Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2022) 14(2), 844-860.

10.29137/umagd.1038899

Cilt/Volume:14 Sayı/Issue:2 Temmuz/July 2022

Araştırma Makalesi / Research Article

A Survey on Lip-Reading with Deep Learning

Ali Erbey[1], **Necaattin Barışçı** [2]

[1]*Department of Computer Programming, Distance Education Vocational School, Usak University, 64200, Usak, TURKEY*
[2]*Department of Computer Engineering, Faculty of Technology, Gazi University, 06500, Ankara, TURKEY*

***Başvuru/Received**: 22/12/2021* ***Kabul / Accepted**: 29/07/2022* ***Çevrimiçi Basım / Published Online:** 31/07/2022*

***Son Versiyon/Final Version**: 31/07/2022*

Abstract

Very successful results have been obtained in areas such as computer vision and voice recognition when applying deep learning methods. Technologies that facilitate the lives of people have been developed as a result of the successes of deep learning within these areas. One of these technologies is voice recognition devices. Research has shown that these devices do not give good results in noisy environments; although, they do give good results in silent environments. With deep learning methods, voice recognition in noisy environments can be achieved using visual signals. Thanks to computerized vision, the success of voice recognition devices can be increased with the analysis of human lips in order to determine what the speaker is saying. In this study, lip-reading studies using deep learning methods published between 2017 and 2020 were examined and data sets were introduced. As a result of the study, it is seen that CNN and LSTM architectures are used more intensively in lip-reading studies, hybrid models are preferred more and the success rates are increasing day by day. In this context, it is seen that technologies that can be used in line with the need can be developed by conducting more academic studies on lip reading.

Key Words

“Lipreading, Deep Learning, Convolutional Neural Networks, Artificial Neural Networks”

***Corresponding Author:** ali.erbey@usak.edu.tr*

1. Introduction

Technology is used to solve many modern problems in the daily lives of society. With technology, individuals have had increasingly easy and more comfortable living conditions. Technology is now used in nearly every domain including the household, health, and industry. The most current advancements in technology can be seen within technological devices. These devices have been developed over time and focus on solving problems for the user. The quality of life for individuals is enhanced, especially in regard to devices which are made for the health field. Although quality of life does include health and wellness, it can be further defined as a standard of excellence or quality in all arenas of life, over a lifetime (Doğan et al. 2016). From this point of view, the aim of technology is to increase social well-being. From this point of view, the quality of life is increased by developing solutions for negative situations in people's health with technology. People's health problems seem to increase as they age (Olgun et al. 2013). As a person ages, there is a higher probability of hearing loss, which is one of the three most common diseases observed in old age (Erdoğan, 2016). Hearing loss has major negative impacts on quality of life (Mulrow et al. 1990). People who had once had clear hearing but later in life encounter hearing loss often isolate themselves from society. It has also been shown that hearing loss among the elderly has a strong correlation with depression (Mulrow et al. 1990). Hearing loss, which can lead to isolation and depression, undoubtedly decreases the quality of life in the elderly. This decrease related to hearing loss can cause further social, emotional, and communication difficulties (Yueh et al. 2003). On the other hand, it is also known that individuals with innate hearing-impairment invariably experience these difficulties. In this context, it is imperative that new technological solutions are developed that are relevant for individuals with hearing loss.

Three current solutions which have been developed for aiding individuals with hearing loss include in-ear hearing aids, sign language recognition technology, and lip-reading. When studies concerning the effects of hearing aids on quality of life are examined, it is seen that the devices have a positive impact (Mulrow et al. 1992),(Hamurcu et al. 2012). In a study made by Hamurcu et al. in 2012, it was observed that 82% of individuals using hearing aids had seen a rise in their quality of life (Hamurcu et al. 2012). It is also known that hearing aids which make speech audible, lose their clarity—notably, in noisy environments (Gogate et al. 2020) In the study by Kahveci et al., it was concluded that the noise issues experienced while using the hearing aid devices negatively affected the satisfaction of the individuals (Kahveci et al. 2011). It can be said that sign language is a tool which facilitates communication between those who do not have hearing loss and those who have innate hearing-impairment. Sign language is useful because of its ability to provide communication within a community (Cheok et al. 2019).

The inadequacies of hearing aids led people to sign language to continue communication. Here, in order to be useful to people, systems that recognize sign language are being developed. These systems are used with existing computers, game consoles and mobile applications. Computer vision, machine learning and deep learning are used in the infrastructure of sign language recognition systems. The upper component of all these structures is artificial intelligence. The fact that it has achieved significant success these days is due to the academic studies carried out for many years.

Turing, in his work in 1950, "Can machines think?" raised the question which led to the discussion of machine intelligence (Turing, 1950). Discussions on machine intelligence began and the concept of artificial intelligence emerged just 6 years later. The concept of "artificial intelligence" was used for the first time in 1956 by John McCarthy (Russell, 2016). The concept of machine learning emerged along with the concept of artificial intelligence. Alpaydın (2020) defines machine learning as computers which have been programmed to optimize a performance criterion by using past experiences or data (Alpaydin, 2020). As a result of these optimizations, many methods of machine learning have emerged to date. Artificial neural networks (ANN), is one of the most important areas of machine learning (Ergezer et al. 2003). In 1957, Rosenblatt defined the concept of "perceptron", which forms the basis of ANN (Rosenbaltt, 1957). As a result of Minsky and Papert (1969) stating that the XOR problem could not be solved with perceptron, progress in the field of artificial intelligence was disrupted for a while (Minsky & Papert, 1969). However, Rumelhart et al. (1988) proposed that Multi-Layer Perceptron (MLP) could also solve the XOR problem and the importance of ANN increased once again (Rumelhartet al. 1986). Occasionally, work in the field of artificial intelligence has been stagnate, such as the period of time when the XOR problem was unsolvable. These periods have been labelled "artificial intelligence winter", and has occurred throughout the late 1960s to the early 1970s and during the late 1980s (Bollier, 2017).

In 1998, Yann LeCun et al. used Convolutional Neural Networks (CNN) in the MNIST dataset to classify characters; as a result of their success, CNN came to the forefront of the field (LeCun et al. 1998). CNN (or ConvNet) is a specialized form of MLP—inspired by ANN—used in classification, image recognition, object recognition, and natural language processing. CNN is a form of deep learning, which is a method that has grown in popularity in recent years (Arı & Hanbay, 2019). The real breakthrough in the field of deep learning occurred in 2012. Ranking first in the ImageNet competition of that year, a research group headed by Hinton, had been able to reduce the Top-5 error rate from 25.8% (of the previous year) to 16.4% (Krizhevsky et al. 2012). With this new architecture, called AlexNet, deep learning studies earnestly resumed. Not only powerful hardware, large datasets, and large models but also the development of architectures and algorithms played an important role in increasing the progress of deep learning research (Szegedy et al. 2015). Architectures continued to emerge as new algorithms were developed; in later years, at the ImageNet competition groups such as VGGNet (Simonyan & Zisserman, 2014), GoogleNet (Szegedy et al. 2015), ResNet (He et al. 2016) persisted in gradually decreasing the error rate.

The architectures used in deep learning methods and the results that improve day by day have led many scientists to this field. With more contributions from scientists, computer vision, machine learning and deep learning have started to get very successful results in the last few decades. In this process, it is seen that significant improvements have been made especially in the field of health. Thanks to computer vision, systems are being developed to assist doctors on MRI, ultrasound and X-ray images. As these systems develop, people will be able to achieve health and improve their quality of life. Lip reading is also one of the areas that will improve the quality of life of people with hearing impairments and better results can be achieved as technological developments such as artificial intelligence continue. In this article, lip-reading studies which use deep learning methods are examined. Lip-reading is the process of understanding speech from lip movements. Due to the fact that images of the mouth region are more successfully analyzed by researchers when utilizing deep learning methods, lip-reading studies have become a subject of interest within deep learning. After 2012, the number of studies on deep learning increased and new architectures began to emerge. Lip-reading studies which have been published in recent years, include methods that use different architectures and hybrid structures with have the highest performance rate. Lip-reading research using deep learning methods, that have been conducted between 2017 and 2020, have been included in this study.

Following this introduction, ANN which is a pioneer for many technological developments and provides the basis for increased success rates—are introduced in the second section. The third section will examine deep learning architectures which are also a sub-field of machine learning and is an additional reason why this study was conducted. Lip-reading will be discussed in the fourth section while different datasets will be reviewed in the fifth section. In the final section, lip-reading studies using deep learning architectures are presented.

2. Artificial Neural Networks

ANN are structures that are created by modelling the human brain and how it utilizes its connections, which allow us to process information in tandem or sporadically (Uğur & Kınacı, 2006). Each individual structure such as neurons and dendrites existing in the human brain have its counterpart in ANN.

McCulloch and Pitts conducted the first study on ANN in 1943 (McCulloch & Pitts, 1943). This study has been used in many different areas and this subject has developed substantially since its first conception. ANN is used frequently in fields such as voice recognition, optical character recognition, robot controls, image processing, and face recognition (Ergezer et al. 2003). Artificial neural networks have achieved very successful results in these areas. In the following years, deep learning architectures were created by increasing the number of layers in artificial neural networks.

3. Deep Learning

The widespread use of the Internet has led to the frequent use of social media and the production of data—such as photographs, videos, text sharing, and internet records—which has led to the creation of datasets called "big data" (Ertam & Aydın, 2017). Artificial Intelligence studies have gained momentum in obtaining sufficient data through big data and increasing computing power parallel to GPUs. Computing power has enabled more data layers to be employed in ANN. With the addition of such numerous layers, deep learning methods have emerged.

Deep learning falls under the category of machine learning, which falls under the field of artificial intelligence. Unlike machine learning, deep learning performs feature extraction processes within its data layers. In Figure 1, machine learning and deep learning are shown.

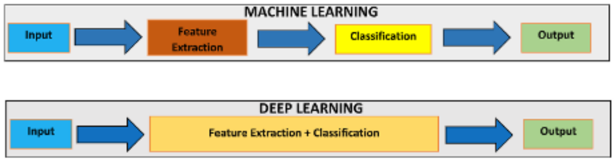

Figure 1. Machine learning and deep learning (Amanullah et al. 2020)

As seen above, the fact that deep learning is not overly involved with feature extraction is a major factor in its increase in popularity. Another factor is that it allows multiple applications to be feasible. Deep learning can easily perform applications such as classification, regression, segmentation, transfer learning, and adaptation. In these applications, different deep learning architectures are used. Architectures frequently used in deep learning include Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), Long-Short Term Memory (LSTM), and Deep Belief Networks (DBN).

3.1. Recurrent Neural Network

RNN, originally introduced by Hopfield in 1984 (Hopfield, 1984) is an ANN class in which connections between units form a unified directional loop (Pang et al. 2020). As opposed to Forward Networks, where inputs are not dependent on previous input, RNN evaluates inputs not only instantaneously but also based on previous inputs. Thus, a temporal structure is formed dynamically in the network. The temporal structure is updated according to the values of the new state and of past states (Bacciu et al. 2020). Thanks to this structure, RNN is capable of understanding the structure of multiple data such as text, speech, and various sensors depending on time or statistical data.

3.2. Convolutional Neural Network

Looking at multiple definitions of CNN, it is emphasized that it is a deep learning tool (Yang et al. 2020) for image analysis as well as a feed forward ANN class (Sam et al. 2019) for analysing visual images in machine learning. With these definitions, CNN is a deep feed forward neural network class. CNN was first introduced in 1998 by LeCun and his team (LeCun et al. 1998). After this date, it has been widely used in image processing studies. In a network within CNN, an input image is taken and the image is classified into certain categories. It performs these operations on the matrix which allows the computers to see the image. Throughout these operations, matrices are subjected to a series of convolution processes. Thanks to the convolution process, qualitative features in the image appear, such as sharp edges and corners. This is done by using filters such as 3 x 3, 5 x 5, or 7 x 7 (UNIT) on the image. A basic CNN structure consists of convolution, activation, pooling, and fully connected layers.

3.3. Long-Short Term Memory

LSTM is an RNN structure which models time sequences more accurately than traditional a RNN (Sak et al. 2014). LSTM is a special class of RNN structures (Pang et al. 2020); it is often used for time series tasks (Chen et al. 2020). LSTM was developed by Hochreiter and Schmidhuber in 1997 to edit the vanishing gradient problem in RNN (Hochreiter & Schmidhuber, 1997). By adding a gate to the cell, they increased the capacity of the standard cell (Yu et al. 2020).

3.4. Deep Belief Networks

DBN is a component of Restricted Boltzmann Machines (RBM) that use Gibbs sampling to train parameters (Farsal et al. 2018). Within RBM that use a single hidden layer, it was not easy to capture hidden properties of data (Chen et al. 2015). Therefore, DBN was created with a probability model, which was added to the stack, for multiple restricted Boltzmann machines (Lv & Qiao, 2020). DBN has a strong structure thanks to its layered learning method whereby it reconstructs both learning-capable vectors and input vectors (Saif et al. 2018). DBN has many characteristics such as feature extraction and classification which are applied in image processing and speaking (Keyvanrad & Homayounpour, 2014).

4. Lip-Reading

Image processing, which is used in many fields, has been applied with deep learning methods in recent years. This process has been a helpful tool—especially in the medical field—because it not only assists in understanding the genetics of diseases but also recommends to doctors correct diagnosis and treatment methods (Esteva et al. 2019). It is widely used in autonomous car technologies, which is the way of car technology today (Fayjie et al. 2018). It is also seen in systems such as face and voice recognition, license plate recognition (Bayram, 2020), and fingerprint readers (Akmese et al. 2019).

Lip-reading problems are issues that can be solved with image processing. In the earliest studies on lip-reading, the region of interest (ROI) was extracted and attempts were made to develop speech defining models (Dupont & Luettin, 2000; Potamianos et al. 2004; Matthews et al. 2002). Moreover, when we look at studies that came later, it is clear that machine learning and deep learning methods were used. CNN and LSTM architectures in particular are principally used in lip-reading research associated with deep learning methods.

Lip-reading is a way for visual and audio signals to support the understanding of a speaker. Lip-reading studies utilizing deep learning methods, which have been conducted with in recent years, are examined in this article. These include surveys conducted by Potamianos et al. in 2004 (Potamianos et al. 2004) Zhou et al. in 2014 (Zhou et al. 2014), and Fernandaz and Sukno in 2018 (Fernandez-Lopez & Sukno, 2018).

For deep learning architectures, it is essential that there is a sufficient amount of data and that it can be interpreted. The success of deep learning architectures is affected by the amount of data within a dataset, the difference that are able to reflect the real world, and the fact that the data is clear and understandable. In order to compare the success of architectures, it is observed that certain datasets are used more often in some subjects, similar to the ImageNet datasets that were used in the image recognition competition described earlier. Similarly, it is seen that certain datasets are used more in lip-reading research.

5. Datasets

In lip-reading studies, it has been noted that some research uses its own datasets, albeit a small number (Lu & Yan, 2020; Goh et al. 2019). Generally, studies are conducted on commonly used datasets (Petridis et al. 2020; Mesbah et al. 2019), known to be OuluVS2 (Anina et al. 2015), AvLetters (Matthews et al. 2002) and GRID. The datasets examined in this section are the most used datasets in the articles reviewed in this study. These datasets differ in content and quality. The content, that can be seen in the datasets, consists of videos in which letters, numbers, or sentences are spoken out loud. In terms of quality, the resolution of the videos and the number of frames per second (fps) are revealed.

5.1. AvLetters2

The AvLetters2 dataset was created by Cox et al. in 2008 (Cox et al. 2008). In order to increase image resolution and number of repetitions, from what was seen in the AvLetters dataset, the videos for AvLetters2 were recorded at 50 fps at a 1920 x 1080 px resolution. Additionally, the speakers were required to make 3 to 7 repetitions of the 26 letters of the English alphabet (Fernandez-Lopez & Sukno, 2018). The videos were taken in one session to avoid any illumination differences and also recorded at an anterior view (Cox et al. 2008). The AvLetters2 dataset has 28 videos, all between 47 and 58 seconds long—producing between 1169 to 1499 frames each. Since single letters were included, it contributed to the controlled experiments (Bear & Harvey, 2017).

5.2. OuluVS2

In this dataset, 52 speakers were required to repeat a series of numbers 10 times; the series includes 10 numbers such as 1, 7, 3, 5, 1, 6, 2, 6, 6, 7. The videos in the dataset have been clipped to display only the mouth region and were shot from different angles (Anina et al. 2015) exhibited in Figure 2.

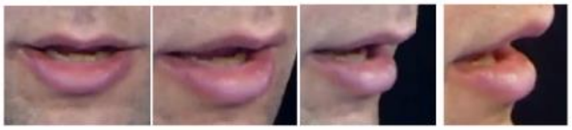

Figure 2. OuluVS2 dataset video frames (Mesbah et al. 2019)

5.3. LRW

In 2016, Oxford-BBC published the LRW dataset, including 500 different words, each spoken up to 1000 times by multiple speakers. Contained within, is metadata which includes location and time information of the words (Chung & Zisserman, 2016). The LRW dataset is challenging in that it contains a large amount of words set close together and from different word classes (Zhao et al. 2020). Frames can be seen below, in Figure 3.

Figure 3. LRW dataset video frames (Stafylakis & Tzimiropoulos, 2017)

5.4. GRID

The GRID dataset consists of 34 speakers who were required to say 1,000 sentences each. Approximately 34,000 pieces of data (34*1,000) were obtained and the speakers were recorded from an anterior view (Cooke et al. 2006). Out of the 34 speakers, 18 are male and 16 are female. Each sentence consists of a 6-word string: command, colour, preposition, letter, number, adverb. "Put Red at G Nine Now", is an example of this sentence structure. The videos are each 3 seconds long and have a frame rate of 25 fps (with a total of 75 frames each) (Qu et al. 2019), making the total length 28 hours. The dataset contains a total of 51 different words; however, the letter W was excluded due to its difficult pronunciation compared to other letters (Wand et al. 2016).

5.5. Cuave

This dataset was presented by Patterson et al. in 2002 (Patterson et al. 2002). Its videos are 29.97 fps with a 720 x 480 px resolution. More than 7,000 pieces of data were collected from 37 different speakers (Patterson et al. 2002).

The CUAVE dataset consists of both digits and numbers. The speakers (17 women and 19 men) were recorded in front of a green background and no special visual aid for face or lip segmentation was used. Obvious erroneous data was removed during recording; however, pauses and speech errors were preserved for realistic test purposes (Patterson et al. 2002).

5.6. MIRACL-VC1

This dataset includes 15 speakers repeating 10 words, 10 different times; forming a total of 1,500 words (15*10*10). The dataset was recorded at a 640 x 480 px resolution under good lighting (Rekiket al. 2014). Words from the MIRACL-VC1 dataset are shown in Table 1.

Table 1. MIRACL-VC1 dataset words (Sindhura et al. 2018)

No	Word
1	Begin
2	Choose
3	Connection
4	Navigation
5	Next
6	Previous
7	Start
8	Stop
9	Hello
10	Web

Aside from the words seen in Table 1, there were also phrases such as “Nice to meet you” and “I love this game”. A Kinect camera was set one meter away from the speakers in order to record depth maps in the dataset (Rekik et al. 2014).

5.7. LRW-1000

With 57 hours of video, this dataset consists of 1,000 classes/sentences in total. It was created with speakers who have a Mandarin dialect. There are over 70,000 data points within this dataset, which was not prepared in any specific format in order to avoid difficulties encountered during practical applications (Yang et al. 2019).

6. Lip-Reading Through Deep Learning

With the continuous development of technology, ease of use for deep learning architectures increases; these architectures differ from time to time in practice. In this study, lip-reading studies using deep learning architectures between 2017-2020 were examined. It is aimed to reach the maximum number of articles by scanning the obtained articles in various databases. Although the reviewed articles are generally on English lip-reading studies, no language restrictions were imposed. The keywords selected while scanning the articles were determined as lip reading, visual speech decoding, and audio-visual speech recognition. The articles reached by using these keywords are shown in Table 2.

Table 2. The performance values obtained from the algorithm as a result of test processes.

Year	Source	Architecture	Dataset	Accuracy (%)
2020	Xu et al. (Xu et al. 2020)	EleAtt-GRU	LRS3-TED	77.5
			LRW	84.8
2020	Lu et al. (Lu & Yan, 2020)	CNN, BiLSTM	Own	
2020	Petridis et al. (Petridis et al. 2020)	BiLSTM	OuluVS2	95.6
			CUAVE	88.4
			AvLetters	69.2
			AvLetters2	42.6
2020	Mamatha et al. (Mamatha et al. 2020)	LSTM - CNN	LRW	88.2
2020	Martinez et al. (Martinez et al. 2020)	DenseNet (conv) + resBi + LSTM	LRW	
			LRW1000	
2020	Chen et al. (Chen et al. 2020)	3DCNN, DenseNET	LRW1000	
2020	Xiao et al. (Xiao et al. 2020)	DFNN	LRW	84.13
			LRW1000	41.3
2020	Adeel et al. (Adeel et al. 2020)		ChiME3	
			GRID	

Table 2 (continued). The performance values obtained from the algorithm as a result of test processes.

Year	Source	Architecture	Dataset	Accuracy (%)
2019	Mesbah et al. (Mesbah et al. 2019)	CNN	AvLetters OuluVS2 LRW	58.02
2019	Goh et al. (Goh et al. 2019)	RNN	Own	89
2019	Zhou et al. (Zhou et al. 2019)	Seq2seq	Own	
2019	Jang et al. (Jang et al. 2019)	CFI + QVGG + Committee	OuluVS2	90.90
2019	Li et al. (Li et al. 2019)	RNN	GRID	83.83
2019	Oliveira et al. (Oliveira et al. 2019)	CNN	GRID AVICAR OuluVS2	
2019	Muljono et al. (Muljono et al. 2019)			
2019	Bi et al. (Bi et al. 2019)	CNN + E3D-LSTM	LRW-1000	38.96
2019	Ozcan and Basturk (Ozcan & Basturk, 2019)	CNN	AvLetters	54.62
2018	Kumar et al. (Kumar et al. 2018)	STCNN + BiGRU	OuluVS2	
2018	Koumparoulis and Potamianos (Koumparoulis & Potamianos, 2018)	CNN	OuluVS2	86.39
2018	Xu et al. (Xu et al. 2018)	3D-CNN + Highway + Bi-GRU	GRID	97.10
2018	Wand et al. (Wand et al. 2018)	Feed – Forward + LSTM	GRID	
2018	Petridis et al. (Petridis et al. 2018)	AutoEncoder + BiLSTM	Av Digits	69.70
2018	Petridis et al. (Petridis et al. 2017)	3D-CNN + BiGRU	LRW	82.00
2018	Fung and Mak (Fung & Mak, 2018)	3D-CNN + BiLSTM	OuluVS2	87.60
2018	Afouras et al. (Afouras et al. 2018)	3D-CNN + BiLSTM	LRS	50.00
2017	Feng et al.	CNN + LSTM + RNN	AvLetters	57.70
2017	Wand and Schmidhuber (Wand & Schmidhuber, 2017)	Feed – forward + LSTM	GRID	42.40
2017	Thangthai et al. (Thangthai et al. 2018)	Eigenlips + DNN-HMM	TCD – TIMIT	42.97
2017	Thangthai and Harvey (Thangthai & Harvey, 2017)	PCA + LDA + MLLT + DNN-HMM	TCD – TIMIT	43.61
2017	Sui et al. (Sui et al. 2017)	CHAVF + SVM	OuluVS	68.90
2017	Stafylakis and Tzimiropulos (Stafylakis & Tzimiropoulos, 2017)	3D-CNN + BiLSTM	LRW	83.00
2017	Rahmani and Almasganj (Rahmani & Almasganj, 2017)	DBNF + DNN-HMM	CUAVE	64.90
2017	Petridis et al. (Petridis et al. 2017)	Autoencoder + LSTM	OuluVS2	94.70
2017	Petridis et al. (Petridis et al. 2018)	Autoencoder + LSTM	OuluVS2	91.80
2017	Petridis et al. (Petridis et al. 2017)	Autoencoder + LSTM	OuluVS2	84.50
2017	Fernandez – Lopez and Sukno Fernandez-Lopez & Sukno, 2017)	DCT + SIFT + LDA	VLRF	23.00
2017	Fernandez – Lopez et al. (Fernandez-Lopez et al. 2017)	DCT + SIFT + LDA	AVICAR	20.00
2017	Chung et al. (Chung et al. 2017)	CNN + LSTM	LRW GRID LRS	76.20 97.00 49.80
2017	Chung and Zisserman (Chung et al. 2017)	CNN + LSTM	OuluVS2 MV-LRS	91.10 43.60

As seen in Table 2, the architectures used in lip reading research and the datasets in which these architectures were applied are given.

In 2020, Xu et al. used the Pseudo-3D Residual Network (P3D) instead of using 3D-CNN or 2D-ResNet for a more accurate detection of the mouth region, recorded from the anterior view. They suggested that the P3D network is better at obtaining spatial-temporal properties within videos. They achieved the highest success on the LRS3-TED (Afouras et al. 2018) and LRW datasets with their proposed method. The LRW dataset is commonly used in lip-reading studies. The P3D study achieved an 84.8% success rate—the highest that has been achieved among studies with the LRW dataset (Xu et al. 2020).

Lu and Yan (2017) developed a hybrid model using CNN and Bidirectional Long-Short Term Memory (BiLSTM). The model was originally trained with CNN, visual features were extracted, and BiLSTM was used to evaluate sequential features among the video frames. This dataset includes 3 male and 3 female speakers; they were required to say the numbers between 0 and 9, in English. For classification accuracy, the Softmax function was chosen and produced an 85.7% accuracy rate. With the model they developed, they achieved better performance than traditional methods including the Hidden Markov Model (HMM) and the Active Contour Model (ACM) (Lu & Yan, 2020).

In 2020, Petridis et al. reached the highest rates of improvement in the datasets OuluVS2, CUAVE, AvLetters, and AvLetters2 at 0.6%; 3.4%; 3.9%; and 11.4%, respectively. The structure they created takes input from two streams; in one of the streams, the original and the differentiated inputs are taken. BiLSTM was used in both streams and when the streams were combined, BiLSTM was used again. The Softmax function was then applied to classify the outputs (Petridis et al. 2020).

Mamatha et al. used the LRW dataset in their 2020 study. They aimed to develop an application that converts movements of the mouth region into text for hearing-impaired individuals with the use of smartphone cameras. For this purpose, they designed an architecture connected to VGG19 by using CNN and LSTM together. The model they presented had an 88.2% accuracy rate, which is 3.3% higher than the success rate of CNN-RNN architecture (Mamatha et al. 2020).

In 2020, Martinez et al. achieved some of the largest successes in recent studies by making improvements on both the Residual Network (ResNet) and the Bidirectional Gated Recurrent Unit (BiGRU). They developed BiGRU with Temporal Convolutional Networks (TCN) and simplified the training phase of the model. When they tested the model in the LRW and the LRW-1000 datasets, they achieved success rates of 1.2% and 3.2%, respectively (Martinez et al. 2020).

Chen et al. (2020) used the LipNet and ResNet architectures in their study with the LRW-1000 dataset and produced improvements of 13.91% and 4.68%, respectively. Since the LRW-1000 dataset consists of sentences in the Mandarin dialect, the architecture was applied on the Chinese language. The network they created consists of two steps using 3D-CNN, DenseNet, 2D resBiLSTM, and a transform structure (Chen et al. 2020).

In 2020, Xiao et al. published their work on the LRW and LRW-1000 datasets, which achieved a high rate of success. In the study, they proposed a Deformation Flow Based Two-Stream Network (DFTN) to capture facial movements in videos. They performed self-supervised learning without the need for tagged data. In their models, they used both raw inputs and corrupted inputs. They used resNet-18 on the front layers and GRU on the back layers, while a 3-step method was used in the training phase. The Adam optimization algorithm was used with its default values and was reduced by half when the model converged; thus, reducing the problem of over-fitting. With this model, they reached an accuracy rate of 84.13% for LRW and 41.3% for LRW-1000 (Xiao et al. 2020).

The study conducted by Zhao et al. in 2020, utilized the datasets LRW, LRW-1000, LRS2, and LRS3. While creating their model, 3D-CNN and ResNet-18 were used on the front layers and BiGRU on the back layers. In this study, they proposed a new architecture in which the training phase of the model was easier. With their work, success was attained in LRW and LRW-1000 datasets (Zhao et al. 2020).

Luo et al. (2020) addressed lip-reading problems as a sequence to sequence issue (seq2seq). Thus, they aimed to convert speech sequences into a text sequence. They proposed a new evolutionary model for the two problems faced by the string-to-string method. When comparing the proposed model to other datasets, significant improvements were achieved. The accuracy rate of 83.5% for the LRW dataset and 38.70% for the LRW-1000 dataset was realized. For the GRID dataset, 11.2% word error rate was reached (Luo et al. 2020).

Feng et al. (2017) have proposed a Multimodal Recurrent Neural Networks (Multimodal RNN). They used LSTM and RNN for voice and CNN, LSTM, and RNN for lip-reading. They used AvLetters as the dataset and achieved a 57.7% success rate (Feng et al. 2017).

Mesbah et al., in 2019, introduced the Hahn Convolutional Neural Network (HCNN); a new architecture for lip-reading. They used Hahn moments—a mathematical method—in their architecture as the first steps within the CNN. The Hahn moments are an orthogonal moment set, based on Hahn polynomials which are defined on the image coordinate area. With the proposed architecture, HCNN has been shown to reduce video dimensions and training time. They tested the model in the AvLetters, OuluVS2, and LRW datasets that contain letters, numbers, and video images. A 20% improvement was achieved in the AvLetters dataset and better results were obtained from the OuluVS2 dataset than both GoogleNet and SyncNet. Architectures provide important solutions by occasionally removing basic and even useful features to ensure the effective classification of an image (Mesbah et al. 2019). In this study, the HCNN architecture was applied on the most commonly used datasets. The OuluVS2 dataset has been used more frequently in recent years. The usage rates of different datasets are shown in Figure 4.

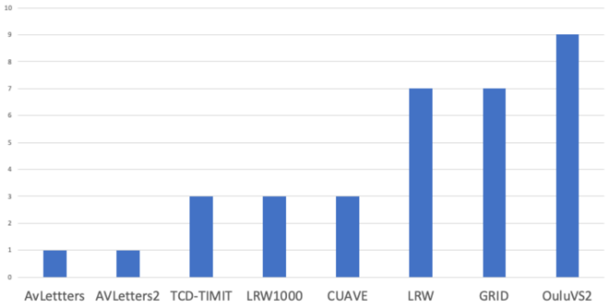

Figure 4. Usage rates of multiple datasets

Based on the commonly used datasets in Figure 4, the OuluVS2 dataset contains numbers, sentences, and expressions; the GRID dataset contains expressions; the LRW dataset contains words; and the CUAVE dataset contains numbers. The datasets can contain one of the sets (sentences, words, digits, expressions, or letters) (Matthews et al. 2002; Chung & Zisserman, 2016; Patterson et al. 2002) or multiple sets (Anina et al. 2015; Petridis et al. 2018). Table 3 delineates the content density of the datasets.

Table 3. Content table of dataset

	AvLettters	**AVLetters2**	**TCD-TIMIT**	**CUAVE**	**LRW**	**GRID**	**OuluVS2**
Sentence			1				1
Word					1		
Digit				1			1
Phrase						1	1
Letter	1	1					

When studies conducted in previous years are examined, it is seen that a Kinect camera is used for depth perception in datasets and research (Rekik et al. 2014; Yargıç & Doğan 2013; Wang et al. 2015). Equally as important as the perception of depth, the angle from which the image is captured is crucial for lip-reading. In 2012, Lan et al. pointed out that lip-reading studies usually include faces from an anterior view. Lip-reading experts stated that their preference when lip-reading is not in fact, an anterior view, but an angular view. Based on this incongruity, the researchers seek answers to two questions. First, is the anterior or the angular view more advantageous for the computer; second, from which angles can a computer still define words independently. From this perspective, they simultaneous recorded a speaker from 5 different angles. They used the LiLIR dataset, which contains images from 0, 90, 30, 45, and 60 degree angles. They concluded that lip-reading is actually easier from different angles than an anterior view. The study concluded that lips provide more information from side angles (Lan et al. 2012).

For studies conducted in the English language, Wand et al. (2016) emphasized that the pronunciation of some letters is extremely close to others. According to this study, the letters ‘p’ and ‘b’ have a similar letter structure as well as little recognizable distinctions (Wand et al. 2016). Although it is seen that the majority of research is in English, there are also studies in other languages such as Chinese (Martinez et al 2020; Chen et al. 2020; Xiao et al. 2020), Malay (Fook et al. 2012), Spanish (Fernandez-Lopez & Sukno, 2017) and Turkish (Yargıç & Doğan, 2013).

In 2013, Yargıç and Doğan created a Turkish dataset with a Kinect camera. Table 4 shows the words found in the dataset along with their English translation.

When creating this dataset, the Kinect camera was positioned 90 cm away from the speaker. The 10 volunteers were required to repeat these 15 words, 5 times. Thus, there is a total of 750 videos (10*15*5) in this dataset. By utilizing the Manhattan Distance, the researchers classified the data using the k-nearest neighbors classifier with the Manhattan and Euclidean Distance and found the best result as 77.8% (Yargıç & Doğan, 2013).

Table 4. Content word set of the dataset (Yargıç & Doğan, 2013)

Turkish	English
Beyaz	White
Bordo	Burgundy
Gri	Gray
Kahverengi	Brown
Kırmızı	Red
Lacivert	Navy blue
Mavi	Blue
Menekşe	Violet
Mor	Purple
Pembe	Pink
Sarı	Yellow
Siyah	Black
Turkuaz	Turquoise
Turuncu	Orange
Yeşil	Green

In the studies examined, it is seen that there are many studies using OuluVs data. In these studies, it is seen that the success is low when models such as CNN are used alone (Koumparoulis & Potamianos, 2018). It is observed that success increases when hybrid models are preferred instead of stand-alone studies (Jang et al. 2019, Petridis et al. 2017, Chung et al. 2017). For the GRID dataset, 7 studies were examined. When examined in these studies, the study in which CNN – LSTM architecture is used stands out with its success rate (Chung et al. 2017). It is seen that the architecture used alone (Wand & Schmidhuber, 2017) has bad results as in the OuluVs dataset (Koumparoulis & Potamianos, 2018). It is seen that hybrid models have a high success rate in the GRID dataset (Chung et al. 2017, Xu et al. 2018). In this context, it is seen that hybrid models give more successful results. In challenging datasets such as the LRW1000 dataset, it is seen to be low in hybrid models (Bi et al. 2019, Xiao et al. 2020). Real-life data sets are more challenging than other data sets. Better architectures need to be developed on these data sets.

In the studies, it was observed that no significant improvements were observed in the datasets using the 3DCNN architecture in hybrid models. The success rate in architecture using CNN – LSTM on the LRS dataset (Chung et al. 2017) is 49.80, while the success rate in architecture using 3D-CNN + BiLSTM (Afouras et al. 2018) is 50.00%. While the success rate is 91.10% in the CNN – LSTM architecture used on another data set OuluVS2, the success rate is 87.60% in the architecture using 3D-CNN + BiLSTM (Fung & Mak, 2018). In this context, it cannot be said that the 3DCNN architecture, which enables to use the spatio-temporal structure for lip-reading studies, will always increase success.

In recent years, it has been observed that there has been a tendency to LRW, LRW1000 datasets, which are more challenging datasets in recent years (Martinez et al. 2020, Chen et al. 2020, Xu et al. 2020). It is seen that the success rates in previous datasets have now reached very high levels (Xu et al. 2018, Chung et al. 2017, Petridis et al. 2017). In this context, it can be said that academic studies on lip reading are progressing to develop new technologies that can be adapted to real life.

7. Conclusion

In deep learning studies, utilizing different optimizations, hybrid structures, and new methods—machines are approaching human capabilities in lip-reading. As a result of these developments, there are also some specific situations where machines have surpassed human abilities. While it is widely acknowledged that it takes many years to become an expert in any subject, machines can easily achieve similar levels as long as they are provided with good data and good models.

Lip-reading is a subject that can be specialized in as a result of many years of experience. This expertise, which is based entirely on visual perception, provides benefits to the hearing impaired and could be used in other various fields as machines continue to advance. The study of lip-reading can also improve the results of voice recognition devices.

In this article, research conducted between 2017 to 2020 on lip-reading through deep learning methods were examined and information about the methods and datasets used was presented. Apart from deep learning methods, it has been seen in recent years that research has also been carried out with traditional models in the Hidden Markov Model (HMM) (Bear & Harvey, 2017).

Recently, the preference for lip-reading research has been methods based on CNN and LSTM architectures. These methods, which have a high contribution to success rate including RNN and LSTM, are used in studies which emphasize that lip-reading should be evaluated with the use of time series methods. The emergence of hybrid methods led to the differentiation of research. The variety of hybrid architectures increased as a result of numerous deep learning studies. While architectures grow in diversity, costs must also be taken into consideration. In light of the studies examined, it was noted that CNN and LSTM were used intensively in hybrid structures in the past few years; however, it became clear that this trend has shifted towards 3D-CNN and BiLSTM in more recent years. It cannot be said that the 3DCNN architecture, which enables to use the spatio-temporal structure in the context of studies, will always increase success. Again, in the studies examined, it is seen that there is a tendency towards datasets in which examples from the real world are formed in recent years, and it is seen that hybrid studies are intensified on these datasets. It is seen that the data set presenting real-world sections for new lip reading studies will contribute more to the literature. Hybrid models and high achievements can pave the way for technological tools that will be beneficial to human life.

With an increase the in costs of new architectures, difficulties may occur in real-time studies. Lip-reading in real-time can increase its functionality for those who rely on it as a form of communication. Therefore, real-time results can be improved with hybrid structures that are used with deep learning techniques such as Yolo. Combining lip-reading and voice recognition techniques in real-time could also greatly advance human-computer interactions.

While issues such as determining the movements of the mouth region were primarily important in early lip-reading studies, the issue concerning current researchers today is the recognition of speech. Problems such as mouth region detection are now largely overcome in computer vision studies. Yet, it is seen that some letters and words cause difficulties during the definition of speech. As soon as those words and letters are used, lip movements cannot be detected and the sequence is distorted. By conducting a separate study for difficult words and letters, movement, speed, and sound bursts can be analysed. There are also research that demonstrate how the emotional state is related to the verbal expression of the speaker. Hybrid studies could be created where emotional states are taken into account.

In the studies examined, compared to the datasets that have been created, datasets should contain more samples as well as including samples from real-life situations. More specifically, the datasets which were created in front of a green screen and those created only displaying the mouth region are far from what would be encountered in daily life. The correct determination of the mouth region is the first stage of study and is very successful when frames are without complexity. However, in complex frames, algorithms that can correctly detect the mouth region could be improved and real-life difficulties could be overcome.

Currently, the majority of lip-reading studies have been conducted in the English language, while studies in the Turkish language are limited. These English studies cannot be directly applied to Turkish due to their linguistic differences and structures. It is recommended that Turkish-specific lip-reading studies with voice recognition be conducted. There is a need for lip-reading techniques—in the Turkish language—to be studied academically.

Referanslar

Adeel, A., Gogate, M., & Hussain, A. (2020). Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. Information Fusion, 59, 163-170.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). Deep lip reading: a comparison of models and an online application. arXiv preprint arXiv:1806.06053.

Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496.

Akmese Ö.F., Erbay H., Kör H., (2019). Derin Ögrenme ile Görüntü Kümeleme. In: 5th International Management Information Systems Conference, Ankara.

Alpaydin, E. (2020). Introduction to machine learning. MIT press.

Amanullah, M. A., Habeeb, R. A. A., Nasaruddin, F. H., Gani, A., Ahmed, E., Nainar, A. S. M., ... & Imran, M. (2020). Deep learning and big data technologies for IoT security. Computer Communications, 151, 495-517.

Anina, I., Zhou, Z., Zhao, G., & Pietikäinen, M. (2015, May). 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) (Vol. 1, pp. 1-5). IEEE.

Arı, A., & Hanbay, D. (2019). Tumor detection in MR images of regional convolutional neural networks. Journal of the Faculty of Engineering and Architecture of Gazi University, 34(3), 1395-1408.

Bacciu, D., Micheli, A., & Podda, M. (2020). Edge-based sequential graph generation with recurrent neural networks. Neurocomputing, 416, 177-189.

Bayram, F. (2020). Derin öğrenme tabanlı otomatik plaka tanıma. Politeknik Dergisi, 23(4), 955-960.

Bear, H. L., & Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. Speech Communication, 95, 40-67.

Bi, C., Zhang, D., Yang, L., & Chen, P. (2019, November). An Lipreading Modle with DenseNet and E3D-LSTM. In 2019 6th International Conference on Systems and Informatics (ICSAI) (pp. 511-515). IEEE.

Bollier, D. (2017). Artificial intelligence comes of age. The promise and challenge of integrating AI into cars, healthcare and journalism. The Aspen Institute Communications and Society Program. Washington, DC.

Chen, L., Xu, G., Zhang, S., Yan, W., & Wu, Q. (2020). Health indicator construction of machinery based on end-to-end trainable convolution recurrent neural networks. Journal of Manufacturing Systems, 54, 1-11.

Chen, X., Du, J., & Zhang, H. (2020). Lipreading with DenseNet and resBi-LSTM. Signal, Image and Video Processing, 14(5), 981-989.

Chen, Y., Zhao, X., & Jia, X. (2015). Spectral–spatial classification of hyperspectral data based on deep belief network. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(6), 2381-2392.

Cheok, M. J., Omar, Z., & Jaward, M. H. (2019). A review of hand gesture and sign language recognition techniques. International Journal of Machine Learning and Cybernetics, 10(1), 131-153.

Chung, J. S., & Zisserman, A. (2016, November). Lip reading in the wild. In Asian conference on computer vision (pp. 87-103). Springer, Cham.

Chung, J. S., & Zisserman, A. P. (2017). Lip reading in profile.

Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017, July). Lip reading sentences in the wild. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3444-3453). IEEE.

Cooke, M., Barker, J., Cunningham, S., & Shao, X. (2006). An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5), 2421-2424.

Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., & Theobald, B. J. (2008, September). The challenge of multispeaker lip-reading. In AVSP (pp. 179-184).

Doğan, M., Nemli, O. N., Yüksel, O. M., Bayramoğlu, İ., & Kemaloğlu, Y. K. (2008). İşitme Kaybının Yaşam Kalitesine Etkisini İnceleyen Anket Çalışmalarına Ait Bir Derleme. Turkiye Klinikleri J Int Med Sci, 4, 33.

Dupont, S., & Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. IEEE transactions on multimedia, 2(3), 141-151.

Erdoğan A.A., (2016). Hearing Loss and Approaches to Hearing Loss in Elderly, The Turkish Journal of Family Medicine and Primary Care, 10 (1): 25-33, (2016). doi:10.5455/tjfmpc.204524

Ergezer, H., Dikmen, M., & Özdemir, E. (2003). Yapay sinir ağları ve tanıma sistemleri. PiVOLKA, 2(6), 14-17.

Ertam, F., & Aydın, G. (2017, October). Data classification with deep learning using Tensorflow. In 2017 international conference on computer science and engineering (UBMK) (pp. 755-758). IEEE.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. Nature medicine, 25(1), 24-29.

Farsal, W., Anter, S., & Ramdani, M. (2018, October). Deep learning: An overview. In Proceedings of the 12th International Conference on Intelligent Systems: Theories and Applications (pp. 1-6).

Fayjie, A. R., Hossain, S., Oualid, D., & Lee, D. J. (2018, June). Driverless car: Autonomous driving using deep reinforcement learning in urban environment. In 2018 15th International Conference on Ubiquitous Robots (UR) (pp. 896-901). IEEE.

Feng, W., Guan, N., Li, Y., Zhang, X., & Luo, Z. (2017, May). Audio visual speech recognition with multimodal recurrent neural networks. In 2017 International Joint Conference on Neural Networks (IJCNN) (pp. 681-688). IEEE.

Fernandez-Lopez, A., & Sukno, F. M. (2017). Automatic viseme vocabulary construction to enhance continuous lip-reading. arXiv preprint arXiv:1704.08035.

Fernandez-Lopez, A., & Sukno, F. M. (2017, February). Optimizing Phoneme-to-Viseme Mapping for Continuous Lip-Reading in Spanish. In International Joint Conference on Computer Vision, Imaging and Computer Graphics (pp. 305-328). Springer, Cham.

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 78, 53-72.

Fernandez-Lopez, A., Martinez, O., & Sukno, F. M. (2017, May). 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) (pp. 208-215). IEEE.

Fook, C. Y., Hariharan, M., Yaacob, S., & Adom, A. H. (2012, February). A review: Malay speech recognition and audio visual speech recognition. In 2012 International Conference on Biomedical Engineering (ICoBE) (pp. 479-484). IEEE.

Fung, I., & Mak, B. (2018, April). End-to-end low-resource lip-reading with maxout CNN and LSTM. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2511-2515). IEEE.

Gogate, M., Dashtipour, K., Adeel, A., & Hussain, A. (2020). CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement. Information Fusion, 63, 273-285.

Goh, Y. H., Lau, K. X., & Lee, Y. K. (2019, October). Audio-Visual Speech Recognition System Using Recurrent Neural Network. In 2019 4th International Conference on Information Technology (InCIT) (pp. 38-43). IEEE.

Hamurcu, M., Şener, B. M., Ataş, A., Atalay, R. B., Bora, F., & Yiğit, Ö. (2012). İşitme cihazı kullanan hastalarda memnuniyetin değerlendirilmesi.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. Proceedings of the national academy of sciences, 81(10), 3088-3092.

Jang, D. W., Kim, H. I., Je, C., Park, R. H., & Park, H. M. (2019). Lip reading using committee networks with two different types of concatenated frame images. IEEE Access, 7, 90125-90131.

Kahveci, O. K., Miman, M. C., Okur, E., Ayçiçek, A., Sevinç, S., & Altuntaş, A. (2011). Hearing aid use and patient satisfaction. Kulak burun bogaz ihtisas dergisi: KBB= Journal of ear, nose, and throat, 21(3), 117-121.

Keyvanrad, M. A., & Homayounpour, M. M. (2014). A brief survey on deep belief networks and introducing a new object oriented toolbox (DeeBNet). arXiv preprint arXiv:1408.3264.

Koumparoulis, A., & Potamianos, G. (2018, December). Deep view2view mapping for view-invariant lipreading. In 2018 IEEE Spoken Language Technology Workshop (SLT) (pp. 588-594). IEEE.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

Kumar, Y., Jain, R., Salik, M., ratn Shah, R., Zimmermann, R., & Yin, Y. (2018, December). Mylipper: A personalized system for speech reconstruction using multi-view visual feeds. In 2018 IEEE International Symposium on Multimedia (ISM) (pp. 159-166). IEEE.

Lan, Y., Theobald, B. J., & Harvey, R. (2012, July). View independent computer lip-reading. In 2012 IEEE International Conference on Multimedia and Expo (pp. 432-437). IEEE.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

Li, X., Neil, D., Delbruck, T., & Liu, S. C. (2019, May). Lip reading deep network exploiting multi-modal spiking visual and auditory sensors. In 2019 IEEE International Symposium on Circuits and Systems (ISCAS) (pp. 1-5). IEEE.

Lu, Y., & Yan, J. (2020). Automatic lip reading using convolution neural network and bidirectional long short-term memory. International Journal of Pattern Recognition and Artificial Intelligence, 34(01), 2054003.

Luo, M., Yang, S., Shan, S., & Chen, X. (2020, November). Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 273-280). IEEE.

Lv, Z., & Qiao, L. (2020). Deep belief network and linear perceptron based cognitive computing for collaborative robots. Applied Soft Computing, 92, 106300.

Mamatha G., Roshan B.B.R., Vasudha S.R., (2020). Lip Reading to Text using Artificial Intelligence, International Journal of Engineering Research & Technology (IJERT), 9 (01): 483-484.

Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020, May). Lipreading using temporal convolutional networks. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6319-6323). IEEE.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2), 198-213.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

Mesbah, A., Berrahou, A., Hammouchi, H., Berbia, H., Qjidaa, H., & Daoudi, M. (2019). Lip reading with Hahn convolutional neural networks. Image and Vision Computing, 88, 76-83.

Minsky, M., & Papert, S. (1969). An introduction to computational geometry. Cambridge tiass., HIT.

Muljono, M., Saraswati, G., Winarsih, N., Rokhman, N., Supriyanto, C., & Pujiono, P. (2019). Developing BacaBicara: An Indonesian Lipreading System as an Independent Communication Learning for the Deaf and Hard-of-Hearing. International Journal of Emerging Technologies in Learning (iJET), 14(4), 44-57.

Mulrow, C. D., Aguilar, C., Endicott, J. E., Tuley, M. R., Velez, R., Charlip, W. S., ... & DeNino, L. A. (1990). Quality-of-life changes and hearing impairment: a randomized trial. Annals of internal medicine, 113(3), 188-194.

Mulrow, C. D., Aguilar, C., Endicott, J. E., Velez, R., Tuley, M. R., Charlip, W. S., & Hill, J. A. (1990). Association between hearing impairment and the quality of life of elderly individuals. Journal of the American Geriatrics Society, 38(1), 45-50.

Mulrow, C. D., Tuley, M. R., & Aguilar, C. (1992). Sustained benefits of hearing aids. Journal of Speech, Language, and Hearing Research, 35(6), 1402-1405.

Oliveira, D. A. B., Mattos, A. B., & da Silva Morais, E. (2019, May). Improving Viseme Recognition with GAN-based Muti-view Mapping. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019) (pp. 1-8). IEEE.

Olgun, N., Aslan, F. E., Yücel, N., Öntürk, Z. K., & Laçin, Z. (2013). Yaşlıların sağlık durumlarının değerlendirilmesi. Acıbadem Üniversitesi Sağlık Bilimleri Dergisi, (2), 72-78.

Ozcan, T., & Basturk, A. (2019). Lip reading using convolutional neural networks with and without pre-trained models. Balkan Journal of Electrical and Computer Engineering, 7(2), 195-201.

Pang, Z., Niu, F., & O'Neill, Z. (2020). Solar radiation prediction using recurrent neural network and artificial neural network: A case study with comparisons. Renewable Energy, 156, 279-289.

Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002). Moving-talker, speaker-independent feature study, and baseline results using the CUAVE multimodal speech corpus. EURASIP Journal on Advances in Signal Processing, 2002(11), 1-13.

Patterson, E. K., Gurbuz, S., Tufekci, Z., & Gowdy, J. N. (2002, May). CUAVE: A new audio-visual database for multimodal human-computer interface research. In 2002 IEEE International conference on acoustics, speech, and signal processing (Vol. 2, pp. II-2017). IEEE.

Petridis, S., Li, Z., & Pantic, M. (2017, March). End-to-end visual speech recognition with LSTMs. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2592-2596). IEEE.

Petridis, S., Shen, J., Cetin, D., & Pantic, M. (2018, April). Visual-only recognition of normal, whispered and silent speech. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6219-6223). IEEE.

Petridis, S., Stafylakis, T., Ma, P., Cai, F., Tzimiropoulos, G., & Pantic, M. (2018, April). End-to-end audiovisual speech recognition. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 6548-6552). IEEE.

Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end audiovisual fusion with LSTMs. arXiv preprint arXiv:1709.04343.

Petridis, S., Wang, Y., Li, Z., & Pantic, M. (2017). End-to-end multi-view lipreading. arXiv preprint arXiv:1709.00443.

Petridis, S., Wang, Y., Ma, P., Li, Z., & Pantic, M. (2020). End-to-end visual speech recognition for small-scale datasets. Pattern Recognition Letters, 131, 421-427.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91(9), 1306-1326.

Potamianos, G., Neti, C., Luettin, J., & Matthews, I. (2004). Audio-visual automatic speech recognition: An overview. Issues in visual and audio-visual speech processing, 22, 23.

Qu, L., Weber, C., & Wermter, S. (2019, September). LipSound: Neural Mel-Spectrogram Reconstruction for Lip Reading. In INTERSPEECH (pp. 2768-2772).

Rahmani, M. H., & Almasganj, F. (2017, April). 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) (pp. 195-199). IEEE.

Rekik, A., Ben-Hamadou, A., & Mahdi, W. (2014, October). A new visual speech recognition approach for RGB-D cameras. In International conference image analysis and recognition (pp. 21-28). Springer, Cham.

Rosenbaltt, F. (1957). The perceptron–a perciving and recognizing automation. Cornell Aeronautical Laboratory.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Russell, S. J., & Norvig, P. Artificial intelligence: a modern approach. 2016: Malaysia.

Saif, D., El-Gokhy, S. M., & Sallam, E. (2018). Deep Belief Networks-based framework for malware detection in Android systems. Alexandria engineering journal, 57(4), 4049-4057.

Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

Sam, S. M., Kamardin, K., Sjarif, N. N. A., & Mohamed, N. (2019). Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet Inception-v1 and Inception-v3. Procedia Computer Science, 161, 475-483.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Sindhura, P. V., Preethi, S. J., & Niranjana, K. B. (2018, December). Convolutional neural networks for predicting words: A lip-reading system. In 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) (pp. 929-933). IEEE.

Stafylakis, T., & Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105.

Sui, C., Togneri, R., & Bennamoun, M. (2017). A cascade gray-stereo visual feature extraction method for visual and audio-visual speech recognition. Speech Communication, 90, 26-38.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Thangthai, K., & Harvey, R. (2017, August). Improving computer lipreading via DNN sequence discriminative training techniques. ISCA.

Thangthai, K., Bear, H. L., & Harvey, R. (2018). Comparing phonemes and visemes with DNN-based lipreading. arXiv preprint arXiv:1805.02924.

Turing A.M., “Computing Machinery and Intelligence”, Mind Journal, 49: 433-460, (1950).

Uğur, A., & Kınacı, A. C. (2006). Yapay zeka teknikleri ve yapay sinir ağları kullanılarak web sayfalarının sınıflandırılması. XI. Türkiye'de İnternet Konferansı (inet-tr'06), Ankara, 1-4.

Wand, M., & Schmidhuber, J. (2017). Improving speaker-independent lipreading with domain-adversarial training. arXiv preprint arXiv:1708.01565.

Wand, M., Koutník, J., & Schmidhuber, J. (2016, March). Lipreading with long short-term memory. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6115-6119). IEEE.

Wand, M., Schmidhuber, J., & Vu, N. T. (2018, April). Investigations on end-to-end audiovisual fusion. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3041-3045). IEEE.

Wang, J., Gao, Y., Zhang, J., Wei, J., & Dang, J. (2015). Lipreading using profile lips rebuilt by 3D data from the Kinect. Journal of Computational Information Systems, 11(7), 2429-2438.

Xiao, J., Yang, S., Zhang, Y., Shan, S., & Chen, X. (2020, November). Deformation flow based two-stream network for lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 364-370). IEEE.

Xu, B., Wang, J., Lu, C., & Guo, Y. (2020). Watch to listen clearly: Visual speech enhancement driven multi-modality speech recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1637-1646).

Xu, K., Li, D., Cassimatis, N., & Wang, X. (2018, May). LCANet: End-to-end lipreading with cascaded attention-CTC. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) (pp. 548-555). IEEE.

Yang, R., Singh, S. K., Tavakkoli, M., Amiri, N., Yang, Y., Karami, M. A., & Rai, R. (2020). CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. Mechanical Systems and signal processing, 144, 106885.

Yang, S., Zhang, Y., Feng, D., Yang, M., Wang, C., Xiao, J., ... & Chen, X. (2019, May). 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) (pp. 1-8). IEEE.

Yargıç, A., & Doğan, M. (2013, June). A lip reading application on MS Kinect camera. In 2013 IEEE INISTA (pp. 1-5). IEEE.

Yu, Y., Hu, C., Si, X., Zheng, J., & Zhang, J. (2020). Averaged Bi-LSTM networks for RUL prognostics with non-life-cycle labeled dataset. Neurocomputing, 402, 134-147.

Yueh, B., Shapiro, N., MacLean, C. H., & Shekelle, P. G. (2003). Screening and management of adult hearing loss in primary care: scientific review. Jama, 289(15), 1976-1985.

Zhao, X., Yang, S., Shan, S., & Chen, X. (2020, November). Mutual information maximization for effective lip reading. In 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (pp. 420-427). IEEE.

Zhou, P., Yang, W., Chen, W., Wang, Y., & Jia, J. (2019, May). Modality attention for end-to-end audio-visual speech recognition. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6565-6569). IEEE.

Zhou, Z., Zhao, G., Hong, X., & Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. Image and vision computing, 32(9), 590-605.