**ABSTRACT** Although automatic speech recognition (ASR) technology is mature, there are still some unsolved problems, such as how to accurately identify what the speaker is saying in a noisy environment. Lipreading is a visual speech recognition technology that recognizes the speech content based on the motion characteristics of the speaker’s lips without speech signals. Therefore, lipreading can detect the speaker’s content in a noisy environment, even without a voice signal. This article summarizes the main research from traditional methods to deep learning methods on lipreading. Traditional lipreading methods are mainly discussed from three aspects: lip detection and extraction, lip feature extraction, and classification. Traditional feature extraction methods focus on handmade features, which are, however, not very reliable under unconstrained conditions. In recent years, traditional lipreading methods have been gradually replaced by deep learning methods. The advantage of deep learning methods is that they can learn the best features from large databases. This article analyzes typical deep learning methods in detail according to their structural characteristics, and lists existing lipreading databases, including their detailed information and the methods applied to these databases. Finally, the problems and challenges of current lipreading methods are discussed, and the future research direction has prospected.

**INDEX TERMS** Visual speech recognition, lipreading, deep learning, feature extraction.

**I. INTRODUCTION**

People often communicate through hearing and vision, that is, through voice signals and visual signals. Speech signals often contain more information than visual signals, so many studies have focused on Automatic Speech Recognition (ASR). At present, the ASR can reach a very high recognition rate without the speech signal being severely damaged and has been put into practical use in many fields. Visual speech recognition is a technology that recognizes the speech content by the lip movement characteristics of the speaker based on no speech signal. The information received by the visual channel is two-dimensional. Compared with the one-dimensional voice information received by the voice channel, the visual information often contains more redundant information. So visual speech recognition has always been a difficult problem to solve. Visual speech technology is also known as Automatic Lipreading (ALR), which infers the speech content according to the movement of lips in the process of speaking.

In the real world, there are many people with hearing impairment. They often communicate through sign language or observing other people’s lip movements. But gesture language has problems such as being difficult to learn and understand, and inadequate expression skills. Therefore, ALR technology can help people with hearing impairment communicate with others better to some extent [1], [2]. In noisy environments, the speech signal is easily interfered with by the surrounding noise, resulting in the reduction of recognition rate. However, the visual information needed for ALR will not be affected, so ALR can improve the recognition effect of speech recognition in noisy environments [3], [4]. In the field of security, first of all, with the popularity of face recognition technology, there are many
attacks against face recognition system, such as photos, video playback, and 3D modeling, etc. adding lip features can further improve the security and stability of the security system [5]. In the field of vision synthesis, traditional speech synthesis can only synthesize a single voice, and lipreading technology can generate high-resolution speech scene video of specific people [6]. Besides, in sign language recognition, lip movements are also combined to better understand the content of sign language or improve the accuracy of sign language recognition [7], [8].

The basic theory of lipreading in 1954 was proposed by Sumby and Pollack [9], and it was first proposed that the features of lip motion could be used to identify the speaker’s speech content. In 1984, Petajan [10] successfully extracted features from lip movement and combined them with speech recognition to form an Audio-Visual Automatic Speech Recognition (AV-ASR) system. The results show that the system is more robust than ordinary speech recognition systems. In 1994, Goldschen et al. [11] realized lipreading by using the extracted motion features as the input of the Hidden Markov Models (HMMs). In 1997, Goldschen et al. [12] used HMM to model features and achieved good recognition results. In 2007, Zhao et al. [13] used a new feature representation method based on the spatiotemporal local binary pattern to solve the problem of isolated phrase recognition.

Over the past years, as deep learning technology has obtained extraordinary achievements in various fields, the focus of lipreading has also changed. Instead of trying to design some feature extraction algorithms manually to extract features, researchers used the deep network’s powerful representation learning ability to automatically learn good features according to the task objectives. These features often have good generalization ability and can achieve good performance in a variety of scenarios. In 2011, Ngiam et al. [14] proposed an AV-ASR system based on depth autoencoder and Restricted Boltzmann Machines (RBMs) [15]. The visual feature extraction method based on the deep learning method is introduced into multimodal speech recognition for the first time. In 2014, Noda et al. [16] of Waseda University used CNN as a feature extraction tool for lip image. The experimental results show that the visual features obtained by Convolutional Neural Network (CNN) are significantly better than the traditional methods including principal component analysis. In 2016, Wand et al. [17] used Long Short-Term Memory (LSTM) for lipreading and achieved a recognition rate of 79.6% on GRID. In 2016, Chung and Zisserman [18] established the first large-scale English lipreading database LRW under natural conditions according to the BBC program. Assael et al. [19] proposed LipNet based on the spatial-temporal convolution network and recurrent neural network in 2017 and used CTC as a network loss function in the LipNet network. The WLAS network proposed by Chung et al. in 2017, which is composed of CNN and Recurrent Neural Networks (RNN), obtains a 46.8% sentence accuracy rate on the LRS database with 10000 samples sentences. In 2019, Yang et al. [20] established the largest Chinese lipreading database LRW-1000 under natural conditions according to China CCTV programs.

At present, lipreading methods are divided into two categories according to different feature extraction methods: 1) Lipreading based on traditional manual feature extraction method; 2) Lipreading based on deep learning feature extraction method. For the traditional manual feature extraction method, the lip region should be extracted firstly; then, the feature extraction algorithm designed by the researchers extracting the bottom moving features of the lip region; and then through some linear functions such as Principal Component Analysis (PCA) and Discrete Cosine Transform (DCT) is used to process the extracted features and encode them into equal length feature vectors. Finally, suitable classifiers such as Artificial Neural Network (ANN), HMM are used for classification. In the deep learning method, it can be used iterative learning method to automatically extract more features than traditional methods from the video or image sequence; Then obtain the scores of each category through the deep model, and then adjust the network model parameters by way of backpropagation according to the labels of the training data, and finally achieve a good classification effect.

In this survey, we review the main methods in the development of lipreading. It is divided into traditional methods and deep learning methods. It also summarizes the current lipreading databases, including why these databases were created, the size of the corpus, the number of speakers, the number of samples and resolution, and the highest recognition rate achieved on these databases. In each task, representative databases are selected and compared with some of its classical algorithms. Then some problems and solutions to the lipreading task are summarized. Finally, the future directions of the lipreading task are discussed and prospected.

Although there have been other literature reviews on lipreading. For example, Zhou et al. [4] summarized three problems of visual feature extraction: speaker dependence, head pose, and temporal feature extraction. The report [21] summarized the lipreading database with recognition task-oriented, and introduces the traditional methods of lipreading and the method of deep learning. However, they only summarized the methods based on deep learning architecture up to 2018. Different deep learning architectures have different functions, such as the front-end of feature extraction and the back-end of sequence modeling. It is necessary to carefully review the network structure of the front-end and back-end as the one presented here.

The organization of the survey is as follows:
1) Some problems of lipreading technology faced currently were analyzed;
2) According to the steps of traditional lipreading, the methods of lip detection and extraction, lip feature extraction, and classification methods were introduced and summarized;
According to their functional characteristics, lipreading methods under deep learning were divided into front-end networks and back-end networks;
4) Some influential lipreading databases were introduced and summarized according to their properties, and some classical methods were compared under the same database;
5) The expansion field of lipreading and its future development directions were discussed and prospected.

II. TRADITIONAL FEATURE EXTRACTION AND RECOGNITION METHODS
Lipreading research has a history of nearly 70 years. Some early researchers focused on how to extract better lip movement features and how to recognize speech content in these features. The main steps are as follows: lip detection extraction, feature extraction transformation, and classification, as shown in Figure 1.

1) Lip detection and extraction: The first step in lipreading is to locate and extract the region of interest (ROI) from raw data, that is, detect the face and extract the lip region from the video image.
2) Feature extraction: It indicates to extract some effective features from lip image, which is the key link in lipreading. The representation of the extracted features directly affects the final recognition efficiency. The main methods are color information-based, model-based, and mixed feature extraction.
3) Feature transformation: The features extracted in the second step are often very dimensional. The purpose of feature transformation is to reduce the dimensionality of features. Common methods include PCA, Linear Discriminant Analysis (LDA), DCT, and Discrete Wavelet Transform (DWT).
4) Classification: The last step of lipreading is classification. Common methods include Template Matching, DTW, HMM, Support Vector Machine (SVM), and Time-Delay Neural Network (TDNN).

A. LIP DETECTION AND EXTRACTION
The first step of the traditional lipreading method is to detect the lip region from the raw video. Because lipreading is to recognize the speech content through the visual information of lips, it only needs to pay attention to the visual information of lips. The quality of ROI extraction will also affect the performance of recognition. The methods of lip detection include the color information-based method, face structure-based method, and model-based method.

1) PIXEL INFORMATION-BASED METHODS
The method based on pixel information is to detect the lips by the difference between the lips and the surrounding skin color. Wark et al. [22] put forward a positioning method based on the R/G ratio, which is mainly judged by the ratio of red and green pixels. The points with the ratio in a certain range are considered to be the points in the lip region. The discrimination formula is shown in equation (1), in which \( R \) is the red component and \( G \) is the green component.

\[
L_{lim} \leq \frac{R}{G} \leq U_{lim}
\]  

Lewis and Powers [23] proposed the red exclusion (R-E) algorithm. They believed that the skin color and lip color of the human face generally contain red, and the difference between skin color and lip color is mainly reflected in the green component and the blue component. Based on this, they detect and locate the lip through the following formula, as shown in equation (2), where \( G \) is the green component, \( B \) is the blue component, and \( b \) is the threshold.

\[
\log\left(\frac{G}{B}\right) < b
\]  

In the research of Skodras and Fakotakis [24], RGB image was transformed into \( L^*a^*b^* \) color space to increase the color contrast of lips and skin color, and then K-means clustering method based on color is used to extract key points of lips. Ghaleh and Behrad [25] used RGB color space and fuzzy c-means clustering method to segment lip shape. Gritzman et al. [26] proposed 33 color transformations for lip
segmentation: 21 color channels from 7 color spaces (RGB, HSV [27], YCbCr, YIQ, CIEXYZ, CIELUV, and CIELAB) plus another 12 color transformation methods. The results show that HSV based color transformation is the best for lip segmentation.

2) FACE STRUCTURE-BASED METHODS
The method based on face structure is mainly to locate the lip region according to the distribution characteristics of each organ of the face. The relative positions of eyes, nose, and mouth of different people are fixed, and the lip region can be located according to the proportion of the face. Firstly, the face can be detected by the Haar feature and AdaBoost cascade classifier [28]. Puviasaran and Palanivel [29] locate lip region according to face width and height. See equation (3) for location formula, where $W_f$ represents face width, $W_m$ represents lip width, $H_f$ represents face height, $H_m$ represents lip height. As shown in Figure 2.

$$
\begin{align*}
\frac{1}{4} W_f \leq W_m & \leq \frac{3}{4} W_f \\
\frac{2}{3} H_f \leq H_m & \leq \frac{1}{15} H_f
\end{align*}
$$

![Figure 2. Face structure-based method.](image)

3) MODEL-BASED METHODS
The model-based method is to locate the ROI of lips according to the shape and appearance of the detected lips. The main methods are Snake, Active Shape Models (ASM), and Active Appearance Models (AAM). The Snake algorithm proposed by Kass et al. [30] is also called Active Contour Model (ACM). The algorithm first obtains several lip key points through certain constraint conditions, and then defines a deformable curve, which is fitted to the lip key points under the joint action of internal constraint energy coefficient (the smoothness of constraint curve) and external constraint energy coefficient (the definition of contour features). Dinh and Milgram [31] proposed a multi-feature ASM which combines the normal contour, gray block, and Gabor wavelet to locate the lip, aiming at the fact that the ASM with a single feature will fall into local minimum value under noise environment (such as beard, wrinkles, lip color and low contrast of skin color). Rothkrantz [32] used the AAM model to track lips. AAM model not only establishes shape statistical model reflecting shape change but also establishes the global gray model reflecting texture change. The results of lip detection are better under controlled conditions, but there are many other interferences under natural conditions, such as internal changes (speaker’s beard, wrinkles, etc.) and external changes (changes in light, background, etc.) will affect the results of lip detection. Table 1 lip detection methods are summarized.

B. FEATURE EXTRACTION
After the lip detection and extraction, the next step is to extract the feature of the ROI area for the subsequent classification. Because the lip image sequence contains a lot of redundant information (such as posture, light, skin color, etc.) which is not related to the lipreading task. So how to extract features related to the lipreading task from the lip image sequence is still the key. Ideally, the extracted visual features should have the following characteristics:

1) The number of extracted features and dimensions are as small as possible, but it must be ensured that it can represent the content of the speaker.
2) Speaker independence, which means that the extracted features must be independent of the speaker.
3) Dynamic, that is, the extracted visual features represent the process of speaking, not a static image.
4) Features should be distinguishable and reliable, that is, between different categories, features should be distinguishable, and features between the same categories should be as similar as possible.

There are many methods for lip feature extraction, which can be roughly divided into three categories: pixel-based method, shape-based method, and mixed feature extraction.

1) PIXEL-BASED METHODS
According to the pixel-based method, all pixels in the ROI region of the lip represent the visual speech information [33]. Therefore, the pixel-based method takes all the pixel values in the ROI region of the lip as the original feature space and uses different methods to reduce the dimension of the original feature space to get the expressive features.

a: LINEAR TRANSFORMATION METHOD
Because all pixels need to be regarded as one feature, the feature dimension is often very high, and the linear transformation method is usually used to reduce the feature dimension. These linear transformation methods include PCA [34], [35], DCT [36], DWT [29], [37], [38] and LDA [39], Local Sensitive Discriminant Analysis (LSDA) [38], [40], Maximum Likelihood Linear Transformation (MLLT) [41]. These algorithms transform the lip feature vector, remove the invalid information, and reduce the dimension of the feature vector. Most pixel-based methods are composed of multi-level linear transforms, which are divided into intra-frame linear transforms and inter-frame linear transforms. In essence, the intra-frame linear transformation is to extract the visual language information of a single image; the inter-frame linear
transformation is to extract the dynamic information of lips between video frames, and the combination of this linear transformation can effectively represent the space-time information. Potamianos et al. [42] proposed that Hierarchical Linear Discriminant Analysis (HILDA) is one of the representative algorithms. However, due to the use of all the pixel information on the image, the change of light, the rotation and scaling of lip region, and the change of skin color will have a great impact on the results. And because the pixel-based methods all adopt the linear transformation dimension reduction method, and for the lipreading task, its spatiotemporal characteristics do not meet the linear spatial distribution, and the feature representation ability of linear change extraction is limited, so the pixel-based method limits the improvement of recognition accuracy.

b: OPTICAL FLOW METHOD
Optical flow is the instantaneous velocity of the pixel motion of a space moving object on the observation imaging plane. The optical flow method is a method to find the corresponding relationship between the previous frame and the current frame by using the changes of pixels in the time domain and the correlation between adjacent frames in the image sequence, to calculate the motion information of objects between adjacent frames. In [11], [43], [44], optical flow is used as a feature of the lipreading task. However, the method based on the optical flow itself has a large amount of computation and is sensitive to the illumination and the change of the speaker’s posture.

c: LOCAL PIXEL FEATURE METHOD
The linear change method is to perform a linear transformation on the original pixel value to extract features. Because of its sensitivity to illumination and skin color changes, local features of pixels are introduced to solve these problems. Local Binary Patterns (LBP) is one of the most representative algorithms for pixel local feature extraction. However, LBP can only process a single two-dimensional image, and the input for the lipreading task is image sequence, so Zhou et al. [45] introduced LBP-TOP (Local Binary Patterns from Three Original Planes) [46] to extract spatiotemporal information. The Histogram of Directional Gradients (HOG) is a statistical value for calculating the direction information of local image gradients. Rekik et al. [47] combined HOG features with Motion Boundary Histogram (MBH) feature to extract spatiotemporal information.

2) SHAPE-BASED METHODS
The shape-based feature is to build a model according to the contour of lips when speaking, and a series of parameters that make up the model constitutes the visual feature. Shape-based methods are mainly divided into geometric features and contour features.

Geometric features: the commonly used features of this method are the height, width, perimeter, area, and contour shape of the lips inside and outside. Ma et al. [48] selected six lip points, each of which is recorded as $P_i(x_i, y_i)$, as shown in Fig. 3, and then calculated five geometric features on the lips according to these six lip points, and the calculation formula is shown in equation (4).

$$
\begin{align*}
    f_1 &= \min(x_1, x_5) \\
    f_2 &= |y_6| \\
    f_3 &= y_3 \\
    f_4 &= \max(y_2, y_4) \\
    f_5 &= \max(|x_2|, x_4)
\end{align*}
$$

Contour feature: ACM, also known as the Snake model [30], is an algorithm based on the point distribution model. The essence of this method is to extract some key feature points at the lip edge and then connect the coordinates of these key feature points into a vector, which is used to represent the object. ASM algorithm needs to label the training set with artificial feature points in advance, then obtain the shape model through training, and then realize the recognition of specific objects through the matching of feature points.

---

**TABLE 1. Summary of lip detection methods.**

| Method                      | Algorithm                      | Describe                                                                 |
|-----------------------------|--------------------------------|--------------------------------------------------------------------------|
| Pixel information-based     | Determine location based on R/G | Extracting lip ROI in RGB color space                                    |
|                             | ratio                          |                                                                          |
|                             | Red exclusion (R-E)            |                                                                          |
|                             | Color space conversion         | Transfer the image to YIQ, HSI color space to segment lips               |
|                             | K-means clustering based on color | Key points of lip region extracted by K-means clustering                 |
|                             | Fuzzy c-means clustering based on color | Lip contour extraction in RGB color space by fuzzy c-means clustering |
|                             | Based on 33 color transformations | A comparative analysis of the results of lip extraction with 33 color transformations |
| Face structure-based method | Face structure                  | Lip region detection based on face proportion                            |
| Model-based method          | ACM                            | Use energy function to constrain curve coefficient                        |
|                             | MF-ASM                         | ASM using a combination of three features                                  |
|                             | AAM                            | Use contour appearance and grayscale appearance to detect lip area        |
TABLE 2. Summary of traditional lip feature extraction methods.

| Method               | Algorithm | Advantage                                                                 | Disadvantage                                                                 |
|---------------------|-----------|---------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Pixel-based method  | Linear transformation | 1) All pixels of the image are used to represent the visual features, with less information loss; 2) No complex manual modeling is required. | 1) High feature dimension; 2) Sensitive to image rotation, scale, illumination, and skin color; 3) The generality of different speakers is poor. |
|                     | Optical flow |                                                                           |                                                                            |
|                     | Local pixel feature |                                                                           |                                                                            |
| Shape-based method  | Geometric features | 1) Low feature dimension; 2) Good interpretability.                        | 1) High requirements for image quality; 2) Information on lip movement is incomplete; 3) Accurate manual marking is required. |
|                     | ACM        |                                                                           |                                                                            |
| Mixed feature       | AAM       | 1) Strong ability of feature expression; 2) Different speakers have good generalization | 1) High algorithm complexity; 2) Accurate manual marking is required.         |

Luettin and Thacker [49] applied the ASM model to the lipreading task for the first time.

Compared with the pixel-based method, the shape feature-based method has the advantages of good controllability and interpretability. The more feature points are selected, the more accurate the model is, the stronger the representation ability is, and will not be affected by light, lip rotation and scaling, or skin color. However, the shape model also has defects: firstly, most of the features extracted by the shape model are on the lip contour, which will cause information loss; secondly, the shape model mainly relies on manual feature point annotation, and the accuracy of annotation directly affects the recognition effect; finally, the demand for image quality is high and the calculation is complex, and the calculation of big data is time-consuming.

3) MIXED FEATURE EXTRACTION

The pixel-based method and shape-based method are different. In a sense, the extracted features belong to low-level features, while the latter belongs to high-level features. Mixed feature extraction is a combination of the above two methods, which has complementary advantages and disadvantages. It contains not only lip contour information, but also texture, brightness, and other pixel information. The classic mixed feature is the AAM. Cootes et al. [50] proposed the AAM algorithm in 2001, which combines the gray-scale features of lip region with lip shape features, taking into account not only local feature information but also global contour and texture feature information. In 2016, Watanabe et al. [51] constructed 3D AAM features from three different perspectives (front, left, and right), which can recognize lip pictures from any angle.

AAM algorithm combines the characteristics of pixels and shapes, which has a strong ability of feature expression and is still widely used in the following lipreading research [52]–[57]. However, although this method can accurately represent the lip features, the AAM model still requires high accuracy for manual feature points, and it needs many iterations to get the feature parameters, which often leads to the problem of local optimization. Table 2 summarizes the three methods of traditional lip feature extraction and their advantages and disadvantages.

C. CLASSIFICATION

Some classification methods are used to recognize the previously extracted lip visual features. Previous research methods include the Template matching method, DTW, HMM, Random Forests (RF), Support Vector Machines (SVM), and Time-Delay Neural Network (TDNN). The description of each classifier is shown in Table 3.

1) TEMPLATE MATCHING METHOD

The principle of template matching is to extract features from static images and then match with existing templates. Petajan [10] put each dynamic feature vector of pronunciation into the database in the training stage, match the feature vector of input word with the template in the database in the recognition stage, and the one with the highest correlation coefficient is the recognition result. This algorithm is relatively simple, but it has great defects, because each person speaks at a different speed, resulting in different dynamic features of pronunciation. Petajan just normalizes the sequence of each feature vector and then calculates the correlation coefficient.
TABLE 3. Classifier description.

| Classifier       | Advantage                                                                 | Disadvantage                                                                 |
|------------------|---------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Template Matching| Extracting features from static images and matching them with existing templates. | Ignoring dynamic features.                                                   |
| DTW              | It solves the problem that the features ignored in template matching change with time. | The recognition effect of a large number of words and continuous speech is not ideal. |
| TDNN             | The multi-layer network has a strong abstract ability to features, can express the relationship between visual features in time, and has a strong classification ability. | It cannot handle long-distance dependencies.                                 |
| HMM              | It considers that the lip movement is linear in a short time. It is expressed by the parameters of the linear model, and then many linear models are linked into a Markov chain in time. | HMM only depends on each state and its corresponding observation object.       |

length manually, so the recognition effect is not ideal. Later, Petajan et al. [58] introduced DTW to solve this problem. But DTW can only alleviate this defect. When the isolated words become continuous pronunciation, the corpus becomes large or the specific person becomes the unspecified person, the recognition effect is not very good.

2) ARTIFICIAL NEURAL NETWORK
Due to the limitation of hardware equipment, the only shallow neural network can be designed in the early stages, also known as ANN. ANN has the ability of anti-jamming, self-adaptive learning, and powerful classification. Its classical algorithm is Time-Delay Neural Network (TDNN). TDNN adopts a multi-layer network, each layer has a strong abstract ability to features. Its input is a time window changing with time, so it is more capable of expressing the relationship of features in time than template matching. Compared with the traditional artificial neural network, it can easily learn by sharing weights [59], [60].

3) HIDDEN MARKOV MODEL
HMM is the most widely used method in early lipreading. HMM was proposed by Baum in 1960. It was originally applied to speech recognition. Since 1990, people began to apply HMM to lipreading. The basic idea is that the lip movement is linear in a short time, which is represented by linear model parameters, and then many linear models are linked into a Markov chain in time. Moreover, the lip movement sequence is observable, which represents semantic information, so it has certain syntax rules, which coincides with the double random process of HMM. Sujatha and Krishnan [61] used DCT to extract moving visual features, and then use these visual features as input to estimate the parameters of HMM, when testing, input the features of the test image sequence to get the prediction results. Thangthai et al. [62] used DNN-HMM to model the extracted features in the time sequence. Compared with traditional GMM-HMM, the only difference is that the generation probability of each state is estimated by replacing GMM with DNN. The output of each unit of DNN represents the posterior probability of the state.

In addition to the above classifiers, other classifiers also show good classification effects, such as Support Vector Machine (SVM) [63]–[66], Random Forests (RF) [67], and Multi-Layer Perceptron (MLP) [68], etc.

III. DEEP NEURAL NETWORK BASED METHODS
As the database becomes more and more complex, there are many problems, such as the increase in the number of speakers, the diversity of posture, and the change of illumination conditions and background environment. The traditional manual features are not universal. The researchers found that the Deep Neural Network (DNN) method can learn deeper features from the experimental data, which show good robustness in the case of big data. Different from the traditional lipreading method, the lipreading flow framework based on the deep learning method is shown in Figure 4.

The lipreading method based on deep neural networks adopts the end-to-end approaches, which can automatically learn the characteristics of lip movement information from the video to achieve classification. The first is lip location and extraction, which mainly extracts the lip region from the original video. Then the deep neural network is used to process the processed data. According to the functional characteristics of the module, it can be divided into the following two parts:

1) Front-end: It mainly uses different forms of depth neural network to extract lip movement features from lip images. The feature extraction of the front-end network greatly affects the classification effect. The front-end model includes Feedforward Neural Network (FNN), Deep Belief Network (DBN), Boltzmann Machines (BM), Auto-encoder, and Convolutional Neural Network (CNN).

2) Back-end: It mainly models the features extracted from the front-end network in time. The main models include Recurrent Neural Network (RNN), Temporal Convolutional Network (TCN), and Transformers.

A. LIP DETECTION AND EXTRACTION
As with traditional methods, lipreading based on deep learning methods also needs to extract lip ROI. However, with
FIGURE 4. Deep learning lipreading process: Firstly, the lip region is located and extracted from the video, then the temporal and spatial features are extracted by the front-end, and finally, these features are used as input to the back-end and are concatenated at each time step for classification.

B. FRONT-END AND BACK-END

The front-end network is for feature extraction and representation learning of lip image. In the beginning, although many other deep neural network structures are applied to the front-end network, such as Feedforward Neural Network [17], [70], [71], Autoencoder [72], Boltzmann Machine [73], etc. However, Convolutional Neural Networks (CNN) has always been the most common and effective network architecture for feature extraction. CNN is a kind of feedforward neural network with deep structure and convolution computation. It is one of the representative methods of deep learning. LeNet [74] was born in 1998. But then, the advantages of CNN began to be overshadowed by hand-designed functions. With the proposal of ReLU [75] and Dropout [76], as well as the historical opportunities brought by GPU and big data, CNN ushered in a historical breakthrough in 2012 – AlexNet [77]. Since then, CNN has shown explosive development. The main classic structures of CNN include LeNet, AlexNet, VGGNet [78], GoogLeNet [79], ResNet [80], DenseNet [81], etc. In the Convolution Neural Network, the neurons only need to perceive the local area, and then synthesize the local information at a higher level to get global information. Therefore, the overlay of the convolution layer on the structure is conducive to image recognition, and can effectively extract and combine the obtained features. Later, 3D CNN [82] was proposed to deal with the information of time dimension in video. The main structure of front-end network based on CNN is 2D CNN, 3D CNN and 3D + 2D CNN. The three network structures are shown in Figure 6.

The back-end network mainly models the features extracted from the front-end network in time, because lip movement is a process of movement. The back-end is to learn the long-term dependence and then predict. The main structure of this part is the Recurrent Neural Network (RNN). The RNN is mainly used to deal with the timing problem. The current information mainly depends on the previous state or the first n time steps. However, the common RNN has difficulties in learning the long-term dependence problem, which often leads to the problem that the gradient disappears, resulting in the RNN only has a short-term memory. The Long Short-Term Memory (LSTM) [84] is improved based on RNN. Three “Gates” are used to control the state and output at different times, namely the input gate, forgetting gate, and output gate. LSTM uses ‘gate’ structure to combine long-term memory and short-term memory, which can alleviate the problem of gradient disappearance. The Gate Control Unit (GRU) [85] improves the structure of LSTM. Compared with the three gates of LSTM, GRU simplifies it into two gates: update gate and reset gate. In the classical RNN, the transmission of the state is from front to back. However, in some problems, the current output depends not only on the

the maturity of face detection technology, many pre-trained models can be used now. For example, Dlib library [69] is used to detect 68 landmarks of the face, as shown in Figure 5. Among them, the landmarks of 49 - 68 are lip points, and the area containing only these 20 lip landmarks is cut down as the input of the front-end network.
previous factors, but also on the subsequent sequence factors, which requires the bi-directional recurrent neural network (Bi-RNN) to solve such problems. The structure of Bi-RNN is a combination of two unidirectional RNNs, the input will be provided to the two opposite RNNs at the same time, and the output is determined by the two unidirectional RNNs together (such as Bi-LSTM and Bi-GRU).

This section mainly discusses the front-end network mechanism based on CNN, mainly divided into four subsections: 2D CNN, 3D CNN, 3D + 2D CNN, and other deep learning network architecture.

1) 2D CONVOLUTIONAL NEURAL NETWORK

The common feature of the 2D convolution neural network is that all the features involved are obtained by 2D CNN. A simple way to use 2D CNN in lipreading is to use 2D convolution for every frame in the video. In the 2D convolution network, the convolution layer of 2D convolution extracts features from the adjacent feature map of the previous layer, and then brings the value after adding a deviation into the activation function. The following equation represents the value in the j feature map at the position (x, y) in the i layer.

\[
v_{ij}^{xy} = \tanh(b_{ij} + \sum_p \sum_m \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} w_{pq}^{jm} (x-p)(y-q))
\]

Noda et al. [86] used 2D CNN in the feature extraction of lip movement in 2014, and input its output as a visual feature sequence into GMM-HMM, classifying 300 Japanese words spoken by six people. The experimental results show that the visual features obtained by the convolutional neural network are better than the traditional methods including PCA. Later, Noda et al. [87] added the voice module to form audio-visual speech recognition. Garg et al. [88] used 2D CNN based on VGG, they spliced the sequence pictures of lip movement into a picture as the input of front-end by the network and then used LSTM to process the features extracted from the front-end network. Li et al. [89] calculated the first-order regression coefficient of the current and ± 3 frames of lip images to obtain the lip dynamic feature images, and then used these dynamic feature images as the input of 2D CNN, and then used HMM to classify the extracted features.

Chung and Zisserman proposed SyncNet [90], which consists of five convolution layers and two full connection layers. Firstly, the gray image is input, and the feature vector is input into a single layer LSTM after the single frame feature is extracted from the convolution layer. LSTM models the input feature in time sequence and returns to the prediction class. As a comparison network, the VGG-M, which is pre-trained in ImageNet [91], consists of five layers of convolution layer and three layers of full connection layer, followed by a single layer of LSTM. The weight of VGG-M will not be updated when training LSTM. The recognition rate of SyncNet on OuluV2 is 92.8%, while the pre-trained VGG-M is only 25.4% because training the convolutional network directly on the lipreading database performs better than pre-training on ImageNet.

Lee et al. [92] proposed a multi-perspective lip language recognition system. They did three groups of experiments, the first group was single angle recognition, testing the recognition effect of each angle separately; the second group was cross angle recognition, that is, training all angle data first, and then testing the recognition effect of each angle separately; the third group spliced five angle pictures at the same time into one picture as the input of convolution network, and then verified the recognition effect. All input pictures are RGB images. The convolutional neural network is composed of two convolutions layers and a full connection layer and uses two-layer LSTM to model the features extracted by CNN and return the prediction category.
Saitoh et al. [93] proposed a new sequential image representation method called cascaded frame image (CFI). All frames of a video are connected to one image as the input of the convolutional neural network. CFI can contain the spatiotemporal information of the entire image sequence. They used three CNN models to extract features from CFI, which are Network in Network (NIN) [94], AlexNet, and GoogLeNet. NIN is to construct a micro-network Mlpconv in the network. A Mlpconv layer is equivalent to a convolutional layer plus a multi-layer perceptron. The author uses four Mlpconv plus space maximum pooling layer; AlexNet is composed of five convolution layers and three full connection layers; GoogLeNet is a 22-layer deep network, using sparse connection architecture to avoid computing bottlenecks. Experimental results show that GoogLeNet performs best.

Chung and Zisserman [95] used VGG-based CNN pairs to extract features. They proposed four structures: Early Fusion (EF), Multiple Tower (MT), 3D Convolution with Early Fusion (EF-3), and 3D Convolution with Multiple Tower (MT-3). Among them, EF uses the smallest unit modeled as a single picture into a series of consecutive frames, which are used as the input of the network after the channel dimensions are stitched. MT adopts to first pass each single frame picture through a convolution layer and a pooling layer, and then stitch each feature map in the channel dimension as the input of the subsequent convolution layer. EF-3 and MT-3 use 3D convolution kernels for EF and MT structures. Experiments show that the MT structure of the 2D structure performs best.

Mesbah et al. [96] proposed a new CNN structure (HCNN) based on Hahn moments. This new structure combines Hahn moments to effectively extract and retain the most useful information in the picture to reduce redundancy and CNN’s superior performance in image classification. Hahn moments are used in the first layer to extract moments and input them to CNN. It can effectively reduce the dimensionality of video images and can express images with less information. CNN takes a moment matrix as input, it consists of three convolutional layers and two fully connected layers, and finally, the soft classification outputs the score of the prediction classification.

Zhang et al. [97] proposed LipCH-Net, an end-to-end visual speech recognition architecture, to recognize Chinese sentences. The architecture consists of two deep neural network modules. The first part is the recognition of sequence pictures to Pinyin. The input is a fixed-size grayscale image. The lip space features are extracted through a CNN based on the VGG-M model, and then passed through a 14-layer ResNet, the residual block is composed of two convolutional layers, plus BN and ReLU, and its jump connection help the spread of information. The output of ResNet then passes through two LSTM layers of the back-end network to solve the problem of the motion characteristics of the image at each moment. At the same time, CTC is used as the loss function of the entire network.

2) 3D CONVOLUTIONAL NEURAL NETWORK

2D CNN can only extract the spatial features of the image, it cannot capture the temporal information well. Lip is a continuous process in the speaking process, so the timing information is essential. Referring to the research on behavior recognition [82], [83], [98], [99], researchers introduced 3D CNN into lipreading. 3D CNN and 2D CNN are similar in structure. A feature image can be obtained by a two-dimensional convolution of an image. As a result, 2D CNN will lose the time information of input data after each convolution operation. However, 3D CNN stacks the continuous frames of the video into cubes in the convolution process and uses a 3D convolution kernel on the cubes to get the feature cubes. In this way, the feature map of the current layer can be connected with multiple consecutive video frames of the previous layer. For 3D convolution, the following equation represents the value at the j feature map (x, y, z) of the i layer.

\[ v_{ij}^{\text{3DCNN}} = \text{tanh}(b_{ij} + \sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_j-1} \sum_{r=0}^{R_k-1} W_{pmq} v_{(x-p)(y-q)(z-r)}^{m(i-1)m}) \]

Assael et al. [19] proposed an end-to-end sentence-level lipreading architecture (LipNet). The front-end network is a three-layer 3D convolution layer, which extracts spatiotemporal features from RGB image sequences with fixed input length. The back-end network is two Bi-GRU [100] layers, which models the characteristics of the front-end network in the time domain. At last, softmax is used for classification, and connectionist temporal classification (CTC) [101] loss
function is used for network training. But CTC also has an obvious defect that the input sequence must be larger than the output sequence. Fung and Mak [102] also proposed a network with the same architecture, but they used eight 3D convolution layers followed by a max-out activation function, instead of using the pooling layer, to input the results into the Bi-LSTM network, and use softmax to obtain the final output at the last time step of the sequence.

Torfi et al. [103] proposed a 3D CNN based on coupling for an audio-visual speech recognition system. The visual module uses 3D CNN. Its input is a fixed size 60 × 100 grayscale image. After four layers of 3D convolution, the first three layers are connected with 3D pooling, and the last layer is connected with two layers of full connection. All layers are followed by a PReLU activation function. In the audio module, only the first layer uses a 3D convolution layer to extract spatiotemporal features after extracting MFCC features from speech signals, and the second layer uses 2D convolution to extract spatiotemporal features, and the second layer is also followed by two full connection layers. Then, the two modules are mapped into a representation space, and the corresponding relationship between the audio-visual streams can be judged by the learned multimodal features.

Chung et al. [104] constructed a 3D CNN architecture based on VGG. This architecture is mainly divided into four parts: Watch, Listen, Attend, and Spell (WLAS). Watch and Listen for the front-end network, and Attn and Spell for the back-end network. The Watch module uses five consecutive grayscale images as input, including five 3D convolution layers, one full connection layer, and three LSTM layers. The encoder LSTM network encodes the characteristics of each time step. The structure of the Listen module is the same except that there is no volume layer. The Spell module of the back-end network consists of three LSTMs, two Attention mechanisms [105], and a multi-layer perceptron (MLP). The two attention mechanisms process the context information of watch and listen respectively to generate the context vectors of Watch and Listen. The decoder LSTM network in Spell uses the output characters of the previous step, and the decoder LSTM state of the previous step and the context vectors of Watch and Listen of the previous step generate the decoder state and output vectors. Finally, the probability distribution of output characters is generated by MLP with softmax. The author emphasizes that the attention mechanism is very important. If there is no attention mechanism, the output will have little correlation with input except for the first and second characters. Moreover, the author discusses the necessity of Bi-LSTM. They say that they have tried Bi-LSTM, but the results have not been improved and the training time is longer. Because the attention mechanism has provided additional local focus, there is no need for a bi-directional encoder to provide context information. The visual speech system proposed in [106] used the same structure to explore the problem of side lipreading, but there is no Listen module.

Xu et al. [107] proposed an end-to-end lipreading system named LCANet. The front-end network of the system consists of three parts: 3D CNN, highway network [108], and Bi-GRU. The back-end network is a cascaded attention CTC network. 3D CNN extracts the spatial information and short-term time information features between video frames and then passes through the two-layer HW network. The HW network and inspiration are the same as the LSTM in the structure of adding “door”. It introduces the transform gate T and the carry gate C, and the formula is as follows:

\[ y = H(x, W_T) \cdot T(x, W_T) \cdot + x \cdot (1 - T(x, W_T)) \] (7)

where x is the input and the carry gate is 1 - T(x, W_T). It can be seen from the formula that the HW network is actually to transform one part of the input, that is, it is the same as the traditional neural network, and the other part directly passes through. Finally, Bi-GRU is used to encode long time information. To capture information clearly from a longer context, LCANet inputs the front-end encoded information into the cascaded Attention-CTC network. The attention mechanism weakens the constraint of the conditional independence assumption in CTC loss. Therefore, this method improves the modeling ability of lipreading problem and provides better prediction at the same time. On the other hand, compared with the separate attention mechanism, the cascaded CTC loss can reduce the uneven alignment in the training stage, and CTC will guide attention to eliminate the unnecessary non-sequential prediction between the hypothesis and the reference.

Yang et al. [20] proposed a D3D model based on DenseNet. The front-end network of D3D is based on the network structure of DenseNet. Firstly, the lip sequence image is passed through a 3D convolution layer and a max-pooling. Then through the combination of three DenseBlocks and TransBlock, each DenseBlocks contains several DenseLayer, each DenseLayer is composed of two layers of 3D convolution layer, the first layer convolution uses 1 × 1 × 1 convolution core to fuse the upper layer features and the number of compression channels, and the second layer convolution uses 3 × 3 × 3 to learn the short-time dependence of features. The cross-layer feature comes from dense connection, which is different from RESNET in the form of splicing. Transblock uses a 1 × 1 × 1 convolution to integrate the features of cross-layer splicing through dense connection, and also plays a role in feature compression. Through the alternating superposition of DenseLayers and TransBlock, the feature is mapped into a 256-dimensional vector and sent to the Bi-GRU behind the second layer for long-term dependency modeling.

3) MIXTURE OF 3D + 2D CONVOLUTIONAL NEURAL NETWORK

2D CNN has been proved to be able to extract spatial features effectively, but the disadvantage is that it cannot extract time dynamic features in the sequence. Although 3D CNN can extract the time dynamic features in the sequence, it has a
high demand for hardware performance because of its high computation and storage cost. Therefore, combining the characteristics of both sides, the researchers constructed the structure of a 3D + 2D convolution neural network, which can not only extract the time dynamic features through 3D CNN but also effectively classify the spatial features extracted by 2D CNN. The structure diagram is shown in Figure 7 (c).

Stafylakis and Tzimiropoulos [109] proposed a word-level end-to-end visual speech system. The input of the system is a gray-scale sequence picture, and the duration is fixed to 1 second. The front-end network consists of 3D CNN and 2D ResNet, in which 3D CNN has only one layer to extract short-term movement features of lips. ResNet uses 34 layers, and the max-pooling layer gradually reduces the spatial dimension until its output is a one-dimensional vector in each time step. The back-end network is a two-layer Bi-LSTM network, which takes the one-dimensional feature vector of the front-end as the input to model the long-term dependence. Then they proposed a word embedding deep learning structure [110] for visual speech recognition. Embeddings summarized the lip region information related to lipreading problems and suppressed other irrelevant information, such as the speaker, posture, and lighting. Petridis et al. [111] also refer to the structure of 3D CNN + 2D ResNet, but they use Bi-GRU in the back-end network. Also, an audio stream is added to form an audio-visual voice system.

Margam et al. [112] constructed a 3D-2D CNN architecture for character-level recognition. The front-end of the network consists of two layers of 3D and two layers of 2D. Each 3D and 2D layers have a batch of standardization layers behind it. 3D convolution is followed by a 3D max-pooling layer, and the output of the 2D convolution layer is directly linearized and then input to the back-end network. The back-end network is a two-tier Bi-LSTM. The whole network uses CTC loss for end-to-end training. The input of the network is $100 \times 50 \times 3$ RGB sequence images.

Xiao et al. [113] used Deformation Flow Network (DFN) to learn the deformation flow between adjacent frames. The learned deformation flow is then combined with the original grayscale frames with a two-stream network to perform lipreading. On the one hand, Zhao et al. [114] used local mutual information maximization constraint (LMIM) to constrain the features generated in each time step to maintain a strong relationship with the voice content. On the other hand, they introduced a mutual information maximization constraint at the global sequence level (GMIM) to make the model pay more attention to distinguish the keyframes related to speech content, and pay less attention to the various noises in the speech process. Inspired by the principle of convolution operation, Luo et al. [115] proposed a pseudo-convolutional policy gradient (PCPG) based seq2seq model for the lipreading task. Zhang et al. [116] studied the effects of different facial regions on recognition, including the mouth, the whole face, the upper face, and even the cheeks.

### C. OTHER DEEP LEARNING NETWORK

There are also some end-to-end architectures proposed by researchers that are not based on CNN or RNN. For example, Wand et al. proposed that three network structures were composed of feedforward neural networks and LSTM [17], [70], [71]. The first system is composed of a feed-forward layer, a two-layer LSTM layer, and a softmax layer [17]. The second system is composed of two feed-forward layers, one LSTM layer, and one softmax layer, and combined with domain confrontation training to solve the difference between known and unknown speakers [70]. The third system is composed of three feed-forward layers, one LSTM layer, and one softmax layer [71].

Four end-to-end systems based on sentence recognition proposed by Petridis et al. were composed of an autoencoder and LSTM [72], [111], [117], [118]. The first system is based on two independent streams, the first stream is to extract features directly from a single frame image, and the second stream is to extract features from the differences between two consecutive frames. Both flows pass through three hidden layers and a linear bottleneck layer. Besides, the first derivative and second derivative characteristics are calculated and added to the bottleneck layer. The output of each stream is input to a separate LSTM layer, and then the LSTM output of each stream is input in series to Bi-LSTM to fuse features, and finally, the output of softmax is used to predict classification [72]. The structure of the second system is the same as that of the first system, except that the audio stream replaces the difference stream in the first system, and the unidirectional LSTM after each stream is replaced by Bi-LSTM [117]. The third system focuses on multi-perspective lipreading, which consists of three streams: front picture sequence, 45° angle picture sequence, and side picture sequence. Features are extracted from these three perspectives, and the structure of each flow is the same as that of the second system. The fourth system only has a single video stream, rather than the previous multi-stream fusion [119].

Afousas et al. [120] proposed three back-end network architectures and compared their performance. These three back-end networks have the same front-end network based on [109], while the back-end network is different. The first is composed of three stacked Bi-LSTM layers. The first layer Bi-LSTM receives the visual feature vector, and the last layer Bi-LSTM sends a prediction character probability for each input frame. CTC is used to train the network. Decoding is a beam search that combines the prior information of the external language model. The second is composed of 15 depthwise separable convolution [121], which is a separate convolution of each channel delay dimension, and then $1 \times 1$ channel convolution. As the first method, the network is trained by CTC and decoded by combining the prior information of the external language model. The third is the transformer model [122], which is an encoder-decoder model built by a pure attention mechanism. They use a basic transformer model consisting of six encoders, six decoders, and eight
attention heads. TM does not need an explicit language model to decode, because it learns the implicit language model in the training of visual sequence. However, the author also suggests that it is beneficial to integrate the external language model in the decoding process. In [123], the model proposed by Afouras and others is similar to the third method above. The author uses two loss functions to compare the performance, namely CTC loss and seq2seq loss.

The ConvLSTM network [124] was designed based on the LSTM by adding convolution operation, it can consider the correlation of time and space in the video at the same time, and effectively integrate time and space features, and it has been applied in action recognition [125], [126] and gesture recognition [127], [128]. Courtney and Sreenivas [129] replaced the convolution layer in ResNet with the ConvLSTM layer to expand the learning time and space characteristics of ResNet structure and used CTC to train the network. Wang [130] built a multi-grained spatiotemporal model for lipreading. The author thinks that simple 2D convolution and 3D convolution are not the best methods for lipreading, because some actions may be weak, so they will be lost in the process of pooling. Therefore, the learning process is divided into three sub-modules, which are fine-grained module, medium-grained module, and self-adaptive fusion and coarse-grained module which gather the features from the first two modules. 2D convolution has been proved to be an effective image feature extraction tool in many fields, and there are many words with similar mouth shapes that need to be distinguished by nuances. Therefore, the author uses 2D convolution to extract fine-grained features of each frame of image, which is based on 2D convolution ResNet-18 model; 3D convolution is widely used in video recognition, which can extract the temporal and spatial information between frames. In addition, although some words must be distinguished by subtle differences, most of them have different pronunciation modes, which requires the model to be able to model short-term dynamics with medium granularity. This work is more suitable for 3D convolution, which is based on the 52-layer Densenet-3D model. Everyone has his own speaking style and habits, such as turning his head or nodding his head when speaking at the same time, because of lighting, posture, accent and other factors, even the image sequence of the same word has several different styles. Considering the diversity of appearance factors, a robust lipreading pattern must model the global latent pattern of the sequence at a high level to highlight the representative information in the sequence and cover the subtle style changes in the sequence. The author uses a two-layer two-way ConvLSTM model to fuse the information of the first two modules and model the global latent pattern of the whole sequence, which is called coarse-grained module.

In this section, some lipreading systems based on deep learning are reviewed. Table 4 lists some mainstream lipreading methods from 2011 to the present. Among them, it shows the year of the lip-reading system, the model used (mainly divided into front-end and back-end), what database is used, the recognized pattern (VSR or AVSR) and task (Alphabet, Digits, Words, Phrases or Sentence), and the performance of the system. This article uses the word recognition rate (WRR) to represent the performance of the reading system.

IV. DATABASES AND PERFORMANCE COMPARISON

A. DATABASES

A major obstacle in the study of lipreading is the lack of a standard and practical database. The size and quality of the database determine the training effect of this model, and a perfect database will also promote the discovery and solution of more and more complex and difficult problems in lipreading tasks. There are many limitations in some early databases, the number of speakers and corpus are relatively small, often are simple to identify letters and numbers, and these data are collected under experimental conditions. With the rapid development of deep learning, the database has gone through the process from static to dynamic, from simple to complex, continuous progress and improvement, such as LRW, LRS, and LRW-1000, which contain thousands of speakers, with a total sample number of hundreds of thousands.

In the following sections, we will introduce and summarize the database’s recognition tasks (identifying letters, numbers, words, phrases, sentences, or multi-views). The summary includes the year of database release, the language used, the number of speakers, classes, task, recording environment resolution, and best accuracy. Figure 8 shows some sample databases.

1) LETTER AND NUMBER RECOGNITION

The process of research is always from easy to difficult, and lipreading is no exception. Most of the early lipreading databases were used to recognize simple English letters and numbers. Compared with the recognition of words and sentences, they need to distinguish fewer categories and do not need to consider the influence of the preceding and the following. Table 5 summarizes the database for letter and number recognition.

The earliest database of letter recognition was AVLetters database [148], which was published by Matthews et al. in 2002. There were 10 people (5 men and 5 women) who read 26 letters A-Z three times in disorder. The video was recorded at 25 frames per second, and the audio sampling frequency was 22.5kHz. The database also provided 80 × 60 pixels images containing only lips. Lee et al. published the audio-visual voice database AVICAR [149] collected in the automobile in 2004. They installed four cameras on the dashboard to record video, resulting in four different views of the synchronous video stream. There were 100 participants (50 men and 50 women) in the recording, with 13 numbers, letters, and 20 TIMIT sentences [150]. Avletters2 [151] improved the resolution to 1920 × 1080 compared with AVLetters, and used 50 frames per second recording, but the number of speakers was reduced to 5. Also, the corpus of AV@CAR [152] also has the part of word spelling.
| Year | Reference | Model | Front-end | Back-end | Pattern | Database | Task | WRR (%) |
|------|-----------|-------|-----------|----------|---------|----------|------|---------|
| 2011 | Ngiann et al. [14] | Autoencoder + RBMs | CNN | VSR | AVLetters | Alphabet | 64.40% |
| 2014 | Noda et al. [86] | GMM-HMM | VSR | Alphabet | 68.70% |
| 2015 | Noda et al. [87] | M5-HMM | VSR | ATR | 37.00% |
| 2015 | Sui et al. [11] | DBN-M+DCT+LDA | VSR | AusTalk | 22.50% |
| 2015 | Sui et al. [73] | DBN | VSR | 67.80% |
| 2015 | Ninnonya et al. [132] | DBN | VSR | CENSREC-1 | 69.10% |
| 2016 | Petidis et al. [133] | DBNs + DCT | LSTM | VSR | AVLetters | 39.30% |
| 2016 | Hu et al. [134] | RTM | SVM | VSR/AVSR | AVLetters | 81.10% |
| 2016 | Takashima et al. [135] | CBNs | VSR | Alphabet | 58.40% |
| 2016 | Garg et al. [88] | CF+VGG | LSTM | VSR | MIRACL-VSC | 64.63%/66.04% |
| 2016 | Li et al. [89] | Feed Forward | LSTM | VSR | GRID | 31.21%/74.77% |
| 2016 | Chung and Zisserman [90] | VGG-M | LSTM | VSR | ATR | 76.00% |
| 2016 | Lee et al. [92] | CNN | LSTM | VSR | AVSR | 71.76% |
| 2016 | Zimmermann et al. [136] | PCAnet plus LSTM | LSTM | VSR | OuluVS2 | 39.10%/25.40% |
| 2016 | Satoh et al. [93] | NIN, AlexNet, GoogLeNet, EF | LSTM | VSR | OuluVS2 | 94.10%/92.80% |
| 2016 | Chung and Zisserman [18] | MT EF-3 MT-2 | LSTM | VSR | LRW | 57.00% |
| 2016 | Assael et al. [19] | 3D CNN | VSR | GRID | 61.10% |
| 2017 | Rahmani and Almasgani [137] | DBN | LSTM | VSR | CUAVE | 43.90% |
| 2017 | Gutierrez and Roberts [138] | VGG-16 | VSR | MIRACL-VSC | 46.20% |
| 2017 | Wand et al. [70] | Feed Forward | LSTM | VSR | GRID | 57.00% |
| 2017 | Petidis et al. [72] | Autoencoder + LSTM | Bi-LSTM | VSR | OuluVS2 | 57.00% |
| 2017 | Petidis et al. [117] | Autoencoder + Bi-LSTM | Bi-LSTM | VSR/AVSR | OuluVS2 | 57.00% |
| 2017 | Petidis et al. [118] | Autoencoder + Bi-LSTM | Bi-LSTM | VSR | OuluVS2 | 91.80%/98.60% |
| 2017 | Chung et al. [104] | CNN + LSTM | LSTM + Attention | VSR | LRW | 96.90% |
| 2017 | Chung and Zisserman [106] | CNN + LSTM | LSTM + Attention | VSR | LRW | 97.00% |
| 2017 | Torfi et al. [103] | 3D CNN | LSTM | VSR | LRW | 76.20% |
| 2017 | Stafylakis and Tzimiropoulos [109] | 3D CNN + ResNet | LSTM | VSR | LRW | 49.80%/86.10% |
| 2017 | Stafylakis and Tzimiropoulos [110] | 3D CNN + ResNet + word boundaries | LSTM | VSR | LRW | 91.10% |
| 2017 | Wand et al. [71] | Feed-forward | LSTM | VSR/AVSR | LRW | 93.60% |
| 2017 | Petidis et al. [119] | Autoencoder | Bi-LSTM | VSR | LRW | Words | 83.00% |
| 2018 | Chung and Zisserman [139] | CNN | LSTM | VSR | LRW | 88.08% |
| 2018 | Hashmi et al. [140] | CF + CNN + BN + Softmax | LSTM | VSR | LRW | 69.70% |
| 2018 | Fung and Mak [102] | 3D CNN | LSTM | VSR | LRW | 68.00% |
| 2018 | Shillingford et al. [141] | 3D CNN + Bi-LSTM | LSTM | VSR | LRW | 71.50% |
| 2018 | Stafylakis et al. [142] | 3D CNN + ResNet | LSTM | VSR | LRW | 52.90% |
| 2018 | Petidis et al. [111] | 3D CNN + ResNet | LSTM | VSR | LRW | 88.08%/86.10% |
| 2018 | Petidis et al. [143] | 3D CNN + ResNet-18+Bi-LSTM (EF) | LSTM | VSR/AVSR | LRW | 82.00%/86.00% | 57.90%/96.40% | 57.90%/95.30% |
TABLE 4. (Continued.) Lipreading system based on deep learning-Part II.

| Year | Reference | Model | Pattern | Database | Task      | WRR (%) |
|------|-----------|-------|---------|----------|-----------|---------|
| 2018 | Afouras et al. [120] | 3D CNN + ResNet 2D CNN DenseNet3D 2D+3D CNN | Bi-GRU + Attention-CTC | VSR | LRS2-BBC Sentences | 37.80% |
| 2018 | Afouras et al. [123] | 3D CNN + ResNet Transformer TM-CTC TM-seq2seq | VSR/AVSR | LRS2-BBC Sentences | 45.30%/91.80% |
| 2018 | Xu et al. [107] | 3D CNN+HW Bi-GRU + Attention-CTC | VSR | GRID Sentences | 97.10% |
| 2018 | Yang et al. [20] | 3D CNN+HW Bi-GRU | VSR | LRW-1000 Words | 25.76% |
| 2019 | Courtney and Sreenivas [129] | Res-BiConvLSTM | VSR | LRW Words | 85.20% |
| 2019 | Fengbour et al. [144] | Decoder-Encoder LSTM CFI + QVGG + Committee | VSR | TIMIT Sentences | 72.20% |
| 2019 | Jang et al. [145] | CNN Bi-LSTM+MD_ATT_MC | AVSR | Chinese TV Sentences | 93.15% |
| 2019 | Moshbah et al. [96] | CFI + Hahn CNN | VSR | OuluVS2 Phrases | 38.19% |
| 2019 | Margam et al. [112] | 3D CNN+2D CNN Bi-LSTM-CTC | VSR | Indian English Sentences | 87.70% |
| 2019 | Weng and Kitani [147] | Two-Stream+13D VGG-M + ResNet + Bi-LSTM+CTC ResNet- DenseNet3D BiConvLSTM | VSR | LRW Words | 84.11% |
| 2019 | Zhang et al. [97] | GRU-Attention | VSR | CCTV Sentences | 50.20% |
| 2019 | Wang [130] | BiConvLSTM | VSR | LRW LRW-1000 Words | 83.34% |
| 2020 | Zhang et al. [116] | Bi-GRU | VSR | LRW LRW-1000 Words | 45.24% |
| 2020 | Xiao et al. [113] | Bi-GRU | VSR | LRW LRW-1000 Words | 41.93% |
| 2020 | Luo et al. [115] | Bi-GRU | VSR | LRW LRW-1000 Words | 85.30% |
| 2020 | Zhao et al. [114] | Bi-GRU | VSR | LRW LRW-1000 Words | 84.41% |

The number recognition database was originally the Tulips database [153] released in 1995. It consisted of 12 people (9 men and 3 women). Each speaker reads 1-4 English numbers twice, a total of 96 samples. The video was recorded at 30 frames per second, the resolution of lip picture is 100 × 75 pixels, the frequency of audio sampling was...
11.127kHz, 8-bit quantization. The M2VTS database [154] contained 37 people (25 men and 12 women). The corpus was 0-9 consecutive French numerals, which were spoken five times by each person. They were collected every other week to observe the influence of subtle facial changes on the results. There were 1850 sentences in total. The XM2VTSDB database [155] contained 295 people participating in the recording. The corpus included: "0 1 2 3 4 5 6 7 8 9", "5 0 6 9 2 8 1 3 7 4" and "Joe too parents green shoe bench out". To capture the natural changes caused by the changes in the speaker’s physical condition, hairstyle, clothing, and mood, the researchers divided the speaker into four independent stages for recording in five months. BANCA database [156] and XM2VTSDB database were collected in one project. However, the BANCA database was specially designed for authentication in three different environments (controlled, degraded, and converse). It contained 208 speakers and covered four different languages (English, French, Italian, and Spanish). The VALID [157] database was designed to study lipreading tasks under different light and noise conditions. It contained 530 digital speech videos recorded by 106 speakers in five different environments. The corpus was isolated or continuous. The speaker was divided into two parts. The first part was 36 people. The experiment was divided into four groups: the first group said 50 isolated numbers for each person; the second group said 30 isolated numbers for each person while their head moves; the third group recorded both the front and the side when each person says 20 isolated numbers; the fourth group said 60 numbers for each person in succession, of which the first 30 heads were fixed, and the head moved when the last 30 numbers were said.

The second part was about 20 people (10 pairs, one pair is A and the other is B) to solve the problem of multi-objective speech. There were also three groups of experiments: In the first experiment, A said a continuous number sequence first, and then B went on; The second experiment was B first, then A; the third experiment was A and B say different number sequences together. In recent years, OuluVS2 [173] database has established a multi-perspective lipreading database, which contained 53 speakers. 159 digital speech sentences were collected, and its resolution was 1920 × 1080 pixels. The AV Digits [119] database collected in 2018 also adopted three perspectives, including 53 speakers and nearly 800 digital sentences.

Besides, other databases are dealing with digit recognition tasks, such as AV@CAR, which also contained digit recognition in Spanish; AVICAR, IBMIH [158], IBMSR [159] and QuLips [162] databases for English; CENSREC-1-AV [166] databases for Japanese; NDUTAVSC [163] databases for German; LTS5 [170] databases for French; AGH AV [166] databases were created for polish.

2) WORD, PHRASE, AND SENTENCE RECOGNITION

Early lipreading databases were mostly alphabet-level and number-level databases. The first reason is that the collection is simple and does not require too much time, and the second is because the recognition task is simple and does not need to consider the problem of different homonyms. But the goal of the lipreading task is not to recognize letters and numbers, but to recognize statements of any length under any conditions. Therefore, later researchers focused on more

| Name               | Year | Language | Task | Classes | Utterances | Speakers | Resolution    | Best Acc. |
|--------------------|------|----------|------|---------|------------|----------|---------------|-----------|
| Tulips [153]       | 1995 | English  | Digits| 4       | 96         | 12       | Lip:100×75, 30fps | 89.58% [153] |
| M2VTS [154]        | 1999 | French   | Digits| 10      | 2920       | 37       | 286×350, 25fps    | 76.60% [154] |
| XM2VTSDB [155]     | 1999 | English  | Digits| 10      | 1064       | 295      | 720×576, 25fps    | N/A       |
| AVLetters [148]    | 2002 | English  | Alphabet | 26 | 780       | 10       | Lip:80×60, 25fps   | 69.60% [159] |
| CUAVE [158]        | 2002 | English  | Digits| 10      | 7,000      | 36       | 720×480, 30fps    | 83.00% [160] |
| BANCA [156]        | 2003 | Multiple | Digits| 10      | 29,952     | 208      | 720×576, 25fps    | N/A       |
| AV@CAR [152]       | 2004 | Spanish  | Alphabet | 26 | 800      | 10       | 768×576, 25fps   | 23.00% [161] |
| AVICAR [149]       | 2004 | English  | Digits| 10      | 59,000     | 100      | 720×480, 30fps    | 37.87% [162] |
| IBMIFH [163]       | 2004 | English  | Digits| 10      | 16,197     | 79       | 720×480, 30fps    | 65.00% [163] |
| VALID [157]        | 2005 | English  | Digits| 10      | 1590       | 106      | 720×576, 25fps    | 63.21% [157] |
| AVLetter2 [151]    | 2008 | English  | Alphabet | 26 | 910      | 5        | 1920×1080, 50fps  | 91.80% [159] |
| IBMSR [164]        | 2008 | English  | Alphabet | 10 | 1661     | 38       | 368×240, 30fps    | 68.58% [165] |
| CENSREC-1-AV [166] | 2010 | Japanese | Digits| 10      | 3234       | 93       | 720×480, 30fps    | 39.30% [132] |
| QuLips [167]       | 2010 | English  | Digits| 10      | 3600       | 2        | 780×576, 25fps    | 98.60% [167] |
| NDUTAVSC [168]     | 2010 | German   | Digits| 10      | 6907       | 66       | 640×480, 100fps   | 84.24% [169] |
| LTS5 [170]         | 2011 | French   | Digits| 10      | 180        | 20       | 1920×1080, 25fps  | 72.30% [170] |
| AGH AV [171]       | 2012 | Polish   | Digits| 10      | 2000       | N/A      | 1920×1080, 50fps  | N/A       |
| AUS-Talk [172]     | 2014 | English  | Digits| 10      | 24,000     | 1000     | 640×480          | 69.18% [65] |
| OuluVS2 [173]      | 2015 | English  | Digits| 10      | 1590       | 53       | 1920×1080, 30fps  | 96.90% [118] |
| AV Digits [119]    | 2018 | English  | Digits| 10      | 795        | 53       | 1280×780, 30 fps  | 68.00% [119] |
TABLE 6. Databases for word, phrase, and sentence recognition.

| Name                  | Year | Language | Task               | Classes | Utterances | Speakers | Resolution | Best Acc. |
|-----------------------|------|----------|--------------------|---------|------------|----------|------------|-----------|
| XM2VTSDB [155]        | 1999 | English  | Sentences         | 7       | 885        | 295      | 720×576, 25fps | N/A       |
| IBMViaVoice [175]     | 2000 | English  | Sentences         | 10,500  | 24,325     | 290      | 704×480, 30fps | 87.05%    |
| CAVSR1.0 [174]        | 2000 | Chinese  | Words             | 78      | 3120       | 20       | 352×288, 25fps | 98.20%    |
| VidTIMTIM [176]       | 2002 | English  | Sentences         | 346     | 430        | 43       | 512×264, 25fps | 57.00%    |
| AV@CAR [152]          | 2004 | Spanish  | Sentences         | 250     | 6000       | 20       | 768×576, 25fps | 23.00%    |
| AVICAR [149]          | 2004 | English  | Sentences         | 1317    | 59,000     | 100      | 720×480, 30fps | 37.87%    |
| AV-TIMTIM [177]       | 2004 | English  | Sentences         | 450     | 4660       | 223      | 720×480, 30fps | 64.80%    |
| IBMHII [163]          | 2004 | English  | Sentences         | N/A     | 16,197     | 113      | 720×480, 30fps | 65.00%    |
| HIT Bi-CAV [186]      | 2005 | Chinese  | Sentences         | 200     | 6000       | 10       | 256×256, 25fps | 78.30%    |
| UWB-05-HSVC [187]     | 2005 | Czech    | Sentences         | 200     | 20,000     | 100      | 720×576, 25fps | 84.20%    |
| GRID [189]            | 2006 | English  | Sentences         | 51      | 34,000     | 34       | 360×288, 25fps | 98.70%    |
| CMU AVPFB [190]       | 2007 | English  | Words             | 150     | 15,000     | 10       | 640×480, 30fps | 32.39%    |
| HIT-AVDB-II [191]     | 2008 | Chinese  | Sentences         | 15      | 4500       | 300      | 780×576, 25fps | N/A       |
| IV2 [192]             | 2008 | French   | Sentences         | 15      | 4500       | 300      | 640×480       | N/A       |
| UWB-07-ICAV [193]     | 2008 | Chinese  | Sentences         | 15      | 4500       | 300      | 640×480       | N/A       |
| OnluVS [178]          | 2009 | English  | Sentences         | 15      | 4500       | 300      | 640×480       | N/A       |
| NDUTAVASC [168]       | 2010 | English  | Sentences         | 15      | 4500       | 300      | 640×480       | N/A       |
| WAPUSK20 [195]        | 2010 | English  | Phrases           | 52      | 2000       | 20       | 640×480, 32fps | N/A       |
| LILIR [179]           | 2010 | English  | Sentences         | 200     | 2400       | 12       | 720×576, 25fps | 61.80%    |
| BL [196]              | 2011 | French   | Sentences         | 238     | 4046       | 17       | 640×480, 30fps | N/A       |
| UNMC-VIER [197]       | 2011 | English  | Sentences         | 12      | 2460       | 123      | 708×640, 29fps | 91.51%    |
| MOBIO [180]           | 2012 | English  | Sentences         | N/A     | 150        | 150      | 640×480, 16fps | 87.40%    |
| MIRACL-VC1 [47]       | 2014 | English  | Words             | 10      | 1500       | 15       | 640×480, 15fps | 95.40%    |
| AusTalk [172]         | 2014 | English  | Words             | 966     | 966,000    | 1000     | 640×480       | 69.18%    |
| OnluVS2 [173]         | 2015 | English  | Phrases           | 530     | 7950       | 53       | 1920×1080, 30fps | 96.90%    |
| RM-3000 [182]         | 2015 | English  | Sentences         | 3000    | 3000       | 1        | 360×640, 60fps | 75.58%    |
| IBM AV-ASR [181]      | 2015 | English  | Sentences         | 10,400a | N/A        | 262      | 704×480, 30fps | 30.64%    |
| TCD-TIMTIM [184]      | 2015 | English  | Sentences         | 5954    | 6913       | 62       | 1920×1080, 30fps | 43.61%    |
| HAVRUS [196]          | 2016 | Russian  | Phrases           | 200     | 4000       | 20       | 640×460, 200fps | N/A       |
| LRW [18]              | 2016 | English  | Words             | 500     | 538,766    | 1000+    | 256×256, 25fps | 83.00%    |
| MODALITY [183]        | 2017 | English  | Sentences         | 182*    | 5880       | 35       | 1920×1080, 100fps | 54.00%    |
| LRS2-BBC [104]        | 2017 | English  | Sentences         | 17,428a | 118,116    | 1000+    | 160×160, 25fps | 66.50%    |
| MV-LRS [106]          | 2017 | English  | Sentences         | 14,960a | 74,564     | 1000+    | 160×160, 25fps | 43.60%    |
| VLRF [202]            | 2017 | Spanish  | Sentences         | 1374a   | 10,200     | 24       | 1280×720, 50fps | 20.00%    |
| AV Digits [119]       | 2018 | English  | Phrases           | 10      | 5850       | 39       | 1280×780, 30fps | 69.70%    |
| LRS3-TED [203]        | 2018 | English  | Sentences         | 70,000a | 165,000    | 1000+    | 224×224, 25fps | 62.80%    |
| LRW-1000 [20]         | 2018 | Chinese  | Words             | 1000a   | 718,018    | 2000+    | 128×128, 23-30fps | 38.19%    |
| LSVR [141]            | 2018 | English  | Sentences         | 127,055a| 2,934,899  | 1000+    | Multiple     | 59.10%    |

* Number of words.

challenging word recognition, phrase recognition, and sentence recognition. The difficulty of word recognition is due to a large number of words, while the difficulty of phrases and sentences is not only to consider the number of words but also to consider the relationship between the words before and after. And then comes the database of words, phrases, and sentences. Table 6 summarizes the databases for word, phrase, and sentence recognition.

The word-level database has CAVSR1.0 [174], which was a Chinese database published by Xu et al. in 2000. It included 20 people. The corpus consisted of 78 kinds of Chinese characters. The syllables contained all initials and finals in Chinese. Each person read them twice. MIRACL-VC1 [47] database was released in 2014 by Rekik et al. There were 15 participants. The corpus consisted of 10 words and 10 phrases. Each person reads the words and phrases 10 times. Because the database used an RGBD camera for recording, the images were divided into color picture and depth picture, which contained more depth information. LRW [18] was the most widely used and challenging word-level English lipreading database. The videos in the database were no longer recorded in a laboratory environment but were collected from BBC TV news programs in the UK. There were more than 1000 speakers in the database, and
each speaker’s posture and lighting were different, including 500 words, each word controlled in 5 to 10 letters, each word contains at least 800 samples. LRW-1000 [20] was collected from 51 TV programs of 26 TV stations in China. It differed from the above databases in that its sample resolution was not fixed. The database contained more than 1000 different classes, and the total number of samples reached 718,018. Each sample contained different numbers of Chinese characters. Because the database considered many conditions, such as the speaker’s posture, age, illumination, and so on, it was the largest Chinese lipreading database so far.

The earliest database of phrase and sentence level is XM2VTSDB [155] database, but it only contained the sentence “Joe too parents green shoe bench out”, so it was not very appropriate to call it sentence level. Then there was the IBMViaVoice [175] database, which contained 290 speakers, 10,500 words, and 24,325 sentences utterances. Unfortunately, this database is not publicly available. VID-TIMIT [176] database contained 43 speakers, each speaking 10 sentences. These sentences were selected from the TIMIT corpus, so these sentences were voiced balanced. Similarly, the sentences of AVICAR [149] database and AV-TIMIT [177] database were also from TIMIT corpus, but the number of people and sentences had increased a lot. AVICAR contained 100 people who speak a total of 10,000 sentences, while AV-TIMIT was 223 people and 450 different sentences. OuluVS [178] database was a widely used database at that time. It contained 20 people, and the corpus was 10 English phrases used in daily life, a total of 1000 sentences. The MIRACL-VC1 [47] database contained a total of 3000 statements for 15 people and provided depth pictures. MOBIO [180] database was shot by mobile phone because the scene of shooting people was not fixed, so the speaker’s head position, background, lighting, and so on are different. A total of 150 people recorded 31,000 sentences. IBM AV-ASR [181] database contained 262 people, and sentences contain more than 10,000 words.

There are also databases created for specific tasks. For example, RM-3000 [182] database, in which only one person participated in the recording, said a total of 3000 sentences. The MODALITY [183] database had a high requirement for video quality. It was recorded at 100 frames per second with a resolution of 1920 × 1080 pixels. Similarly, the TCD-TIMIT [184] database also used 1920 × 1080 pixels for recording. OuluVS2 [173] database used five cameras to record at five angles at the same time, including a total of 1060 sentences by 53 speakers.

Besides English, there are databases in other languages. For example, HIT Bi-CAV [186] database was the first sentence level Chinese database, involving 10 people (5 men and 5 women). The corpus consisted of 200 common Chinese phrases, which cover all the vowels in Chinese. During the acquisition process, each speaker read each sentence three times at an even speed. Then HIT-AVDB-II [191] database was collected from Chinese poems, which contained 30 people, each reading 11 Chinese poems. IV2 [192] database was a sentence level database based on French, with 300 people participating in the recording, each speaking 15 French sentences. There are also Czech databases UWB-05-HSVC [187] and UWB-07-ICAV [193], NDU-TAVSC [168] database for German, HAVRUS [199] database for Russian, BL [196] database for French, and VLRF [202] database for Spanish.

With the great achievement of deep learning in the field of computer vision, the application of neural networks in lipreading has become more and more extensive. However, the training of deep learning often needs a lot of data, so a large-scale lipreading database was also emerging. Although the GRID [189] database was created earlier, it has been used more and more times this year. There were 34 participants, each of whom said 1000 sentences. The sentences in this database were six fixed words, and the sentence form was < verb > + < color > + < position > + < digit > + < letter > + < adverb >. Among them, verbs, colors, prepositions, and adverbs were all four words, the numbers were ten numbers from 0 to 9, the letters were 25 (excluding w because W is longer than other letters), and the total number of words was 51. Each speaker said 1000 sentences. During the recording, each sentence was finished in 3 seconds. Similar to this is WAPUSK20 [195] database, which added ‘ W ’ to the building word as well. LRS2-BBC [104] database was the same as LRW, which collected 246 hours of videos from 2010 to 2016, more than 10,000 sentences. The data contained 17,428 different words and 118,116 sentences. The MV-LRS [106] database was also collected in the BBC program. Unlike LRW and LRS2-BBC, which face angle deviation was small. The face angle of MV-LRS was 0° to 90°. LRS3-TED [203] database was a collection of 150,000 sentences from TED programs. Finally, the LSVSR [141] database collected data from the video website YouTube, yielding 140,000 hours of audio segments, including 2,934,899 speech statements and more than 127,000 words, which is the largest database so far. Video sequences are represented as identically-sized frames (128 × 128) stacked in the time-dimension.

3) MULTI VIEW DATABASES

Most lipreading tasks are performed from the front of the speaker. But in reality, we can’t make sure that all the input data is from the front shooting. When there is only a fixed camera, and the speaker is accompanied by face rotation in the process of speaking, then the data we get will exist in the image sequence of multiple angles in the same speech. Therefore, some databases provide data from different perspectives during the speaker’s speaking process. And the researchers found that the front shooting is not necessarily the best angle for lipreading. A slight angle deviation proved to be beneficial because the protrusion and fillet of the lip can be better observed [204]. In this section, we review some databases that are not positive data collection. Table 7 summarizes the multi-perspective database.
The collection of a multi-perspective database is diversified. In this article, the collection methods are summarized into four kinds: (1) only collect the front and side view; (2) place multiple cameras between 0 and 90 degrees of the speaker; (3) the face of the speaker is rotating when collecting data; (4) the angle reflected in each sample is different.

CMU AVPFV [190], IV2 [192], BL [196], and UNMC-VIER [197] all place a camera on the front and single side of the speaker, and collect the front and side views at the same time. The IBMSR [164] database placed three cameras on the front and both sides of the speaker and collects views from three angles at the same time, but unfortunately, the IBMSR database was not public. The AVICAR [149] database placed four cameras on the dashboard of the car and simultaneously collects views from four angles on the front.

AV Digital [119] database placed the camera at three angles of 0°, 45°, and 90°. The HIT-AVDB-II [191] and LTS5 [170] collected view data at 0°, 30°, 60°, and 90°. LILiR [179] and OuluVS2 [173] collected view data at 0°, 30°, 45°, 60°, and 90°. LRS2-BBC [104] collected view data at 0°, 30°, 60°, and 90°.

LRW [18], LRS2-BBC [104], MV-LRS [106], LRSTED [203], LRW-1000 [20], and LSVER [141] were not collected in a controlled laboratory environment like previous databases. They were all collected on TV shows. Because the speakers in the video were uncontrolled, they talked freely to each other. Therefore, the angles collected were also variable and exist between 0° and 90°, and the angle of view between the speaker and the camera was not explicitly given.

### B. PERFORMANCE COMPARISON

In this section, we will compare the performance of some lipreading algorithms. Because each algorithm aimed at different recognition databases, and the experimental conditions of databases were different, it is difficult to make a completely fair and just comparison. We select the most widely used databases in the alphabet, numbers, words, phrases, and sentences as well as multi-perspective recognition tasks, which are AVLetters, CUAVE, GRID, OuluVS2, and LRW, and list the most representative lipreading algorithms under each database. Table 8 lists the comparison of representative algorithms under these databases. Table 9 summarizes and compares the front-end and back-end structures of the DL architecture.

For alphabet recognition, the most famous is the AVLetters databases. Zhao et al. [178] used LBP-TOP to extract features and SVM to classify, reaching 62.80% WRR. The RFMA based system proposed by Pei et al. [205] had reached the highest WRR of 69.60%. Petridis and Pantic [133] mixed DBNFs features and DCT features, and then used LSTM to model the time series, and achieved 58.10% classification accuracy. Hu and Li [134] proposed a system based on multimodal RBMs (MRBMs), named Recurrent Temporal Multimodal Restricted Boltzmann Machines (RTMRBMs), and achieved 64.63% WRR.

CUAVE databases are the most widely used for digital recognition. Papandreou et al. [160] used AAM to extract lip features and HMM to model time series, reaching a maximum
of 83.00% WRR. Ngiam et al. [14] achieved 68.70% WRR based on the structure of Auto-encoder and limited Boltzmann Machine. Rahmani and Almasganj [137] extracted the deep bottleneck features, and used GMM-HMM for timing modeling, reaching 63.40% WRR. Petridis et al. [72] achieved 78.60% WRR using the dual flow method.

GRID database was created earlier, but it is still used more in recent years. Wand et al. [17] used Eigenlips, HOG, and feedforward neural networks to extract features respectively. The first two features were classified using SVM, and the latter used LSTM time series modeling. The experimental results show that. The combination of feedforward networks and LSTM achieved the best recognition effect. Then Assael et al. [19], Xu et al. [107] and Margam et al. [112] obtained 95.20%, 97.10%, and 98.70% WRR respectively by using the combination of 3D convolution and bi-directional recurrent network.

OuluVS2 database is the most widely used multi-view database. Lee et al. [92] used the combination of DCT and PCA to extract features, and then input the features into HMM to get 63.00% phrase WRR. Then they used CNN to extract features and input LSTM to model the time sequence and got 83.80% phrase WRR. Wu et al. [66] extracted the mixed features of SDF and STLP and classified them with SVM, which achieved 87.55% phrase classification accuracy. Petridis et al. [118] obtained 96.90% of the phrase WRR based on the three-stream method.

LRW databases and LRS2-BBC databases were collected from the BBC. Because the speaker was out of control, the recognition tasks are full of challenges. Chung and Zisserman [18] obtained 61.10% WRR using MT structure to extract spatiotemporal features. Later, Torfi et al. [103] used the coupled 3D convolutional neural network architecture to build the audio-visual speech recognition system and achieved 98.50% WRR. Stafylakis and Tzimiropoulos [109] used 3D convolution plus RESNET to extract features and used Bi-LSTM to model the extracted features, obtaining 83.00% WRR. In LRS2-BBC databases, Chung et al. [104] proposed a WAS system that achieved 49.80% WRR, plus the voice stream WLAS system achieved 86.10% WRR. Then Afouras et al. [123] used 3D convolution plus ResNet front-end and transformer back-end to train with CTC loss and seq2seq loss respectively and obtained 45.30% and 51.70% WRR respectively.

It can be seen from table 8 that the traditional methods were still mainly used in the task of letter and number recognition. The main reason is that the current research focuses on more categories of word classification and continuous sentence recognition, and the previous research on letter and number recognition has declined. Another point is that the recognition categories of letters and numbers are less, and the number of samples in the database is only hundreds or even dozens, which results in that the method based on deep learning has no advantage over the traditional method. On GRID, OuluVS2, and LRW databases, deep learning methods show.

| Name   | Year | Front-end | Back-end | Task | Utterances | Patter | Best Acc     |
|--------|------|-----------|----------|------|------------|--------|--------------|
| AVLetters | 2002 | LBP-TOP   | SVM      | Alphabet | 780     | VSR   | 62.80% [178] |
| 2013   |      | RFMA      |          |        |           |        |              |
| 2016   |      | DBNF + DCT| LSTM     |        |           |        |              |
| 2016   |      | RTM-RBM   | SVM      |        |           |        | 58.10% [133] |
| CUAVE  | 2009 | AAM       | HMM      |        |           |        | 83.00% [160] |
| 2011   |      | Autoencoder + RBMs | GMM-HMM | Digits | 7000     | VSR   | 68.70%/82.20% [14] |
| 2017   |      | DBNF      |          |        |           |        |              |
| 2017   |      | Autoencoder + LSTM | Bi-LSTM  |        |           |        | 63.40% [137] |
| GRID   | 2016 | Eigenlips | SVM      | Sentences | 34,000  | VSR   | 78.60% [72]  |
| 2016   |      | HOG       |          |        |           |        |              |
| 2016   |      | Feed Forward | LSTM   |        |           |        |              |
| 2018   |      | 3D CNN    | Bi-LSTM + CTC |        |           |        |              |
| 2019   |      | 3D CNN + 2D CNN | Bi-LSTM-CTC |        |           |        |              |
| OuluVS2| 2016 | DCT + PCA | HMM      | Phrases  | 7950     | VSR   | 63.00% [92]  |
| 2016   |      | SDF + STLF | SVM      |        |           |        |              |
| 2016   |      | CNN       | LSTM     |        |           |        |              |
| 2017   |      | Autoencoder + Bi-LSTM | Bi-LSTM  |        |           |        | 83.80% [92]  |
| LRW    | 2016 | CNN + LSTM | LSTM + Attention |        |           |        | 76.20% [104] |
| 2017   |      | 3D CNN    | Bi-LSTM  | Words    | 538,766  | VSR   | 83.00% [109] |
| 2017   |      | 3D CNN + ResNet | Bi-LSTM  |        |           |        |              |
| 2017   |      | 3D CNN + ResNet + word boundaries | Bi-LSTM  |        |           |        |              |
| LRS2-BBC| 2017 | CNN + LSTM | LSTM + Attention |        |           |        | 88.08% [110] |
| 2018   |      | 3D CNN + ResNet | TM-CTC  |        |           |        |              |
| 2018   |      | 3D CNN + ResNet | TM-seq2seq |        |           |        |              |

It can be seen from table 8 that the traditional methods were still mainly used in the task of letter and number recognition. The main reason is that the current research focuses on more categories of word classification and continuous sentence recognition, and the previous research on letter and number recognition has declined. Another point is that the recognition categories of letters and numbers are less, and the number of samples in the database is only hundreds or even dozens, which results in that the method based on deep learning has no advantage over the traditional method. On GRID, OuluVS2, and LRW databases, deep learning methods show
TABLE 9. Comparison of front-end and back-end networks.

| Network | Model       | Advantage                                                                 | Disadvantage                                                                 |
|---------|-------------|---------------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Front-end | 2D CNN      | It can extract fine-grained spatial features with fewer parameters.       | It cannot extract the spatial information between frames in the video.         |
| Front-end | 3D CNN      | It can extract the short-term temporal information between video frames and extract the spatial features in the frames. | It has a large number of parameters and cannot extract fine-grained spatial features. |
| Front-end | 3D + 2D CNN | It can simultaneously extract fine-grained spatial features and short-term temporal information between frames with fewer parameters. | It can only extract the short time information between frames. |
| Back-end | LSTM/GRU    | It can model a long-time unidirectional sequence.                         | The transmission of its state is unidirectional from front to back. The sequence must be processed one by one. |
| Back-end | Bi-LSTM/Bi-GRU | It can model a long-time unidirectional sequence.                       | The current state is predicted by the before and after states. The sequence must be processed one by one. |
|          | Transformer | It processes sequences in a non-sequential manner, that is, parallel processing. | It can extract short-term time information on the video. |

their advantages. For LRS2-BBC, LRW-1000, and LSVSR database, the traditional methods are almost invalid, and the deep learning method cannot show good results and needs to further improve the neural network feature extraction ability and time sequence coding ability.

V. DIFFICULTIES AND CHALLENGES OF LIPREADING

The main reason why lipreading is challenging is that its input is the video (also known as image sequence), and most of the image content is unchanged. The main difference is the change of lip movement. However, action recognition, which belongs to video classification, can be classified through a single image. While lipreading often needs to extract the features related to the speech content from a single image and analyze the time relationship between the whole sequence of images to infer the content. Therefore, the main difficulties of lipreading are as follows:

**External factors:** The diversity of external influencing factors such as illumination, skin color, and beards. Because different speakers have skin color, wrinkles on the skin, beard or not, as well as external light and background changes, which will cause interference in lipreading. It has a great influence on feature extraction. The traditional lipreading method adopts the shape-based method, because the extracted features only include the shape of the lip [48], [49], and there is no other irrelevant information such as illumination, skin color, and beard. In the method based on deep learning, big data training, and deep neural network are used to extract deep temporal and spatial features related to lip movement.

**Visual ambiguity:** First of all, in the process of pronunciation, there are some different phonemes with the same mouth shape. For instance, the consonant phonemes /P/ and /B/, /D/ and /T/ in English are visually indistinguishable, so it is difficult to distinguish them without context. The phenomenon of weak sound and continuous reading in English will also lead to the loss of viseme. And the speaker’s accent problem will also lead to a phoneme, but the mouth shape of the pronunciation is different. Some researchers used different phoneme-to-viseme mappings [41], [161], [206]. Also, some authors use adjacent characters or words to solve the problem of visual ambiguity [18], [104], [109].

**The difference of speaker’s pose:** In reality, people may be accompanied by the rotation of their heads in the process of speaking. This change results in a change in lip angle. Because of the different postures of different speakers, there are different angle samples. It is very hard to extract features bound up with speech content from these samples. The current DL based methods are still rarely targeted to solve these problems. Only rely on the support of large-scale databases. The multi-view database contains multiple views of the speaker when speaking. For example, LILIR and OuluVS2 both contain five perspectives of the speaker. After that, LRW, LRS2-BBC, LRW-1000, and other databases contained more diversified perspectives. These databases are of great help to solve the problem of the face angle of speakers.

**Speaker-dependent:** In some databases, the number of speakers is very small, and it is often necessary to identify the content of new speakers in practical applications. There are differences in different people’s speaking habits, and lipreading is to extract features from lip images. These features often contain the speaker’s information, which is irrelevant to what is said. So how to extract speaker-independent information in lip pictures is also a problem. In traditional methods, LDA is widely used to deal with speaker dependency since it tries to pull the class means away from each other and push data points of the same class together at the same time [39]. The shape-based approach is also used to solve the problem of speaker dependence. Another solution is to include more speakers in the training data. In recent years, the database has become more and more perfect. For example, LRW and LRW-1000 have thousands of speakers. The influence of the speaker’s dependence is less and less.

**Database problems:** Some early databases are monotonous in number, corpus, sample, or background. This inevitably
limits the limitations of the lipreading system. In recent years, both LRW, LRS, and LSVSR are striving to increase the number of speakers and the content of their speeches in the database. However, these databases still have some common problems. For example, these videos are collected from TV programs, and the background, environment, illumination, and other conditions are relatively stable, and the language content is relatively limited. A large-scale database with multiple speakers and diverse postural backgrounds plays a major role in the development of lipreading technology.

VI. CONCLUSION
This article presents the development of lipreading including traditional lipreading focuses on finding hand features with strong expressive ability, so far DCT and AAM features are still the most widely used features. As well as the selection of classifiers, HMM and its variants are also the best models for context modeling. The method based on deep learning focuses on the construction of DNN architecture, from the beginning of the front-end only 2D CNN or 3D CNN, to the later add ResNet or HW network to build a deeper network to extract more expressive features. Then the researchers found that the performance of 3D + 2D CNN is better than that of single 2D CNN or 3D CNN. For the back-end, the use of Unidirectional RNN from the beginning to the back, as well as the use of a transformer in the past two years, all indicate that researchers apply the latest deep learning technology to lipreading tasks. With the deepening of lipreading research, the existing lipreading methods still cannot meet the actual needs, and the research of lipreading still has a long way to go. There are many existing problems and possible research directions in the future.

The construction of a large-scale audio-visual database. In the actual scene, there are scene transformation and various noises. Although the deep learning model has a strong ability of data expression, the quality of the model is still based on the scale of the database. Unfortunately, most of the current databases still have great defects, even the largest LRW, LRW-1000, and LSVSR databases. Because they are collected in TV programs, their background and illumination are relatively fixed. Besides, what the speaker said may also have certain limitations. How to build a more comprehensive and realistic database is still an important problem.

The choice of lip ROI. Nowadays, most lipreading research is to extract a fixed size of lip ROI as input, and the size of this size is still an open problem. Koumparoulis et al. [207] proved in the experiment that the selection of ROI of different sizes of lips will have an impact on the final recognition results, but still cannot determine the optimal ROI size selection scheme. Many people talk. At present, the database provides a single person speaking scene, in the actual scene, there are often many people speaking at the same time. How to find the speaker in the multi-person scene and identify the content of each person’s speech has not been studied. Besides, at present, most of the research focuses on off-line recognition, and the bi-directional RNN used can refer to context information. The real-time recognition in the real scene can only refer to the previous information, so the recognition accuracy will be reduced.

The development of audio-visual speech recognition. Lipreading is an interdisciplinary subject, which is closely related to computer vision and natural language processing. In the process of lipreading research, other emerging research directions also have great application value. For example, more voice generated lip motion video or generated voice based on lip motion video, which is very helpful for animation. It also provides a new research idea to add lipreading methods to the cocktail party problem (speech separation task) [208] and speech enhancement task [209].

REFERENCES
[1] E. Antonakos, A. Roussos, and S. Zafeiriou, “A survey on mouth modeling and analysis for sign language recognition,” in Proc. 11th IEEE Int. Conf. Workshops Automat. Face Gesture Recognit. (FG), vol. 1, May 2015, pp. 1–7.
[2] R. Seymour, D. Stewart, and J. Ming, “Comparison of image transformation-based features for visual speech recognition in clean and corrupted videos,” EURASIP J. Image Video Process., vol. 2008, no. 1, 2008, art. no. 810362.
[3] G. Potamianos, “Audiovisual automatic speech recognition: Progress and challenges,” J. Acoust. Soc. Amer., vol. 123, p. 3939, Jan. 2008.
[4] 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.
[5] Z. Akhtar, C. Micheloni and G. L. Foresti, “Biometric liveness detection: Challenges and research opportunities,” IEEE Security Privacy, vol. 13, no. 5, pp. 63–72, Sept./Oct. 2015, doi:10.1109/MSP.2015.116.
[6] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesizing Obama: Learning lip sync from audio,” ACM Trans. Graph., vol. 36, no. 4, p. 95, Jul. 2017.
[7] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, “Weakly supervised learning with multi-stream CNN-LSTM-HMs to discover sequential parallelism in sign language videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 9, pp. 2306–2320, Sep. 2020.
[8] H. Zhou, W. Zhou, Y. Zhou, and H. Li, “Spatial-temporal multi-cue network for continuous sign language recognition,” in Proc. AAAI Conf. Comput. Intell., vol. 34, Mar. 2020, pp. 13009–13016.
[9] W. H. Sumby and I. Pollack, “Visual contribution to speech intelligibility in noise,” J. Acoust. Soc. Amer., vol. 26, no. 2, pp. 212–215, Mar. 1954.
[10] E. D. Petajan, “Automatic lipreading to enhance speech recognition,” in Proc. Global Telecommun. Conf., 1984, pp. 265–272.
[11] A. J. Goldschen, O. N. Garcia, and E. Petajan, “Continuous optical automatic speech recognition by lipreading,” in Proc. 28th Asilomar Conf. Signals, Syst. Comput., 2002, pp. 572–577.
[12] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, “Continuous automatic speech recognition by lipreading,” in Motion-Based Recognition, M. Shah and R. Jain, Eds. Dordrecht, The Netherlands: Springer, 1997, pp. 321–343.
[13] G. Zhao, M. Pietikäinen, and A. Hadid, “Local spatiotemporal descriptors for visual recognition of spoken phrases,” in Proc. ACM Int. Multimedia Conf. Exhib., 2007, pp. 57–66.
[14] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in Proc. 28th Int. Conf. Mach. Learn. (ICML), 2011, pp. 689–696.
[15] H. Lee, C. Ekanadham, and A. Y. Ng, “Sparse deep belief net model for visual area V2,” in Proc. Adv. Neural Inf. Process. Syst., 2008, pp. 873–880.
[16] K. Noda, Y. Yamaguchi, K. Nakada, H. G. Okuno, and T. Ogata, “Lipreading using convolutional neural network,” in Proc. Conf. Int. speech Commun. Assoc., 2014, pp. 1149–1153.
[17] M. Wand, J. Koutnik, and J. Schmidhuber, “Lipreading with long short-term memory,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2016, pp. 6115–6119.
[18] J. S. Chung and A. Zisserman, “Lip reading in the wild,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 87–103.
A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemim, and J. Gubbi, “Lip reading using optical flow and support vector machines,” in Proc. 3rd Int. Congr. Image Signal Process., vol. 1, Oct. 2010, pp. 327–330.

L. Cappellotta and N. Harte, “Viseme definitions comparison for visual-only speech recognition,” in Proc. 19th Eur. Signal Process. Conf., Nov. 2011, pp. 2109–2113.

Z. Zhou, G. Zhao, and M. Pietikainen, “Towards a practical lipreading system,” in Proc. CVPR, Jun. 2011, pp. 137–144.

G. Zhao and M. Pietikainen, “Dynamic texture recognition using local binary patterns with an application to facial expressions,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, Jun. 2007.

A. Rekk, A. Ben-Hamadou, and W. Mahdi, “A new visual speech recognition approach for RGB-D cameras,” in Image Analysis and Recognition, S. Chum, S. Matas, and M. Sminchisescu, Eds., Cham, Switzerland: Springer, 2014, pp. 21–28.

X. Ma, Y. Lan, and Q. Zhong, “Lip feature extraction based on improved jumping-snake model,” in Proc. 35th Chin. Control Conf. (CCC), Jul. 2016, pp. 6928–6933.

J. Luetten and N. A. Thacker, “Speechreading using probabilistic models,” Comput. Vis. Image Understanding, vol. 65, no. 2, pp. 163–178, Feb. 1997.

Y. C. Cotes, G. J. Edwards, and C. J. Taylor, “Active appearance models,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.

T. Watanabe, K. Katsurada, and Y. Kanazawa, “Lip reading from multi view facial images using 3D-AAM,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 303–316.

H. L. Bear, S. J. Cox, and R. W. Harvey, “Speaker-independent machine lip-reading with speaker-dependent viseme classifiers,” 2017, arXiv:1710.01127.

H. L. Bear, R. W. Harvey, and Y. Lan, “Finding phonemes: Improving machine lip-reading,” 2017, arXiv:1710.01142. [Online]. Available: http://arxiv.org/abs/1710.01142

A. Biswas, P. K. Sahu, and M. Chandra, “Multiple camera in car audio-visual speech recognition using phonetic and viseme information,” Comput. Electr. Eng., vol. 47, pp. 35–50, Oct. 2015.

H. L. Bear and R. Harvey, “Decoding visemes: Improving machine lip-reading,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Mar. 2016, pp. 2009–2013.

D. Howell, S. Cox, and B. Theobald, “Visual units and confusion modelling for automatic lip-reading,” Image Vis. Comput., vol. 51, pp. 1–12, Jul. 2016.

H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: The good, the bad, and the ugly,” Speech Commun., vol. 95, pp. 40–67, Dec. 2017.

E. Petajan, B. Bischoff, D. Bodof, and N. Brooke, “An improved automatic lipreading system to enhance speech recognition,” in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Jan. 1988, pp. 19–25.

Y. Bennani and P. Gallinari, “On the use of TDNN-extracted features information in talker identification,” in Proc. Int. Conf. Acoust., Speech, Signal Process., vol. 1, 1991, pp. 385–388.

C. Bregler and Y. Konig, “Eigenlips’ for robust speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 2, Apr. 1994, pp. II696–II672.

P. Sujatha and M. R. Krishnan, “Lip feature extraction for visual speech recognition using hidden Markov model,” in Proc. Int. Conf. Comput. Commun., Amman, Apr. 2012, pp. 1–5.

K. Thangthai, H. L. Bear, and R. Harvey, “Comparing phonemes and visemes with DNN-based lipreading,” 2018, arXiv:1805.02924.

Y. Bennani, F. G. C. C. Dinis, and A. G. A. D. Pereira, “A new feature extraction method for visual speech recognition,” IEEE Trans. Multimedia, vol. 12, no. 1, pp. 40–46, Feb. 2010.

S. Pathan and A. Ghoktar, “Recognition of spoken English phrases using visual features extraction and classification,” Int. J. Comput. Sci. Inf. Technol., vol. 6, pp. 3716–3719, Jan. 2015.

A. A. Shaikh, D. K. Kumar, W. C. Yau, M. Z. C. Azemim, and J. Gubbi, “Lip reading using optical flow and support vector machines,” in Proc. 3rd Int. Congr. Image Signal Process., vol. 1, Oct. 2010, pp. 327–330.

H. L. Bear, S. J. Cox, and R. W. Harvey, “Speaker-independent machine lip-reading with speaker-dependent viseme classifiers,” 2017, arXiv:1710.01127.

H. L. Bear, R. W. Harvey, and Y. Lan, “Finding phonemes: Improving machine lip-reading,” 2017, arXiv:1710.01142. [Online]. Available: http://arxiv.org/abs/1710.01142

A. Biswas, P. K. Sahu, and M. Chandra, “Multiple camera in car audio-visual speech recognition using phonetic and viseme information,” Comput. Electr. Eng., vol. 47, pp. 35–50, Oct. 2015.

H. L. Bear and R. Harvey, “Decoding visemes: Improving machine lip-reading,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), Mar. 2016, pp. 2009–2013.

D. Howell, S. Cox, and B. Theobald, “Visual units and confusion modeling for automatic lip-reading,” Image Vis. Comput., vol. 51, pp. 1–12, Jul. 2016.

H. L. Bear and R. Harvey, “Phoneme-to-viseme mappings: The good, the bad, and the ugly,” Speech Commun., vol. 95, pp. 40–67, Dec. 2017.

E. Petajan, B. Bischoff, D. Bodof, and N. Brooke, “An improved automatic lipreading system to enhance speech recognition,” in Proc. SIGCHI Conf. Hum. Factors Comput. Syst., Jan. 1988, pp. 19–25.

Y. Bennani and P. Gallinari, “On the use of TDNN-extracted features information in talker identification,” in Proc. Int. Conf. Acoust., Speech, Signal Process., vol. 1, 1991, pp. 385–388.

C. Bregler and Y. Konig, “Eigenlips’ for robust speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), vol. 2, Apr. 1994, pp. II696–II672.

P. Sujatha and M. R. Krishnan, “Lip feature extraction for visual speech recognition using hidden Markov model,” in Proc. Int. Conf. Comput. Commun., Amman, Apr. 2012, pp. 1–5.

K. Thangthai, H. L. Bear, and R. Harvey, “Comparing phonemes and visemes with DNN-based lipreading,” 2018, arXiv:1805.02924.
A. Nasuha, F. Arifin, T. Sardjono, H. Takahashi, and M. H. Purnomo, “Automatic lip reading for daily Indonesian words based on frame difference and horizontal-vertical image projection,” J. Theor. Appl. Inf. Technol., vol. 95, pp. 393–402, Jan. 2017.

D. E. King, “Dlib-ml: A machine learning toolkit,” J. Mach. Learn. Res., vol. 13, pp. 2811–2814, Jul. 2012.

M. Wand and J. Schmidhuber, “Improving speaker-independent lipreading with domain-adversarial training,” 2017, arXiv:1708.01565. [Online]. Available: http://arxiv.org/abs/1708.01565

M. Wand, J. Schmidhuber, and N. T. Vu, “Investigations on end-to-end audiovisual fusion,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 3041–3045.

S. Petridis, Z. Li, and M. Pantic, “End-to-end visual speech recognition with LSTMs,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. 2017, pp. 2592–2596.

C. Sui, M. Bennamoun, and R. Togneri, “Listening with your eyes: Towards a practical visual speech recognition system using deep Boltzmann machines,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 154–162.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in Proc. 14th Int. Conf. Artif. Intell. Statist., 2011, pp. 315–323.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A simple way to prevent neural networks from overfitting,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 1929–1958, 2014.

A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2012, pp. 1097–1105.

A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), vol. 25, Jun. 2012, pp. 1097–1105.

D. Lee, J. Lee, and K.-E. Kim, “Multi-view automatic lip-reading using neural network,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 290–302.

T. Saitoh, Z. Zhou, G. Zhao, and M. Pietikäinen, “Concatenated frame image based cnn for visual speech recognition,” in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 277–289.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 1–9.

A. Nasuha, F. Arifin, T. Sardjono, H. Takahashi, and M. H. Purnomo, “Lip reading in the wild,” in Proc. Asian Conf. Comput. Vis. (ACCV), Oct. 2015, pp. 1543–1554.

J. Chung, A. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” 2014, arXiv:1412.3555. [Online]. Available: http://arxiv.org/abs/1412.3555

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks,” in Proc. 23rd Int. Conf. Mach. Learn. (ICML), 2006, pp. 369–376.

I. Fung and B. Mak, “End-to-end low-resource lip-reading with maxout CNN and LSTM,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 2511–2515.

A. Torfi, S. M. Iramanesh, N. Nasrabadi, and J. Dawson, “3D convolutional neural networks for cross-visual-audio matching recognition,” IEEE Access, vol. 5, pp. 22081–22091, 2017.

J. Chung, A. Senior, T. Stafylakis, M. Schlaman, and A. Zisserman, “Lip reading sentences in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3444–3453.

D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv:1409.0473. [Online]. Available: https://arxiv.org/abs/1409.0473

J. S. Chung and A. Zisserman, “Lip reading in profile,” in Proc. Brit. Mach. Vis. Conf., 2017, pp. 1–11.

K. Xu, D. Li, N. Casmantis, and X. Wang, “LCANet: End-to-end lipreading with cascaded attention-CTC,” in Proc. 13th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), May 2018, pp. 548–555.

R. K. Srivastava, K. Greff, and J. Schmidhuber, “Training very deep networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 2377–2385.

T. Stafylakis and G. Tzimiropoulos, “Combining residual networks with LSTMs for lipreading,” presented at the Interspeech, 2017.

T. Staflakas and G. Tzimiropoulos, “Deep word embeddings for visual speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 4974–4978.

S. Petridis, T. Staflakas, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, “End-to-end audiovisual speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 6548–6552.

D. Kumar Margam, R. Avalakatti, T. Sharma, A. Thanda, P. A. K. Roy, and M. Venkatesh, “End-to-end lip reading with 3D-2D-CNN BLSTM-HMM and word-CTC models,” 2019, arXiv:1906.12170. [Online]. Available: http://arxiv.org/abs/1906.12170

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), Mar. 2020, pp. 836–842.

X. Zhao, S. Yang, S. Shan, and X. Chen, “Mutual information maximization for effective lip reading,” in Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), Mar. 2020, pp. 843–850.

M. Luo, S. Yang, S. Shan, and X. Chen, “Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading,” in Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), Mar. 2020, pp. 69–76.

Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, “Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition,” in Proc. 15th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), May 2020, pp. 851–858.
G. Papandreou, A. Katsamanis, V. Pitkäsaari, and P. Maragos, “Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition,” IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 3, pp. 423–435, Mar. 2009.

A. Fernandez-Lopez and F. M. Sukno, “Automatic viseme vocabulary completion to enhance continuous lip-reading,” 2017, arXiv:1704.08035. [Online]. Available: http://arxiv.org/abs/1704.08035

Y. Xu, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T. S. Huang, “Lipreading by locality discriminant graph,” in Proc. IEEE Int. Conf. Image Process., vol. 3, Sep./Oct. 2007, pp. III-325–III-328.

J. Huang, G. Potamianos, J. Connell, and C. Neti, “Audio-visual speech recognition using an infrared headset,” Speech Commun., vol. 44, nos. 1–4, pp. 83–96, Oct. 2004.

P. J. Lucey, G. Potamianos, and S. Sridharan, “Patch-based analysis of visual speech from multiple views,” in Proc. Int. Conf. Auditory-Vis. Speech Process. (AVSP), 2008, pp. 69–73.

P. Lucey, G. Potamianos, and S. Sridharan, “A unified approach to multi- pose audio-visual ASR,” in Proc. 8th Annua. Conf. Int. Speech Commun. Assoc. (Interspeech), Antwerp, Belgium, 2007, pp. 650–653.

S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, S. Matsuda, T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura, “CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition,” in Proc. AVSP, 2010, pp. 85–88.

A. Pass, J. Zhang, and D. Stewart, “AN investigation into features for multi-view lipreading,” in Proc. IEEE Int. Conf. Image Process., Sep. 2010, pp. 2417–2420.

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. Berlin, Germany: Springer, 2010, pp. 259–266.

A. Chiù and L. J. Rothkrantz, “Automatic visual speech recognition,” in Proc. Speech Enhancement, Modeling Recognit. Algorithms Appl., 2012, p. 95.

V. Estellers and J.-P. Thiran, “Multipose audio-visual speech recognition,” in Proc. 19th Eur. Signal Process. Conf., 2011, pp. 1065–1069.

M. Igras, B. Ziółko, and T. Jadczyk, “Audiovisual database of Polish speech recordings,” Studia Inf., vol. 33, no. 2B, pp. 163–172, 2012.

D. Estival et al., “Austalk: An audio-visual corpus of Australian English,” in Proc. 9th Lang. Resour. Eval. Conf. (LREC), 2014, pp. 3105–3109.

I. Anina, 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.

X. Yanjun, D. limin, L. guoqiang, Z. zhi, “Chinese audiovisual bimodal speech database CAVSR1.0,” Acta Acustica, vol. 25, no. 1, pp. 42–49, 2000.

C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, “Audio-visual speech recognition,” Center Lang. Speech Process., The Johns Hopkins Univ., Baltimore, MD, USA, Final Workshop Rep., Oct. 2000.

C. Sanderson, “The VidTIMIT database,” IDIAP Commun., Martigny, Switzerland, Tech. Rep., 2002.

T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, “A segment-based audio-visual speech recognition: Data collection, development, and initial experiments,” in Proc. 6th Int. Conf. Multimodal Interfaces (ICMI), 2004, pp. 235–242.

G. Zhao, M. Barnard, and M. Pietikäinen, “Lipreading with local spatiotemporal descriptors,” IEEE Trans. Multimedia, vol. 11, no. 7, pp. 1254–1265, Nov. 2009.

Y. Lan, B. Theobald, R. Harvey, E. Ong, and R. Bowden, “Improving visual features for lip-reading,” in Proc. Int. Conf. Auditory-Vis. Speech Process. (AVSP), 2010, pp. 142–147.

C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernock, N. Poh, J. Kittler, A. Larcher, C. Levy, D. Matrouf, J.-F. Bonastre, P. Tresadern, and T. Cootes, “Bi-modal person recognition on a mobile phone: Using mobile phone data,” in Proc. Int. Conf. Auditory-Vis. Speech Process. Expo Workshops, Jul. 2012, pp. 635–640.

Y. Mroueh, E. Marcheret, and V. Goel, “Deep multimodal learning for audio-visual speech recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2015, pp. 2130–2134.

D. L. Howell, “Confusion modelling for lip-reading,” Ph.D. dissertation, Univ. East Anglia, Norwich, UK, 2014.
MINGFENG HAO was born in Yangzhou, Jiangsu, China, in 1992. He received the bachelor’s degree in electronic information engineering from the School of Electronic and Information Engineering, Changshu Institute of Technology, China, in 2015. He is currently pursuing the M.S. degree with the Department of Information Science and Engineering, Xinjiang University, China. His research interests include biometrics, image processing, and pattern recognition.

MUTALLIP MAMUT was born in Kuerle, Xinjiang, China, in 1975. He received the B.S. and M.S. degrees in computer application technology from Xinjiang University, China, in 1998 and 2009, respectively. He is currently an Engineer at the Library, Xinjiang University. He has published more than 20 articles. His research interests include biometrics, image processing, and pattern recognition. He is a member of the China Computer Federation (CCF).

NURBIYA YADIKAR was born in Unumqi, Xinjiang, China, in 1970. She received the B.S. degree in electronic information technology from Xinjiang University, China, in 1993, and the M.S. degree in advanced cyber technology from the Kyoto Institute of Technology, Japan, in 2004. She is currently an Assistant Professor (Lecturer) with the School of Information Science and Engineering, Xinjiang University. She has authored more than 30 articles. Her research interests include image processing, pattern recognition, and signal processing. She is a member of the Pattern Recognition Committee for China Artificial Intelligence Association (CAAI-PR), and the China Computer Federation (CCF).

ALIMJIAN AYSA was born in Kashi, Xinjiang, China, in 1973. He received the B.S. degree from Sichuan University, in 1997, and the M.S. and Ph.D. degrees from Xinjiang University, Xinjiang, in 2008 and 2013, respectively, all in computer application technology. He is currently a Professor with the Network and Information Technology Center, Xinjiang University. He has also been as a Researcher with the Xinjiang Laboratory of Multi-language Information Technology, since 2002. He has authored one book and more than 60 articles. His research interests include natural language processing, pattern recognition, and digital signal processing. He is a member of the IEEE Computer Society and the China Computer Federation.

KURBAN UBUL (Member, IEEE) was born in Bachu, Xinjiang, China, in 1974. He received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, in 1997, and the M.S. degree in signal and information processing and the Ph.D. degree in computer application technology from Xinjiang University, Xinjiang, in 2009 and 2018, respectively. He has studied as a Visiting Scholar with the Department of Systems and Computer Engineering, Carleton University, Canada, from 2011 to 2012. He has been a Researcher with the Xinjiang Laboratory of Multi-Language Information Technology, since 2000. He is currently a Professor with the School of Information Science and Engineering, Xinjiang University. He has authored three books, one book chapter, and more than 110 articles, presented at numerous international journals and conferences. His research interests include image processing, pattern recognition, speech signal processing, digital signal processing, and education technology. He is a member of the International Association of Pattern Recognition, the IEEE Computer Society, the Association for Computing Machinery, and the China Computer Federation. He is the General Chair of NLPAI2020, the TPC Chair of CCBR2018, and the Area Chair of PRCV2019. He has served as a Technical Committee Member/Reviewer of many international conferences such as the ICPR, ICadar, CCFAI, CCDM, and CCBR. He is a Committee Member of the Pattern Recognition Professional Committee for China Artificial Intelligence Association and the Pattern Recognition and Machine Intelligence Professional Committee for China Automation Association. He currently serves as an Editor for the Singapore Journal of Scientific Research, the Cyber Journals, the Image and Signal Processing, and the Journal of Security and Safety Technology. He was a Reviewer of Neurocomputing, IEEE Intelligent Hospital Management System, the IET Biometrics, and the IJITM.