Sumer Journal for Pure Science (ISSN: 2790-7031)
SJPS 2023 (December), Vol. 2, No. 2, p.p.: 68-87

A Survey on Speech Recognition from Lip Movement

Ahmed S. Ketab[1,2]*, Nidhal K. El abbadi[3]

[1] *Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Najaf, Iraq*

[2]*Directorate of Education of Thi-Qar, Department of Vocational Education, Thi-Qar, Iraq*

[3]*Computer Engineering Techniques, college of Engineering Techniques, Al-Mustaqbal University, Babylon, Iraq*

Abstract

Speech recognition from lip movement is fascinating and challenging research area that has garnered substantial interest in recent times. The ability to recognize speech from lip movement has many potential applications, such as aiding those with hearing impairments and enhancing speech recognition in video conferencing systems and / or in noisy environments. The purpose of this survey is to present a comprehensive summary of the latest cutting-edge techniques of speech recognition from lip movement. It will cover various approaches used in this field, including deep learning-based methods, traditional machine-learning techniques, and hybrid approaches. The paper will also explore the existing datasets, which can be valuable resources for researchers in their future work. Overall, the survey can be beneficial for researchers who are interested in the field of speech recognition using lip movement.

Keywords: Lip-reading, Visual Speech Recognition, Lip movement, deep learning.

استطلاع حول تعرف الكلام من حركة الشفاه

أحمد سعود كتاب[1, 2]*, نضال خضير العبادي[3]

[1]قسم علوم الحاسبات, كلية علوم الحاسوب والرياضيات, جامعة الكوفة, النجف, العراق

[2] قسم التعليم المهني, مديرية تربية ذي قار, ذي قار, العراق

[3] تقنيات هندسه الحاسوب, كلية الهندسة والتقنيات, جامعة المستقبل, بابل, العراق

الخلاصة

يعد التعرف على الكلام من خلال حركة الشفاه مجالًا بحثيًا رائعًا ومليئًا بالتحديات وقد حظي باهتمام كبير في الآونة الأخيرة. إن القدرة على التعرف على الكلام من خلال حركة الشفاه لها العديد من التطبيقات المحتملة، مثل مساعدة الأشخاص الذين يعانون من ضعف السمع وتعزيز التعرف على الكلام في أنظمة مؤتمرات الفيديو و/أو في البيئات الصاخبة. الغرض من هذا الاستطلاع هو تقديم ملخص شامل لأحدث التقنيات المتطورة للتعرف على الكلام من خلال حركة الشفاه. وسيغطي مختلف الأساليب المستخدمة في هذا المجال، بما في ذلك الأساليب القائمة على التعلم العميق، وتقنيات التعلم الآلي التقليدية، والأساليب الهجينة. وستستكشف الورقة أيضًا مجموعات البيانات الموجودة، والتي يمكن أن تكون موارد قيمة للباحثين في عملهم المستقبلي. بشكل عام، يمكن أن يكون الاستطلاع مفيدًا للباحثين المهتمين بمجال التعرف على الكلام باستخدام حركة الشفاه.

* ahmedsaudketab@utq.edu.iq

1. Introduction

Communication often involves the use of voice and vision, with speech signals being more information-rich than visual signals. As a result, researches have primarily concentrated on Automatic Speech Recognition (ASR), which has achieved high recognition rates and is widely used in various fields. However, Visual Speech Recognition (VSR), that focuses on recognizing the speech depend on shape of lips only without any speech signals, has emerged as a promising technology. The VSR faces some challenging problems due to the 2D nature of visual information, which contains higher degree of redundant information compared to one-dimensional voice information [1].

Despite some promising advancements, VSR still lags behind its counterpart, ASR. The later approach, which aims to decode spoken text, can be considered as a heterogeneous modality that shares the same underlying distribution as lip reading. Despite having the same training data and model architecture, there is a noticeable performance gap between the two techniques. For instance, the character error rates for speech recognition and lip reading are 10.4% and 39.5% respectively, highlighting the challenges associated with lip reading in achieving comparable accuracy to traditional speech recognition[2]. The observed performance difference can be primarily attributed to the inherent ambiguity of lip movements. Multiple lip movements can appear similar but correspond to different words, making it challenging to extract distinct features from the video and accurately predict the corresponding text output. This ambiguity poses a significant hurdle in achieving high accuracy in lip reading tasks [3].

In the 1950s, Sumby puts forward the notation that observing lip movements during speech could serve as a means to gather information, thereby introducing the concept of lip reading and igniting fresh exploration in the field. This marked the beginning of early research on lip reading. Today, lip-reading systems are divided into traditional and deep learning-based methods. Traditional lip reading involves feature extraction of lips using a pixel based approach, such as Discrete Cosine Transform (DCT) and Principal Component Analysis(PCA), and lip movement recognition using Hidden Markov Models (HMMs). HMMs represent lip movements as linear parametric models and combine them in series to form a Markov chain, allowing for maximum probability recognition results [4]. Figure 1 summarizes the traditional process, which involves locating and extracting the lip, followed by extracting significant features of the lip images. Subsequently, methods of feature transformation may employ to decrease the size of obtained features. Finally, a classifier is ultimately utilized to determine the class of them.

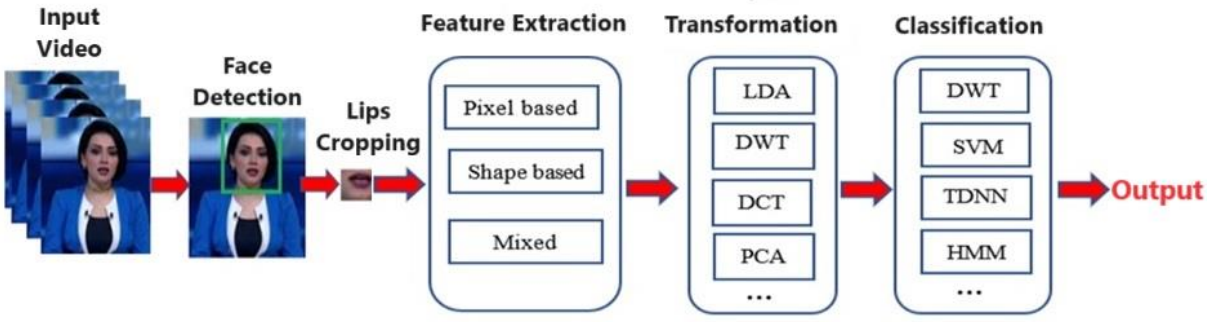

Figure -1 Traditional Lipreading process.

The poor performance of human lipreading is a significant concern, with hearing-impaired individuals achieving approximately 17±12% accuracy for a restricted set comprising 30 single words and 21±11% accuracy for 30 phrases. So automating reading of lips has become a crucial objective. Machine lipreaders hold significant practical potential and can be applied in various domains. Some potential applications include enhancing hearing aids, enhancing security measures, facilitating speech recognition in environments that are noisy, enabling identification by biometric through lip movements, and aiding in the processing of silent movies. However, Machine lipreading is a complex task that entails extraction of spatiotemporal features from video data, as both the motion and position of the lips are crucial for accurate interpretation. Recent approaches of deep learning aim to extract these spatiotemporal features in an end-to-end method. However, the majority of existing research in this field primarily focuses on word classification tasks rather than sentence [5][6]. Figure 2 Summarize the deep learning process. In the first, the lip region determined and extracted from the video. Subsequently, front-end network is utilized for extracting of spatial and temporal features, which are further utilized as input of the back-end. Finally, these features are concatenated and used for classification.

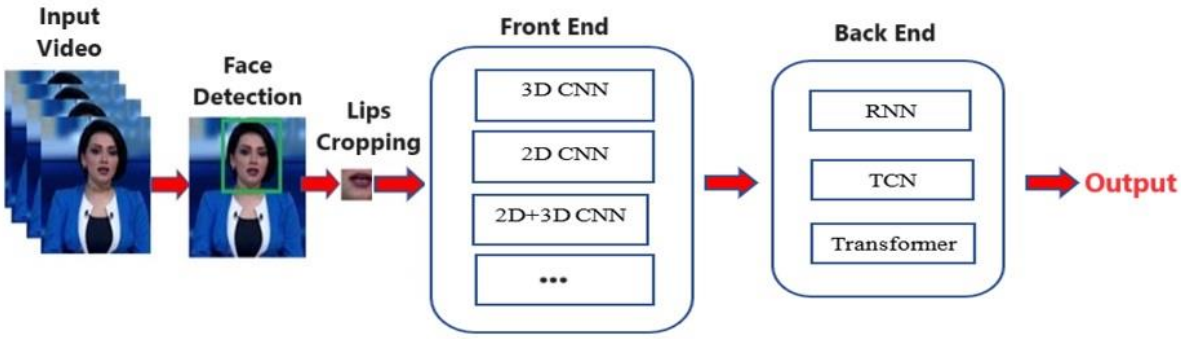

Figure -2 VSR by Deep Learning.

The survey is structured into different sections. In Section Two, the focus is on the challenges and problems related to lip reading. Section Three introduces the application of the lip reading process. A summary of related works is provided in Section Four. Section Five presents the general steps involved in lip reading. Section Six provides a comprehensive discussion on the commonly used datasets. Section Seven provides a comparison of various techniques used in lip reading. A comparative analysis of various techniques employed in the lip reading devoted to section seven. Finally, Section Eight presents the conclusion.

2. Lip Reading Problems and Challenges

Lip reading or Visual Speech Recognition faces several problems and challenges that affect its accuracy and effectiveness. Firstly, traditional ASR systems can be unreliable in environments having unrelated sound commonly referred to noise, leading to misunderstandings due to extraneous sounds [7]. Secondly, millions of individuals with speech-related medical conditions, such as Vocal Cord Paralysis, Spasmodic Dysphonia, Dysarthria, and Laryngeal Cancer, face challenges in communicating effectively[8]. Laryngeal cancer alone, which is treated by Total laryngectomy, causes approximately 800,000 deaths

[9]. Thirdly, hearing loss is a major problem, particularly among the elderly population, affecting over 5% of people worldwide according to the World Health Organization [8].

Also, lip reading poses several challenges that make it a difficult process for both machines and humans. The most significant challenges include the similarity between some phonemes or visemes, such as "p" and "b" [10][2], variations in speakers' mouth shape, mustache, makeup, skin color, and age [11], differences in accents, languages, and speaking speeds [12], and variations in facial direction, lighting conditions, poor temporal resolution, and intensities [13].

3. Applications of VSR

VSR has very wide applications. In noisy environments, it aids ASR for recognizing speech in multi-talker environments [8] or Target Speech Separation [14]. This is known as audiovisual speech recognition (AVSR).

It is also used to give instructions [15]and User Authentication[16] to systems or machines like smartphones. VSR has important medical applications. It helps to understand speech-impaired individuals (ex: Lou Gehrig's disease speak) [17] and patients with diseases who lose the ability to speak (laryngeal cancer, vocal cord paralysis) [11]. Additionally, it can serve as a complement tool in virtual environments, enhancing the immersive experience and interaction with virtual characters or avatars [18]. Biometric authentication [19], analysis of silent movies [20], analysis of facial expressions [21], synthesizing talking faces, speaker recognition [22], forensic video analysis [23]and deep fake detection [24] are other important applications.

4. Literature Review

In this part, a review has been conducted on last research related to the topic of VSR. In 2018, Yuto Koguchi et al. [15] had to use lip reading for input instruction in smartphones instead of the touch input technique. Here, first, the camera of the smartphone captures the face and then sends the facial image to the server. The server that utilizes CNN as feature extraction is responsible to recognize the Japanese vowel. Finally, the server returns the recognized vowel to the smartphone. In [25] introduced a segmentation method to determine the region of lips and then extract visual features. The data set collected by themselves contains 26 English letters from different people. In the classification, had used NVG-RAM and KNN separately. The recognition performance when using NVG-RAM was 94.679% and for KNN was 92.628%. Triantafyllos Afouras, et al. [26] introduced three architectures: first a model using LSTM; second a fully convolution model; third the recently proposed transformer model. The models were tested on BBC-Oxford Lip Reading Sentences2 (LRS2) dataset. The researchers show when the models are applied for online lip reading of continuous speech, the fully convolutional model achieves high performance with low latency. In [10] a deep neural network for lipreading called LCANet was introduced. LCANet encodes the video by a stacked 3D-CNN, highway network, and big network. both long-term and short-term spatiotemporal information used for the encoder. A cascaded attention CTC is utilized to extract the result text at decoder part. The experiments show that LCANet obtains a 3.0% WER and 1.3% CER on the GRID

dataset and achieves a 12.3% enhancement compared to other methods. What distinguishes [51] [27] is that they provided databases with three speech modes: normal and whispered and silent. The results prove that training on one speech mode and testing on another one, leads to a decrease in classification accuracy. In [52] [28], a deep learning technique called DenseNets was proposed for extracting visual representations from RGB image. Additionally, physiologic feature of lips , which captures the position and structure of muscles of face involved in speech production, was extracted to represent the similarity of individuals' speech patterns. These features were combined in the final in a fully-connected layer of the DenseNets architecture. According to experimental findings, DenseNets are capable of handling the spatial-temporal information in the entire image sequence and integrating the proposed 3D geometric-physiological feature led to an improvement up to 3.91%.

In 2019, Yuanyao Lu and Hongbo Li [29] applied CNN in the feature extraction stage and RNN depend on the attention in the recognition stage in their lip-reading system. This system was tested on a database created by themselves and containing numbers from zero to nine. The highest accuracy obtained was 88.2%. The lipreading system suggested in [30] is called LipCH-Net. It is for the Chinese language and works on unconstrained sentences by two DNNs. They collected the dataset from China Central Television (CCTV) website by themselves. The results show that this system has reduced overfitting and accelerated training and overcome the syntactic ambiguity. Chenhao Wang [20] introduces a lip-reading model using the multi-grained spatiotemporal model. first, it extracts fine-grained features and then short-term medium-grained features. Next, a BiConvLSTM model enhanced with temporal attention, is employed to gather and integrate spatiotemporal information from the entire input sequence. The researcher shows the effectiveness of his method by testing on two datasets LRW-1000 and LRW. In [11] proposed a CNN architecture named HCNN that utilizes Hahn moments. The HCNN tested on OuluVS2, AVLetter, and LRW datasets and the results show significant progress in comparison with other methods.

In 2020, in addition to a survey, Souheil Fenghour et al. [31], developed a mechanism for reading lips use neural networks. The system is compared with other works and achieves an increased performance with a 15% decreased word error rate using LRS2 dataset. In [32], Viola -Jons algorithm was executed for face detection, and the mouth area was isolated based on a proposed skin color segmentation. Then a collection of visual shape feathers (eight pixels) that pointed to the edge of the lips was extracted. Finally, a neural network is used for classification. These steps are implanted in with hardware based on FPGA to distinguish twenty-six English letters. The dualLip system was developed by Weicong Chen et al. [33] it includes two models: 1) the lip generation model used to generate lip video form text, 2) the lip-reading model used to recognize the speech. For the experiment, they used two datasets, TCD-TIMIT, and GRID that prove the effectiveness of the system. In [2] propose a method, called Lip by Speech (LIBS), that aims to increase the efficiency of the system by using learning from speech recognizers. The most important techniques used are CNN and RNN. The method gets cutting-edge performance on LRS2 and CMLR datasets. A study to diagnose utterances was performed in [34]. Luminance values, lip width, and height of the mouth were utilized as lip movement feature for vowel identification. The experiment illustrated that the luminance values and the lip height are helpful for the identification of the vowels “i” and “u”. The proposed system in

[19] utilizes local and global features to discover the relation in speech and increases the model recognition ability. The system evaluated LRW-1000 and LRW datasets. The results proved the effectiveness of the method on both datasets. A refined network is a DNN developed by [35] for lip reading. The experiments show that the refined network obtain a clear improvement from 38.2% to 55.7% on the LRW-1000 dataset and from 83.7% to 88.4% on the LRW dataset. In [36] a model to distinguish vowel phonemes of Indonesian from lip movement was proposed. The model used 3D CNNs. The highest accuracy rate reached was 84%. Adriana Kurniawan and Suyanto Suyanto [37] proposed a model based on syllables that allow the creation of a newly coined term that doesn't appear in their dictionary. The model is based on a 3D Deep NN and applied in the Indonesian language. The results of the experiments gave an accuracy of 80%.

In 2021, a lip-reading system that uses a learnable module called ALSOS proposed by [38]. The ALSOS includes spatiotemporal 3D and spatial 2D convolutions and two components for conversion (2D-to-3D and 3D-to-2D). The performance of the system was evaluated through multiple ALSOS and ResNet pairings for Greek language and LRW-500 datasets. The results demonstrated that the integration of ALSOS improved the performance of the system. In [39] presents a dataset for lipreading in the Russian language (LRWR). They also introduce a comprehensive comparison of some lip-reading systems on LRWR. Zhijie Lin et al. [40] devised SimulLR transducer that uses attention-guided adaptive memory. The SimlLR also uses different effective training methods like CTC pre-training model and curriculum learning that is designed to motivate lip-reading transducer training. The results of the trails demonstrate that the proposed method's effect is achieved via SimulLR. Pingchuan Ma et al. [41] utilize DC-TCN, Squeeze-and-Excitation blocks, and lightweight attention for lip reading. Their method obtained 88.36% accuracy by using LRW-1000 dataset. The Hybrid lip-reading (HLR-Net) which is a deep CNN model is developed by [18]. It contains a decoder and encoder as main stages. In the decoder, gradient, inception, and BiGRU layers are utilized while in the decoder the fully connected, attention, and activation function layers are utilized. In comparison with two other models, LCANet and A-ACA, the HLR-Net model achieved significant improvements on the GRID corpus dataset. In [42], the focus was on the Turkish language, and for this purpose, due to the lack of Turkish databases, the researcher created two databases, one for words containing 111 words and the other for sentences containing 113 Turkish phrases. The model employs CNN and BI-LSTM and got an accuracy of 84.5% in words and 88.55% in sentences. Dweik et al. [8] propose a system of lipreading able to distinguish ten Arabic words. In this system after preprocessing of input video, converts it to grayscale image frames. Now, the two image frames, RGB and grayscale frames enter CNN, TD-CNN-BiLSTM and TD CNN-LSTM severally. So, the system gets six results. Finally, a voting model that combines these six results to give a final classification is used. The highest testing accuracy that was evaluated on a locally collected Arabic dataset was 82.84%.

In 2022, K R Prajwal et al. [17] introduce a pooling mechanism based on attention to collect visual speech representations. Additionally, they introduced a visual speech detection model that was trained from a lip reading network. This model achieves a 22.6% word error rate using the LRS2. The proposed system in [22] includes a 3D convolutional vision transformer (3DCvT) in the front end and Bidirectional Gated Recurrent Unit (BiGRU) in the back end.

The test was conducted on LRW-1000 and LRW datasets. The best accuracy they got was 88.5%.

5. The General Steps of Lip Reading Process

Lip-reading system is complex and involves multiple steps and techniques, each of which plays a crucial role in achieving accurate and reliable results [24]. Figure 3 Summarize the general steps for the lip-reading system.

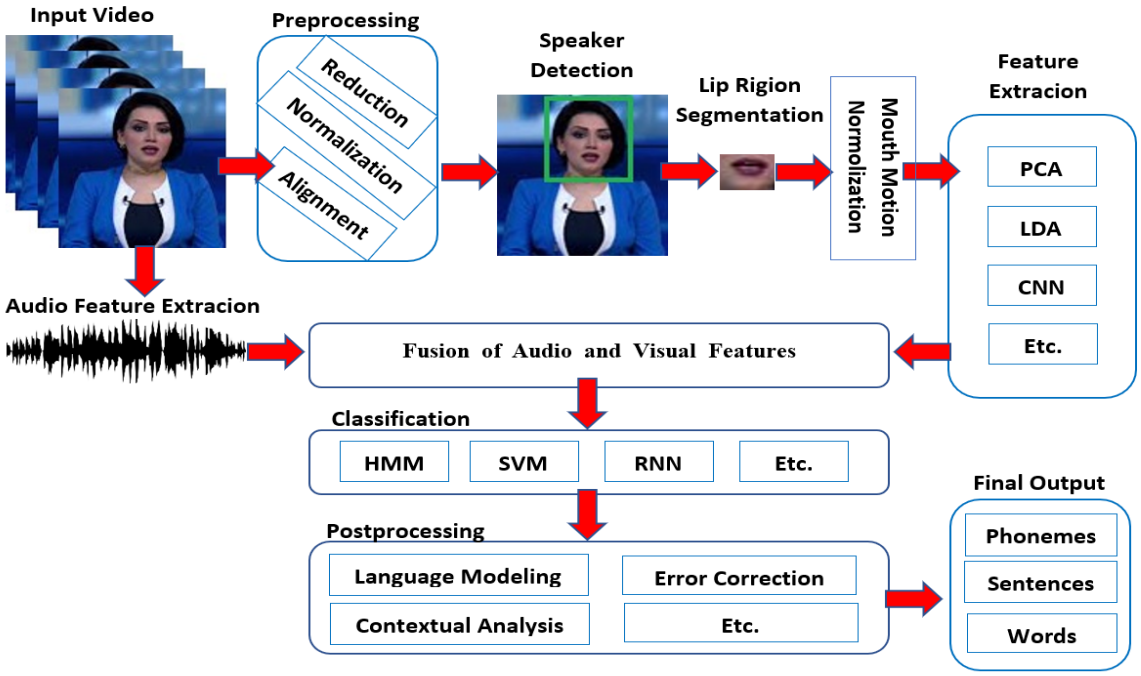

Figure -3 Main Steps of lip-reading system.

5.1 Feature Extraction Techniques

After detecting and extracting the lip region, the next crucial step is the process of taking features that are helpful for classification. As lip image sequences typically contain a significant amount of irrelevant information, it is essential to identify and extract features that are specifically related to lip reading. Therefore, the feature extraction stage has a major role in the success of the overall visual speech recognition system. There exist numerous techniques for extracting lip features, that broadly fall into three categories include: pixel-based approaches, shape-based approaches, and Artificial Neural network-based approaches. Note that there is a method that uses a combination of these approaches.

5.1.1 Pixel-Based Approach

In this approach, the feature space consists of all values of pixels within the lip region, and various methods are employed to decrease the feature space dimensionality in order to obtain the most significant features. Commonly, a linear transformation such as PCA [43], DCT [44],

Linear Discriminant Analysis (LDA) [45], and Discrete Wavelet Transform (DWT) [46] is employed to decrease the feature dimension.

Optical flow which is a computer vision method for estimating the motion of an object within a video, also can be used as a feature extraction method to capture the subtle movements of the lips as they form different phonemes. By analyzing the changes in the optical flow between frames, it is possible to extract features that are highly relevant to lip reading, such as the velocity and acceleration of the lip movements. Optical flow has been employed as a feature for lipreading [47]. The local pixel feature method is a commonly used approach in lip reading systems. It involves extracting lip features based on the intensity values within the region of interest (ROI). The ROI is typically defined as lips area. This method can extract a vast number of features and can be employed for the train Machine Learning (ML) models. In [48] introduced LBP-TOP [49] for extracting spatiotemporal information, as Local Binary Patterns (LBP) is not suitable for processing a single image while the lipreading task requires an image sequence input.

5.1.2 Shape-Based Approach

This approach involves constructing a model or pattern depending on the shape of the lips with some parameters that represents the visual features and can be separated into two groups: geometric and contour features. Geometric features typically include width, perimeter, area, height and etc. In [50], seven high-level geometric features were defined for visual speech recognition. These features are calculated by measuring the Euclidean distances and areas between specific key points. These key points include height and width of mouth, width and height of mouth aperture, mouth area, aperture area, and the distance from chin to nose as Figure 5.

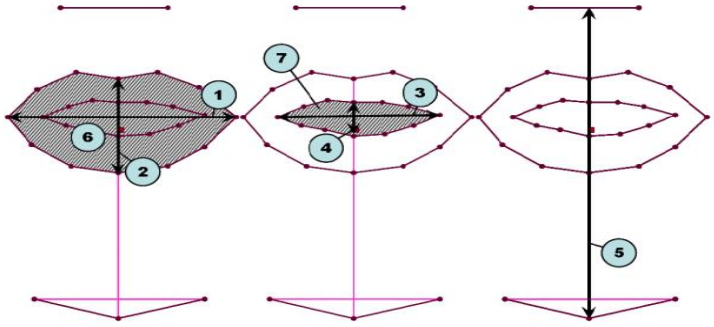

Figure -5 The High-Level Geometric Features [50].

Article [51] extracts three features of lips after determining their locations. These features consist of one measurement of lip width and two measurements of height. Figure 6 shows these features.

Figure -6 Geometric Features in [51].

The contour feature is an algorithm (ACM) that works by extracting key features along with the lip edge. then it converts their coordinates into a vector, which represents the lips. It's known as the Snake model [52]. The STASM library is utilized to employ the Active Shape Model (ASM) for locating landmarks on the lips. After locating the landmarks, the landmarks' coordinates of both the inside and outside of lips are concatenated to create the feature vector. Additionally, the feature vector expanded to include the lips' height and width [53]. Figure 7 illustrates these features.

Figure -6 (a) Landmarks of the Lips (b) Width and Height as Lips Feature [53].

5.1.3 Artificial Neural Network Approach

This approach primarily utilizes different types of Deep Neural Networks(DNN) to extract features of lip movement from a sequence of images. it is also known as a front-end network and may have various deep neural network structures. Feedforward Neural Networks [54], Autoencoders [55], and Boltzmann Machines [56] are some examples of front-end networks. However, after the emergence of CNNs which revolutionized the field of feature extraction, has become the most commonly used and effective network architecture for feature extraction. At the outset, different CNNs were introduced. Some examples of CNNs are AlexNet, LeNet, and VGGNet [57] Subsequently, a 3D CNN [58] to handle the temporal information of videos was proposed.

5.2 Classification Techniques

After extracting the lip features, the classification stage begins. The accuracy of the classification is directly affects on the overall performance of the systems. Therefore, developing efficient feature extraction and classification methods is essential to improve the accuracy and efficiency of lip-reading systems. Many classification methods can be categorized into three main approaches.

5.2.1 Machine Learning Approach

A type of artificial intelligence known as machine learning techniques enables computers to automatically learn from experience and get better over time without explicit programming. Instead of relying on explicit rules or programming, machine learning algorithms use statistical and mathematical models to analyze and learn from data. The Support Vector Machine (SVM) [59], K Nearest Neighbors (K-NN) [60], Multi-Layer Perceptron (MLP) [61] and etc used in different research.

5.2.2 Deep Learning (DL) Approach

DL is a branch of ML that leverages ANNs to learn and make predictions from vast volumes of data. It extracts complex features from raw data automatically and has demonstrated remarkable success across various domains. DL models consists of multiple layers of interconnected nodes, allowing them to learn hierarchical representations of the input data. CNNs, RNNs and their variations, including Gated Recurrent Units(GRU) and Long-Short-Term Memorys (LSTM) networks, are among the most commonly utilized deep learning models. There are also bidirectional versions of LSTM, RNN, and GRU. The success of DL is largely attributed to the existence of large datasets and the advancement of hardware technologies such as CPUs and GPUs, which have enabled faster training of deep models.

5.2.3 Other Approach

In addition to deep learning and machine learning methods, there're other approaches used in data analysis and modelling. The majority of the techniques employed in earlier research are Hidden Markov models (HMM) [62], Template Matching and etc. HMM are statistical models which are utilized to model time series data, particularly sequential data, where each observation is dependent on the previous ones. It is a type of generative probabilistic model that can be utilized to model observational sequences.

Template matching is a technique that involves finding the best match between a template image and a larger image. The template image is a small portion of an image that is known to contain the wanted object, while the larger image is the scene in which the object is located. The purpose of template matching is to identify the location of the template image within the larger image. In Petajan's research [63], the dynamic feature vectors of the pronunciation are saved in the dataset during the training. During the recognition stage, the feature vector is compared with the templates saved in the dataset. Finally, the recognition result is determined by selecting the template with the highest correlation coefficient.

6. Datasets

To train and test VSR systems, large amounts of annotated data are required, which can be challenging to obtain. The creation and utilization of VSR datasets, which contain recorded videos of people speaking and corresponding transcriptions or annotations, have seen an increase in interest in recent years. These datasets are essential to improve and assess of VSR algorithms, as enabling the researchers to train and test their models and then compare their performance with that of other models. The existence of a good ready dataset that meets the search requirements greatly reduces effort and time. The most popular datasets utilized by the researchers for lipreading are mentioned in Table I. of this section. The properties mentioned are the name of the database, the language used, the year of creation, The context of speech continuous or separate words (isolated), segments (alphabets, digits, phrases, words or sentences), number of speakers, number of classes, utterance, photography angle, resolution and frame rate of the video.

Table 1- The Summary of Used Datasets

Dataset name	Language	Year	Isolated/ Continuous	Segment	Speaker	Classes	Frame Rate	Pose
IBMViaVoice [64]	English	2000	Isolated	Sentence	290	10500	30	Frontal
UNMC VIER [65]		2002		Sentence	123	12	29	0, 90
BANCA [66]	Multiple	2003		Digit	209	10	25	frontal
AVICAR [65]	English	2004	Continuous	Alphabet	86	26	30	4 views
				Digit	86	10		
			Isolated	Sentence	86	20		
VALID [67]		2005		Digit	106	10	25	frontal
GRID [68]		2006		Phrase	34	34000	30	
GMU AVPFV [51]		2007		Word	10	150	50	0,90
IV2 [69]	French	2008		Sentence	300	15	<50	
Oulu VS2 [70]	English	2010	Continuous	Digit	53	10	30	0,30,45,60,90
			Isolated	Phrase	53	10		
				Sentence	53	540		
TCD TIMIT [71]		2011		Sentence	20	62		0,30
AGH AV [72]	Polish	2012		Digit	20	10	50	frontal
AVAS [73]	Arabic	2013		Digit	50	10	30	-90,-45,0,45,90
				Word	50	24		
				Phrase	50	13		
LSVSR [74]	English	2014		Sentence	>1000	127055		-30,30
HAVRUS [75]	Russian	2016		Sentence	20	1530	200	
LRS2 [3]	English	2017		Sentence	>1000	17428	25	-30,30
LRS3 [76]		2018		Sentence	>1000	70000		-90,90
AVSD [77]	Arabic	2019		Phrase	22	10	30	frontal
NSTDB [78]	Chinese	2020	Continues	Word	N/A	349	25	-90,90
LRWR [39]	Russian	2021	Isolated	Word	153	235	N/A	0 to 20
Ümit Atila et al [42]	Turkish	2022		Word	24	111	30	frontal
				Sentence	24	113	60	
Waleed et.al [8]	Arabic			Word	73	10	30	

There are many differences between datasets used for VSR according to the resources mentioned in this table, including:

- Language: Some datasets may contain speech and mouth movements in a particular language or dialect, while others may include multiple languages or accents.
- The number of speakers: Some datasets may have a larger number of speakers than others, which can impact the diversity of speech and mouth movements captured.
- Recording conditions: Differences in recording equipment, lighting, background noise, and other factors can affect the quality and consistency of the data.
- Speaking style: Some datasets may include scripted speech, while others may contain spontaneous speech, which can impact the variability of mouth movements.
- Annotations: Datasets may have different types and levels of annotations, such as phonetic transcriptions or labels indicating the location of the mouth in each frame.

These differences can affect the performance of VSR models trained on different datasets, as well as the generalizability of these models to real-world scenarios.

7. Comparing the Popular Techniques

As we mentioned previously there are many techniques proposed for lip reading. Table 2 contains the performance of many popular methods, the names of the researchers, techniques of feature extraction and classification used, Language of speakers, what level (alphabets, digits, phrases, words or sentences), the dataset used, the year of publication, for comparing.

Table 1- The Summary of Used Techniques

Researchers	year	Level	Data set	techniques	performance
Kumar et al. [51]	2007	Words	CMU AVPFV	Geometry of Mouth + HMM	32.39%
Lucey et al. [79]	2007	Digits	IBMSR	DCT + LDA + HMM	68.58%
Lucey et al. [80]	2008	Digits	IBMSR	DCT + PCA + HMM	66.21%
Papandreou et al. [81]	2008	Digits	CUAVE	AAM + HMM	75.7%
Shao et al. [82]	2008	Phrases	GRID	DCT + HMM	58.4%
Hilder et al. [83]	2009	Alphabet	AVLetters2	AAM + HMM	75.24%
Zhao et al. [84]	2009	Alphabet	AVLetters	LBP-TOP + SVM	62.8%
Pass et al. [85]	2010	Digits	QuLips	DCT + HMM	98%
Saitoh et al. [86]	2010	Words	Own data	L2 between key points + HMM	68.93%
Cappelletta et al. [87]	2011	Sentence	VIDTIMIT	Optical flow + HMM	57%
	2011	Sentence	VIDTIMIT	PCA + HMM	60.1%
Estelleers et al. [88]	2012	Digits	Own data	DCT + LDA + HMM	71%

Lan et al. [89]	2012	Sentence	LILiR	AAM + HMM	33%
Pei et al. [90]	2013	Alphabet	AVLetters	RFMA	69.6%
	2013	Phrases	OuluVS	RFMA	89.7%
Noda et al. [91]	2014	Words	ATR	CNN + GMM-HMM	37%
Stewart et al. [92]	2014	Digits	XM2VTS	DCT + MS-HMM	70%
Biswas et al. [93]	2015	Sentence	AVICAR	AAM + HMM	28.23
Sui et al. [56]	2015	Digits	AusTalk	DBM + DCT + LDA + HMM	69.1%
Assael et al. [5]	2016	Phrases	GRID	3D-CNN + BiGRU + CTC	93.4%
Chung et al. [94]	2016	Phrases	OuluVS2	VGG-M + LSTM	31.9%
	2016	Phrases	OuluVS2	SyncNet + LSTM	94.1%
Sui et al. [62]	2017	Phrases	OuluVS	CHAVF + SVM	68.9%
	2017	Digits	AusTalk	CHAVF + HMM	69.18%
Petridis et al. [55]	2017	Phrases	OuluVS2	Autoencoder + LSTM	84.5.8%
wand et al. [95]	2017	Phrases	GRID	Feed-Forward + LSTM	42.4%
Petridis et al. [96]	2018	Words	LRW	3D CNN +ResNet +BiGRU	82%
Estival et al. [99]	2018	Phrases	AV Digits	Autoencoder + BiLSTM	69.7%
Wand et al. [97]	2018	Phrases	GRID	Feed-Forward + LSTM	84.7%
Mesbah et al. [11]	2019	Words	LRW	CFI + Hahn CNN	58.2%
Zhang et al. [30]	2019	Sentence	CCTC	VGG-M + ResNet +BiLSTM +CTC + GRU + Attention	50.2%
Wang et al. [20]	2019	Words	LRW	3D CNN + Bi-Conv-LSTM	83.34%
Zhang et al. [98]	2020	Words	LRW	3D CNN +ResNet + BiGRU	85.2%
Ma et al. [99]	2021	Sentence	LRS2	3D CNN + ResNet + Conformer Encoder	62.1%
Tsourounis at al. [38]	2021	word	LRW-500	MS-TCN +CNN	87.01%
Prajwal at al. [17]	2022	sub-word	LRS2	VTP +CNN	88.2
Dweik at al. [8]	2022	Words	Own data	CNN + LSTM	82.84%
Atila and Sabaz [42]	2022	word	Own data	CNN + Bi-LSTM	84.5%

8. Conclusion

Based on the survey paper, lip reading is an active topic of research with many techniques used for improving performance. Some of the techniques include feature extraction, classification, a fusion of visual and audio features, and error correction. The performance of these techniques varies depending on the specific approach and application.

VSR faces many challenges. The most important of these challenges is the similarity of some phonemes and the high variability of lip movements caused by factors such as accents, speech rate, facial expressions, lighting conditions, and camera angles. Additionally, limited access to large-scale annotated datasets and a lack of standard evaluation metrics makes comparing the effectiveness of various approaches challenging.

Despite these challenges, recent literature indicates significant progress in the field, with notable achievements in areas like recognition of speech, identification of the speaker, and recognition of emotions. Several large-scale datasets have also been developed, including the OuluVS2, LRW, GRID, and LRS datasets, which have contributed to the advancement of lip-reading research.

In conclusion, lip reading is an important and challenging area of research with potential applications in different fields like recognition of speech, human-computer interaction, and security. While there are still many challenges to overcome, recent progress in the field is promising and suggests that further advances are possible.

References

[1] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, "A survey of research on lipreading technology," *IEEE Access*, vol. 8. Institute of Electrical and Electronics Engineers Inc., pp. 204518–204544, 2020. doi: 10.1109/ACCESS.2020.3036865.

[2] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, "Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers." [Online]. Available: www.aaai.org

[3] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip Reading Sentences in the Wild."

[4] G. Zhang and Y. Lu, "Research on a Lip Reading Algorithm Based on Efficient-GhostNet," *Electronics (Switzerland)*, vol. 12, no. 5, Mar. 2023, doi: 10.3390/electronics12051151.

[5] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-End Sentence-level Lipreading," Nov. 2016, [Online]. Available: http://arxiv.org/abs/1611.01599

[6] A. B. A. Hassanat, "4 Visual Speech Recognition." [Online]. Available: www.intechopen.com

[7] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, "Deep Learning-Based Automated Lip-Reading: A Survey," *IEEE Access*, vol. 9. Institute of Electrical and Electronics Engineers Inc., pp. 121184–121205, 2021. doi: 10.1109/ACCESS.2021.3107946.

[8] W. Dweik, S. Altorman, and S. Ashour, "Read my lips: Artificial intelligence word-level arabic lipreading system," *Egyptian Informatics Journal*, vol. 23, no. 4, pp. 1–12, Dec. 2022, doi: 10.1016/j.eij.2022.06.001.

[9] B. S. Lin, Y. H. Yao, C. F. Liu, C. F. Lien, and B. S. Lin, "Development of novel lip-reading recognition algorithm," *IEEE Access*, vol. 5, pp. 794–801, 2017, doi: 10.1109/ACCESS.2017.2649838.

[10] K. Xu, D. Li, N. Cassimatis, and X. Wang, "LCANet: End-to-End Lipreading with Cascaded Attention-CTC," Mar. 2018, [Online]. Available: http://arxiv.org/abs/1803.04988

[11] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, "Lip reading with Hahn Convolutional Neural Networks," *Image Vis Comput*, vol. 88, pp. 76–83, Aug. 2019, doi: 10.1016/j.imavis.2019.04.010.

[12] Dharin Parekh, Ankitesh Gupta, Shharrnam Chhatpar, Anmol Yash, and Manasi Kulkarni, "2019 IEEE 5th International Conference for Convergence in Technology (I2CT).," 2019.

[13] A. Garg, J. Noyola, and S. Bagadia, "Lip reading using CNN and LSTM."

[14] R. Gu, S.-X. Zhang, Y. Xu, L. Chen, Y. Zou, and D. Yu, "Multi-modal Multi-channel Target Speech Separation," Mar. 2020, doi: 10.1109/JSTSP.2020.2980956.

[15] Y. Koguchi, K. Oharada, Y. Takagi, Y. Sawada, B. Shizuki, and S. Takahashi, "A mobile command input through vowel lip shape recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 297–305. doi: 10.1007/978-3-319-91250-9_23.

[16] Li Lu *et al.*, "LipPass: Lip Reading-based User Authentication on Smartphones Leveraging Acoustic Signals," *IEEE Communications Society Institute of Electrical and Electronics Engineers*, 2018.

[17] K. R. Prajwal, T. Afouras, and A. Zisserman, "Sub-word Level Lip Reading With Visual Attention." [Online]. Available: https://www.robots.ox.ac.uk/

[18] A. M. Sarhan, N. M. Elshennawy, and D. M. Ibrahim, "HLR-Net: A hybrid lip-reading model based on deep convolutional neural networks," *Computers, Materials and Continua*, vol. 68, no. 2, pp. 1531–1549, Apr. 2021, doi: 10.32604/cmc.2021.016509.

[19] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual Information Maximization for Effective Lip Reading," in *Proceedings - 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 420–427. doi: 10.1109/FG47880.2020.00133.

[20] C. Wang, "Multi-Grained Spatio-temporal Modeling for Lip-reading," Aug. 2019, [Online]. Available: http://arxiv.org/abs/1908.11618

[21] Leticia Ria Aran, Farrah Wong, and Lim Pei Yi, "A Review on Methods and Classifiers in Lip Reading Proceedings, 2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS 2017) : Malaysia, 21 October 2017," *IEEE Control Systems Society. Chapter Malaysia Institute of Electrical and Electronics Engineers*.

[22] H. Wang, G. Pu, and T. Chen, "A Lip Reading Method Based on 3D Convolutional Vision Transformer," *IEEE Access*, vol. 10, pp. 77205–77212, 2022, doi: 10.1109/ACCESS.2022.3193231.

[23] M. Oghbaie, A. Sabaghi, K. Hashemifard, and M. Akbari, "Advances and Challenges in Deep Lip Reading," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.07879

[24] A. Elhassan, M. Al-Fawa'reh, M. T. Jafar, M. Ababneh, and S. T. Jafar, "DFT-MF: Enhanced deepfake detection using mouth movement and transfer learning," *SoftwareX*, vol. 19, Jul. 2022, doi: 10.1016/j.softx.2022.101115.

[25] M. Mahmmed, T. Saeed, and W. Ali, "Robust Visual Lips Feature Extraction Method for Improved Visual Speech Recognition System," *Engineering and Technology Journal*, vol. 36, no. 2A, pp. 136–145, Feb. 2018, doi: 10.30684/etj.36.2a.4.

[26] T. Afouras, J. S. Chung, and A. Zisserman, "Deep Lip Reading: a comparison of models and an online application," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.06053

[27] S. Petridis, J. Shen, D. Cetin, and M. Pantic, "Visual-only recognition of normal, whispered and silent speech," in *2018 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2018, pp. 6219–6223.

[28] J. Wei, F. Yang, J. Zhang, R. Yu, M. Yu, and J. Wang, "Three-dimensional joint geometric-physiologic feature for lip-reading," in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, IEEE Computer Society, Dec. 2018, pp. 1007–1012. doi: 10.1109/ICTAI.2018.00155.

[29] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Applied Sciences (Switzerland)*, vol. 9, no. 8, Apr. 2019, doi: 10.3390/app9081599.

[30] X. Zhang, H. Gong, X. Dai, F. Yang, N. Liu, and M. Liu, "Understanding Pictograph with Facial Features: End-to-End Sentence-Level Lip Reading of Chinese." [Online]. Available: www.aaai.org

[31] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip Reading Sentences Using Deep Learning with only Visual Cues," *IEEE Access*, vol. 8, pp. 215516–215530, 2020, doi: 10.1109/ACCESS.2020.3040906.

[32] W. H. Ali, T. R. Saeed, and M. H. Al-Muifraje, "FPGA Implementation of Visual Speech Recognition System based on NVGRAM-WNN," in *2020 International Conference on Computer Science and Software Engineering (CSASE)*, IEEE, 2020, pp. 132–137.

[33] W. Chen, X. Tan, Y. Xia, T. Qin, Y. Wang, and T. Y. Liu, "DualLip: A System for Joint Lip Reading and Generation," in *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2020, pp. 1985–1993. doi: 10.1145/3394171.3413623.

[34] M. Tanaka, E. Nakamura, Y. Kageyama, M. Shirasu, M. Ishii, and M. Nishida, "Identification of Utterance Content Using Lip Movement Features," in *LifeTech 2020 - 2020 IEEE 2nd Global Conference on Life Sciences and Technologies*, Institute of Electrical and Electronics Engineers Inc., Mar. 2020, pp. 167–168. doi: 10.1109/LifeTech48969.2020.1570618728.

[35] D. Feng, S. Yang, S. Shan, and X. Chen, "Learn an Effective Lip Reading Model without Pains," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.07557

[36] Maxalmina, S. Kahfi, K. N. Ramadhani, and A. Arifianto, "Lip Motion Recognition for Indonesian Vowel Phonemes Using 3D Convolutional Neural Networks," in *2020 3rd International Conference on Computer and Informatics Engineering, IC2IE 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 157–161. doi: 10.1109/IC2IE50715.2020.9274562.

[37] A. Kurniawan and S. Suyanto, "Syllable-based Indonesian lip reading model," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 2020, pp. 1–6.

[38] D. Tsourounis, D. Kastaniotis, and S. Fotopoulos, "Lip reading by alternating between spatiotemporal and spatial convolutions," *J Imaging*, vol. 7, no. 5, May 2021, doi: 10.3390/jimaging7050091.

[39] E. Egorov, V. Kostyumov, M. Konyk, and S. Kolesnikov, "LRWR: Large-Scale Benchmark for Lip Reading in Russian language," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.06692

[40] Z. Lin *et al.*, "SimulLR: Simultaneous Lip Reading Transducer with Attention-Guided Adaptive Memory," in *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, Association for Computing Machinery, Inc, Oct. 2021, pp. 1359–1367. doi: 10.1145/3474085.3475220.

[41] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with Densely Connected Temporal Convolutional Networks."

[42] Ü. Atila and F. Sabaz, "Turkish lip-reading using Bi-LSTM and deep learning models," *Engineering Science and Technology, an International Journal*, vol. 35, Nov. 2022, doi: 10.1016/j.jestch.2022.101206.

[43] C. G. Lee, E. S. Lee, S. T. Jung, and S. S. Lee, "Design and implementation of a real-time lipreading system using PCA and HMM," *Journal of Korea Multimedia Society*, vol. 7, no. 11, pp. 1597–1609, 2004.

[44] G. Sterpu and N. Harte, "Towards lipreading sentences with active appearance models," *arXiv preprint arXiv:1805.11688*, 2018.

[45] J. He, H. Zhang, and J. Z. Liu, "LDA based feature extraction method in DCT domain in lipreading," *Computer Engineering and Applications*, vol. 45, no. 32, pp. 150–155, 2009.

[46] N. Puviarasan and S. Palanivel, "Lip reading of hearing impaired persons using HMM," *Expert Syst Appl*, vol. 38, no. 4, pp. 4477–4481, Apr. 2011, doi: 10.1016/j.eswa.2010.09.119.

[47] A. J. Goldschen, O. N. Garcia, and E. Petajan, "Continuous optical automatic speech recognition by lipreading," in *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, Ieee, 1994, pp. 572–577.

[48] Z. Zhou, G. Zhao, and M. Pietikäinen, "Towards a practical lipreading system," in *CVPR 2011*, IEEE, 2011, pp. 137–144.

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

[50] S. Ramakrishnan, *Speech Enhancement, Modeling and Recognition-Algorithms and Applications*. BoD–Books on Demand, 2012.

[51] K. Kumar, T. Chen, and R. M. Stern, "Profile view lip reading," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, IEEE, 2007, pp. IV–429.

[52] M. Kass and A. Witkin, "Snakes: Active Contour Models," KIuwer Academic Publishers, 1988.

[53] M. H. Rahmani and F. Almasganj, "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)*, IEEE, 2017, pp. 195–199.

[54] M. Wand, J. Koutník, and J. Schmidhuber, "Lipreading with long short-term memory," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2016, pp. 6115–6119.

[55] S. Petridis, Z. Li, and M. Pantic, "End-to-end visual speech recognition with LSTMs," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 2592–2596.

[56] C. Sui, M. Bennamoun, and R. Togneri, "Listening with your eyes: Towards a practical visual speech recognition system using deep boltzmann machines," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 154–162.

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[58] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 1, pp. 221–231, 2012.

[59] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Trans Multimedia*, vol. 18, no. 3, pp. 326–338, 2016.

[60] A. Rekik, A. Ben-Hamadou, and W. Mahdi, "An adaptive approach for lip-reading using image and depth data," *Multimed Tools Appl*, vol. 75, pp. 8609–8636, 2016.

[61] A. Nasuha, F. A. T. A. Sardjono, H. Takahashi, and M. H. Purnomo, "Automatic lip reading for daily Indonesian words based on frame difference and horizontal-vertical image projection," *J Theor Appl Inf Technol*, vol. 95, no. 2, p. 393, 2017.

[62] C. Sui, R. Togneri, and M. Bennamoun, "A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition," *Speech Commun*, vol. 90, pp. 26–38, 2017.

[63] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. IEEE-CS Conference on Computer Vision and Pattern Recognition*, 1985, pp. 40–47.

[64] C. Neti *et al.*, "Audio visual speech recognition," IDIAP, 2000.

[65] Y. W. Wong *et al.*, "A new multi-purpose audio-visual UNMC-VIER database with multiple variabilities," *Pattern Recognit Lett*, vol. 32, no. 13, pp. 1503–1510, 2011.

[66] E. Bailly-Bailliére *et al.*, "The BANCA database and evaluation protocol," in *Audio-and Video-Based Biometric Person Authentication: 4th International Conference, AVBPA 2003 Guildford, UK, June 9–11, 2003 Proceedings 4*, Springer, 2003, pp. 625–638.

[67] N. A. Fox, B. A. O'Mullane, and R. B. Reilly, "VALID: A new practical audio-visual database, and comparative results," in *Audio-and Video-Based Biometric Person Authentication: 5th International Conference, AVBPA 2005, Hilton Rye Town, NY, USA, July 20-22, 2005. Proceedings 5*, Springer, 2005, pp. 777–786.

[68] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J Acoust Soc Am*, vol. 120, no. 5, pp. 2421–2424, 2006.

[69] D. Petrovska-Delacrétaz *et al.*, "The iv 2 multimodal biometric database (including iris, 2d, 3d, stereoscopic, and talking face data), and the iv 2-2007 evaluation campaign," in *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, IEEE, 2008, pp. 1–7.

[70] I. Anina, Z. Zhou, G. Zhao, and M. Pietikäinen, "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)*, IEEE, 2015, pp. 1–5.

[71] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Trans Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[72] M. Igras, B. Ziółko, and T. Jadczyk, "Audiovisual database of Polish speech recordings," *Studia Informatica*, vol. 33, no. 2B, pp. 163–172, 2012.

[73] S. Antar and A. Sagheer, "Audio visual Arabic speech (AVAS) database for human-computer interaction applications," *The International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no. 9, 2013.

[74] B. Shillingford *et al.*, "Large-scale visual speech recognition," *arXiv preprint arXiv:1807.05162*, 2018.

[75] V. Verkhodanova, A. Ronzhin, I. Kipyatkova, D. Ivanko, A. Karpov, and M. Železný, "HAVRUS corpus: high-speed recordings of audio-visual Russian speech," in *Speech and Computer: 18th International Conference, SPECOM 2016, Budapest, Hungary, August 23-27, 2016, Proceedings 18*, Springer, 2016, pp. 338–345.

[76] T. Afouras, J. S. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[77] L. A. Elrefaei, T. Q. Alhassan, and S. S. Omar, "An Arabic visual dataset for visual speech recognition," *Procedia Comput Sci*, vol. 163, pp. 400–409, 2019.

[78] X. Chen, J. Du, and H. Zhang, "Lipreading with DenseNet and resBi-LSTM," *Signal Image Video Process*, vol. 14, pp. 981–989, 2020.

[79] P. Lucey, G. Potamianos, and S. Sridharan, "A unified approach to multi-pose audio-visual ASR," in *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, Causal Productions Pty Ltd, 2007, pp. 650–653.

[80] P. Lucey, G. Potamianos, and S. Sridharan, "Patch-based analysis of visual speech from multiple views.," in *AVSP*, 2008, pp. 69–74.

[81] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audiovisual speech recognition," *IEEE Trans Audio Speech Lang Process*, vol. 17, no. 3, pp. 423–435, 2009.

[82] X. Shao and J. Barker, "Stream weight estimation for multistream audio–visual speech recognition in a multispeaker environment," *Speech Commun*, vol. 50, no. 4, pp. 337–353, 2008.

[83] S. Hilder, R. W. Harvey, and B.-J. Theobald, "Comparison of human and machine-based lip-reading.," in *AVSP*, 2009, pp. 86–89.

[84] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Trans Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[85] A. Pass, J. Zhang, and D. Stewart, "An investigation into features for multi-view lipreading," in *2010 IEEE International Conference on Image Processing*, IEEE, 2010, pp. 2417–2420.

[86] T. Saitoh and R. Konishi, "Profile lip reading for vowel and word recognition," in *2010 20th International conference on pattern recognition*, IEEE, 2010, pp. 1356–1359.

[87] L. Cappelletta and N. Harte, "Viseme definitions comparison for visual-only speech recognition," in *2011 19th European Signal Processing Conference*, IEEE, 2011, pp. 2109–2113.

[88] V. Estellers and J.-P. Thiran, "Multi-pose lipreading and audio-visual speech recognition," *EURASIP J Adv Signal Process*, vol. 2012, pp. 1–23, 2012.

[89] Y. Lan, B.-J. Theobald, and R. Harvey, "View independent computer lip-reading," in *2012 IEEE International Conference on Multimedia and Expo*, IEEE, 2012, pp. 432–437.

[90] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 129–136.

[91] K. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata, "Lipreading using convolutional neural network," in *fifteenth annual conference of the international speech communication association*, 2014.

[92] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE Trans Cybern*, vol. 44, no. 2, pp. 175–184, 2013.

[93] A. Biswas, P. K. Sahu, and M. Chandra, "Multiple camera in car audio–visual speech recognition using phonetic and visemic information," *Computers & Electrical Engineering*, vol. 47, pp. 35–50, 2015.

[94] J. S. Chung and A. Zisserman, "Out of time: automated lip sync in the wild," in *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, Springer, 2017, pp. 251–263.

[95] M. Wand and J. Schmidhuber, "Improving speaker-independent lipreading with domain-adversarial training," *arXiv preprint arXiv:1708.01565*, 2017.

[96] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 6548–6552.

[97] M. Wand, J. Schmidhuber, and N. T. Vu, "Investigations on end-to-end audiovisual fusion," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 3041–3045.

[98] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, 2020, pp. 356–363.

[99] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 7613–7617.