model called Lip Slow-Fast (LSF) based on the Slow-Fast Net. To obtain subtle lip motion features, two streams with different channel capacities are used to extract dynamic features and static features, respectively.

III. DATASET COLLECTION

Early deep learning models relied on massive amounts of data for training to prevent serious overfitting problems when
learning some features with strong generalization capabilities. In recent years, researchers have used deep learning tools to collect large-scale datasets for lip reading. Existing public datasets can be roughly divided into two categories: word-level lip reading datasets and sentence-level lip reading datasets. There are few public Chinese Mandarin lip reading datasets, and so far there is no public Cantonese lip reading dataset. For better Cantonese lip reading, we propose a large scale word-level Cantonese lip reading dataset in this paper. In this section, we first introduce the principles of data collection, then give the detailed process of data processing procedures, and finally make statistics on CLRW.

A. PRINCIPLES OF DATA COLLECTION
In the process of initial data collection, we follow the following principles:

1. The collected videos should include a talking person facing the camera. Moreover, video clips should not contain invalid frames (no speakers or multiple speakers).
2. The video source should contain clear speech and must be discarded if the ambient noise is too large.
3. The examples of speakers in our dataset are shown in Fig.1, in order to make our dataset closer to the real scene distribution, we don’t make too many restrictions on gender, age, speaking speed and light conditions, etc.

We have collected various forms of video sources including Cantonese news programs, Cantonese variety shows, Cantonese talk shows, Cantonese vlogs, and Cantonese character interviews, to achieve data diversity. The video resolutions are distributed between 1920 × 1080 and 1024 × 576, and the video format is 25fps with H.264 encoding. The Fig. 2 shows the automated data collection pipeline in our dataset.

B. SCENE BOUNDARY DETECTION
We need to perform scene boundary detection on the collected video source, then we preserve the scene with only one speaker facing the camera, and remove the other invalid scenes. We use the global histogram of the image to judge the switching between a single speaker and other scenes in the video and obtain a rough single speaker video clip. The global histogram calculates the difference between adjacent frames according to Formula 1 by counting the number of all pixels in the frame at each gray level.

$$D(i, i + 1) = \sum_{i=1}^{M} |H_i(j) - H_{i+1}(j)|$$  \hspace{1cm} (1)

where $H_i(j)$ represents the value of the histogram of level j in the i frame, and M is the total number of levels of the histogram. We mark the locations where the rate of change is greater than 0.5 as the shot boundary to obtain a rough video clip of a single speaker. Finally, we manually clip the video to generate valid video samples, which should contain a complete sentence and a single speaker.

C. AUDIO-VIDEO-SYNCHRONIZATION
We download videos from Bilibili, YouTube, Guangzhou Radio and Television, TVB, and many other websites. The video and audio stream are inevitably out of sync in the process of repeated encoding.

We first manually filter out the video samples which audios and videos are obviously out of sync. But for small out-of-sync videos, we introduce the SyncNet Model [26] to solve this problem, which is shown in Fig. 3. We take the 5 frames of the video sequence and the 0.2s MFCC features as input, then we use 3D VGG-M and 2D VGG-M to extract visual features and audio features respectively. Within the range of ±15 frames, the model search for an offset that minimizes the L2 distance of the fc6 layer features of the two streams. We calculate the offset of each video sample and average the distances among these video samples as a basis for synchronization. If the offset is greater than ±7 frames, we will discard these video samples.

D. AUDIO INFORMATION ANNOTATION
There are few Cantonese programs with precise subtitles, we thus use an audio stream to generate annotations. In this paper, iFLYTEK commercial-grade Cantonese speech transcription service is used to obtain the text content, word segmentation results and timestamps of the valid video samples.

We transcribed the audio-video-synchronized samples to obtain audio information. However, the phonetic transcription process is not word-for-word, and may have pauses.
or lack of modal particles. After automatically generating the annotated text, we must manually verify the annotation text. During the verification process, the annotation must be strictly based on the voice content. If the ambient sound is too large to affect the labelled text, the sample will be discarded directly.

### E. FACE DETECTION AND MOUTH REGION EXTRACTION

We have obtained valid video samples after audio and video synchronization, we then convert the video clip into image frames and use the mediapipe toolkit to face detection and face tracking. We can control the minimum resolution to detect faces and the sensitivity of face tracking through parameter adjustment. After processing by the mediapipe toolkit, we get the face image sequences and the corresponding facial landmarks. To get the mouth regions, we first rotate the face to horizontal based on the eye positions and normalize the face size to 110 × 110. We mark the mean face rotation angle of each video sample into the text information to facilitate subsequent sample classification. We provide not only aligned face images but also cropped lip frames in the CLRW dataset.

As shown in Fig. 4, the center point of the mouth regions is obtained by calculating the coordinates of the center points of the two lip corners. The distance from the tip of the nose to the center point is \(d_{MN}\), so we choose the larger one as the side length of the mouth region. The length of the cropped area between 2\(d_{MN}\) and the distance of the two corners of the mouth expand outward by 12% along the x-axis respectively. Where \(X_1\) and \(X_2\) are the left and right mouth corners, and \(L\) is the side length of the mouth regions.

\[
L = \max\{2d_{MN}, 1.12X_1 - 0.88X_2\} \quad (2)
\]

### F. CLRW DATASET STATISTICS

CLRW is a challenging large-scale word-level lip reading dataset, which contains large variations in gender, age, head pose, resolutions, backgrounds, light conditions, etc. They are all important factors to make our dataset to be closer to the real scene distributions. We download a total of 417 hours of original videos from Bilibili, YouTube, TVB, Guangzhou Broadcasting Network and many other websites including news, interviews, talk shows and other forms. We end up with about 65 hours of valid video with 30,000 face image sequences and 400,000 video clips in total. Finally, we keep the 800 most frequent word classes, with an average of 500 samples per word class. The distribution of program sources in CLRW is shown in Tab.3.

The sample lengths are distributed between 0.01s and 2s with an average of 0.25s. This is because many words and modal particles are inherently short in pronunciation. Moreover, different speakers have different speech rates so that the lengths of different samples in the same word classes are usually different. The ratio of samples in the training set, test set, and validation set is 8:1:1. For the validity of the dataset, duplicate data and speakers are not allowed in each subset. The sample length can be seen in Fig. 5.

### IV. METHODS

In this section, we will first describe the overall pipeline of lip reading models. Then we will illustrate the design motivation of our two branch network. Finally, we introduce a bidirectional knowledge distillation loss for jointly training the two branches.

#### A. THE ARCHITECTURE OF TBGL

An overview of the pipeline is shown in Fig.6. The pipeline consists of a front-end network and a back-end network. The front-end network is designed to extract visual spatiotemporal features representing lip reading dynamics. The back-end decodes the feature sequences and predicts the probability of each word class.

For our front-end network, we first send the lip image sequences as inputs into a 3D CNN layer to perform an initial spatial-temporal alignment in the sequence. Then we compact the features in the spatial domain with a spatial max pooling. In the next step, we employ a ResNet-18 module to extract discriminative features. All these features obtained from the ResNet-18 module would be fed into a global average pooling layer to further compress. To fully exploit the global spatial information and the subtle local lip motion, we propose a two branch back-end network, which consists of a global branch and a local branch.

**Global branch**: we take the whole feature as the input, aiming at modelling the global motion information of the lip. The global branch back-end contains a MS-TCN for increasing the receptive field to mix up the short term and long term information during the feature encoding.

**Local branch**: we want to enhance the robustness of the system and optimize the fitting degree of models by learning
from the separate lip parts. As shown in Fig. 7, two different speakers have similar lip shapes in two consecutive frames and the lips are divided into three parts: right corner, center, and right corner. For each speaker, the same part from the consecutive frames is put together for comparison. For speaker 1, the difference in the center of the lip is more obvious than in the corners. For speaker 2, the difference in the center of the lip and the left corner are both obvious. We can conclude that the variation of some characteristics occurs in a certain part of the lip, so modeling the whole lip would ignore some fine-grained features. In this paper, we propose a local branch back-end network to focus on subtle local lip motion. Firstly, we divide the whole feature map into three parts according to the real space relationship they have. For each part of the local branch, we use identical MS-TCNs and a softmax layer to independently model their temporal varieties. Finally, we summed the loss of the three parts up as the final loss, which is as follows:

$$\text{Loss}_{ce}^L = -\sum_{n=1}^{N} \sum_{d=1}^{D} \frac{1}{D} [y_d \log(p_n^d) + (1-y_d) \log(1-p_n^d)]$$

(3)

where $\text{loss}_{ce}^L$ stands the final loss of the local branch, $N$ is the number of lip parts, $N=3$. $D$ is the number of total classes, $p_n^d$ is the result on the $d^{th}$ frame from the $n^{th}$ part local branch, $y_d$ is the label target.

### B. THE BIDIRECTIONAL KNOWLEDGE DISTILLATION LOSS

Fusion strategies of two branch networks have been widely used in the field of video analysis tasks for many years. In this section, we employ a bidirectional knowledge distillation loss to provide additional supervision for jointly training the two branches of TBGL. The outputs of the fully connected layers of the global branch are denoted as $z_g$. We set the mean of the outputs of the fully connected layers of the three parts of the local branch as $z_l$. We then obtain the predicted probability distribution over all classes, $q_g$ and $q_l$ as:

$$q^{(i)}_g = \frac{\exp(c^{(i)}/T)}{\sum_j \exp(c^{(j)}/T)}$$

$$q^{(i)}_l = \frac{\exp(c^{(i)}/T)}{\sum_j \exp(c^{(j)}/T)}$$

(4)

where $T$ is a parameter known as temperature, we set $T$ to 20 in this work. The knowledge distillation loss can be
improve the performance of lip reading. The sample length of LRW has already been fixed at 29 frames, and the target word is in the center.

**B. PARAMETERS SETTING**

Our experiment is based on PyTorch and the model is trained on a single NVIDIA 3090 GPU, with 24GB memory. We use the Adam Optimizer with default hyper-parameters, the initial learning rate $\eta = 3e-4$, the weight decay is $1e-4$ and the batch size is set to 32. As for the weight of bidirectional knowledge distillation loss, we set it to 100 at first, and we reduce it by half every time when the validation loss stagnates. We train for 80 epochs using a cosine scheduler, and the learning rate $\eta_t$ at epoch $t$ is calculated as follows:

$$\eta_t = \frac{\eta}{2}(1 + \cos(\frac{t}{T}))$$

(8)

**C. EVALUATION CRITERIA**

We adopt the recognition accuracy of overall word classes on all datasets as the evaluation criterion in our experiments.

**D. EVALUATION OF THE TBGL MODEL**

In this subsection, we will first perform a thorough evaluation of TBGL and compare our method with other word-level lip reading methods on both LRW and CAS-VSR-W1K datasets. Finally, we will discuss the effectiveness of the bidirectional knowledge distillation loss.

1) **EVALUATION OF THE TBGL MODEL**

In TBGL, the global branch models the whole features from the front-end network, while the local branch aims to learn from the separate partial features. The results are shown in Tab.4. On LRW, our model has achieved an accuracy of 87.5% (global branch), 87.5% (local branch) and 88.4% (TBGL). The performance of TBGL is only 0.1% less than MS-TCN-Ensemble [27], which is the state-of-the-art method. Meanwhile, our model produces an accuracy of 47.5% (global branch), 47.7% (local branch) and 49.1% (TBGL), which is higher than the best method [27] (MS-TCN-Ensemble, 46.6%) on the CAS-VSR-W1K dataset. From the above results, we can see that our method has similar performance on the LRW dataset compared to the previous best methods, but has a significant improvement on the LRW-1000 dataset. However, MS-TCN-Ensemble uses multi-stage distillation, which requires huge computational cost and multiple training time for better performance. In contrast, our TBGL model is jointly trained in only one stage, and both global branch and local branch can be computed in parallel. This is where our advantage lies. In addition, the local branch always has a better performance than the global branch on both LRW and CAS-VSR-W1K datasets, the reason lies in that the global branch ignore the local subtle information, which is important for lip reading. Our TBGL model fully utilize both global and local information, and thus achieves impressive performance.

---

**TABLE 5. Evaluation of the bidirectional knowledge distillation loss on LRW.**

| Model                  | LRW  |
|------------------------|------|
| Global branch (w/o $\text{Loss}_{kd}$) | 86.5% |
| Local branch (w/o $\text{Loss}_{kd}$)   | 86.3% |
| Global branch ($\text{Loss}_{kd}$)       | 87.5% |
| Local branch ($\text{Loss}_{kd}$)        | 87.5% |
| TBGL (w/o $\text{Loss}_{kd}$)            | 87.3% |
| TBGL                                   | 88.4% |

**TABLE 6. Different difficult levels on CLRW.**

| Factors/Degree | Easy-level | Medium-level | Hard-level |
|----------------|------------|--------------|------------|
| Sample length  | ≤25        | ≤15          | ≤8         |

defined as:

$$\text{Loss}_{kd}(q_l, q_s) = -\sum_{i=1}^{N} q_l^{(i)} \log q_s^{(i)}$$

(5)

where $q_l$ and $q_s$ denote the soft probability distributions of the teacher network and the student network, respectively. We employ a bidirectional knowledge loss for jointly training TBGL, and the final objective function of TBGL is:

$$\text{Loss}_{Bkd}(q_s, q_l) = \text{Loss}_{kd}(q_s, q_l) + \text{Loss}_{kd}(q_l, q_s)$$

(6)

$$\text{Loss} = \gamma \text{Loss}_{Bkd}(q_s, q_l) + \text{Loss}_{ce}^{g} + \text{Loss}_{ce}^{l}$$

(7)

where $\text{Loss}_{ce}^{g}$ is the standard cross-entropy loss of the global branch and the $\text{Loss}_{ce}^{l}$ is the final loss of the local branch. $\gamma$ is a hyper-parameter indicating the weight of $\text{Loss}_{Bkd}$

**V. EXPERIMENTS**

In this section, we present the evaluation results of popular lip reading methods on LRW, CAS-VSR-W1K and CLRW, respectively. Then we will perform a thorough analysis to illustrate the characteristics and challenges of the proposed benchmark.

**A. DATA PRE-PROCESSING**

For LRW, we shuffle the order of the input video at each epoch and then do face detection and face alignment. We align each frame to the reference mean face shape for cropping a fixed 96 × 96 ROI from the aligned face image so that the mouth region is always roughly centered of the image crop. For CAS-VSR-W1K and CLRW, the mouth ROIs have been prepared already. We use a random crop of 88 × 88 pixels and flip all the frames horizontally with a probability of 0.5. There is no evidence to prove the performance is different when using RGB images, so we convert all frames to grayscale, and normalized them to [0,1]. We choose 29 frames for each word on both CAS-VSR-W1K and CLRW. The target word is put at the center to make the data provide more context and
TABLE 7. Recognition results on CLRW dataset.

| Methods/Degree | Easy-level | Medium-level | Hard-level | Overall   |
|----------------|------------|--------------|------------|-----------|
| The global branch | 45.7% | 43.2% | 40.5% | 42.9% |
| The local branch | 45.9% | 43.2% | 40.9% | 43.2% |
| TBGL           | 46.3% | 43.8% | 41.4% | 44.5% |

2) EVALUATION OF THE BIDIRECTIONAL DISTILLATION KNOWLEDGE LOSS

To train the two branches jointly and further improve the performance of TBGL, we introduce the bidirectional distillation knowledge loss. In this paper, we perform an ablation study on LRW to verify the effectiveness of the bidirectional distillation knowledge loss. We test TBGL and each branch of TBGL with and without (w/o) the bidirectional distillation knowledge loss, respectively. The results are presented in Tab.5, and each model with bidirectional distillation knowledge loss outperforms the same model without bidirectional distillation knowledge loss by a clear improvement. We can thus get the conclusion that the bidirectional distillation knowledge loss not only improve the performance of TBGL, but also improve the accuracy of each single branch of TBGL.

E. DATASET BENCHMARKING

To evaluate the effects of different sample lengths on lip-reading, we split the CLRW dataset into three difficulty levels, as shown in Tab.5.

We divided the CLRW dataset into three levels according to the sample length. The performance with different sample lengths is shown in Tab.6. As expected, we get a better performance on the easy-level dataset and these models perform a similar trend at all levels. As the sample length increases, the recognition accuracy of our models also increases and tends to be stable. It may be because long length samples carry more textual information, and short length samples increase the difficulty of recognition due to their short duration. From our statistics on the sample length in Fig. 5, we can see that there are a certain number of short samples in the CLRW dataset, which are related to the speaking habits of Cantonese speakers. Cantonese speakers generally speak faster than Mandarin speakers in the same context and Cantonese has a lot of modal particles. This is why we have a certain number of short samples in our dataset, which makes our dataset closer to the real scene distribution and full of challenges.

VI. CONCLUSION

In this paper, we have proposed a large scale word-level Cantonese lip reading dataset, named CLRW which contains large variations in gender, age, head pose, resolutions, backgrounds and light conditions, etc. To make CLRW a reliable dataset, we make manual corrections in the scene detection link, sample annotations link and the mouth ROIs extracted link. To fully exploit the global spatial information and the subtle local lip motion for effective lip reading, we propose a two-branch network, named TBGL. The global branch of TBGL models the whole motion of the lip and the local branch divides the feature into three parts to focus on subtle local lip motion. We jointly train these two branches and achieve comparable performance on these three challenging datasets. Finally, we benchmark our dataset and divide the dataset by sample lengths. The consistency and challenge of the experimental results proved that our dataset has research significance and application value. We look forward to the CLRW can fill the gap of automatic Cantonese lip reading and capturing the Interest of more community researchers.

REFERENCES

[1] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, “Lip-interact: Improving mobile device interaction with silent speech commands,” in Proc. 31st Annu. ACM Symp. User Interface Softw. Technol., 2018, pp. 581–593.
[2] Y. Zhao, R. Xu, and M. Song, “A cascade sequence-to-sequence model for Chinese Mandarin lip reading,” in Proc. ACM Multimedia Asia, 2019, pp. 1–6.
[3] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, “LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild,” in Proc. 14th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), May 2019, pp. 1–8.
[4] W. Qian, “Comparison of the differences between Mandarin and Cantonese,” J. Chin., vol. 6, 2010.
[5] C. Sheng, X. Zhu, H. Xu, M. Pietikäinen, and L. Liu, “Adaptive semantic-spatio-temporal graph convolutional network for lip reading,” IEEE Trans. Multimedia, vol. 24, pp. 3545–3557, 2021.
[6] Z. Miao, H. Liu, and B. Yang, “Part-based lipreading for audio-visual speech recognition,” in Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC), Oct. 2020, pp. 2722–2726.
[7] Y. Gao and J. Wen, “A commentary on the Cantonese (Cantonese) pinyin project,” J. Beibu Gulf Univ., 2020.
[8] A. Ortega, F. Sukno, E. Lleda, A. F. Frangi, A. Miguel, L. Buera, and E. Zaccu, “AV@ CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition,” in Proc. LREC. 2004.
[9] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, “Extraction of visual features for lipreading,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 2, pp. 198–213, Feb. 2002.
[10] G. Zhao, M. Barnard, and M. Pietikäinen, “Lipreading with local spatiotemporal descriptors,” IEEE Trans. Multimedia, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.
[11] I. Amin, Z. Zhou, G. Zhao, and M. Pietikäinen, “OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis,” in Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit. (FG), May 2015, pp. 1–5.
[12] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in Proc. Asian Conf. Comput. Vis. Springer, 2016, pp. 87–103.
[13] S. Antar and A. Sagheer, “Audio visual Arabic speech (AVAS) database for human-computer interaction applications,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 3, no. 9, 2013.
[14] A. G. Chitu, K. Driel, and L. J. Rothkrantz, “Automatic lip reading in the Dutch language using active appearance models on high speed recordings,” in Proc. Int. Conf. Text. Speech Dialogue. Springer, 2010, pp. 259–266.
[15] Y. Benezech, G. Bachman, G. Le-Jan, N. Souviraï-Labastie, and F. Bimbout, “BL-database: A French audiovisual database for speech-driven lip animation systems,” INRIA, Tech. Rep., 2011.
[16] S. Tamura, C. Miyajima, N. Kitakoa, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, and M. Fujimoto, “CENSREC-1: AV: An audio-visual corpus for noisy bimodal speech recognition,” in Auditory-Visual Speech Processing. 2010.
[17] Y. Xu, L. Du, G. Li, P. Wu, and X. Zhang, “Chinese audiovisual bimodal speech database CAVSR1. 0,” in Proc. Int. Symp. Chin. Spoken Lang. Process., 2000, pp. 98–101.
[18] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, “A review of recent advances in visual speech decoding,” Image Vis. Comput., vol. 32, no. 9, pp. 590–605, Sep. 2014.
[19] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, “Lipreading using convolutional neural network,” in Proc. 15th Annu. Conf. Int. Speech Commun. Assoc., 2014.
[20] T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” Tech. Rep., 2017.
[21] B. Martinez, P. Ma, S. Petridis, and M. Pantic, “Lipreading using temporal convolutional networks,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2020, pp. 6319–6323.
[22] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, “Lip-reading with densely connected temporal convolutional networks,” in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., Jan. 2021, pp. 2857–2866.
[23] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, “How to use time information effectively? Combining with time shift module for lipreading,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Jun. 2021, pp. 7988–7992.
[24] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, “Deformation flow based two-stream network for lip reading,” in Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), Nov. 2020, pp. 364–370.
[25] H. Li, N. Yadikar, Y. Zhu, M. Mamut, and K. Ubul, “Learning the relative dynamic features for word-level lipreading,” Sensors, vol. 22, no. 10, p. 3732, 2022.
[26] J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in Proc. Asian Conf. Comput. Vis. Springer, 2016, pp. 251–263.

YEWEI XIAO received the B.S. and M.S. degrees from Xiangtan University, Xiangtan, China, in 2000 and 2004, respectively. He is currently an Associate Professor with Xiangtan University. His research interests include deep learning, intelligent information processing, and multi-sensor information fusion.

LIANWEI TENG is currently pursuing the master’s degree in control engineering with Xiangtan University. His research interests include image processing, pattern recognition, and deep learning.

AOSU ZHU is currently pursuing the master’s degree in control engineering with Xiangtan University. His research interests include image processing, pattern recognition, and deep learning.

YEWEI XIAO

LIANWEI TENG

XUANMING LIU is currently pursuing the master’s degree in control engineering with Xiangtan University. His research interests include image processing, pattern recognition, and deep learning.

PICHENG TIAN is currently pursuing the master’s degree in control engineering with Xiangtan University. His research interests include image processing, pattern recognition, and deep learning.

***