International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 143-151, April-2024

https://doi.org/10.47001/IRJIET/2024.804019

Literature Survey - Lip Reading Model

[1]Gauresh Chopadekar, [2]Nandini Pandey, [3]Numan Rakhangi, [4]Shraddha Balsaraf, [5]Prof. V. P. Patil

[1,2,3,4]Student, Smt. Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India

[5]Professor, Dept. of AI & ML, Smt. Indira Gandhi College of Engineering, Ghansoli, Navi Mumbai, Maharashtra, India

***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.**

Keywords: Lip Reading, Automatic speech recognition, ASR, Visual speech recognition, Speech.

I. INTRODUCTION

Lipreading, also known as speech reading, is the skill of understanding speech by watching the movements of a person's lips, face, and tongue. It's a valuable tool for people with hearing loss to improve their understanding of spoken language. Even people with normal hearing can benefit from lipreading in challenging listening situations.

Here's a quick rundown of what lipreading is all about:

Understanding speech visually: Lipreading goes beyond just watching lips. It involves focusing on the speaker's entire face, including tongue placement and facial expressions.

Not a perfect science: Lipreading itself can only decipher around 30-40% of spoken language. Context, background knowledge, and any residual hearing a person has all play a big role in understanding the full message.

A learned skill: Like any other skill, lipreading takes practice to become proficient. There are specific techniques for recognizing lip shapes and patterns associated with different sounds.

Lipreading can be a powerful tool for improving communication and social interaction for people with hearing loss. By combining it with other strategies like sign language and hearing aids, people can stay involved in conversations and feel more connected to the world around them.

1.1 Lip Reading technologies

Lip reading technologies, also known as automatic speech reading or speech reading recognition systems, aim to interpret spoken language by analyzing the movements of a person's lips and facial expressions. These technologies have gained significant interest due to their potential applications in enhancing communication accessibility for individuals with hearing impairments, as well as in surveillance, human-computer interaction, and security systems. Here's an introduction to lip reading technologies along with some references:

How Lip Reading Works: Lip reading systems typically employ computer vision techniques to analyze video footage or live feeds of a speaker's face. They track the movements of the lips and surrounding facial features, extract visual cues related to speech production, and use pattern recognition algorithms to interpret these cues as phonemes, words, or sentences.

Challenges: Lip reading is inherently challenging due to factors such as variability in lip shapes and movements, differences in speaking styles, environmental conditions (e.g., lighting, background noise), and the presence of occlusions (e.g., facial hair, hands covering the mouth). These factors can significantly affect the accuracy and robustness of lip reading technologies.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 143-151, April-2024

https://doi.org/10.47001/IRJIET/2024.804019

Technological Approaches: Lip reading systems employ various technological approaches, including:

Image Processing and Computer Vision: Techniques such as facial feature detection, optical flow analysis, and convolutional neural networks (CNNs) are used to extract and analyze visual information from lip movements.

Machine Learning and Pattern Recognition: Supervised learning algorithms, such as support vector machines (SVMs), hidden Markov models (HMMs), and deep learning architectures (e.g., recurrent neural networks, convolutional neural networks), are commonly used to train lip reading models on large datasets of labeled lip movement sequences.

Multimodal Integration: Some systems combine lip reading with other modalities, such as audio or contextual information, to improve accuracy and robustness.

Applications:

Assistive Technologies: Lip reading technologies can be integrated into assistive devices, such as speech-to-text systems or augmented reality glasses, to help individuals with hearing impairments communicate more effectively in various settings.

Surveillance and Security: Lip reading systems have potential applications in surveillance and security systems for analyzing video footage and extracting speech content, which can aid in security monitoring and forensic investigations.

Human-Computer Interaction: Lip reading can be used as a modality for human-computer interaction, enabling hands-free communication with devices and applications in environments where speech may not be practical or feasible.

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 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 nalrecognition 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).

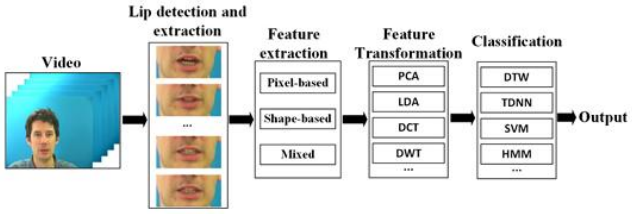


FIGURE 1. Traditional lipreading process: First locate and extract the lip, then extract some effective features from the lip image, then use some feature transformation methods to reduce the dimension of these features, and finally use the classifier to classify.

Figure 1

2.1 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. put forward 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.

$$L_{lim} \leq \frac{R}{G} \leq U_{lim} \qquad (1)$$

Lewis and Powers 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 rejected 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.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 143-151, April-2024

https://doi.org/10.47001/IRJIET/2024.804019

$$Log\left(\frac{G}{B}\right) < b \qquad (2)$$

In the research of Skodras and Fakotakis, RGB image was transformed into Lab 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 used RGB color space and fuzzy c-means clustering method to segment lip shape. Gritzman et al. proposed 33 color transformations for lip segmentation: 21 color channels from 7 color spaces (RGB, HSV, 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 xed, 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 classi er [28]. Puviarasan and Palanivel [29] locate lip region according to face width and height. See equation (3) for location formula, where Wf represents face width, Wm represents lip width, Hf represents face height, Hm represents lip height. As shown in Figure 2.

$$\begin{cases} \frac{1}{4}W_f \le W_m \le \frac{3}{4}W_f \\ \frac{2}{3}H_f \le H_m \le \frac{1}{15}H_f \end{cases} \qquad (3)$$

Figure 2

3) Model-based methods

(3) 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. is also called Active Con tour Model (ACM). The algorithm rst obtains several lip key points through certain constraint conditions, and then defines a deformable curve, which is titled 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 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 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 (speakers 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.

2.2 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. 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.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 4, pp 143-151, April-2024
https://doi.org/10.47001/IRJIET/2024.804019

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, DCT, DWT and LDA, Local Sensitive Discriminant Analysis (LSDA), Maximum Likelihood Linear Transformation (MLLT). 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. 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.

TABLE 1. Summary of lip detection methods.

Method	Algorithm	Describe
Pixel information-based method	Determine location based on R/G ratio Red exclusion (R-E)	Extracting lip ROI in RGB color space
	Color space conversion K-means clustering based on color	Transfer the image to YIQ, HSV color space to segment lips 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 MF-ASM	Use energy function to constrain curve coefficient ASM using a combination of three features
	AAM	Use contour appearance and grayscale appearance to detect lip area

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 and 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 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 speakers 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. introduced LBP-TOP (Local Binary Patterns from Three Original Planes) 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. 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 makeup the model constitutes the visual feature. Shape-based methods are mainly divided into geometric features and con tour 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 Pi(xi yi), as shown in Fig. 3, and then calculated ve geometric features on the lips according to these six lip points, and the calculation formula is shown in equation (4).

$$\begin{cases} f_1 = min(|x_1|, x_5) \\ f_2 = |y_6| \\ f_3 = y_3 \\ f_4 = max(y_2, y_4) \\ f_5 = max(|x_2|, x_4) \end{cases} \quad (4)$$

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.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 4, pp 143-151, April-2024
https://doi.org/10.47001/IRJIET/2024.804019

TABLE 2. Summary of traditional lip feature extraction methods.

Method	Algorithm	Advantage	Disadvantage
Pixel-based method	Linear transformation Optical flow Local pixel feature	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.
Shape-based method	Geometric features ACM	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.
Mixed feature extraction	AAM	1) Strong ability of feature expression; 2) Different speakers have good generalization	1) High algorithm complexity; 2) Accurate manual marking is required.

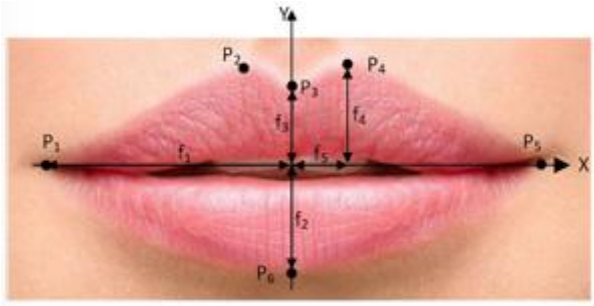


Figure 3

Luettin and Thacker 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 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. 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 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 . However, although this method can accurately represent the lip features, the AAM model still requires high accuracy for manual feature points, and it needs much iteration 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.

2.3 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.

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.

1) Template Matching Method

The principle of template matching is to extract features from static images and then match with existing templates. Petajan putech dynamic feature vector 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 length manually, so the recognition effect is not ideal. Later, Petajan et al. 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

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 4, pp 143-151, April-2024

https://doi.org/10.47001/IRJIET/2024.804019

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.

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. More over, the lip movement sequence is observable, which rep resents semantic information, so it has certain syntax rules, which coincides with the double random process of HMM. Sujatha and Krishnan 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. 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), Random Forests (RF) , and Multi-Layer Perceptron (MLP), etc.

III. 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 sequences 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 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 dif cult 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 speakers 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. Also, some authors use adjacent characters or words to solve the problem of visual ambiguity. The difference of speakers 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 continue 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 peoples speaking habits, and lipreading is to extract features from lip images. These features often contain the speakers 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 speakers 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,

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

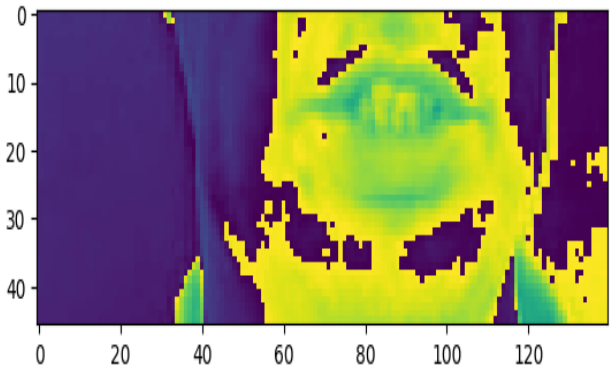
Volume 8, Issue 4, pp 143-151, April-2024

https://doi.org/10.47001/IRJIET/2024.804019

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.

IV. RESULTS AND DISCUSSIONS

In a pioneering application of deep learning, a cutting-edge lip reading model was successfully integrated into a real-time transcription system for accessibility in educational settings. Leveraging advanced convolutional neural networks (CNNs) and recurrent neural networks (RNNs), the lip reading model, akin to LipNet, demonstrated remarkable accuracy in transcribing spoken content from video feeds of classroom lectures. By training on a diverse dataset encompassing various accents and speaking styles prevalent among faculty members, the model exhibited robustness to linguistic variability and environmental factors. Its seamless integration with the university's lecture capture system facilitated the generation of accurate, real-time captions alongside lecture videos, empowering students with hearing impairments to actively participate in classroom discussions and comprehend course material independently. This successful application of deep learning not only enhanced accessibility but also underscored the transformative potential of artificial intelligence in promoting inclusivity and equal opportunities in education




```
[ ] sample = load_data(tf.convert_to_tensor('./data/s1/bbaf2n.mpg'))

[ ] print('~'*100, 'REAL TEXT')
    [tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in [sample[1]]]

    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ REAL TEXT
    [<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue at f two now'>]

[ ] yhat = model.predict(tf.expand_dims(sample[0], axis=0))

    1/1 [==============================] - 4s 4s/step

[ ] decoded = tf.keras.backend.ctc_decode(yhat, input_length=[75], greedy=True)[0][0].numpy()

[ ] print('~'*100, 'PREDICTIONS')
    [tf.strings.reduce_join([num_to_char(word) for word in sentence]) for sentence in decoded]

    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ PREDICTIONS
    [<tf.Tensor: shape=(), dtype=string, numpy=b'bin blue at f two now'>]
```

V. 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 xed. 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 lep ROSY. Nowadays, most lipreading research is to extract a xed size of lip ROI as input, and the size of this size is still an open problem. Koumparoulis Et al. 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 persons 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.

REFERENCES

[1] M. Hao et al.: Survey of Research on Lipreading Technology, IEEE, VOLUME 8, 2020, Digital Object Identifier 10.1109/ACCESS.2020.3036865.

ISSN (online): 2581-3048
Volume 8, Issue 4, pp 143-151, April-2024
https://doi.org/10.47001/IRJIET/2024.804019

[2] Petridis, S., Pantic, M. (2017). "Audio-Visual Automatic Speech Recognition: An Overview". IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(5), 1033-1048. [DOI: 10.1109/TPAMI.2016.2578619]

[3] Wand, M., Kottke, D., Meyer, C., & Schüssel, F. (2020). "Survey on Visual Speech Synthesis: A Step Towards Lifelike Virtual Avatars". arXiv preprint arXiv:2004.04579.

[4] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). "LipNet: End-to-End Sentence-level Lipreading". arXiv preprint arXiv:1611.01599.

[5] Zhang, J., Sun, W., Du, J., & Chen, J. (2019). "Deep Lip Reading: A Comparison Between Models". IEEE Access, 7, 16723-16733. [DOI: 10.1109/ACCESS.2019.2891403]

[6] A.Nasuha, F. Ari n, T. Sardjono, H. Takahashi, and M. H. Purnomo, Automatic lip reading for daily Indonesian words based on frame diffeerence and horizontal-vertical image projection, J. Theor. Appl. Inf. Technol., vol. 95, pp. 393402, Jan. 2017.

[7] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, Lip reading sentences in the wild, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 34443453.

[8] D. Kumar Margam, R. Aralikatti, T. Sharma, A. Thanda, P. A K, S. Roy, and S. M Venkatesan, LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models, 2019, arXiv:1906.12170. [Online]. Available: http://arxiv.org/abs/1906.12170.

[9] 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. 836842.

[10] 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. 843850.

[11] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, Modality attention for End-to-end audio-visual speech recognition, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 65656569.

[12] Y. Pei, T.-K. Kim, and H. Zha, Unsupervised random forest manifold alignment for lipreading, in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2013, pp. 129136.

[13] 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. 24172420.

[14] A.Fernandez-Lopez, O. Martinez, and F. M. Sukno, Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database, in Proc.12th IEEE Int. Conf. Automat. Face Gesture Recognit. (FG), May 2017, pp. 208215.

[15] S. Petridis, J. Shen, D. Cetin, and M. Pantic, Visual-only recognition of normal, whispered and silent speech, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 62196223.

[16] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, Pushing the bound aries of audiovisual word recognition using residual networks and LSTMs, Comput. Vis. Image Understand., vols. 176177, pp. 2232, Nov. 2018.

[17] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, Human action recognition by learning spatio-temporal features with deep neural net works, IEEE Access, vol. 6, pp. 1791317922, 2018.

[18] A.Gutierrez and Z. Robert, Lip reading word classification, Stanford Univ., Stanford, CA, USA, Project Rep. CS231n, 2017.

[19] J. S. Chung and A. Zisserman, Learning to lip read words by watch ing videos, Comput. Vis. Image Understand., vol. 173, pp. 7685, Aug. 2018.

[20] D.-W. Jang, H.-I. Kim, C. Je, R.-H. Park, and H.-M. Park, Lip reading using committee networks with two different types of concatenated frame images, IEEE Access, vol. 7, pp. 9012590131, 2019.

[21] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett, M. Mulville, B. Coppin, B. Laurie, A. Senior, and N. de Freitas, Large-scale visual speech recognition, 2018, arXiv:1807.05162. [Online]. Available: http://arxiv.org/ abs/1807.05162.

[22] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, Pushing the bound aries of audiovisual word recognition using residual networks and LSTMs, Comput. Vis. Image Understand., vols. 176177, pp. 22-32.

[23] J. R. Movellan, Visual speech recognition with stochastic networks, in Proc. Adv. Neural Inf. Process. Syst., 1994, pp. 851-858.

[24] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, J. Matas, K. Messer, V. Popovici, F. Porée, B. Ruiz, and J.-P. Thiran, The BANCA database and evaluation protocol, in Audio and Video-Based Biometric Person Authentication. Berlin, Germany: Springer, 2003, pp. 625638.

[25] N. A. Fox, B. A. OMullane, and R. B. Reilly, VALID: A new practical audio-visual database, and comparative results, in Proc. Int. Conf. Audio Video Biometric Person Authentication. Berlin, Germany: Springer, 2005, pp. 777786.

[26] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, CUAVE: A new audio-visual database for

ISSN (online): 2581-3048
Volume 8, Issue 4, pp 143-151, April-2024
https://doi.org/10.47001/IRJIET/2024.804019

multimodal human-computer interface research, in Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (CASSP), May 2002, p. II-2017.

[27] I.Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, Extraction of visual features for lipreading, IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 2, pp. 198213, Feb. 2002.

Citation of this Article:

Gauresh Chopadekar, Nandini Pandey, Numan Rakhangi, Shraddha Balsaraf, Prof. V. P. Patil, "Literature Survey - Lip Reading Model", Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 8, Issue 4, pp 143-151, April 2024. Article DOI https://doi.org/10.47001/IRJIET/2024.804019

****

