Computer Vision and Image Understanding 233 (2023) 103738

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Analyzing lower half facial gestures for lip reading applications: Survey on vision techniques

Preethi S.J. *, Niranjana Krupa B.

Department of ECE, PES University, Bangalore, 560085, India

ARTICLE INFO

Communicated by Nikos Paragios

Keywords:
Lip reading
Audio visual speech recognition
Silent speech recognition
Voice from lips
Lip HCI
Machine learning
Deep learning

ABSTRACT

Lip reading has gained popularity due to the proliferation of emerging real-world applications. This article provides a comprehensive review of benchmark datasets available for lip-reading applications and pioneering works that analyze lower facial cues for lip-reading applications. A comprehensive review of lip reading applications is broadly classified into five distinct applications: Lip Reading Biometrics (LRB), Audio Visual Speech Recognition (AVSR), Silent Speech Recognition (SSR), Voice from Lips, and Lip HCI (Human–computer interaction). LRB entails extensive research in the fields of authentication and liveness detection. AVSR covers key findings that have contributed significantly to applications such as voice assistants, video-to-text transcription, hearing aids, and pronunciation-correcting systems. SSR analyzes the efforts made for silent-video-to-text transcription and surveillance camera applications. The voice from lips section discusses applications such as voice for the voiceless and vision-infused speech inpainting. In lip HCI, LR-HCI for smartphones, smart TVs, computers, robots, and musical instruments is reviewed in detail. Comprehensive coverage is given to cutting-edge techniques in computer vision, signal processing, machine learning, and deep learning. The advancements that aid the system in learning to lip-read and authenticate lip gestures, generate text transcription, synthesize voice based on lip movements, and control systems via lip movements (lip HCI) are covered. The work concludes by highlighting the limitations of existing frameworks, the road maps of each application illustrating the evolution of techniques employed over time, and future research avenues in lip-reading applications.

1. Lip reading approaches and applications

Applications in Lip Reading recognize speech by analyzing the tongue, teeth, lips, and lower face when speech sound is unavailable or unclear (Fernandez-Lopez and Sukno, 2018). Continuous automatic real-time Lip reading systems have aroused a great deal of interest in recent years due to their vast applications in authentication, human–computer interaction devices, a hearing aid for the deaf, voice assistants in noisy environments, spy surveillance cameras, pronunciation correcting systems, and health care applications such as synthesizing voice for unable-to-talk patients from their mouth gestures. Due to homophonemes such as </p/, and/b/>, </k/, and/g/>, that possess distinct sounds but the same mouth shape, lip reading is a particularly difficult task. Mapping from many to one among phonemes and visemes makes mapping viseme to phoneme extremely difficult. Additionally, head position variations of the speaker make lip reading even more challenging. Consequently, pose-invariant lip reading is difficult to achieve. Changes in lighting conditions and poor spatial and temporal resolutions also impact the performance of lip-reading applications. The duration of each word utterance varies, so speech detection must be independent of the rate at which words are spoken. Additionally, lip movements during respiration and laughter can affect lip-reading accuracy. Furthermore, speaking styles vary between individuals, rendering speaker-independent lip-reading impractical. Recognizing words and silence in continuous speech is challenging. Numerous organs, including the tongue, teeth, lips, jaw, vocal cords, and lungs, contribute to human speech production. It is impossible to depict all these organs' motions using only the face's lower half. Table 1 provides a list of all abbreviations used in the manuscript.

The evolution of lip reading is analyzed in three streams based on a thorough literature review. The first stream leverages electromyography(EMG) (Shaikh et al., 2010), electropalatography (EPG) (Gilbert et al., 2010), electromagnetic articulography (EMA) (Heracleous and Hagita, 2011), triboelectric sensors (Lu et al., 2022), and RFID sensors (Zhang et al., 2022a) to capture the lip movements. The second stream examines the Doppler effect of inaudible acoustic (Gao et al., 2020) and ultrasonic (Tan et al., 2017) signals' influence on the articulation movements of speech (also known as acoustic echo and ultrasonic echo). Smartphones primarily employ the second method, while the

* Corresponding author.
E-mail address: preethisj@pes.edu (Preethi S.J.).

https://doi.org/10.1016/j.cviu.2023.103738
Received 14 November 2022; Received in revised form 26 April 2023; Accepted 18 May 2023
Available online 24 May 2023

Table 1
Glossary of the abbreviations used in the manuscript.

Abbreviation	Definition	Abbreviation	Definition
2D	Two dimensional	LDA	Linear Discriminant Analysis
3D	Three dimensional	LF-MMI	Lattice-free maximum mutual information
3DMM	3D Morphable Model	LMFB	Log mel-scale filterbank
AAM	Active Appearance Models	LMFE	Local motion feature extraction
ACN	Acoustic-Conditional-Network	LPC	Linear-Predictive-Coding coefficients
ALSOS	Alternating Spatiotemporal and Spatial Convolutions	LPCC	Linear-Prediction-Cepstral-Coefficients
ANNs	Artificial Neural Networks	LR	Lip-reading
ASGM	Adaptive spatial graph Convolutional network	LRB	Lip Reading Biometrics
ASM	Active-Shape-Model	LSF	Line spectral frequencies
ASST-GCN	Adaptive-Semantic-Spatio Temporal Graph Convolutional Network	LSP	Line Spectral Pairs
AV	Audio-Visual	LSTM	Long short-term memory
AVSR	Audio Visual Speech Recognition	MBH	Motion Boundary Histograms
BERT	Bidirectional-Encoder-Representations from Transformers	MCD	Mel Cepstral Distance.
Bi	Bidirectional	MCT	Modified Census Transform
Bi-GRU	Bidirectional Gated Recurrent Unit	MFCC	Mel-frequency cepstral coefficient
Bi-VAE	Bimodal Variational Autoencoder	MFM	Message Flow module
C3D–P3D	Deep 3DCNNs (C3D) - pseudo-3D (P3D)	ML	Machine learning
CFD	Chicago Face Dataset	MLLT	Maximum likelihood linear transform
CGAN	Coupled GAN	MLP	Multi-Layer Perceptron
CLM	Constrained Local Model	MMST	Multimodal sparse transformer network
Conv2D	2D Convolutional layer	MS	Multi-scale
Conv3D	3D Convolutional layer	MSTP	Multi-grained ST Features Perceived Network
ConvNet or CNN	Convolutional Neural Networks	MVM	Multi-head VA memory
Corr2D	2D correlation between the actual spectrogram and the reconstructed spectrogram	NN	Neural Network
CSAM	Combined Shape and Appearance Model	OF	Optical flow
CTC	Connectionist Temporal Connection	OOV	Out of Vocabulary
DAE	Convolutional Deep Autoencoders	PCA	Principal Component Analysis
DAE-Net	Dynamic Enhanced Authentication Network	PER	Phoneme Error Rate,
DBNFs	Deep Bottle Neck Features	PESQ	Perceptual Evaluation of Speech Quality,
DBNs	Dynamic Bayesian Networks	PGGAN	Progressive Growing GAN
DCGAN	Deep Convolutional GAN	PLP	Perceptual-Linear-Prediction coefficients
DCT	Discrete Cosine Transform	PPLFA	Penn Phonetics Lab Forced Aligner
DC-TCN	Densely connected temporal Convolutional networks	RC-Net	Representative Descriptor extraction Classification Network
DDBNs	Deep Dynamic Bayesian Networks	RFCs	Random Forest Classifier
DFN	Deformation flow network	RM	Resource Management
DINO	Dual Inverse Network Optimization	RNN	Recurrent Neural Network
DL	Deep learning	ROI	Region of Interest
DNN	Deep Neural Networks	RSM	Random Subspace Method
DSS	Data sharing scheme	SA	Squeeze and attention
DS-TCN	Depthwise-Separable Temporal ConvNet	SA-DTH-Net	Speaker-Authentication Dynamic-Talking-Habit Network
DTW	Dynamic-Time-Wrapping	SCS	Shifted Cosine Similarity
DWT	Discrete Wavelet Transform	SD	Speaker dependent
EER	Equal-Error-Rate	seq2seq	sequence to sequence
EMA	Electromagnetic articulography	SID	Speaker independent
EMD	Empirical Mode Decomposition	SII	Speech Intelligibility Index,
EMG	Electromyography	SPE	Spatial position encodings
EPG	Electropalatography	SSR	Silent Speech Recognition
E-TDCNN	Extended time delay neural network	ST	Spatio-temporal
FAR	False-Accept-Rate	STFM	ST-fusion framework
Fast-LTS	Fast-Lip-to-Speech	STFT	Short-time Fourier transform
FFE-Net	Feature Extraction Network	ST-GCN	Spatial graph Convolutional network followed by Temporal graph Convolutional network
fMLLR	feature space Maximum-Likelihood-Linear-Regression	STMI	Spectro Temporal Modulation Index,
FO	Fundamental frequency	STOI	Short Time Objective Intelligibility, ESTOI
FRR	False-Rejection-Rate	SURF	Speeded-Up-Robust Features
FSTs	Finite-state-transducers	SVD	Singular Value Decomposition
GANs	Generative adversarial networks	SVM	Support Vector Machine
GAP	Global-Average-Pooling	TCN	Temporal convolutional network
GFCC	Gammatone Frequency Cepstral Coefficients	TDNN-F	Factored time delay NNs
GMMs	Gaussian Mixture Models	TF block	Temporal-Focal block
HCI	Human–computer interaction	TPS	Thin plate spline transformation
HiLDA	Hierarchical Linear Discriminant Analysis	U/V	Unvoiced/Voiced
HMMs	Hidden Markov Models	VAE	Variational Autoencoder
HOG	Histogram-of-Oriented-Gradients	VAM	Visual Audio Memory
HOGTOP	HOGs extracted from X-Y, X-Time, and Y-Time axis	VCA-GAN	Visual Context Attention GAN
HTER	Half-Total-Error-Rate	ViT	Vision transformer
I3D	Deep-inflated-3D-CNN	VQ	Vector quantization
ISA	independent subspace analysis	VSR	Visual speech recognition
KD	Knowledge Distillation	VTM	Variational Temporal Mask module
KLD	Kullback–Leibler divergence	VTP	Visual Transformer Pooling
KNN	k-Nearest Neighbors	VV	Visual Voice
LBP	Local Binary Pattern	WER	Word Error Rate,
LBPTOP	LBPs extracted from X-Y, X-Time, and Y-Time axis	WLAS	Watch-Listen-Attend-Spell
LCFE	Local co-ordinate feature extraction	XCS-LBP	Extended-Center-Symmetric Local Binary Pattern

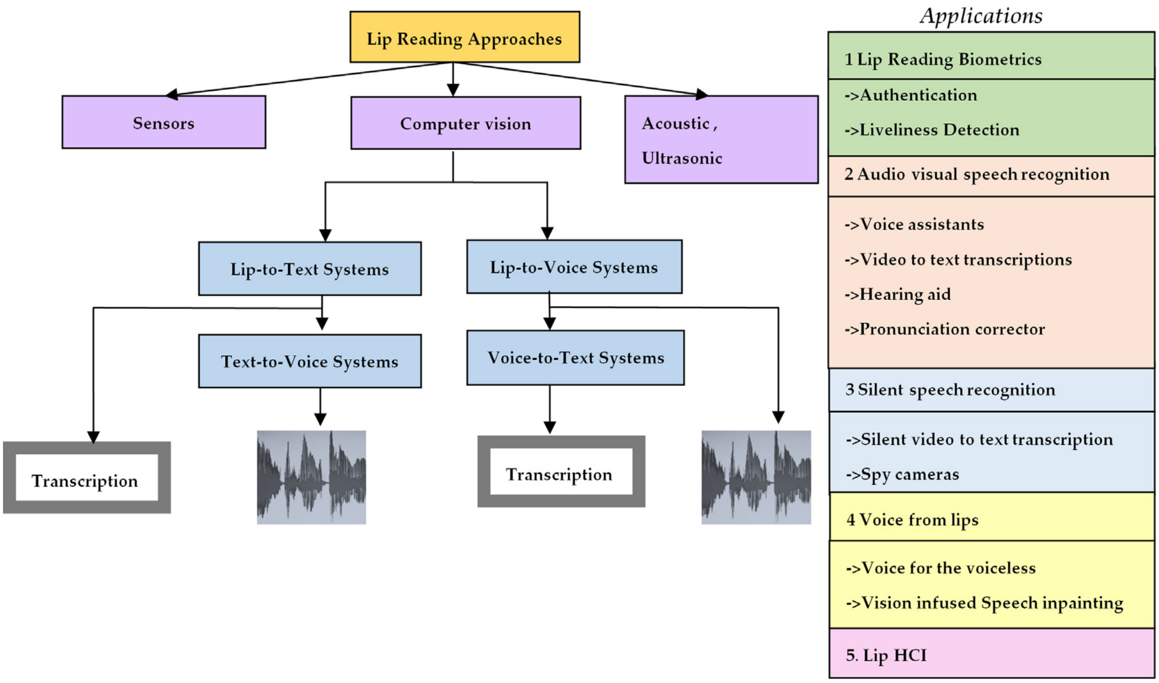

Fig. 1. Introducing taxonomy for lip-reading approaches and applications.

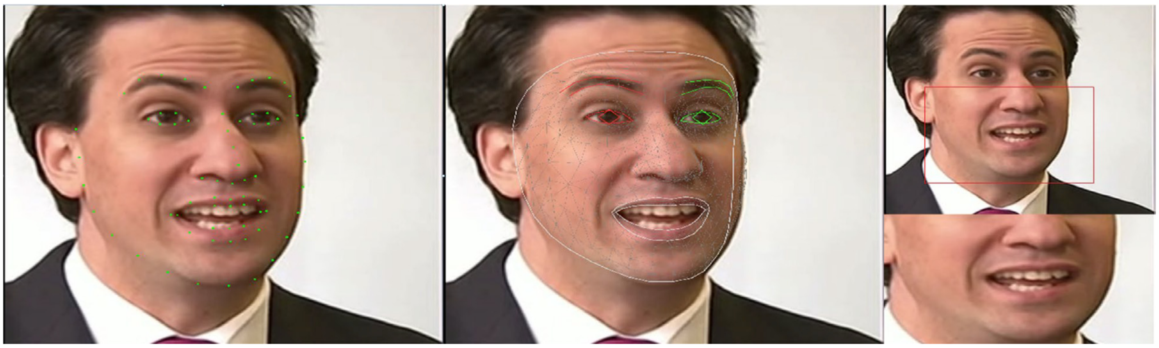

Fig. 2. (a) Dlib landmarks; (b) Mediapipe landmarks; (c) Extraction of ROI (Lower half face).

third stream is a computer-vision-based approach in which a camera captures and analyzes the movements of the lips to interpret speech.

Fig. 1 depicts the taxonomy of approaches and applications for lip-reading. Work on lip reading is broadly categorized into five distinct applications: Lip Reading Biometrics (LRB), Audio Visual Speech Recognition (AVSR), Silent Speech Recognition (SSR), and Voice from lips and Lip HCI (Human–computer interaction). Computer or machine vision, signal processing, machine learning, and DNN-based tasks are considered for each application. Lip-to-Text systems, which receive video as input and output text transcription, and Lip-to-Voice systems, which receive video as input and then anticipate voice, have made significant progress in machine vision for lip reading. The output of Lip-to-Text systems can be converted to voice using Text-to-Voice systems, and the output of Lip-to-Voice systems can be converted to text transcription.

For the computer-vision-based approach, the camera captures the movements of lips and is analyzed to recognize the speech. Extraction of Region of Interest (ROI) after mouth detection is the initial step in almost all vision-based methodological approaches. Fig. 2 depicts the extraction of ROI from a sample LRW500 video frame. Pretrained models such as DLIB's 68 points 2D face landmarks (King, 2009) and Mediapipe's 468 points 3D face landmarks (Lugaresi et al., 2019) assist lip reading researchers in extracting ROI. Points 48 to 68 on the DLIB landmarks map the mouth region.

In some works, ROIs are augmented in order to perform illumination- and pose-invariant lip readings. Lip motion video frames can be input into lip reading frameworks in multiple ways. In a 2D image format, video frames can be cascaded into a single frame or fused into a single dynamic image that describes the motions in all frames. Deep lip reading frameworks can accept RGB frames, optical flow, Extended-Center-Symmetric Local Binary Pattern (XCS-LBP), contour images, saliency maps, or any transformed RGB frames. In machine learning approaches, feature extraction, dimensionality reductions, and classification follow ROI extraction. The preprocessing and feature extraction phases utilize numerous computer vision and signal processing techniques. In deep learning, feature extraction is implicit during the classification stage. Synthesis of speech is possible with deep learning networks, including Generative adversarial networks (GANs) and Autoencoders. The literary works of Shorten and Khoshgoftaar (2019) looked into the current data augmentation techniques used to avoid overfitting problems in deep learning along with other methods like dropout, batch-normalization, transfer learning, pre-training, and one-shot learning. This work summarizes image data augmentations using basic image manipulations, ML, and DL techniques. Basic image manipulation techniques include rotation, flipping, shearing, color jitter, random cropping, translation, noise injection, contrast manipulation, brightness manipulation, kernel filters (image blurring or sharpening), linear and nonlinear image mixing, and random erasure techniques. Employing k-Nearest Neighbors, the Synthetic Minority Over Sampling technique interpolates new synthetic samples from existing samples. Deep learning strategies for data augmentations consist of Generative adversarial techniques and Neural-style transfer. GANs comprehend the trained images very well and have developed an ability to imagine alterations to them in the same way humans do, and they generate new images with alterations. Neural style transfers identify the style of an image and then transfer it to another image while preserving its original content, thereby generating a new augmented image.

The impressive survey articles enumerated in Table 2 provide the overall summary of computer vision methodologies, signal processing, ML, and DL approaches involved in lip reading. However, their research has yet to investigate the extensive literature on all lip-reading techniques and applications. Some articles have discussed only ML or DL types. Their survey only includes studies conducted in the year 2022, except two recent survey papers that include 2022 articles but only cover some lip reading applications. In most review papers, the number of articles cited is also limited. Consequently, we intend to continue the survey by incorporating five lip reading applications. From those mentioned above, it is found to be helpful to review literature works that are employed in all lip-reading applications and exploit various signal processing, machine vision, machine learning, and DNN approaches. It is of utmost importance to carefully analyze the progress of lip-reading technologies in all applications and identify their primary challenges and unresolved issues impeding their development. Our work is guided by the shortfall of similar survey articles found in the literature. Thus, the objective is to compensate for this incompleteness. This review article's primary goal is to present an overall strategy of the innovative approaches that carry out the lip-reading tasks in several lip-reading application areas and to provide a pathway for further research in this challenging area. Research works are all sorted year-wise along with paradigm tailored, including several ways to input modalities into the model, audio–video features, fusion strategies, front-end/back-end (classifiers/encoders/decoders), benchmark datasets used for training and testing, and results and limitations.

In early literary works, Kandagal and Udayashankara (2014) analyzes past ML research and development in AVSR technologies, the fundamental block diagram of an AVSR, and a few datasets with small vocabulary sizes. Katsaggelos et al. (2015) describes audio and video feature extractions and AV fusion techniques, such as SVMs, DBNs, HMMs, and Kalman filters. They also discuss the application of deep learning (DL) to AV fusion, such as multimodal fusion learning that combines features or scores from audio and video modalities to improve speech recognition performance, cross-modality learning that uses a single input autoencoder and transfers knowledge learned from the video to enhance audio performance, and shared-representation learning that uses two input autoencoders to learn a joint representation that is compatible with both audio and video modalities. Burton et al. (2018) conduct a systematic study of experiments with the TCD-TIMIT benchmark employing traditional and DL methods to provide a series of SI benchmarks and demonstrate that convolutional features classified by SVM outperform others. Authors Fernandez-Lopez and Sukno (2018) list all audio-visual corpora, describing their applications, popularity, and key features like the total subjects, vocabulary size, recording environments, and recording length. They compare AVSR and SSR systems research from 2007 to 2017, showing the shift from traditional approaches to end-to-end DL architectures. Addarrazi et al. (2020) summarizes early, intermediate, and late fusion strategies for AVSR. In Hao et al. (2020) extensive review is carried out, and the frameworks for both conventional and DL methods are presented and discussed systematically. A comprehensive analysis of LR databases were presented, including the motivation behind their creation, the dataset size, the total number of participants, the total number of instances, camera resolution, and the highest accuracy attained with these databases. LR issues, their solutions, and research directions are analyzed.

In recent review papers, Pujari et al. (2021) provide an overview of very few DL techniques and datasets, highlighting their performance and limitations. Radha et al. (2021) talks about various visual features such as image pixel-based, transform-based, shape-based, appearance-based, motion-based, and color-based features. The paper also focuses on AV fusion strategies such as feature-level fusion (early fusion), decision-level fusion (late fusion), model-level fusion (middle-level fusion), hybrid fusion (combination of early, middle-level, and late fusion), and very few AV corpora. The article Fenghour et al. (2021) describes the AV-benchmark datasets and the various DL architectures used for lip reading, including autoencoders, RBMs, 2DCNN like VGG Network, AlexNet, Network in Network, and GoogleLeNet, 3DCNN, (2D+3D) CNN, LSTM, GRU, Attention, CTC, transformers, and TCN. New frameworks such as visual voice memory, knowledge distillation methods, graph neural network, and two-stream networks still needed to be included. Pu and Wang (2022) explain LR datasets, traditional techniques, and deep LR techniques up until 2021. And they have discussed a few aspects to focus on while developing LR models, such as which ROI to take into account (full face, lip, or half face), building datasets closer to the real-world environment, and combining LR with the language model. The authors of Choudhury et al. (2023) have investigated AVSR's audio features, visual features, feature fusion, and decision fusion techniques. In Sheng et al. (2022a), the authors summarize deep learning-based approaches for silent speech recognition and silent speech generation (speech-to-video) until 2022. It discusses the primary challenges, data benchmarks, performance metrics, architecture of renowned methods, unresolved issues, and potential future research directions. Yang et al. (2023) presents three latest AVSR models with the highest performance on the LSR2 and LRS3 benchmarks.

To summarize the key contributions in this article, this is the first review paper to systematically and exhaustively analyze all LR-related applications, including their significant challenges, frameworks, benchmark datasets, performance measures, and limitations. Frameworks for each lip-reading application are presented. Every application is unique and requires a unique approach; each has its own challenges and limitations. Open research queries, road maps to show the growth of techniques implemented across time, and potential future research directions in lip-reading applications are discussed.

Various keyword searches were performed on the Google Scholar electronic database to locate the pertinent literature. Only pages written in the English language were investigated. In addition, only doctoral dissertations, articles published in Scopus and Web of Science-indexed journals, and conference proceedings were considered. A comprehensive screening of titles and abstracts followed. Studies that were not closely related to the LR applications were excluded. LR using sensors, ultrasound, or acoustic signals was excluded, and only the articles that employ the computer vision methodology were included. Investigations were conducted to determine the application areas for which the authors proposed their work, such as authentication (Yang et al., 2020), liveness detection (Zhou et al., 2021), voice assistants (Ivanko et al., 2022b), video-to-text transcriptions (Chung and Zisserman, 2017), hearing aid (Abrar et al., 2019; Kumar et al., 2022), pronunciation corrector (Yu, 2022), silent video-to-text transcription (Chen et al., 2020), spy cameras (Rothkrantz, 2017), voice for the voiceless (Hong et al., 2021), vision infused speech inpainting (Morrone et al., 2021), and lip HCI (Shin et al., 2015). After analyzing the applications, it was found that LR applications can be grouped into five categories: Lip Reading Biometrics (LRB), Audio Visual Speech Recognition (AVSR), Silent Speech Recognition (SSR), Voice from lips, and Lip HCI (Human–computer interaction). Then, for each application type's computer vision or machine vision, signal processing, machine learning, and DNN-based tasks were pursued in detail. To accomplish this, a search was conducted to identify five application-related articles and datasets.

To locate works related to LR benchmark datasets, the search terms [“audio” AND “visual speech recognition” AND “dataset” OR “corpus”

Table 2
Summary of review papers.

Year	Ref.	Apps.	Scope	Datasets discussed	ML types discussed	DL types discussed	Survey till year	No. of papers ref.	Key findings
Nov 2014	Kandagal and Udayashankara (2014)	ASR, AVSR	Audio video feature extraction techniques and classifiers	Very few with little detail	Very few	no	2014	30	The HMM-based AVSR system is the subject of extensive investigation.
Sep 2015	Katsaggelos et al. (2015)	AVSR	Audio visual fusion techniques	Very few with little detail	moderate	moderate	2015	171	Weighting AV streams, and synchronizing A & V helps to improve accuracy.
Dec 2018	Burton et al. (2018)	SI-SSR	Speaker independent lip reading models	Only TCD-TIMIT	Very few	Only ResNet50 and LSTM discussed	2018	50	Convolution features classified by SVM is the best model.
Oct 2018	Fernandez-Lopez and Sukno (2018)	SSR	Traditional and DL approaches to automatic lip reading.	Comprehensive analysis of the datasets.	Extensive analysis of ML types.	moderate	2018	178	DL architectures vastly enhance the performance of SSR.
April 2020	Addarrazi et al. (2020)	AVSR	Audio visual fusion techniques	no	Very few	no	2018	22	AVSR systems fusion stages.
Nov 2020	Hao et al. (2020)	SSR	ML and DL types for lip reading.	Comprehensive analysis of the datasets.	Extensive analysis of ML types.	Extensive analysis of DL types.	2020	209	The development of a realistic database and multi-person scene LR is an important area for future research. Current research is focused on offline LR, Accuracy of online LR will be less.
Feb 2021	Pujari et al. (2021)	SSR	DL types for lip reading	Very few	One or two	Very few	2020	25	CNN+BiLSTM is the most advanced model available for LR.
March 2021	Radha et al. (2021)	AVSR	Visual feature extraction methods and AV fusion at different levels.	Very few	Very few	no	2019	55	The majority of articles utilize decision-level fusion as opposed to feature-level fusion.
Aug 2021	Fenghour et al. (2021)	SSR	DL types for lip reading	Comprehensive analysis of the datasets.	no	Extensive analysis of DL types.	2021	141	2D+3D CNNs are the preferred front-end; Transformers and TCNs are replacing LSTMs for the back-end; and the prediction of unseen words, SI-LR, and the generalization of videos with varying frame rates and resolutions are unresolved issues.
June 2022	Pu and Wang (2022)	SSR	Traditional and DL LR techniques.	Comprehensive analysis of the datasets.	few	moderate	2021	118	Future research would focus on the development of lightweight LR models and multimodal LR models with language models.
Feb 2023	Choudhury et al. (2023)	AVSR	Extraction of audio and visual features, as well as their integration techniques.	Moderate review on datasets.	Extensive analysis of ML types.	few	2022	77	Deep learning seems to be a potential AVSR solution.
May 2022	Sheng et al. (2022a)	SSR,Visual speech generation(VSG)	Visual speech recognition and visual speech generation.	Review of frequently employed datasets.	no	Exhaustive analysis of the most recent DL types.	May 2022	200	Future research will focus on developing contact visual speech sensors for LR, learning models with fewer labels, and multilingual LR.
April 2023	Yang et al. (2023)	AVSR	Three latest AVSR models with the highest performance on the LSR2 and LRS3 benchmarks.	LRW, LRS2, LRS3	no	Three latest DL types for AVSR	March 2023	25	MoCo v2 + word2vec is the best AVSR model.
Ours	–	LRB, AVSR, SSR, Voice from lips, Lip HCI	Five lip reading applications, LRB, AVSR, SSR, Voice from lips, and Lip HCI.	Comprehensive analysis of the datasets.	moderate	Extensive analysis of DL types.	March 2023	213	They are mentioned in the Discussion and Summary sections.

OR “benchmark” AND “English”] were employed. Similarly, for LR applications the articles published between 2012 and 2022 were searched for. The keywords [-“ultrasonic” AND “lip reading” OR “visual speech recognition” AND “authentication” OR “biometrics”] were utilized to locate articles on Lip Reading Biometrics. The terms [“audiovisual speech recognition” OR “audio visual speech recognition”] were used to find articles about AVSR. The search terms [“lip reading” OR “silent speech recognition” OR (“visual speech recognition” -“audio”)] were used to locate articles pertaining to silent speech recognition. The search terms [“lip to speech” OR “speech reconstruction from the silent video”] were used to locate articles relating to Voice from Lips. Using the search terms [“lip reading” AND “HCI” OR “silent speech interface”], articles pertaining to lip HCI were identified.

The structure of the review paper proceeds, as indicated in Fig. 1, Section 2, provides an overall summary of the available benchmark datasets used for Lip reading applications. Section 3 reveals the studies on lip-reading authentication and liveness detection systems. Section 4 summarizes the works on how lip-reading and speech aids speech detection in noisy environments for applications like voice assistants, video-to-text transcriptions, hearing aid, and pronunciation correcting systems. Section 5 provides an overview of how lip-reading alone deals with silent video text transcription and spy camera systems. Section 6 represents the cutting-edge strategies used in the voice for unable-to-talk patients and vision-infused speech inpainting. Section 7 analyzes how lip reading is used for Human–computer interaction (HCI) in smartphones, smart TVs, computers, robots, and musical instruments. Finally, Section 8 provides a discussion and recommendations for lip reading applications, while Section 9 presents trends for future advancements in the lip reading field and the review’s conclusions.

2. Lip-reading benchmark datasets

Over the last few years, most eminent lip-reading and audiovisual speech recognition research have been undertaken using supervised learning algorithms, where large annotated data is required for training. Furthermore, deep learning algorithms are more data-hungry. Several benchmark datasets are available for recognizing alphabet, digits, words, and sentences. Most of the datasets are constrained and contain recorded frontal views. Only some datasets contain multi-view video recordings using multi-cameras, which are placed in different directions to capture several views of the visual speech. Some of the unconstrained datasets are constructed from youtube videos, BBC News, and TED talk videos which contain many speakers in multiple-head poses. Some datasets consist of data captured using other types of cameras. For example, the MIRACL-VC1 benchmark includes the depth images recorded from the depth camera along with the RGB images, the SpeakingFaces dataset includes thermal images along with video and audio data, and the DVS-lip benchmark consists of data gathered using an event camera. Benchmark datasets available for lip-reading and audio-visual-speech in English are tabulated in Table 3. Several research works have used their custom datasets.

3. Lip reading biometrics

Authentication can utilize the physiology and behavior of the face’s lower half, including the mouth, cheeks, and chin. The physiology of the face’s lower half has no temporal information. The distinct geometric shape, texture, intensity, and color of the visible wrinkles on the chin, cheeks, and lips serve as defining physiological features. Mouth prints can be utilized for the biometric trait (Chowdhury et al., 2022). Still, illumination conditions, applied cosmetics such as lipstick, certain deformations caused by medical conditions such as dry lips, and deformations due to aging can affect the physiology of the mouth. The presence of a mustache and beard can alter the appearance of the face’s lower half. In one of the fascinating works (Farrukh and van der Haar, 2023), the physiology of the lip was validated for biometric authentication using the Chicago Face Dataset (CFD), which consists of 597 individuals of various ethnicities. Using the VGG16 and VGG19 frameworks, they attained 91 and 93 percent validation accuracy, respectively. Changes in mouth shape and texture due to movements of the lips, tongue, teeth, chin, and cheeks form the behavioral lower-half face features. Behavioral features are a function of time and thus include temporal information. The horizontal separation between the user and the camera and variations in head pose can affect behavioral features. A lip-motion password is a behavioral feature used for human authentication. It consists of a password that is latent in the speaker’s utterances.

As the lip password is effective against man-in-behind, brute force, spoofing, guessing, and user impersonation attacks, it is gaining momentum in the authentication process. When using a lip gesture as a password, an authentication system must deal with four scenarios. The authorized user whispers the correct password in the first scenario. In the second scenario, the authorized user whispers the incorrect password. In the third scenario, the unauthorized user whispers the valid password, while in the fourth scenario, the unauthorized user whispers the wrong password. Then, the system must perform strong authentication only for the first scenario. Fig. 3 depicts the procedure for lip reading biometrics. During the training phase, a lip-password feature extractor or classifier template is created for each client based on video samples obtained from the client. The extracted Client-N features, or classifiers, are then stored on the template server. During the testing phase, a video of the client’s lip password is fed to the Client-N feature extractor or classifier, whose access must be authenticated. A decision is made according to matching criteria. If both templates match, the client is authenticated; otherwise, access to the system is denied. Most studies have employed an equal error rate (EER) and a half-total error rate (HTER) to evaluate the performance of authentication models. At EER, the false-accept rate (FAR) is the same as the false-rejection rate (FRR); the mean of FAR and FRR is HTER, and few works have used the word “accuracy” or “error rate” as a metric for evaluation.

In earlier works, lip reading primarily relied on motion estimation. A study of block match algorithms (Alghathbar and Mahmoud, 2009), three-step search, four-step search, and diamond search on an audio-visual tulips-1 dataset were carried out to estimate eight motion curves in eight geographical directions to represent features of the utterance of a word. The database stored motion features extracted during the training phase from the corpus of word utterances. During recognition, the test video’s motion features were compared with the corpus’s motion features in the database using the mean square error. Lip-reading password system was built by Sengupta et al. (2012) without requiring laborious training procedures. Sixteen RGB frames of the client’s silent password were encrypted and saved in the database during the registration. During the authentication, thresholding was applied to the red planes of captured reference images and decrypted stored images. Then, their normalized difference image was compared to the minimum allowable difference threshold, and access was granted if the value was lower than a threshold. However, it failed to function in various lighting conditions. Furthermore, lip-reading was used to authenticate a mobile banking application that recognizes characters and then combines them to identify passwords. However, they could not achieve good accuracy on the LiLiR dataset (Lesani et al., 2015). Using a thresholding operation, one of the earliest biometric authentications (Shubhangi and Balbhim, 2017) extracted the pink portion of the lips and created a lips template for authentication. Then the closest match between the user’s spoken password template and lips templates were determined using hamming distance.

A few fascinating works employ ML strategies. Liu and Cheung (2014) used multi-boosted Hidden Markov Models (HMMs) with Adaboost learning to authenticate lip-password digits. The HMMs were trained on principal component analysis (PCA) features, discrete cosine transform (DCT) features, and geometric features. Here, the Random Subspace Method (RSM) scheme was utilized for feature selection, and

Table 3
Benchmark English lip-reading/audio-visual-speech dataset.

Dataset Names	Description and Relevance	Contents	Multiview	Data Capture
Tulips (Movellan, 1994; Potamianos et al., 2003)	It is among the earliest and simplest corpus created for digit recognition. It includes 96 grayscale frame sequences of 12 speakers (9M, 3F) pronouncing the 1st four English digits two times each. Videos were sampled at 30 fps with the mouth region cropped to 100 × 75 pixels.	Digits	Frontal	Camera
XM2VTS (Messer et al., 1999)	It includes digital video recordings of resolution 720 × 576 pixels and 25 fps of 295 individuals of varying ages. To capture variation in hairstyle, mood, and attire, data were collected in four sessions separated by one month, each recording two instances of ten digits. Various head positions were used to collect data. The dataset contains monophonic 16-bit 48-kHz audio recordings. XM2VTS is more focused on individual authentication.	Digits	−90,−45, 0, 45, 90,look up, and look down	Camera (Diverse head positions of the speaker)
AVLetters (Matthews et al., 2002)	Consists of three instances of 26 isolated alphabets (A-Z) spoken by ten individuals (5M+5F). Each instance of the dataset contains cropped grayscale lip regions with a resolution of 80 by 60 pixels, including a time resolution of 25 fps. The sound was recorded with a sampling frequency of 22.05 kHz and a bit depth of 16 bits.	Alphabets	Frontal	Camera
CUAVE (Patterson et al., 2002)	CUAVE contains 720 × 480, 30 fps, single-camera recordings with 36 speakers (17 female, 19 male) saying digit sequences in frontal and two profile views. The recording was carried out in two sessions. In sessions with a single speaker, the speaker stood naturally in front of the camera while pronouncing 50 isolated digits. The speaker was then recorded from two profile angles while pronouncing twenty-265 isolated digits. This was followed by capturing 60 combined digits utterances in a frontal pose. Two speakers were recorded simultaneously for dual-speaker sessions; while one spoke, the other remained silent, but the camera captured both. Speakers were required to repeat two combined-digit sequences twice alternately. The dataset may be utilized for multi-speaker and speaker-independent testing. The dataset contains monophonic 16-bit 16-kHz audio recordings.	Digits	−90, 0, 90	Camera
VIDTIMIT (Sanderson, 2002)	It consists of AV recordings of 43 speakers (24 M,19 F) reciting 10 TIMIT sentences each. The recordings were carried out with head pose variation towards the right and towards the left. The data were recorded throughout three sessions, separated by one week to accommodate changes in the speaker's voice, makeup, dress, and mood. The recording was conducted with varying camera zoom, acoustic noise, and office atmosphere. The videos were captured at a frame rate of 25 fps with a resolution of 512 by 384 pixels and 16-bit mono audio sampled at 32 kHz.	sentences	(Diverse head positions of the speaker)	Camera
AVICAR (Lee et al., 2004)	AVICAR was recorded in moving vehicles and addressed shortcomings of AVLetters, like low resolution data and limited number of subjects. Four cameras captured near-frontal perspectives of the vehicle's driver from four angles. Videos were captured with a 720-by-480-pixel resolution and a frame rate of 30 per second. Eight channels of 48 kHz audio are down sampled to 16 kHz. The dataset has 86 speakers. Subjects are required to speak isolated 26 letters twice under each noise condition. Subjects are required to speak isolated digits two times under each noise condition. The dataset also includes phone numbers. Each script set includes 20 10-digit phone numbers. Each of the 100 speakers in the dataset reads the same script repeatedly under five distinct noise conditions. This script's vocabulary consists of 13 digits, 26 isolated alphabets, and words from TIMIT sentences, totaling 1317 words. There are a total of 118 instances for each script set. Consequently, the total number of utterances is 59000 (118 × 5 × 100).	Alphabets Digits Sentences	4 View cameras	Camera
AVOZES (Millar and Goecke, 2004)	Consists of 720 × 480, 30 fps video recordings of 20 (10F, 10M) speakers speaking ten digits and three sentences containing nearly every phoneme and viseme in Australian English. The dataset contains 16-bit 48-kHz monophonic audio recordings. The dataset has different view angles with the speaker's head turning, different illumination levels, and different acoustic noise levels.	Digits Sentences	−45, 0, 45 (Diverse head positions of the speaker)	Camera
AV TIMIT (Hazen et al., 2004)	For SI-AVSR, the AV-TIMIT database was created and the dataset includes 233 (117M, 106F) speakers' videos, speaking 450 TIMIT-SX sentences that are intended to provide thorough coverage of phonetic contexts. At least nine different speakers were required to utter each of the 20 sentences, and one sentence had to be uttered by every speaker. Video files had a resolution of 720 × 480 and 16 kHz audio.	sentences	Frontal	Camera
VALID (Fox et al., 2005)	VALID was created to compare AV-speaker identification systems. In a realistic office setting, the database was captured in five sessions with varying illumination, background, and noise. The recording process spanned a month to account for skin tone, facial hair, and visual background variations. The VALID database contains 106 (77 M, 29 F) subject recordings with video resolution of (576 × 720, 25 fps). The dataset comprises 16-bit stereo 32-kHz audio recordings. Per session, three utterances were recorded, including the speaker's full name, ten connected digits, and sentences identical to those in XM2VTS.	Digits	15 degrees above, below, left, and right of the camera (Diverse head positions of the speaker)	Camera
GRID (Cooke et al., 2006)	This corpus includes videos collected from 34 speakers (18M, 16F) who uttered 1000 constrained sentences at a video resolution of 720 × 576 at 25 frames per second and an audio resolution of 25 kHz at 16 bits. Each speaker speaks all possible combinations for the fixed structure of the sentence: "[command: bin; lay; place; set] [color: blue; green; red; white] [preposition: at; by; in; with] [25 alphabets a-z, excluding w] [digits:0 to 9] [adverb: again; now; please; soon]." It includes a 51-word vocabulary.	Phrases	Frontal	Camera
CMU AVPFV (Kumar et al., 2007)	The CMU AVPFV database was built for SD lip-reading. It consists of cameras that record the front and profile views at the same time. It includes the data of ten participants, each of whom repeated 150 possible words from the "Modified Rhyme Test" 10 times. The video was captured in 640 × 480 resolution at 30 fps.	Words	0, 90	Camera

(continued on next page)

Table 3 (*continued*).

AV Letters2 (Cox et al., 2008)	Contains HD videos (1920 × 1080) with 50 frames per sec, of five subjects reciting the 26 alphabets seven times each. Recording environment: frontal pose, constant illumination, constant head pose, and neutral face expression without emotion. Monophonic 16-bit 48-kHz audio recordings.	Alphabets	Frontal	Camera
QuLips (Pass et al., 2010)	QuLips consists of recordings of two speakers set against a blue background and with constant illumination settings. Two cameras were used to record each speaker reciting a pair of ten-digit sequences. The first camera's initial position was maintained at all times. On the other hand, the subject and the second camera could rotate in 10-degree steps, enabling the capture of two distinct angles simultaneously. It includes digital video recordings with a resolution of 720 × 576 and a frame rate of 25 and is helpful for SD testing.	Digits	0, 10, 20, .., 90	Camera
WAPUSK20 (Vorwerk et al., 2010)	It adopts the same sentence structure as GRID. 20 (11M, 9F) speakers recorded 100 GRID-type sentences using four audio channels recording 16kHz, 32-bit audio, and a stereo camera recording videos with 640 × 480, 32 fps resolution. These recordings were made in an office setting. These corpora provide a significant number of examples per class.	Phrases	Frontal images captured partially at an angle from the side using stereo cameras.	Stereo camera
LILir (Lan et al., 2010)	LILiR consists of 12 speakers (7M, 5F), each of whom recites one instance of 200 sentences drawn from a corpus of 1000 words called the Resource Management Corpus. Recordings were made under constant illumination using five cameras positioned at 0, 30, 45, 60, and 90 degrees.	sentences	0, 30, 45, 60, 90	Camera
UNMC-VIER (Wong et al., 2011)	With different resolution cameras (<708 × 640, 25 fps>,<320 × 240, 29 fps>, and <320 × 240, 15 fps>). , UNMC-VIER adds appearance variations to video recordings, such as lighting, facial expression, head poses, and image quality. The utterances are spoken both slowly and at a regular speaking pace. The 11 sentences used come from the XM2VTS. The database contains recordings of 123 (74M, 49F) speakers in two sessions, one in a controlled environment and the other in an uncontrolled setting.	sentences	(Diverse head positions of the speaker) 0, 30, 60, 90, 120, 180,up,down and front	Camera
MOBIO (McCool et al., 2012)	MOBIO was proposed for speaker authentication, and 12 separate recording sessions were conducted, typically separated by several weeks. Recoding was performed using mobile phones in an unrestricted setting. There are 192 distinct samples for each of the 150 participants. The sentences included several answers to short questions and responses in free-speech and predefined text. Video recordings have a 640 × 480 resolution, a 16 fps frame rate, and a 48 kHz audio sampling rate.	sentences	Uncontrolled settings	Mobile, laptop
AusTalk (Estival et al., 2014)	It is a large AV corpus of Australian English consisting of 640 × 480 resolution video recordings of 1000 speakers speaking 322 words, 12-digit strings, and 59 sentences.	Digits words Sentences	Frontal	Camera
MIRACL-VC1 (Rekik et al., 2014)	The MIRACL-VC1 dataset was proposed for SI and SD lip-reading. An RGBD camera with a resolution of 640 × 480 pixels and a frame rate of 15 fps was used to record the videos. RGB frames and depth frames were sampled from the videos. There are 15 participants who each utter ten words ten times and ten phrases ten times, resulting in 1500 word and 1500 phrase instances of RGB and Depth.	Words Phrases	Frontal	RGBD camera
CREMA-D (Cao et al., 2014)	The dataset consists of 12 sentences expressing diverse emotions, including happiness, sadness, anger, fear, disgust, and neutrality. The dataset contains 91 speakers (48M, 43F). Videos have a resolution of 960 by 720 pixels, and audio has a sampling rate of 48 kHz.	sentences	Frontal	Camera
OuluVS2 (Anina et al., 2015)	Consists of 53 (40M, 13F) speakers uttering ten digit sequences (3 repetitions), ten phrases (3 repetitions), and ten randomly selected TIMIT sentences (1 repetition) in OuluVS2. Each speaker had a separate set of sentences. Each utterance was captured simultaneously by six cameras from five distinct angles. The dataset includes audio recordings with a 128 kbps bitrate and video recordings with a 1920 × 1080 resolution and 30 frames per second.	Digits Phrases Sentences	0, 30, 45, 60, 90	Camera
RM-3000 (Howell, 2015)	RM-3000 is a collection of 3000 sentences with 1000 words uttered by a single male subject. The videos were recorded at a frame rate of 60 fps with a resolution of 1920 × 1080 pixels which was then downsampled to 360 × 640, and 16-bit mono-recorded audio was sampled at 48 kHz.	sentences	Frontal	Camera
TCD-TIMIT (Harte and Gillen, 2015)	The TCDTIMIT database contains videos with a resolution of 1920 × 1080 pixels, a frame rate of 30 fps, and 16-bit, 16 kHz audio. The data set comprises 62 (30M, 32F) speakers, three expert lip readers, and 59 volunteers. Each of the three experts speaks for 377 sentences, compared to the others, who speak 98 sentences. Two cameras, one fixed at a frontal angle of 30 degrees and the other at a frontal view, were used to record in a controlled lab environment with constant illumination.	sentences	0, 30	Camera
LRW500 (Chung and Zisserman, 2017)	LRW is a dataset of 500 words designed for SI lip reading. A multistage pipeline for the automatic collection of data from BBC television shows between 2010 and 2016 was set up to extract the video clips and align them with transcriptions of the spoken words. It includes over thousand-word instances said by over a thousand individuals in realistic scenarios with varying lighting conditions and head poses. All videos are of resolution 256 × 256 and 25 fps. There are 29 frames in each video. And the audio was recorded at 16 kHz (single channel).	Words	Many views	BBC TV shows
MODALITY (Czyzewski et al., 2017)	The dataset consists of 35 (26M, 9F) speakers pronouncing 182 distinct words to simulate 168 voice control commands for laptops and mobile devices. Each speaker took part in 12 recording sessions, six of which were conducted in quiet settings and the remaining six with three types of sounds (traffic, conversation, and factory sounds) added via four speakers. The cameras were set up to record 1080 × 1920 video streams at 100 frames per second. 8 microphones captured audio at 44.1 kHz and 16 bits.	Words Sentences	Frontal images captured partially at an angle from the side using stereo cameras.	Camera (stereo, time of flight depth camera)

(*continued on next page*)

Table 3 (*continued*).

VoxCeleb1 (Nagrani et al., 2017)	It is a vast AV corpus without text transcriptions gathered from YouTube. It includes over 100,000 utterances from 1251 celebrities. Its primary purpose is speaker identification.	utterances	Many views	youtube videos
MV-LRS (Son and Zisserman, 2017)	The MV-LRS comprises 74, 564 video recordings from BBC shows containing 14,960 words.	sentences	0 to 90 degrees	BBC TV shows
AV-Digits (Petridis et al., 2018)	In three speech modes, regular, whisper, and mute, 53 (41M, 12F) participants pronounced ten digits in 0-9 in random order five times, and 39 (32M, 7F) participants pronounced short phrases identical to ouluVS2 five times, from three different angles (frontal, 45, and profile) and in three speech modes, regular, whisper, and mute. The 1280 × 780 pixel videos were recorded at 30 fps with a 44.1 kHz audio sampling rate.	Digits Phrases	0, 45, 90	Camera
CRSS-4ENGLISH-14 (Tao and Busso, 2018)	This AVSR corpus was created by 442 contributors (217 female and 225 male speakers) with four distinct English accents: American, Australian, Indian, and Hispanic. Five microphones with a sampling frequency of 44.1 kHz and two cameras with resolutions of 1280 × 720 at 24 fps and 1440 × 1080 at 29.97 fps were used to capture data against a uniform green screen background. The content contained continuous sentences, questions, brief phrases or commands, continuous numbers, single words, and city names. The first phase of data collection did not include noisy environments (malls, houses, offices, and restaurants), while the second phase did.	sentences	Frontal	Camera
LRS2 (Afouras et al., 2018a)	LRS2 is sentence-level data beneficial for SI lip reading. Similar to LRW500, a multistage pipeline was built to automatically collect data from BBC News between 2010 and 2016 to extract video clips and align them with transcriptions of the spoken sentences. It includes 17,428 unique words within 118,116 utterances spoken by over a thousand individuals in realistic settings with varying lighting conditions and head positions. In addition to subtitling, the dataset also includes word alignment boundaries. All videos are 160 × 160 and 25 frames per second. The sampling frequency of the available audio is 16 kHz (single channel).	sentences	Many views	BBC News
LRS3 (Afouras et al., 2018b)	LRS3 is sentence-level data beneficial for SI lip reading. Like LRW500 and LRS2, a multistage pipeline was developed to automatically collect data from 5594 TED and TEDx talks to extract video clips and align them with spoken-sentence transcriptions. It includes approximately 71.1k unique words in 151.8k utterances spoken by over 5000 people in realistic settings with varying lighting conditions and head positions. The dataset also contains word alignment boundaries in addition to subtitles. Every video is 224 × 224 pixels and 25 fps. The sampling rate of the available audio is 16 kHz (single channel).	sentences	Many views	TED and TEDX talks
LSVSR (Shillingford et al., 2019)	LSVSR is speaker independent lip-reading dataset extracted from public YouTube videos. It includes approximately 127,055 unique words in 2,934,899 utterances. The dataset includes paired phonemes and video sequences. Every video has a 128 × 128 pixels resolution and a sampling rate between 23 and 30 fps.	sentences	Many views	Youtube videos
YTDEV18 (Makino et al., 2019)	YTDEV18 was collected for large-scale AVSR. It contains 25 h of transcribed YouTube videos uploaded by individuals. A total of twenty thousand segmented utterances are available. YTDEV18 utterances were artificially corrupted with babble noise and randomly selected noise from the NoiseX database. Every video sample has a resolution of 128 × 128 pixels and a sampling rate between 24 and 30 frames per second. The available audio has a sampling rate of 16 kHz (single channel).	sentences	Many views	youtube videos
VoxCeleb2 (Chung et al., 2018; Nagrani et al., 2020)	VoxCeleb2 is an improvement over VoxCeleb1 that features a broader range of ethnicities. Six thousand speakers are contributing one million utterances. The dataset includes a speech from speakers of 145 nationalities, representing various accents, ages, ethnicities, and languages.	utterances	Many views	youtube videos
SpeakingFaces (Abdrakhmanova et al., 2021)	The SpeakingFaces dataset was proposed for HCI and authentication systems. It consists of aligned, high-resolution thermal, visual image streams of face images synchronized with audio recordings of each subject speaking approximately 100 essential phrases selected from the digital assistant database at Stanford University and command sets of the virtual assistant Siri. In the first session, 142 (71M, 71F) subjects were instructed to maintain silence, and in the second session, they were instructed to read commands. The nine camera positions collected visual, thermal, and audio data during both sessions. Each subject participated in two trials consisting of two sessions, conducted on separate days at least two weeks apart to account for variations in hairstyle, clothing, and glasses. The recordings were conducted with a thermal video resolution of 464 × 348 at 28 frames per second and a visual camera resolution of 768 × 512 at 28 frames per second. A sampling frequency of 44.1kHz was used to record two-channel audio.	Phrases	9 different camera positions	Thermal and visual camera
LIPAR (Nemani et al., 2022)	LIPAR consists of 35 (31M, 4F) speakers uttering ten words of MIRACL-VC1 only once. The video recordings have a high resolution of 3840 × 2160 pixels.	Words	Frontal	Camera
DVS-lip (Tan et al., 2022)	DVS-Lip dataset contains a total of 100 words, including the 25 most frequently confused word pairs and 50 words randomly selected from the LRW500 vocabulary. Participating were 40 (20M, 20F) speakers, whose intensity frames were recorded at 346 × 260 resolution and 25 fps, while event frames were recorded at 25FPS and 200FPS.	Words	Frontal	Recorded by event camera.

the Data Sharing Scheme (DSS) was employed to increase training samples. Hassanat (2014) conducted a pilot study using nine descriptors, including height, the width of the mouth, the mutual information between Discrete Wavelet Transform (DWT) of current and previous mouth images, the image quality factor, the vertical and horizontal features ratio derived from DWT, the ratio of vertical edges to horizontal edges detected using Sobel edge detector, the amount of red color indicating tongue, the amount of visible white teeth, and the chi-square test on histograms of first mouth image and the current image, and catered k-nearest neighbors (KNN) for word classification.

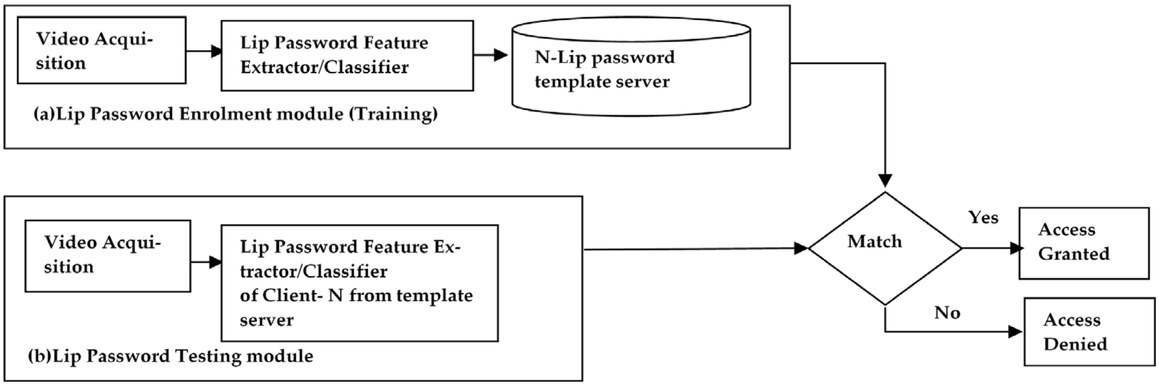
Video Acquisition
Lip Password Feature Extractor/Classifier
N-Lip password template server
(a)Lip Password Enrolment module (Training)
Video Acquisition
Lip Password Feature Extractor/Classifier of Client- N from template server
(b)Lip Password Testing module
Match
Yes
No
Access Granted
Access Denied

Fig. 3. General pipeline for Lip reading biometrics.

The system then grants access only if the distance between the features of the test password and the authenticated visual password is below a threshold. A survey article Mathulaprangsan et al. (2015) describes the lip-reading techniques used for password verification. Radha et al. (2015) presented the person identification system by late fusing the scores of the face recognition system, which uses PCA for extracting features and KNN for classification, and the lip-password recognition system, which uses DWT for extracting features and HMM for classification. Integrating face recognition and lip-reading systems improved the accuracy. The authors Shi et al. (2016) extracted a word transition segment known as a "mute" segment from the utterance of four-digit numbers. The word segments were further input into the HMM-Universal background model for word classification. And the mute segment's closed lips were input into a linear Support Vector Machine (SVM) to identify a specific individual. In Lai et al. (2016), Lip's video was viewed as a spatiotemporal cube that may be further subdivided into overlapping 3D spatiotemporal cells. Sparse coding and hierarchical max-pooling were then applied to obtain the final features, which were then classified using a linear SVM.

Lip passwords function as a double security system. The first level of security is introduced by speaker verification, and the second level of security is introduced by granting access only when the target speaker utters the right password. In a fascinating study (Cheung and Zhou, 2018), a language-independent lip password was investigated. The authors Shang et al. (2018) introduced extreme learning machines for smartphone lip-reading authentication. Lip motion image frames were examined in the XY, YT, and XT planes, and histogram images were derived from Local Binary Pattern (LBP) images. The Singular Value Decomposition (SVD) of histogram images was computed to determine the projection vector. The final step involved multiplying the histogram image by the projection vector to obtain the feature vector. Their experimental results on the ouluVS benchmark demonstrate that combining face identification and lip-reading system scores improves the system's accuracy. The authors Ezz et al. (2020) favored HMM over long short-term memory (LSTM) because training HMMs requires less training data. They have developed a dual authentication system. The first authentication layer identifies the user, and the second layer extracts the Arabic digit password from the three speaker-dependent lip-reading HMM models using a voting scheme. These HMM models were trained on the lip-password video of an authenticated user using handcrafted features such as Speeded-Up-Robust Features (SURF), Histogram-of-Oriented-Gradients (HOG), and Haar features. Kapkar and Bharkad (2021) suggested a double identification system. Face identification was performed using KAZE features and KNN classifier, as well as alphabet lip-reading, by extracting the ratio of lips, height, and width features that were classified using multiclass SVM.

The following works investigate deep learning techniques. Lee and Myung (2017) developed Lip-Reader, a smart device login system, by training a network consisting of a dense layer, two LSTM layers, and a drop-out layer with the GRID-digit dataset. In Cheng et al. (2018), they have used a dual-task ConvNet (CNN) module on the GRID benchmark to identify the speaker and the uttered content for authentication and liveliness detection with random prompt strings. The model consists of a lip feature network with multiple conv3D layers, an identity network with time-distributed conv2D, and a content network with Bi-directional Gated-Recurrent-Unit (Bi-GRU). Augmentations were employed to make the model less sensitive to the horizontal separation between the user and camera sensor and different profile views. Initially, the user was required to utter 25 sequences for prompt strings during training, and subsequently, the baseline model was fine-tuned using 25 additional samples. Half the total error rate (HTER) was mitigated as the number of new samples increased from 25, 50, 100, and 200.

Few researchers have employed one-shot learning to develop an open-set model. The open set authentication model proposed by Wright and Stewart (2019) employs a modified version of 2D CNN referred to as Spatio-temporal (ST) CNN that processes frames of the XM2VTS dataset over time, along with Bi-GRU as an embedding branch of the Siamese network and triplet loss was used for training. Further, the work was enhanced in LipAuth (Wright and Stewart, 2020), a Deep Siamese network consisting of STCNN and Bi-GRU was trained with semi-hard triplet loss to extract one-shot embeddings for user authentication, and the model showed promising results on the XM2VTS, qFace, and FAVLIPS benchmarks.

Ma et al. (2020) introduced feature extraction using a Difference block and Dynamic response block for Dynamic Enhanced Authentication Network (DAE-Net) classification. The Difference block eliminates static information by calculating a weighted sum of difference frame maps. The Dynamic response block calculates a weighted sum of feature responses within the same spatial location and different time locations to determine differences between pixels over time. The authentication trend again once again shifted to one-shot learning. Ruengprateepsang et al. (2020) proposed one-shot learning using a hybrid model trained with semi-hard triplet loss that combines the sentence and speaker models. Although trained with a single training sample of the new user, the model's performance on ten phrases from the AVDigits dataset was excellent.

Researchers (Yang et al., 2020) have attempted to develop an authentication system named Speaker-Authentication Dynamic-Talking-Habit Network (SA-DTH-Net) that is robust against imposters and deep fake attacks such as Generative adversarial network (GAN)-generated talking videos of clients. SA-DTH-Net consists of a Feature Extraction Network (FFE-Net), which uses time-distributed convolutions to extract features across time, and a Representative Descriptor extraction Classification Network (RC-Net) with two branches, a classification branch that determines whether the authenticated user utters the word or not, and a Reconstruction branch that reconstructs the lip motions from the extracted features. In the proposed model, one model was required for

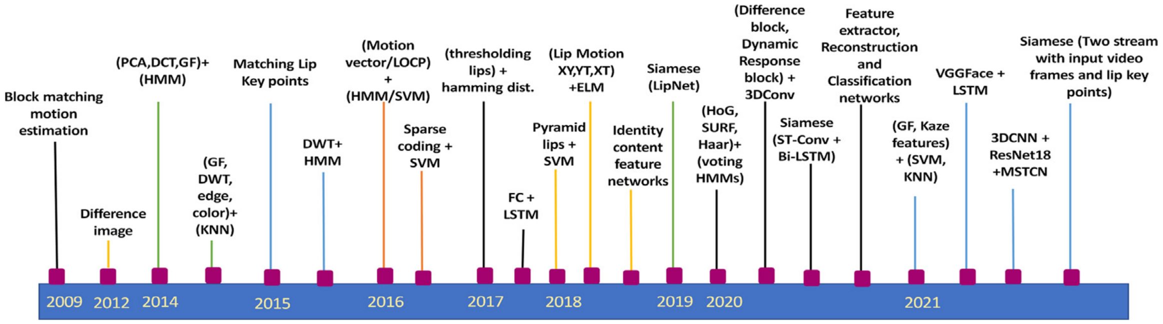
Block matching motion estimation
Difference image
(PCA,DCT,GF)+ (HMM)
(GF, DWT, edge, color)+ (KNN)
Matching Lip Key points
DWT+ HMM
(Motion vector/LOCP) + (HMM/SVM)
Sparse coding + SVM
(thresholding lips) + hamming dist.
FC + LSTM
Pyramid lips + SVM
(Lip Motion XY,YT,XT) +ELM
Identity content feature networks
Siamese (LipNet)
(HoG, SURF, Haar)+ (voting HMMs)
(Difference block, Dynamic Response block) + 3DConv
Siamese (ST-Conv + Bi-LSTM)
Feature extractor, Reconstruction and Classification networks
(GF, Kaze features) + (SVM, KNN)
VGGFace + LSTM
3DCNN + ResNet18 +MSTCN
Siamese (Two stream with input video frames and lip key points)
2009
2012
2014
2015
2016
2017
2018
2019
2020
2021

Fig. 4. Road map of Lip reading biometrics.

Table 4
Lip reading in biometrics using computer vision and signal processing approaches.

Computer Vision and Signal processing approaches								
Year	Ref.	Input type	Feature extraction Technique	Classifier	Benchmark	Task Type	Result	Limitations
May 2009	Alghathbar and Mahmoud (2009)	Grayscale mouth Images	Estimating eight motion curves in 8 directions using block match algorithms	Motion feature vectors comparison using MSE	Tulips 1.0	1st four digits Worked with three speakers	Average errors (30%)	Designed and tested only for 1st four digits.
Sep 2012	Sengupta et al. (2012)	RGB-frames of the user's lip	A threshold was applied on the red plane of both images.	The cutoff was set to the sum of the Difference image's pixel values.	own	Four words (hello, yes, no, yellow) with three speakers	100% (without varying lightening conditions)	Dint work for changing lightening conditions
Apr 2015	Lesani et al. (2015)	RGB-frames of the user's lip	Coordinates around the lip are extracted from all the video frames and cascaded.	Nearest distance between the extracted features and the stored lip features	LiLiR(English Characters)	Characters Ten subjects	Lip reading accuracy: 70%	Accuracy is low
Jul 2017	Shubhangi and Balbhim (2017)	RGB-frames of the user's lip	A threshold was applied on an extracted pink area of lips to create a lips template.	The hamming distance was used to classify	own	Three words (camera, water, sugar) One subject	Correct password (95%) and detecting wrong password (100%)	Verified only for one subject

every user word, and clients became impatient when asked to record additional training videos during registration. Authnet is an open-set model introduced by Raghavendra et al. (2021). It utilizes VGGFace to extract face features and the LSTM network to determine if the input test video contains the correct client-word combination. It was evaluated on MIRACL-VC1 and a collated dataset. Applying cosine difference to feature descriptors revealed a significant difference between descriptors obtained through VGGFace for different clients and the word vocabulary spoken by the same client.

Instead of a vision-based approach, Lu et al. (2019) attempted acoustic signal analysis, wherein a smartphone speaker continuously emits a 20 kHz acoustic signal. User lip movements uttering password phrases induce the Doppler effect. This signal is transformed into the Fourier domain to calculate the signal gradient. It is then divided into small chunks corresponding to individual words. These chunks are then input to a three-layer Autoencoder to extract efficient features for training one-against-all SVM, thereby constructing a spoofer detector and binary classifiers for each newly registered user to distinguish them from previously registered users. Forty-eight volunteers participated in experiments conducted in four distinct environments: the bright lab, the dark lab, the noisy station, and the pub. The outcomes demonstrated an increase in accuracy for user identification and imposter detection, making the system robust to noisy and nighttime environments. FaceLip was developed for Android phones (Zhou et al., 2021) for liveness detection to counter 3D silicon masks attack and 3D projection attacks during face authentication. Several inaudible tones between 18 kHz and 21 kHz were produced by the phone speaker placed 18 cm from the client's mouth. The microphone was used to record the lip-reflecting acoustic signals. The captured lip movements are unique and unforgeable. These are then fed to the CNN-LSTM network to validate the authorized client's four-character passcode containing the client's digits and alphabet. Our research has yet to explore similar work on lip-reading using sensors, ultrasound, or acoustic signals but has focused solely on computer vision methods.

In recent works, Liu et al. (2021) investigated speech signals and user lip dynamics for biometrics. For the lip reading biometrics stream, a grayscale lower face ROI was fed to 3DCNN, followed by ResNet18 and Multi-scale TCN. Extended time delay neural network (E-TDCNN) was fed a biometrics stream spectrogram for speech recognition. The intermediate fusion, late fusion, and decision-level fusion of audio-visual feature vectors were performed, and then cosine similarity was tailored to assess the match between two test embeddings. The training set consisted of phrases spoken by 30 female speakers from TCD-TIMIT, the validation set consisted of phrases spoken by ten female speakers from Lombard Grid, and the test set consisted of phrases spoken by 20 female speakers from Lombard Grid and 16 female speakers from Grid. According to their findings, the deep learning approach outperforms the conventional approach. All experiments, however, were conducted on female speakers only. Using triplet loss, a Siamese network with a two-stream base network that accepts RGB mouth frames and 24 lip landmarks was trained (Zakeri and Hassanpour, 2021). For training the model, scholars used 66 subjects from the CREMA-D dataset consisting of 12 sentences uttered with varying emotions and 11 different subjects to validate the model. Fig. 4 depicts the progression of Lip reading biometrics techniques over time. Tables 4–6 summarizes the discussed lip-reading biometric approaches.

4. Audio visual speech recognition

Because audio signals are frequently distorted by noise, using lip cues together with speech recognition improves speech recognition

Table 5
Lip reading in biometrics using ML approaches.

Machine learning approaches								
Year	Ref.	Input type	Feature extraction Technique	Classifier	Benchmark	Task Type	Result	limitations
Feb 2014	Liu and Cheung (2014)	RGB-frames of the user's lip A data-sharing scheme (DSS) was applied to increase training samples.	PCA features, DCT features, and Geometric features were extracted. Random Subspace Method (RSM) scheme was tailored for the feature selection.	Multi-boosted Hidden Markov Models (HMMs) with Adaboost learning	own	Digital password phrase uttered by 46 speakers	Authorized user with the right password EER=3.91 Imposter with correct password EER=4.06 Imposter with wrong password EER=3.37	Different illumination conditions and variations in speaking pace should have been considered.
Sep 2014	Hassanat (2014)	RGB-frames of the user's lip For a stronger authentication system, the system can ask part of the password (each time a different sequence), like 1st and 3rd visual password.	Height, the width of the mouth, DWT, Sobel edge detector, amount of red color of the tongue and visible white teeth, chi-square test on histograms.	k-nearest neighbors (KNN)	in-house video database	20 subjects 0-9 digits	Accuracy: Single-digit-known: 79.5%- Single-digit-unknown: 85.8%- Double-digit-known: 84.6%- Double-digit-unknown: 92.4%-	Time complexity is more.
Dec 2015	Radha et al. (2015)	RGB-frames of the user's lip	DWT	HMM	own	Speaker specific ten phonetically different words, ten subjects	Accuracy (72.2%)	Videos were shot under the same lighting conditions. Thirty-five instances were considered for training.
Sep 2016	Shi et al. (2016)	RGB-frames of the user's lip	The motion vector of 14 points on lips, Local Ordinal Contrast Pattern (LOCP)	HMM-Universal background model SVM	Dataset of Wang and Liew (2012)	Four-digit sequence, 40 subjects	HTER: 1.26%	Tested only for 40 subjects
Dec 2016	Lai et al. (2016)	RGB-frames of the user's lower half face	Sparse coding and Hierarchical max-pooling	Linear SVM	Dataset of Wang and Liew (2012)	Four-digit sequence, 40 subjects	HTER: 0.46%	Tested only for 40 subjects, To extend the work to "prompt-text" password
Jul 2018	Cheung and Zhou (2018)	RGB-frames of the user's lower half face	Diagonal-like pooling lip representation, 3-layer pyramid structure of lips	Linear SVM	Own dataset	English and Chinese languages, 43 subjects, eight lip passwords	EER:0.41% HTER:8.86%	Tested only for 43 subjects
Nov 2018	Shang et al. (2018)	Gray frames of the user's lip and face	Linearity Preserving Projection face features, Features extracted from lip movements in XY, YT, and XT planes	Extreme learning machine	OuluVS	10 Phrases 20 subjects	LR accuracy:71% Face+LR:93.25%	Tested only for 20 subjects
Mar 2020	Ezz et al. (2020)	Gray frames of the user's lip	HoG, SURF, Haar	HMMs with the voting scheme	Own (Arabic lip)	10 Arabic digits (0-9), 20 subjects	Accuracy: 96.2%	The voting scheme increases the complexity of the model.
Feb 2021	Kapkar and Bharkad (2021)	RGB-frames of the user's lip and face	The ratio of lips, height, and width (for lips) Kaze features (Alcantarilla et al., 2012) (for face)	Multiclass SVM(for lips) KNN (for face)	own	18 subjects Characters	Only LR accuracy (70%) Face recognition+ LR accuracy (80%)	Less number of subjects

systems' precision. Lip-reading systems can improve voice-driven applications (voice assistants) (Ivanko et al., 2022a). Using frameworks for lip reading and speech recognition, hearing aids that output transcribed text is created (Kumar et al., 2022). Additionally, it can improve the precision of audio-based speech transcriptions. The performance of the Watch-Listen-Attend-Spell (WLAS) architecture developed by Chung et al. (2017) demonstrated that visual lip cues improve the precision of speech recognition models in disturbing, noisy environments. Here, LSTM was used to classify the Visual Geometry Group VGGM extracted features. Their trained WLAS network converts the videos of talking lips into text transcription. Although there has been a significant revolution in the study of audio speech recognition, there are still many problems to solve, such as multiple interfering speakers in a cocktail party scenario, background noise in audio and video conferencing, and voice assistants with background noise (Sahu et al., 2018). Fig. 5 depicts the AVSR pipeline, illustrating early and late audio and video fusion. In the case of early fusion, the AVSR classifier generates text transcription using cascaded AV features extracted from the audio feature extractor

Table 6
Lip reading in biometrics using DL approaches.

Deep learning approaches								
Year	Ref.	Input type	Front End	Backend	Benchmark	Task Type	Result	limitations
Jan 2017	Lee and Myung (2017)	Gray frames of the user's lip	Fully connected(FC) layers	LSTM +FC	Grid	Only digits (0-9) Speakers(1-9)	Speaker dependent accuracy: 93.8% Speaker independent accuracy: 22.5%	Tested on 9 Closed set users
Nov 2018	Cheng et al. (2018)	RGB-frames of the user's lip	Lip-Feature-Network	Identity-Network Content-Network	Grid	Phrase 34 speakers 200 samples for training	HTER: 2.03	Tested on 34 closed-set users with more training samples
Oct 2019	Wright and Stewart (2019)	RGB-frames of the user's lip	The Siamese model was trained using triplet loss	Lip-Net (Assael et al., 2016) as a twin network	XM2VTS(295)	A sequence of 20 Digits	Closed set EER: 1.03%	Computation complexity is more compared to a single network.
Mar 2020	Wright and Stewart (2020)	RGB-frames of the user's lip	The siamese model was trained using triplet loss	Lip-Net (Assael et al., 2016) as a twin network	FAVLIPS (42), qFace (10), XM2VTS(295)	Sequence of Digits	EER:22% EER:2.88% Open set EER:1.65%	Not robust to illumination variations
Oct 2020	Ma et al. (2020)	RGB-frames of the user's lip	Difference block Dynamic Response block	3D Conv layers, FC layers	Grid	Selected 22 Words containing digit, color, adverb, command 33 speakers	HTER: 1.38	The model needs a lot of training examples. Designed for closed-set users
Nov 2020	Ruengpra-teepsang et al. (2020)	RGB-frames of the user's lip	A Siamese neural network trained with semi-hard triplets loss	Spatio-temporal convolution and Bi-LSTM as twin network	AVDigits	Ten phrases, 39 subjects	EER:9.0%	The model must be pretrained before one-shot learning.
Dec 2020	Yang et al. (2020)	RGB-frames of the user's lip	Feature Extraction Network	Reconstruction Classification network	Grid, Mobio subset (6-digits and 59 speakers)	Grid: Selected 22 Words containing digit, color, adverb, command 33 speakers	HTER:0.6 for Grid and HTER: 2.4 for Mobio	One model for one open set user.
May 2021	Raghavendra et al. (2021)	RGB-frames of the user's face	Pre-trained VGGFace model (Parkhi et al., 2019) +Stacked LSTM layers	Cosine difference applied to feature descriptors.	MIRACL-VC1	Ten words Ten speakers	Accuracy: 98%	One model for one open set user.
Dec 2021	Liu et al. (2021)	Gray scale-frames of the user's lower half face	3D CNN + Resnet18	Multi-scale TCN	TCD-TIMIT (30 F)-train set Lombard GRID (10 F)-val set GRID (16 F)-test set	sentences	EER: 8.65	Scholars have experimented only with female speakers.
Dec 2021	Zakeri and Hassanpour (2021)	RGB-frames of the user's lip + 24 Mouth landmark Points	Siamese network with triplet loss with two-stream base networks.	Video frames Stream 1: STCNN +GRU Lip landmarks Stream 2: three-time distributed FC layers + LSTM	CREMA-D	12 Sentences Six emotional states Total 88 subjects. During the training process, 66 subjects were involved.	The model was validated and tested using 11 different subjects. Test acc: 95.41%	Variations in the subject's face due to facial hair was not considered.

and the video feature extractor, respectively. In the case of late fusion, the decision scores from the AV classifiers are combined to produce the final decision score used for generating text transcription.

In Borde et al. (2022), Ivanko et al. (2021), Katsaggelos et al. (2015) and Kandagal and Udayashankara (2014) steps to designing an AVSR system are outlined. In the first step, unimodal audio and video recognizers are built. In the second step, both modalities are fused. Audio features like spectrogram, MFCC coefficients, Perceptual-Linear-Prediction coefficients (PLP), Linear-Predictive-Coding coefficients (LPC), Linear-Prediction-Cepstral-Coefficients (LPCC), DWT, and Dynamic-Time-Wrapping (DTW) were extracted from audio. Video features like Pixel-Based features obtained after applying image transformations like DCT, DWT, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Hierarchical Linear Discriminant Analysis (HiLDA: where LDA is applied to cascaded feature vectors of first, middle, and last frames in a 15-frame window, then maximum likelihood linear transform (MLLT) maps the features to suppress within class distance and maximize between class distance), Geometry-Based features of lips motion like histogram, height, width, area, and shape, Texture-Based features like Local Binary Patterns (LBP), Model-Based features like Active-Appearance-Model features (AAM), Active-Shape-Model (ASM) were extracted from the video. For the fusion strategies of AV models, Early-fusion, where audio and video features are combined and given to the classifier, and Late-fusion, where the classification

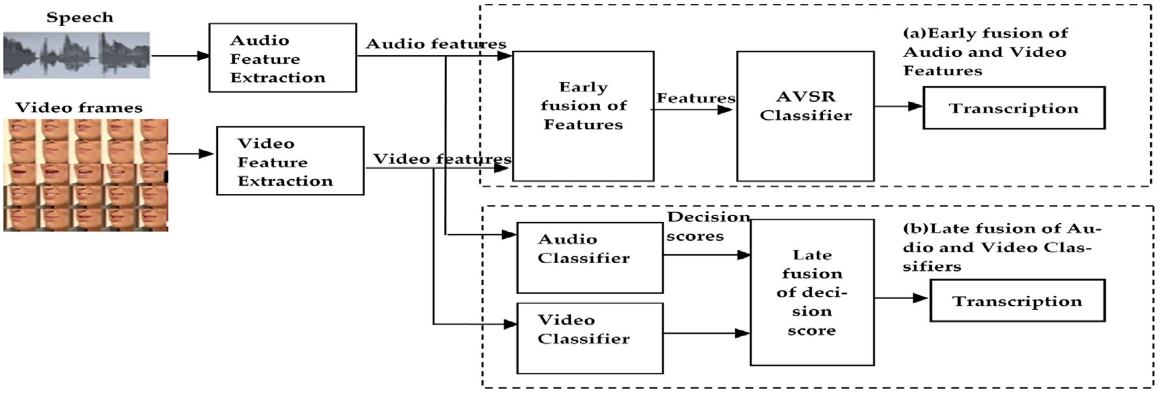

Fig. 5. General pipeline of AVSR.

scores of both models are merged to make decisions. And Hybrid-fusion, which combines both features and decision scores, was employed. Earlier scholars tailored the classifiers like HMMs, Gaussian Mixture Models (GMMs), DTW, K-NN, SVMs, Artificial Neural Networks (ANNs), Dynamic Bayesian Networks (DBNs), Deep Dynamic Bayesian Networks (DDBNs), and Random Forest Classifier (RFCs).

Several techniques attempted to combine speech recognition and lip reading networks employing deep learning approaches. Deep belief networks were evaluated by Huang and Kingsbury (2013), and Paleček (2018) and Noda et al. (2015) employed multi-stream HMM. Yang et al. (2017) proposed a novel Correlational RNN (CorrRNN) to fuse AV features. Their unsupervised CorrRNN trained with reconstruction losses consists of an encoder with a dynamic weighting framework that extracts joint audio-visual features and a decoder that reconstructs audio and video simultaneously from joint features. Finally, extracted features were classified utilizing SVM. Chung and Zisserman (2018) adopted SyncNet trained using contrastive loss. They have developed a fully automated pipeline to collect BBC News lip-reading data. They employed two-stream ConvNets capable of generating joint features between sound and mouth movements from the unlabeled dataset for active speaker detection and audio–video synchronization. To classify the dataset words, a VGG-M LSTM network has been used. In an inspiring study (Paleček, 2018), AVSR was tested on half a million Czech vocabulary terms. Several features, including Mel frequency cepstral coefficients (MFCC), DCT2D, Depth Active Appearance Models (AAM), LBPTOP (LBPs extracted from X-Y, X-Time, and Y-Time axis), DCT3D, HOGTOP (HOGs extracted from X-Y, X-Time, and Y-Time axis), and ST-DNN features of depth maps and RGB frames, were classified using Multi-stream HMM models. Their findings demonstrate that spatiotemporal DNN features extracted from RGB video and linearly interpolated Depth video, along with MFCC features, outperform other classification features.

Further, Bi-LSTMs were used by Stafylakis et al. (2018), transformer and seq2seq networks were utilized by Afouras et al. (2018a), Makino et al. (2019) created the RNN-Transducer framework to carry out AVSR, and it was tested on the LRS3 and YTDEV18 dataset comprised of YouTube videos. Martinez et al. (2020) tried to recognize LRW-500 words by extracting features through 3D-convolution and ResNet18 and classifying them using a Multi-scale Temporal convolutional network (TCN) back end. Because every individual has unique lip movements, the speaker-independent lip reading model is extremely difficult to achieve. Motivated by this, the authors Fernandez-Lopez et al. (2020) investigated the adaptation approach for an unknown subject by constructing a speaker-independent AVSR framework utilizing “Coupled GAN (CGAN)”, which can generate the joint weights of multi-speaker images. Their model, evaluated on the TCD TIMIT dataset, showed a 10 percent average improvement in AVSR recognition. Yu et al. (2020) proposed a gated fusion technique employing a stack of factored time delay NNs (TDNN-F) to integrate AV features and improve the lattice-free maximum mutual information (LF-MMI) audio recognition framework. It is difficult to accurately determine the fusion weights in the case of late fusion. To circumvent this issue, Wand and Schmidhuber (2020) proposed two variants of trainable fusion architectures and evaluated them on 51 GRID benchmark words. Liu et al. (2020) on LRW data, performed word-level AVSR and proposed a Bidirectional SYNC network for AV-feature fusion. Their work employed a hybrid visual encoder, in which one stream accepts video frames and the other stream ingests a lip graph of 2D locations of twenty DLIB mouth points, further ST-GCN (spatial graph Convolutional network followed by temporal graph Convolutional network) extracts shape and motion features from the lip graph. On the LRW-500 corpus, Liu et al. (2021b, 2020) evaluated the performance of Bi-GRU using both feature and decision-level fusions (hybrid). Their proposed model consists of three components for feature-level fusion: an audio encoder, a video encoder, and an audio–video encoder. The final decision was made by combining the decisions of the three encoders. Ma et al. (2021b) conducted experiments with multilayer perceptron.

Sayed et al. (2021) came up with a novel Bimodal Variational Autoencoder (Bi-VAE) to fuse MFCC/GFCC (Gammatone Frequency Cepstral Coefficients) audio embeddings and video embeddings extracted from a pre-trained VGGFace2 network. Three models, LSTM, DNN, and SVM, were used to classify the obtained ASVR features, with SVM outperforming the other two on the CUAVE benchmark. Serdyuk et al. (2021) proposed a video transformer as the front end as an alternative to the 3DCNN front end. Their model outperformed CNNs when trained on YouTube videos with transcriptions and tested on the LRS3 benchmark. ViT extracts small image patches and sends them along with their positional embedding to a linear transformer. A fascinating work adopted AVSR to ensure correct pronunciation (Yu, 2022). The duration of the expert’s speech and the learner’s speech commonly vary. Consequently, following the extraction of MFCC features from speech, Empirical Mode Decomposition (EMD) and conditional maximum matching dynamic planning were used for matching features and evaluating the similarity score between the expert and learner’s speech. The optical flow captures the movements of the lip’s key points. A stacked convolutional independent subspace analysis (ISA) network was leveraged to extract video features and classify digits. The scoring levels for pronunciation ranged from 1 (highest correlation) to 4 (lowest correlation). Pronunciation correctness was tested on the YCLIP 6-digit Putonghua lip language database. This model highly relies on synchronization between lip movement and speech.

Recent research (Shashidhar et al., 2022) uses DNN to combine audio and video features. Shi et al. (2022) adopted the Audio-Visual-Hidden-Unit BERT for the AVSR. Their architecture employs self-supervised training to learn speech by watching lip cues and speech signals. The effectiveness of a multimodal sparse transformer network has been demonstrated by Song et al. (2022). Yang et al. (2022) attempted to carry out six cross/intra-modality functions, video-to-video, video-to-voice, video-to-text, voice-to-voice, voice-to-video, and

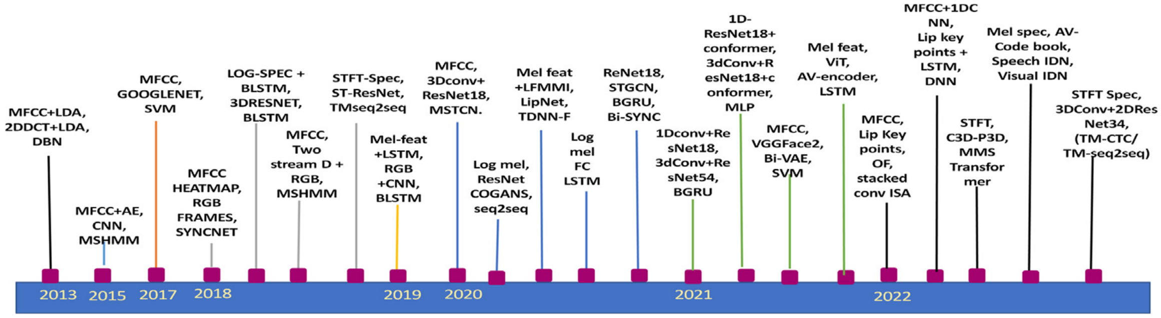

Fig. 6. Road map of AVSR DL-Types.

voice-to-text, in a single framework and proposed cross-modal mutual learning strategy using modality invariant codebook, which generates modality agnostic linguistic features that are required to carry out AVSR. Afouras et al. (2022) proposed two AVSR architectures and demonstrated that the performance of transformer seq2seq architecture surpasses that of transformer CTC architecture. Fig. 6 presents the remarkable progress in AVSR techniques employed over the years. Table 7 lists the overall summary of Deep learning approaches used in AVSR.

5. Silent speech recognition

This section investigates how powerful lip-reading can be without speech recognition and how it can contribute to various applications, such as silent video text transcription and spy cameras (Rothkrantz, 2017). The general ML pipeline is depicted in Fig. 7(a). First, the feature descriptor for each ROI of the video frame is extracted by applying transformations like DCT, DWT, PCA, and LDA or by extracting geometric features like height, width, area, the shape of lips, histogram, or texture features like LBP, or by finding out model-based features like AAM, ASM then these features are cascaded and fed to classifiers like HMMs, GMMs, DTW, ANN, KNN, SVMs, DBNs, so on for the classification. In the case of deep learning approaches, as depicted in Fig. 7(b), 2D CNN, 3D CNN, or Autoencoder is placed as a front-end to extract significant visual features. To depict the temporal relationship between video frames, the GRU, Bi-GRU, LSTM, Bi-LSTM, Transformer, and TCN are utilized. For classification, the softmax or CTC layer is used. Some works could perform SSR without ML and DL types, but the results were subpar. Al-Ghanim et al. (2013) created Arabic lip-reading software using motion estimation by extracting differences between successive frames and then finding first, second, and third-order moments for each viseme video frame. The documents are then saved in the viseme database. During recognition, the error in the first, second, and third moments of the entered video is computed using the stored database. Then the viseme with the minimum error function is considered. Grayscale images and image subtraction produced better results than RGB images. Nandini et al. (2020) employed the Active Shape Model's shape features corresponding to Kannada consonants. During testing, the similarity of shape features is calculated using Dice, Jaccard, and Cosine similarity functions, which describe the accuracy of recognized shapes, followed by the recognition of words from shapes.

The advancement of machine learning techniques facilitated the growth of SSR systems. Initially, Harte and Gillen (2015) applied DCT to lip images, then found its first and second derivatives, which were then cascaded and fed to HMM for viseme recognition. Almajai et al. (2016) investigated speaker-independent lip-reading on the Resource Management (RM) benchmark. They extracted Combined Shape and Appearance Model (CSAM) features, then applied MLLT and feature space Maximum-Likelihood-Linear Regression (fMLLR) with LDA to obtain adaptive speaker features, and later used these features to train GMM-HMM and context-dependent DNN-HMM models. The most effective model was a DNN-HMM model in which GMM likelihoods were replaced by DNN posterior probabilities. In Rekik et al. (2016), HMM, KNN, and SVM classifiers were trained on combinations of Histogram of Oriented Gradient (HOG) features extracted from Grayscale and depth images and Motion Boundary Histograms (MBH), where HOG2D+HOGdepth+MBH features with SVM provided the best performance on the MIRACL-VC1 benchmark.

Phoneme and viseme, speaker-dependent, and speaker-independent TCD-TIMIT word classification models employing GMM-HMM and DNN-HMM networks were proposed by Thangthai et al. (2017). Even though the speaker-dependent and speaker-independent phoneme models' accuracy was less than the viseme models' accuracy, the word accuracy from viseme based system was less compared to phoneme based system because the complexity in the viseme pronunciation dictionary was greater compared to the phoneme pronunciation dictionary. Bear et al. (2017) used HMMs to analyze how similar visual speech is across speakers and concluded that subjects have an identical collection of mouth gestures but use them differently. Modified Census Transform (MCT), MBH, DCT, HOG, and LBP were used for feature extraction, and SVM, KNN, and Sub-space KNN (using Decision tree and ensemble learning) were used for classification of 10 MIRACL-VC1 words (Sheshpoli and Nadian-Ghomsheh, 2018). In thesis work (Thangthai, 2018), A Deep denoising autoencoder was employed to obtain bottleneck layer features, and the extracted features were then analyzed using the t-SNE algorithm. DNN-HMM lip-reading was trained on the TCD-TIMIT benchmark. It was observed that visual features were clustered by speaker similarity as opposed to word similarity, whereas audio features were clustered word-wise as opposed to speaker-wise. Therefore, they also employed the feature transformation techniques LDA, MLLT, and FMLLR for adaptive speaker training. Researchers discovered that phoneme and bi-gram language models trained from TCD-TIMIT-provided text improved lip-reading accuracy.

Reviewing the tremendous success of DL frameworks in other applications, such as action recognition, researchers began implementing deep learning strategies for SSR. On MIRACLE-VC1, Garg et al. (2016) presented the face concatenation concept and VGGNet transfer learning technique. All the videos were stretched to contain 25 faces, and then the concatenated image was classified using pre-trained VGGNet. In Assael et al. (2016), the spatiotemporal convolution neural network (STCNN) that can convolve video frames across time and space was combined with Bi-GRU, and the framework was trained on the GRID corpus using Connectionist Temporal Classification (CTC) loss. In Chung and Zisserman (2017), activations from multiple towers were transmitted to VGGM to recognize 500 LRW words. Rahmani and Almasganj (2017) adopted Deep Bottle Neck Features (DBNFs) and a hybrid of DNN and HMM for CUAVE classification. Furthermore, the lattice-generating decoder combined the scores from the visual and language models. Stafylakis and Tzimiropoulos (2017) experimented with the 3D-ST-Convolutional network, followed by the un-pre-trained residual network with 34 layers, followed by the two-layer Bi-LSTM,

Table 7
Lip reading in AVSR using DL approaches.

Deep learning approaches								
Year	Ref.	Audio features	Video features	Classifier	Benchmark	Task Type	Result	limitations
May 2013	Huang and Kingsbury (2013)	MFCC+LDA	2D-DCT+LDA	Deep belief networks	IBM infrared headset dataset	continuous digits	WER:1.4% (noiseless) WER:12.2% (noisy)	Performance drops with noise.
Jun 2015	Noda et al. (2015)	MFCCs, log mel-scale filterbank (LMFB), final audio features extracted from deep denoising autoencoder	CNN	Multi-stream HMM	Japanese audiovisual dataset	400 words	WRR: 92%	Performance drops with noise.
July 2017	Yang et al. (2017)	MFCC features	GoogleNet	SVM classifier	AVLetters CUAVE	Alpha. digits	Acc: 83.4% Acc: 95.9%	Closed-set alphabet and digit identification
Aug 2018	Chung and Zisserman (2018)	MFCC represented as heat-map	Raw grayscale lower-face video frames	SyncNet Trained using contrastive loss	OuluVS2	Phrases	Accuracy: 94.1%	Cannot classify confusing words like million and billion
Nov 2018	Stafylakis et al. (2018)	Log Spectral features +BiLSTM	Frames within word boundaries are fed to 3D ResNet.	Bi-LSTMs	LRW-500 DEMAND dataset (for noise)	Words	MCR:1.9%	Closed-set word identification
Dec 2018	Paleček (2018)	MFCC features	ST-DNN features of depth maps and RGB frames	middle fusion (MSHMM)	TULAVD (50 words, 100 sentence)	word recognition (Czech lang.)	Accuracy: 94.9%	Robustness of work not tested on hundreds of hours of speech.
Dec 2018	Afouras et al. (2018a)	Spectrograms obtained after taking STFT	ST- ResNet	Transformer seq2seq	LRS2-BBC	sentence	WER:9.4% (without noise) WER: 5.9% (with noise)	Performance drops with noise.
Nov 2019	Makino et al. (2019)	Mel features fed (through an audio switch) Graphemic symbol fed to LSTM decoder	Cropped RGB mouth video frames fed to a 5-layer CNN to extract video features. (using video switch)	Fusion: a 5-layer stack of Bi-LSTM	YTDEV18 LRS3	Sentence	WER: 20.5% (without noise) WER: 7.2% (without noise)	Performance drops with noise.
May 2020	Martinez et al. (2020)	MFCC features	3D-Conv, ResNet-18	Multi-Scale TCN	LRW-500	Word recognition	1.04% error rate (without noise) 6.53% error rate (with SNR: -5 dB)	Performance drops with noise.
May 2020	Fernandez-Lopez et al. (2020)	log Mel features	ResNet for feature extraction and CoGANs are used to adapt to an unseen speaker.	Seq2seq network to fuse audio video network	TCD-TIMIT	sentence	10% improvement in character error rate	Adapting to an unseen speaker is challenging.
May 2020	Yu et al. (2020)	Mel features AudioNet: lattice-free maximum mutual information (LFMMI)	LipNet (Assael et al., 2016)	Stack of factored time delay neural network (TDNN-F)	LRS2	sentence	WER: 5.93%	Computationally complex.
Oct 2020	Wand and Schmidhuber (2020)	Log Mel-scale features extracted from OPENSMILE toolkit. Audio encoder: Stack of FC layers + LSTM +linear layer with Logits representing 51 neurons	Gray scale cropped mouth video frames fed to video encoder with a stack of FC layers + LSTM + linear layer with Logits representing 51 neurons	The model outputs the Audio Logits, and Video Logits, which are further processed through the FC layer to get merged Logits.	GRID	51 words	Acc: 98%	Closed-set word recognition.
Oct 2020	Liu et al. (2020)	Audio encoder: Speech wave fed to ResNet18+Bi-GRU	Stream1: gray-scale mouth video frames fed to 3DConv + ResNet34 + Bi-GRU Stream2: Lip graph fed to ST-GCN + Bi-GRU	Bi-directional Sync Block to fuse AV-features	LRW-500	words	Acc: 98.49% (without noise) Acc: 92.73% (with SNR: -5 dB)	Performance drops with noise.
Jan 2021	Liu et al. (2021b)	5 ms of raw audio with 0.25 ms stride. 1D Conv, ResNet-18, BGRU	3D Conv, ResNet-34, BGRU	Audio-visual late feature fusion using BGRU	LRW-500	Words	Accuracy: 98.26%(without noise) 92.10%(with SNR: -5 dB)	Word-level classification models. Using raw audio increases computation complexity.

(continued on next page)

Table 7 (*continued*).

Jan 2021	Liu et al. (2021a)	5 ms of raw audio with 0.25 ms stride. Audio encoder: 1D Conv, ResNet-18, BGRU	3D Conv, ResNet-34, BGRU	AV Encoder feature fusion: [FC + BN + Drop out + RELU layers with residual connection]Decision-level fusion of A+V+AV encoders	LRW-500	Words	Acc: 98.91%	Model constructed only for word-level classification. Using raw audio increases computation complexity.
Jun 2021	Ma et al. (2021b)	1D-ResNet-18, Conformer (Gulati et al., 2020) Encoder	Conv 3D, ResNet-18, Conformer Encoder	Multi-Layer Perceptron (MLP)	LRS2 LRW(for pre-training)+LRS2 LRW(for pre-training)+LRS3	sentence	WER: 4.2% WER: 3.7% WER:1.2%	To try the adaptive fusion method, which can learn to weigh audio and video modality based on noise levels
Nov 2021	Sayed et al. (2021)	MFCC/GFCC	7 × 7 stretched and cascade RGB video frames fed to pre-trained VGGFace2	Bi-VAE for feature fusion. Classifier: SVM	CUVAE	Isolated and connected digits	Acc: 99.42 (clean) 98.6% (with SNR: -5 dB)	The model implements word-level classification but not continuous speech recognition.
Dec 2021	Serdyuk et al. (2021)	Mel features	Vision transformer (ViT)	Feature concatenation fed to a 14-layer A/V encoder and LSTM decoder.	YouTube videos (training) LRS3 (only for testing)	sentence	WER: 23.1%	Requires more floating point operations compared to CNNs
Feb 2022	Yu (2022)	MFCC features (Empirical Mode Decomposition (EMD) and conditional maximum matching dynamic planning for feature matching)	The optical flow was considered to calculate lip motion information between neighboring frames using the key points on the upper and lower lip contours.	Stacked convolutional independent subspace analysis (ISA) network for digit classification	Own (YCLIP)	six digit string	Pronunciation RR for 2, 5, and 8 exceeded 90%, whereas for numbers 0, 4, and 7 recognition rate is meager.	EMD algorithm gives genuine feedback on the learner's pronunciation level but consumes more time. It relies on synchronization between lip movements and speech.
Feb 2022	Shashidhar et al. (2022)	MFCCS, CHROMA, MEL, CONTRAST, and TONNETZ cascaded and then fed to 1D CNN	Location of the outer lip coordinates extracted from DLIB provided to LSTM network.	Concatenated audio and video features are passed to DNN.	Own	24 spk. × 7 words × 240 instances	Accuracy: 91%	Experimented only on seven words.
Mar 2022	Shi et al. (2022)	MFCC features	ResNet	Audio-Visual-Hidden-Unit BERT (Transformer-LARGE +sequence to sequence)	LRS3 VoxCeleb2 (pretraining)	sentence	WER: 26.9%	Requires pre-training
Apr 2022	Song et al. (2022)	STFT	C3D–P3D network to extract visual features and optical flow features (pretrained with LRW)	multimodal sparse transformer network (MMST)	LRS2 LRS3	sentence	WER: 7.6% WER: 5.5%	Requires pre-training
June 2022	Yang et al. (2022)	Mel-spectrogram features Speech ID N/W: 1D Conv N/W+ temporal pooling Speech encoder: 1D Conv N/W 1D-Deconv N/W to generate mel-spectrogram	Visual ID N/W: ResNet18 Visual encoder: 3DResNet18 2D-Deconv N/W to reconstruct video	Modality invariant codebook to generate AVSR features (trained with KL-Divergence loss)	LRW-500	words	Acc: 98.5%	Cross-modal training ingests more time.
Dec 2022	Afouras et al. (2022)	STFT spectrogram	ST ResNet (3D Conv +2D ResNet34)	Transformer CTC model Transformer seq2seq model	LRS2 LRS3 LRS2 LRS3	sentence	WER: 3.7% WER: 12.3 WER: 9.4% WER: 8%	CTC is non-auto-regressive and thus faster, but seq2seq is autoregressive and, therefore, slower.

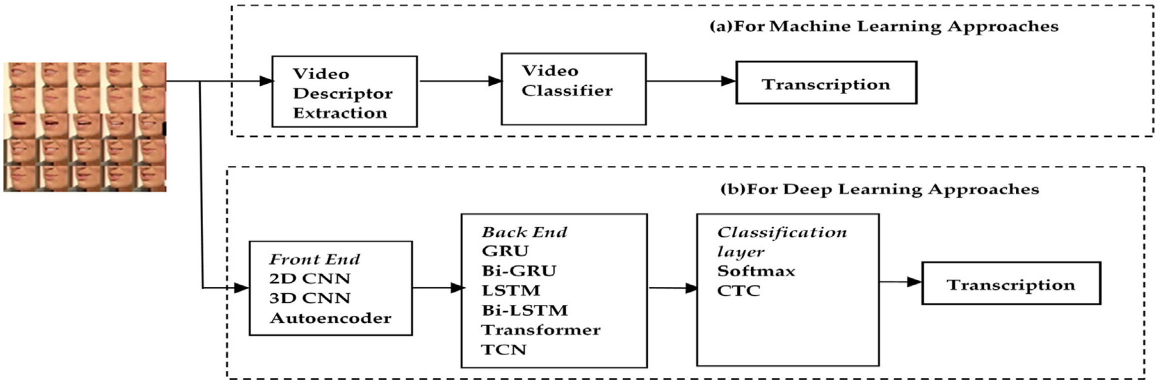

Fig. 7. General pipeline of SSR.

to extract features and classify the LRW500 dataset without specifying word boundaries.

In Stafylakis and Tzimiropoulos (2018), the 3D-ST-Convolutional network was followed by ResNet-18 to extract word-level embeddings for 500 LRW-seen words. They were classified using LSTM, and for few-shot learning classification, Probabilistic LDA Analysis was used, with the module being trained on 350 words and validated on the remaining 150 unseen words. Using CNN and Bi-GRU, Zimmermann (2018) calculated the average speaker-dependent TCD-TIMIT viseme accuracy of the seventeen selected speakers. NadeemHashmi et al. (2018) classified 7 × 7 stretched and concatenated MIRACL-VC1 phrase and word images by employing thirteen CNN layers and two batch normalization layers. Sterpu et al. (2018) used visual frontend 3D-ConvNet and 2D-ConvNet to extract features from the TCD-TIMIT dataset. These features were then fed to a sequence-to-sequence network consisting of two RNNs (Encoder and Decoder) trained with CTC loss. The model outperformed hybrid HMMs in viseme classification. Sindhura et al. (2018) cropped lips and cascaded them into a 4 × 4 frame. If the number of frames was less than 16, the last frame was appended repeatedly for word classification through Alexnet and InceptionV3.

The authors Lu and Li (2019) demonstrate that CNN-attention-based-LSTM outperforms CNN-LSTM for digit recognition. In Xiao (2019), researchers constructed Spatiotemporal feature pyramids to investigate multi-scale details in both ST dimensions and classified LRW-500 words and Grid sentences using them. Due to disparities in size, shape, and skin color between individual lips, speaker-independent lip-reading accuracy will be affected. So Fenghour et al. (2019) pre-processed GRID dataset lips by contour mapping using Sobel edge detection, and all lips from different individuals were made to look uniform by eliminating skin hues. This was then fed to LipNet architecture consisting of 3D CNN followed by GRU and was trained using CTC loss, and they achieved good results for unseen speakers. Parekh et al. (2019) used LSTM to classify descriptors extracted from Convolutional Autoencoders in a study involving the MIRACL and GRID benchmarks. In Large-scale VSR (LSVSR) (Shillingford et al., 2019), video frames were fed to spatiotemporal convolutions followed by Bi-LSTM, and the entire framework was trained with CTC loss to generate phoneme sequences. Finally, finite-state transducers (FSTs) were used to generate word sequences from phoneme sequences. Instead of 2D-CNN or shallow 3D-CNN, Weng and Kitani (2019) used two-round pre-trained Deep-inflated-3D-CNN (I3D) to extract LRW500 embeddings, followed by Bi-LSTM for word classification. Their research revealed that two-stream fused networks trained on grayscale and optical flow images outperformed a single-stream network trained on grayscale images. In Wang (2019), the coarse 2D-ResNet fine-grained model and the 52-layer 3D-DenseNet medium-grained model were used to capture features, and then a coarse-grained model was used to fuse the captured features from the two models using an attention mechanism. These features were input into a two-layer Bi-ConvLSTM for classifying LRW-500. In the MobiVSR (Shrivastava et al., 2019) architecture, LSTM was replaced by 3D depth-wise separable convolution layers and one-dimensional temporal convolution to mitigate the computations involved in CNN for developing a smartphone lip-reading interface. Scholars provided users with the option to select the MobiVSR Depth. Global-Average-Pooling (GAP) effectively compresses features along spatial dimensions but fails to preserve spatial details. To tackle this issue, Zhang et al. (2019) employed an ST-fusion framework (STFM) and Temporal-Focal block (TF block) to capture short-time temporal relations in convolutional seq2seq networks. Yao et al. (2019) proposed a simple sliding window method for generating sentence embedding for lip-reading using word models. Android application "Deep lip reader" was created for deaf people using VGGNet (Abrar et al., 2019), where 5 × 5 concatenated image frames were used as input for classification, and training was conducted using the MIRACL-VC1 benchmark.

In the 2020s, outstanding SSR systems-related research emerged. Deep Bi Convolutional LSTM network pre-trained with LRS2 and LRS3 dataset was used for classification (Courtney and Sreenivas, 2020), pre-training on LRS, then fine-tuning on LRW dataset was found to enhance model performance. Cheng et al. (2020) constructed a large database of synthetic poses for lip-reading tailoring robust 3D Morphable Model (3DMM) fitting that transforms any pose to a frontal image. The 3DMM augmentation expands the LRW dataset to include LRW-in-Large-Poses. The authors employed a cross-database experiment in which LRW words were used for training 3Dconv, ResNet-18, two-layer Bi-GRU, and the LRS2 database, with the same LRW words being used for testing. Researchers trained a continuous lip-reading system without using ground-truth annotations (Afouras et al., 2020a). For this purpose, knowledge was extracted from two neural networks, the speech-to-text network (teacher) and the lip-reading network (student), where the teacher's supervision was used to train the student. The proposed distillation technique was pre-trained with the VoxCeleb2 benchmark and further trained to evaluate the LRS2 and LRS3 benchmarks.

ST-visual frontend and attention-based transformers were used to classify visemes (Fenghour et al., 2020) Then, for word classification, the perplexity metric was used to map each visemes cluster to a word in the CMU dictionary. Even though viseme classification accuracy was high for the proposed model, word classification accuracy on the LRS2 benchmark decreased significantly. Shreekumar et al. (2020) presented Conditional Deep Convolutional GAN (DCGAN), conditioned on twelve viseme classes, and Twelve Unconditional Progressive Growing GAN (PGGAN), which can progressively train the generator and discriminator by initially training with low-resolution images and then gradually increasing the resolution to learn the finer details. Models were used to generate images of twelve viseme classes, each containing 200 images of size 256 × 256 × 3, and VGG16 was used to classify visemes with and without GAN-generated images. GAN-generated images used for training improved viseme recognition accuracy by 3.695% for models trained with PGGAN-augmented images and by 2.59% for conditional DCGAN. However, their model was only trained on 01M and 11F speakers from the TCD-Limit corpus. Xiao et al. (2020) proposed a deformation flow network (DFN) with an encoder that accepts the current and next frame of video and encodes it to two feature embeddings, followed by a decoder that accepts the concatenation of them to produce a deformation flow image. In order to reconstruct the next frame, the current frame was warped by deformation flow. DFN was trained using the absolute pixel-by-pixel deviations between the actual next frame and the subsequent reconstructed frames. Furthermore, the authors discussed the significance of a two-stream network and bidirectional knowledge distillation loss for jointly training original grayscale and deformation flow images. Experiments conducted on LRW500 demonstrated that deformation flow captures more information than optical flow. Interesting research work carried out by Zhang et al. (2020) on the LRW500, LRW1000, and Grid datasets identified that in addition to lip movements, the upper face and cheek deformations aid in lip reading. The results of the LR model will improve if the face is considered instead of the lips when lip-reading.

Zhao et al. (2021b) employed Lip-Corrector for sentence-level lip-reading on LRS2 and LRS3 benchmarks. The front end of the 3D and 2D ConvNets extracts features from the mouth region, the middle-end leverages the seq2seq transformer network to predict sentences, and the back end employs a pre-trained BERT (Bidirectional-Encoder-Representations from Transformers) model for sentence correction. In accordance with his work Pandey and Arif (2021) present the Repair model. It consists of a light image enhancement network that accepts dark images and brightens them to overcome poor lighting and an error corrector language model that alleviates error rates. In another work, ST convolutions and Spatial ConvNets were alternately placed to capture effective features (Tsourounis et al., 2021), which was then followed by Multi-scale temporal ConvNet for the LRW-500 benchmark classification. Researchers (Ma et al., 2021a) employed ShuffleNet v2 and the novel Depthwise-Separate Temporal ConvNet (DS-TCN) to reduce computation cost, but this resulted in a drop in LRW word recognition accuracy. Due to this, they employed the Knowledge Distillation (KD) framework to recover the efficiency loss. In KD, a student from the current generation is regarded as the teacher for the next generation, and the self-distillation procedure is carried out until no further progress in results is observed. Finally, an ensemble combines the predictions from many generations.

In a captivating paper (Sheng et al., 2022c), the authors propose the Adaptive-Semantic-Spatial-Temporal Graph Convolutional Network (ASST-GCN) to tackle SSR. Lower-half-face video frames were sent to the global stream, composed of 3DCNN+ 2DResNet, to extract global video features. The local stream was built using three modules: local motion feature extraction (LMFE), which accepts 32 × 32 grayscale patch sequences centered around 38 lip landmark points and forwards it to three 3D CNN layers to extract motion of lip contour features; local coordinate feature extraction (LCFE), which accepts Lip Landmark coordinates and delivers it to 1D CNN to fetch Landmark points features; and ASST-GCN framework, which takes concatenated lip contour feature vectors and landmark point feature vectors and explores local video features. Back-end networks consisting of multi-scale dilated TCN layers for word classification and seq2seq transformer networks for sentence classification concatenate and process the video features from the local and global streams. The proposed ASST-GCN was evaluated using the LRW500, LRS2, and LRS3 benchmarks. AVSR systems only work when both visual and audio inputs are present, but in the case of SSR, audio information is not available during testing; still, audio information can be used during training, so Kim et al. (2022a) adapted Visual Audio Memory (VAM) as a back-end module and demonstrated that the accuracy of SSR systems is improved when VAM is used in conjunction with the frontend baseline modules of LRW500, LRW1000, GRID, and LRS2.

Instead of going through entire video frames and then predicting the whole sentence, Lin et al. (2021) attempted simultaneous silent speech recognition, where a portion of the video clip was simultaneously processed to generate the corresponding sentence. For this purpose, the authors constructed a transducer with an adaptive memory containing 20 memory slots for storing visual features guided by dot product attention. If the current video clip differed from the one in the memory bank, the memory slot that was less often used was replaced with the most recent video clip's features. The visual encoder comprised a temporally truncated 3DCNN followed by a transformer seq2seq model with time-masked self-attention. Adaptive memory was leveraged to enhance the extracted visual features. With the aid of a transformer-based language model, a subsequent cross-model decoder predicts the following sentence based on partial video clip-enhanced features. On the GRID and TCD-TIMIT datasets, their model was trained and evaluated. Then Zhao et al. (2021a) adopted video-to-speech cross-modal pre-training utilizing LRW-500, audio was predicted from the concatenated disentangled lip movement feature that was obtained by k-means clustering of video features extracted from video encoder and speaker identity feature obtained from speaker encoder with speech as its input, and further this pre-trained video encoder was used for SSR of LRW and OuluVS2 benchmarks. Zhang et al. (2021b) developed a network to magnify lip movements and used it to improve SSR performance. Ren et al. (2021) proposed a collaborative distillation framework that is trained with a dynamic fusion loss to replace the traditional knowledge distillation framework in which a pre-trained audio teacher was used to transfer knowledge to a student via distillation. The framework includes a trainable teacher who can be modified based on student feedback. Annotation, a pre-trained audio tutor, and a pre-trained video tutor were integrated into the student framework for learning. The trainable teacher was trained using audio, video, and student feedback. The training consisted of two stages. During the first stage, the student was optimized using the teacher's knowledge and annotations, while the master was optimized using annotations and student feedback in the second stage. Benchmarks LRW500, LRS2, and LRS3 were used to evaluate the proposed framework. RGB lower half face frames were fed to the video stream network with 3D-ResNet18 front-end and Transformer seq2seq back-end. Raw audio was fed to the Audio stream network with 1D-ResNet18 front-end and Transformer seq2seq back-end. Instead of selecting training samples at random, a curriculum learning strategy was implemented. Here, both the sample size per batch and the difficulty level of classifying samples increase gradually. Later Takashima et al. (2021) adopted a pseudo labeling and knowledge distillation framework for one hundred LRW-500 out-of-class vocabularies by employing intermediate layer embeddings rather than output embeddings of audio and video encoders. First, the audio encoder was trained to estimate word probability using mel-spectrograms. Then, its weights were seized in order to train the video encoder using knowledge distillation loss and domain adaptation loss (loss between intermediate layer embedding of audio and video encoder).

Huang et al. (2022), leveraged a transformer encoder and transformer decoder to lip-read Grid corpus. Ma et al. (2022b) work investigates several training strategies to improve the word recognition accuracy of LRW-500, like using word boundary indicator vector, pre-training model with LRS2+LRS3+AVSpeech, time masking data augmentation followed by mix-up, utilizing Densely connected temporal Convolutional networks (DC-TCN) for back-end, the ensemble of all students and teacher models using self distillation framework. In Tan et al. (2022), a special type of sensing camera known as an event camera was leveraged to create the DVS-Lip dataset of 100 words and 40 speakers for lip-reading. As network inputs, the authors have utilized slow-rate event frames, which contain subtle spatial details but coarse temporal details, and fast-rate event frames, which contain subtle temporal details but coarse spatial details. Two branches of Multi-grained ST Features Perceived Networks (MSTP) accept slow-rate and fast-rate event frames. The Message Flow module (MFM) then combines these features to generate ST features, which are then forwarded to Bi-GRU for word classification. Kim et al. (2022b) enhanced their baseline (Kim et al., 2022a) by using multi-head VA memory (MVM) to resolve the many-to-one mapping between Viseme and Phoneme. The MVM outputs saved 'h' audio embeddings for the given visual embedding. MVM featured 112 memory slots with eight heads and was placed at four multiple temporal levels.

In Wang et al. (2022), the front-end framework 3DCvT, which consists of 3DCNN and three-stage vision transformers, was used to extract the best features, and Bi-GRU was employed as the back-end to classify LRW500 and LRW1000 words. The experts improved the word accuracy by providing the word boundary input, utilizing mix-up data augmentation to generate virtual training examples, and employing Label smoothing in the cross-entropy loss function. Two front-end streams were used to obtain Spatio-temporal (ST) and spatial features, which were then processed by two MSTCN back-ends in a fascinating study (Zhang et al., 2022b). Ultimately, features from the two streams were combined using decision-level fusion. Adaptive spatial graph Convolutional network (ASGM) was used to obtain spatial features from features extracted from 2Dconv+ResNet18 and the arrangement of lip landmarks (heat-map) obtained from the heat-map regression network. In order to extract ST features, five adjacent frames were fed through 3Dconv+ResNet18. This two-stream fusion network was evaluated using LRW500 and OuluVs2 phrases. Then, Feng et al. (2022) utilized audio data to enhance SSR. By introducing audio-driven 128 × 128 deformation flow (DF) images generated by the encoder–decoder module, they addressed the domain gap issue between the AV modalities. This prevents the trained SSR from focusing on skin color, lighting, and pose variations. The training consisted of three phases. In the first stage, the DF encoder–decoder network was trained with MFCC features and video frames to generate DF images, and in the second stage, the DF-SSR (teacher) was trained with DF images. In the third stage, DF-SSR transfers its knowledge to student SSR trained using video frames through knowledge distillation. In a separate study, Prajwal et al. (2022) created the TEDxext dataset using imperfect audio-to-text transcriptions. STCNNs were fed sub-video clips of five frames with stride one in order to extract visual features. Then, in Visual Transformer Pooling (VTP), spatial position encodings (SPE) were added to these features and given to the transformer encoder. Instead of using GAP on features, the researchers propose using attention-weighted average pooling and cascading those features to extract temporal embeddings, which are then passed to the encoder–decoder transformer to predict word probabilities for a single token at a time, and beam search was leveraged to infer complete sentence. Their model was trained using a combination of LRS2, LRS3, MV-LRS, and TEDxext and evaluated using LRS2 and LRS3 benchmarks for sentence-level SSR.

Sheng et al. (2022b) proposed a Variational Temporal Mask module (VTM) with an information bottleneck framework to extract keyframe features and discard unwanted frame features. VTM can be inserted with any baseline SSR to improve its performance. VTM is trained using KL loss to output a binary mask, so the masked features produce similar predictions as the trained baseline SSR. The KL divergence loss matches the output class probability distributions of the trained baseline SSR and the SSR+VTM network. Furthermore, Ivanko et al. (2022b) utilized SSR to assist vehicle drivers. Using the VOSK module, they first eliminated the redundant silent video frames without audio. The media pipe was employed to extract mouth ROI. Mix-up augmentation with label soothing was leveraged to alleviate overfitting, and an altered 3DResNet18 with squeeze and attention (SA) network followed by Bi-LSTMs was utilized to classify LRW-500 words. In addition, they evaluated the model using their Russian custom data set, RUSAVIC. Then, Ma et al. (2022a) supervised the SSR model with pre-trained SSR and ASR models to improve its accuracy. Using generated transcriptions, they trained the model in several languages, including Spanish, English, Italian, Portuguese, and French. Later, Huang and Zhang (2022) presented a novel dual tower ST-feature extraction framework that combined the features from both spatial and temporal towers and evaluated their framework on ten phrases from the OuluVS2 benchmark. Fig. 8 (a), (b), and (c) illustrate the progression of SSR techniques over time.

The progression of work in silent speech recognition using machine vision, machine learning, and deep learning are tabulated in Tables 8–10.

6. Voice from lips

Liopa (Liam McQuillan, 2022) is a startup that focuses on lip reading technology for applications such as voice-to-voiceless patients, keyword detection in CCTV, and speech activity detection in the absence of audio. Patients with tracheostomies and laryngectomies with speaking difficulty have benefited from their services. They offer facilities for individual client training. Their application can deal with varying lighting conditions and slight variations in the head pose. Fig. 9 depicts a general ‘‘voice for the voiceless’’ system pipeline. It aims to perform cross-domain mapping, in which video frames of a clip are mapped to their corresponding speech clip. During training, the loss function attempts to suppress the disparity between actual speech features (e.g., Mel features) and reconstructed speech features from the speech decoder. Here, training is conducted in a self-supervised manner, with the feature vector of the speech signal serving as a natural label for silent video clips. The speech encoder extracts encoded speech features from the original speech features, such as the Mel spectrogram, and uses LSTM/BGRU to learn temporal information. A speech decoder (mirrored version of the speech encoder) accepts encoded speech features from an LSTM and outputs the original speech features (Mel spectrogram). Using a speech reconstruction algorithm such as Griffin-Lim, it is possible to reconstruct speech signals. In parallel, the region of interest is extracted from video frames and then fed to a video encoder that extracts the Spatio-Temporal characteristics of video frames. The Speech decoder then converts these visual features into Mel features, and the speech reconstruction algorithm is employed to reconstruct speech. The training intends to suppress the reconstruction gap between original Mel features and revamped Mel features.

The quality of speech is measured using various performance metrics, like PESQ (Perceptual Evaluation of Speech Quality), WER (Word Error Rate), PER (Phoneme Error Rate), SII (Speech Intelligibility Index), STOI (Short Time Objective Intelligibility), ESTOI (Extended Short Time Objective Intelligibility), ViSQOL (Virtual Speech Quality Objective Listener), STMI (Spectro Temporal Modulation Index), Corr2D: 2D correlation between the actual spectrogram and the reconstructed spectrogram, SCS (Shifted Cosine Similarity) and MCD (Mel Cepstral Distance).

Milner and Le Cornu (2015) used GMM and DNN for speech reconstruction from GRID speaker 2 video features. Using the expectation–maximization algorithm, GMM was employed to model the joint distribution of LPC audio features and AAM video features, while in another method, DNN was used for regression. It accepted 2D-DCT audio features; a linear vocoder was then used to reconstruct speech, and it was found that GMM outperformed DNN. Ephrat and Peleg (2017) utilized CNN for silent video-to-speech conversion. In Le Cornu and Milner (2017) Regression model and Classification model using the trained Vector Quantization (VQ), a codebook was applied to two speakers of the Grid dataset, and the classification model outperformed the regression model. Ephrat et al. (2017) proposed the Two-Tower ResNet Encoder-Decoder framework, where the first tower accepts grayscale face frames and the second tower takes Optical-Flow images. Output feature embeddings are concatenated and fed to a decoder with dense layers capable of producing Mel-spectrogram output. Using a post-processing framework, consecutive Mel-scale spectrogram vectors are transformed into a linear-scale spectrogram that can synthesize a sound waveform based on example-based synthesis.

In Ephrat and Peleg (2017), Le Cornu and Milner (2017), Ephrat et al. (2017) and Akbari et al. (2018), speech spectrogram features

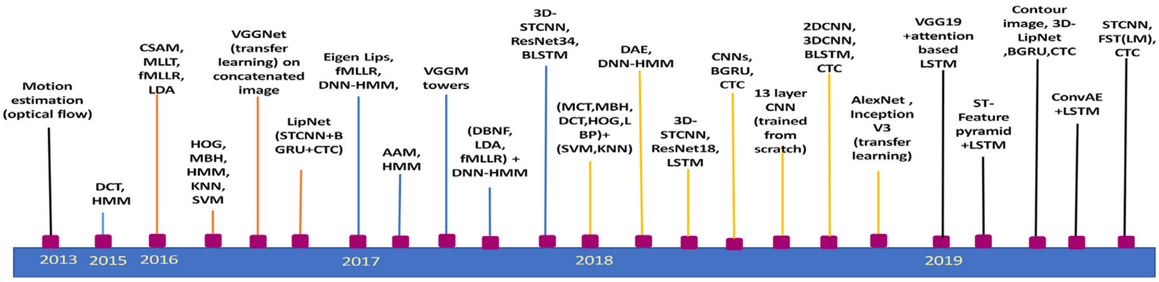
Motion estimation (optical flow)
DCT, HMM
CSAM, MLLT, fMLLR, LDA
HOG, MBH, HMM, KNN, SVM
VGGNet (transfer learning) on concatenated image
LipNet (STCNN+B GRU+CTC)
Eigen Lips, fMLLR, DNN-HMM,
AAM, HMM
VGGM towers
(DBNF, LDA, fMLLR) + DNN-HMM
3D-STCNN, ResNet34, BLSTM
(MCT,MBH, DCT,HOG,L BP)+ (SVM,KNN)
DAE, DNN-HMM
3D-STCNN, ResNet18, LSTM
CNNs, BGRU, CTC
13 layer CNN (trained from scratch)
2DCNN, 3DCNN, BLSTM, CTC
AlexNet , Inception V3 (transfer learning)
VGG19 +attention based LSTM
ST-Feature pyramid +LSTM
Contour image, 3D-LipNet ,BGRU,CTC
ConvAE +LSTM
STCNN, FST(LM), CTC
2013
2015
2016
2017
2018
2019

Fig. 8a. Road map of SSR (2013–2019).

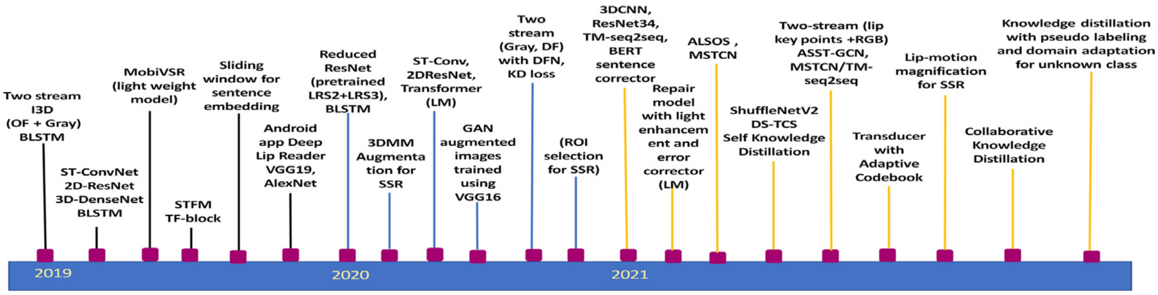
Two stream I3D (OF + Gray) BLSTM
ST-ConvNet 2D-ResNet 3D-DenseNet BLSTM
MobiVSR (light weight model)
STFM TF-block
Sliding window for sentence embedding
Android app Deep Lip Reader VGG19, AlexNet
Reduced ResNet (pretrained LRS2+LRS3), BLSTM
3DMM Augmentation for SSR
ST-Conv, 2DResNet, Transformer (LM)
GAN augmented images trained using VGG16
Two stream (Gray, DF) with DFN, KD loss
(ROI selection for SSR)
3DCNN, ResNet34, TM-seq2seq, BERT sentence corrector
Repair model with light enhancement and error corrector (LM)
ALSOS , MSTCN
ShuffleNetV2 DS-TCS Self Knowledge Distillation
Two-stream (lip key points +RGB) ASST-GCN, MSTCN/TM-seq2seq
Transducer with Adaptive Codebook
Lip-motion magnification for SSR
Collaborative Knowledge Distillation
Knowledge distillation with pseudo labeling and domain adaptation for unknown class
2019
2020
2021

Fig. 8b. Road map of SSR (2019–2021).

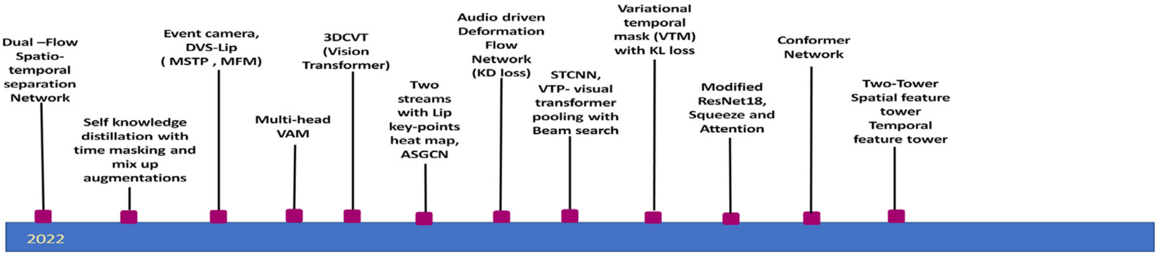
Dual –Flow Spatio-temporal separation Network
Self knowledge distillation with time masking and mix up augmentations
Event camera, DVS-Lip (MSTP , MFM)
Multi-head VAM
3DCVT (Vision Transformer)
Two streams with Lip key-points heat map, ASGCN
Audio driven Deformation Flow Network (KD loss)
STCNN, VTP- visual transformer pooling with Beam search
Variational temporal mask (VTM) with KL loss
Modified ResNet18, Squeeze and Attention
Conformer Network
Two-Tower Spatial feature tower Temporal feature tower
2022

Fig. 8c. Road map of SSR (2022).

Table 8
Lip reading in SSR using Computer vision approaches.

Computer vision approaches								
Year	Ref.	Input type	Video features	Classifier	Benchmark	Task Type	Result	Limitations
Dec 2013	Al-Ghanim et al. (2013)	Gray scale-cropped mouth video frames	Motion estimation using optical flow	Comparing 2nd and 3rd order moment of difference image with stored viseme features	ISWYS database (own)	34 phonemes (Arabic alphabets)	Accuracy:70%	Only phonemes were recognized.
Nov 2020	Nandini et al. (2020)	RGB-cropped mouth video frames	Active shape models (ASM)	ASM similarity using dice, Jaccard, and cosine similarity.	own	Kannada words	Accuracy:82.83%	Difficult to classify the words with the same homophenes

were compressed utilizing an autoencoder. To retrieve Spatio-temporal features from input video sequences, a framework consisting of seven-layer 3DCNN followed by LSTM was used, which was then fed to the decoder of autoencoder that was pre-trained to synthesize the speech spectrogram, and then NSRtools was used to generate waveform from the spectrogram, MSE, and correlation loss were used during training. Still, training was carried out in multiple stages. If lip-reading is considered a classification task, it is impossible to reconstruct speech from unseen phrases. Therefore, Kumar et al. (2019) have modeled the lip reading task as a regression task. They used multiview lip-reading images of the OuluVS2 benchmark for the first time to develop a language- and vocabulary-independent speech reconstruction system named Lipper. Lipper consisted of a VGG16 view classifier with decoding logic to test all possible combinations of 0,30,45,60 and 90-degree speaker views. The selected views were then fed to the appropriate speech reconstruction model, which consisted of 7 layers of Spatio-temporal CNN (STCNN) and two layers of Bi-GRU, which extracted LPC and LSP (Line Spectral Pairs) encoded audio vectors to generate speech. Experiments revealed that 0, 45, and 60 degrees provided the optimal combination of views. However, Lipper performed poorly in speaker-independent settings, and the generated sound was robotic. Later Vougioukas et al. (2019a) adopted Wasserstein's GAN to synthesize speech from video frames.

In Prajwal et al. (2020), scholars utilized chemistry lecture videos to compile their dataset. As a video encoder, stacked 3D-Conv layers with skip connections were used to draw out significant features of the talking face. LSTMs decoded these visual features to generate Mel-spectrograms and teacher enforcement was employed to accelerate the

Table 9
Lip reading in SSR using ML approaches.

Machine learning approaches								
Year	Ref.	Input type	Video features	Classifier	Benchmark	Task Type	Result	Limitations
May 2015	Harte and Gillen (2015)	Gray scale-cropped mouth video frames	DCT, its 1st and 2nd order derivatives	HMM	TCD-TIMIT	Visemes	Lip speaker acc: 52.74% Volunteers acc: 3454%	Only straight frames were used, 30 deg frames were not tried for classification.
Mar 2016	Almajai et al. (2016)	RGB-cropped mouth video frames	CSAM,MLLT, fMLLR, LDA	Context-Dependent DNN-HMM	(RM corpus-only front view)	Viseme, Phoneme, word	SI Word accuracy: 54%	A small amount of training data was used.
Jul 2016	Rekik et al. (2016)	RGB and depth-cropped mouth video frames	HOG2D+HOGdepth +MBH	HMM, KNN, SVM	MIRACL-VC1	Phrases Word	Phrases: Best SD acc: 96% SI acc:79.2% Words: Best SD acc: 95.3% SI acc:63.1%	There is still scope to improve Speaker independent accuracy.
Sep 2017	Thangthai et al. (2017)	Gray scale-cropped mouth video frames	Eigen lips: fMLLR	DNN-HMM	TCD-TIMIT	Word recognition	SD word accuracy: 48.74% SID word accuracy: 42.97%	The Eigen lip feature needs training.
Oct 2017	Bear et al. (2017)	RGB-cropped mouth video frames	Active Appearance Model features	HMM (P2V: Phoneme to Viseme maps (Bear et al., 2014) based on confusion matrices allows the variation in pronunciation)	AVLetters2	alphabets	SD word recognition Spk1: 15.9% Spk2: 25% Spk3: 45% Spk4: 38.4%	The accuracy obtained is less.
Jul 2018	Sheshpoli and Nadian-Ghomsheh (2018)	RGB and depth-cropped mouth video frames	MCT, MBH, DCT, HOG, and LBP	SVM, KNN, Sub-space KNN	MIRACL-VC1	Ten words	Best Accuracy SD: 95% SI: 52%	DCT+Cubic SVM was best for SI, and MBH with LBP+Sub-space KNN was best for SD.
Nov 2018	Thangthai (2018)	Gray scale-cropped mouth video frames	30-dimensional deep autoencoder (DAE) features followed by LDA-MLLT and fMLLR feature transformation	DNN-HMM	TCD-TIMIT	6000-word vocabulary	Best Accuracy SD: 61.82% SI: 59.42%	Lip reading is carried out in a studio environment, Requires correct labels to train.

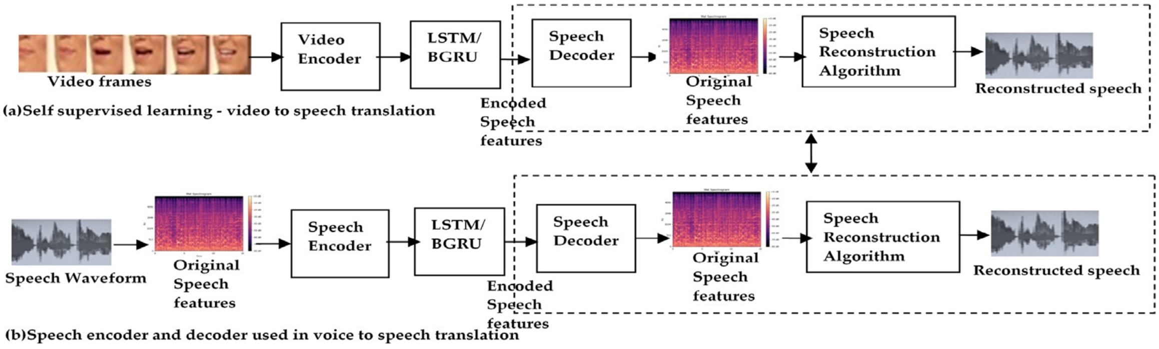

Fig. 9. General pipeline of Voice from Lips.

training process. In Michelsanti et al. (2020), a video encoder consisting of five 3D convolutions followed by GRU was used to extract video features from the cropped mouth or cropped face ROIs. To predict WORLD Vocoder features, the extracted video features were then fed to five decoders, SP, AP, UVU, F0, and Visual speech recognition (VSR) decoder. Then the speech was synthesized using the WORLD synthesis algorithm. Chen et al. (2020) leverage the duality of text-to-lip and lip-to-text networks to boost the performance of dual networks. Both networks are initially trained independently during supervised training to reduce the loss between ground truth and classifier prediction. In the second phase, during unsupervised training, the unlabeled text is provided to the lip video synthesis model to generate lip video. Then this pair of text and generated lip video is used to improve the LR model. Similarly, an unlabeled lip video is provided to the LR model to generate text, and this pair of video and text is then used to improve the lip generation model. A text-to-speech converter was used to generate voice, and two methods were employed to align text and videos. In the first method, PPLFA (Penn Phonetics Lab Forced Aligner) expanded text based on frame duration; in the second method, the attention mechanism was tested. To align output video and generated speech

Table 10
Lip reading in SSR using DL approaches.

Deep learning approaches								
Year	Ref.	Input type	Front End	Back End	Benchmark	Task Type	Result	Limitations
2016	Garg et al. (2016)	(a) 5 × 5 Concatenated RGB lower face images (stretched 25 frames)	VGGNet Transfer Learning technique	With-out LSTM	MIRACLE-VC1	Words, Phrases	Best Accuracy SD (concatenated frames) Words: 56% Phrases: 33%	Tested on a constrained dataset.
		(b) Sequence of RGB-cropped face video frames	VGGNet Transfer Learning technique	With LSTMs				
Dec 2016	Assael et al. (2016)	RGB-cropped mouth video frames	spatiotemporal convolution neural network (ST-CNN) (LipNet)	Bi-GRU, trained using CTC loss	GRID	28775 sentences	SD Acc: 95.2% Unseen speaker WER: 11.4% Overlapped speakers WER: 4.8%	It was tested on a constrained dataset with 51 words of vocabulary.
Mar 2017	Chung and Zisserman (2017)	RGB-cropped lower-half-face video frames	Concatenated activations from 25 towers and activations at pool1	Modified VGGM base network	LRW-500	500 words	Accuracy: 61.1%	Failed to classify ambiguous words like "report" and "reports"
Apr 2017	Rahmani and Almasganj (2017)	Gray scale-cropped mouth video frames	Deep Bottle Neck Features (DBNFs) using autoencoder followed by LDA and fMLLR feature transformation	Hybrid DNN and HMM visual model, followed by lattice generating decoder to fuse language and visual models.	CUAVE	Isolated digits and connected digits (0-9)	PHONEME ERROR RATES (PER): 35.1%	Digit recognition carried out at the phoneme level
Aug 2017	Stafylakis and Tzimiropoulos (2017)	RGB-cropped mouth video frames	3D-STCNN+ ResNet34	Bi-LSTM	LRW-500	words	Accuracy: 83%	Failed to distinguish words like spend and spent, living and giving.
Apr 2018	Stafylakis and Tzimiropoulos (2018)	RGB-cropped mouth video frames	3D-STCNN+ ResNet18	LSTM	LRW-500	words	For seen 500 words EER:11.92% For low shot learning (150 unseen words), error: 43.8%	To attempt sentence recognition by utilizing word recognition.
July 2018	Zimmermann (2018)	RGB-cropped mouth video frames	CNN	Bi-GRU with CTC	TCD-TIMIT	Viseme acc. for selected 17 spks.	Accuracy: For straight frames: 52.28% For 300 frames (53.64%)	Computational cost and training time is proportional to the number of views.
Aug 2018	(NadeemHashmi et al., 2018)	Concatenated Grayscale image consisting of stretched 49 frames (7 × 7)	Thirteen layers CNN	–	MIRACLE-VC1	Words and phrases	Accuracy: 52.9%	The model was trained from scratch.
Oct 2018	Sterpu et al. (2018)	Gray scale -cropped mouth video frames	2D-CNN/3D-CNN	Bi-LSTM with CTC	TCD-TIMIT	Viseme accuracy for 59 subjects	Speaker-dependent viseme accuracy: 66.27%	Poor accuracy for phonemes/oy/,/ao/, and/aw/.
Dec 2018	Sindhura et al. (2018)	Grayscale cascaded lips (4 × 4)	Alexnet and InceptionV3	–	MIRACLEVC1	Ten words	Best SD Acc: 86.6% SI Acc:37.1%	Poor speaker-independent accuracy.
Jan 2019	Lu and Li (2019)	RGB-cropped mouth video frames	CNN-VGG19	Attention-based LSTM	Own	(0-9) Digits, 6 spks.	SD Accuracy: 88.2%	The speaker-independent model was not explored.
Jan 2019	Xiao (2019)	RGB-cropped mouth video frames	Embedding 3D Feature Pyramid Attention (3D-FPA) module inside LipNet or ResNet+LSTM		Grid (LipNet) LRW-500 (ResNet+LSTM)	Sentences words	Grid: WER:0.141 LRW: Accuracy is 79.2%	Some words' accuracy in LRW is poor (e.g., Under, These, etc.)
Mar 2019	Fenghour et al. (2019)	Contour images of grayscale-cropped mouth video frames	LipNet (3D CNN)	LipNet(Bi-GRU) trained with CTC loss	GRID dataset	Sentences	Unseen speaker 1 WER: 2 0.6% Unseen speaker 2 WER: 17%	It is necessary to eliminate noise from contour images. It does not account for the varying sizes of the subjects' lips.

(continued on next page)

Table 10 (*continued*).

Mar 2019	Parekh et al. (2019)	Grayscale-cropped mouth video frames	Convolutional Autoencoders	LSTM	MIRACL-VC1 GRID BBC-27	Words Sentences (5spk) Words (LRW-27spk)	SD Acc: 98% SI Acc: 63.22% Acc: 84.80% Acc: 77.421%	CAE+LSTM is superior to CNN+LSTM/ CNN+HMM, but CAE must be trained separately to learn optimal features.
Sep 2019	Shillingford et al. (2019)	RGB-cropped mouth video frames	ST-CNN	FST (LM) with CTC loss	LSVSR	Sentence	WER: 40.9%	Requires Language model for decoding
Sep 2019	Weng and Kitani (2019)	(Grayscale +optical flow (OF))-cropped lower face video frames	Inflated 3D ConvNets (pretrained with imagenet and Kinetics)	Bi- LSTM	LRW-500	words	Accuracy: 84.07%	When employing Optical Flow, illumination conditions should not vary.
Sep 2019	Wang (2019)	RGB-cropped lower-face video frames	ST ConvNet followed by dual branch,2D ResNet, and 52 layers 3D-DenseNet	The coarse grain model consists of an attention mechanism and Bi-ConvLSTM	LRW-500	words	Accuracy: 83.34%	Model complexity is more
Sep 2019	Shrivastava et al. (2019)	Grayscale cropped lower face video frames	Two 3D depthwise separable convolutions followed by Residual blocks whose depth can be selected.	Two temporal convolution layers	LRW-500	words	Accuracy: 70.8% with 6x fewer parameters	Confusion with similar visemes, such as Billions-Millions and General-Several.
Oct 2019	Zhang et al. (2019)	Lower half-face video frames	3DConv + ResNet18 + STFM	Conv-seq2seq model with TF-block	LRW-500 LRS2 LRS3	Words Sentences	WER: 16.3% WER: 51.7% WER: 60.1%	The obtained WER can be reduced further.
Oct 2019	Yao et al. (2019)	RGB-cropped lower-face video frames (Sliding Windows for Sentence Embedding)	3D spatio temporal convolution followed by 2D ResNet-34	Bi-LSTM	Mandarin LRW-1000	words	Accuracy: 38.87%	Visual keyword spotting is ineffective for words of varying lengths.
Nov 2019	Abrar et al. (2019)	5 × 5 Concatenated RGB lower face images consisting of stretched 25 frames	VGG19 transfer learning Alexnet transfer learning	- -	MIRACLE-VC1	Ten words	SI Accuracy: 60% SI Accuracy: 41%	The entire cascaded image must be resized to the input size of CNN, resulting in low resolution images when resized.
Feb 2020	Courtney and Sreenivas (2020)	RGB-cropped lower-face video frames	Reduced-ResNet (pretrained with LRS2+LRS3)	Deep Bi Convolutional LSTM(pretrained with LRS2+LRS3)	LRW-500	words	Accuracy: 83.4%	Pretraining models consume time.
May 2020	Cheng et al. (2020)	Grayscale-cropped lower-face video frames	3D Conv+ResNet18	Bi-GRU	LRW-500, LRS2-500 words	Words (Cross-modal experiment) Words	Accuracy: 83.20% Accuracy: 59.6%	Large training data employed (LRW500 + 3DMM Augmented) LRS2(same 500 words as in LRW))
May 2020	Afouras et al. (2020a)	Audio input to Teacher ASR model Grayscale-cropped lower face video frames input to student VSR model	Teacher ASR model and Student VSR model	Cross-modal distillation of an ASR teacher into a student VSR model.	LRS2 LRS3	Sentences	WER:58.2% WER:65.6%	No annotations are necessary, but a misalignment in synchronization between the teacher and student timings will cause a problem.
Nov 2020	Fenghour et al. (2020)	ST-3D ConvNet with filter dimension for five frames+2D ResNet	Viseme classifier (transformer)	word detector (Generative Pre-Training Transformer) for perplexity calculation	LRS2	sentences	WER:35.4% SAR(Sentence accuracy rate): 33.4%	Under varying illumination, SAR decreases.

(continued on next page)

Table 10 (*continued*).

Nov 2020	Shreekumar et al. (2020)	RGB-cropped mouth video frames (synthetic images using DCGAN, PGGAN)	VGG16	-	TCD-TIMIT	12 Viseme classification (speakers 01M and 11F)	Best Viseme accuracy: 01M: 53.5% 11F: 54.24%	Focuses on speaker-dependent viseme classification; does not address sentence recognition or speaker-independent models.
Nov 2020	Xiao et al. (2020)	Grayscale and deformation flow cropped lower half face video frames fed to two stream network	3D Conv + ResNet18 (for grayscale) 2D Conv + ResNet18 (for deformation flow) DF network: encoder and decoder	Bi-GRU (trained using bi-directional knowledge distillation loss)	LRW-500	words	Accuracy: 84.13%	Two streams increase the computational complexity.
Nov 2020	Zhang et al. (2020)	Different types of ROIs for lip reading, including full face, cropped lips, and face cutout, are investigated.	3D-ResNet18 for LRW LipNet for Grid	3D-ResNet18 for LRW LipNet for Grid	LRW-500 LRW-1000 Grid	words	Best Accuracy for face cutout: 85.02% 45.24% WER: 2.9%	Using detected eye and nose landmarks, facial alignment is performed. Future efforts can eliminate face alignment.
Apr 2021	Zhao et al. (2021b)	Grayscale-cropped lower-face video frames	3D CNN+ ResNet34	Transformer-based seq2seq model and pre-trained BERT (LM) for sentence prediction and correction	LRS2 LSR3	sentence	WER: 47.1% WER: 58.2%	Combining lip reading with a Language Model trained on a billion-character benchmark improves the accuracy of lip reading. This model was not trained explicitly with LRS sentences.
May 2021	Pandey and Arif (2021)	Preprocessing: Light enhancement network (GLADNet)	3D CNN+ squeeze and excitation ResNet34	Bi-GRU The whole model was trained using CTC loss Postprocessing: Error correction using a deep de-noising autoencoder, and bi-directional count-based trigram LM	GRID	Sentence	WER: 20.5% By using the repair model for seen data, there was a 61.9% reduction, and for unseen data, there was a 16.3% reduction in WER.	Poor performance on unseen data.
May 2021	Tsourounis et al. (2021)	Grayscale-cropped mouth video frames (Aligned face images leveraging mean face alignment technique of Martinez et al., 2020)	Modified ResNet18 includes Alternating Spatiotemporal and Spatial Convolutions (ALSOS) modules in between ResNet (pre-trained on LRW) blocks	Multi-Scale Temporal Convolutional Network (MSTCN) pretrained on LRW	LRW-500	words	Accuracy: 87.01%	Model complexity is more (41.23M parameters)
Jun 2021	Ma et al. (2021a)	Grayscale-cropped lower-face video frames	3D conv+ ShuffleNet v2 (Knowledge distillation in generations of models)	Depthwise separable TCN (Knowledge distillation in generations of models)	LRW-500	words	ResNet18 + MS-TCN KD ensemble Acc: 88.5% ShuffleNet V2+ DS-TCN KD ensemble Acc: 85.5%	Training is carried out in generations (need an ensemble of models)
Aug 2021	Sheng et al. (2022c)	RGB-cropped lower face video frames + 32 × 32 grayscale patch sequences centered around 38 lip landmark points +landmark points	Two stream networks with Local stream with LMFE, LMCE, ASST-GCN, and Global stream with 3DCNN+2DResNet18	Multiscale dilated TCN (for word classification) Seq2seq transformer network (for sentence classification)	LRW-500 LRS2 LRS3	Words Sentences	Acc: 85.7% WER: 55.7% WER: 62.7%	Model complexity is greater. Landmark points do not work for profile views.

(*continued on next page*)

Table 10 (*continued*).

Oct 2021	Kim et al. (2022a)	Grayscale-cropped lower-face video frames Mel-spectrogram for audio encoder	LRW500,1000 (ResNet18+BiGRU) GRID (LipNet) LRS2 (WAS) Audio encoder (conv2D+Residual blocks)	Visual Audio Memory (VAM) trained with (classification loss + audio imprinting loss + address matching loss)	LRW-500 LRW1000 GRID LRS2	Words sentences	Acc: 85.4% Acc: 50.83% WER: 2.7% WER: 46.2%	Good accuracy without using a language model, but requires pre-training of a model with large datasets, and the model will fail if audio and video are not synchronized.
Oct 2021	Lin et al. (2021)	Gray scale-cropped mouth video frames	Truncated 3D-CNN Seq2seq transformer network with time-masked self-attention	Attention-guided adaptive memory to enhance visual features Cross-model decoder with a transformer-based language model	GRID TCD-TIMIT	Sentence	WER: 2.738% PER: 56.02%	Utilizes language model. For the restricted-grammar GRID dataset, accuracy is higher, but TCD-TIMIT sentences perform poorly.
Oct 2021	Zhang et al. (2021b)	Five adjacent video frames (RGB cropped lower face video frames)	Lip motion magnification encoder–decoder framework to magnify lip movements SSR: ST-Conv + 2DResNet18	Bi-GRU	OuluVS2	Ten phrases	Acc: 92.5% (with magnification) Acc: 91.1% (without magnification)	Extra preprocessing step (Magnification) added to SSR
Oct 2021	Zhao et al. (2021a)	Grayscale-cropped lower-face video frames Mel-spectrogram (for speech)	[Video encoder: (3DCCN+2DCNN)+Codebook+ Speaker identity encoder] pre-trained on LRW to predict speech using speech decoder (1D transposed convolutions+ Residual blocks)	Trained Video encoder+ LSTM or Bi-GRU	OuluVS2 LRW-500 LRW-1000	Phrases	Acc: 96% Acc: 75.8% Acc: 36.5%	Needs pre-training
Nov 2021	Ren et al. (2021)	Raw audio for the audio stream and RGB cropped lower face video frames for the video stream	1D ResNet18 for audio stream and 3D ResNet18 for video stream. (Collaborative distillation with dynamic fusion loss employed for training)	Transformer seq2seq	LRW-500 LRS2 LRS3	Words Sentence Sentence	WER: 14.3% WER: 49.2% WER: 59%	Model complexity is more.
Dec 2021	Takashima et al. (2021)	Grayscale-cropped mouth video frames Mel-spectrogram (for speech)	Stacked Conv layers (Audio-Teacher) Stacked Conv layers (Video-Student)	Knowledge distillation loss with pseudo labeling and unsupervised domain adaptation for unknown classes	LRW-500	Words	Word accuracy (unknown classes): 52.46%	Word accuracy can be improved using more efficient video encoder architecture.
Jan 2022	Huang et al. (2022)	RGB-cropped mouth video frames	Pretrained VGG16 (separate ST stream)	Transformer encoder–decoder network	GRID	words	Word accuracy: 45.81%	Poor word accuracy
Apr 2022	Ma et al. (2022b)	(Grayscale lower half face) five consecutive video frames Time masking and mix-up augmentations	3D Conv +2DResNet18 Pre-trained with LRS2+LRS3+AVSpeech	DC-TCN (ensemble of all students and teacher models using self-distillation)	LRW-500	words	Word accuracy: 94.1%	Training is carried out in generations (need an ensemble of models). Requires pre-training of a model with a large dataset
June 2022	Kim et al. (2022b)	Grayscale-cropped lower-face video frames Mel-spectrogram for audio encoder	LRW500,1000 (ResNet18+BiGRU) GRID (LipNet) LRS2 (WAS) Audio encoder (conv2D+Residual blocks)	Multi-head VA memory (MVM) trained with (cross-entropy loss/CTC loss + cosine similarity-based reconstruction loss + contrastive loss) MVM was applied at multiple temporal levels	LRW-500 LRW1000 LRS2	Words sentences	Acc: 88.5% Acc: 53.82% WER: 44.5%	Without using a language model, the accuracy is high. Solve homophenes but requires pre-training of a model with large datasets, and the model fails if audio and video are not correctly synchronized.

(*continued on next page*)

Table 10 (*continued*).

June 2022	Tan et al. (2022)	High-rate and low-rate event frames of lip movements captured by event camera	MSTP with Message flow module	Bi-GRU	DVS-Lip	words	Unseen speaker Accuracy: 72.10%	Requires Event camera for lip reading
July 2022	Wang et al. (2022)	Grayscale cropped lower face video frames	3DCVT	Bi-GRU	LRW-500 LRW-1000	words	Word accuracy: 88.5% 57.5%	Adding transformer blocks increases the computational complexity.
Aug 2022	Zhang et al. (2022b)	Grayscale cropped lower face video frames	Two streams: 1. Features from 3Dconv+ResNet18 2. Features from (2Dconv+ResNet18) and Heat map Regression network fed to ASGM	MSTCN's	LRW-500 OuluVS2	Words Phrases	Acc: 87.4% Recognition Acc: 94.2%	Two streams increased computational complexity.
Aug 2022	Feng et al. (2022)	*DF generator:* *Audio encoder:* MFCC input to modified VGGM *Video encoder:* RGB lower half face video frames input to ResNet18 *AV-decoder* (Deconv layers) accepts concatenated AV features and outputs 128 × 128 DF	DF-SSR (teacher) Frontend: modified ResNet18 Input: DF images Note: Student SSR architecture is the same as DF-SSR, but input is video frames	DF-SSR (teacher) Backend: Bi-GRU (for word)/ conformer Network(for sentence) Trained using KD-loss, CTC loss, and cross-entropy loss	LRW-500 LRS2	Words Sentence	Acc: 89.9% WER: 59.8%	Involves a three-stage training process.
Sep 2022	Prajwal et al. (2022)	Sub-video clips (RGB lower half face) of five frames with stride one	STCNN + VTP	Transformer Encoder-Decoder with beam search	LRS2 LRS3	sentences	WER: 22.6% WER:30.7%	The study employed the Language Model and computationally complex modules, and training was conducted in two phases.
Sep 2022	Sheng et al. (2022b)	RGB lower half-face video frames	Squeeze and extract + Conv3D +Reset18	VTM+GRU	LRW-500 LRW-1000	Words	accuracy: 86.92% accuracy: 49.74%	Introducing VTM has little effect on the model performance.
Oct 2022	Ivanko et al. (2022b)	RGB-cropped mouth video frames	Altered 3DResNet18 with Squeeze and Attention (SA)	BiLSTM	LRW-500 RUSAVIC	Words Driver request sentences	accuracy: 88.7% accuracy: 64.09%	Although the model achieved good accuracy for LRW-500 word classification, it could have worked better on the RUSAVIC dataset.
Nov 2022	Ma et al. (2022a)	Grayscale lower half face video frames (for SSR) Speech signal (for ASR)	SSR: 3DConv + 2DResNet18 + conformer (12-layer) ASR: 1DResNet18 + conformer (12-layer)	Transformer Decoder (CTC loss + cross-entropy loss)	LRS2 LRS3	sentences	WER: 25.5% WER: 31.5%	The model requires pre-training with large datasets. (LRW, AVSpeech, etc.)
Dec 2022	Huang and Zhang (2022)	Grayscale-cropped mouth video frames	Temporal feature tower: shallow 3DCNN+ResNet18+ MSTCN or BiGRU Spatial feature tower: 3DCNN+ResNet18+GAP		OuluVS2	10 Phrases	Ten runs mean accuracy: 94.5%	Having two towers increases the complexity of the model.

for talking face generation, the same duration is input to both the lips-to-face model and the text-to-speech model.

Using ultrasound-tongue images and lip images, line spectral frequencies (LSF), Fundamental frequency (FO), and Unvoiced/Voiced (U/V) flag features of speech are predicted using three separate Convolutional Deep Autoencoders (DAE), and then they are used to generate a speech wave utilizing a CNN-based vocoder (Zhang et al., 2021a). Articulatory-to-acoustic mapping is preferred over articulatory-to-text mapping because acoustic speech contains more details that cannot be embedded in the text, such as emotion, prosody, and word emphasis. In contrast to video-to-text, video-to-speech can be trained using an unconstrained data dictionary under natural supervision without labeling or segmentation. The feature vector of the associated audio signal will serve as a natural label for the short video segment. Thus, in another study Vougioukas et al. (2021), mapping silent video frames to speech was attempted using Dual Inverse Network Optimization (DINO) forward architecture consisting of a video encoder, single-layer GRU, and audio decoder. The generator receives video as input and converts it to a speech waveform. A Discriminator has two encoders, an audio encoder proceeded by a two-layer GRU, and an identity encoder of a person. The embeddings from the two encoders are cascaded and fed to a frame decoder, which generates a video sequence. As a result, Discriminator accepts speech waveform and person identification photo as input and converts it to video; tests were conducted on 33 speakers of the GRID benchmark. Their method captured audio and visual mapping effectively. Converting video-to-text followed by text-to-speech conversion increases the network's complexity, Hence, authors Yadav et al. (2021) have employed a variational autoencoder composed of an audio decoder that generates audio based on audio features, an audio encoder that outputs audio features, a video frame encoder that outputs video features, and LSTMs that learn temporal information present in audio and video frame streams. The audio and video feature vectors' mean and variance are determined to obtain the posterior distribution, and their KL divergence loss is minimized during training. The posterior distribution is obtained from the video frame encoder network during the test and then fed to the audio decoder network to generate the audio waveform.

Lip reading is also used for speech inpainting (Morrone et al., 2021). In the speaker-independent Speech inpainting system, the acoustic features and 68 facial landmarks motion vectors were cascaded and given to the Bi-LSTM and CTC networks in order to generate an inpainted spectrogram. Their model was evaluated using the GRID benchmark. The authors Hong et al. (2021) propose a visual voice memory consisting of value-memory slots for storing speech descriptors and key-memory slots for storing visual lip motion descriptors corresponding to the Mel-spectrogram speech features. VV-memory is trained to remember lip motion descriptors and their corresponding speech descriptor pairs using the cosine similarity function. When VV-memory receives a visual descriptor, it searches the key memory for the slot of the stored speech descriptor and retrieves the matching speech descriptor. The authors Kim et al. (2021) believe that when predicting the Mel-spectrogram, focusing on a long video clip, that is, global level lip motions, can make the mapping of viseme to phoneme clearer than focusing on a short local window of the lip motion video clip. Thus they propose Visual Context Attention GAN that considers both local and global lip motion features to tackle a lip-to-speech task such as lip-video to Mel-spectrum image translation. VCA-GAN comprises multiple generators that can generate mel-spectrograms from low to high resolution, as well as their discriminators. The first generator synthesizes coarse speech Mel-spectrogram, while other generators improve this by tailoring global lip motion features. The synchronization loss function was used with GAN loss and Reconstruction loss (Mel loss) to ensure audio-visual synchronization.

Previous works are based on autoregressive architecture, in which speech generation depends not only on the current frame but also on previous frames. This introduces a delay in inference time that increases linearly with video length. So authors He et al. (2022) propose GlowLTS, which consists of a visual encoder to extract visual features, a non-auto regressive condition module that generates coarse Mel-spectrogram, and a flow-based decoder for a more natural and realistic Mel-spectrogram synthesis from coarse Mel-spectrogram. In order to bring realism and clarity to synthesized speech, authors Mira et al. (2022) have used Wasserstein GAN loss based on waveform judgment and spectral power judgment and trained this end-to-end GANs framework along with three other losses like perceptual loss, spectral power loss, and MFCC loss. Following this, Zeng and Xiong (2022) utilized a pre-trained speaker encoder module to improve the reconstructed speech quality. To generate Mel-spectrogram from talking video frames, the Tacotron2 seq2seq network was given concatenated visual embedding extracted from the Content encoder and speaker identity embedding from the speaker encoder. Then Varshney et al. (2022) integrated transformers into Variational Autoencoder (VAE) to perform speaker-dependent silent video-to-speech synthesis on the GRID and Lip2Wav benchmarks. The framework includes a video transformer, an audio transformer, and a speech decoder transformer. Face encoders extract the visual feature input to the visual transformer and positional embedding for each video frame. For each MFCC feature frame, the audio encoder simultaneously extracts the audio feature input to an audio transformer and embeds positional information. The KL loss between video feature distribution and audio feature distribution is suppressed at each training iteration. The speech decoder transformer generates the MFCC feature at each iteration from the sample point chosen from the latent space.

Leveraging Variational Autoencoder (VAE) and Wasserstein-GAN (Hegde et al., 2022) tackled the ill-posed multi-speaker lip-to-speech problem. Mel-spectrogram audio features were extracted using a pre-trained ASR module whose weights had been frozen. The 3DConv network concurrently extracted video features from five face-ROI frames with stride one. The speech features are then mapped to the shared speech-video latent feature space using VAE. Training minimizes both global and local KL-divergence losses. The features of the visual encoder are mapped to the same shared hidden feature space so that the model can function without speech input during testing. Consequently, a sample point from the hidden feature space and speaker identity embedding are sent to the speech decoder during speech reconstruction to generate intelligent and realistic speech. The framework performed admirably on the GRID, TCDTIMIT, LRW, and LRS2 benchmarks. Recent work Wang and Zhao (2022) has constructed a light model called Fast-Lip-to-Speech (Fast-LTS) by optimizing the Lip-to-Speech non-autoregressive model with respect to memory requirements and output latency. The Acoustic-Conditional-Network (ACN) was fed visual features extracted from the Spatio Temporal Factorized Visual Transformer to generate speech. The speech decoder employs Hifi-GAN. Mel-spectrograms were only utilized for training and not for testing. The training procedure included two stages. During the first stage, the waveform generator was removed, and only the visual network and ACN got trained to synthesize Mel-spectrograms from the Auxiliary-Mel layer. In the second stage of training, the Auxiliary-Mel layer was excluded, the trained visual network and ACN had their weights frozen, and only the waveform generator was trained to synthesize speech and Mel-spectrogram by suppressing adversarial, Mel-spectrogram, and feature-matching losses. Fig. 10 presents the progression in Voice from Lips techniques tailored over time. Table 11 demonstrates the significant advancements made in Voice from Lips systems.

7. Lip HCI

There are a few examples of a silent speech recognizer being used to identify control commands for HCI devices such as computers, smart TVs, smartphones, robots, and musical instruments. Fig. 11 depicts a generic pipeline for employing a silent speech recognizer as an HCI. First, an SSR is trained on the necessary command vocabulary, and then it is used to detect silent lip commands. Further steps are taken to control HCI devices based on the identified commands. When controlling access to a smart TV (Su et al., 2021), audible speech commands are not preferred because they interfere with the TV sound. Because the music produced by instruments such as the guitar interferes with audible speech commands (Shin et al., 2015), audible speech commands are not preferred for musical instruments. Sound from the computer's speakers and other external noise sources will interfere with a user's attempts to control a computer via voice commands. Lip HCI is also beneficial for disabled users because it provides hands-free computer access for Chinese text entry (Hwang et al., 2021) and hands-free access to web applications (Pandey and Arif, 2022). For providing secure access to devices such as smartphones, speaker-dependent and multispeaker-dependent models are preferred over speaker-independent models (Su et al., 2023; Sun et al., 2018).

Shin et al. (2015) used HCI to control musical instruments such as a guitar, where guitar effector hardware was controlled by speaker-dependent lip commands, and a constrained local model was used to

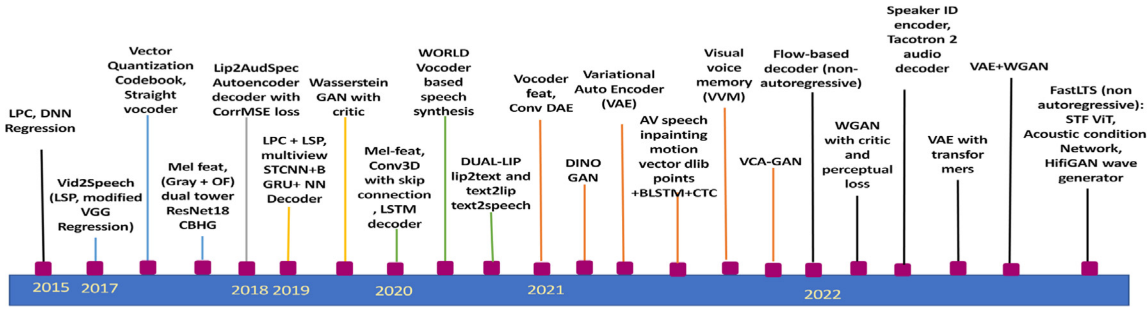
LPC, DNN Regression
Vid2Speech (LSP, modified VGG Regression)
Vector Quantization Codebook, Straight vocoder
Mel feat, (Gray + OF) dual tower ResNet18 CBHG
Lip2AudSpec Autoencoder decoder with CorrMSE loss
LPC + LSP, multiview STCNN+B GRU+ NN Decoder
Wasserstein GAN with critic
Mel-feat, Conv3D with skip connection , LSTM decoder
WORLD Vocoder based speech synthesis
DUAL-LIP lip2text and text2lip text2speech
Vocoder feat, Conv DAE
DINO GAN
Variational Auto Encoder (VAE)
AV speech inpainting motion vector dlib points +BLSTM+CTC
Visual voice memory (VVM)
VCA-GAN
Flow-based decoder (non-autoregressive)
WGAN with critic and perceptual loss
Speaker ID encoder, Tacotron 2 audio decoder
VAE with transfor mers
VAE+WGAN
FastLTS (non autoregressive): STF ViT, Acoustic conditional Network, HifiGAN wave generator
2015
2017
2018 2019
2020
2021
2022

Fig. 10. Road map of Voice from Lips.

Table 11
Lip reading in Voice from Lips using DL approaches.

Deep learning approaches								
Year	Ref.	Audio features/ Audio encoder	Video Encoder	Reconstruction (Audio Decoder)	Loss functions	Benchmark	Result	limitations
Sep 2015	Milner and Le Cornu (2015)	(1) 8th-order LPC audio features. (2) 20-channel filter bank audio features.	(1) Expectation Maximization applied to GMM to model the joint AV features (AAM) (2) DNN is used for regression. It accepts visual features (2D-DCT) as input and outputs audio features on its final layer.	STRAIGHT vocoder: is used for speech reconstruction.	Mean squared error (MSE) loss	*GRID* (constrained dataset) spk4 GMM Intelligibility DNN intelligibility	GMM outperformed DNN. 60.46% 54.24%	Only the S4 SD model was evaluated. The achieved acc. has room for improvement.
Mar 2017	Ephrat and Peleg (2017)	18 Line spectral pairs (LSP) coefficients	VGG, like CNN, is used. The length of the last CNN layer is 18, which corresponds to 18 LSP.	Source-filter speech synthesizer of Klatt and Klatt (1990)	Mean squared error (MSE) loss.	*GRID* Grid spk2 Grid spk4	79% intelligibility	SD models reconstructed speech quality was poor on constrained dataset.
Sep 2017	Le Cornu and Milner (2017)	spectral envelope surface from Mel-filterbank features	Visual features are extracted from the AAM, cubic interpolation is employed to upsample video features and match their frame rate with audio frames, and then the features are fed into LSTM. The Linear regression layer is then used for the regression method, while the softmax layer is used for the classification method.	Regression method: DNN is used to map input visual feature v_i to output audio feature a_i. Classification method: Trained vector quantization (VQ) codebook is used to map v_i to a_i by minimizing Euclidean distance between audio label c_i and feature vector a_i STRAIGHT vocoder: is used for speech reconstruction.	mean squared error (MSE) (Note: All models created are SD and designed using constrained dataset GRID.)	*GRID* (Classification) Spk 3 Spk 4 (Regression) Spk3 Spk4	PESQ 2.0551.686 PESQ 1.959 1.620	The classification model reported a word accuracy of 85%, but the generated sounds lacked naturalness.
Oct 2017	Ephrat et al. (2017)	Mel frequency coefficients	(Grayscale +optical flow (OF)) cropped lower face video frames fed to dual tower ResNet18	Visual features are fed to two fully connected layers which output Mel frequency coefficients (80 for each window). These are then fed into a framework containing a CBHG module (convolutional highway+Bi-GRU) that increases temporal resolution and converts a Mel spectrogram to a linear-scale spectrogram. It is then converted to speech by the Griffin-Lim phase reconstruction algo.	1) Reconstruction loss (Decoder output). 2)Post-processing network output loss.	*GRID* Spk1 Spk2 Spk3 Spk4 *TCD-TIMIT* LipSpk 1 LipSpk 2 LipSpk 3	PESQ 1.952 2.136 1.904 1.922 PESQ 1.364 1.158 1.612	Future research should focus on enhancing the intelligibility and naturalness of speech reconstruction for the unconstrained dataset. Developed models are SD.

(*continued on next page*)

Table 11 (*continued*).

Apr 2018	Akbari et al. (2018)	An auditory spectrogram was generated from speech using NSR tools (Chi et al., 2005). 32 Bottleneck features of 128 frequency bin auditory spectrogram were extracted from a pre-trained autoencoder.	Gray scale face images fed to a seven-layer conv3D network.	The Decoder part of the pretrained autoencoder will reconstruct the auditory spectrogram. And the speech waveform was rebuilt from the output spectrogram using NSRtools (Chi et al., 2005).	CorrMSE: a combination of correlation loss and MSE.	*GRID* Spk1 Spk2 Spk4 Spk29	PESQ 2.07 2.01 1.61 1.84	Dint leverage end-end network training The model was evaluated on a constrained dataset. Emotional prosody not included in the reconstructed speech.
July 2019	Kumar et al. (2019)	Linear predictive coding (LPC of order 24) + Line spectral pairs (LSP) encoded audio vector	0, 30, 45, 60 and 90-degree view classifier models used VGG16. During training, each cropped lip image, and its label (audio vector) were fed to 7 layers of STCNN and two layers of Bi-GRU, with the audio vector being the output of Bi-GRU. The SD-best was (0, 45, and 60°) views.	A neural network is employed to synthesize speech from an encoded audio vector.	Cross entropy loss for the view classifier MSE loss for Regression	*OuluVS2(10 Phrases)* PESQ : *SD-OOV* (the model was tested on one unseen phrase)	*SD-best model* 2.315 PESQ Range: 1.46 to 1.84	Failed for SI models. The generated sound was robotic and lacked emotion, prosody, and modulation.
Sep 2019	Vougioukas et al. (2019a)	Speech encoder network form (Vougioukas et al., 2019b) is used to extract speech features.	Visual encoder consisting of five layers of conv3D to generate video features.	An audio decoder receives video features and generates audio samples.	Wasserstein GAN loss: Wasserstein's difference between real and synthesized speech is suppressed using a critic network (discriminator). Also, during training, the perceptual loss is reduced.	*GRID* PESQ: STOI: WER: *GRID* PESQ: STOI: WER:	*(SD)* 1.71 0.518 26.6 *(SI)* 1.24 0.445 40.5	The generated speech was intelligible, but the PESQ needed to be better.
Jun 2020	Prajwal et al. (2020)	Mel spectrograms	Short video face images are fed to a Spatiotemporal face encoder consisting of stacked 3D-Conv layers with skip connections.	Attention-based speech Decoder using LSTMs was taken from Shen et al. (2018)	L1 reconstruction loss between the actual and the generated mel spectrograms.	*Own (lip2wav)* PESQ: STOI: *GRID* *(4 spk's)* PESQ: STOI: WER: *TCD-TIMIT* PESQ: STOI: WER:	(lec. videos) 1.3 0.416 *(Unseen Spk.)* 1.772 0.731 14.08 *(3 lip spk-Unseen)* 1.35 0.558 31.26	Model tested for a few speakers. The model is slow because of auto-regression.
Oct 2020	Michelsanti et al. (2020)	WORLD vocoder features 1) SP: spectral envelope 2) AP: aperiodic param. 3)VUV: voice/unvoiced 4)F0: fundamental freq.	Lower half cropped face or cropped mouth regions from video frames were fed to a video encoder consisting of five 3D convolutions followed by GRU.	Features from GRU are decoded by five decoders that predict SP, AP, UVU, F0, and Visual speech recognition (VSR) decoder.	(combination of five losses derived from five decoders) CTC loss was used to train VSR, while MSE loss was used to train SP, AP, UVU, and F0 decoders. The WORLD synthesis algorithm was used to reconstruct speech from vocoder features (Morise et al., 2016).	*GRID* PESQ: ESTOI: WER: PESQ: ESTOI: WER:	*(Seen Spk.)* 1.9 0.45515.1 *(Unseen Spk.)* 1.23 0.227 51.6	Tested on a constrained dataset Adding a VSR decoder may improve performance but adds complexity to the model.

(*continued on next page*)

Table 11 (*continued*).

Oct 2020	Chen et al. (2020)	(Audio encoder not used) Dual Model 1)Lip-to-text 2)Text-to-lip	Dual Model: M1 and M2 M1) Lip-to-text models: Employed LipNet (Assael et al., 2016) and LipResNet (Stafylakis and Tzimiropoulos, 2017) for GRID and TCDTIMIT, respectively.	M2) Text-to-Lip Model: Features extracted from the text encoder and image encoder are concatenated and sent to the RNN, and then the image decoder is utilized to obtain a reconstructed lip image. The text-to-speech converter consists of the FastSpeech model and WaveNet vocoder.	Lip-to-text training utilized CTC loss Text to Lip was trained with L1 loss between the original and reconstructed images.	*GRID* WER: *TCD-TIMIT* PER:	3.10 46.2	Converting lip motions to text and vice versa increases the model's complexity.
Jan 2021	Zhang et al. (2021a)	line spectral frequencies (LSF), Fundamental frequency (FO), and Unvoiced/Voiced (U/V) flag	1) Resized ROIs of lip and tongue are fed to a Conv-DAE whose bottleneck values are used to predict 12 LSF values (12, 100-class classifier) and U/V (binary classifier). 2) Three regression output layers predict F0, Amplitude, and DC offset of the speech.	Based on the predicted features, the ConvNet-based Vocoder generates a song. *The model was evaluated on its own data set (five-minute-long Latin and Corsican songs).*	1) CNN Vocoders, authors have used distortion loss (MSE). 2) DAE is trained using the sum of MSE loss of generated lip and tongue images. 3)LSF and U/V classification using cross-entropy loss 4)F0 prediction with MSE loss.	LSF distortion F0 MSE: U/V accuracy: SCS:	3.5 dB 0.008 93% 0.209	The model's performance can be improved further.
May 2021	Vougioukas et al. (2021)	Five layers of conv1D followed by two GRUs	1) Video encoder: Five layers of conv3D The identity encoder and frame decoder take U-net-like architecture. 2) Identity encoder: Five layers of conv2D 3) Frame decoder: 2D transposed convolution layers.	Audio decoder: Five layers of 1D transposed convolution with self-attention GAN Network: consists of Forward and Reverse networks 1) The Forward network accepts lower-face video frames as input and translates them into speech. It includes a video encoder and an audio decoder.	2) Reverse network architecture resembles U-Net. It cascades features from the identity network and audio encoder and forwards them to the Frame decoder to generate lip motions corresponding to speech. *Forward and reverse networks* form GAN, which is trained using adversarial loss.	*GRID* PESQ: STOI: WER:	*unseen subjects* 1.21 0.51 32.6	Voice generated is robotic.
Jun 2021	Yadav et al. (2021)	Mel spectrograms fed to 3-layered fully connected network +LSTM.	VGG16 architecture + LSTM	Mirrored version of the audio encoder	KL divergence loss of variational autoencoder that minimizes the distance between audio feature distribution and video feature distribution.	*GRID* *(4 spk.)* PESQ: STOI:	*(SD)* 1.932 0.724	There is no identity encoder, so all samples share the same robotic voice.
Jun 2021	Morrone et al. (2021)	Log of STFT magnitude Local Weighted Sum (LWS) technique retrieved the missing phase.	By subtracting the current frame landmarks from the previous frame landmarks, the motion vectors of 68 x-y DLIB landmarks were calculated. In order to match the audio rate, video features were upsampled from 25 fps to 83.33 fps.	Bi-LSTMs are trained with CTC loss in order to reconstruct the inpainted spectrogram, which is then added to the masked spectrogram to restore the final spectrogram.	CTC loss	*GRID* PER:	*(SI)* 0.137	GRID contains only 33 phonemes, while TIMIT contains 61. Thus, the model was evaluated using fewer phonemes.

(*continued on next page*)

Table 11 (*continued*).

Nov 2021	Hong et al. (2021)	Mel spectrograms fed to three 2D convolutional layers+ residual block	Video encoder: 3D conv layers with skip connections + LSTM Cosine similarity is used to train VV memory, and after training, it contains the visual descriptor and matching speech descriptor in key and value memory slots, respectively.	Concatenated audio and visual descriptors are passed through the PreNet (2 layers of LSTM and linear projection layer), followed by PostNet, consisting of five convolutional layers. Postnet predicts residual and uses it to improve the mel-spectrogram. Speech is reconstructed using Griffin-Lim or Wavenet Vocoder.	The L2 loss (real and pseudo audio feature extracted from VV memory) KL divergence loss between value and key address vectors L2 loss between predicted and real mel spectrograms measured before Postnet and after Postnet	*GRID* PESQ: STOI: WER: *GRID* PESQ: STOI: WER:	*(SD)* 1.984 0.738 13.83% *(SI)* 1.332 0.6 37.96%	Even though the model improves intelligibility, obtained speech quality can be further enhanced.
Dec 2021	Kim et al. (2021)	Mel spectrogram Audio encoder: two Conv2D + 2D Residual block	Visual context attention (VCA) GAN with three 2D GANs where visual encoder consists of 3D convolution layer + ResNet-18	Postnet: 1D Residual blocks and two Conv1D layers (maps Mel spectrogram to linear spectrogram) Discriminator: Residual blocks. Griffin-Lim is used to synthesize speech.	Combination of three losses: 1) Adversarial loss for GAN training. 2) L1 loss between real and generated Mel spectrogram. 3) audio–video synchronization loss	*GRID* PESQ: STOI: WER: *TCD-TIMIT* PESQ: STOI: *GRID* PESQ: STOI: WER: *LRW* PESQ: STOI: WER:	*(SD)* 2.008 0.724 12.25 *(3 lip spk-SD)* 1.425 0.584 *(SI)* 1.372 0.569 23.45 *(SI)* 1.337 0.565 29.95	The performance of the unseen speaker's model is not good.
Feb 2022	He et al. (2022)	Mel spectrogram	Visual encoder: Four layers of conv3D with skip connections + BiLSTM	Condition module: generates a coarse Mel-spectrogram. Flow-Based Decoder: Conditioned on Coarse spectrogram, which generates a realistic Mel-spectrogram. It comprises the squeeze, actnorm, invertible 1 × 1 conv, affine coupling, and unsqueeze layers.	The visual encoder and condition module are trained using the MSE loss between the real and coarse Mel-spectrogram. The flow-based decoder is trained using the negative log-likelihood of a spherical Gaussian and Jacobian log-determinant of actnorm, affine coupling, and invertible 1 × 1 conv layers.	*lip2wav* dataset. *Chemistry Le.* STOI: ESTOI: *Chess Analysis.* STOI: ESTOI:	By Prajwal et al. (2020) 0.47 0.328 0.394 0.259	The model size is huge. Not efficient in Memory usage. Advanced Vocoder can be employed for more realistic speech
Apr 2022	Mira et al. (2022)	(Mel Frequency Cepstral Coefficients)	Cropped mouth video fed to 3D Conv+ResNet18	Six stacked transposed convolutional layers	(Wasserstein GAN loss) Based on waveform judgment and power judgment along with three other L1 losses, perceptual loss (Lpase-between perceptual features of real and generated speech, pretrained PASE model was used to extract perceptual feature), power loss (Lpower- between STFT magnitudes of real and generated speech) and MFCC loss (Lmfcc- between 25 Mel-frequency cepstral coefficients of real and generated waveform)	*GRID* PESQ: STOI: WER: *GRID* PESQ: STOI: WER: *TCD-TIMIT* PESQ: STOI: *LRW* PESQ: STOI: WER:	*Unseen spk.* 1.47 0.523 23.13 *Seen spk.* 2.02 0.601 8.03 *(3 lip spk-SD)* 1.61 0.295 *(SI)* 1.45 0.556 42.51	Loss improves generated speech's realism and clarity, but the model fails to model high-frequency contents.
June 2022	Zeng and Xiong (2022)	Pretrained speaker ID encoder network (Wan et al., 2018)	Video encoder: a stack of 3DConv + Bi-LSTM	Tocotron2 decoder (Shen et al., 2018)	L1 loss between real and generated Mel-spectrogram.	*GRID* (4 spk.) PESQ: STOI: *lip2wav* dataset. PESQ: STOI:	*(SD)* 1.781 0.756 *(Chemistry Le.)* 1.313 0.441	Poor performance in multi-speaker environments and unconstrained Lip2Wav datasets.

(*continued on next page*)

Table 11 (*continued*).

Aug 2022	Varshney et al. (2022)	MFCC features are input to audio encoders that are processed by an audio transformer encoder.	Features extracted from face encoders are added with positional embedding and processed by a visual transformer encoder.	VAE module built with audio transformer encoder, visual transformer encoder, and speech transformer decoder.	VAE trained with reconstruction loss, KL-loss, and Deep metric learning loss.	*GRID* PESQ: STOI: *Lip2Wav* chem. Lec.) PESQ: STOI:	*(SD)* 2.022 0.711 *(SD)* 1.233 0.49	The training of the Transformer decoder is auto-regressive. This introduces output latency.
Oct 2022	Hegde et al. (2022)	Mel-spectrum processed by a pre-trained ASR	Cropped face video with five frame window and stride one processed by 3DConv Network.	VAE maps audio and video features to the same space of latent features. Audio Decoder: Wasserstein GAN	VAE trained with local and global KL loss WGAN trained with adversarial loss	*GRID* PESQ: STOI: *TCD-TIMIT* PESQ: STOI: *LRW* PESQ: STOI: *LRS2* PESQ: STOI:	1.28 0.45 1.35 0.55 0.78 0.15 0.6 0.34	The pre-trained framework for multiple speakers can be fine-tuned for a single speaker. Still, there is space to improve PESQ and STOI.
Oct 2022	Wang and Zhao (2022)	Mel-spectrum (utilized only for training)	Cropped face video frames processed by a spatiotemporal factored visual transformer	Acoustic Conditional Network + Waveform Generator (Hifi-GAN) + Mel-spectrogram Generator (Auxiliary Mel Layer)	They are trained using adversarial, Mel-spec, and feature-matching losses.	*GRID* PESQ: *Lip2Wav*	1.939 MOS: Approximately: 3.8%	The training procedure involves two stages. The model experimented only with single-speaker settings.

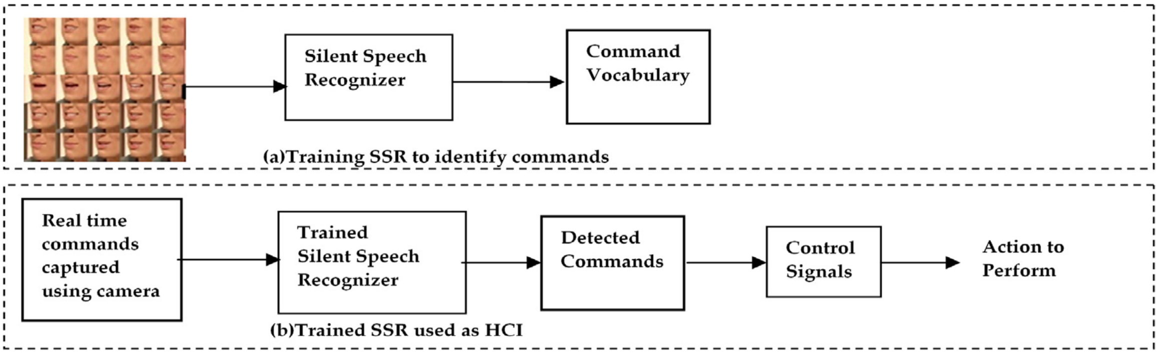
Silent Speech Recognizer
Command Vocabulary
(a)Training SSR to identify commands
Real time commands captured using camera
Trained Silent Speech Recognizer
Detected Commands
Control Signals
Action to Perform
(b)Trained SSR used as HCI

Fig. 11. General pipeline of Lip HCI.

track facial features. While playing a musical instrument, the user's head pose would change; therefore, pose normalization was performed by normalizing the facial landmarks to the frontal view by eliminating rigid global transformations. Dynamic time warping (DTW) was used for word recognition. DTW determines the optimal time warping to match two temporal sequences. This helps to align words spoken at varying speeds. Lip HCI was employed on the smartphone. The touchscreen mobile display size will be limited. Multiple operations are required to reach the desired item on the menu. Using gestures can be helpful for small command sets, but they are not feasible for large command sets. Although voice assistants like Siri can open apps, people are concerned about privacy issues because they do not want their actions to be noticed by others. Lip reading using various sensors, such as EMG, ultrasound imaging, and EEG, involves an invasive setup. Therefore, authors Sun et al. (2018) present lip-Interact, which opens the user-facing camera to capture silent, exaggerated lip movements and recognize the issued 44 commands with a spatial transformer, 3D spatio-temporal CNN to extract features that are input to Bi-GRU in order to predict the lip command class, here spatial transformer uses bilinear sampling and 4 × 4 thin plate spline transformation (TPS) with a regular control point grid as the transformation function in order to correct the mouth view towards the camera. Their implementation uses an Android accessibility service, which runs in the background and retrieves window content in real-time. The computations were performed on a GPU host. They have also tailored the model's customization to specific end-users.

Researchers developed a Russian visual speech recognition system for human–robot interfaces (Ivanko et al., 2019). They investigated PCA and AAM feature extraction techniques. The authors utilized the Gaussian mixture HMM for classification because more training data is required to train neural networks. They created ten speaker-specific VSR systems for ten speakers in the HAVRUS dataset. The authors Su et al. (2021) experimented with gaze and lips to develop a hands-free interface for controlling a smart TV using YouTube's command set. To determine whether or not the user is issuing a command, the Dlib face detector library was utilized to identify mouth landmarks and calculate the degree of mouth opening. The spatiotemporal network was used to classify words, and the softmax layer makes predictions by considering only candidates falling within the range of gaze commands and ignoring other commands. The authors suggest choosing command words with various visemes in order to avoid confusion between commands on the same menu. Hwang et al. (2021) provided disabled users with hands-free computer access by integrating eye-tracking technology with lip motion recognition HCI interface. The user's eye gaze directs the cursor to a specific command item on the screen, after which the user's spoken command is decoded, and the corresponding action is executed. On the CANDIDE-3 face model, the Kinect face tracking SDK detects the

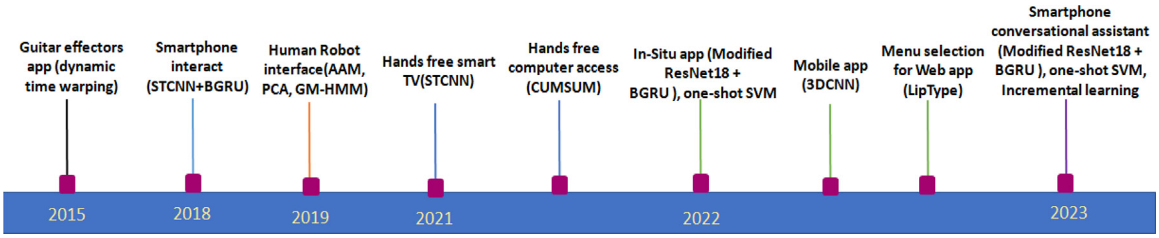
Guitar effectors app (dynamic time warping)
Smartphone interact (STCNN+BGRU)
Human Robot interface(AAM, PCA, GM-HMM)
Hands free smart TV(STCNN)
Hands free computer access (CUMSUM)
In-Situ app (Modified ResNet18 + BGRU), one-shot SVM
Mobile app (3DCNN)
Menu selection for Web app (LipType)
Smartphone conversational assistant (Modified ResNet18 + BGRU), one-shot SVM, Incremental learning
2015
2018
2019
2021
2022
2023

Fig. 12. Road map of Lip HCI.

face and extracts ten lip points. The CUMSUM control chart algorithm identifies lip commands using seven normalized Euclidean distances between these points.

In recent works, Su et al. (2023, 2022) used contrastive learning with InfoNCE loss using cosine similarity to pre-train the SSR model using the LRW500 corpus to fetch the optimal feature vector for one-shot learning. In addition, the real-time model was fine-tuned for unseen commands using a single training example per command. During the new command registration phase, speech recognition was used to annotate the voice input with text. The registered user's command features were utilized to train a one-shot SVM (classifier), which was then used to make the final decision. They created custom datasets in six recording environments to ensure that their lip-HCI approach is robust to environmental and interaction variations, including illumination conditions, pose variations, and various mobile holding angles. They leveraged keyword spotting to identify the beginning and end of the command. They employed incremental and few-shot learning by evaluating the accuracy of their incremental model with 30 commands for one shot, three shots, and five shots. In another work, Nemani et al. (2022) carried out speaker-independent SSR using 3DCNN. In this context, researchers created LIPAR, a custom word dataset similar to MIRACL-VC1 that consists of high-resolution videos of ten words spoken ten times by thirty-five orators. Since not all videotapes had the same number of frames, they concatenated the video with fewer frames with blank or silent frames to make the total number of frames equal to 70. The authors deployed their trained module on an Azure cloud A2V2 virtual machine to support mobile edge applications, and then the entire framework was tested in real-time.

Eye-gaze is less accurate in the screen's top and bottom corners than in the center of the screen, and the short time dwell selection procedure leads to more false positives. In contrast, long-time dwell selection leads to eye fatigue. Therefore, Pandey and Arif (2022) adopted a contactless HCI, Eye-gaze with SSR, for menu selection in the custom web application. They leveraged the LipType (Pandey and Arif, 2021) architecture to detect and select menu commands. Even though the literature contains numerous works on SSR, only a few studies on Lip HCI exist. The above demonstrates the challenge and difficulty of implementing SSR in real-time. Even though the work performed on state-of-the-art benchmarks yields satisfactory results, it fails when tested on real-time systems. Fig. 12 presents the progression in Lip HCI techniques employed over the years. Table 12 compiles the advancements in lip HCI.

8. Discussion and recommendations

This section includes a summary of the analysis of existing works, a list of the limitations for developing each LR application, and a recommendation for the best existing model.

8.1. Discussion and suggestion on lip-reading benchmark datasets

In order to evaluate the performance of well-designed frameworks, datasets are crucial. There are numerous benchmark datasets for recognizing the alphabet, digits, words, and sentences. Initially, limited vocabulary digits and alphabet datasets were constructed with a frontal view. They were recorded in a studio environment with a uniform background and a fixed distance between the user and the low-resolution camera. Also, they should have taken into account changes in facial features such as facial hair, sunglasses, or clothing (Matthews et al., 2002; Movellan, 1994). Then, gradually, datasets were recorded with high-resolution cameras and began to include head pose variations, either by the speaker adopting diverse head positions (Fox et al., 2005; Messer et al., 1999; Millar and Goecke, 2004) or by employing multiple cameras (Anina et al., 2015; Lee et al., 2004). For the AVSR dataset, noisy conditions were also recorded to determine the impact of noise on speech recognition (Czyzewski et al., 2017; Lee et al., 2004; Millar and Goecke, 2004; Sanderson, 2002; Tao and Busso, 2018). Also, researchers began to record datasets under various lighting conditions (Millar and Goecke, 2004; Rekik et al., 2014). The datasets (Abdrakhmanova et al., 2021; Fox et al., 2005) were recorded in multiple sessions separated by a month or a week to account for facial features, background, and attire variations. Some datasets include data captured by specialized cameras such as the depth camera (Rekik et al., 2014), stereo camera (Czyzewski et al., 2017; Vorwerk et al., 2010), thermal camera (Abdrakhmanova et al., 2021), and event camera (Tan et al., 2022). Researchers began constructing the unconstrained datasets from existing videos such as YouTube videos (Makino et al., 2019; Nagrani et al., 2020, 2017; Shillingford et al., 2019), BBC News (Afouras et al., 2018a; Chung and Zisserman, 2018), Ted talks (Afouras et al., 2018b), and lecture videos. These unconstrained datasets can be used to build speaker-independent models. In contrast, the only unconstrained dataset currently available for building speaker-dependent models is the Lip2Wav (Prajwal et al., 2020) dataset, which consists of chess analysis and lecture videos on chemistry, deep learning, hardware security, and ethical computing. Unconstrained datasets with an extensive vocabulary available in the literature prevent overfitting in the models and aid in developing more effective speaker-independent models. These SI-LR models can be used as pretrained models and further customized to an individual speaker, enabling the creation of speaker-dependent models with few training examples (Su et al., 2023). To develop biometric speaker-dependent models, it is necessary to record samples in unrestricted environments for every individual. In addition to the datasets listed in Table 3, there are several AV datasets available in different languages, like LRW-1000 (Yang et al., 2019) (Mandarin language), HAVRUS (Verkhodanova et al., 2016) (Russian language), Visual LR Feasibility (Fernandez-Lopez et al., 2017) (Spanish language), and the 14-language dataset (multilanguage LR dataset) LRS3-lang (Afouras et al., 2020b), which is out of the scope of our survey. Constrained datasets are unsuitable for real-time applications, so unconstrained datasets are preferred over constrained ones.

8.2. Discussion and suggestion on lip-reading biometrics

Motion estimation, cascaded lip coordinates, and mouth area served as classification features in the beginning era. The classification was accomplished without the use of laborious training procedures. Simple distance metrics and thresholding were used to identify feature matches. These techniques could not be more resistant to varying illumination conditions, camera-to-speaker distance, pose variations,

Table 12
Lip HCI summary.

Lip HCI approaches								
Year	Ref.	Video features	Classifier	HCI Application	HCI Commands	Benchmark	Result	Limitations
Sep 2015	Shin et al. (2015)	Chehra, Constrained Local Model (CLM)-based face feature tracker with pose normalization.	Dynamic Time Warping	Guitar effectors application	Distortion, mode change, music score, next page, previous page, home, let it be, jamming.	OuluVS (SD) AVLetters2 (SD)	WRR (word recognition rate) 85.0% 75.4%	Research should be conducted on speaker-independent (SI) command recognition.
Oct 2018	Sun et al. (2018, p.)	Cropped mouth lip command frames	Spatial Transformer (ST) + 3DCNN +Bi-GRU	Smart phone interaction (Huawei P9 Plus)	Forty-four frequently used commands like back, home, copy, cut, paste, etc.	(Own) Classification accuracy for a set of 20 commands	Multi-SD models: Single user: 95.464% Many users: 98.9%	Lip interaction must be further personalized for each user. Their implementation computes using a host server. Future work may integrate lip-interact into a standalone smartphone.
Aug 2019	Ivanko et al. (2019)	PCA and AAM	Gaussian mixture -hidden Markov models	human–robot interface	Isolated commands for human–robot interface	HAVRUS (Own) Russian 1) Geometry-based features 2) Pixel-based features	WCR SD models (41.54% to 39.80%) (37.27% to 32.09%).	Word correctly recognized (WCR) is low. The proposed lip-reading method can be extended to SI-SSR cases.
May 2021	Su et al. (2021)	To determine if a user is uttering a command, Dlib face landmarks are used to determine the degree of mouth opening.	Cropped mouth images fed to Spatiotemporal network	Hands-free Smart TV access	27 frequently-used Youtube commands: watch later, Play, Stop, Go back, Go forward, Scroll up, Scroll down, etc.	Own SD model	Gaze + SSR mean classification accuracy 98.42%	Commands with distinct visemes improve precision. Whether the user was speaking or laughing, commands were recognized when they opened their mouth.
Jun 2021	Hwang et al. (2021)	On the CANDIDE-3 face model, the Kinect face tracking SDK extracts ten lip points. The CUMSUM control chart algorithm uses these points.	CUMSUM uses seven normalized Euclidean distances between these points to identify lip commands.	Hands-free computer access: Chinese text entry	Eye tracking was used to move the pointer on the screen, while lip commands were used to select buttons.	Own Button selection error rate Button selection error rate in words per minute	2.167% 4.41 wpm	Increasing text entry speed with rapid word prediction.
Oct 2022	Su et al. (2022)	Embeddings (500 -dimensional vector) are extracted from grayscale lower half face ROI using Contrastive Learning	ResNet18 with modified 3D conv (size 5 × 7x7), followed by three layers of BGRU. Personalized SVM one-shot feature classifier	in-situ SSR	The command comprised ten phrases from the OuluVS. Five repetitions of ten sentences by three subjects	LRW500 for contrastive learning Testing in real-time with their custom data set	Average Acc. 91.33%	The proposed lip-reading method can be extended to SI-SSR cases.
Nov 2022	Nemani et al. (2022)	Grayscale frames of mouth ROI	3DCNN	Mobile application	Ten speaker-independent commands (LIPAR dataset)	LIPAR MIRACL-VC1	SI Acc: 70.2% SI Acc: 77.9%	As the computation is performed in the cloud, Uptime, downtime, and inference time cause delays in cloud computing.

(continued on next page)

Table 12 (*continued*).

Nov 2022	Pandey and Arif (2022)	RGB -cropped mouth video frames	LipType architecture (Pandey and Arif, 2021)	Menu selection in custom web application	Training: (10 commands × 20 subjects x10 instances) Testing: (10 commands × 12 subjects × 10 instances)	Custom dataset	Error Rate: 1.94% (Eye-gaze +SSR)	Menu selection time:5 s
April 2023	Su et al. (2023)	Embeddings (500 -dimensional vector) are extracted from grayscale lower half face ROI using Contrastive Learning	ResNet18 with modified 3D conv (size 5 × 7 × 7) followed by three layers of BGRU. Personalized SVM one-shot feature classifier	Smartphone conversational assistant	Training commands: 25 phonetically balanced sentence commands selected from Alexa and Siri commands. Recorded with variations in lighting, posture, and mobile grasp gestures. Training dataset: 11 subjects (7M+ 4F) × 6 recording conditions × 25 commands × 5 instances for each command	LRW500 for contrastive learning Real-time testing with a custom dataset (with incremental learning)	For testing, 30 commands and five instances were used. *Accur.* 1-shot: 81.7% 3-shot: 96.5% 5-shot: 98.8%	While recording the dataset, they should have accounted for variations in facial hair and other facial features. Keyword spotting determines whether the user is uttering a command or not but does not distinguish between command utterance and everyday speech.

and other LR challenges. In addition, the LR systems' accuracy could have been better, and determining the optimal threshold took a lot of work. With the advancements in pattern recognition and machine learning, researchers began exploring the most effective feature extraction, feature selection, dimensionality reduction, and classification techniques to improve the accuracy of LR systems. However, the rise of deep learning technology has paved the way for a new method of reading lips that surpasses human, conventional, and ML performances. Early research focused on word-level LR models and then gradually progressed to sentence-level LR models. LR was combined with face recognition technology for biometrics.

In summary, there are several limitations to lip-reading authentication systems. Currently, systems are designed for a closed set of users (Hassanat, 2014; Lee and Myung, 2017; Raghavendra et al., 2021; Sengupta et al., 2012), and adding a new user, deleting an old user, or dealing with an unknown user is a pressing concern. Some systems extract handcrafted features (Alghathbar and Mahmoud, 2009; Hassanat, 2014; Shubhangi and Balbhim, 2017) and use a similarity index to determine the best match. Some systems are intended for an open set of users. They implement one-shot learning using a Siamese network (Ruengprateepsang et al., 2020; Wright and Stewart, 2020; Zakeri and Hassanpour, 2021). Few open-set module systems employ one-class or binary classifiers for each user enrolling in the authentication system, increasing the model's complexity. Examples include one-class HMM models (Ezz et al., 2020) and SVM frameworks (Kapkar and Bharkad, 2021; Lai et al., 2016) trained on handcrafted embeddings. The "benchmark" and "task type" columns of Table 4 reveal that existing systems are designed to recognize digits, alphabets, alphanumeric passwords, sequences of four-digit words, limited words in Miracl, and sentences in datasets such as AV-digit phrases, Grid, and Lilir. Currently, they are being validated for a smaller number of speakers. Most research papers and results are based on closed benchmark datasets with no real-time implementations. A one-shot or few-shot learning requirement is necessary for authentication purposes. If this is not the case, clients become impatient when asked to record additional training samples (Yang et al., 2020). Even though lip-reading authentication systems take advantage of the fact that speaker-independent accuracy is poor compared to speaker-dependent models, they have to deal with their own challenges and limitations like one-shot or few-shot learning, designing open-set models, and dealing with an unknown user.

According to the survey, the model developed by Yang et al. (2020) SA-DTH-Net with FFE-Net and RC-Net is the best LR-biometric framework, as it has been proven to be robust against imposters and deep fake attacks, and it represents the state-of-the-art with HTER 0.6 on the Grid dataset and HTER 2.4 on the Mobio dataset. The second-best model is proposed by Ruengprateepsang et al. (2020), which uses a Siamese network with ST-Conv + Bi-LSTM as its twin network for one-shot learning, and it represents the state-of-the-art with an EER of 9% on AVDigits.

8.3. Discussion and suggestion on audio visual speech recognition

The performance of each module, audio feature extraction module, video feature extraction module, and audio video fusion module, contribute to the efficiency of an AVSR system. It is evident from the information presented in Table 7 that DL-AVSR frameworks perform better on constrained datasets. Consequently, researchers are currently focusing more on unconstrained datasets. All AVSR research focuses solely on improving recognition accuracy, ignoring real-time applications. Thus, developed models have more network parameters. The performance of AVSR systems degrades in the presence of noise, and they all struggle to classify confusing words. AVSR systems require pre-training with an enormous lip-reading corpus. Whether it is a word recognizer or sentence recognizer, their recognition is constrained to a fixed number of possible classes. In reality, however, the system must be capable of predicting any English word from the dictionary. Even though there are corpora with depth and multiview images, they still need to be explored in the research work. Many breakthrough techniques have emerged for AVSR systems.

In summary, researchers have investigated various audio features such as Mel, Log-Mel, STFT, log spectral features, MFCC, GFCC, LDA, features extracted by auto-encoders, 1DResNet18+Bi-GRU, and conformer encoder in addition to using raw audio. AVSR researchers have experimented with numerous backbone architectures for video encoders, including GoogleNet, VGGFace2, CNN layers (own network), front-ends (ResNet or ST-ResNet), with back-ends (Bi-GRU, transformer CTC, or transformer seq2seq) networks, ST-DNN (Depth+RGB), ST-GCN with Bi-GRU, CoGANs, C3D-P3D network, conformer networks, and ViT. Multiple AV-fusion strategies employing DBF, multi-stream HMM, SVMs, decision level fusion utilizing CNN layers or MLP, Bi-LSTMs, SyncNet trained with contrastive loss, Multiscale-TCN, Bi-VAE, ISA network, HUBERT, MMST, Modality Invariant Codebook, Transformer CTC, and Transformer seq2seq are investigated. According to the survey, the model developed by Liu et al. (2021a) is the most accurate word-level AVSR framework; despite using raw audio instead of speech features, which increases computation complexity, it achieves 98.91 percent accuracy on the LRW-500 benchmark. The second-best word model (Yang et al., 2022) achieves 98.5 percent accuracy on the LRW-500 benchmark. They used a modality-invariant codebook to generate AVSR features. The transformer seq2seq and transformer CTC architectures put forth by Afouras et al. (2022) are the best sentence-level AVSR frameworks. The transformer CTC model achieves a WER of 3.7% on the LRS2 benchmark, and the transformer seq2seq model achieves a WER of 8% on the LRS3 benchmark. AVSR applications favor speaker-independent lip-reading models as opposed to authentication, where speaker-dependent models are the go-to approach. The vocabulary of the English language encompasses approximately 250,000 unique words, and all datasets developed are constrained to a small number of vocabulary terms and limited subjects. Voice assistants, video text transcription systems, and hearing aids for deaf people can perform better when speech recognition and lip-reading models are integrated. In the absence of noise, the audio recognition module should be prioritized, while in the presence of noise, the video module (SSR) should be prioritized because its performance is unaffected by noise. Consequently, adaptive weighting is essential for optimizing audio and video integration.

8.4. Discussion and suggestion on silent speech recognition

It is evident from the literature review that SSR researchers have achieved satisfactory performance on constrained datasets with relative ease; however, they currently need help to achieve good performance on unconstrained datasets. According to the information presented in Tables 8–10, traditional approaches classified video features such as optical flow, ASM, DCT, CSAM + MLLT + fMLLR + LDA, HoG, MBH, Eigenlips, AAM, MCT, and LBP using cosine similarities or ML classifiers such as HMM, KNN, subspace-KNN, and SVM. However, their results were subpar, so the trend shifted towards DL strategies. In DL approaches, researchers utilized numerous front-end and back-end backbone frameworks. For the front-end 2D-CNN, 3D-CNN, truncated 3D-CNN networks, AlexNet, InceptionV3, VGG16, VGG19, 3D-DenseNet, 3Dconv + ShuffleNet, VGGNet, LipNet (ST-CNN), lip motion magnification encoder–decoder with STCNN, STCNN + VTP, Autoencoder, 3D-STCNN with ResNet34 or ResNet18, Reduced-ResNet, squeeze and excitation ResNet, ALSOS modules in between ResNet, two stream networks like Inflated 3D ConvNets, ASST-GCN network, MSTP with message flow module, 3DCVT, ASGM network, conformer network, DF network, audio-driven DF network, Teacher ASR - Student VSR models using cross-modal distillation, Teacher VSR - Student VSR models using self distillation, trained using KD-loss have been explored. In the beginning, for back-end networks, GRU, Bi-GRU, LSTM, Bi-LSTM, and FST language models were employed. Later, they came up with new back-end architectures to improve the SRR accuracy, like Bi-Conv LSTM, TCN, multiscale TCN, depth-wise separable TCN, densely connected TCN, and back-end integration with pre-trained language models like transformer CTC, transformer seq2seq with pretrained BERT, cross-modal decoder with LM model, and GPT word detector model. Recent research employs only audio for SSR training, necessitating the development of back-ends such as VAM and multi-head VA memory (MVM).

Analyzing the top SSR works at present, Since annotations are expensive and time-consuming, self-supervised learning techniques that use speech features as natural annotations for lip reading are a growing trend in these systems. The frameworks for knowledge distillation are based on the principle that audio and video modalities can share latent space. Thus, SSR can function even without an audio modality. Due to the gap between audio and video modalities, it is difficult to map them onto a shared latent space. Therefore, using VAM can bridge this gap between the two modalities. After training is complete, VAM is able to store in memory the audio embeddings that correspond to the short video clip embeddings. During testing, the VAM searches the codebook and retrieves the corresponding audio embedding for a given visual embedding. Even though researchers exert great effort to achieve sentence-level lip reading and improve word recognition rates, the word error rates achieved in the works could be even better. There are still many opportunities and room for innovation to improve the accuracy of the existing lip-reading system. Even though good results were obtained with speaker-dependent models, working with speaker-independent models remains challenging. Existing lip reading systems have the limitations of demanding massive training data, requiring a great deal of training time, and being too cumbersome for real-time applications. Cloud deployment is necessary for lip-reading applications. The SSR literature review reveals that the model developed by Ma et al. (2022b) employing techniques like pre-training with large datasets, data augmentation, including word boundary, a densely connected TCN backend instead of Bi-GRU, and self-distillation with KL loss is the best SSR-word model, surpassing the state-of-the-art accuracy on the LRW-500 benchmark with an accuracy of 94.1%. The SSR framework with visual transformer pooling (VTP) and attention-weighted average pooling proposed by Prajwal et al. (2022) is the best SSR-sentence model, achieving the state-of-the-art WER on LRS2 with a WER of 22.6% and on LRS3 with a WER of 30.7%.

8.5. Discussion and suggestion on voice from lips

Converting silent video to text requires extensive manual text labeling, language models, and word boundary detection. On the other hand, training in silent video-to-speech conversion can be conducted under natural supervision without the need for text labeling. In this case, the feature vector of the audio signal will serve as a natural label for the silent video segment. Furthermore, it is impossible to reconstruct speech from unseen sentences if lip reading is considered a classification task. It is a good idea to convert silent video to speech and then speech to text to identify prediction errors in a sentence-level continuous LR system. According to the information presented in Table 11, numerous researchers have exerted significant effort to translate lip movements into their corresponding sounds. Nevertheless, their work must synthesize natural voices and achieve high SI model precision. These matters may be the focus of future investigations. In addition, most of the proposed works depend on perfect audio and video synchronization. Using voice from lips systems in surveillance systems will raise privacy concerns (Yadav et al., 2021).

Several audio features are evaluated for their ability to represent speech information, including log STFT, LPC, LSP, SP + AP + VUV + F0, Mel features, speech features extracted using 1DCNN-LSTM networks, autoencoder extracting Mel features, 2DConv+ResNet18 extracting features from the 2D-Mel spectrogram, and pretrained ASR extracting Mel spectrum features, and then MFCCs, and MFFCs + audio transformer encoder. Typically, the video frame rate does not match the audio frame rate; therefore, the video frame rate must be upsampled to match the audio frame rate. Then, to map these silent lip video features to their corresponding speech or audio features, Researchers have proposed several architectures. These regression architectures include GMM, DNN, and VGG like CNN, AAM + LSTM, trained vector quantization code book to map video features to corresponding audio features, variational autoencoder, (grayscale-OF) dual tower ResNet18, 3DConv, 3DConv + Bi-LSTM + Tacotron2 decoder, 3DConv with skip connections + Bi-LSTM+ flow-based decoder (GlowLTS), visual voice memory (VVM), Wasserstein GAN with (3DConv + ResNet18), VGG16 view classifier + STCNN +Bi-GRU, motion vectors of 68 Dlib key points +Bi-LSTM, VCA-GAN, VAE (with visual transformer), VAE+ Wassertein GAN, and ST factored visual transformer network (fastLTS). If Autoencoders are employed, the audio decoder is a mirrored version of the audio encoder that consists of 1D transposed conv layers. Several additional modules are used to generate speech from audio encodings, including the STRAIGHT vocoder, source-filter speech synthesizer, WORLD vocoder synthesis algorithm, Wasserstein GAN audio decoder, Neural network speech synthesizer, HiFi-GAN waveform generator, ConvNet-based vocoder, attention based speech decoder network, and Griffin-Lim algorithm. Loss functions such as MSE, CorrMSE, Wasserstein GAN loss with perceptual loss, power loss, and MFCC loss, KL-divergence loss for VAE to minimize the distance between audio and video feature distribution, audio–video synchronization loss, negative log-likelihood of a spherical Gaussian, and Jacobian loss to train flow-based decoders are used to train these regression networks to improve the quality of speech generated.

If a short video clip is considered, it will be challenging to distinguish homophones. Therefore, predicting the Mel-spectrogram of the short video clip by sliding window of fixed length may not help distinguish homophones; thus, using VCA-GAN improved results. Autoregressive models, where generated speech looks into current frames and previous ones, introduce latency. Therefore non-autoregressive models like flow-based speech generation are preferred. Trained visual voice memory where predicted speech features are calculated as a weighted sum over all memory slots, and Wasserstein GAN also contributes to improving voice from lips systems. The Voice from Lips literature review reveals three competing SD models, which are evaluated using the constrained dataset GRID. The VAE module proposed by Varshney et al. (2022) is constructed with an audio transformer encoder, visual transformer encoder, and speech transformer decoder. The second model is proposed by Zeng and Xiong (2022) and employs a video encoder with a stack of 3Dconv+Bi-LSTM and a pretrained Tacotron2 decoder. The third one is the VVMemory model proposed by Hong et al. (2021). They all achieve PESQ values of 2.022, 1.781, and 1.984 and STOI values of 0.711, 0.756, and 0.738, respectively, on the GRID benchmark (Kim et al., 2021). VCA-GAN achieves state-of-the-art performance on the LRW dataset for the SI model, with a PESQ of 1.337, STOI of 0.565, and WER of 29.95.

8.6. Discussion and suggestion on lip HCI

In summary, there are very few works in the literature related to lip-HCI, as shown in Table 12, where they have been used to control the Guitar effector application, 44 frequently used commands such as back, home, copy, cut, and paste for smartphone interaction, isolated command recognition for human–robot interaction, lip-HCI for in-situ applications, and smartphone conversational assistants. Few works combine eye gaze and SRR, such as sending 27 frequently used YouTube commands, such as play, pause, etc., to control a smart TV, entering Chinese text on a computer, and selecting menu options in a custom web application. Even though there are available, high-performing backbone architectures designed for SSR, they are not used for HCI. Table 12 shows that Researchers have employed simple SSR classifiers such as DTW, GMM-HMM, the CUMSUM algorithm, 3DCNN, spatial transformer + 3DCNN + Bi-GRU, and 3DCNN with ResNet18 or ResNet34. Thus their performances with out eye gaze could be better. According to the survey, the model developed by Su et al. (2023) LipLearner is the best Lip-HCI framework, as it includes a pre-training model employing contrastive learning, key-word spotting, incremental learning, few-shot learning, and a one-shot SVM classifier for each user command.

9. Conclusions and future research directions

Lip reading applications discussed in the survey can be of a great helping hand to society by providing security to systems, voice assistants for devices, a hearing aid for deaf people, for generating video text transcriptions, for pronunciation correction or evaluation, aiding forensics with spy cameras, synthesizing voice for unable to talk patients, attempting speech inpainting during noisy video conferencing, and by providing handsfree HCI for intelligent devices like Smart Phone's, Smart TV, computers, robots, and musical instruments. In this review paper, benchmark datasets available for LR applications, the growth of five lip reading application systems is reviewed, highlighting their progression in the techniques they have employed over time, covering all signal processing, machine vision, machine learning, and deep learning techniques to date.

Lip-reading applications are still in the budding stage. A vast untapped potential exists to improve lip-reading applications using the latest concepts of deep learning, computer vision, machine learning, and signal processing research. Still, other than enhancing LR accuracy, LR application researchers have to face a lot of challenges, like learning from fewer labeled data, incremental learning of networks even after deployment, dealing with unknown classes, adding new classes, and deleting old classes from the network after training, and reducing the complexities of the network for real-time applications. Continuous LR in unconstrained environments still needs to be solved. LR applications face dilemmas because of homophenes, illumination condition variations, spatial and temporal resolution changes, varying word utterance durations, and lip movements during breathing and laughing. Speaker-independent LR is substantially more challenging than speaker-dependent LR, as the mouth gestures of a speech vary among individuals. Also, the speech accent varies from region to region. Speech production results from the collaboration between many organs like the tongue, teeth, lips, jaw, vocal cords, and lungs. Thus, it is very complex to capture speech by just analyzing lower-half facial gestures.

Conventional approaches to LR applications involve extracting the best feature using computer vision techniques and then employing the best ML classifier to improve the accuracy of LR systems. These systems do not seek as much training data as DL systems, and their training and computation times are also shorter. In contrast, deep learning architectures extract the best features during training, and their classification results outperform conventional lip-reading techniques. But DL systems are data-hungry. They demand vast amounts of labeled data for training, training time, computation time, and expensive resources such as GPUs. If they employ neural networks, they are susceptible to catastrophic forgetting, which reduces their compatibility with real-time applications. Most research articles have focused on closed-set systems with constrained datasets, with the results outlined on benchmark datasets. Significantly few research articles have come up with innovative real-time products. Researchers face many real challenges while building real-time products for lip-reading applications.

Among all lip-reading corpus analyzed, unconstrained datasets LRW-500, LRS2, LRS3, LSVSR, and VoxCeleb2 outperform others. Suppose researchers are using specialized cameras to collect databases for LR applications; in that case, the LR applications developed will also require specialized cameras like depth cameras, thermal cameras, or event cameras, which can make LR applications costly. It is extremely difficult to create an annotated dataset by considering all environmental variations, variations in utterance speed, and variations in facial features.

However, compared to AVSR, SSR, and voice-from-lips applications, there needs to be more literature on LR biometrics and lip HCI applications. Analyzing the limitations of existing works in the field of LR biometrics, there is a need to develop a biometric system capable of performing one-class classification with few-shot learning that can effectively detect unknown classes as opposed to incorrectly classifying them into a group of known types. This will provide a pathway for future research in this area. Future works can use customized autoencoders, GANs, or Siamese networks with the best-handcrafted feature extracted using the computer vision approach and the best CNN feature extraction for one-class classification. LR biometrics can be combined with other techniques, such as liveness detection and face recognition, to provide a high level of security. For SSR, numerous frameworks that achieve accurate word recognition are proposed, which can be incorporated into LR biometrics.

Future directions in AVSR and SSR could include designing backbone architectures such as front-end and back-end networks, video encoders, audio encoders, and audio–video integration networks to enhance performance. The greater the network depth, the greater the model complexity regarding time and memory requirements, necessitating high-end resources such as GPUs. Thus, future work can be carried out to construct lightweight models without sacrificing performance. Since pretraining the SSR model demands more time and annotations, future work can concentrate on eliminating them and encouraging the model to learn with limited annotations. Future models that detect unseen or out-of-vocabulary classes with sufficient accuracy could be developed. Even though the accuracy of SSR models for speaker-dependent settings is quite good, SSR models for speaker-independent settings still require substantial improvement. Instead of a random sampling strategy, exploring alternative training methods such as curriculum learning is possible.

Scholars can develop novel loss functions that make the model learns better than the proposed ones. New preprocessing techniques like Keyframe extraction, image enhancement, and augmentations can be explored to boost SSR performance. Multi-steam or hybrid SSR can be investigated. Novel paradigms to perform sentence-level end-end classifications should be researched. Language models and dictionaries are available to build SSR for the English language, but work can be carried out to build SSR for other languages. Dealing with homophenes is the greatest challenge in SSR, which has to be tackled in future works. Although many frameworks perform well on existing constrained datasets, future directions will be to achieve better performance on unconstrained datasets and real-time SSR. For AVSR, researchers can develop more novel DL feature fusion strategies.

As per the review of the results from an existing voice from lips application systems, for constrained corpus like GRID, TCD-TIMIT, OuluVS2, and unconstrained small corpora like LRW-500 and lip2wav, models can generate intelligible speech but fail to have naturalness in speech quality, researchers have put their best effort to bring realism in the speech by training system with a variety of loss functions but still fail to bring the expected quality in produced speech, and they sound more robotic and lack prosody and emotions. Along with improving speech quality, future research should focus on improving the accuracy of predicting unseen words or out-of-vocabulary words and unseen speaker words in an unconstrained environment. Although voice-from-lip systems have achieved better performance in speaker-dependent, multi-speaker-dependent settings, significant breakthrough research must still happen to carry out the task in speaker-independent settings and detect out-of-vocabulary words.

Currently, lip HCI is utilized in very few devices. In the future, researchers can use them to control smart home devices for home automation and provide HCI for disabled people to control and access devices. If the SSR models are bulky, they can be deployed on the cloud, or if they are tiny, they can be easily deployed on edge devices. For real-time implementation, the system's response time has to be considerably alleviated. In these systems, enough care should be taken that a user laughing or uttering other speech phrases should not be interpreted as commands.

CRediT authorship contribution statement

Preethi S.J.: Data curation, Conceptualization, Methodology, Visualization, Investigation, Writing – original draft, Writing – review, and editing. **Niranjana Krupa B.:** Conceptualization, Visualization, Investigation, Formal analysis, Supervision, Writing – review, and editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

The authors are grateful to PES University, Bangalore, for supporting the ongoing research work. We express our gratitude to the Reviewers. Their careful reading, insightful comments, and valuable suggestions greatly assisted us in enhancing the quality of the paper.

References

Abdrakhmanova, M., Kuzdeuov, A., Jarju, S., Khassanov, Y., Lewis, M., Varol, H.A., 2021. SpeakingFaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. Sensors 21 (3465), http://dx.doi.org/10.3390/s21103465.

Abrar, M.A., Islam, A.N.M.N., Hassan, M.M., Islam, M.T., Shahnaz, C., Fattah, S.A., 2019. Deep lip reading-a deep learning based lip-reading software for the hearing impaired. In: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129). Presented at the 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129). pp. 40–44. http://dx.doi.org/10.1109/R10-HTC47129.2019.9042439.

Addarrazi, I., Satori, H., Satori, K., 2020. A follow-up survey of audiovisual speech integration strategies. In: Bhateja, V., Satapathy, S.C., Satori, H. (Eds.), Embedded Systems and Artificial Intelligence, Advances in Intelligent Systems and Computing. Springer, Singapore, pp. 635–643. http://dx.doi.org/10.1007/978-981-15-0947-6_60.

Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A., 2018a. Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. 1. http://dx.doi.org/10.1109/TPAMI.2018.2889052.

Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A., 2022. Deep audio-visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. 44, 8717–8727. http://dx.doi.org/10.1109/TPAMI.2018.2889052.

Afouras, T., Chung, J.S., Zisserman, A., 2018b. LRS3-TED: A large-scale dataset for visual speech recognition. arXiv:1809.00496 [Cs].

Afouras, Triantafyllos, Chung, J.S., Zisserman, A., 2020a. ASR is all you need: Cross-modal distillation for lip reading. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 2143–2147. http://dx.doi.org/10.1109/ICASSP40776.2020.9054253.

Afouras, T., Chung, J.S., Zisserman, A., 2020b. Now you're speaking my language: visual language identification. In: Proceedings of ISCA 2020.

Akbari, H., Arora, H., Cao, L., Mesgarani, N., 2018. Lip2Audspec: Speech reconstruction from silent lip movements video. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 2516–2520. http://dx.doi.org/10.1109/ICASSP.2018.8461856.

Al-Ghanim, A., AL-Oboud, N., Al-Haidary, R., Al-Zeer, S., Altammami, S., Mahmoud, H.A., 2013. I see what you say (ISWYS): Arabic lip reading system. In: 2013 International Conference on Current Trends in Information Technology (CTIT). Presented at the 2013 International Conference on Current Trends in Information Technology. CTIT, pp. 11–17. http://dx.doi.org/10.1109/CTIT.2013.6749470.

Alcantarilla, P.F., Bartoli, A., Davison, A.J., 2012. KAZE features. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (Eds.), Computer Vision – ECCV 2012. In: Lecture Notes in Computer Science., Springer, Berlin, Heidelberg, pp. 214–227. http://dx.doi.org/10.1007/978-3-642-33783-3_16.

Alghathbar, K., Mahmoud, H.A., 2009. Block-based motion estimation analysis for lip reading user authentication systems. WSEAS Trans. Info. Sci. Appl. 6, 829–838.

Almajai, I., Cox, S., Harvey, R., Lan, Y., 2016. Improved speaker independent lip reading using speaker adaptive training and deep neural networks. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 2722–2726. http://dx.doi.org/10.1109/ICASSP.2016.7472172.

Anina, I., Zhou, Z., Zhao, G., Pietikäinen, M., 2015. OuluVS2: A multi-view audiovisual database for non-rigid mouth motion analysis. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Presented at the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. FG, pp. 1–5. http://dx.doi.org/10.1109/FG.2015.7163155.

Assael, Y.M., Shillingford, B., Whiteson, S., de Freitas, N., 2016. LipNet: End-to-end sentence-level lipreading. arXiv:1611.01599 [Cs].

Bear, H.L., Cox, S.J., Harvey, R.W., 2017. Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. arXiv:1710.01122 [Cs, Eess].

Bear, H.L., Harvey, R.W., Theobald, B.-J., Lan, Y., 2014. Which phoneme-to-viseme maps best improve visual-only computer lip-reading? In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., McMahan, R., Jerald, J., Zhang, H., Drucker, S.M., Kambhamettu, C., El Choubassi, M., Deng, Z., Carlson, M. (Eds.), Advances in Visual Computing, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 230–239. http://dx.doi.org/10.1007/978-3-319-14364-4_22.

Borde, P., Kulkarni, S., Gawali, B., Yannawar, P., 2022. Recognition of isolated digit using random forest for audio-visual speech recognition. Proc. Natl. Acad. Sci. India, Sect. A Phys. Sci 92, 103–110. http://dx.doi.org/10.1007/s40010-020-00724-7.

Burton, J., Frank, D., Saleh, M., Navab, N., Bear, H.L., 2018. The speaker-independent lipreading play-off; a survey of lipreading machines. In: 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS). Presented at the 2018 IEEE International Conference on Image Processing, Applications and Systems. IPAS, pp. 125–130. http://dx.doi.org/10.1109/IPAS.2018.8708874.

Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. IEEE Trans. Affect. Comput. 5, 377–390. http://dx.doi.org/10.1109/TAFFC.2014.2336244.

Chen, W., Tan, X., Xia, Y., Qin, T., Wang, Y., Liu, T.-Y., 2020. DualLip: A system for joint lip reading and generation. In: Proceedings of the 28th ACM International Conference on Multimedia. MM '20, Association for Computing Machinery, New York, NY, USA, pp. 1985–1993. http://dx.doi.org/10.1145/3394171.3413623.

Cheng, S., Ma, P., Tzimiropoulos, G., Petridis, S., Bulat, A., Shen, J., Pantic, M., 2020. Towards pose-invariant lip-reading. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 4357–4361. http://dx.doi.org/10.1109/ICASSP40776.2020.9054384.

Cheng, F., Wang, S.-L., Liew, A.W.-C., 2018. Visual speaker authentication with random prompt texts by a dual-task CNN framework. Pattern Recognit. 83, 340–352. http://dx.doi.org/10.1016/j.patcog.2018.06.005.

Cheung, Y., Zhou, Y., 2018. Lip password-based speaker verification without a priori knowledge of speech language. In: Li, K., Li, W., Chen, Z., Liu, Y. (Eds.), Computational Intelligence and Intelligent Systems, Communications in Computer and Information Science. Springer, Singapore, pp. 461–472. http://dx.doi.org/10.1007/978-981-13-1651-7_41.

Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118, 887–906. http://dx.doi.org/10.1121/1.1945807.

Choudhury, A., Roy, P., Bandyopadhyay, S., 2023. Review of various machine learning and deep learning techniques for audio visual automatic speech recognition. In: 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC). Presented at the 2023 International Conference on Intelligent Systems, Advanced Computing and Communication. ISACC, pp. 1–10.

Chowdhury, D.P., Kumari, R., Bakshi, S., Sahoo, M.N., Das, A., 2022. Lip as biometric and beyond: A survey. Multimed Tools Appl. 81, 3831–3865. http://dx.doi.org/10.1007/s11042-021-11613-5.

Chung, J.S., Nagrani, A., Zisserman, A., 2018. VoxCeleb2: Deep speaker recognition. In: Interspeech 2018. Presented at the Interspeech 2018. ISCA, pp. 1086–1090. http://dx.doi.org/10.21437/Interspeech.2018-1929.

Chung, J.S., Senior, A., Vinyals, O., Zisserman, A., 2017. Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3444–3453. http://dx.doi.org/10.1109/CVPR.2017.367.

Chung, J.S., Zisserman, A., 2017. Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (Eds.), Computer Vision – ACCV 2016. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 87–103. http://dx.doi.org/10.1007/978-3-319-54184-6_6.

Chung, J.S., Zisserman, A., 2018. Learning to lip read words by watching videos. Comput. Vis. Image Underst. 173, 76–85. http://dx.doi.org/10.1016/j.cviu.2018.02.001.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120, 2421–2424. http://dx.doi.org/10.1121/1.2229005.

Courtney, L., Sreenivas, R., 2020. Using deep convolutional LSTM networks for learning spatiotemporal features. In: Palaiahnakote, S., Sanniti Di Baja, G., Wang, L., Yan, W.Q. (Eds.), Pattern Recognition, Lecture Notes in Computer Science. Presented at the Asian Conference on Pattern Recognition. Springer International Publishing, Cham, pp. 307–320. http://dx.doi.org/10.1007/978-3-030-41299-9_24.

Cox, S., Harvey, R., Lan, Y., Newman, J., Theobald, B., 2008. The challenge of multispeaker lip-reading. In: International Conference on AuditoryVisual Speech Processing.

Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., Szykulski, M., 2017. An audio-visual corpus for multimodal automatic speech recognition. J. Intell. Inf. Syst. 49, 167–192. http://dx.doi.org/10.1007/s10844-016-0438-z.

Ephrat, A., Halperin, T., Peleg, S., 2017. Improved speech reconstruction from silent video. In: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Presented at the 2017 IEEE International Conference on Computer Vision Workshops. ICCVW, pp. 455–462. http://dx.doi.org/10.1109/ICCVW.2017.61.

Ephrat, A., Peleg, S., 2017. Vid2speech: Speech reconstruction from silent video. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 5095–5099. http://dx.doi.org/10.1109/ICASSP.2017.7953127.

Estival, D., Cassidy, S., Cox, F., Burnham, D.K., 2014. AusTalk : An audio-visual corpus of Australian English. In: Proceedings of the 9th International Conference on Language Resources and Evaluation. LREC 2014, 26-31 2014, Reykjavik, Iceland, pp. 3105–3109.

Ezz, M., Mostafa, A.M., Nasr, A.A., 2020. A silent password recognition framework based on lip analysis. IEEE Access 8, 55354–55371. http://dx.doi.org/10.1109/ACCESS.2020.2982359.

Farrukh, W., van der Haar, D., 2023. Lip print-based identification using traditional and deep learning. IET Biometrics http://dx.doi.org/10.1049/bme2.12073.

Feng, D., Yang, S., Shan, S., Chen, X., 2022. Audio-driven deformation flow for effective lip reading. In: 2022 26th International Conference on Pattern Recognition (ICPR). Presented at the 2022 26th International Conference on Pattern Recognition. ICPR, pp. 274–280. http://dx.doi.org/10.1109/ICPR56361.2022.9956316.

Fenghour, S., Chen, D., Guo, K., Li, B., Xiao, P., 2021. Deep learning-based automated lip-reading: A survey. IEEE Access 9, 121184–121205. http://dx.doi.org/10.1109/ACCESS.2021.3107946.

Fenghour, S., Chen, D., Guo, K., Xiao, P., 2020. Lip reading sentences using deep learning with only visual cues. IEEE Access 8, 215516–215530. http://dx.doi.org/10.1109/ACCESS.2020.3040906.

Fenghour, S., Chen, D., Xiao, P., 2019. Contour mapping for speaker-independent lip reading system. In: Eleventh International Conference on Machine Vision (ICMV 2018). Presented at the Eleventh International Conference on Machine Vision. ICMV 2018, SPIE, pp. 282–289. http://dx.doi.org/10.1117/12.2522936.

Fernandez-Lopez, A., Karaali, A., Harte, N., Sukno, F.M., 2020. Cogans for unsupervised visual speech adaptation to new speakers. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6294–6298. http://dx.doi.org/10.1109/ICASSP40776.2020.9053299.

Fernandez-Lopez, A., Martinez, O., Sukno, F.M., 2017. Towards estimating the upper bound of visual-speech recognition: The visual lip-reading feasibility database. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). Presented at the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition. FG 2017, pp. 208–215. http://dx.doi.org/10.1109/FG.2017.34.

Fernandez-Lopez, A., Sukno, F.M., 2018. Survey on automatic lip-reading in the era of deep learning. Image Vis. Comput. 78, 53–72. http://dx.doi.org/10.1016/j.imavis.2018.07.002.

Fox, N.A., O'Mullane, B.A., Reilly, R.B., 2005. VALID: A new practical audio-visual database, and comparative results. In: Kanade, T., Jain, A., Ratha, N.K. (Eds.), Audio- and Video-Based Biometric Person Authentication, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 777–786. http://dx.doi.org/10.1007/11527923_81.

Gao, Y., Jin, Y., Li, J., Choi, S., Jin, Z., 2020. EchoWhisper: Exploring an acoustic-based silent speech interface for smartphone users. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 4 80:1–80:27. http://dx.doi.org/10.1145/3411830.

Garg, A., Noyola, J., Bagadia, S., 2016. Lip Reading using CNN and LSTM. Technical report, Stanford University, CS231 n project report 9.

Gilbert, J.M., Rybchenko, S.I., Hofe, R., Ell, S.R., Fagan, M.J., Moore, R.K., Green, P., 2010. Isolated word recognition of silent speech using magnetic implants and sensors. Med. Eng. Phys. 32, 1189–1197. http://dx.doi.org/10.1016/j.medengphy.2010.08.011.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., Pang, R., 2020. Conformer: Convolution-augmented transformer for speech recognition. http://dx.doi.org/10.48550/arXiv.2005.08100.

Hao, M., Mamut, M., Yadikar, N., Aysa, A., Ubul, K., 2020. A survey of research on lipreading technology. IEEE Access 8, 204518–204544. http://dx.doi.org/10.1109/ACCESS.2020.3036865.

Harte, N., Gillen, E., 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. IEEE Trans. Multimed. 17, 603–615. http://dx.doi.org/10.1109/TMM.2015.2407694.

Hassanat, A.B., 2014. Visual passwords using automatic lip reading. arXiv:1409.0924 [Cs].

Hazen, T.J., Saenko, K., La, C.-H., Glass, J.R., 2004. A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments. In: Proceedings of the 6th International Conference on Multimodal Interfaces. ICMI '04, Association for Computing Machinery, New York, NY, USA, pp. 235–242. http://dx.doi.org/10.1145/1027933.1027972.

He, J., Zhao, Z., Ren, Y., Liu, J., Huai, B., Yuan, N., 2022. Flow-based unconstrained lip to speech generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. p. 9.

Hegde, S.B., Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.V., 2022. Lip-to-speech synthesis for arbitrary speakers in the wild. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, Association for Computing Machinery, New York, NY, USA, pp. 6250–6258. http://dx.doi.org/10.1145/3503161.3548081.

Heracleous, P., Hagita, N., 2011. Automatic recognition of speech without any audio information. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 2392–2395. http://dx.doi.org/10.1109/ICASSP.2011.5946965.

Hong, J., Kim, M., Park, S.J., Ro, Y.M., 2021. Speech reconstruction with reminiscent sound via visual voice memory. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3654–3667. http://dx.doi.org/10.1109/TASLP.2021.3126925.

Howell, D., 2015. Confusion Modelling for Lip-Reading (Ph.D.). University of East Anglia.

Huang, J., Kingsbury, B., 2013. Audio-visual deep learning for noise robust speech recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Presented at the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 7596–7599. http://dx.doi.org/10.1109/ICASSP.2013.6639140.

Huang, H., Song, C., Ting, J., Tian, T., Hong, C., Di, Z., Gao, D., 2022. A novel machine lip reading model. In: Procedia Computer Science, the 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021): Developing Global Digital Economy After COVID-19. 199, pp. 1432–1437. http://dx.doi.org/10.1016/j.procs.2022.01.181.

Huang, A., Zhang, X., 2022. Dual-flow spatio-temporal separation network for lip reading. J. Phys.: Conf. Ser. 2400, 012028. http://dx.doi.org/10.1088/1742-6596/2400/1/012028.

Hwang, I.-S., Tsai, Y.-Y., Zeng, B.-H., Lin, C.-M., Shiue, H.-S., Chang, G.-C., 2021. Integration of eye tracking and lip motion for hands-free computer access. Univ. Access. Inf. Soc. 20, 405–416. http://dx.doi.org/10.1007/s10209-020-00723-w.

Ivanko, D., Ryumin, D., Karpov, A., 2021. An experimental analysis of different approaches to audio–Visual speech recognition and lip-reading. In: Ronzhin, A., Shishlakov, V. (Eds.), Proceedings of 15th International Conference on Electromechanics and Robotics Zavalishin's Readings, Smart Innovation, Systems and Technologies. Springer, Singapore, pp. 197–209. http://dx.doi.org/10.1007/978-981-15-5580-0_16.

Ivanko, D., Ryumin, D., Karpov, A., 2022a. Developing of a software–hardware complex for automatic audio–Visual speech recognition in human–robot interfaces. In: Ronzhin, A., Shishlakov, V. (Eds.), Electromechanics and Robotics, Smart Innovation, Systems and Technologies. Springer, Singapore, pp. 259–270. http://dx.doi.org/10.1007/978-981-16-2814-6_23.

Ivanko, D., Ryumin, D., Kashevnik, A., Axyonov, A., Karnov, A., 2022b. Visual speech recognition in a driver assistance system. In: 2022 30th European Signal Processing Conference (EUSIPCO). Presented at the 2022 30th European Signal Processing Conference. EUSIPCO, pp. 1131–1135. http://dx.doi.org/10.23919/EUSIPCO55093.2022.9909819.

Ivanko, D., Ryumin, D., Kipyatkova, I., Axyonov, A., Karpov, A., 2019. Lip-reading using pixel-based and geometry-based features for multimodal human–robot interfaces. In: Ronzhin, A., Shishlakov, V. (Eds.), Proceedings of 14th International Conference on Electromechanics and Robotics Zavalishin's Readings, Smart Innovation, Systems and Technologies. Springer, Singapore, pp. 477–486. http://dx.doi.org/10.1007/978-981-13-9267-2_39.

Kandagal, A.P., Udayashankara, V., 2014. Automatic bimodal audiovisual speech recognition: A review. In: 2014 International Conference on Contemporary Computing and Informatics (IC3I). Presented at the 2014 International Conference on Contemporary Computing and Informatics. IC3I, pp. 940–945. http://dx.doi.org/10.1109/IC3I.2014.7019673.

Kapkar, P.P., Bharkad, S.D., 2021. Double authentication system based on face identification and lipreading. In: Santosh, K.C., Gawali, B. (Eds.), Recent Trends in Image Processing and Pattern Recognition, Communications in Computer and Information Science. Springer, Singapore, pp. 214–224. http://dx.doi.org/10.1007/978-981-16-0507-9_19.

Katsaggelos, A.K., Bahaadini, S., Molina, R., 2015. Audiovisual fusion: Challenges and new approaches. Proc. IEEE 103, 1635–1653. http://dx.doi.org/10.1109/JPROC.2015.2459017.

Kim, M., Hong, J., Park, S.J., Ro, Y.M., 2022a. CroMM-VSR: Cross-modal memory augmented visual speech recognition. IEEE Trans. Multimed. 24, 4342–4355. http://dx.doi.org/10.1109/TMM.2021.3115626.

Kim, M., Hong, J., Ro, Y.M., 2021. Lip to speech synthesis with visual context attentional GAN. In: Advances in Neural Information Processing Systems. Curran Associates, Inc., pp. 2758–2770.

Kim, M., Yeo, J.H., Ro, Y.M., 2022b. Distinguishing homophenes using multi-head visual-audio memory for lip reading. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. pp. 1174–1182. http://dx.doi.org/10.1609/aaai.v36i1.20003.

King, D.E., 2009. Dlib-ml: A machine learning toolkit. J. Mach. Learn. Res 10, 1755–1758.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am. 87, 820–857. http://dx.doi.org/10.1121/1.398894.

Kumar, K., Chen, T., Stern, R.M., 2007. Profile view lip reading. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. Presented at the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP '07, pp. IV–429–IV–432. http://dx.doi.org/10.1109/ICASSP.2007.366941.

Kumar, Y., Jain, R., Salik, K.M., Shah, R.R., Yin, Y., Zimmermann, R., 2019. Lipper: Synthesizing thy speech using multi-view lipreading. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33. pp. 2588–2595. http://dx.doi.org/10.1609/aaai.v33i01.33012588.

Kumar, L.A., Renuka, D.K., Rose, S.L., Shunmuga priya, M.C., Wartana, I.M., 2022. Deep learning based assistive technology on audio visual speech recognition for hearing impaired. Int. J. Cogn. Comput. Eng. 3, 24–30. http://dx.doi.org/10.1016/j.ijcce.2022.01.003.

Lai, J.-Y., Wang, S.-L., Liew, A.W.-C., Shi, X.-J., 2016. Visual speaker identification and authentication by joint spatiotemporal sparse coding and hierarchical pooling. Inform. Sci. 373, 219–232. http://dx.doi.org/10.1016/j.ins.2016.09.015.

Lan, Y., Theobald, B., Harvey, R., Ong, E., 2010. Improving visual features for lip-reading. In: International Conference on Auditory-Visual Speech Processing. Presented at the International Conference on Auditory-Visual Speech Processing, Volterra.

Le Cornu, T., Milner, B., 2017. Generating intelligible audio speech from visual speech. IEEE/ACM Trans. Audio Speech Lang. Process. 25, 1751–1761. http://dx.doi.org/10.1109/TASLP.2017.2716178.

Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T., 2004. AVICAR: Audio-visual speech corpus in a car environment. In: Interspeech 2004. Presented at the Interspeech 2004. ISCA, pp. 2489–2492. http://dx.doi.org/10.21437/Interspeech.2004-424.

Lee, D., Myung, K., 2017. Read my lips, login to the virtual world. In: 2017 IEEE International Conference on Consumer Electronics (ICCE). Presented at the 2017 IEEE International Conference on Consumer Electronics. ICCE, pp. 434–435. http://dx.doi.org/10.1109/ICCE.2017.7889386.

Lesani, F.S., Ghazvini, F.F., Dianat, R., 2015. Mobile phone security using automatic lip reading. In: 2015 9th International Conference on E-Commerce in Developing Countries: with Focus on E-Business (ECDC). Presented at the 2015 9th International Conference on E-Commerce in Developing Countries: with Focus on E-Business. ECDC, pp. 1–5. http://dx.doi.org/10.1109/ECDC.2015.7156322.

Liam McQuillan, L., 2022. Liopa - The world's only startup focused on automated lipreading via visual speech recognition. URL https://liopa.ai/. (Accessed 29 July 2022).

Lin, Z., Zhao, Z., Li, H., Liu, J., Zhang, M., Zeng, X., He, X., 2021. SimulLR: Simultaneous lip reading transducer with attention-guided adaptive memory. In: Proceedings of the 29th ACM International Conference on Multimedia. MM '21, Association for Computing Machinery, New York, NY, USA, pp. 1359–1367. http://dx.doi.org/10.1145/3474085.3475220.

Liu, H., Chen, Z., Yang, B., 2020. Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion. In: Interspeech 2020. Presented at the Interspeech 2020. ISCA, pp. 3520–3524. http://dx.doi.org/10.21437/Interspeech.2020-3146.

Liu, X., Cheung, Y., 2014. Learning multi-boosted HMMs for lip-password based speaker verification. IEEE Trans. Inf. Forensics Secur. 9, 233–246. http://dx.doi.org/10.1109/TIFS.2013.2293025.

Liu, H., Li, W., Yang, B., 2021a. Robust audio-visual speech recognition based on hybrid fusion. In: 2020 25th International Conference on Pattern Recognition (ICPR). Presented at the 2020 25th International Conference on Pattern Recognition. ICPR, pp. 7580–7586. http://dx.doi.org/10.1109/ICPR48806.2021.9412817.

Liu, M., Wang, L., Lee, K.A., Zhang, H., Zeng, C., Dang, J., 2021. DeepLip: A benchmark for deep learning-based audio-visual lip biometrics. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Presented at the 2021 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 122–129. http://dx.doi.org/10.1109/ASRU51503.2021.9688240.

Liu, H., Xu, W., Yang, B., 2021b. Audio-visual speech recognition using a two-step feature fusion strategy. In: 2020 25th International Conference on Pattern Recognition (ICPR). Presented at the 2020 25th International Conference on Pattern Recognition. ICPR, pp. 1896–1903. http://dx.doi.org/10.1109/ICPR48806.2021.9412454.

Lu, Y., Li, H., 2019. Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory. Appl. Sci. 9 (1599), http://dx.doi.org/10.3390/app9081599.

Lu, Y., Tian, H., Cheng, J., Zhu, F., Liu, B., Wei, S., Ji, L., Wang, Z.L., 2022. Decoding lip language using triboelectric sensors with deep learning. Nature Commun. 13, 1401. http://dx.doi.org/10.1038/s41467-022-29083-0.

Lu, L., Yu, J., Chen, Y., Liu, H., Zhu, Y., Kong, L., Li, M., 2019. Lip reading-based user authentication through acoustic sensing on smartphones. IEEE/ACM Trans. Netw. 27, 447–460. http://dx.doi.org/10.1109/TNET.2019.2891733.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M.G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M., 2019. MediaPipe: A framework for building perception pipelines. arXiv e-prints.

Ma, P., Martinez, B., Petridis, S., Pantic, M., 2021a. Towards practical lipreading with distilled and efficient models. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7608–7612. http://dx.doi.org/10.1109/ICASSP39728.2021.9415063.

Ma, P., Petridis, S., Pantic, M., 2021b. End-to-end audio-visual speech recognition with conformers. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7613–7617. http://dx.doi.org/10.1109/ICASSP39728.2021.9414567.

Ma, P., Petridis, S., Pantic, M., 2022a. Visual speech recognition for multiple languages in the wild. Nat. Mach. Intell. 4, 930–939. http://dx.doi.org/10.1038/s42256-022-00550-z.

Ma, P., Wang, Y., Petridis, S., Shen, J., Pantic, M., 2022b. Training strategies for improved lip-reading. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 8472–8476. http://dx.doi.org/10.1109/ICASSP43922.2022.9746706.

Ma, J., Wang, S., Zhang, A., Liew, A.W.-C., 2020. Feature extraction for visual speaker authentication against computer-generated video attacks. In: 2020 IEEE International Conference on Image Processing (ICIP). Presented at the 2020 IEEE International Conference on Image Processing. ICIP, pp. 1326–1330. http://dx.doi.org/10.1109/ICIP40778.2020.9190976.

Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., Siohan, O., 2019. Recurrent neural network transducer for audio-visual speech recognition. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Presented at the 2019 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 905–912. http://dx.doi.org/10.1109/ASRU46091.2019.9004036.

Martinez, B., Ma, P., Petridis, S., Pantic, M., 2020. Lipreading using temporal convolutional networks. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6319–6323. http://dx.doi.org/10.1109/ICASSP40776.2020.9053841.

Mathulaprangsan, S., Wang, C.-Y., Kusum, A.Z., Tai, T.-C., Wang, J.-C., 2015. A survey of visual lip reading and lip-password verification. In: 2015 International Conference on Orange Technologies (ICOT). Presented at the 2015 International Conference on Orange Technologies. ICOT, pp. 22–25. http://dx.doi.org/10.1109/ICOT.2015.7498485.

Matthews, I., Cootes, T.F., Bangham, J.A., Cox, S., Harvey, R., 2002. Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. 24, 198–213. http://dx.doi.org/10.1109/34.982900.

McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matejka, P., Cernocký, J., Poh, N., Kittler, J., Larcher, A., Lévy, C., Matrouf, D., Bonastre, J.-F., Tresadern, P., Cootes, T., 2012. Bi-modal person recognition on a mobile phone: using mobile phone data. In: 2012 IEEE International Conference on Multimedia and Expo Workshops. Presented at the 2012 IEEE International Conference on Multimedia and Expo Workshops. pp. 635–640. http://dx.doi.org/10.1109/ICMEW.2012.116.

Messer, K., Matas, J., Kittler, J., Lüttin, J., Maitre, G., 1999. XM2vtsdb: The extended M2VTS database. In: Second International Conference on Audio and Video-Based Biometric Person Authentication. pp. 72–77.

Michelsanti, D., Slizovskaia, O., Haro, G., Gómez, E., Tan, Z.-H., Jensen, J., 2020. Vocoder-based speech synthesis from silent videos. In: Interspeech 2020. Presented at the Interspeech 2020. ISCA, pp. 3530–3534. http://dx.doi.org/10.21437/Interspeech.2020-1026.

Millar, J.B., Goecke, R., 2004. The audio-video australian english speech data corpus AVOZES. In: Interspeech 2004. Presented at the Interspeech 2004. ISCA, pp. 2525–2528. http://dx.doi.org/10.21437/Interspeech.2004-433.

Milner, B., Le Cornu, T., 2015. Reconstructing intelligible audio speech from visual speech features. In: Interspeech 2015. Presented at the Interspeech 2015. DEU.

Mira, R., Vougioukas, K., Ma, P., Petridis, S., Schuller, B.W., Pantic, M., 2022. End-to-end video-to-speech synthesis using generative adversarial networks. IEEE Trans. Cybern..

Morise, M., Yokomori, F., Ozawa, K., 2016. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. Inf. Syst. E99-D 1877–1884.

Morrone, G., Michelsanti, D., Tan, Z.-H., Jensen, J., 2021. Audio-visual speech inpainting with deep learning. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6653–6657. http://dx.doi.org/10.1109/ICASSP39728.2021.9413488.

Movellan, J.R., 1994. Visual speech recognition with stochastic networks. In: Proceedings of the 7th International Conference on Neural Information Processing Systems. NIPS'94, MIT Press, Cambridge, MA, USA, pp. 851–858.

NadeemHashmi, S., Gupta, H., Mittal, D., Kumar, K., Nanda, A., Gupta, S., 2018. A lip reading model using CNN with batch normalization. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). Presented at the 2018 Eleventh International Conference on Contemporary Computing. IC3, pp. 1–6. http://dx.doi.org/10.1109/IC3.2018.8530509.

Nagrani, A., Chung, J.S., Xie, W., Zisserman, A., 2020. Voxceleb: Large-scale speaker verification in the wild. Comput. Speech Lang. 60, 101027. http://dx.doi.org/10.1016/j.csl.2019.101027.

Nagrani, A., Chung, J.S., Zisserman, A., 2017. VoxCeleb: A large-scale speaker identification dataset. In: Interspeech 2017. Presented at the Interspeech 2017. ISCA, pp. 2616–2620. http://dx.doi.org/10.21437/Interspeech.2017-950.

Nandini, M.S., Bhajantri, N.U., Nagavi, T.C., 2020. Correlation of visual perceptions and extraction of visual articulators for kannada lip reading. In: Panigrahi, C.R., Pati, B., Mohapatra, P., Buyya, R., Li, K.-C. (Eds.), Progress in Advanced Computing and Intelligent Engineering, Advances in Intelligent Systems and Computing. Springer, Singapore, pp. 252–265. http://dx.doi.org/10.1007/978-981-15-6353-9_23.

Nemani, P., Krishna, G.S., Ramisetty, N., Sai, B.D.S., Kumar, S., 2022. Deep learning based holistic speaker independent visual speech recognition. IEEE Trans. Artif. Intell. 1–10. http://dx.doi.org/10.1109/TAI.2022.3220190.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T., 2015. Audio-visual speech recognition using deep learning. Appl. Intell. 42, 722–737. http://dx.doi.org/10.1007/s10489-014-0629-7.

Paleček, K., 2018. Experimenting with lipreading for large vocabulary continuous speech recognition. J. Multimodal User Interfaces 12, 309–318. http://dx.doi.org/10.1007/s12193-018-0266-2.

Pandey, L., Arif, A.S., 2021. Liptype: a silent speech recognizer augmented with an independent repair model. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–19.

Pandey, L., Arif, A.S., 2022. Design and evaluation of a silent speech-based selection method for eye-gaze pointing. Proc. ACM Hum.-Comput. Interact. 6, 570:328–570:353. http://dx.doi.org/10.1145/3567723.

Parekh, D., Gupta, A., Chhatpar, S., Yash, A., Kulkarni, M., 2019. Lip reading using convolutional auto encoders as feature extractor. In: 2019 IEEE 5th International Conference for Convergence in Technology (I2CT). Presented at the 2019 IEEE 5th International Conference for Convergence in Technology. I2CT, pp. 1–6. http://dx.doi.org/10.1109/I2CT45611.2019.9033664.

Parkhi, O., Vedaldi, A., Zisserman, A., 2019. Deep face recognition. In: BMVC 2015 - Proceedings of the British Machine Vision Conference, Vol. 2015.

Pass, A., Zhang, J., Stewart, D., 2010. AN investigation into features for multi-view lipreading. In: 2010 IEEE International Conference on Image Processing. Presented at the 2010 IEEE International Conference on Image Processing. pp. 2417–2420. http://dx.doi.org/10.1109/ICIP.2010.5650963.

Patterson, E.K., Gurbuz, S., Tufekci, Z., Gowdy, J.N., 2002. CUAVE: A new audio-visual database for multimodal human–computer interface research. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. Presented at the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. II–2017–II–2020. http://dx.doi.org/10.1109/ICASSP.2002.5745028.

Petridis, S., Shen, J., Cetin, D., Pantic, M., 2018. Visual-only recognition of normal, whispered and silent speech. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6219–6223. http://dx.doi.org/10.1109/ICASSP.2018.8461596.

Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W., 2003. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE 91, 1306–1326. http://dx.doi.org/10.1109/JPROC.2003.817150.

Prajwal, K.R., Afouras, T., Zisserman, A., 2022. Sub-word level lip reading with visual attention. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 5152–5162. http://dx.doi.org/10.1109/CVPR52688.2022.00510.

Prajwal, K.R., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.V., 2020. Learning individual speaking styles for accurate lip to speech synthesis. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 13793–13802. http://dx.doi.org/10.1109/CVPR42600.2020.01381.

Pu, G., Wang, H., 2022. Review on research progress of machine lip reading. Vis. Comput. 1–17. http://dx.doi.org/10.1007/s00371-022-02511-4.

Pujari, S., Sneha, S., Vinusha, R., Bhuvaneshwari, P., Yashaswini, C., 2021. A survey on deep learning based lip-reading techniques. In: 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). Presented at the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks. ICICV, pp. 1286–1293. http://dx.doi.org/10.1109/ICICV50876.2021.9388569.

Radha, N., Shahina, A., khan, A.N., 2021. A survey on visual speech recognition approaches. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). Presented at the 2021 International Conference on Artificial Intelligence and Smart Systems. ICAIS, pp. 934–939. http://dx.doi.org/10.1109/ICAIS50930.2021.9395878.

Radha, N., Shahina, A., Nayeemulla Khan, A., 2015. A person identification system combining recognition of face and lip-read passwords. In: 2015 International Conference on Computing and Network Communications (Coconet). Presented at the 2015 International Conference on Computing and Network Communications. CoCoNet, pp. 882–885. http://dx.doi.org/10.1109/CoCoNet.2015.7411294.

Raghavendra, M., Omprakash, P., R, M.B, 2021. AuthNet: A deep learning based authentication mechanism using temporal facial feature movements (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. pp. 15873–15874.

Rahmani, M.H., Almasganj, F., 2017. Lip-reading via a DNN-HMM hybrid system using combination of the image-based and model-based features. In: 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA). Presented at the 2017 3rd International Conference on Pattern Recognition and Image Analysis. IPRIA, pp. 195–199. http://dx.doi.org/10.1109/PRIA.2017.7983045.

Rekik, A., Ben-Hamadou, A., Mahdi, W., 2014. A new visual speech recognition approach for RGB-d cameras. In: Campilho, A., Kamel, M. (Eds.), Image Analysis and Recognition. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 21–28. http://dx.doi.org/10.1007/978-3-319-11755-3_3.

Rekik, A., Ben-Hamadou, A., Mahdi, W., 2016. An adaptive approach for lip-reading using image and depth data. Multimed. Tools Appl. 75, 8609–8636. http://dx.doi.org/10.1007/s11042-015-2774-3.

Ren, S., Du, Y., Lv, J., Han, G., He, S., 2021. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 13320–13328. http://dx.doi.org/10.1109/CVPR46437.2021.01312.

Rothkrantz, L., 2017. Lip-reading by surveillance cameras. In: 2017 Smart City Symposium Prague (SCSP). Presented at the 2017 Smart City Symposium Prague. SCSP, pp. 1–6. http://dx.doi.org/10.1109/SCSP.2017.7973348.

Ruengprateepsang, K., Wangsiripitak, S., Pasupa, K., 2020. Hybrid training of speaker and sentence models for one-shot lip password. In: Yang, H., Pasupa, K., Leung, A.C.-S., Kwok, J.T., Chan, J.H., King, I. (Eds.), Neural Information Processing. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 363–374. http://dx.doi.org/10.1007/978-3-030-63830-6_31.

Sahu, P., Dua, M., Kumar, A., 2018. Challenges and issues in adopting speech recognition. In: Agrawal, S.S., Devi, A., Wason, R., Bansal, P. (Eds.), Speech and Language Processing for Human–Machine Communications, Advances in Intelligent Systems and Computing. Springer, Singapore, pp. 209–215. http://dx.doi.org/10.1007/978-981-10-6626-9_23.

Sanderson, C., 2002. The VidTIMIT database. IDIAP Communication.

Sayed, H.M., ElDeeb, H.E., Taie, S.A., 2021. Bimodal variational autoencoder for audiovisual speech recognition. Mach. Learn. http://dx.doi.org/10.1007/s10994-021-06112-5.

Sengupta, S., Bhattacharya, A., Desai, P., Gupta, A., 2012. Automated lip reading technique for password authentication. IJAIS 4, 18–24. http://dx.doi.org/10.5120/ijais12-450677.

Serdyuk, D., Braga, O., Siohan, O., 2021. Audio-visual speech recognition is worth $32\times 32\times 8$ voxels. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). Presented at the 2021 IEEE Automatic Speech Recognition and Understanding Workshop. ASRU, pp. 796–802. http://dx.doi.org/10.1109/ASRU51503.2021.9688191.

Shaikh, A.A., Kumar, D.K., Yau, W.C., Azemin, M.Z.C., Gubbi, J., 2010. Lip reading using optical flow and support vector machines. In: 2010 3rd International Congress on Image and Signal Processing. Presented at the 2010 3rd International Congress on Image and Signal Processing. pp. 327–330. http://dx.doi.org/10.1109/CISP.2010.5646264.

Shang, D., Zhang, X., Xu, X., 2018. Face and lip-reading authentication system based on android smart phones. In: 2018 Chinese Automation Congress (CAC). Presented at the 2018 Chinese Automation Congress. CAC, pp. 4178–4182. http://dx.doi.org/10.1109/CAC.2018.8623298.

Shashidhar, R., Patilkulkarni, S., Puneeth, S.B., 2022. Combining audio and visual speech recognition using LSTM and deep convolutional neural network. Int. J. Inf. Tecnol. http://dx.doi.org/10.1007/s41870-022-00907-y.

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., Agiomvrgiannakis, Y., Wu, Y., 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 4779–4783. http://dx.doi.org/10.1109/ICASSP.2018.8461368.

Sheng, C., Kuang, G., Bai, L., Hou, C., Guo, Y., Xu, X., Pietikäinen, M., Liu, L., 2022a. Deep learning for visual speech analysis: A survey. http://dx.doi.org/10.48550/arXiv.2205.10839.

Sheng, C., Liu, L., Deng, W., Bai, L., Liu, Z., Lao, S., Kuang, G., Pietikäinen, M., 2022b. Importance-aware information bottleneck learning paradigm for lip reading. IEEE Trans. Multimed. 1–13. http://dx.doi.org/10.1109/TMM.2022.3210761.

Sheng, C., Zhu, X., Xu, H., Pietikäinen, M., Liu, L., 2022c. Adaptive semantic-spatio-temporal graph convolutional network for lip reading. IEEE Trans. Multimed. 24, 3545–3557. http://dx.doi.org/10.1109/TMM.2021.3102433.

Sheshpoli, A.J., Nadian-Ghomsheh, A., 2018. Temporal and spatial features for visual speech recognition. In: Montaser Kouhsari, S. (Ed.), Fundamental Research in Electrical Engineering. In: Lecture Notes in Electrical Engineering, Springer, Singapore, pp. 135–145. http://dx.doi.org/10.1007/978-981-10-8672-4_10.

Shi, B., Hsu, W.-N., Lakhotia, K., Mohamed, A., 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. http://dx.doi.org/10.48550/arXiv.2201.02184.

Shi, X.-X., Wang, S.-L., Lai, J.-Y., 2016. Visual speaker authentication by ensemble learning over static and dynamic lip details. In: 2016 IEEE International Conference on Image Processing (ICIP). Presented at the 2016 IEEE International Conference on Image Processing. ICIP, pp. 3942–3946. http://dx.doi.org/10.1109/ICIP.2016.7533099.

Shillingford, B., Assael, Y., Hoffman, M.W., Paine, T., Hughes, C., Prabhu, U., Liao, H., Sak, H., Rao, K., Bennett, L., Mulville, M., Denil, M., Coppin, B., Laurie, B., Senior, A., Freitas, N. de, 2019. Large-scale visual speech recognition. In: Interspeech 2019. Presented at the Interspeech 2019. ISCA, pp. 4135–4139. http://dx.doi.org/10.21437/Interspeech.2019-1669.

Shin, J., Kim, H.-I., Park, R.-H., 2015. New interface for musical instruments using lip reading. IET Image Process. 9, 770–776. http://dx.doi.org/10.1049/iet-ipr.2014.1014.

Shorten, C., Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. J. Big Data 6, 60. http://dx.doi.org/10.1186/s40537-019-0197-0.

Shreekumar, J., Shet, G.K., N, V.P, J, P.S, Krupa, N., 2020. Improved viseme recognition using generative adversarial networks. In: 2020 IEEE Region 10 Conference (TENCON). Presented at the 2020 IEEE Region 10 Conference. TENCON, pp. 1118–1123. http://dx.doi.org/10.1109/TENCON50793.2020.9293784.

Shrivastava, N., Saxena, A., Kumar, Y., Shah, R.R., Stent, A., Mahata, D., Kaur, P., Zimmermann, R., 2019. MobiVSR: : Efficient and light-weight neural network for visual speech recognition on mobile devices. In: Interspeech 2019. Presented at the Interspeech 2019. ISCA, pp. 2753–2757. http://dx.doi.org/10.21437/Interspeech.2019-3273.

Shubhangi, R.K., Balbhim, N.B., 2017. Lip's movements biometric authentication in electronic devices. In: 2017 International Conference on Computing Methodologies and Communication (ICCMC). Presented at the 2017 International Conference on Computing Methodologies and Communication. ICCMC, pp. 998–1001. http://dx.doi.org/10.1109/ICCMC.2017.8282619.

Sindhura, P., Preethi, S.J., Niranjana, K.B., 2018. Convolutional neural networks for predicting words: A lip-reading system. In: 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT). Presented at the 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques. ICEECCOT, pp. 929–933. http://dx.doi.org/10.1109/ICEECCOT43722.2018.9001505.

Son, J.S., Zisserman, A., 2017. Lip reading in profile. In: Procedings of the British Machine Vision Conference 2017. Presented at the British Machine Vision Conference 2017. British Machine Vision Association, London, UK, p. 155. http://dx.doi.org/10.5244/C.31.155.

Song, Q., Sun, B., Li, S., 2022. Multimodal sparse transformer network for audio-visual speech recognition. IEEE Trans. Neural Netw. Learn. Syst. 1–11. http://dx.doi.org/10.1109/TNNLS.2022.3163771.

Stafylakis, T., Khan, M.H., Tzimiropoulos, G., 2018. Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs. Comput. Vis. Image Understand. 176–177, 22–32. http://dx.doi.org/10.1016/j.cviu.2018.10.003.

Stafylakis, T., Tzimiropoulos, G., 2017. Combining residual networks with LSTMs for lipreading. In: Interspeech 2017. Presented at the Interspeech 2017. ISCA, pp. 3652–3656. http://dx.doi.org/10.21437/Interspeech.2017-85.

Stafylakis, T., Tzimiropoulos, G., 2018. Deep word embeddings for visual speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 4974–4978. http://dx.doi.org/10.1109/ICASSP.2018.8461347.

Sterpu, G., Saam, C., Harte, N., 2018. Can DNNs learn to lipread full sentences? In: 2018 25th IEEE International Conference on Image Processing (ICIP). Presented at the 2018 25th IEEE International Conference on Image Processing. ICIP, pp. 16–20. http://dx.doi.org/10.1109/ICIP.2018.8451388.

Su, Z., Fang, S., Rekimoto, J., 2022. LipLearner: Customizing silent speech commands from voice input using one-shot lipreading. In: Adjunct Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology. UIST '22 Adjunct, Association for Computing Machinery, New York, NY, USA, pp. 1–3. http://dx.doi.org/10.1145/3526114.3558737.

Su, Z., Fang, S., Rekimoto, J., 2023. LipLearner: Customizable silent speech interactions on mobile devices. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. pp. 1–21. http://dx.doi.org/10.1145/3544548.3581465.

Su, Z., Zhang, X., Kimura, N., Rekimoto, J., 2021. Gaze+lip: Rapid, precise and expressive interactions combining gaze input and silent speech commands for hands-free smart TV control. In: ACM Symposium on Eye Tracking Research and Applications. ETRA '21 Short Papers, Association for Computing Machinery, New York, NY, USA, pp. 1–6. http://dx.doi.org/10.1145/3448018.3458011.

Sun, K., Yu, C., Shi, W., Liu, L., Shi, Y., 2018. Lip-interact: Improving mobile device interaction with silent speech commands. In: Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology, UIST '18. Presented at the Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology. Association for Computing Machinery, New York, NY, USA, pp. 581–593. http://dx.doi.org/10.1145/3242587.3242599.

Takashima, Y., Takashima, R., Tsunoda, R., Aihara, R., Takiguchi, T., Ariki, Y., Motoyama, N., 2021. Unsupervised domain adaptation for lip reading based on cross-modal knowledge distillation. J Audio Speech Music Proc. 2021, 44. http://dx.doi.org/10.1186/s13636-021-00232-5.

Tan, J., Nguyen, C.-T., Wang, X., 2017. SilentTalk: Lip reading through ultrasonic sensing on mobile phones. In: IEEE INFOCOM 2017 - IEEE Conference on Computer Communications. Presented at the IEEE INFOCOM 2017 - IEEE Conference on Computer Communications. pp. 1–9. http://dx.doi.org/10.1109/INFOCOM.2017.8057099.

Tan, G., Wang, Y., Han, H., Cao, Y., Wu, F., Zha, Z.-J., 2022. Multi-grained spatio-temporal features perceived network for event-based lip-reading. In: Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20094–20103.

Tao, F., Busso, C., 2018. Gating neural network for large vocabulary audiovisual speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 26, 1290–1302. http://dx.doi.org/10.1109/TASLP.2018.2815268.

Thangthai, K., 2018. Computer Lipreading Via Hybrid Deep Neural Network Hidden Markov Models (Doctoral). University of East Anglia.

Thangthai, K., Bear, Y., Harvey, R., 2017. Comparing phonemes and visemes with DNN-based lipreading.

Tsourounis, D., Kastaniotis, D., Fotopoulos, S., 2021. Lip reading by alternating between spatiotemporal and spatial convolutions. J. Imaging 7 (91), http://dx.doi.org/10.3390/jimaging7050091.

Varshney, M., Yadav, R., Namboodiri, V.P., Hegde, R.M., 2022. Learning speaker-specific lip-to-speech generation. In: 2022 26th International Conference on Pattern Recognition (ICPR). Presented at the 2022 26th International Conference on Pattern Recognition. ICPR, pp. 491–498. http://dx.doi.org/10.1109/ICPR56361.2022.9956600.

Verkhodanova, V., Ronzhin, A., Kipyatkova, I., Ivanko, D., Karpov, A., Železný, M., 2016. HAVRUS corpus: High-speed recordings of audio-visual Russian speech. In: Ronzhin, Andrey, Potapova, R., Németh, G. (Eds.), Speech and Computer. In: Lecture Notes in Computer Science, Springer International Publishing, Cham, pp. 338–345. http://dx.doi.org/10.1007/978-3-319-43958-7_40.

Vorwerk, A., Wang, X., Kolossa, D., Zeiler, S., Orglmeister, R., 2010. WAPUSK20 - a database for robust audiovisual speech recognition. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Presented at the LREC 2010. European Language Resources Association (ELRA), Valletta, Malta.

Vougioukas, K., Ma, P., Petridis, S., Pantic, M., 2019a. Video-driven speech reconstruction using generative adversarial networks. In: Interspeech 2019. Presented at the Interspeech 2019. ISCA, pp. 4125–4129. http://dx.doi.org/10.21437/Interspeech.2019-1445.

Vougioukas, K., Petridis, S., Pantic, M., 2019b. End-to-end speech-driven realistic facial animation with temporal GANs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 37–40.

Vougioukas, K., Petridis, S., Pantic, M., 2021. DINO: A conditional energy-based GAN for domain translation. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May (2021) 3-7. Presented at the 9th International Conference on Learning Representations. ICLR 2021,Virtual Event, Austria, May (2021) 3-7, OpenReview.net.

Wan, L., Wang, Q., Papir, A., Moreno, I.L., 2018. Generalized end-to-end loss for speaker verification. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 4879–4883. http://dx.doi.org/10.1109/ICASSP.2018.8462665.

Wand, M., Schmidhuber, J., 2020. Fusion architectures for word-based audiovisual speech recognition. In: Interspeech 2020. Presented at the Interspeech 2020. ISCA, pp. 3491–3495. http://dx.doi.org/10.21437/Interspeech.2020-2117.

Wang, C., 2019. Multi-grained spatio-temporal modeling for lip-reading. In: Sidorov, K., Hicks, Y. (Eds.), Proceedings of the British Machine Vision Conference. BMVC, BMVA Press, pp. 225.1–225.11. http://dx.doi.org/10.5244/C.33.225.

Wang, S.-L., Liew, A.W.-C., 2012. Physiological and behavioral lip biometrics: A comprehensive study of their discriminative power. In: Pattern Recognition, Best Papers of Iberian Conference on Pattern Recognition and Image Analysis, Vol. 45. IbPRIA'2011, pp. 3328–3335. http://dx.doi.org/10.1016/j.patcog.2012.02.016.

Wang, H., Pu, G., Chen, T., 2022. A lip reading method based on 3D convolutional vision transformer. IEEE Access 10, 77205–77212. http://dx.doi.org/10.1109/ACCESS.2022.3193231.

Wang, Y., Zhao, Z., 2022. FastLTS: Non-autoregressive end-to-end unconstrained lip-to-speech synthesis. In: Proceedings of the 30th ACM International Conference on Multimedia. MM '22, Association for Computing Machinery, New York, NY, USA, pp. 5678–5687. http://dx.doi.org/10.1145/3503161.3548194.

Weng, X., Kitani, K., 2019. Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading. In: Sidorov, K., Hicks, Y. (Eds.), Proceedings of the British Machine Vision Conference. BMVC, BMVA Press, pp. 2.1–2.13. http://dx.doi.org/10.5244/C.33.2.

Wong, Y.W., Ch'ng, S.I., Seng, K.P., Ang, L.-M., Chin, S.W., Chew, W.J., Lim, K.H., 2011. A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities. Pattern Recognit. Lett. 32, 1503–1510. http://dx.doi.org/10.1016/j.patrec.2011.06.011.

Wright, C., Stewart, D., 2019. One-shot-learning for visual lip-based biometric authentication. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Ushizima, D., Chai, S., Sueda, S., Lin, X., Lu, A., Thalmann, D., Wang, C., Xu, P. (Eds.), Advances in Visual Computing. In: Lecture Notes in Computer Science., Springer International Publishing, Cham, pp. 405–417. http://dx.doi.org/10.1007/978-3-030-33720-9_31.

Wright, C., Stewart, D.W., 2020. Understanding visual lip-based biometric authentication for mobile devices. EURASIP J. Inform. Secur. 2020, 3. http://dx.doi.org/10.1186/s13635-020-0102-6.

Xiao, J., 2019. 3D feature pyramid attention module for robust visual speech recognition. arXiv:1810.06178 [Cs].

Xiao, J., Yang, S., Zhang, Y., Shan, S., Chen, X., 2020. Deformation flow based two-stream network for lip reading. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). Presented at the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. FG 2020, pp. 364–370. http://dx.doi.org/10.1109/FG47880.2020.00132.

Yadav, R., Sardana, A., Namboodiri, V.P., Hegde, R.M., 2021. Speech prediction in silent videos using variational autoencoders. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 7048–7052. http://dx.doi.org/10.1109/ICASSP39728.2021.9414040.

Yang, C.-C., Fan, W.-C., Yang, C.-F., Wang, Y.-C.F., 2022. Cross-modal mutual learning for audio-visual speech recognition and manipulation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36. pp. 3036–3044. http://dx.doi.org/10.1609/aaai.v36i3.20210.

Yang, Wenfeng, Li, P., Yang, Wei, Liu, Y., He, Y., Petrosian, O., Davydenko, A., 2023. Research on robust audio-visual speech recognition algorithms. Mathematics 11, 1733.

Yang, C.-Z., Ma, J., Wang, S., Liew, A.W.-C., 2020. Preventing DeepFake attacks on speaker authentication by dynamic lip movement analysis. IEEE Trans. Inf. Forensics Secur. 16, 1841–1854. http://dx.doi.org/10.1109/TIFS.2020.3045937.

Yang, X., Ramesh, P., Chitta, R., Madhvanath, S., Bernal, E.A., Luo, J., 2017. Deep multimodal representation learning from temporal data. In: Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE Computer Society, pp. 5066–5074. http://dx.doi.org/10.1109/CVPR.2017.538.

Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., Long, K., Shan, S., Chen, X., 2019. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition. FG 2019, IEEE Press, Lille, France, pp. 1–8. http://dx.doi.org/10.1109/FG.2019.8756582.

Yao, Y., Wang, T., Du, H., Zheng, L., Gedeon, T., 2019. Spotting visual keywords from temporal sliding windows. In: 2019 International Conference on Multimodal Interaction. ICMI '19, Association for Computing Machinery, New York, NY, USA, pp. 536–539. http://dx.doi.org/10.1145/3340555.3356101.

Yu, R., 2022. Computer-aided english pronunciation accuracy detection based on lip action recognition algorithm. In: 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). Presented at the 2022 Second International Conference on Artificial Intelligence and Smart Energy. ICAIS, pp. 1009–1012. http://dx.doi.org/10.1109/ICAIS53314.2022.9743068.

Yu, J., Zhang, S.-X., Wu, J., Ghorbani, S., Wu, B., Kang, S., Liu, S., Liu, X., Meng, H., Yu, D., 2020. Audio-visual recognition of overlapped speech for the LRS2 dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, pp. 6984–6988. http://dx.doi.org/10.1109/ICASSP40776.2020.9054127.

Zakeri, A., Hassanpour, H., 2021. WhisperNet: Deep siamese network for emotion and speech tempo invariant visual-only lip-based biometric. In: 2021 7th International Conference on Signal Processing and Intelligent Systems (ICSPIS). Presented at the 2021 7th International Conference on Signal Processing and Intelligent Systems. ICSPIS, pp. 1–5. http://dx.doi.org/10.1109/ICSPIS54653.2021.9729394.

Zeng, R., Xiong, S., 2022. Lip to speech synthesis based on speaker characteristics feature fusion. In: Proceedings of the 4th International Conference on Information Technology and Computer Communications. ITCC '22, Association for Computing Machinery, New York, NY, USA, pp. 78–83. http://dx.doi.org/10.1145/3548636.3548648.

Zhang, X., Cheng, F., Shilin, W., 2019. Spatio-temporal fusion based convolutional sequence learning for lip reading. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision. ICCV, pp. 713–722. http://dx.doi.org/10.1109/ICCV.2019.00080.

Zhang, S., Ma, Z., Lu, K., Liu, X., Liu, J., Guo, S., Zomaya, A.Y., Zhang, J., Wang, J., 2022a. HearMe: Accurate and real-time lip reading based on commercial RFID devices. IEEE Trans. Mob. Comput. 1–14.

Zhang, J., Roussel, P., Denby, B., 2021a. Creating song from lip and tongue videos with a convolutional vocoder. IEEE Access 9, 13076–13082. http://dx.doi.org/10.1109/ACCESS.2021.3050843.

Zhang, X., Sheng, C., Liu, L., 2021b. Lip motion magnification network for lip reading. In: 2021 7th International Conference on Big Data and Information Analytics (BigDIA). Presented at the 2021 7th International Conference on Big Data and Information Analytics. BigDIA, pp. 274–279. http://dx.doi.org/10.1109/BigDIA53151.2021.9619626.

Zhang, Y., Yang, S., Xiao, J., Shan, S., Chen, X., 2020. Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). Presented at the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. FG 2020, pp. 356–363. http://dx.doi.org/10.1109/FG47880.2020.00134.

Zhang, X., Zhang, C., Sui, J., Sheng, C., Deng, W., Liu, L., 2022b. Boosting lip reading with a multi-view fusion network. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). Presented at the 2022 IEEE International Conference on Multimedia and Expo. ICME, pp. 1–6. http://dx.doi.org/10.1109/ICME52920.2022.9859810.

Zhao, Y., Ma, C., Feng, Z., Song, M., 2021a. Speech guided disentangled visual representation learning for lip reading. In: Proceedings of the 2021 International Conference on Multimodal Interaction. ICMI '21, Association for Computing Machinery, New York, NY, USA, pp. 687–691. http://dx.doi.org/10.1145/3462244.3479952.

Zhao, H., Zhang, B., Yin, Z., 2021b. Lip-corrector: Application of BERT-based model in sentence-level lipreading. J. Phys.: Conf. Ser. 1871, 012146. http://dx.doi.org/10.1088/1742-6596/1871/1/012146.

Zhou, M., Wang, Q., Li, Q., Jiang, P., Yang, J., Shen, C., Wang, C., Ding, S., 2021. Securing face liveness detection using unforgeable lip motion patterns. arXiv e-prints.

Zimmermann, M., 2018. Visual Speech Recognition: From Traditional to Deep Learning Frameworks (Doctoral Dissertation). École Polytechnique Fédérale de Lausanne.