Advances in Technology Innovation, vol. 11, no. 3, 2026, pp. 277-303

A Review of Lip-Reading: From Datasets and Architectures to Cross-Linguistic Performance Gaps

Qian Hu[1,2], Kuryati Kipli[1,*], Tengku Mohd Afendi Zulcaffle[1], Yuan Liu[2], Jianrong Cao[3]

[1]Faculty of Engineering, University Malaysia Sarawak, Kota Samarahan, Sarawak, Malaysia

[2]Institute of Computer and Information Engineering, Qilu Institute of Technology, Shandong, China

[3]School of Information and Electrical Engineering, Shandong Jianzhu University, Shandong, China

Received 17 April 2026; revised 27 May 2026; accepted 29 May 2026

DOI: https://doi.org/10.46604/aiti.2026.16377

Abstract

This paper aims to systematically review deep learning–driven lip-reading technologies, covering benchmark datasets, advances in model architecture, and cross-linguistic performance. Distinguished from existing surveys that predominantly enumerate methodologies in chronological sequence, this work consolidates the most extensive collection of 79 lip-reading datasets reported to date and establishes a problem-driven analytical framework to delineate the technological evolution from shallow feature extraction to global temporal modeling. Systematic analysis of the state of the art reveals a pronounced cross-linguistic performance disparity: state-of-the-art visual-only models attain 94.1% accuracy on standard English benchmarks, yet only 57.1% accuracy on Chinese benchmarks, representing a substantial performance gap of 37 percentage points rooted in intrinsic linguistic properties. This review identifies critical challenges and outlines promising research directions in lip-reading, therefore offering structured guidance for the continued development of multilingual visual speech recognition systems.

Keywords: lip-reading, visual speech recognition, deep learning, cross-linguistic performance

1. Introduction

Lip-reading, also known as visual speech recognition (VSR), infers spoken content by analyzing lip movements and has broad applications in areas such as hearing assistance, surveillance, military operations, and human–computer interaction. Traditional approaches rely on handcrafted features and classical classifiers, which exhibit limited generalization in complex wild environments and are highly sensitive to variations in illumination, head pose, and other environmental factors. The advent of deep learning, particularly the application of convolutional neural networks (CNN), recurrent neural networks (RNN), and Transformer architectures, has enabled models to automatically learn hierarchical and discriminative features. These advances have achieved substantial improvements in recognition accuracy and generalization performance.

To ensure the comprehensiveness and reproducibility of this review, systematic literature searches were conducted across major academic databases, including Google Scholar, IEEE Xplore, ACM Digital Library, and Web of Science, using the keywords “lip-reading” and “visual speech recognition”. Eligible publications primarily consisted of original research articles and review papers. Original research articles were considered if they introduced novel methodologies, original datasets, or quantifiable performance benchmarks, or reported reliable quantitative metrics on publicly available benchmark datasets. Through a systematic review and critical comparison of existing surveys in lip-reading research, this paper synthesizes and analyzes 11 representative surveys, summarizing their scope, strengths, weaknesses, and research gaps in Table 1.

* Corresponding author. E-mail address: kkuryati@unimas.my

Table 1 Comparison of existing lip-reading surveys

Reference	Year	Core Scope	Strengths	Weaknesses	Research Gaps
Zhou et al. [1]	2014	Traditional visual speech decoding	Systematic review of early geometric features and HMMs.	Limited to conventional methods and controlled environments; no deep learning.	Lacks deep architecture evolution and cross-linguistic analysis.
Mathulaprangsan et al. [2]	2015	Lip-reading and lip-password verification	Introduces lip-password biometric applications.	Shallow analysis; insufficient large-scale dataset comparisons.	No end-to-end deep learning or cross-linguistic evaluation.
Fernandez-Lopez & Sukno [3]	2018	Deep learning-based lip-reading	First comprehensive review of early CNN and LSTM approaches.	Narrow dataset coverage; limited end-to-end architecture analysis.	No problem-driven framework; no English–Chinese performance gap study.
Lu et al. [4]	2018	Lip-reading preprocessing and feature extraction	Clear pipeline overview of feature extraction workflows.	Few framework comparisons; ignore model evolution and multilingual issues.	No architectural progression; no cross-linguistic research.
Hao et al. [5]	2020	Comprehensive lip-reading survey	Broad multilingual coverage, including Chinese.	Descriptive-dominated; insufficient analysis of ROI and complex scenes.	No quantitative cross-linguistic gap; no global temporal modeling.
Chen et al. [6]	2020	Talking-head video generation	Provides benchmarks for speech-driven facial animation.	Focuses primarily on synthesis rather than recognition; limited relevance to core VSR.	Does not address key gaps in visual speech recognition.
Fenghour et al. [7]	2021	Deep learning architectures for lip-reading	In-depth analysis of 3D-CNN, Transformer, and attention mechanisms.	No dataset evolution review; no cross-linguistic comparison.	No problem-driven design; no linguistically induced performance gaps.
Sheng et al. [8]	2024	Visual speech analysis	Covers self-supervised and cross-modal pre-training.	Insufficient discussion of large-model efficiency; limited discussion of multilingual evaluation.	No English–Chinese accuracy gap; no linguistic cause analysis.
Pu & Wang [9]	2023	Machine lip-reading progress	Tracks shift from word-level to sentence-level recognition.	Limited generative AI discussion; insufficient multilingual datasets.	No quantitative cross-linguistic gap; no language-specific analysis.
Rathipriya & Maheswari [10]	2024	Lip-reading and sign language recognition	Integrates lip-reading and sign language; compiles 48 datasets.	Inadequate multimodal fusion analysis; no cross-linguistic comparison.	No English–Chinese performance gap; no linguistic explanation.
Deshpande et al. [11]	2025	Advanced lip-reading trends	Highlights ViT and large language model (LLM) integration trends.	Relies heavily on preprint studies; lacks dataset synthesis.	No unified dataset collection; no cross-linguistic gap exploration.

Based on the above analysis, two major research gaps are identified in this review:

(1) Existing surveys lack a problem-driven analytical framework to trace the evolutionary logic of deep learning models and thus fail to clarify how architectural innovations address the core challenges in lip-reading in a targeted manner.

(2) No prior survey has quantitatively investigated the cross-linguistic performance gap between English and Chinese visual speech recognition, nor has it systematically examined the intrinsic linguistic factors, such as the visual indistinguishability of Chinese lexical tones and the ambiguity in syllable-to-character mapping.

Accordingly, this review aims to address the following explicit research questions:

RQ1: What evolutionary patterns have emerged in lip-reading datasets and deep learning architectures over recent decades?
RQ2: What intrinsic linguistic factors lead to the significant and persistent performance gap between English and Chinese visual speech recognition?
RQ3: What key challenges and future directions exist in multilingual lip-reading research?

To bridge these gaps, this work systematically reviews deep learning-based lip-reading technologies by collecting and analyzing more than 120 peer-reviewed publications. The main contributions are threefold:

(1) The most comprehensive dataset collection to date is compiled, covering 79 datasets published from 1994 to 2026, and their evolutionary patterns are summarized in terms of data scale, language type, and collection environment.
(2) A problem-driven framework is established to systematically delineate the developmental trajectory of model architectures from shallow feature extraction to global temporal modeling.
(3) The cross-linguistic performance gap between English and Chinese visual speech recognition is quantitatively evaluated using standard benchmark datasets (LRW and CAS-VSR-W1k), and the 37-percentage-point discrepancy is analyzed from the perspective of linguistic characteristics, including the visual inaccessibility of Chinese lexical tones and the high ambiguity in syllable-to-character mapping.

The remainder of this review is structured as follows: Section 2 systematically analyzes the evolution of lip-reading datasets. Section 3 provides a comprehensive overview of the architectural evolution of deep learning models for lip -reading. Section 4 conducts a comparative performance analysis of state-of-the-art models on English and Chinese benchmark datasets. Section 5 discusses the current challenges and future research directions. Section 6 concludes the review.

2. Lip-Reading Datasets

Significant evolution has been observed in lip-reading datasets, progressing from small-scale, single-language datasets collected in controlled laboratory settings to large-scale, multi-language datasets collected in wild scenarios. From a macro perspective, existing lip-reading datasets can be classified into two major categories: controlled datasets and wild datasets. Table 2 summarizes 55 lip-reading datasets collected under controlled conditions, with data collection efforts dating back to 1994; Table 3 summarizes 24 lip-reading datasets collected in wild environments, with data collection initiatives commencing in 2014. The evolution of these datasets has played a pivotal role in shaping the development and evaluation of modern lip-reading systems, providing increasingly realistic and challenging benchmarks for model training and assessment. In this section, the lip-reading datasets in Tables 2 and 3 are systematically analyzed from four perspectives: acquisition environment and technology, dataset scale, task difficulty, and language diversity. The correlation between dataset evolution and model performance is further explored in the final part.

Table 2 Controlled lip-reading datasets

Datasets	Year	Task	Language	Speakers	Utterances	Video (Resolution)	View
Tulips [12]	1994	Digits	English	12	96	100 × 75, 30 fps	Frontal
M2VTS [13]	1999	Digits	French	37	2,920	288 × 360, 25 fps	Frontal
XM2VTSDB [14]	1999	Digits	English	295	2,360	720 × 576, 25 fps	Frontal + Left + Right
CAVSR1.0 [15]	2000	Words	Chinese	20	3,120	352 × 288, 25 fps	Frontal
AV Letters [16]	2002	Alphabets	English	10	780	376 × 288, 25 fps	Frontal
VidTIMIT [17]	2002	Sentences	English	43	N/A	384 × 512, 25 fps	Frontal
CUAVE [18]	2002	Digits	English	36	7,200	720 × 480, 30 fps	Frontal + side + tilting the head

Table 2 Controlled lip-reading datasets (continued)

Datasets	Year	Task	Language	Speakers	Utterances	Video (Resolution)	View
BANCA [19]	2003	Digits	English French Italian Spanish	208	29,952	720 × 576, 25 fps	Frontal
AV@CAR [20]	2004	Alphabets Digits Sentences	Spanish	20	800 600 6,000	768 × 576, 25 fps	Frontal
AVICAR [21]	2004	Alphabets Digits Sentences	English	86	59,000 59,000 59,000	720 × 480, 30 fps	Frontal
IBMIH [22]	2004	Digits	English	79	16,197	720 × 480, 30 fps	Frontal
AVOZES [23]	2004	Digits Sentences	English	20	200 60	720 × 480, 30 fps	Frontal + Left + Right
AV-TIMIT [24]	2004	Sentences	English	233	4,600	720 × 480, 30 fps	Frontal
VALID [25]	2005	Digits	English	106	1,590	720 × 576, 25 fps	Frontal + side
HIT Bi-CAV [26]	2005	Sentences	Chinese	10	6,000	256 × 256, 25 fps	Frontal
UWB-05-HSAVC [27]	2005	Sentences	Czech	100	20,000	720 × 576, 25 fps	Frontal
GRID [28]	2006	Sentences	English	34	34,000	720 × 576, 25 fps	Frontal + side
CMU-AVPFV [29]	2007	Words	English	10	15,000	640 × 480, 30 fps	Frontal + Profile
IV2 [30]	2008	Sentences	French	300	4,500	780 × 576, 25 fps	Frontal + Profile
HIT-AVDB-II [31]	2008	Sentences	Chinese	30	1,980	720 × 576, 25 fps	0, 30, 60, and 90
UWB-07-ICAV [32]	2008	Sentences	Czech	50	10,000	720 × 576, 50 fps	Frontal
IBMSR [33]	2008	Digits	English	38	1,661	368 × 240, 30 fps	-90, 0, and 90
AVLetters2 [34]	2008	Alphabets	English	5	910	1920 × 1080, 50 fps	Frontal
OuluVS1 [35]	2009	Digits Phrases Sentences	English	20	1,000	720 × 576, 25 fps	Frontal + side
WAPUSK20 [36]	2010	Phrases	English	20	2,000	640 × 480, 32 fps	Frontal
QuLips [37]	2010	Digits	English	2	3,600	720×576, 25 fps	Frontal + Left + Right
NDUTAVSC [38]	2010	Digits Words Sentences	German	66	6,907	640 × 480, 100 fps	Frontal
LILiR [39]	2010	Sentences	English	12	2,400	720 × 576, 32 fps	0, 30, 45, 60, and 90
CENSREC-1-AV [40]	2010	Digits	Japanese	42	3,234	720 × 480, 30 fps	Frontal
UNMC-VIER [41]	2011	Sentences	English	123	2,460	708 × 640, 29 fps	Frontal+ Left + Right
BL [42]	2011	Sentences	French	17	4,064	640 × 480, 30 fps	Frontal + Profile
LTS5 [43]	2011	Digits	French	20	180	1920 × 1080, 25 fps	0, 30, 60, and 90
AGHAV [44]	2012	Digits	Polish	20	N/A	1920 × 1080, 50 fps	Frontal
MOBIO [45]	2012	Sentences	English	152	N/A	640 × 480, 16 fps	Frontal + side

Table 2 Controlled lip-reading datasets (continued)

Datasets	Year	Task	Language	Speakers	Utterances	Video (Resolution)	View
AVAS [46]	2013	Digits	Arabic	50	13,850	640 × 480, 30 fps	-90, -45, 0, 45, and 90
		Words					
		Phrases					
MIRACL-VC [47]	2014	Words	English	15	1,500	640 × 480, 15 fps	Frontal
		Phrases					
		Words			966,000		
		Sentences			59,000		
TCD-TIMIT [48]	2015	Sentences	English	62	6,913	1920 × 1080, 30 fps	Frontal
OuluVS2 [49]	2015	Digits	English	53	1,590	1920 × 1080, 30 fps	Frontal + side
		Phrases			1,590		
		Sentences			530		
RM-3000 [50]	2015	Sentences	English	1	3,000	360 × 640, 60 fps	Frontal
IBM-AV-ASR [51]	2015	Sentences	English	262	N/A	704 × 480, 30 fps	Frontal
KinectDigits [52]	2016	Digits	English	15	N/A	N/A	Frontal + Slightly side
HAVRUS [53]	2016	Phrases	Russian	20	4,000	640 × 480, 200 fps	Frontal
MODALITY [54]	2017	Sentences	English	35	5,880	1920 × 1080, 100 fps	Frontal
VLRF [55]	2017	Sentences	Spanish	24	10,200	1280 × 720, 50 fps	Frontal
AV Digits [56]	2018	Digits	English	53	795	1280 × 720, 30fps	0, 45, and 90
		Phrases		39	5,850		
Lombard GRID [57]	2018	Sentences	English	54	N/A	704 × 480, 24 fps	Frontal + side
RAVDESS [58]	2018	Words	English	24	N/A	1920 × 1080, 30 fps	Frontal
AVSD [59]	2019	Phrases	Arabic	22	1,100	1920 × 1080, 30 fps	Frontal
ASPIRE [60]	2019	Sentences	English	3	N/A	1920 × 1080, 30 fps	Frontal + Profile
MEAD [61]	2020	Sentences	English	60	N/A	1920 × 1080, 30 fps	Frontal + Slightly side
VR Digits [62]	2020	Digits	English	6	6,000	1920 × 1080, 25 fps	Frontal
DeepLip [63]	2021	Sentences	English	132	N/A	88 × 88,25 fps	Frontal
		Phrases					
Speaking Faces [64]	2021	Words	English	142	N/A	768 × 512, 25 fps	Frontal + side
		Phrases					
MobLip [65]	2024	Words	Greek	30	55,275	160 × 80, 30 fps	Frontal + slight dynamic
LipBengal [66]	2025	Words	Bengali	150	353,150	720 × 1280, 30 fps	Frontal

Table 3 Wild lip-reading datasets

Datasets	Year	Task	Language	Speakers	Utterances	Video (Resolution)	View
AusTalk [67]	2014	Digits	English	1,000	24,000	640 × 480, 25 fps	Frontal
		Words			966,000		Frontal
		Sentences			59,000		Frontal
LRW [68]	2017	Words	English	1,000+	538,766	256 × 256, 25 fps	Frontal + side
VoxCeleb [69]	2017	Sentences	English	1,251	153,516	N/A	Frontal

Table 3 Wild lip-reading datasets (continued)

Datasets	Year	Task	Language	Speakers	Utterances	Video (Resolution)	View
MV-LRS [70]	2017	Sentences	English	1,000+	74,564	160 × 160, 25 fps	0 ~ 90
LRS2-BBC [71]	2018	Sentences Phrases	English	1,693	118,116	160×160, 25fps	-30 ~ 30
LRS3-TED [72]	2018	Sentences Phrases	English	5,000	165,000	224 × 224, 25 fps	Frontal and Profile
AVSpeech [73]	2018	Sentences	English	150,000	N/A	N/A	Different face poses
VoxCeleb2 [74]	2018	Sentences	English	6,112	1,028,246	N/A	Videos from YouTube
LSVSR [75]	2018	Sentences	English	1,000+	2,934,899	N/A	Frontal + dynamic view
CAS-VSR-W1k [76]	2019	Words	Chinese	2,000+	718,018	256× 256, 25 fps	Frontal + side
CMLR [77]	2019	Sentences	Chinese	11	102,072	64 × 128, 25 fps	Frontal + side
YTDEV18 [78]	2019	Sentences Phrases	English	1,000+	N/A	128 × 128, 25 fps	Frontal + side
AVA-Active Speaker [79]	2019	Sentences	English	Several	N/A	N/A	multi-view
Lip2wav [80]	2020	Sentences	English	5	N/A	48 × 48, 25 fps	Frontal + side
NSTDB [81]	2020	Sentences	Chinese	1,000+	N/A	64 × 64, 25 fps	-90 ~ 90
HDTF [82]	2021	Sentences	English	300+	10K+	N/A	Frontal + Natural dynamic
KSC [83]	2021	Words Phrases Sentences	Kazakh	1,000+	N/A	N/A	Frontal + side
LRWR [84]	2021	Words	Russian	1,000+	117,500	1920 × 1080, 25 fps	Frontal + side
VVAD-LRS3 [85]	2021	Sentences Phrases	English	1,000+	N/A	1920 × 1080, 25 fps	Frontal + side
RUSAVIC [86]	2022	phrases	Russian	20	N/A	1920 × 1080, 60 fps	Frontal + Natural dynamic
CANDOR [87]	2022	Conversation	English	7 million	N/A	N/A	Frontal + Natural dynamic
DMCLR [88]	2022	phrases	Chinese	11	102,072	1920 × 1080, 30 fps	Frontal
CN-CVS [89]	2023	Sentences	Chinese	2,681	193,245	N/A	Frontal
MuAViC [90]	2023	Sentences	9 languages	8,000+	N/A	N/A	Frontal + dynamic multi-view

2.1 Collection Environment and Technology

The collection environment for lip-reading datasets has evolved from highly controlled laboratory conditions to unconstrained wild scenarios to support a broader range of application contexts. Early datasets were typically recorded in strictly controlled laboratory environments to ensure consistent lighting, fixed backgrounds, and minimal external noise, thereby maintaining high data quality. The 55 lip-reading datasets listed in Table 2 were all collected under controlled

experimental conditions, including AVLetters [16] and AVOZES [23]. Although some datasets adopted multi-view recording to obtain more comprehensive facial information, their highly controlled environments restrict the generalization ability of models in wild applications.

With the growing demand for robust speech recognition in natural environments, several datasets have incorporated unconstrained conditions, collecting data under varying lighting, backgrounds, and noise levels. For example, RUSAVIC [86] focuses on in-vehicle scenarios, covering diverse lighting conditions, driving states, and background noise, thereby facilitating the application of lip-reading technology in driving safety systems and voice assistants. Similarly, MuAViC [90] spans nine languages and includes indoor, outdoor, and noisy environments, encompassing a variety of complex wild scenarios.

Advances in data collection technologies have also significantly improved data quality. Early datasets were typically recorded at low resolutions, limiting the ability to capture fine lip movements; for example, Tulips [12] has a resolution of only 100×75 pixels. In contrast, modern datasets use high-resolution images to enhance the fidelity of lip movement capture. Notably, datasets such as AVLetters2 [34], LTS5 [43], AGHAV [44], TCD-TIMIT [48], OuluVS2 [49], MODALITY [54], RAVDESS [58], AVSD [59], MEAD [61], VR Digits [62], RUSAVIC [86], and DMCLR [88] provide high-definition recordings at 1920×1080, enabling more precise modeling of subtle lip motion dynamics.

2.2 Dataset Scale

In the early stages of lip-reading research, dataset scales were relatively limited, typically consisting of a small number of speakers, utterances, and categories, and were primarily used for exploratory studies. For example, Tulips [12] includes only 12 speakers and 96 utterances, while M2VTS [13] comprises 37 speakers and 2,920 utterances. These datasets were collected entirely in controlled laboratory environments and thus do not fully reflect the complexity of in-the-wild application scenarios.

As research in lip-reading has progressed, dataset sizes have expanded significantly, encompassing more speakers, larger utterances, and multi-view recording techniques. For example, CUAVE [18] includes 36 speakers and 7,200 utterances, with multi-view recordings to enhance dataset complexity. TCD-TIMIT [48] further improved dataset quality, covering 62 speakers and 6,913 utterances, and used high-resolution video to support continuous sentence recognition.

In recent years, the scale of lip-reading datasets has exhibited a clear increasing trend. LRW [68] focuses on word-level lip-reading in unconstrained environments, comprising 500 word categories and 538,766 utterances. CAS-VSR-W1k [76] further extends the number of categories to 1,000, with 718,018 utterances, making it one of the most comprehensive datasets in the field of Chinese lip-reading. LRS3-TED [72], which includes approximately 5,000 speakers and 165,000 utterances, has become a benchmark dataset for sentence-level lip-reading recognition. Overall, dataset development has been characterized by substantial growth in scale, diversity, and realism.

Fig. 1 depicts the temporal evolution trend of lip-reading datasets from 1994 to 2026. In this figure, blue circles denote datasets collected in controlled laboratory environments, whereas orange triangles denote datasets collected in natural real-world scenarios. From 1994 to 2010, most lip-reading datasets remained at a relatively small scale, generally below 10^4 utterances, with only a few controlled datasets reaching the 10^4–10^5 range. Since 2017, dataset sizes have grown exponentially, with wild-scenario datasets experiencing the most significant expansion. Typical benchmarks such as LRW, LRS3, and CAS-VSR-W1k have successfully expanded to the order of 10^5–10^6 utterances. The continuous accumulation of large-scale data resources has strongly supported the development of deep learning-based visual speech recognition systems. Moreover, large-scale datasets have enabled the transition from handcrafted features to deep neural architectures, significantly improving recognition performance.

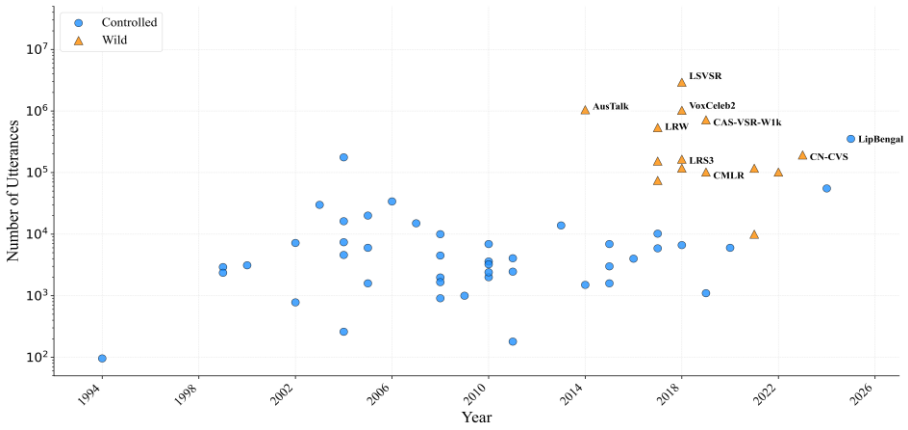


Fig. 1 Growth of lip-reading datasets (1994-2026)

2.3 Task Difficulty

The task design of lip-reading datasets exhibits a clear hierarchical progression, evolving from basic discrete-category recognition to continuous sentence-level understanding, and in some cases extending to complex multimodal tasks involving emotion recognition. The types of lip-reading tasks have evolved from simple to complex, spanning digits, letters, isolated words, phrases, and continuous sentences. Early datasets, such as Tulips [12], contained only four digit classes and served as benchmark resources. As research advanced, the scope of tasks expanded to phrase-level and sentence-level recognition, enabling continuous speech analysis. For instance, TCD-TIMIT [48] comprises 6,913 sentences covering 39 phoneme classes. More recent datasets have further extended task complexity, including emotion recognition, speech enhancement, and speech separation. For example, MEAD [61] provides high-resolution recordings of 60 speakers expressing eight distinct emotions, reflecting the continued escalation of lip-reading research toward more diverse and challenging research objectives.

2.4 Language Types

Existing lip-reading datasets exhibit highly imbalanced language distributions. As illustrated in Fig. 2, English datasets account for 54 instances, representing 68.4% of the total, whereas Chinese datasets comprise only 8 instances, corresponding to 10.1%. Datasets in other languages—including French, Spanish, German, Arabic, Russian, Czech, Japanese, Polish, Kazakh, Greek, and Bengali—collectively constitute the remaining 21.5%. Distinct discrepancies in phonetic structures, syllable formation, and articulatory habits across languages inevitably lead to significant differences in final recognition performance. This pronounced imbalance represents a critical limitation in the field, fundamentally constraining models' cross-linguistic generalization capabilities. Consequently, future research is expected to benefit from the construction of multilingual datasets, with particular emphasis on expanding Chinese corpora. Given that Mandarin is the most widely spoken native language globally, enhancing both the scale and quality of Chinese lip-reading datasets is crucial for advancing the field and enabling more robust cross-linguistic applications.

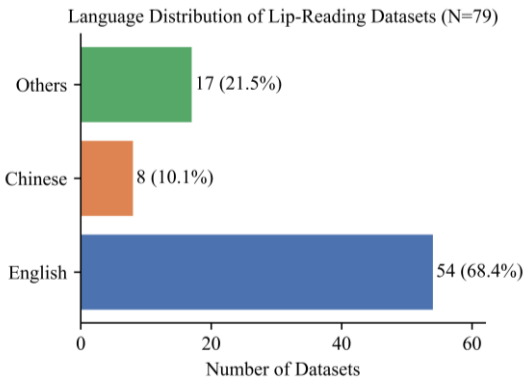


Fig. 2 Language distribution of the 79 lip-reading datasets surveyed

2.5 Relationship between Dataset Evolution and Model Performance

A strong co-evolutionary relationship exists between lip-reading datasets and model performance. On the one hand, the expansion of dataset scale, increased environmental complexity, and elevated task difficulty collectively provide the essential data foundation and standardized evaluation benchmarks for sustained performance breakthroughs. On the other hand, pivotal advances in model architectures—such as the transition from convolutional neural networks to Transformer-based frameworks—have imposed more stringent requirements on datasets, including longer temporal sequences of continuous utterances, higher-resolution recordings for capturing fine-grained lip movements, and precise phoneme-level temporal alignment annotations. Current challenges in lip-reading research extend beyond sheer data volume; they also involve improvements in data quality, sample diversity, annotation granularity, and linguistic coverage, which are essential for supporting the development of cross-lingual, multimodal, and robust lip-reading models.

A retrospective analysis of lip-reading research reveals that prior to 2017, most deep learning models were trained on relatively small-scale, laboratory-controlled datasets such as GRID. In recent years, however, driven by the rapid advancement of large-scale models and self-supervised pre-training techniques, large-scale datasets collected in unconstrained real-world scenarios have become the mainstream. Representative benchmarks include LRW [68], LRS2-BBC [71], LRS3-TED [72], CAS-VSR-W1k [76], and CMLR [77].

3. Evolution of Deep Learning in Lip-Reading

Lip-reading typically comprises four core stages: data preprocessing, feature extraction, sequence modeling, and output decoding. In the data preprocessing stage, videos of the target speaker's lip movements are first collected and converted into valid image sequences for model inputs. In the feature extraction stage, algorithms learn dynamic representations of lip shape changes. The sequence modeling stage utilizes classifiers to establish a mapping between the extracted visual features and linguistic representations. Finally, in the output decoding stage, corresponding text transcription results or further synthesized speech information are generated. The overall lip-reading workflow is illustrated in Fig. 3.

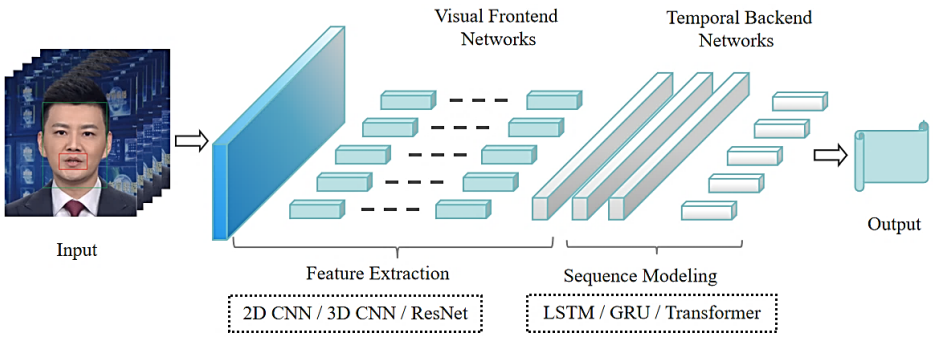


Fig. 3 Overall diagram of lip-reading

Front-end feature extraction in lip-reading primarily relies on CNN-based architectures such as ResNet variants. Back-end sequence modeling is typically implemented using LSTM, GRU, TCN, and Transformer-based frameworks. Overall, the development of lip-reading models has exhibited two main trends: first, the gradual evolution from shallow feature learning to deep spatiotemporal representation learning; second, the shift from local temporal modeling to the capture of global sequence dependencies.

In contrast to previous surveys that primarily organize architectures chronologically, this review adopts a problem-driven analytical framework, elucidating the core challenges addressed by each architecture class, the breakthroughs achieved, and the limitations that remain for subsequent model designs. Table 4 systematically summarizes six representative model categories, highlighting their structural characteristics, key features, and respective advantages and disadvantages. A detailed exposition and analysis along this logical trajectory are provided in the following sections.

Table 4 Analysis of deep learning model architectures

Model Type	Core Feature	Strength	Limitation
2D CNN	Spatial feature extraction per frame	Efficient for static lip shapes	Cannot capture temporal dynamics
3D CNN	Spatiotemporal feature extraction	Captures short-term motion	Limited receptive field for long sequences
3D CNN + ResNet	Deep residual connections	Enhanced abstract feature representation	Still relies primarily on local temporal modeling
RNN/LSTM/GRU	Sequence modeling	Captures long-term dependencies	Sequential computation limits parallelism
TCN	Parallel temporal convolutions	Supports parallel computation, controllable receptive field, and stable training	Receptive field remains dependent on network depth
Transformer	Global attention	Parallel modeling, long-term dependencies	Quadratic computational cost

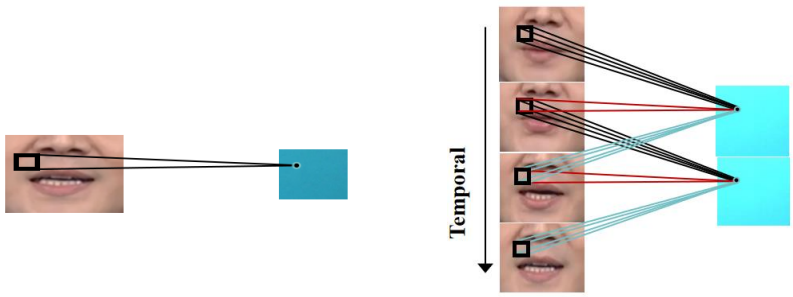
3.1 From Simple Neural Network Feature Extraction to Deep Neural Networks

The primary objective in lip-reading is the extraction of informative visual features from lip movement videos. The evolution along this dimension can be traced from handcrafted feature-based methods to shallow neural networks capable of automatic feature extraction, and ultimately to deep convolutional neural networks that learn hierarchical and discriminative spatiotemporal representations. A central challenge is learning discriminative visual representations that capture both static appearance and dynamic spatiotemporal patterns.

(1) 2D CNN

Early deep learning-based lip-reading methods primarily utilized simple neural networks for feature extraction, while classification still relied on traditional approaches. In 2011, Ngiam et al. [91] first employed deep autoencoders as feature extractors, combined with support vector machines (SVM) for classification. This approach was capable of recognizing digits (0–9) and letters (A–Z), achieving a significant improvement in accuracy compared to traditional methods. This work represents one of the earliest applications of deep learning to lip-reading research and is widely regarded as an important milestone in the field.

With the development of computer vision, 2D CNN has been introduced into the field of lip-reading. 2D CNN can extract semantically rich spatial features from single-frame cropped lip images, thereby significantly enhancing the representation capability of static lip shapes. As shown in Fig. 4(a), a typical 2D CNN uses sliding windows to extract local features from input images of size H×W, where each convolution kernel captures edge, texture, and shape information. Subsequently, the generated feature maps are processed through pooling layers to reduce spatial dimensions, thereby improving computational efficiency and enhancing robustness to positional variations. Finally, fully connected layers are flattened and transformed into vector representations, which serve as inputs for subsequent sequence modeling networks to learn temporal features.



(a) 2D CNN (b) 3D CNN

Fig. 4 Visual front-end network

In 2014, Noda et al. [92] first introduced CNN into the field of visual speech recognition. They used a 2D CNN to extract image features of the lip region, which were then fed into an LSTM sequence model. This approach achieved a recognition accuracy of 58% on a 40-phoneme task, demonstrating the effectiveness of 2D CNN in learning robust representations of lip movements. Building on this work, Wand et al. [93] developed an end-to-end neural network lip-reading system in 2016 by stacking feed-forward neural networks with LSTM. Evaluated on a 51-word recognition task using the GRID dataset, the system achieved a word-level accuracy of 79.6%, which was 11.6 percentage points higher than traditional SVM classifiers based on handcrafted visual features (Eigenlips and HOG). Together, these studies established the distinct advantage of 2D CNN-based features over handcrafted features in modeling visual speech.

However, the inherent frame-by-frame processing of 2D CNN limits their ability to capture temporal dynamics, prompting researchers to develop feature extraction methods capable of modeling both spatial and temporal dimensions simultaneously.

(2) 3D CNN

While 2D CNN has demonstrated significant effectiveness in capturing spatial features from static frames, lip-reading is intrinsically a dynamic temporal process. Reliance on single-frame spatial information alone is insufficient to represent the motion trajectories of lip movements during continuous speech. To address this limitation, 3D CNN was introduced into the lip-reading domain. As illustrated in Fig. 4(b), a 3D CNN accepts consecutive video frames as input and applies convolutional kernels across both spatial and temporal dimensions, enabling the direct extraction of spatiotemporal features from video segments.

In 2015, Tran et al. [94] proposed the C3D architecture, empirically demonstrating that 3D CNN is more suitable than 2D CNN for modeling spatiotemporal video features. Subsequent studies explored diverse applications of 3D CNN in lip-reading. Torfi et al. [95] developed an audio-visual speech recognition system leveraging a coupled 3D CNN to process visual streams. Xu et al. [96] introduced LCANet, whose front end integrates 3D convolutional layers and highway networks, while its back end employs a bidirectional GRU. Chen et al. [81] constructed a network for sentence-level Mandarin lip-reading, combining 3D CNN with DenseNet to enhance feature extraction.

By incorporating the temporal dimension, 3D CNN effectively addresses the limitations of 2D CNN in dynamic sequence modeling, capturing short-term lip motion patterns such as lip closure, opening, and protrusion. However, the temporal receptive field of 3D CNN remains constrained by the kernel depth, limiting their ability to model long-range temporal dependencies spanning dozens of frames. Moreover, early 3D CNN architectures often exhibit restricted representational capacity, hindering the learning of highly abstract lip-reading features. These limitations have motivated the exploration of deeper and more sophisticated network architectures.

(3) Integration of 3D CNN with ResNet

Conventional 3D CNN architectures suffer from notable drawbacks when extracting highly abstract spatiotemporal features. To address these inherent limitations, researchers integrated 3D CNN with ResNet, thereby increasing network depth and enhancing representational capacity.

In 2015, He et al. [97] introduced the ResNet architecture, which substantially accelerated network training while achieving superior classification accuracy. Central to ResNet is the residual module, commonly referred to as a “skip connection” or “shortcut connection,” as illustrated in Fig. 5. The residual module introduces a bypass pathway between convolutional layers, allowing features from earlier layers to be directly propagated as input to deeper layers. This mechanism enables higher layers to learn abstract features while preserving low-level information from earlier layers, mitigating information loss during feature extraction and improving the network’s representational power.

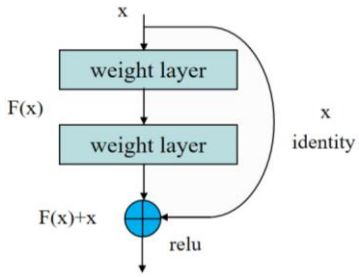


Fig. 5 Residual block

In 2017, Stafylakis and Tzimiropoulos [98] were among the first to introduce residual learning into the field of lip-reading, proposing a framework that integrates 3D CNN, ResNet, and highway networks, achieving 83% accuracy on the LRW dataset. In 2019, Zhang et al. [99] developed LipCH-Net, combining a 14-layer ResNet with a VGG-M, demonstrating superior performance in handling Chinese polyphonic characters and complex lip shape variations. Subsequent studies have further validated this hybrid design. In 2020, Feng et al. [100] adopted an architecture that integrates 3D CNN-ResNet with bidirectional GRU, achieving an accuracy of 88.5% on the LRW.

The combination of 3D CNN and ResNet effectively leverages the short-term dynamic modeling capability of 3D CNN and the deep feature extraction advantage of ResNet, having become a mainstream backbone in lip-reading systems. However, the core operations of both 3D CNN and ResNet are still limited to spatiotemporal convolutions within local receptive fields. For tasks that require parsing long-range dependencies in continuous speech, such as distinguishing homophones, simply increasing network depth is insufficient to achieve true global context awareness. Therefore, developing methods capable of modeling long-range temporal dependencies through feature extraction remains a key challenge in visual speech recognition.

3.2 From Local Spatiotemporal Modeling to Global Sequence Learning

Visual speech recognition inherently involves analyzing continuous sequences of lip movements, rendering effective temporal modeling a central challenge. The key question addressed in this section is how to model frame-level visual feature sequences to capture speech context dependencies spanning tens to hundreds of frames. The evolution along this dimension can be characterized as a progression from RNN for sequential modeling, to TCN enabling parallel computation, and ultimately to Transformer architectures that provide global context awareness.

(1) RNN and Sequence Modeling

When using 3D CNN alone for spatiotemporal modeling, the temporal receptive field is constrained by the kernel depth, limiting the ability to capture long-range dependencies. To address this limitation, RNN and their variants have been introduced into the domain of visual speech recognition.

The RNN is specifically designed for sequential data processing, in which historical information is propagated through recurrent connections to influence the current state. The network architecture of a conventional RNN is illustrated in Fig. 6. However, the RNN is prone to vanishing or exploding gradient problems during backpropagation, which restricts its capacity to model long-range temporal dependencies.

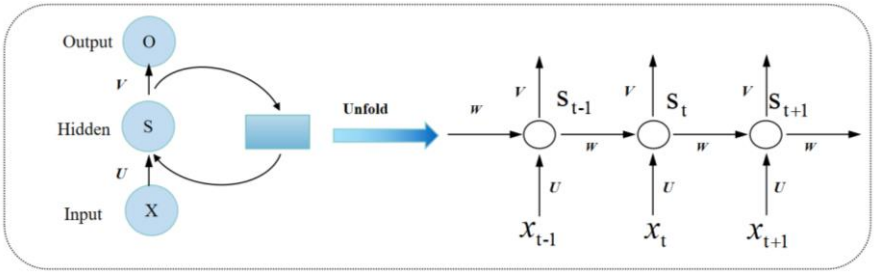


Fig. 6 Structure of RNN

The LSTM represents a significant advancement over traditional RNN by incorporating memory cells and gating mechanisms to mitigate these issues. The architecture of an LSTM is shown in Fig. 7, comprising three gates: the forget gate, which determines which historical information to discard; the input gate, which regulates the incorporation of new information; and the output gate, which controls the hidden state output at each time step. In 2014, Noda et al. [92] proposed an early CNN-LSTM framework for visual speech recognition, using a 2D CNN to extract spatial features from lip images, which were then modeled sequentially by the LSTM. This approach achieved a recognition accuracy of 58% across 40 phonemes, demonstrating the effectiveness of LSTM for temporal modeling. Subsequently, in 2016, Assael et al. [101] introduced LipNet, an end-to-end model integrating 3D CNN, bidirectional LSTM, and the CTC loss function, achieving a word-level accuracy of 95.2% on the GRID and significantly accelerating the adoption of LSTM in lip-reading research.

The GRU is a streamlined variant of LSTM that has garnered attention for its reduced parameters and improved training efficiency. The GRU architecture is illustrated in Fig. 8. By merging the forget gate and input gate into a single update gate and combining the cell state with the hidden state, GRU can efficiently regulate information flow. The synergistic effect of the reset gate and update gate not only achieves effective gradient control but also reduces model complexity, improving training efficiency without sacrificing temporal modeling capability.

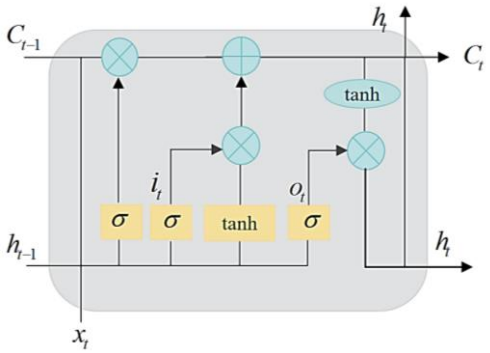


Fig. 7 Structure of LSTM

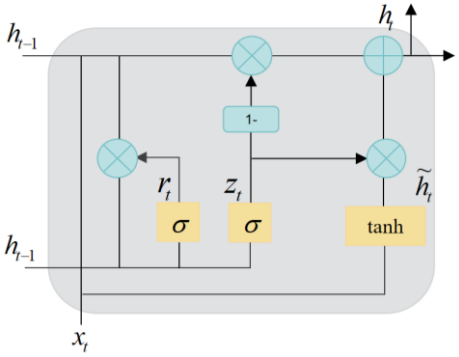


Fig. 8 Structure of GRU

The evolution from RNN to LSTM and GRU architectures reflects the continuous optimization of recurrent networks in balancing long-range dependency modeling, training efficiency, and architectural simplicity. These models provide effective sequence modeling capabilities for visual speech recognition, enabling the extraction of contextual temporal representations from frame-level features. However, unidirectional temporal modeling relies solely on past information, which limits the integration of future context and thereby restricts the ability to capture global semantic dependencies. This limitation has motivated the development of bidirectional sequence modeling approaches, which leverage both past and future information to afford a more comprehensive contextual understanding in lip-reading tasks.

(2) Bidirectional Sequence Modeling and Context Integration

To further enhance temporal modeling capabilities, bidirectional recurrent architectures have been adopted in visual speech recognition. Representative models, such as the bidirectional gated recurrent unit (Bi-GRU), integrate past and future contextual information at each time step through independently computed forward and backward layers. As illustrated in Fig. 9, the Bi-GRU architecture comprises an input layer, a forward computation layer, a backward computation layer, and an output layer. The forward layer processes the sequence from past to future, whereas the backward layer processes it from future to past, and the output layer fuses the hidden states from both directions. This bidirectional design enables the model to capture temporal dependencies in lip movements more comprehensively, particularly for tasks requiring richer contextual understanding. The gating mechanisms in Bi-GRU also help alleviate the vanishing-gradient problem, thereby enhancing adaptability and generalization in complex linguistic environments.

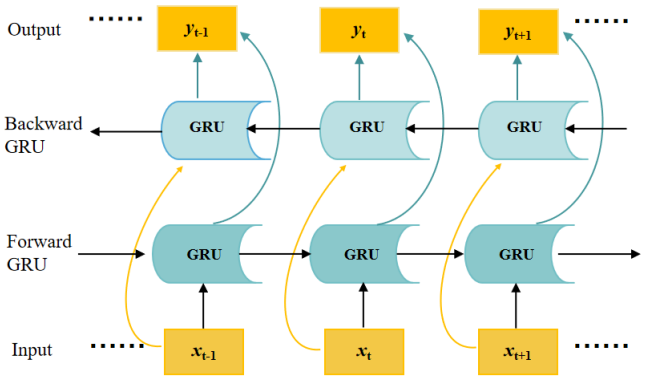


Fig. 9 Architecture of Bi-GRU

In 2017, Xu et al. [96] introduced LCANet, an end-to-end lip-reading system that encodes input video frames through stacked 3D CNN combined with highway networks and captures short- and long-term spatiotemporal features via Bi-GRU. Experiments on the GRID dataset demonstrated a character error rate (CER) of 1.3% and a word error rate (WER) of 3.0%, representing a 12.3% improvement over the previous state-of-the-art. Subsequently, Bi-GRU has been widely adopted for temporal modeling in lip-reading. Feng et al.[100] combined 3D ResNet with Bi-GRU to achieve 88.5% accuracy on LRW. Arakane and Saitoh [102] further optimized network structures to improve temporal modeling efficiency, achieving 94.1% accuracy on LRW. Collectively, these studies underscore the efficacy of Bi-GRU, whose bidirectional context modeling and gating mechanisms have established it as a widely adopted approach for temporal feature extraction in lip-reading.

Despite these advantages, Bi-GRU does not resolve the inherent limitation of sequential computation, which constrains training efficiency. For continuous speech sequences comprising hundreds of frames, the sequential nature of Bi-GRU prevents full exploitation of GPU parallelism. This computational bottleneck becomes increasingly significant in the era of large-scale datasets.

(3) TCN and Parallel Sequence Modeling

TCN was first introduced by Lea et al. [103] in 2016 for action segmentation in video sequences. As illustrated in Fig. 10, the TCN architecture is underpinned by three core mechanisms: causal convolutions, which ensure that the output at each time step depends solely on current and past inputs; dilated convolutions, which exponentially expand the receptive field through an increasing dilation factor; and residual connections, which stabilize the training of deep networks. Unlike recurrent networks, TCN enables parallel computation across all time steps, thereby substantially improving training efficiency.

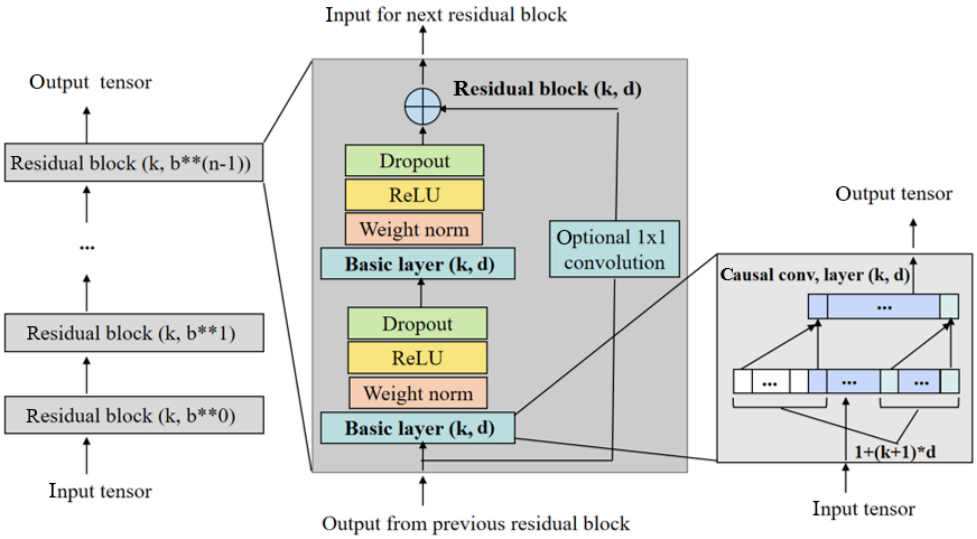


Fig. 10 Architecture of TCN

In visual speech recognition, Martinez et al. [104] pioneered the application of TCN in 2020 and proposed the multi-stage TCN (MS-TCN) architecture. This method replaced Bi-GRU layers in prevailing state-of-the-art models and achieved accuracies of 85.30% on the LRW dataset and 41.40% on the CAS-VSR-W1k dataset.

The TCN framework has been continuously optimized in subsequent studies. Ma et al. [105] proposed the densely connected TCN (DC-TCN) in 2021, achieving an improved accuracy of 88.36% on LRW. To adapt to edge deployment scenarios, their team further optimized the model in 2022 by integrating multiple optimization strategies and depthwise separable convolutions, reducing computational cost to 87.8% of the original version while boosting the LRW accuracy to 94.10%. Based on MS-TCN, Arakane and Saitoh [102] also reached 94.10% LRW accuracy in 2023 via word-boundary supervision and knowledge distillation. These studies collectively demonstrate that TCN has emerged as a competitive temporal modeling paradigm for word-level lipreading tasks.

The adoption of TCN in lipreading marks a paradigm shift in temporal modeling from sequential computation of recurrent networks to parallel computation of convolutional networks, thereby substantially improving training efficiency. Although dilated convolutions enable an exponential expansion of the receptive field, such expansion remains inherently local and constrained by network depth. For Mandarin Chinese lipreading, disambiguating homophones typically relies on long-range semantic context. The convolutional operation of TCN essentially performs sliding weighted summation along the temporal axis, making it difficult to directly capture global dependencies between arbitrary frames, in contrast to attention mechanisms.

(4) Transformer and Global Parallel Modeling

As mentioned earlier, RNN-based models are constrained by sequential computation, resulting in limited training efficiency; whereas TCN can achieve parallel processing, it still exhibits fundamental limitations when applied to tonal languages such as Mandarin Chinese. Specifically, their fixed receptive fields and inherent local convolution operations hinder the modeling of long-range dependencies, which are crucial for resolving homophone ambiguities. The introduction of the Transformer architecture has fundamentally addressed these issues. Its core self-attention mechanism can directly compute the relationships between any two frames in a sequence, thereby enabling fully parallel and globally contextualized modeling. At the same time, it eliminates the recursive structure of RNN and overcomes the inherent receptive field limitations of TCN.

The Transformer model, centered on the attention mechanism, adopts an encoder-decoder framework, as illustrated in Fig. 11. The encoder consists of multiple stacked layers, each including a multi-head self-attention mechanism, a feed-forward neural network (FNN), residual connections, and layer normalization. The decoder structure is similar to that of the encoder but additionally incorporates a cross-attention layer. The positional encoding mechanism is used to provide both absolute and relative positional information for elements within the sequence. By combining global attention with parallel computation, the Transformer significantly improves training efficiency while effectively capturing the long-range temporal dependencies that are crucial for lip-reading tasks.

In 2018, Afouras et al. [71] first introduced the Transformer architecture to lip-reading, achieving a word-level accuracy of 51.7% on the sentence-level LRS2 benchmark, which represented a 22% improvement over the state of the art at that time. Subsequent studies have continued to optimize and extend Transformer-based approaches. In attention mechanism optimization, Prajwal et al. [106] replaced ResNet with a Transformer to learn, track, and integrate lip motion features, achieving 87.4% word-level accuracy on LRS2. For training stability, Li et al. [107] introduced regularization via dropout into Transformer-based lip-reading models. In terms of visual feature extraction, Park et al. [108] proposed the SwinLip model in 2025. This model leverages the hierarchical attention mechanism of the Swin Transformer to capture both local and global lip movement features while maintaining computational efficiency, achieving an accuracy of 90.7% on the LRW dataset. Collectively, these studies highlight the transformative potential of the Transformer architecture in lip-reading.

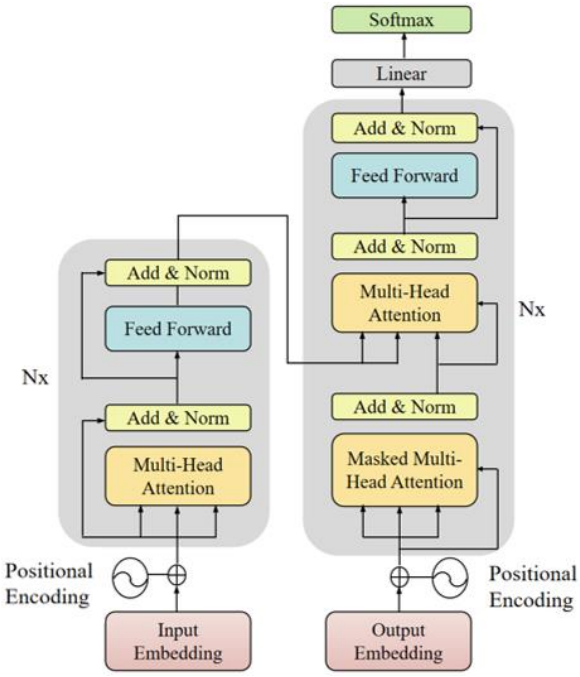


Fig. 11 Architecture of Transformer

Owing to its self-attention mechanism and parallel processing capability, the Transformer has become the dominant architecture for time-series modeling in lip-reading. It performs particularly well in sentence-level tasks and tonal languages such as Mandarin Chinese, where understanding long-range contextual information is crucial. However, the computational complexity of self-attention grows quadratically with sequence length, which remains a major limitation. For high-frame-rate video sequences, the direct application of the Transformer leads to significant computational overhead.

Collectively, each network architecture possesses distinct merits and inherent limitations, and a single model structure fails to simultaneously satisfy the requirements of temporal dynamic perception, computational efficiency, and global contextual modeling. Benefiting from their respective strengths in bidirectional contextual integration, parallel temporal computation, and long-range global dependency capture, BiGRU, TCN, and Transformer have been increasingly adopted for experimental validation in contemporary lip-reading investigations. Such widespread adoption also facilitates comprehensive performance comparisons among state-of-the-art models.

4. Performance in Lip-Reading

Table 5 systematically summarizes and comparatively analyzes the studies published between 2019 and 2025, evaluating the performance of pure visual lip-reading models on the LRW and CAS-VSR-W1k benchmark datasets. It is important to emphasize that these studies differ in aspects such as data augmentation, training strategies, and evaluation protocols, making strict numerical comparison not strictly valid. Therefore, the absolute values reported in Table 5 should be interpreted as references reflecting overall trends rather than definitive rankings. Subsequent analyses will focus on the relative performance improvements and the potential factors contributing to cross-linguistic performance differences, rather than direct numerical ranking.

4.1 Performance Evolution Driven by Model Architecture

Table 5 delineates the evolutionary trajectory of visual-only lip-reading model architectures. Overall, the trend reflects a progression in front-end networks from 2D CNN to 3D CNN combined with ResNet, which is now considered the de facto standard, and a parallel evolution in back-end networks from Bi-GRU to TCN and, subsequently, Transformer-based architectures. The analysis below emphasizes relative performance improvements rather than absolute metrics.

Table 5 Comparative top-1 accuracy of SOTA lip-reading models on the LRW and CAS-VSR-W1k datasets

References	Time	Front-end Network	Back-end Network	LRW	CAS-VSR-W1k
Courtney et al. [109]	2019	2D CNN	Deep ConvLSTM	83.40%	-
Wang et al. [110]	2019	3D + 2D CNN	Bi-ConvLSTM	83.34%	36.91%
Yang et al. [76]	2019	3D Conv + ResNet-34	Bi-GRU	-	38.19%
Luo et al. [111]	2020	3D CNN + ResNet	Bi-GRU	83.50%	38.70%
Zhao et al. [112]	2020	3D + 2D CNN	Bi-GRU	84.41%	38.79%
Martinez et al. [104]	2020	3D CNN + ResNet-18	MS-TCN	85.30%	41.40%
Xiao et al. [113]	2020	3D CNN + ResNet+ DFN	Bi-GRU	84.13%	41.93%
Zhang et al. [114]	2020	3D CNN+ ResNet-18	Bi-GRU	85.02%	45.24%
Feng et al. [100]	2020	3D CNN + ResNet-18	Bi-GRU	88.40%	55.70%
Ma et al. [105]	2021	3D CNN + ResNet-18	DC-TCN	88.36%	43.65%
Ivanko et al. [115]	2022	2D CNN	Bi-LSTM	88.70%	-
Koumparoulis et al. [116]	2022	EfficientNetV2	Transformer	89.52%	-
Ma et al. [117]	2022	3D CNN + ResNet-18	DC-TCN	94.10%	-
Arakane and Saitoh [102]	2023	3D CNN + ResNet-18	MS-TCN	94.10%	55.70%
Chen et al. [118]	2023	3D CNN + ResNet-18	DC-TCN	92.2%	-
Rathipriya and Maheswari [10]	2024	ResNet + CRO-TSM	3D Conv	92.40%	-
Wang et al. [119]	2025	3D CNN + Vision Transformer	Bi-GRU	88.30%	57.10%
Wu et al. [120]	2025	3D CNN	STDNet	90.20%	53.56%
Park et al. [108]	2025	Swin Transformer + 3D Embedding	Bi-GRU/TCN/ Transformer	90.70%	-

(1) Front-end Network Evolution

Early studies, such as Courtney et al. [109], employed 2D CNN as the front-end, achieving 83.40% top-1 accuracy on the LRW dataset. The integration of 3D CNN with ResNet rapidly became the prevailing front-end configuration. For instance, Yang et al. [76] reported a baseline accuracy of 38.19% on CAS-VSR-W1k, while Zhang et al. [114], through a face-cropping strategy, raised word-level accuracy to 85.02%. Recent front-end explorations, including Swin Transformer [108], EfficientNet [116], and Vision Transformer [119], have emerged post-2022. Nevertheless, the 3D CNN plus ResNet combination still achieves the highest performance (94.10% on LRW), demonstrating its effectiveness for word-level lip-reading.

(2) Back-end Network Evolution

Back-end architectures have transitioned from recurrent networks to convolutional networks, and ultimately to Transformer-based models. During 2019–2020, Bi-GRU represented the mainstream choice, with Luo et al. [111] and Zhao et al. [112] reporting 83.50% and 84.41% top-1 accuracy, respectively. In 2020, Martinez et al. [104] replaced Bi-GRU with MS-TCN, achieving an approximate absolute gain of 1.2 percentage points on LRW. In 2021, Ma et al. [105] introduced DC-TCN, attaining 88.36% on LRW, which represents approximately a 4-point improvement over contemporary Bi-GRU models. In 2022, Koumparoulis et al. [116] combined EfficientNetV2 with a Transformer back-end, achieving 89.52% on LRW. Across word-level tasks, TCN architectures generally attain slightly higher performance than Transformer models, reflecting that precise local temporal modeling may be more critical than global context understanding in short-context scenarios. Conversely, Transformer architectures are expected to excel in sentence-level tasks requiring disambiguation of homophones.

(3) Contribution of Training Strategies

Beyond architectural innovations, systematic optimization of training strategies has substantially enhanced performance. Feng et al. [100] retained the 3D CNN + ResNet–BiGRU backbone while incorporating MixUp, label smoothing, cosine

learning rate decay, and word-boundary information, boosting LRW accuracy from approximately 83% to 88.40%, and Chinese accuracy from 38% to 55.70%. Ma et al.[105] conducted comprehensive studies of strategy combinations, identifying time masking as the most impactful augmentation, followed by MixUp. Leveraging combined strategies and additional pretraining, LRW word-level accuracy was elevated to 94.10%, marking a milestone for visual-only models. Arakane and Saitoh [102] further validated the generalizability of the 3D CNN + ResNet18 front-end with an MS-TCN back-end across multiple datasets, consistently maintaining 94.10% top-1 accuracy on LRW.

4.2 Cross-Linguistic Performance Gap: English vs. Chinese

Building on the analysis of architectural evolution, this section further investigates the persistent performance gap between Chinese and English lip-reading. This discrepancy is not merely a difference in quantitative metrics but also reflects fundamental linguistic distinctions in visual speech recognition between the two languages.

4.2.1 Quantitative Analysis of the Performance Gap

As of 2025, the highest reported top-1 accuracy for visual-only lip-reading on the CAS-VSR-W1k dataset reached 57.10%, achieved by the Mini-3DCvT proposed by Wang et al. [119]; whereas the best performance on the LRW dataset was 94.10%, derived from the work of Ma et al. [117], reflecting a gap of approximately 37 percentage points. Historical data shows that in 2019, the highest accuracy for Chinese lip-reading was 38.19%, compared to 83.40% for English lip-reading, yielding a disparity of roughly 45 percentage points. By 2025, this absolute difference had narrowed by about 9 percentage points, primarily driven by an 18.91-percentage-point improvement in Chinese lip-reading performance.

Fig. 12 illustrates the evolution of peak performance for English and Chinese visual lip-reading from 2019 to 2025. English performance increased from 84.41% to 94.10%, approaching saturation, whereas Chinese performance rose from 38.19% to 57.10%, indicating substantial remaining potential. The vertical distance between the two curves, referred to as the performance gap, decreased from approximately 46 to 37 percentage points, demonstrating that research in Chinese lip-reading is making progress, yet a significant disparity remains.

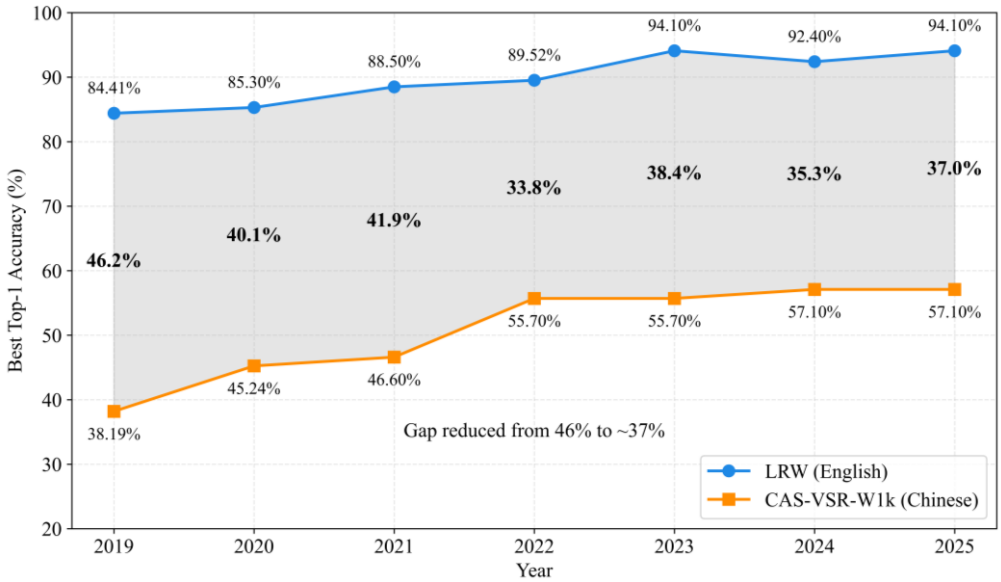


Fig. 12 Cross-linguistic performance gap: English vs. Chinese

Beyond the overall performance gap, Table 5 further reveals that architectural improvements that yielded substantial gains on LRW did not translate proportionally to CAS-VSR-W1k. For instance, Ma et al. [105] achieved 88.36% on LRW, surpassing the Bi-GRU baseline by approximately 4.9 percentage points; however, its performance on the Chinese task was only 43.65%, falling short of the 55.70% attained by Feng et al. [100], which used the same Bi-GRU architecture with

optimized training strategies. Conversely, Wang et al. [119], which integrates 3D convolution with Vision Transformer, reached 88.30% on LRW—below the best TCN model (94.10%)—yet achieved 57.10% on the Chinese dataset, outperforming all reported TCN models for Chinese lip-reading.

4.2.2 Linguistic Factors Underlying the Performance Gap between English and Chinese Lip-Reading

Extensive experimental results consistently reveal a persistent performance gap between English and Chinese lip-reading tasks in existing visual speech recognition systems. This cross-lingual accuracy discrepancy cannot be explained solely by differences in model architectures or training strategies. Instead, the gap can be largely attributed to three linguistic dimensions.

(1) Visually indistinguishable tonal contrasts

In Mandarin Chinese, each syllable consists of an initial consonant, a vowel nucleus, and a lexical tone. Lexical tones—comprising high-level (¯), rising (´), dipping (ˇ), falling (ˋ), and neutral tones—are primarily realized via vocal fold vibration frequency, with minimal observable movement in accessible articulators such as the lips and jaw. Consequently, lip movements alone provide insufficient cues to distinguish tonal variants of the same syllable, e.g., "妈 (mā), 麻 (má), 马 (mǎ), 骂 (mà)."

(2) One-to-many syllable-to-character mapping

Modern Chinese contains far more characters than distinct syllables, resulting in pervasive one-to-many correspondences. For instance, the pinyin "yi" corresponds to 366 distinct characters in Xinhua Zidian, including "一", "衣", and "医". In purely visual lip-reading, identical lip motion sequences may map to hundreds of candidate characters, producing substantially higher ambiguity than alphabetic writing systems.

(3) Context-dependent polysemy

Individual Chinese characters can convey different meanings depending on context, even when pronounced identically. As illustrated in Table 6, the character "中" pronounced as "zhōng" may refer to "middle", "China", or a progressive aspect marker, whereas when pronounced as "zhòng", it may mean "to win a prize" or "to suffer from heatstroke." Resolving such polysemy typically requires contextual information beyond the target word itself.

Table 6 The pronunciation and meaning of "中"

Chinese character	Pin yin	Chinese meaning	English meaning	Combine words
中	zhōng	中间	middle	中间; 集中
		里面	inside	心中; 家中
		中等	intermediate	中才; 中则
		适于	suitable	中用; 中看
		可以	all right	中不中; 真中
		正在进行	ongoing	研究中; 学习中
		中国	China	中文; 古今中外
	zhòng	获得	obtain	中奖; 中选
		受到	suffer	中计; 中毒
		充满	full	中暑

The linguistic analysis above is corroborated by the empirical data presented in Table 5. Despite CAS-VSR-W1k encompassing a larger dataset (718,018 utterances) than LRW (538,766 utterances), its highest reported accuracy remains markedly lower (57.10% versus 94.10%). Even when the same Transformer architecture is applied, LRW achieves 92.40%, whereas CAS-VSR-W1k reaches only 57.10%. Considering both dataset scale and model architecture, these factors alone are unlikely to fully explain the approximately 37-percentage-point performance gap. This discrepancy highlights the intrinsic

challenges posed by the linguistic characteristics of Chinese lip-reading, where the visual inaccessibility of tonal information and the high ambiguity arising from one-to-many syllable-to-character mappings fundamentally constrain purely visual Chinese lip-reading.

Notably, this language-inherent difficulty interacts with model architecture. The highest performance on the English LRW dataset (94.10%) was achieved using a TCN, whereas the best result on Chinese CAS-VSR-W1k (57.10%) employed a Transformer. The TCN's local receptive fields are well-suited for modeling the relatively linear phoneme sequences of English, while the Transformer's global attention mechanism more effectively resolves the long-range ambiguities inherent in Chinese due to missing tonal cues and syllable-to-character mappings. This cross-linguistic difference in optimal architecture underscores that no single temporal modeling framework is universally superior; rather, model selection should consider the specific linguistic characteristics of the target language.

4.2.3 Error Analysis of Visual Speech Recognition

Apart from the aforementioned model architectures and cross-linguistic factors, the performance discrepancy between the LRW and CAS-VSR-W1k datasets in visual speech recognition is primarily attributed to three error sources: viseme ambiguity, speaker variability, and environmental visual interference.

(1) Viseme ambiguity

A viseme is defined as the visual counterpart of a phoneme that characterizes distinct lip morphological patterns. Both English and Chinese exhibit widespread viseme confusion, wherein multiple phonemes correspond to an identical viseme. Typical examples include the English phonemes /p/ and /b/, as well as the Chinese initial consonants /zh/, /ch/, and /sh/, all of which present highly similar lip movements. Pure visual models are consequently prone to misclassifying acoustically similar words, and this ambiguity is particularly pronounced in Chinese visual speech recognition.

(2) Speaker variability

Individual differences in lip morphology, speaking rate, accent, and articulatory habits substantially degrade the generalization capability of recognition models. Speakers exhibit varied lip movement amplitudes, ranging from subtle to exaggerated; even identical words show distinct visual lip characteristics across speakers. Compared with English datasets, Chinese datasets contain more diverse speaker distributions. Models trained on limited speaker samples generally fail to generalize to unseen speakers, thereby further widening the performance gap between English and Chinese visual speech recognition tasks.

(3) Environmental visual noise

Real-world scenarios involve diverse visual disturbances, including uneven illumination, head-pose deflection, motion blur, and partial lip occlusion. These interfering factors distort critical lip features, thereby inevitably reducing recognition accuracy. In contrast to the relatively standardized English benchmark dataset LRW, Chinese datasets such as CAS-VSR-W1k incorporate more complex real-world scenarios. This inherent complexity results in weaker robustness under environmental interference for Chinese visual speech recognition models.

The above analysis explains the performance bottleneck of current visual speech recognition models, as exemplified by the 94.1% accuracy achieved on the LRW dataset. Furthermore, it reveals the essential causes of the persistent cross-linguistic performance disparity. Chinese recognition systems are more susceptible to viseme ambiguity due to abundant tonal homophones, whereas English benefits from greater discriminability between visemes and phonemes. Future research can mitigate the aforementioned recognition errors by implementing multi-view data augmentation, speaker-adaptive training strategies, and anti-noise front-end feature extraction networks.

5. Future Directions and Open Challenges

To advance lip-reading toward real-world viability and linguistic inclusivity, several critical challenges must be addressed. This section outlines four key future directions: dataset construction, model architectures, practical deployment, and cross-linguistic generalization.

(1) Dataset Construction

Existing lip-reading datasets exhibit three major limitations: the absence of fine-grained annotations such as phonemes and visemes, insufficient temporal alignment accuracy, and a highly imbalanced language distribution. Future dataset development should prioritize the creation of hierarchical annotation frameworks spanning phonemes, words, and sentences; the expansion of multilingual corpora; the inclusion of speaker and environmental metadata; and the establishment of standardized protocols for data acquisition and evaluation.

(2) Model Architectures

Lip-reading faces a fundamental trade-off between modeling capacity and computational efficiency. While the CNN is constrained by limited receptive fields, Transformers offer global sequence modeling but incur quadratic complexity O (L2). Promising research directions include sparse or hierarchical attention mechanisms, linear-complexity state-space models, graph neural networks for modeling lip key point relationships, and the integration of lip-reading with LLM. LLM has demonstrated potential in contextual reasoning and cross-modal inference, which can enhance sentence-level disambiguation and zero-shot transfer. Recent studies, such as Mamba [121] and VALLR [122], have begun exploring these directions.

(3) Practical Deployment

High-performance lip-reading models remain computationally intensive, limiting their deployment on edge devices. Lightweight design strategies can be pursued at both architectural and compression levels: using efficient front-end modules such as Swin Transformer, and applying pruning, quantization, and knowledge distillation. To address challenging conditions—including low lighting, occlusion, and head pose variations—future work should develop robust video enhancement techniques and uncertainty-aware fusion mechanisms.

(4) Cross-Linguistic Generalization

Performance disparities between Mandarin and English lip-reading primarily arise from the visual inaccessibility of tonal information and the high ambiguity in mapping syllables to Chinese characters. Future research should investigate: visual proxies for tonal cues; cross-lingual pretraining strategies that leverage large-scale English datasets before adapting to low-resource languages, as demonstrated by Tapu et al. [123], who showed that lightweight adapters mapping visual features into LLM embedding spaces enable effective generalization to low-resource languages; and linguistically informed model design, integrating prior knowledge of phoneme sets, tone systems, and other language-specific features into network architectures.

6. Conclusion

This review comprehensively summarizes the advances of visual speech recognition (VSR) from 2016 to 2026, covering the evolutionary trajectories of public datasets, model architectures, and performance benchmarks. A total of 79 publicly available datasets released between 1994 and 2026 were compiled. Adopting a problem-driven analytical framework, the technical evolution from shallow neural feature extractors to global attention-based architectures was systematically analyzed. The key findings are concluded as follows:

(1) Research has shifted from small-scale controlled datasets to large-scale wild and multilingual resources. The volume of speech samples has increased from below 10^4 before 2010 to between 10^5 and 10^6 after 2017.

(2) Model paradigms have evolved progressively from 2D CNN for static feature extraction, to 3D CNN combined with ResNet for short-term dynamic modeling and deep feature learning, and further advanced from RNN and LSTM toward TCN and Transformer for global context capturing. Each stage has addressed the inherent limitations of previous frameworks.

(3) A persistent performance gap of approximately 37 percentage points remains between English lip-reading (94.1%) and Chinese visual speech recognition (57.1%) under the LRW and CAS-VSR-W1k benchmarks. This gap is primarily attributed to the intrinsic linguistic properties of Chinese, including the visual inaccessibility of tonal cues and severe ambiguity induced by homophones, rather than differences in dataset scale or model representation capability.

Overall, lip-reading research is transitioning from data-centric single-model optimization toward linguistically informed, multimodally integrated, and cross-lingually generalized innovation. Limitations inherent to visual modality must be addressed through the integration of linguistic priors and architectural advances, enabling lip-reading to evolve from benchmark performance toward robust real-world deployment. This paradigmatic reconfiguration will establish a unified framework for multilingual visual speech recognition, further unlocking the potential of lip-reading technologies across diverse linguistic contexts.

Acknowledgments

This work is supervised by professors from Universiti Malaysia Sarawak. This work is also supported by the research program of Qilu Institute of Technology (No.: QIT23NN022).

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Z. Zhou, G. Zhao, X. Hong, and M. Pietikäinen, "A Review of Recent Advances in Visual Speech Decoding," Image and Vision Computing, vol. 32, no. 9, pp. 590-605, 2014.

[2] S. Mathulaprangsan, C.-Y. Wang, A. Z. Kusum, T.-C. Tai, and J.-C. Wang, "A Survey of Visual Lip Reading and Lip-Password Verification," Proceedings of 2015 International Conference on Orange Technologies (ICOT), IEEE, pp. 22-25, 2015.

[3] A. Fernandez-Lopez and F. M. Sukno, "Survey on Automatic Lip-Reading in the Era of Deep Learning," Image and Vision Computing, vol. 78, pp. 53-72, 2018.

[4] Y. Lu, J. Yan, and K. Gu, "Review on Automatic Lip Reading Techniques," International Journal of Pattern Recognition and Artificial Intelligence, vol. 32, no. 7, article no. 1856007, 2018.

[5] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A Survey of Research on Lipreading Technology," IEEE Access, vol. 8, pp. 204518-204544, 2020.

[6] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, "What Comprises a Good Talking-Head Video Generation?: A Survey and Benchmark," arXiv preprint arXiv:2005.03201, 2020.

[7] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep Learning-Based Automated Lip-Reading: A Survey," IEEE Access, vol. 9, pp. 121184-121205, 2021.

[8] C. Sheng, G. Kuang, L. Bai, C. Hou, Y. Guo, X. Xu, et al., "Deep Learning for Visual Speech Analysis: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 9, pp. 6001-6022, 2024.

[9] G. Pu and H. Wang, "Review on Research Progress of Machine Lip Reading," The Visual Computer, vol. 39, no. 7, pp. 3041-3057, 2023.

[10] N. Rathipriya and N. Maheswari, "A Comprehensive Review of Recent Advances in Deep Neural Networks for Lipreading with Sign Language Recognition," IEEE Access, vol. 12, pp. 136846-136879, 2024.

[11] S. Deshpande, K. Shirsath, A. Pashte, P. Loya, S. Shingade, and V. Sambhe, "A Comprehensive Survey of Advancement in Lip Reading Models: Techniques and Future Directions," IET Image Processing, vol. 19, no. 1, article no. e70095, 2025.

[12] J. R. Movellan, "Visual Speech Recognition with Stochastic Networks," Advances in Neural Information Processing Systems, vol. 7, pp. 851-858, 1994.

[13] O. Vanegas, K. Tokuda, and T. Kitamura, "Location Normalization of HMM-Based Lip-Reading: Experiments for the M2VTS Database," Proceedings of 1999 International Conference on Image Processing (Cat. 99CH36348), IEEE, pp. 343-347, 1999.

[14] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The Extended M2VTS Database," Second International Conference on Audio and Video-Based Biometric Person Authentication, Washington, D.C., USA, 1999.

[15] Y. Xu, L. Du, G. Li, X. Zhang, and Z. Zhou, "Chinese Audiovisual Bimodal Speech Database CAVSR1. 0," Acta Acoust. A Sinica, vol. 25, no. 1, pp. 42-49, 2000.

[16] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of Visual Features for Lipreading," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198-213, 2002.

[17] C. Sanderson, "The VidTIMIT Database," IDIAP–Com 02-06, IDIAP, 2002.

[18] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research," Proceedings of 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, IEEE, pp. II-2017-II-2020, 2002.

[19] E. Bailly-Bailliére, S. Bengio, F. Bimbot, M. Hamouz, J. Kittler, J. Mariéthoz, et al., "The BANCA Database and Evaluation Protocol," Proceedings of International Conference on Audio-and Video-Based Biometric Person Authentication, Springer, pp. 625-638, 2003.

[20] A. Ortega, F. Sukno, E. Lleida, A. F. Frangi, A. Miguel, L. Buera, et al., "AV@CAR: A Spanish Multichannel Multimodal Corpus for In-Vehicle Automatic Audio-Visual Speech Recognition," Proceedings of International Conference on Language Resources and Evaluation (LREC), European Language Resources Association, pp. 763-766, 2004.

[21] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, J. Boley, S. Borys, et al., "AVICAR: Audio-Visual Speech Corpus in a Car Environment," Proceedings of Interspeech Conference, ISCA, pp. 2489-2492, 2004.

[22] J. Huang, G. Potamianos, J. Connell, and C. Neti, "Audio-Visual Speech Recognition Using an Infrared Headset," Speech Communication, vol. 44, no. 1-4, pp. 83-96, 2004.

[23] R. Goecke and J. B. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES," Proceedings of the 8th International Conference on Spoken Language Processing, Interspeech, pp. 2525-2528, 2004.

[24] T. J. Hazen, K. Saenko, C.-H. La, and J. R. Glass, "A Segment-Based Audio-Visual Speech Recognizer: Data Collection, Development, and Initial Experiments," Proceedings of the 6th International Conference on Multimodal Interfaces, ACM Press, pp. 235-242, 2004.

[25] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: A New Practical Audio-Visual Database, and Comparative Results," Proceedings of International Conference on Audio- and Video-Based Biometric Person Authentication, Springer, pp. 777-786, 2005.

[26] X Hong, H Yao, M Xu, "Bimodal Database and Its Material Segmentation for Lip-Reading Recognition on Sentence," Computer Engineering and Applications, vol. 3, 2005.

[27] P. Cisar, M. Železný, Z. Krňoul, J. Kanis, J. Zelinka, and L. Müller, "Design and Recording of Czech Speech Corpus for Audio-Visual Continuous Speech Recognition," Proceedings of the Auditory-Visual Speech Processing International Conference 2005 (AVSP2005), pp. 93-96, 2005.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," The Journal of the Acoustical Society of America, vol. 120, no. 5, pp. 2421-2424, 2006.

[29] K. Kumar, T. Chen, and R. M. Stern, "Profile View Lip Reading," Proceedings of 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, IEEE, pp. IV-429-IV-432, 2007.

[30] D. Petrovska-Delacrétaz, S. Lelandais, J. Colineau, L. Chen, B. Dorizzi, M. Ardabilian, et al., "The IV^2 Multimodal Biometric Database (Including Iris, 2D, 3D, Stereoscopic, and Talking Face Data), and the IV^2-2007 Evaluation Campaign," Proceedings of 2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems, IEEE, pp. 1-7, 2008.

[31] X. Lin, H. Yao, X. Hong, and Q. Wang, "HIT-AVDB-II: A New Multi-View and Extreme Feature Cases Contained Audio-Visual Database for Biometrics," Proceedings of the 11th Joint International Conference on Information Sciences (JCIS 2008), Atlantis Press, pp. 357-363, 2008.

[32] J. Trojanová, M. Hrúz, P. Campr, and M. Železný, "Design and Recording of Czech Audio-Visual Database with Impaired Conditions for Continuous Speech Recognition," Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), pp. 1239-1243, 2008.

[33] P. Lucey, G. Potamianos, and S. Sridharan, "Patch-Based Analysis of Visual Speech from Multiple Views," Proceedings of the International Conference on Auditory-Visual Speech Processing 2008, AVISA, pp. 69-74, 2008.

[34] S. J. Cox, R. W. Harvey, Y. Lan, J. L. Newman, and B.-J. Theobald, "The Challenge of Multispeaker Lip-Reading," Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP 2008), AVISA, pp. 179-184, 2008.

[35] G. Zhao, M. Barnard, and M. Pietikäinen, "Lipreading with Local Spatiotemporal Descriptors," IEEE Transactions on Multimedia, vol. 11, no. 7, pp. 1254-1265, 2009.

[36] A. Vorwerk, X. Wang, D. Kolossa, S. Zeiler, and R. Orglmeister, "WAPUSK20—A Database for Robust Audiovisual Speech Recognition," Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA), pp. 3016-3019, 2010.

[37] A. Pass, J. Zhang, and D. Stewart, "An Investigation into Features for Multi-View Lipreading," Proceedings of 2010 17th IEEE International Conference on Image Processing (ICIP), IEEE, pp. 2417-2420, 2010.

[38] A. G. Chitu, K. Driel, and L. J. Rothkrantz, "Automatic Lip Reading in the Dutch Language Using Active Appearance Models on High Speed Recordings," Proceedings of the 3rd International Conference on Text, Speech and Dialogue, Springer, pp. 259-266, 2010.

[39] Y. Lan, B.-J. Theobald, R. W. Harvey, E.-J. Ong, and R. Bowden, "Improving Visual Features for Lip-Reading," Proceedings of Auditory-Visual Speech Processing (AVSP), 2010.

[40] S. Tamura, C. Miyajima, N. Kitaoka, T. Yamada, S. Tsuge, T. Takiguchi, et al., "CENSREC-1-AV: An Audio-Visual Corpus for Noisy Bimodal Speech Recognition," Proceedings of Auditory-Visual Speech Processing (AVSP), ISCA, 2010.

[41] Y. W. Wong, S. I. Ch'ng, K. P. Seng, L.-M. Ang, S. W. Chin, W. J. Chew, et al., "A New Multi-Purpose Audio-Visual UNMC-VIER Database with Multiple Variabilities," Pattern Recognition Letters, vol. 32, no. 13, pp. 1503-1510, 2011.

[42] Y. Benezeth, G. Bachman, G. Le-Jan, N. Souviraà-Labastie, and F. Bimbot, "BL-Database: A French Audiovisual Database for Speech Driven Lip Animation Systems," INRIA, 2011.

[43] V. Estellers and J.-P. Thiran, "Multipose Audio-Visual Speech Recognition," Proceedings of 2011 19th European Signal Processing Conference, IEEE, pp. 1065-1069, 2011.

[44] M. Igras, B. Ziółko, and T. Jadczyk, "Audiovisual Database of Polish Speech Recordings," Studia Informatica, vol. 33, no. 2B, pp. 163-172, 2012.

[45] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernocký, et al., "Bi-modal Person Recognition on a Mobile Phone: Using Mobile Phone Data," Proceedings of 2012 IEEE International Conference on Multimedia and Expo Workshops, IEEE, pp. 635-640, 2012.

[46] S. Antar and A. Sagheer, "Audio-Visual Arabic Speech (AVAS) Database for Human-Computer Interaction Applications," The International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, no. 9, 2013.

[47] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "A New Visual Speech Recognition Approach for RGB-D Cameras," Proceedings of Image Analysis and Recognition: 11th International Conference, Springer, pp. 21-28, 2014.

[48] N. Harte and E. Gillen, "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," IEEE Transactions on Multimedia, vol. 17, no. 5, pp. 603-615, 2015.

[49] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "Ouluvs2: A Multi-view Audiovisual Database for Non-rigid Mouth Motion Analysis," Proceedings of 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, pp. 1-5, 2015.

[50] D. L. Howell, "Confusion Modelling for Lip-Reading," Ph.D. dissertation, School of Computing Sciences, University of East Anglia, Norwich, UK, 2015.

[51] Y. Mroueh, E. Marcheret, and V. Goel, "Deep Multimodal Learning for Audio-Visual Speech Recognition," Proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 2130-2134, 2015.

[52] L. Schönherr, D. Orth, M. Heckmann, and D. Kolossa, "Environmentally Robust Audio-Visual Speaker Identification," Proceedings of 2016 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp. 312-318, 2016.

[53] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, and M. Železný, "HAVRUS Corpus: High-Speed Recordings of Audio-Visual Russian Speech," Proceedings of International Conference on Speech and Computer, Springer, pp. 338-345. 2016.

[54] A. Czyzewski, B. Kostek, P. Bratoszewski, J. Kotus, and M. Szykulski, "An Audio-Visual Corpus for Multimodal Automatic Speech Recognition," Journal of Intelligent Information Systems, vol. 49, no. 2, pp. 167-192, 2017.

[55] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, "Towards Estimating the Upper Bound of Visual-Speech Recognition: The Visual Lip-Reading Feasibility Database," Proceedings of 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), IEEE, pp. 208-215, 2017.

[56] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-Only Recognition of Normal, Whispered and Silent Speech," Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6219-6223, 2018.

[57] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A Corpus of Audio-Visual Lombard Speech with Frontal and Profile Views," Journal of the Acoustical Society of America, vol. 143, no. 6, pp. EL523-EL529, 2018.

[58] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English," PLOS ONE, vol. 13, no. 5, article no. e0196391, 2018.

[59] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, "An Arabic Visual Dataset for Visual Speech Recognition," Procedia Computer Science, vol. 163, pp. 400-409, 2019.

[60] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "CochleaNet: A Robust Language-Independent Audio-Visual Model for Real-Time Speech Enhancement," Information Fusion, vol. 63, pp. 273-285, 2020.

[61] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, et al., "Mead: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation," Proceedings of European Conference on Computer Vision, Springer, pp. 700-717, 2020.

[62] Y. Lu and H. Li, "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory," Applied Sciences, vol. 9, no. 8, article no. 1599, 2019.

[63] M. Liu, L. Wang, K. A. Lee, H. Zhang, C. Zeng, and J. Dang, "Exploring Deep Learning for Joint Audio-Visual Lip Biometrics," arXiv preprint arXiv:2104.08510, 2021.

[64] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "Speakingfaces: A Large-Scale Multimodal Dataset of Voice Commands with Visual and Thermal Video Streams," Sensors, vol. 21, no. 10, article no. 3465, 2021.

[65] T. Exarchos, G. N. Dimitrakopoulos, A. G. Vrahatis, G. Chrysovitsiotis, Z. Zachou, and E. Kyrodimos, "Lip-Reading Advancements: A 3D Convolutional Neural Network/Long Short-Term Memory Fusion for Precise Word Recognition," BioMedInformatics, vol. 4, no. 1, pp. 410-422, 2024.

[66] M. T. R. Sahed, M. T. I. Aronno, H. Nyeem, M. A. Wahed, T. Ahsan, R. R. Islam, et al., "LipBengal: Pioneering Bengali Lip-Reading Dataset for Pronunciation Mapping through Lip Gestures," Data in Brief, vol. 58, article no. 111254, 2025.

[67] D. Estival, S. Cassidy, F. Cox, and D. Burnham, "AusTalk: An Audio-Visual Corpus of Australian English," Proceedings of International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association, pp. 3105-3109, 2014.

[68] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6447-6456, 2017.

[69] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," arXiv preprint arXiv:1706.08612, 2017.

[70] J. S. Chung and A. Zisserman, "Lip Reading in Profile," Proceedings of British Machine Vision Conference (BMVC 2017), BMVA Press, 2017.

[71] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-Visual Speech Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8717-8727, 2018.

[72] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition," arXiv preprint arXiv:1809.00496, 2018.

[73] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, et al., "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," arXiv preprint arXiv:1804.03619, 2018.

[74] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2018), pp. 1086-1090, 2018.

[75] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, et al., "Large-Scale Visual Speech Recognition," Proceedings of Interspeech 2019, pp. 4135-4139, 2019.

[76] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, et al., "LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild," Proceedings of 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, pp. 1-8, 2019.

[77] Y. Zhao, R. Xu, and M. Song, "A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading," Proceedings of the 1st ACM International Conference on Multimedia in Asia, pp. 1-6, 2019.

[78] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, et al., "Recurrent Neural Network Transducer for Audio-Visual Speech Recognition," Proceedings of 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, pp. 905-912, 2019.

[79] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, et al., "Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection," Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4492-4496. 2020.

[80] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis," Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13796-13805, 2020.

[81] X. Chen, J. Du, and H. Zhang, "Lipreading with DenseNet and resBi-LSTM," Signal, Image and Video Processing, vol. 14, no. 5, pp. 981-989, 2020.

[82] Z. Zhang, L. Li, Y. Ding, and C. Fan, "Flow-Guided One-Shot Talking Face Generation with a High-Resolution Audio-Visual Dataset," Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Press, pp. 3661-3670, 2021.

[83] Y. Khassanov, S. Mussakhojayeva, A. Mirzakhmetov, A. Adiyev, M. Nurpeiissov, and H. Varol, "A Crowdsourced Open-Source Kazakh Speech Corpus and Initial Speech Recognition Baseline," arXiv preprint arXiv:2009.10334, 2020.

[84] E. Egorov, V. Kostyumov, M. Konyk, and S. Kolesnikov, "LRWR: Large-Scale Benchmark for Lip Reading in Russian Language," arXiv preprint arXiv:2109.06692, 2021.

[85] A. Lubitz, M. Valdenegro-Toro, and F. Kirchner, "The VVAD-LRS3 Dataset for Visual Voice Activity Detection," arXiv preprint arXiv:2109.13789, 2021.

[86] D. Ivanko, A. Axyonov, D. Ryumin, A. Kashevnik, and A. Karpov, "RUSAVIC Corpus: Russian Audio-Visual Speech in Cars," Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022), pp. 1555-1559, 2022.

[87] A. Reece, G. Cooney, P. Bull, C. Chung, B. Dawson, C. Fitzpatrick, et al., "Advancing an Interdisciplinary Science of Conversation: Insights from a Large Multimodal Corpus of Human Speech," arXiv preprint arXiv:2203.00674, 2022.

[88] M. A. Haq, S.-J. Ruan, W.-J. Cai, and L. P.-H. Li, "Using Lip Reading Recognition to Predict Daily Mandarin Conversation," IEEE Access, vol. 10, pp. 53481-53489, 2022.

[89] C. Chen, D. Wang, and T. F. Zheng, "CN-CVS: A Mandarin Audio-Visual Dataset for Large Vocabulary Continuous Visual to Speech Synthesis," Proceedings of ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 1-5, 2023.

[90] M. Anwar, B. Shi, V. Goswami, W.-N. Hsu, J. Pino, and C. Wang, "Muavic: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-to-Text Translation," arXiv preprint arXiv:2303.00628, 2023.

[91] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal Deep Learning," Proceedings of the 28th International Conference on Machine Learning (ICML 2011), Omnipress, pp. 689-696, 2011.

[92] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading Using Convolutional Neural Network," Interspeech, vol. 1, pp. 1149-1153, 2014.

[93] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with Long Short-Term Memory," Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6115-6119, 2016.

[94] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489-4497, 2015.

[95] A. Torfi, S. M. Iranmanesh, N. Nasrabadi, and J. Dawson, "3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition," IEEE Access, vol. 5, pp. 22081-22091, 2017.

[96] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-End Lipreading with Cascaded Attention-CTC," Proceedings of 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, pp. 548-555, 2018.

[97] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[98] T. Stafylakis and G. Tzimiropoulos, "Combining Residual Networks with LSTMs for Lipreading," arXiv preprint arXiv:1703.04105, 2017.

[99] X. Zhang, H. Gong, X. Dai, F. Yang, N. Liu, and M. Liu, "Understanding Pictograph with Facial Features: End-to-End Sentence-Level Lip Reading of Chinese," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9211-9218, 2019.

[100] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an Effective Lip Reading Model without Pains," arXiv preprint arXiv:2011.07557, 2020.

[101] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "Lipnet: End-to-End Sentence-Level Lipreading," arXiv preprint arXiv:1611.01599, 2016.

[102] T. Arakane and T. Saitoh, "Efficient DNN Model for Word Lip-Reading," Algorithms, vol. 16, no. 6, article no. 269, 2023.

[103] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal Convolutional Networks: A Unified Approach to Action Segmentation," Proceedings of European Conference on Computer Vision, Springer, pp. 47-54, 2016.

[104] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading Using Temporal Convolutional Networks," Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 6319-6323, 2020.

[105] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-Reading with Densely Connected Temporal Convolutional Networks," Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, pp. 2857-2866, 2021.

[106] K. Prajwal, T. Afouras, and A. Zisserman, "Sub-word Level Lip Reading with Visual Attention," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5162-5172, 2022.

[107] Z. Li, T. Lohrenz, M. Dunkelberg, and T. Fingscheidt, "Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention," Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, pp. 723-730, 2023.

[108] Y.-H. Park, R.-H. Park, and H.-M. Park, "Swinlip: An Efficient Visual Speech Encoder for Lip Reading Using Swin Transformer," Neurocomputing, vol. 639, article no. 130289, 2025.

[109] L. Courtney and R. Sreenivas, "Learning from Videos with Deep Convolutional LSTM Networks," arXiv preprint arXiv:1904.04817, 2019.

[110] C. Wang, "Multi-Grained Spatio-Temporal Modeling for Lip-Reading," arXiv preprint arXiv:1908.11618, 2019.

[111] M. Luo, S. Yang, S. Shan, and X. Chen, "Pseudo-Convolutional Policy Gradient for Sequence-to-Sequence Lip-Reading," Proceedings of 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, pp. 273-280, 2020.

[112] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual Information Maximization for Effective Lip Reading," Proceedings of 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, pp. 420-427. 2020.

[113] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation Flow Based Two-Stream Network for Lip Reading," Proceedings of 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, pp. 364-370, 2020.

[114] 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," Proceedings of 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), IEEE, pp. 356-363, 2020.

[115] D. Ivanko, D. Ryumin, A. Kashevnik, A. Axyonov, and A. Karnov, "Visual Speech Recognition in a Driver Assistance System," Proceedings of 2022 30th European Signal Processing Conference (EUSIPCO), IEEE, pp. 1131-1135, 2022.

[116] A. Koumparoulis and G. Potamianos, "Accurate and Resource-Efficient Lipreading with Efficientnetv2 and Transformers," Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 8467-8471, 2022.

[117] P. Ma, Y. Wang, S. Petridis, J. Shen, and M. Pantic, "Training Strategies for Improved Lip-Reading," Proceedings of ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, pp. 8472-8476, 2022.

[118] H. Chen, W. Li, Z. Cheng, X. Liang, and Q. Zhang, "TCS-LipNet: Temporal & Channel & Spatial Attention-Based Lip Reading Network," Proceedings of International Conference on Artificial Neural Networks, Springer Nature Switzerland, 2024.

[119] H. Wang, B. Cui, Q. Yuan, G. Pu, X. Liu, and J. Zhu, "Mini-3DCvT: A Lightweight Lip-Reading Method Based on 3D Convolution Visual Transformer," The Visual Computer, vol. 41, no. 3, pp. 1957-1969, 2025.

[120] X. Wu, Z. Tan, Z. Cheng, and Y. Ru, "STDNet: Improved Lip Reading via Short-Term Temporal Dependency Modeling," Virtual Reality & Intelligent Hardware, vol. 7, no. 2, pp. 173-187, 2025.

[121] X. Zhang, J. Sun, P. Zhu, T. Xiao, M. Lao, and Y. Guo, "Mamba-Based Temporal Modeling for Event-Based Lip Reading," Proceedings of 2025 6th International Conference on Computer Vision, Image and Deep Learning (CVIDL), IEEE, pp. 926-930, 2025.

[122] M. Thomas, E. Fish, and R. Bowden, "VALLR: Visual ASR Language Model for Lip Reading," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, pp. 2846-2856, 2025.

[123] R. Tapu, B. Mocanu, and I.-C. Chiva, "Multimodal Visual Speech Recognition for Under-Resource Languages via Cross-Modal Learning and Large Language Models," Science and Technology (ROMJIST), vol. 29, no. 1, pp. 53-64, 2026.