Deep Learning Techniques for Enhancing Lip-Reading and Speech-Reading: A Survey

Akshay N Patel
Student(GTU-GSET)
akpatel3290@gmail.com

Prof. S.K. Hadia
Associate Professor(GTU-GSET)
asso_s_k_hadia@gtu.edu.in

Abstract:

In our noisy world, understanding spoken words can be challenging. This survey explores how deep learning can unlock the language of lips, making it accessible to a broader audience. Aimed not only at technical enthusiasts but also at those wishing to decipher whispered conversations, it delves into the applications of deep learning for lip-reading and speech-reading. The focus is on practical uses, such as aiding individuals with hearing difficulties and improving communication in noisy settings like streets or classrooms. This survey provides a user-friendly journey through the advancements in this field, empowering researchers and developers to utilize deep learning. By making silent voices heard, it envisions a future where understanding conversations, even in the middle of noise, becomes a reality.

Keywords: Lip-reading, Hearing impairment, Silent communication, Feature extraction, Multimodal fusion, Deep learning

1. Introduction:

Lip reading, also known as speech reading, is a fascinating and challenging domain that intersects both computer vision and natural language processing, encompassing various disciplines such as pattern recognition, speech processing, image classification, and natural language processing. The ability to understand spoken language by visually interpreting the movements and shapes of the lips has significant implications for various applications, ranging from assistive technologies for the hearing-impaired to human-computer interaction and security systems. Over the years, the field of lip reading has witnessed a surge in research and development, driven by advancements in deep learning, computer vision, and the availability of large-scale datasets. This survey paper aims to provide a comprehensive overview of the current state-of-the-art in lip reading techniques and methodologies. By examining the evolution of lip reading approaches, we aim to shed light on the challenges, breakthroughs, and potential avenues for future research in this dynamic field.

Face recognition is an essential component of lip reading, thus highlighting the importance of understanding the various challenges and approaches associated with it [5]. In recent years, despite advancements, automatic speech recognition systems (ASRs) continue to struggle in real-world noise, driving research towards noise-robust technologies. However,

visual information remains unaffected by acoustic disturbances, rendering automatic lip-reading indispensable in noisy environments and enhancing ASR performance, especially in challenging acoustic conditions. Visual language information is vital for speech recognition, particularly in scenarios where audio is corrupted or unavailable [1]. However, practical lip-reading recognition encounters significant challenges due to the diversity and complexity of daily scenarios.

The rapid development of lip reading in recent years has led to the publication of numerous excellent articles and relevant review articles, highlighting the latest advancements in lip reading architecture and datasets [6,7]. This paper focuses on visual feature extraction in lip reading, emphasizing recent developments in the field.

Artificial intelligence is deeply integrated into various fields, with Multimodal Lip Reading (MLR) leveraging a range of technologies, exhibiting strong practicality and broad application. For instance, lip reading technology enhances speech recognition, specifically Multimodal Audio-Visual Speech Recognition [4], to improve accuracy in noisy environments. Additionally, lip reading significantly enhances communication methods for the hearing impaired [4], providing direct access to spoken language without reliance on special equipment or sign language interpreters. This empowerment enables individuals to communicate independently across diverse situations.

The survey is structured as follows: delving into the various methodologies employed, including traditional techniques and the recent surge of deep learning approaches. The paper also reviews existing datasets and evaluates the performance metrics commonly used to assess lip reading systems. Furthermore, we discuss the challenges posed by real-world conditions such as noise, lighting variations, and diverse speaker characteristics.

As we navigate through the literature, it becomes evident that lip reading has evolved beyond its traditional confines. It has emerged as a multidisciplinary research area, incorporating insights from linguistics, psychology, and neuroscience.

In summarizing the existing body of knowledge, this survey aims to serve as a valuable resource for researchers, practitioners, and enthusiasts interested in the expansive domain of lip reading. By understanding the current landscape and identifying gaps in the literature, we hope to inspire future research endeavors that push the boundaries of what is achievable in lip reading technology.

Datasets:

In this section, we have examined the notable datasets utilized by researchers in lip reading, along with their respective properties.

Lrs2 BBC [8]

In 2018, the Oxford-BBC Lip Reading Sentences 2 (LRS2-BBC) dataset emerged as a rich source for advancing lip reading and automatic speech recognition (ASR) technologies. Comprising numerous hours of carefully curated dialogue excerpts from BBC programs, LRS2-BBC offers a unique glimpse into spoken language in the wild. LRS2-BBC is thoughtfully divided into development and test sets. The development set, further encompassing training and validation subsets totaling 17,660 words, facilitates robust algorithm training and refinement. The separate test set of 1,698 words provides an objective platform for evaluation.

Turkish Lip Reading [2]

Turkish lip-reading created two new datasets for scientific research: one with 111 words and

the other with 113 sentences. The word dataset consists of 111 words, and the sentence dataset consists of 113 sentences, all of which are in different lengths and selected from those used in daily life. The video frames used to create the datasets were obtained under the same ambient and light conditions, with each speaker located at 1.5 meters in these images. The properties of the lip-reading datasets are detailed in the study.

The CELR-200 [3]

This dataset is introduced as the first cross-language word-level lipreading dataset. It aims to address the challenge of insufficient generalization ability of lipreading systems in multilingual scenarios. The dataset is labeled at two levels, language, and word level, to comprehensively train and evaluate each lipreading system. This dataset is a comprehensive collection of lip-reading data for training and evaluating deep learning models. With 83,788 video samples across 200-word classes, captured in real-world scenarios at high resolution (96 x 96 pixels), it's an extensive resource. Videos are categorized into language (Chinese and English) and word classes, making it suitable for multilingual lip reading and specific vocabulary recognition studies. This dataset is invaluable for researchers and developers in automatic lip reading and multimodal speech recognition.

LibriSpeech [9]

LibriSpeech is a corpus comprising approximately 1000 hours of 16kHz read English speech, developed by Vassil Panayotov with assistance from Daniel Povey. The dataset originates from audiobooks in the LibriVox project, systematically segmented and aligned for research purposes.

GRID [10]

GRID is a widely used sentence-level dataset for the lip-reading task, comprising 34 speakers, each delivering 1000 sentences, resulting in approximately 34,000 sentence-level videos. All videos in GRID feature a fixed, clean, single-colored background, with speakers instructed to face the camera frontally during the speaking process.

2. Literature Survey:

The ability to decipher unspoken words from lip movements has captivated humanity for centuries. Through an analysis of recent advancements, this review aims to shed light on the current state-of-the-art and future directions in lip reading technology.

Here are several challenges encountered in the practice of lip reading.

Visual ambiguity in lip reading can be caused by a variety of factors, including variations in lip motions among speakers, similarities between visemes, variations in lip movements among speakers, and the influence of lighting and viewing angles etc.

Language-related challenges in lip reading include struggling to recognize silent or non-speech sounds, the difficulties to interpret contextual and grammatical cues, and the effect of co-articulation on lip movement analysis.

Technical constraints include things like the lack of labeled datasets for lip reading model training, the high computational demands associated with deep learning approaches.

Ethical Considerations includes addressing privacy concerns related to data collection and ensuring accessibility for users with disabilities or cultural differences.

Now, we look at the methodologies used by different studies in the field of lip reading to investigate and address the issues.

Methodology:

A. Data Preprocessing

Before the lip detection and extraction, data preprocessing plays an important role. Data preprocessing in lip reading involves transforming raw video data to enhance the accuracy of subsequent processing. This includes noise reduction, image enhancement, face detection and tracking, lip localization, feature extraction, normalization, and data augmentation. These steps ensure clearer lip movements for accurate speech recognition.

Figure 1 displays the primary steps of the traditional lip reading process, while Figure 2 illustrates the deep learning-based approach to lip reading.

Fig.1 Traditional lip reading Architecture

B. Lip detection and extraction:

This is the first step, it isolating the lip area from the video. This is crucial because lipreading relies solely on visual lip information to recognize speech. The accuracy of this extraction directly impacts recognition performance. Lip detection techniques utilize methods based on color analysis, facial structure assessment, and model-driven approaches. In this process, extracting the region of interest (ROI) from raw data is essential. MediaPipe, dlib, and OpenCV, can be used for ROI (Region of Interest) detection

C. Feature Extraction:

Feature extraction in lip reading involves extracting relevant information from the lip region of video frames to represent the visual cues essential for recognizing speech content. Various features are extracted to capture important characteristics of lip movements, including shape, texture, and motion. These features serve as input to machine learning algorithms or neural networks, enabling the system to effectively analyze and interpret lip movements and infer spoken content. Common feature extraction techniques in lip reading include histogram of oriented gradients (HOG), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) etc.

D. Feature Transformation:

Feature transformation in lip reading involves modifying extracted features to improve their effectiveness for subsequent tasks. Techniques such as PCA, LDA, and autoencoders are used to enhance feature representation.

E. Classification:

Classification in lip reading refers to the process of categorizing or identifying spoken words or phrases based on the features extracted from lip movements. This step involves training a machine learning model or neural network to recognize patterns in the extracted features and associate them with specific words or phonemes. Common classification algorithms used in lip reading include support vector machines (SVM), k-nearest neighbors (KNN), decision trees, and deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). By accurately classifying lip movements, the system can infer the spoken content and facilitate communication for individuals with hearing impairments or in noisy environments.

Deep Learning Techniques:

Traditional lip reading relies on manual feature extraction and rule-based methods to interpret lip movements. On the other hand, deep learning architectures use neural networks to automatically learn features from raw data, potentially leading to more robust and accurate lip reading performance without the need for explicit feature engineering. Here given in figure 2, the working flow of deep learning lip reading process.

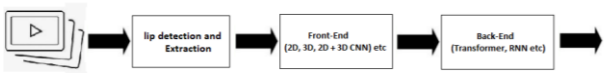

Fig. 2 Deep learning lip reading Architecture

2D CNN:

CNNs, or Convolutional Neural Networks, are a type of deep learning architecture inspired by the biological structure of the visual cortex. They excel at tasks like image recognition and classification due to their ability to automatically extract features from visual data.

In traditional lip reading, manual feature extraction involves geometric, motion, and texture features, followed by classification with SVMs or HMMs. Conversely, 2D CNNs automatically extract features from raw video frames, learning patterns indicative of lip movements through multiple convolutional layers. These networks achieve higher accuracy without the need for manual feature engineering, providing robustness to variations in lighting and speaker appearance. They also offer adaptability, allowing integration with other modalities for multimodal lip reading. 2D CNNs represent a powerful advancement in lip reading technology, promising improved accuracy and accessibility in communication interfaces.

3D CNN:

Traditional 2D Convolutional Neural Networks (CNNs) analyze individual video frames, capturing spatial information about lip shapes and movements, while 3D CNNs process sequences of video frames as 3D volumes, crucial for understanding speech. In lip reading, 3D CNNs employ convolutional filters that operate on the 3D volume of video frames, extracting spatiotemporal features capturing changes over time, such as lip opening/closing patterns or subtle transitions between phonemes. These networks utilize multiple layers to combine and refine information, extracting increasingly complex spatiotemporal features relevant for speech recognition, with final layers typically employing fully connected layers to recognize spoken words or phonemes based on the extracted features. Advantages of 3D CNNs over 2D CNNs include enhanced performance, robustness, and a more natural representation of lip movements, making them particularly effective for speech recognition tasks. Additionally, examples of 3D CNN architectures for lip reading include stacked 2D convolutions and attention mechanisms focusing on informative parts of the 3D volume.

2D + 3D CNN:

Combining 2D and 3D CNNs in lip reading presents a versatile approach to extract both spatial and temporal features from video data. Hybrid architectures blend both types of convolutional layers within a single network, while a staged approach involves using 2D CNNs for spatial analysis followed by 3D CNNs for temporal modeling. This integration capitalizes on the complementary strengths of each approach, leading to superior accuracy and robustness in speech recognition tasks. Examples of combined architectures include models that incorporate 2D convolutions with

3D recurrent layers or attention mechanisms to focus on informative regions of the video. Overall, leveraging both 2D and 3D CNNs offers a promising avenue for advancing lip reading systems, providing flexibility and improved performance in various applications. When high accuracy and robustness are crucial, such as in applications for people with hearing loss or in noisy environments, then it helps.

LSTM:

LSTMs, a type of RNN, excel at lip reading thanks to their ability to handle sequence data. Unlike regular neural networks that focus on individual frames, LSTMs "remember" past lip movements, crucial for understanding the flow of speech. They analyze extracted features like lip corners and opening patterns over time, capturing the subtle transitions between sounds and ultimately "decoding" the spoken word, like deciphering a silent movie with subtitles you can actually hear. This temporal understanding surpasses even simpler RNNs, leading to impressive accuracy in recognizing words and sentences from silent lip movements.

Here, we have explored the findings and observation of different research work (Table I).

Table 1: Comparative analysis of lip reading research work

	Paper	Techniques	Findings	Observation
	Lip Reading Sentences Using Deep Learning With Only Visual Cues[1]	VISEME CLASSIFIER + spatial-temporal (3D) convolution + 2D ResNet	System has achieved a significantly improved performance with 15% lower word error rate.	Classification accuracy of visemes achieved by the proposed system was very high (over 95%). Classification accuracy of word was significantly dropped after the conversion (65.5%). Individual viseme classification shows less satisfactory performance compared to word classification.
	Turkish lip-reading using Bi-LSTM and deep learning models[2]	Bi-LSTM + ResNet-18	Researchers try to detect what a person says from video frames without sound(Turkish), ResNet-18 model achieved higher classification success than other models.	Word and sentence datasets with accuracy values 84.5% and 88.55%, respectively. Study focus on Turkish lip-reading restricts its generalizability to other languages.
	Cross-language lipreading by reconstructing Spatio-Temporal relations in 3D convolution[3]	Serial-STRNet18, Parallel-STRNet18	Achieved an absolute improvement of 2.56% over the state-of-the-art model, Presented cross-language lip reading at the word-level.	Accuracy: 66.35%, 65.68%. Spatio-Temporal reconstruction model using 3D convolutional kernels may require significant computational resources for training and processing. Sophisticated techniques can increase the risk of overfitting.
	Deep learning	RNN-GRU	Achieved a lowered	Accuracy: 95%. Used common metric,

	based assistive technology on audio visual speech recognition for hearing impaired[4]	Speech-to-Text Model & CNN	word error rate of about 6.59% for ASR system, Introduced multimodal fusion.	single metric doesn't capture all aspects of performance. Ensuring the generalizability of the models to diverse real-world settings could be a challenge. Model achieved good performance, but further research is needed to address the challenge of seamlessly integrating audio and visual cues to avoid information loss or redundancy in real-world scenarios.

3. Discussion:

The Discussion section aims to synthesize and analyze the findings of the five key papers reviewed in this survey on lip reading. Each of these papers contributes unique insights into the field, offering valuable perspectives on the methodologies, challenges, and advancements in lip reading research. In the following sections, we provide a summary of the main findings from each paper, followed by a comprehensive analysis of the collective insights derived from the reviewed literature:

The study 'Lip Reading Sentences Using Deep Learning With Only Visual Cues' [1] introduces a novel system integrating a viseme classifier with spatial-temporal (3D) convolution and 2D ResNet architectures, trained on the LRS2 dataset. The proposed lip reading system is lexicon-free and uses purely visual cues, enabling recognition of words not presented during training and generalization to different languages. While achieving a significant improvement in performance with a 15% reduction in word error rate, the system exhibits exceptional viseme classification accuracy exceeding 95%. However, the general classification performance for individual segmented visemes has been less satisfactory compared to word classification. Additionally, the conversion process leads to a notable drop in word classification accuracy, decreasing to 65.5%. The study 'Turkish Lip-Reading Using Bi-LSTM and Deep Learning Models' [2] explores the application of Bi-LSTM and ResNet-18 architectures on a custom dataset comprising 111 words and 113 sentences for Turkish speech recognition from visual cues alone. The study uses advanced deep learning techniques, such as Bi-LSTM and ResNet-18, to achieve high accuracy in lip-reading. Additionally, the study provides two new datasets for scientific research, one with 111 words and the other with 113 sentences, which can be used for further research in the field. Researchers aim to understand what is being said by analyzing video frames without sound. Results indicate that the ResNet-18 model outperforms other models in classification accuracy. Specifically, the word and sentence datasets achieve accuracy values of 84.5% and 88.55%, respectively. However, the study only focuses on Turkish lip-reading, which limits its generalizability to other languages.

The study 'Cross-language lipreading by reconstructing Spatio-Temporal relations in 3D convolution' [3] investigates the effectiveness of Serial-STRNet18 and Parallel-STRNet18 models trained on the CELR-200 dataset. The findings reveal promising results, with an absolute improvement of 2.56% over the state-of-the-art model. Notably, the study presents successful cross-language lip reading at the word level. Statistical analysis indicates an accuracy of 66.35% for Serial-STRNet18 and 65.68% for Parallel-STRNet18. However, a limitation of the study lies in the computational

cost and model training difficulty associated with the original 3D convolutional kernel, which the proposed Spatio-Temporal reconstructed convolutional kernel aims to alleviate. On the other hand, the study introduces the novel cross-language lipreading dataset CELR-200, addressing the problem of insufficient generalization ability of lipreading systems in multilingual scenarios. Another study 'Deep Learning Based Assistive Technology on Audio-Visual Speech Recognition for Hearing Impaired' [4] employs RNN-GRU Speech-to-Text and CNN models, trained on datasets including LibriSpeech and GRID. The findings reveal a notable reduction in word error rate, reaching approximately 6.59% for the ASR system, alongside the introduction of multimodal fusion techniques. Moreover, the system achieves a high accuracy rate of 95%. The study's limitations include concerns about dataset size and diversity, insufficient discussion on evaluation metrics, and a lack of exploration into real-world implementation challenges.

4. Conclusion:

Deep learning architectures are driving impressive advancements in lip reading, with models achieving high accuracy in word and sentence recognition (up to 95%). Lexicon-free approaches and cross-language systems offer exciting possibilities for multilingual communication. However, limitations remain in generalizability, computational cost, and real-world robustness. Future research should focus on overcoming these challenges by exploring cross-lingual learning, enhancing robustness, and expanding datasets. Inspired by successful techniques such as viseme classifier combined with spatial-temporal (3D) convolution and 2D ResNet, further exploration of similar approaches can contribute to enhancing lip reading and speech reading accuracy, particularly in noisy environments. Additionally, deeper multimodal fusion and real-world deployment hold immense potential for transforming lip reading technology into a powerful tool for accessibility and human-computer interaction. By addressing these areas, researchers can unlock the full potential of deep learning for transformative lip reading systems, impacting communication accessibility and human-computer interaction.

5. Abbreviations paper:

3D	3-dimensional
2D	2-dimensional
ASR	Automatic Speech recognition
AV	Audio Visual
AVSR	Audio visual speech recognition
CNN	Convolutional neural network
RNN	Recurrent neural network
DL	Deep Learning
GRU	Gated recurrent unit
LRS2	Lip Reading Sentence 2
BBC	British Broadcasting Corporation
CELR	Chinese and English lipreading
ResNet	Residual Neural Network.
Bi-LSTM	Bidirectional long short-term memory
LSTM	Long short-term memory
ROI	Region of Interest
HOG	Histogram of Oriented Gradients
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
KNN	K- Nearest Neighbors
SVM	Support Vector Machine
HMM	Hidden Markov Model

Refference:

[1] Souheil Fenghour, Daqing Chen, Kun Guo, Perry Xiao "Lip Reading Sentences Using Deep Learning With Only Visual Cues"

[2] Umit Atila a , Furkan Sabaz b, "Turkish lip-reading using Bi-LSTM and deep learning models"

[3] Jiangfan Feng, Renhua Long, "Cross-language lipreading by reconstructing Spatio-Temporal relations in 3D convolution"

[4] Kumar, L. A., Renuka, D. K., Rose, S. L., & Wartana, I. M. (2022). Deep learning based assistive technology on audio visual speech recognition for hearing impaired. International Journal of Cognitive Computing in Engineering, 3, 24-30.

[5] Bowyer, Kevin W., Kyong Chang, and Patrick Flynn. "A survey of approaches and challenges in 3D and multi-modal 3D+ 2D face recognition." Computer vision and image understanding 101, no. 1 (2006): 1-15.

[6] SOUHEIL FENGHOUR, DAQING CHEN, KUN GUO, BO LI, AND PERRY XIAO, "Deep Learning-Based Automated Lip-Reading: A Survey"

[7] Sheetal Pujari, SK Sneha, R Vinusha, P Bhuvaneshwari, C Yashaswini, "A Survey on Deep Learning based Lip-Reading Techniques"

[8] T. Afouras, J. S. Chung, A. Senior, O. Vinyals and A. Zisserman, "Deep Audio-Visual Speech Recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8717-8727, 1 Dec. 2022, doi: 10.1109/TPAMI.2018.2889052.

[9] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 2015, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964.

[10] Cooke, Martin, Jon Barker, Stuart Cunningham, and Xu Shao. "The Grid Audio-visual Speech Corpus". Zenodo, January 1, 2006. https://doi.org/10.5281/zenodo.3625687.