POPULAR AND DEMOCRATIC REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY MOHAMED BOUDIAF - M'SILA
FACULTY OF TECHNOLOGY

ELECTRONICS DEPARTMENT
DOMAIN : SCIENCE AND TECHNOLOGY
FILIERE: ELECTRONICS
OPTION : EMBEDDED SYSTEMS

ELECTRICAL ENGINEERING DEPARTMENT
DOMAIN : SCIENCE AND TECHNOLOGY
FILIERE: AUTOMATIC
OPTION : ROBOTICS

Dissertation Submitted in partial fulfilment of the requirements
For the Master Academic Degree
By:

BARKA Riyadh

&

FERHAT Zeyd

Entitled

Automatic System for Lip Reading Transcription

Diplôme de Master dans le cadre du décret ministériel 1275

Presented on: June 12th, 2024, in front of the jury composed of :

Pr. BOURAS Mounir	University of Mohamed Boudiaf - M'sila	President
Dr. KHENNOUF Salah	University of Mohamed Boudiaf - M'sila	Supervisor
Dr. BENYOUNES Abdelhafid	University of Mohamed Boudiaf - M'sila	Examiner
Dr. BRIK Youcef	University of Mohamed Boudiaf - M'sila	CATI Representative
Dr. ATTALAH Billal	University of Mohamed Boudiaf - M'sila	INCUBATEUR Representative
Dr. BRIKI Zineb	Directorate of Social Action and Solidarity.	Socio-Economic Partner

Academic year : 2023 / 2024

Acknowledgments

The completion of this thesis was made possible through the support of several individuals to whom we would like to express our gratitude.

First and foremost, this work would not be as rich and would not have come to fruition without the help and guidance of Dr. Salah KHENNOUF, whom we thank for his exceptional supervision, patience, rigor, and availability during the preparation of this thesis.

We also wish to thank the head of the electronics department at the University of M'sila, Dr. TABBAKH Mostefa, and the head of the electrical engineering department, Dr. ZEGHLACHE Samir, for their invaluable support.

Our sincere thanks go to all the professors, contributors, and everyone who, through their words, writings, advice, and critiques, guided our reflections and agreed to meet with us and answer our questions during our research.

Dedication

First of all, I thank Almighty God for giving me the courage, will, and patience to carry out this work despite all the difficulties I encountered.

*To my beloved mother **Fatima**, who endured countless sleepless nights and weary days, sacrificing her own comfort for mine since childhood, your unwavering love and support have shaped me into who I am today.*

*To my dear father **Hachemi**,* for all the sacrifices you made and the hardships you endured, your unwavering support and guidance have been a constant source of inspiration for me.

*To my siblings, **Wissam**, **Djamila**, and **Youcef**, you have been the best siblings one could ever ask for, always there with unwavering support and love.*

*to my entire family, **Ferhat** and **Lounas***

To all my friends, for your unwavering support and delightful company, thank you for being there for me.

To the serene adventures, radiance surrounds, resonating in my life.

*To my colleague in this project, **Riyadh**.*

Zeyd

Dedication

First of all, I thank Almighty God for giving me the courage, will, and patience to carry out this work despite all the difficulties I encountered. Indeed, the path of God is good.

*To my dear mother, **Salima**, for the sacrifices she made from my childhood until today;*

*To my dear father, **Bendjdou**, who has always supported and helped me face challenges, and who has always been and continues to be a role model for me.*

*To my siblings, **Salah**, **Saber**, **Aisha**, and **Araj**, for being a refuge and a shelter from life's worries and problems.*

to my entire family, Barka and Ouali;

To all my brothers not born of my mother, Habib, Rayan, Haider, Lina, Haythem, Abbad, Chourouk, Isra, Nour, Choukri, for their support and their good and pleasant company in life.

*To my colleague in this project, **Zeyd**.*
To all the teachers in the Department of Electrical Engineering;

To all the students.

Riyadh

List of Abbreviations

3DCNN :	3D Convolutional Neural Network
ASR :	Automatic Speech Recognition
CNNs :	Convolutional Neural Networks
Conv3D :	3D Convolution
DCT :	Discrete Cosine Transform
GND :	Ground
HDMI :	High-Definition Multimedia Interface
LAN :	Local Area Network
LR :	Lip Reading
LRT :	Lip Reading Transcription
LSTM :	Long Short-Term Memory
MFCCs :	Mel-Frequency Cepstral Coefficients
NPY :	NumPy
PCA :	Principal Component Analysis
PC :	Personal Computer
RAM :	Random-access memory
ReLU :	Rectified Linear Unit
RNNs :	Recurrent Neural Networks
ROC :	Receiver Operating Characteristic
SB-ASR :	Sound-based Automatic Speech Recognition
SWR :	True Wireless Stereo
VB-ASR :	Vision-based Automatic Speech Recognition

List of Figures

N° Figure	Titre	Page
Figure 1.1 :	Sign language	09
Figure 1.2 :	Cochlear implants	10
Figure 1.3 :	LSTM architecture	12
Figure 2.1 :	Count of each Arabic word label	20
Figure 2.2 :	Data collection	21
Figure 2.3 :	Count plot of English lables	21
Figure 2.4 :	Images from every video lables	22
Figure 2.5 :	Flowchart of database	23
Figure 2.6 :	Dlib lip tracking	25
Figure 2.7 :	3D CNN Architecture	28
Figure 3.1 :	Raspberry Pi 4 Model B General Structure	31
Figure 3.2 :	GPIO and 40-pin header	33
Figure 3.3 :	Bone Conduction Earphones	34
Figure 3.4 :	Flowchart of work	35
Figure 3.5 :	Distribution of Classes in Training and Test Sets	36
Figure 3.6 :	Training and Validation Loss and Accuracy Curve for Experiment-1	39
Figure 3.7 :	Connecting layers used in model composition	42
Figure 3.8 :	Training and Validation Loss and Accuracy Curve	44
Figure 3.9 :	Confusion Matrix	46
Figure 3.10 :	ROC and AUC Curve	47

List of Tables

N° table	Titre	Page
Table 3.1 :	Characteristics and features for Raspberry Pi 4	32
Table 3.2 :	Model structure components	37
Table 3.3 :	Hyperparameters for Experiment 1	38
Table 3.4 :	Results for Experiment 1	38
Table 3.5 :	Model structure components for Experiment 2	40
Table 3.6 :	Hyperparameters for Experiment 2	40
Table 3.7 :	Results for Experiment 2	41
Table 3.8 :	Model structure components	42
Table 3.9 :	Hyperparameters for the Model	43
Table 3.10 :	Final results for Each Class	45

Table of Contents

Acknowledgment i

Dedication ii

List of Abbreviations, Figures and Tables iii

INTRODUCTION 2

CHAPTER-1 :*GENERALITIES ABOUT LIP-READING SYSTEMS AND TRANSCRIPTION METHODS*

1.1 Introduction 5

1.2 Overview of Automatic Speech Recognition (ASR) 5

1.2.1 Sound-Based Automatic Speech Recognition (SB-ASR) 6

1.2.2 Vision-Based Automatic Speech Recognition (VB-ASR) 7

1.3 Evolution of Lip-Reading Technology 7

1.3.1 Historical background 8

1.3.2 Lip-Reading from Manual to Automatic Systems 10

1.3.3 Key advancements and breakthroughs in the field 11

1.4 Lip-Reading (LR) as an alternative method to speech recognition 12

1.4.1 Lip-Reading in helping communication for deaf people 13

1.4.2 Lip-Reading in Assistive Technologies 14

1.5 Lip-Reading Transcription 14

1.5.1 Methods of Lip-Reading Transcription: 15

1.5.2 Applications of Lip Reading Transcription (LRT) 15

1.5.3 Challenges & limitations of current (LRT) methods 17

1.6 Conclusion 17

CHAPTER-2 : CONCIEVED DATABASES AND PROPOSED METHODS

2.1 Introduction 19

2.2 Database Collection, Cleaning, and Parameters 19
2.2.1 Data Collection Techniques 19
2.2.2.1 First Dataset 21
2.2.2.2 Second Dataset 22
2.2.2 Data Cleaning Processes 22
2.2.3 Considered Parameters and their Significance 23
2.3 Used Materials 24
2.3.1 Laptop Camera 24
2.3.2 Laptop 24
2.4 Used Libraries 25
2.4.1 Ddlib 25
2.4.2 Opencv 26
2.5 3D Convolution Operation 27
2.5.1 3D Convolutional Kernels 27
2.5.2 3D CNN Architecture 27
2.5.3 3D Convolutional Neural Networks : Classification 28
2.6 Conclusion 28

CHAPTER-3 : PROPOSED AND REALIZED SYSTEM FOR LIP READING TRANSCRIPTIONS

3.1 Introduction 30
3.2 Used Materials 30
3.2.1 Raspberry Pi 4 Model B 30
3.2.2 Input and output 33
3.2.3 Bone Conduction Earphones 33
3.3 Flowchart and Workflow Overview 35
3.3.1 Load Data 35
3.3.2 Data Preprocessing 36
3.4 Experimental work and Results 37
3.4.1 Experiment 1: Using the Arabic data set 37
3.4.2 Results of Experiment 1 38

3.4.3 Experiment 2: Use Different Structures with Final Dataset 40
3.4.4 Experiment 3: Using Final Model Architecture and Results 41
3.5 Results of Experiment 3 44
3.6 Conclusion 48
CONCLUSION 50
RERERENCES 53

INTRODUCTION

INTRODUCTION

In recent years, advancements in artificial intelligence (AI) and computer vision have led to remarkable progress in the field of lip reading. Lip reading, also known as speech reading, is the practice of understanding speech by observing the movements of a speaker's lips. Traditionally, lip reading has been a crucial skill for individuals with hearing impairments, enabling them to adapt and engage with the world around them.

However, with the advance of AI technologies, the scope and potential of lip reading have transcended its initial boundaries. Today, lip reading is not merely a tool for the hearing impaired; it has found applications in diverse domains such as surveillance, human-computer interaction, and even forensic analysis. The ability to decode spoken words solely through visual cues opens doors to many possibilities, promising enhanced communication, security, and accessibility.[1]

At the heart of this transformative shift are deep learning techniques, particularly the use of three-dimensional Convolutional Neural Networks (3D CNN). These advanced neural networks are capable of extracting complex temporal and spatial features from video sequences, enabling more accurate and robust lip reading systems than ever before. By joining the power of deep learning, researchers aim to bridge the gap between visual input captured through lip movements and textual output, thereby revolutionizing the way we interact with and understand spoken language.[2]

In this study, we investigate into the details of lip-reading systems and transcription methods, exploring the fusion of AI, computer vision, and linguistics to unlock the potential of visual speech recognition. .

As we embark on this journey into the realm of visual speech recognition, we envision a future where words need not be heard to be understood a future where the power of AI enables seamless communication and empowers individuals, regardless of their hearing abilities. [3]

In this project, we aimed to develop a smart system that turns lip movements into text and sound using advanced artificial intelligence and deep learning techniques. The importance of this work is to promote communication for deaf or hearing-impaired individuals and to provide them with an effective tool to improve integration into society.

This document is set up as follows : The first chapter provides a exploring the theoretical foundations of speech recognition, lip-reading techniques, and examining developments and limitations in current systems. This includes a detailed overview of the algorithms and deep learning models used in these technologies, paving the way for development.

In the second chapter, we delve into the design and development phase of our proposed databases. This involves outlining the structure of the databases, as well as the cleaning and preprocessing steps taken, along with the different databases collected.

Finally, we delve into the design and development phase of our proposed system, describing the integration of hardware components with software solutions that use deep learning in real-time and voice conversion from lip movements. This phase also covers the different stages of prototype testing and redundancy to ensure the system's reliability and efficiency we also present the results of our experimental studies, including analysis of the system's performance and accuracy in interpreting lip movements.

CHAPTER-1

GENERALITIES ABOUT LIP-READING SYSTEMS AND TRANSCRIPTION METHODS

CHAPTER-1

GENERALITIES ABOUT LIP-READING SYSTEMS AND TRANSCRIPTION METHODS

1.1 Introduction

In this chapter, we investigate into the transformative technologies of Automatic Speech Recognition (ASR) and Lip-Reading (LR), exploring their evolution, challenges, applications, and breakthroughs. The ASR, also known as speech-to-text or computer speech recognition, has revolutionized human-machine interactions, while Lip-Reading plays a crucial role in aiding individuals with hearing impairments.

We will navigate through the foundational principles of Sound-Based ASR (SB-ASR), the emerging edge of Vision-Based ASR (VB-ASR), and the advancements in Lip-Reading Transcription (LRT). Through this journey, we will uncover the complexities, potentials, and ethical considerations surrounding these pivotal communication technologies.

1.2 Overview of Automatic Speech Recognition (ASR)

The ASR systems operate by taking an audio input, typically in the form of a wave file or microphone recording, and processing it through several key stages. Initially, ASR systems extract relevant acoustic features such as Mel-frequency Cepstral Coefficients (MFCCs), which represent the spectral characteristics of speech. These features are then passed through statistical models during the acoustic modeling phase, mapping the features to phonemes or words. Language modeling further refines the output by considering the grammatical structure and probabilities of word sequences.

The advancement of deep learning, particularly through models like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has significantly enhanced ASR accuracy. These models enable simultaneous learning of feature extraction and acoustic modeling, leading to more accurate recognition results. More

recently, architectures such as Transformers and end-to-end models have also been applied, pushing the boundaries of ASR performance even further.

ASR technology finds broad applications across various domains, including voice assistants (e.g., Siri, Alexa, Google Assistant), dictation software, and automated captioning, demonstrating its versatile utility in modern-day communication and accessibility efforts.

1.2.1 Sound-Based Automatic Speech Recognition (SB-ASR)

Sound-Based Automatic Speech Recognition (SB-ASR) forms the foundational technology underpinning modern Automatic Speech Recognition (ASR) systems prevalent today. It functions by leveraging acoustic cues within audio signals to transcribe spoken words into textual form. This technology centers on the extraction of crucial features from speech audio, such as Mel-frequency cepstral coefficients (MFCCs) or spectrogram representations, which encapsulate the spectral evolution of speech sounds over time. These features encapsulate valuable data pertaining to vocal characteristics like pitch, formant frequencies delineating vocal tract resonances, and the energy distribution across the speech spectrum.

Through meticulous analysis of these features, SB-ASR systems endeavor to discern the most plausible sequence of phonemes or words responsible for generating the given audio input. Nonetheless, SB-ASR encounters several challenges and limitations, including:

- **Susceptibility to Background Noise**: Background noise can distort the signal, making accurate transcription difficult.
- **Variability in Speech Characteristics**: Differences in age, gender, and accent can affect speech recognition accuracy.
- **Ambiguity of Sounds**: Certain sounds may represent multiple phonemes, necessitating contextual cues for accurate disambiguation.

Despite these challenges, SB-ASR offers a multitude of benefits, including:

- **Broad Applicability**: It can be applied across different languages and speech styles given adequate training data.
- **Independence from Visual Cues**: It operates independently of visual cues, thus mitigating privacy concerns associated with technologies reliant on lip-reading capabilities.

1.2.2 Vision-Based Automatic Speech Recognition (VB-ASR)

Vision-Based Automatic Speech Recognition (VB-ASR) is an emerging technology aimed at recognizing spoken language by analyzing visual cues such as lip movements and facial expressions. Despite the dominance of Sound-Based ASR (SB-ASR), VB-ASR is gaining traction as an active research area, offering unique advantages. VB-ASR focuses on extracting features from video recordings of a speaker's face, including lip shapes, facial muscle movements, and head orientation. These visual cues complement the audio signal, improving recognition in challenging environments where SB-ASR may struggle. Potential applications of VB-ASR include:

- Enhancing lip-reading accuracy in noisy settings.

However, VB-ASR faces challenges such as:

- Achieving comparable accuracy to SB-ASR.
- Increased computational costs due to processing video data.
- Privacy concerns regarding the collection of visual information.

Responsible development and deployment of VB-ASR require careful ethical considerations and user consent to address these challenges effectively. [4]

1.3 Evolution of Lip-Reading Technology

Lip-reading technology has come a long way, evolving from its early days of visual feature extraction techniques such as Discrete Cosine Transform (DCT) and Principal Component Analysis (PCA). These early efforts were focused on extracting visual information from the mouth region to identify basic lip shapes, albeit with limited

accuracy due to the complexities of lip movement and variations among individuals. However, the advent of deep learning algorithms marked a significant revolution in this field. Convolutional Neural Networks (CNNs) were introduced to learn intricate visual features from lip movements, while Long Short-Term Memory (LSTM) networks captured temporal dependencies within lip sequences. Consequently, deep learning models have surpassed traditional methods by a substantial margin, enhancing accuracy and opening doors to diverse applications. These applications include enhancing speech recognition for hearing-impaired individuals by supplementing audio with visual cues, analyzing silent video footage for surveillance and security purposes. Despite these advancements, challenges remain, such as improving accuracy in noisy environments or with obscured faces, addressing language variations and speaker-specific nuances, and navigating ethical considerations surrounding privacy and the potential for misuse of this technology.

1.3.1 Historical background

Lip-Reading, the communication technique of visually interpreting the spoken word by observing the movements of the speaker's lips, face, and tongue, has a rich historical background spanning centuries. It is has been found that the LR was originated in Spain as far back as the 16th century and gradually spread throughout Europe to countries such as England, Belgium, the Netherlands, and France. [7]

In the XVIII, a French priest called Charles Michel de Lepe founded the first school for the deaf in Paris in 1760 and invented the sign language system still in use today.[8]

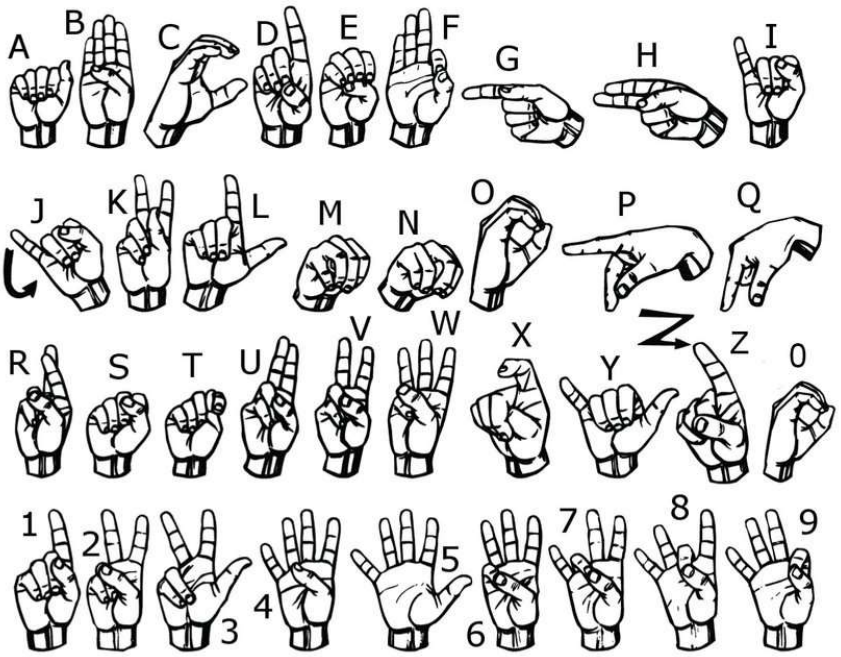

Figure 1.1 : *sign language* **Source**:[17]

Samuel Heinicke (1727-1790), a German educator founded the first German school for the deaf in 1778 and pioneered the teaching of speaking and lip-reading to deaf children.

The 19th century saw the emergence of specialized instructors teaching lip-reading to adults and the start of conferences and events focused on lip-reading.[9] The late 19th and early 20th centuries witnessed the rise of organizations for the hearing impaired, specifically addressing the challenges that parents of deaf children faced in learning to lip-read.

Technological advances had a significant impact on lipreading:

- Hearing aids: The invention and improvement of hearing aids enhanced hearing function and complemented lipreading.
- Cochlear implants: The development of cochlear implants revolutionized hearing restoration and further improved communication for the deaf.[11]

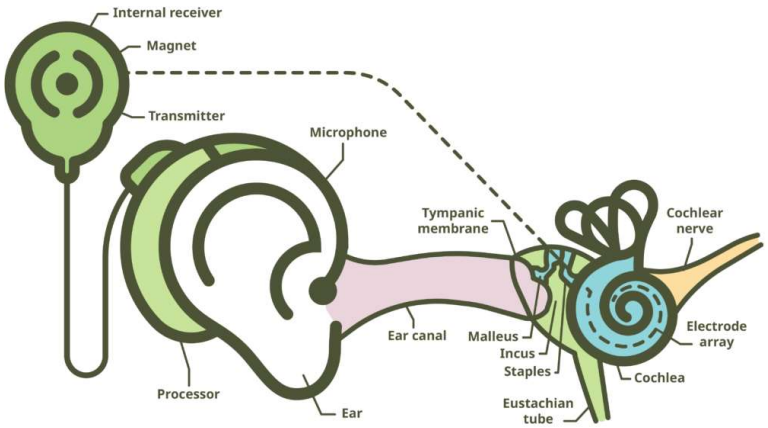

Figure 1.2 :*Cochlear implants* ***Source****:[18]*

Despite technological advances, lipreading remains a valuable skill, especially in noisy environments and when conversing with individuals from different language backgrounds. Lipreading continues to empower the deaf and hard of hearing to communicate effectively.

1.3.2 Lip-Reading from Manual to Automatic Systems

The shift from manual lip reading to automated systems has been facilitated by technological advancements, particularly the use of deep convolutional neural networks (CNNs) and attention-based long short-term memory (LSTM) in automatic lip-reading systems [6]. These advanced systems detect lip movements and extract key frames from videos to enhance the accuracy and efficiency of lip reading, thereby making this technology more accessible to a broader audience.

In automatic lip-reading systems, deep convolutional neural networks (CNNs) play a crucial role in processing visual information by analyzing the spatial features of lip movements. CNNs are adept at detecting patterns and shapes in the visual data, allowing them to identify key visual cues associated with speech sounds. By leveraging the hierarchical structure of CNNs, these systems can extract meaningful representations of lip movements from video frames, enabling them to effectively recognize and interpret spoken language.

On the other hand, attention-based long short-term memory (LSTM) models are utilized to capture the temporal dynamics of lip movements. LSTMs are well-suited for modeling sequential data and are particularly effective in capturing long-range dependencies in time series data. By incorporating attention mechanisms into LSTM architectures, automatic lip-reading systems can focus on relevant parts of the input sequence, enhancing their ability to understand the sequential nature of speech production and improve the accuracy of speech recognition.

The integration of CNNs and LSTMs in automatic lip-reading systems enables these systems to extract and analyze both spatial and temporal information from videos, leading to more accurate and efficient lip reading capabilities. This technological advancement has significantly enhanced the accessibility of lip reading for various applications, including communication aids for individuals with hearing impairments, surveillance systems, and human-computer interaction interfaces.

1.3.3 Key advancements and breakthroughs in the field

Recent years have witnessed remarkable strides in lip-reading technology, largely propelled by the advent of deep learning methodologies. Among the pivotal advancements contributing to its evolution are sophisticated neural network architectures.

Convolutional Neural Networks (CNNs) have emerged as adept tools for discerning nuanced lip movements, as they can autonomously learn intricate features from extensive datasets coupling lip videos with corresponding audio.

Concurrently, Long Short-Term Memory (LSTM) networks have excelled in unraveling temporal dependencies within lip movement sequences, a pivotal aspect for decoding speech nuances affected by preceding and subsequent movements.

Moreover, the paradigm of end-to-end learning has revolutionized traditional workflows, as deep learning models now seamlessly integrate feature extraction and classification, leading to enhanced accuracy in lip-reading tasks.

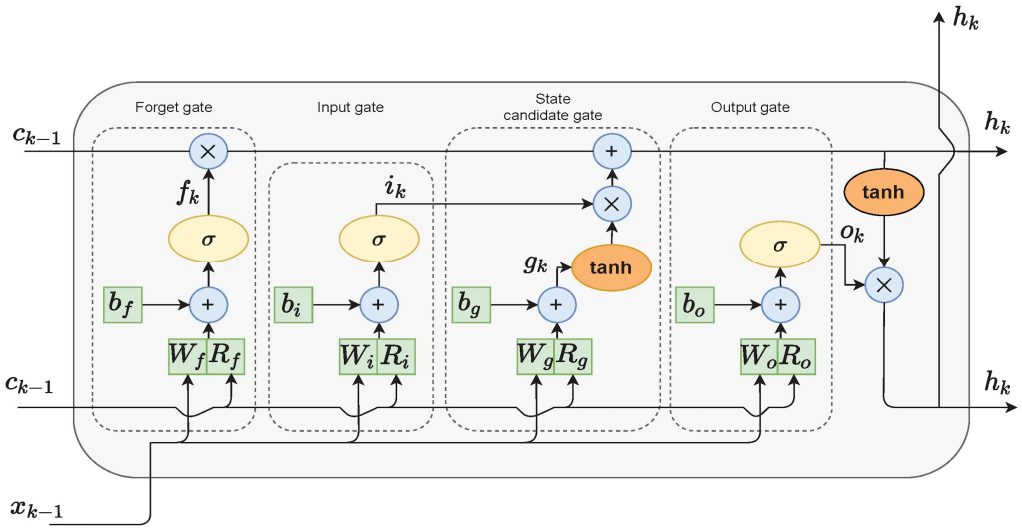

***Figure* 1.3:** LSTM architecture Source :[12]

The surge in available labeled data, facilitated by the creation of expansive lip-reading datasets, coupled with advancements in transfer learning techniques from related domains like facial recognition, has further bolstered model efficacy even with limited specific lip-reading data.

Particularly, ongoing research endeavors are also prioritizing the development of speaker-independent recognition capabilities, overcoming historical challenges posed by varying lip shapes and speaking styles, thus enabling these technologies to comprehend speech from a broader spectrum of individuals.

1.4 Lip-Reading (LR) as an alternative method to speech recognition

Lip-reading (LR) and speech recognition are two distinct methods of understanding human communication. LR involves visually interpreting lip, mouth, and facial movements to comprehend speech, while speech recognition relies on audio input to analyze spoken language. However, LR faces numerous challenges that hinder its effectiveness compared to speech recognition.

These challenges include difficulties in group settings or with multiple speakers, variations in speaking styles, limited contextual cues, and obstacles to clear visual access.

Instances where multiple speakers are present or the speaker is not directly facing the deaf or hard-of-hearing individual pose significant hurdles for LR. Moreover, factors like facial hair obstructing lip movements, hand gestures covering the mouth, downward gazes, or erratic pacing further complicate lip-reading. Furthermore, LR can be mentally taxing for the deaf or hard-of-hearing, as they constantly strain to follow conversations and may struggle to discern appropriate moments for participation. [5]

Additionally, LR may miss non-visible information, such as when the camera angle or microphone placement obstructs the speaker's mouth. Despite these obstacles, LR can still prove valuable in specific scenarios, particularly when the speaker directly faces the deaf or hard-of-hearing individual without hindrances. On the other hand, speech recognition stands out as a more dependable method for comprehending spoken language. Studies indicate its superiority over LR, boasting a character error rate of 3.9% for speech recognition compared to 39.5% for lip-reading. Speech recognition exhibits versatility in handling speech variations like accents and dialects, a challenge for LR. In conclusion, while LR holds utility in certain contexts, its overall efficacy pales in comparison to speech recognition for understanding human communication. Speech recognition emerges as the more reliable option, especially in scenarios where direct visual access to the speaker is unseen or multiple barriers exist. [6]

1.4.1 Lip-Reading in helping communication for deaf people

Lip-reading, also known as speech reading, plays a crucial role in communication for many deaf individuals by supplementing auditory information. Even with hearing aids, some sounds can be challenging to distinguish, but lip-reading provides visual cues about mouth movements, enhancing the understanding of spoken language.

Additionally, lip-reading increases context awareness as facial expressions and body language complement lip movements, providing context to the spoken word. This combined information aids comprehension, especially in situations with background noise or unclear audio.

Moreover, lip-reading facilitates social interactions by empowering deaf and hard-of-hearing individuals to participate more actively in conversations, fostering social inclusion, and reducing feelings of isolation.

1.4.2 Lip-Reading in Assistive Technologies

Lip-reading transcription is a crucial component of assistive technology, aiding individuals with hearing impairments in communication. These technologies find applications in communication aids, education for students with hearing impairments, and forensic investigations [14].

Lip reading involves transforming visual lip movements into spoken or written language. While essential, it requires excellent visual acuity and extensive training. Factors like distinctness of lip movement, speaker's voice, and background noise affect transcription accuracy [10].

In education, lip-reading assists students with hearing disabilities in comprehending lectures and participating in discussions. It can be combined with spoken or sign language to enhance communication abilities in children with disabilities [15].

Recent advancements in AI-based lip-reading systems have revolutionized the industry, enhancing transcription precision and reliability. These advancements complement existing transcription software [16].

Overall, lip-reading transcriptions hold great value in assistive technologies, facilitating communication for individuals with hearing impairments in various settings. Advanced technology has made lip-reading transcription more precise and convenient, enhancing the communication experience for those with hearing impairments.

1.5 Lip-Reading Transcription

Lip reading transcription involves converting visual information obtained from lip movements into written or spoken language. This valuable skill serves to bridge the gap

between oral communication and individuals with hearing loss. This process can be performed either manually by a trained lip reader or automatically using advanced AI technology.

The accuracy of lip reading transcription relies on several factors, including the clarity of lip movements, speaker tone, and the presence of background noise [13].

1.5.1 Methods of Lip-Reading Transcription

There are two primary methods of lip-reading transcription: manual and automatic . Manual transcription relies on skilled lip readers to observe and translate lip movements into written or spoken words, while automatic transcription employs AI algorithms to analyze video or image data for lip pattern recognition.

Several factors impact the accuracy of both methods [14]. Clear enunciation and good lighting enhance accuracy, while regional accents and background noise can pose challenges for both human and AI transcription systems.

1.5.2 Applications of Lip Reading Transcription (LRT)

Lip reading transcription (LRT) has diverse applications across various fields, enhancing communication and technology in unique ways: Assistive Technology for Hearing Impairments: LRT plays a crucial role in assisting individuals with hearing impairments by providing a visual interpretation of speech through lip movements, enabling better communication.

- **Forensic Investigations:** Forensic lip reading is utilized in forensic investigations to collect information or evidence from video footage where audio is unclear or unavailable. This technique has been instrumental in cases like the Arlene Fraser murder, where forensic lip reading helped convict individuals based on lip movements captured in recordings.
- **Improving Speech Recognition Accuracy in Noisy Environments**: LRT can enhance speech recognition accuracy in noisy environments by providing visual

cues that complement audio information, particularly useful in scenarios with background noise or audio disturbances.

- **Language Learning and Classroom Accessibility:** LRT helps in language learning by providing visual hints for correct pronunciation. Also, LRT helps students with hearing impairments in following their studies effectively.
- **Subtitling, captioning, dubbing and voiceover:** LRT is used for accurate subtitling and captioning, enhancing accessibility for viewers with hearing impairments. It can also, ensure lip movements match audio for a natural viewing experience
- **Biometric Identification:** LRT is used in biometric identification systems for authentication.
- **Cognitive Science:** LRT aids in cognitive science research for studying human processing of visual and auditory information during communication
- **Interrogations and Interviews:** LRT aids in analyzing non-verbal cues during interrogations and interviews.
- **Courtroom Proceedings:** LRT assists in transcribing and interpreting lip movements during courtroom proceedings, enhancing accuracy in documenting testimonies
- **Call Centers:** LRT improves customer service interactions by capturing and analyzing lip movements for better understanding and response.
- **Multilingual Communication:** LRT facilitates multilingual communication by providing visual cues that aid in understanding accents and dialects
- **Coaching and Training:** LRT is used in sports coaching to analyze lip movements for better communication between coaches and athletes.
- **Broadcasting:** LRT assists in providing accurate commentary by capturing and transcribing lip movements of commentators for live broadcasts

- **Therapeutic Applications:** LRT has therapeutic applications in mental health settings by helping individuals improve communication skills through visual feedback.

1.5.3 Challenges & limitations of current Lip-Reading transcription methods

Despite the advancements in LR technology, there are challenges and limitations to overcome. One major issue is accuracy; even with deep learning models, lip-reading transcription is not perfect. Factors such as lighting conditions, speaker appearance (like facial hair), and speaking style (like accents or fast speech) can all affect the accuracy of lip-reading models. Additionally, there is a limitation in vocabulary, as current models may struggle with specialized terms or uncommon words. Real-time applications also pose challenges, especially for extended conversations, and speaker dependence is another issue, with many models needing specific training on individuals for optimal performance. Furthermore, privacy fears arise due to the potential for extracting information from silent videos, requiring careful ethical frameworks for in charge use.

1.6 Conclusion

In this chapter, we covered the evolution, challenges, applications, and breakthroughs of Automatic Speech Recognition (ASR), Vision-Based ASR (VB-ASR), and Lip-Reading Transcription (LRT). We discussed their foundational principles, innovative frontiers, and transformative advancements, highlighting their impact on human-machine interactions.

In the next chapter, we will delve into the methods used for data collection, processing techniques, parameters, and databases utilized in our research.

CHAPTER-2

CONCIEVED DATABASES AND PROPOSED METHODS

CHAPTER-2

CONCIEVED DATABASES AND PROPOSED METHODS

2.1 Introduction

In this chapter, we delve into the comprehensive methodologies utilized in our data collection and database creation. Our focus encompasses the intricate processes of data collection, meticulous data cleaning, and the strategic selection of parameters pivotal for optimizing model performance. We also explore the essential materials and libraries that facilitated our research, highlighting their significance in achieving accurate and reliable results in lip reading.

2.2 Database Collection, Cleaning, and Parameters

In this section, we outline the methodologies employed for data collection, the subsequent data cleaning processes, the parameters considered, and their significance within the context of our lip reading project using 3D CNN.

2.2.1 Data Collection Techniques

Our data collection methodology underwent several iterations and refinements to ensure comprehensive coverage and quality. Initially, we employed a Python script to capture video data, focusing on Arabic language words with 12 frames per video, encompassing three distinct Arabic labels. Each label was associated with 50 videos, resulting in a dataset of 1800 images. Subsequent adjustments were made to explore various frame rates and labels to enhance the model's performance and achieve better results.

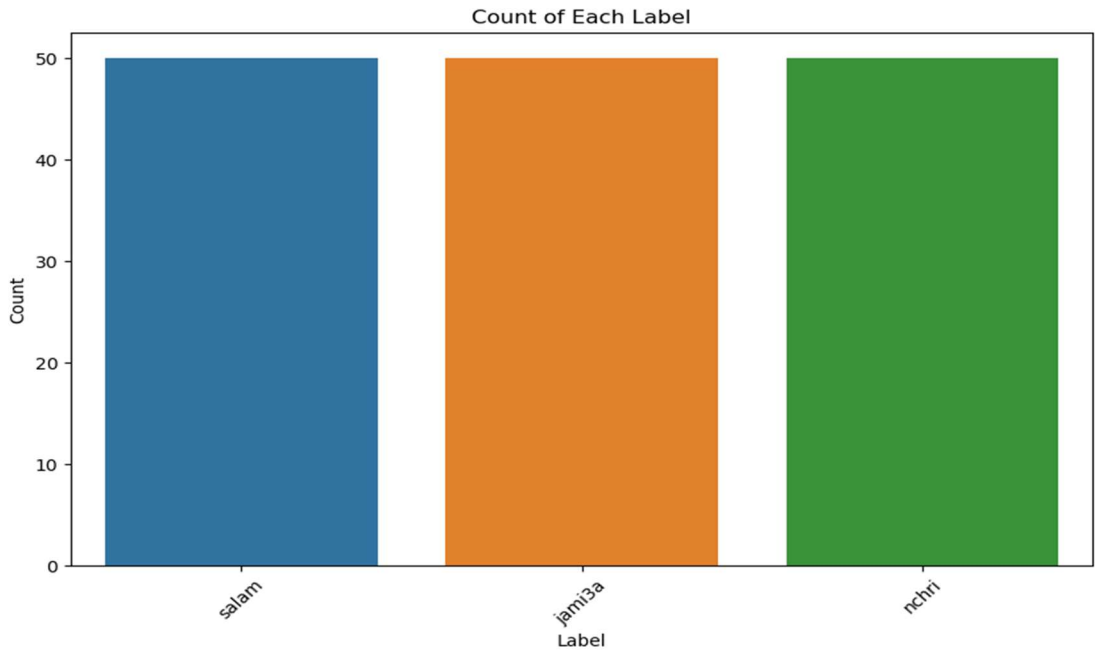

***Figure* 2.1 :** Count of each Arabic word label

A pivotal shift in our approach involved transitioning to English language words, driven by English's global prevalence and the inherent ease in articulating lip movements. We settled on a frame rate of 20 frames per video, as this rate effectively captured most word formations accurately within a reasonable temporal window, aligning seamlessly with the project's objectives.

Acknowledging the variability in lip movements among individuals, we diversified our dataset by collecting data from two distinct individuals, each contributing unique word labels. This effort resulted in a comprehensive dataset, enriching the dataset's diversity and robustness.

For our final approach, we employed a Python script to capture data concurrently with a secondary camera recording the person saying each word 20 times. This dual-camera setup facilitated the creation of two distinct datasets, each serving specific purposes and enriching the overall dataset diversity and quality.

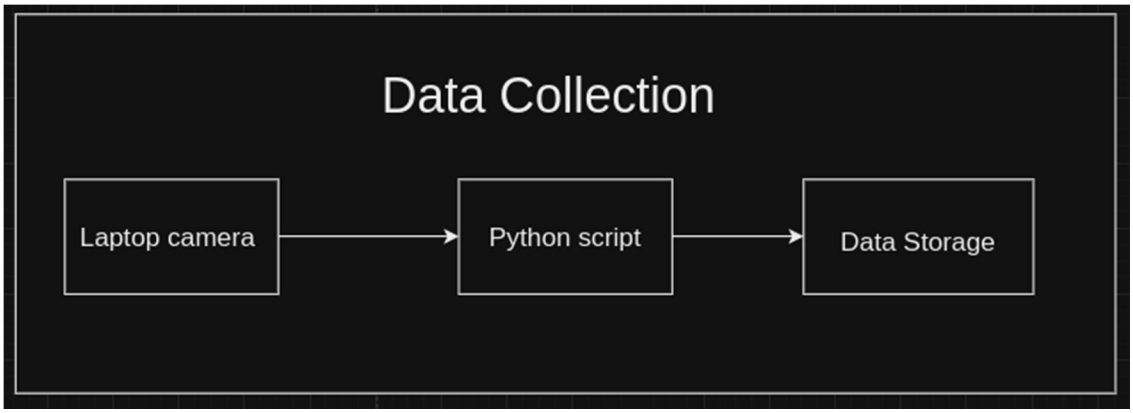

***Figure* 2.2 :** Data collection

2.2.1.1 First Dataset

The first dataset, captured using a laptop camera, included ten unique labels, with data collected from two individuals, both of them were men. Each person contributed 20 videos per label, with each video containing 20 frames capturing the lip region. We utilized Dlib and a face weight model to isolate and extract the lip region within each frame, enhancing the dataset's suitability for lip reading analysis. The labels collected for this dataset included ["help", "emergency", "danger", "hospital", "hurt", "hungry", "airport", "train", "restaurant", "lost"].

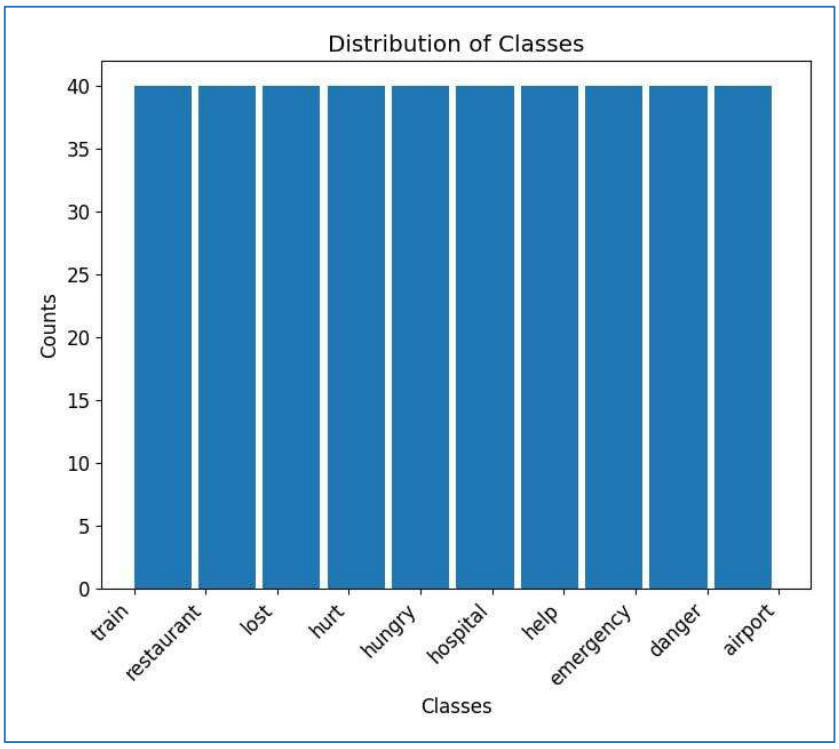

***Figure* 2.3 :** Count plot of English labels

***Figure* 2.4 :** Images from every video labels

2.2.1.2 Second Dataset

The second dataset captured using a mobile phone camera, also featured ten unique labels, with data collected from two male individuals. Each person repeated the same label 20 times in a single video, contributing to a diverse dataset. The labels for this dataset mirrored those of the first dataset, providing consistency and comparability across both datasets.

2.2.2 Data Cleaning Processes

The decision-making process regarding data cleaning revolved around optimizing the dataset's quality while maintaining its authenticity and real-world applicability. We primarily focused on Dataset 1 due to its favorable condition, attributed to the meticulous data collection facilitated by the Python script. This script efficiently captured each image, emphasizing the lip region, which inherently reduced the need for extensive cleaning processes.

Our approach to data cleaning involved a deliberate choice to refrain from extensive cleaning techniques that might compromise the dataset's real-world resemblance. Unlike idealized datasets often used in controlled environments, real-world video data can be inherently noisy and imperfect. Therefore, we opted not to employ aggressive cleaning methods or exclude videos unless necessary.

The rationale behind this decision stemmed from the understanding that real-world scenarios often involve imperfect or messy video data, reflecting the challenges faced in practical lip reading applications. By retaining such variability in the dataset, our

model could learn to adapt and generalize better to diverse real-world scenarios, enhancing its robustness and applicability in practical settings.

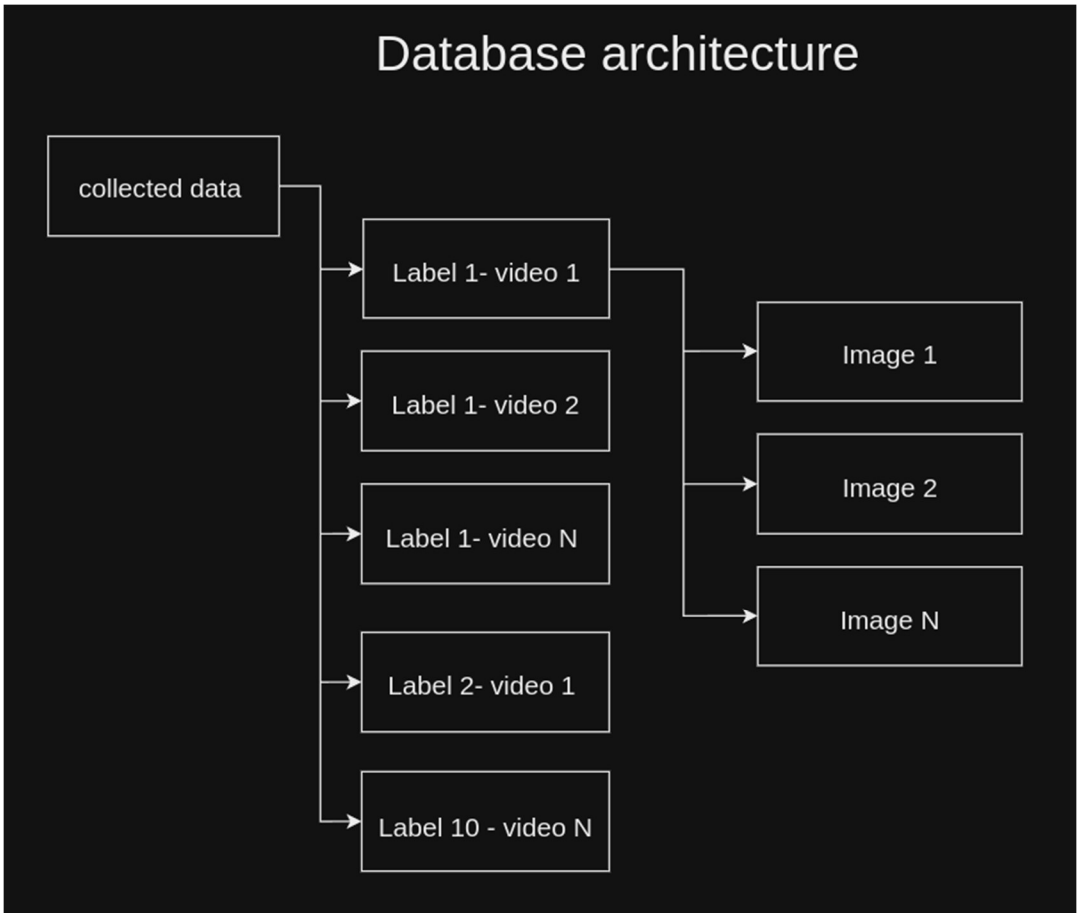

***Figure* 2.5 :** Flowchart of database

2.2.3 Considered Parameters and their Significance

Several critical parameters were incorporated into our data processing pipeline to optimize model performance:

- TOTAL_FRAMES (Frame Count): Set at 20 frames per video, striking a balance between capturing sufficient temporal information and computational efficiency.

- VALID_WORD_THRESHOLD: Configured at 1, ensuring that only words with a minimum occurrence threshold were retained, promoting robustness in word recognition.

- NOT_TALKING_THRESHOLD: Defined as 10, delineating periods of non-speech to facilitate accurate word segmentation.

- PAST_BUFFER_SIZE: Established at 4 frames, facilitating context-awareness by considering preceding lip movements in word prediction.

- LIP_WIDTH and LIP_HEIGHT: Dimensions set to 112x80 pixels, optimizing lip region extraction for feature representation.

Additionally, the utilization of the Dlib shape predictor ("face_weights.dat") served as a crucial component in accurately localizing and extracting the lip region within each frame, streamlining the focus on lip movements for subsequent analysis.

These parameters collectively contributed to refining the dataset quality, enhancing model training efficacy, and ultimately enhancing the lip reading system's performance.

2.3 Used Materials

In this section, we explore the essential materials utilized in our study, each material playing a crucial role in the data collection, processing, and final outcome of the project. These materials encompass a range of technological components carefully selected to ensure optimal performance and accuracy throughout the research process.

2.3.1 Laptop Camera

The foundation of our data collection phase relies on a high-resolution laptop camera. This camera serves as the primary tool for capturing videos of facial movements and lip patterns.

Through its advanced imaging capabilities, intricate details of lip movements are recorded, providing a rich dataset for analysis and model development.

2.3.2 Laptop

Working in Collaboration with the camera, the laptop serves as the hub for data collection and initial processing. Its computational power enables real-time video processing, facilitating the extraction of key features necessary for training the lip reading model. Additionally, the laptop acts as the interface for data storage and management, ensuring seamless workflow integration.

2.4 Used Libraries

2.4.1 Dlib library

The Dlib library is a versatile and comprehensive C++ library that offers a range of machine learning algorithms and tools designed to assist in the creation of sophisticated software systems. Initially developed by Davis King, Dlib has gained substantial popularity within the computer vision community due to its efficiency, ease of use, and the high performance of its algorithms. The library's design focuses on providing a robust set of tools for developers to implement and experiment with machine learning models and computer vision techniques, without delving deeply into the complex mathematics behind them.

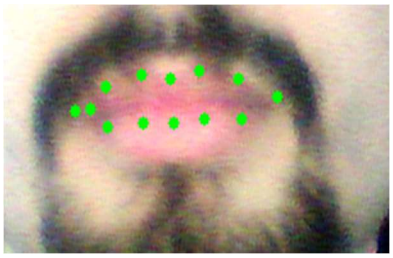

***Figure* 2.6 :** Dlib lip tracking

One of Dlib's standout features is its powerful facial landmark detection functionality. This capability is particularly relevant for lip reading applications, where precise localization of facial features is crucial. Facial landmark detection involves identifying key points on a human face, such as the eyes, nose, mouth, and jawline. Dlib achieves this through a pre-trained model that uses an ensemble of regression trees, a machine learning technique known for its speed and accuracy. The model has been trained on a large dataset of annotated facial images, allowing it to generalize well across different faces, lighting conditions, and orientations.

When integrated with Python, Dlib becomes even more accessible and versatile. Python's simple syntax and **extensive** ecosystem of libraries make it a preferred language for many researchers and developers working in the field of computer vision.

In the context of lip reading using 3D Convolutional Neural Networks (3DCNN), Dlib plays a critical role in the preprocessing pipeline. Lip reading systems need to focus

on the mouth region to accurately interpret the movements of the lips, which correspond to spoken words or sounds. Dlib's facial landmark detector helps achieve this by providing precise coordinates of the lips among other facial features. The typical workflow involves capturing video frames of a speaking individual, using Dlib to detect facial landmarks in each frame, and then extracting the region of interest (ROI) that contains the lips. This ROI is then normalized and fed into the 3DCNN for further processing.

By isolating the mouth region, Dlib not only reduces the amount of data that the 3DCNN needs to process but also enhances the quality of the input data. This focused approach helps the neural network concentrate on the relevant features of lip movement, thereby improving the accuracy of the lip reading system. Moreover, Dlib's landmark detection algorithm is robust against various challenges such as occlusions (e.g., when the hand partially covers the face), different facial expressions, and slight head movements, ensuring consistent performance across diverse conditions.

Additionally, Dlib's performance extends beyond just facial landmark detection. The library includes other computer vision tools, such as object detection, image processing, and deep learning frameworks, which can be leveraged to build more comprehensive and sophisticated lip reading systems. For instance, Dlib's support for deep learning frameworks allows for seamless integration with other popular libraries like TensorFlow or PyTorch, enabling the development of custom neural network architectures tailored to specific lip reading tasks.

2.4.2 Opencv

OpenCV (Open Source Computer Vision Library) is an extensive open-source computer vision and machine learning software library. Designed for computational efficiency and with a strong **focus** on real-time applications, it is one of the most widely used libraries for computer vision tasks. OpenCV was initially developed by Intel and is now supported by the community. It provides a comprehensive set of tools for various computer vision applications, including image and video processing, object detection, facial recognition, and more. With its extensive functionalities and active community,

OpenCV has become a staple in the toolbox of developers and researchers working on computer vision projects.

One of the key strengths of OpenCV is its versatility. The library supports multiple programming languages, including Python, C++, Java, and MATLAB, with Python being particularly popular due to its simplicity and readability.

2.5 3D Convolution Operation

The 3D convolution **operation** is a fundamental component of 3D CNNs. It applies a 3D filter to the input data, moving in three dimensions (x, y, z) to extract spatial and temporal features. The mathematical representation of this operation can be expressed as:

$$Output = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{p} Filter(i,j,k) \times Input(i,j,k) \qquad (2.1)$$

Output is the output feature map.

Filter(i,j,k)is the 3D filter applied at position(i,j,k)

Input(i,j,k)is the input data at position (i,j,k)

n,m,p are the dimensions of the filter and input data.

2.5.1 3D Convolutional Kernels

3D convolutional kernels are used **to** extract features from the input data. These kernels are applied to overlapping 3D cubes in the input video to extract motion features. The size of the convolution kernel in the temporal dimension is critical in capturing motion information.

2.5.2 3D CNN Architecture

A 3D Convolutional Neural **Network** (3D CNN) refers to neural network architectures with multiple layers that can learn hierarchical data representations. Each layer learns increasingly complex spatial features of data. These representations can then be used for various tasks such as classification, regression, or generation. Each type of layer has a definite function.

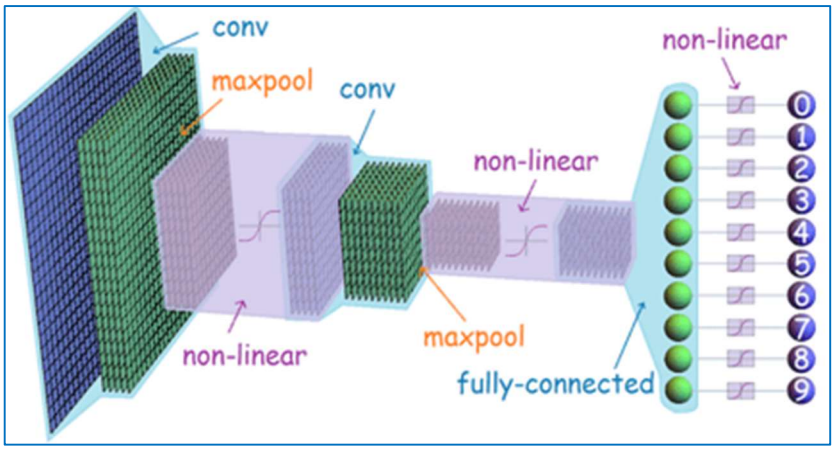

***Figure* 2.7 :** 3D CNN Architecture[19]

For example, the max-pooling layer is a "downsampler" for feature maps. A feature map is generated by convolving a **filter** over an image reducing its spatial size and computational burden. Another form of pooling is global average pooling.

2.5.3 3D Convolutional Neural Networks: Classification

A 3D Convolutional Neural Network is a deep learning model used in different applications, such as computer **vision** or medical imaging.

In these scenarios, we aim for AI (deep learning) to learn how to respond to inputs rather than programming AI according to a predetermined pattern. The result of this learning process is a predictive model.

2.6 Conclusion

The methodologies outlined in this chapter underscore the importance of meticulous data collection and cleaning processes in developing a reliable lip reading model. By leveraging diverse datasets and focusing on real-world applicability, we have enhanced the model's ability to generalize across different scenarios. In addition, the incorporation of advanced tools like the Dlib shape predictor for lip localization highlights the technical precision involved in this research. These efforts collectively contribute to the field of automatic lip reading, paving the way for further innovations and practical applications in communication technologies and accessibility solutions.

In the next chapter, we will describe the different steps of the realization of our project.

CHAPTER-3

PROPOSED AND REALIZED SYSTEM FOR LIP READING TRANSCRIPTIONS

CHAPTER-3
PROPOSED AND REALIZED SYSTEM FOR LIP READING TRANSCRIPTION

3.1 Introduction

In this chapter, we focus on achieving the system used to transcribe lip movement. At the beginning, we offer a comprehensive overview of the project plan, while clarifying its key elements; we also highlight the materials used. The organization of our work is then determined, with emphasis on the structure, techniques, models and results obtained during our study. Detailed tables and specific benchmarks are also provided, allowing for an accurate assessment of our system and its performance. Finally, we conclude this chapter by summarizing the main points, highlighting the main contributions of our work and discussing the future implications of our results.

3.2 Used Materials

3.2.1 Raspberry Pi 4 Model B

The Raspberry Pi 4 Model B is a small, versatile and powerful one-panel computer. Featuring a 64-bit high-performance quad-core processor, it offers desktop-like performance similar to entry-level PC x86 systems.

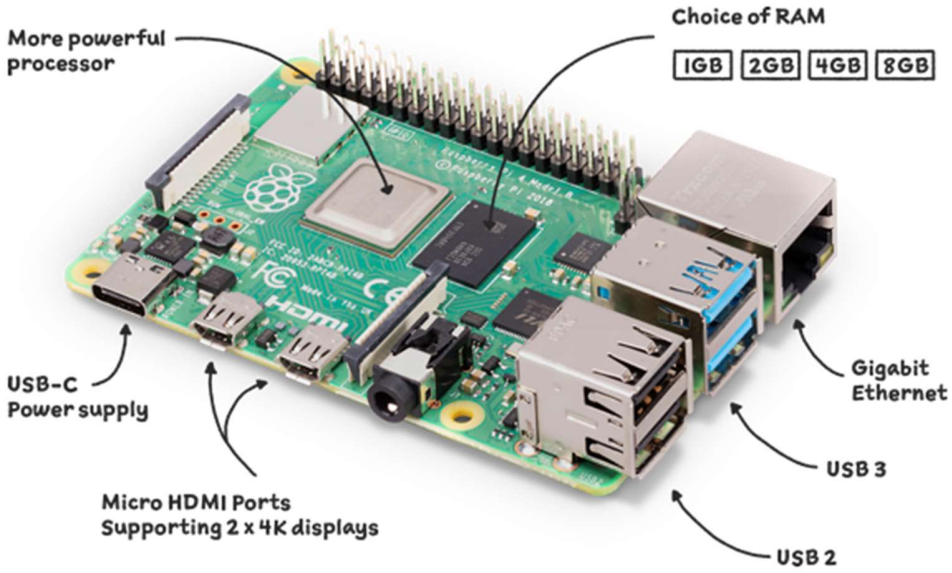

Figure 3.1 : Raspberry Pi 4 Model B General Structure

With dual display support with resolution up to 4K via small HDMI ports and hardware video decoding capabilities up to 4Kp60, it provides a multimedia experience. In terms of connectivity, the Raspberry Pi 4 comes with Gigabit Ethernet network, 2.4/5.0 GHz wireless LAN network, and Bluetooth 5.0, ensuring fast and reliable network options. In addition, it offers abundant memory configurations ranging from 1GB to 8GB of RAM, meeting different computing needs.

Despite its strong performance, the Raspberry Pi 4 maintains a silent and energy-efficient process, featuring multiple USB ports, including USB 3.0 and USB 2.0, as well as PoE capability via a separate addition.

In short, the Raspberry Pi 4, with its 8GB RAM variant, is a compact and powerful computer solution, suitable for a wide range of applications, from desktop computing to multimedia projects and beyond. Its affordability, diversity and community support make it an attractive choice for both amateurs, teachers and professionals.

Table 3.1: Characteristics and features for Raspberry Pi 4

Processor	Broadcom BCM2711, quad-core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
Memory	1GB, 2GB, 4GB or 8GB LPDDR4 (depending on model) with on-die ECC
Connectivity	2.4 GHz and 5.0 GHz IEEE 802.11b/g/n/ac wireless LAN, Bluetooth 5.0, BLE Gigabit Ethernet 2 × USB 3.0 ports 2 × USB 2.0 ports.
GPIO	Standard 40-pin GPIO header (fully backwards-compatible with previous boards)
Video & sound	2 × micro HDMI ports (up to 4Kp60 supported) 2-lane MIPI DSI display port 2-lane MIPI CSI camera port 4-pole stereo audio and composite video port
Multimedia	H.265 (4Kp60 decode); H.264 (1080p60 decode, 1080p30 encode); OpenGL ES, 3.0 graphics
SD card support	Micro SD card slot for loading operating system and data storage
Input power	5V DC via USB-C connector (minimum 3A1) 5V DC via GPIO header (minimum 3A1) Power over Ethernet (PoE)–enabled (requires separate PoE HAT)
Environment	Operating temperature 0–50ºC
Production lifetime	Raspberry Pi 4 Model B will remain in production until at least January 2031.
Compliance	For a full list of local and regional product approvals, please visit pip.raspberrypi.com

3.2.2 Input and output

One of the prominent features of Raspberry Pi is the GPIO pin set (inlet/outlet), located along the top edge of the panel. Each Raspberry Pi features a 40-pin GPIO head. These heads have 0.1 inch (2.54 mm) spaced pins.

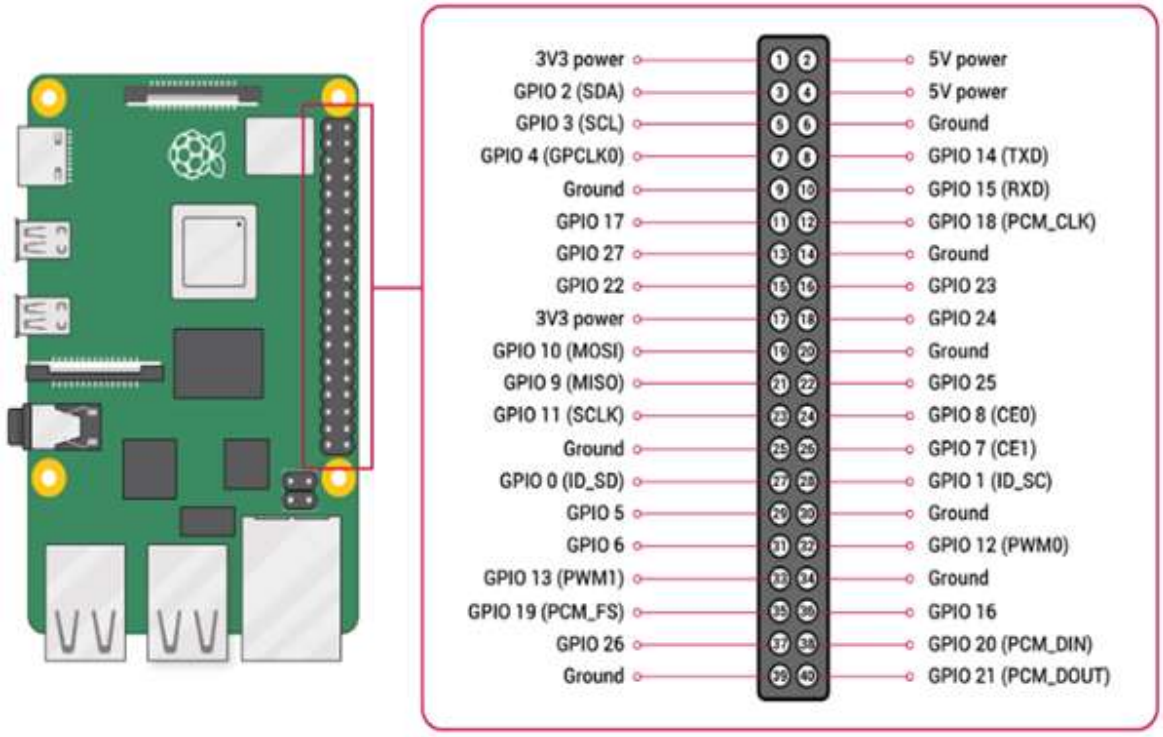

Figure 3.2: GPIO and 40-pin header

The board includes 5V pins and specific 3.3V, along with ground pins (GND), which can be reconfigured. The rest of the versatile pins are 3.3V general-purpose pins, designed to handle input and 3.3V output.

3.2.3 Bone Conduction Earphones

Advanced technology transmits sound directly to the inner ear through the bones of the skull, allowing you to hear surrounding sounds. Wireless with Bluetooth 5.3: Stable and fast connection, compatible with iPhone, iPad, iOS, and Android devices.

- **Ergonomic design :** Ear hook style for a secure and comfortable fit even during intense physical activities.

- **Waterproof:** Designed to be water-resistant, making them perfect for outdoor activities in all weather conditions.
- **Advanced features:**
  - **HiFi sound quality :** Delivers high-fidelity audio.
  - **Active noise cancellation:** Immersive listening experience.
  - **Built-in microphone:** Hands-free calls and voice commands.
  - **True wireless stereo (TWS):** Wireless stereo sound.
  - **Voice assistant support:** Compatible with Apple Siri and Google Assistant for easy hands-free use.
  - **Volume control and control buttons:** Easy management of music and calls.
  - **Sensitivity and frequency range:** 86dB sensitivity and 20Hz to 20kHz frequency response range for exceptional audio quality.
  - **Impedance and resistance:** Up to 32 Ω resistance with a corresponding impedance range.
  - **Charging method:** USB cable charging, charging cable included.
  - **Certifications and origin:** CE certified and made in China (CN).

Figure 3.3: Bone Conduction Earphones

3.3 Flowchart and Workflow Overview

This section provides an overview of the workflow from the beginning of the data upload until the evaluation of the trained model. Every step of the work is explained in detail, a flowchart will be used to clarify the data flow and modal process.

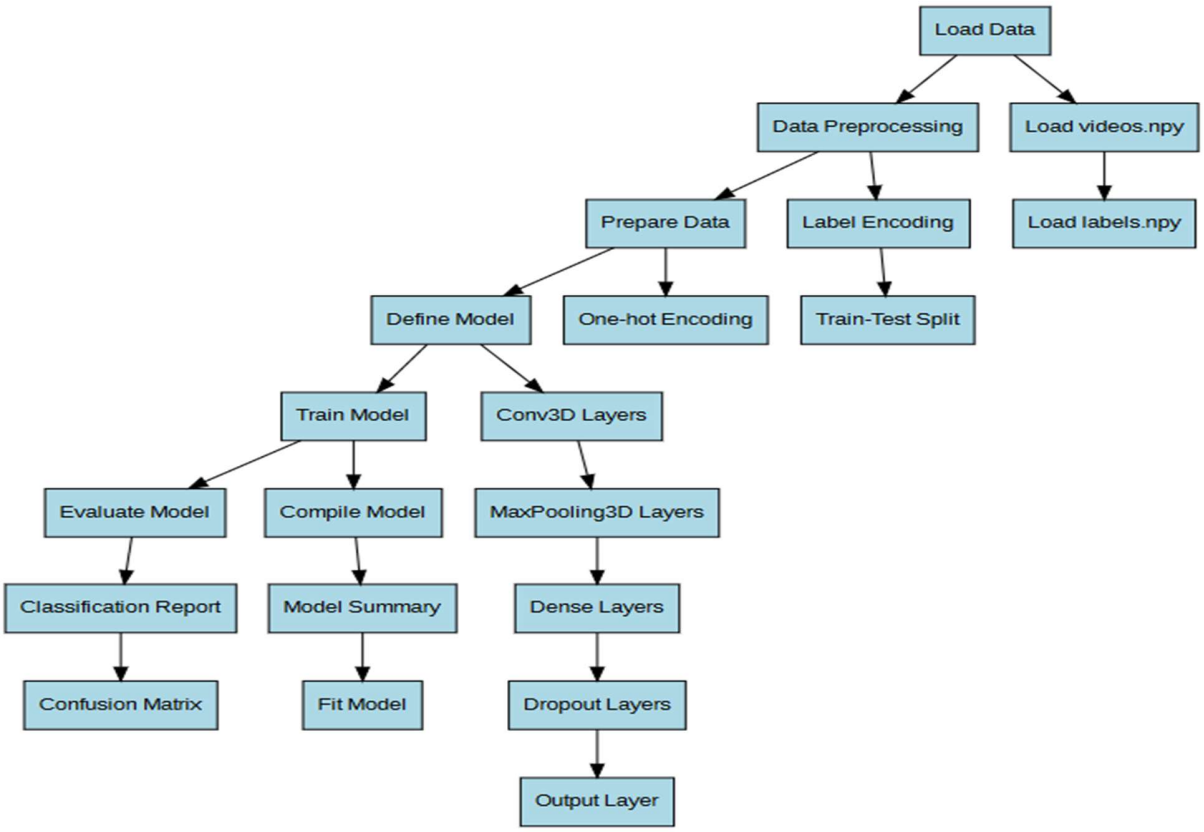

Figure 3.4 : Flowchart of workflow

3.3.1 Load Data

Initially, the necessary data must be uploaded to train and evaluate the model, the data in this context includes videos and associated labels, the videos and labels are stored in files in NumPy format (.npy). This step is essential because it provides the raw materials that the model will work on later.

3.3.2 Data Preprocessing

Data processing is a vital phase aimed at effectively preparing data to feed the neural model. This step includes several processes:

- Converting labels into numbers using LabelEncoder to convert text labels into correct numbers that the model can handle.
- Dividing data into training and testing sets using the train_test_split function of the scikit-learn library. This allows a portion of the data to be used to train the model and another part to assess its performance on invisible data during training.
- The data is broken down to 80% for training and 20% for testing with random_state=42 value used to ensure reproduction of results.

Figure 3.5: Distribution of Classes in Training and Test Sets

- Reconfiguration of data to comply with model requirements, where videos are converted to the desired format in terms of dimensions (number of frames, height, width, number of channels);

Where dimensions (20, 80, 112, 3) are as follows :

- 20 : Is the number of frames in the video clip;
- 80 : Is the height of each frame;
- 112 : Is the width of each frame;
- 3 : Is the number of channels (in this case, representing the three colors: red, green, and blue).

3.4 Experimental work and Results

In this section, we explain the preliminary experiments conducted using different models of the same dataset as well as another experiment using a different dataset.

These diverse experiences have played a crucial role in selecting the data set and shaping the final model structure and training strategy.

3.4.1 Experiment 1 : Using the Arabic data set

We will explain in this section, the obtained results using the Arabic dataset for training and demonstrating the user form mentioned below (**Table 3.2**) with the different Hyper parameters.

Table 3.2: Model structure components

Layer (type)	**Output Shape**	**Parameters**
Conv3D	(None, 18, 78, 110, 16)	1312
MaxPooling3D	(None, 9, 39, 55, 16)	0
Conv3D	(None, 7, 37, 53, 64)	27712
MaxPooling3D	(None, 3, 18, 26, 64)	0
Flatten	(None, 89856)	0
Dense	(None, 128)	11501696
Dropout	(None, 128)	0
Dense	(None, 64)	8256
Dropout	(None, 64)	0
Dense	(None, 10)	650

Table 3.3: Hyperparameters for Experiment 1

Hyperparameter	Value
Dropout Rates	0,5 0,5
Epochs	10
Batch Size	2
Random State	42
optimizer	Adam
Loss Function	Categorical Crossentropy

3.4.2 Results of Experiment 1

Table 3.4: Results for Experiment 1

Class	Precision	Recall	F1-Score	Support
salam	0.71	0.83	0.77	6
jamiaa	0.93	0.93	0.93	14
nchri	1.00	0.9	0.95	10

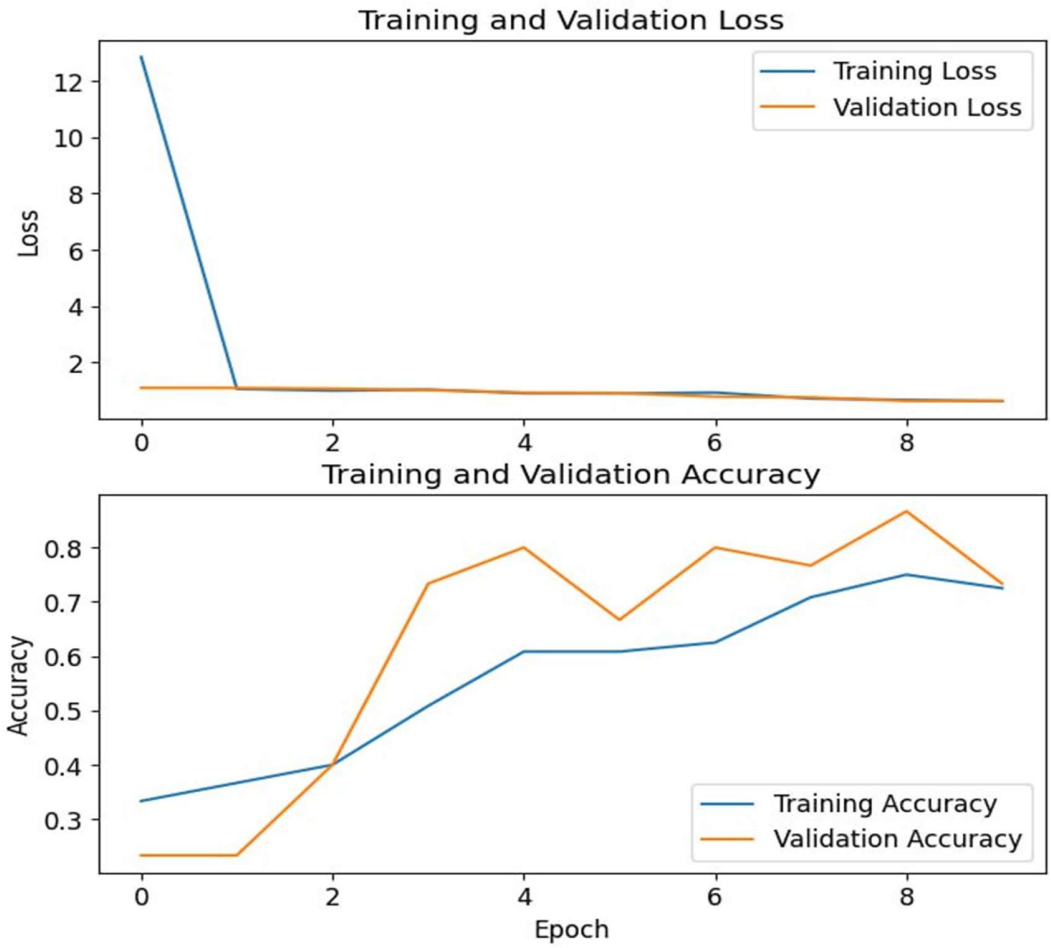

***Figure* 3.6:** Training and Validation Loss and Accuracy Curve for Experiment 1

The highest rate of classification error is between “salam” and “jamiaa”, suggesting that these categories may share some similarities that make it difficult to distinguish between them.

In addition to the difficulty of distinguishing words, we encountered another problem where it is difficult to track the movement of lips to certain Arabic letters, which led to the discontinuation of work in the Arabic dataset and the change to another dataset.

3.4.3 Experiment 2: Use Different Structures with Final Dataset

Table 3.5: Model structure components for Experiment 2

Layer (type)	Output Shape	Parameters
Conv3D	(None, 18, 78, 110, 16)	1312
MaxPooling3D	(None, 9, 39, 55, 16)	0
Conv3D	(None, 7, 37, 53, 16)	6928
MaxPooling3D	(None, 3, 18, 26, 16)	0
Flatten	(None, 22464)	0
Dense	(None, 128)	2875520
Dropout	(None, 128)	0
Dense	(None, 64)	8256
Dropout	(None, 64)	0
Dense	(None, 10)	650

Table 3.6: Hyperparameters for Experiment 2

Hyperparameter	Value
Dropout Rates	0,2 0,2
Epochs	20
Batch Size	32
Random State	42
optimizer	Adam
Loss Function	Categorical Crossentropy

3.4.4 Results of Experiment 2

Table 3.7: Results for Experiment 2

Class	Precision	recall	F1-score	support
train	1.0	0.5	0.67	12
restaurant	0.48	1.0	0.65	10
lost	0.78	0.88	0.82	8
hurt	1.0	0.83	0.91	6
hungry	0.5	0.17	0.25	6
hospital	0.25	0.6	0.35	5
help	0.89	1.0	0.94	8
emergency	1.0	0.5	0.67	6
danger	0.7	0.88	0.78	8
airport	1.0	0.27	0.43	11

Generally, this model shows unsatisfactory results as we observe that the categories "train," "hurt," "help," and "emergency" have high precision (1.0) but low recall for some. The "restaurant" category has excellent recall (1.0) but low precision (0.48). Categories "lost," "hurt," "help," and "danger" have high F1-scores, while "hungry" and "airport" have low F1-scores. Categories with a small number of samples, such as "hospital" and "hurt," might affect the accuracy of the conclusions.

3.4.5 Experiment 3 : Using final Model architecture and Results

In this context, we will learn about the final structure of the model and the layers used in its composition, as well as discuss the final results obtained, The following picture shows Connecting layers used in model composition

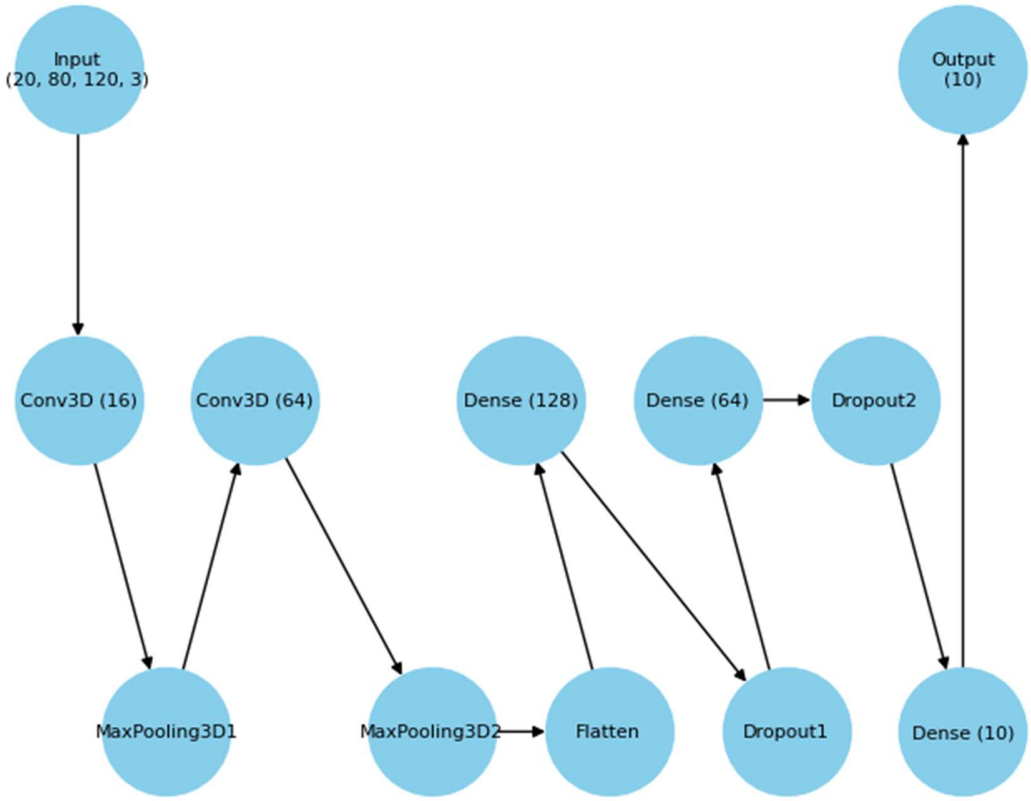

Figure 3.7 : Connecting layers used in model composition

Table 3.8: Model structure components

Layer (type)	Output Shape	Parameters
Conv3D	(None, 18, 78, 110, 16)	1312
MaxPooling3D	(None, 9, 39, 55, 16)	0
Conv3D	(None, 7, 37, 53, 64)	27712
MaxPooling3D	(None, 3, 18, 26, 64)	0
Flatten	(None, 89856)	0
Dense	(None, 128)	11501696
Dropout	(None, 128)	0
Dense	(None, 64)	8256
Dropout	(None, 64)	0
Dense	(None, 10)	650

- **Conv3D layers :** to extract time and spatial features from videos. Several layers with an increasing number of filters can be used to extract features that are more complex.
- **MaxPooling3D layers :** to reduce spatial and time dimensions and maintain the most important features.
- **Dense layers :** to convert extracted features into classifiable outputs. Dense layers are added with the activation of ReLU function to enhance the discriminatory capacity of the model.
- **Dropout layers:** to prevent over-generalization by dropping some neutrons randomly during training.

Table 3.9: Hyperparameters for the Model

Hyperparameter	Value
Dropout Rates	0,2 0.31
Epochs	20
Batch Size	16
Random State	42
optimizer	Adam
Loss Function	Categorical Crossentropy

- **Dropout Rates :** Two dropout layers with rates of 0.2 and 0.31 are used to reduce **overfitting.**
- **Epochs :** The model is trained for 20 epochs, providing ample opportunity to learn from the data.
- **Batch Size :** A batch size of 16 is used to balance memory usage and computational efficiency.

- **Random State:** A random state of 42 ensures reproducibility when splitting the data into training and test sets.
- **Optimizer:** The Adam optimizer is used for its efficiency and ability to handle sparse gradients.
- **Loss Function:** Categorical Crossentropy is chosen for its suitability in multi-class classification problems.

3.4.6 Results of experiment 3

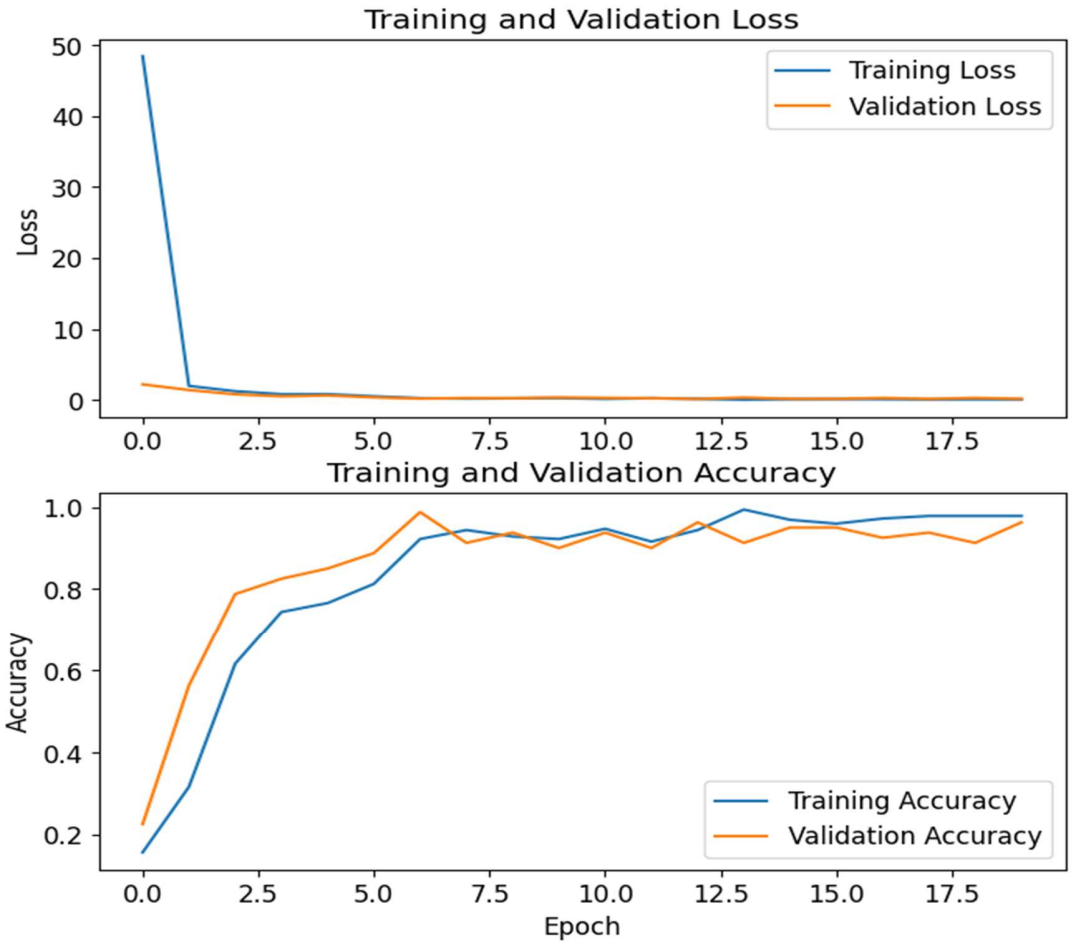

***Figure* 3.8:** Training and Validation Loss and Accuracy Curve

❖ **A Loss Curve Analysis**

Rapid Decrease in Loss: Both the training and validation loss decrease suddenly within the first few epochs and then stabilize, indicating that the model quickly converged to a good solution.

Stable Loss Values: The training and validation losses remain low and stable after the initial decrease, indicating no signs of overfitting or underfitting.

Increasing Accuracy: Both training and validation accuracy increase rapidly and then stabilize around high values (approximately 0.90 to 1.00).

Close Performance: The training and validation accuracy curves are close to each other, suggesting a well-generalized model without significant overfitting.

Table 3.10: Results for Each Class

Class	Precision	Recall	F1-Score	Support
train	1.0	1.0	1.0	12
restaurant	0.9	0.9	.09	10
lost	1.0	1.0	1.0	8
hurt	1.0	1.0	1.0	6
hungry	1.0	1.0	1.0	6
hospital	0.83	1.0	0.91	5
help	0.88	0.88	.088	8
emergency	1.0	0.83	0.91	6
danger	1.0	1.0	1.0	8
airport	1.0	1.0	1.0	11

The precision, recall, and F1-score metrics provide further insights into the classifier's performance:

- **High Precision and Recall:** Most classes have perfect precision and recall scores, leading to an F1-score of 1.0. This indicates that the classifier is highly reliable for these classes.
- **Lower Scores for Some Classes:** The hospital, help, and emergency classes have slightly lower precision, recall, and F1-scores, suggesting room for improvement in distinguishing these classes from others.

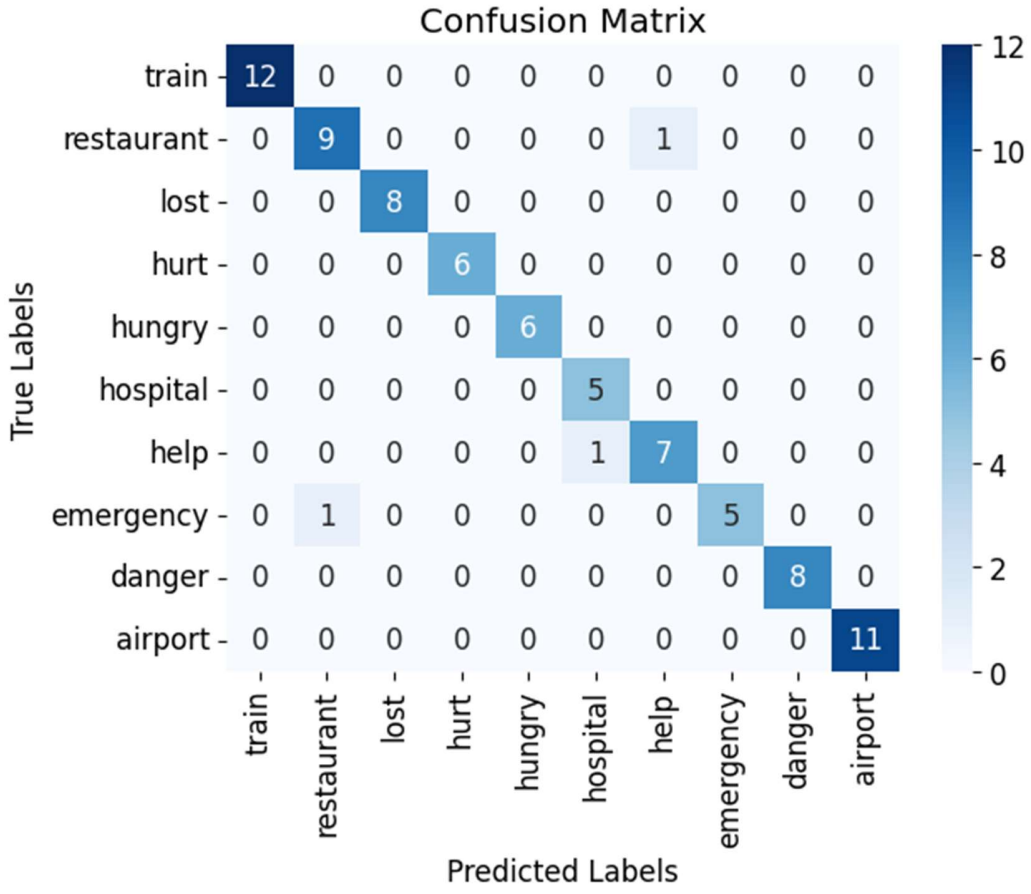

***Figure* 3.9:** Confusion Matrix

The confusion matrix provides a detailed breakdown of the classifier's performance for each class by showing the number of true positive, false positive, false negative, and true negative predictions. Here are the key points:

- **High Accuracy for Most Classes:** The majority of the classes, such as train, lost, hurt, hungry, danger, and airport, have perfect precision, recall, and F1-scores of 1.0. This indicates that the model is very accurate in predicting these classes.
- **Restaurant and Emergency Classes:** The restaurant class has a slight confusion with the help class, as shown by the 1 misclassification in the confusion matrix. Similarly, the emergency class has a slight confusion with the restaurant class.
- **Class 'help':** The class help has a recall of 0.88, indicating that 12% of the help instances were misclassified.
- **Class 'hospital':** The hospital class has a recall of 1.0 but a precision of 0.83, indicating that some instances from other classes were misclassified as hospital.

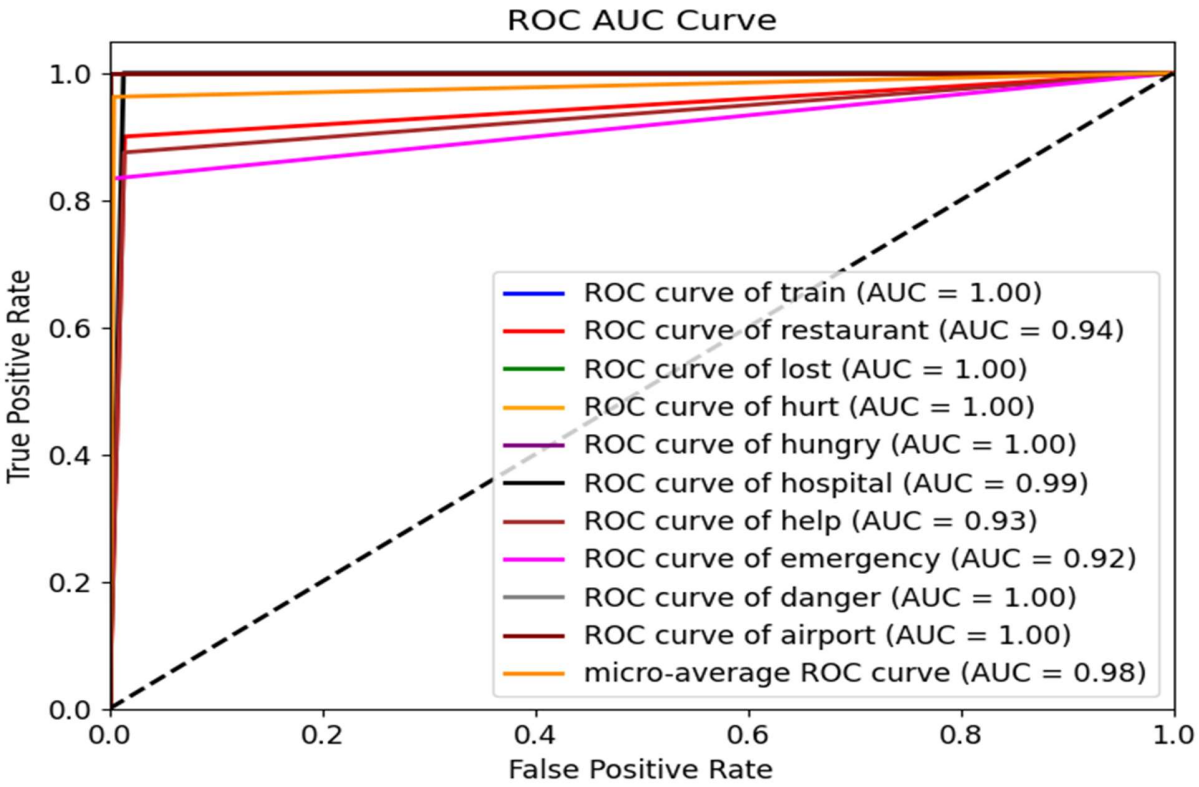

***Figure* 3.10:**ROC and AUC Curve

- **High AUC Scores:** Most classes have an AUC close to 1.0, indicating excellent classification performance. Classes like train, lost, hurt, hungry, danger, and airport have perfect AUC scores of 1.0. Classes like restaurant, help, and emergency have slightly lower AUC scores but still indicate good performance.

- **Micro-Average ROC:** The micro-average ROC curve, which aggregates the contributions of all classes, has an AUC of 0.98, showing overall excellent performance.

3.5 Conclusion

In this chapter, we started by introducing the materials and devices used in the realization of our lip-reading system. We then gave a detailed explanation of the pre-processing steps of the data and the structure of the model used, showing how each component contributes to the overall functionality of our system. The obtained results are discussed in details in the following section.

CONCLUSION

CONCLUSION

This research focused on developing a lip-reading transcription system, examining various aspects of data collection and processing, as well as the design and evaluation of the model used. The previous chapters detailed the methods and technologies employed, with a particular emphasis on practical implementation and the results obtained.

The first chapter provided an overview of Automatic Speech Recognition (ASR) technologies, vision-based systems (VB-ASR), and lip-reading transcription systems (LRT). It explored the theoretical foundations, the challenges faced, and the potential applications of these technologies, highlighting the advancements made and the ethical issues associated with their development and use.

The second chapter presented the methodologies for data collection and cleaning, as well as the essential parameters for optimizing the performance of the lip-reading model. Several data collection techniques were explored, from Arabic to English videos, to diversify and enrich the dataset. Emphasis was placed on the importance of data diversity and the necessary cleaning processes to ensure reliable and accurate results.

The third chapter, which is the core of our work, focused on the practical implementation of the lip-reading transcription system. It began with a detailed description of the materials used, particularly the Raspberry Pi 4 Model B, which served as the hardware platform for the system. The chapter then detailed the steps of data preprocessing, the model architecture, and the modeling techniques employed.

Preliminary experiments identified specific challenges related to the use of Arabic data, leading to the exploration of alternative methods to improve model performance. The final results showed significant improvements in the system's performance, with high

precision and recall scores for most categories, although some still require improvements. Confusion matrices and ROC curves provided a detailed evaluation of the model's performance, confirming the effectiveness of our methodology.

Through preliminary experiments, we identified the challenges associated with using the Arabic dataset and explored alternative approaches to enhance the model's performance. Our final experiences have shown significant improvements, as demonstrated by detailed results.

Our system has achieved high accuracy and reliability in most categories, with secondary areas for improvement identified in specific categories. The confusion matrix and ROC curves have shown a comprehensive assessment of the model's performance,

The results confirm the effectiveness of our proposed methods and the resilience of our system. High accuracy, F1-scores and recall in the different categories validate our approach potential for practical applications. Moving forward, future work can focus on addressing specific weaknesses, expanding the dataset, and exploring more advanced models to further enhance the system's capabilities.

Overall, this project demonstrated the feasibility of our lip-reading system, showcasing its potential for real-world publishing and paving the way for future progress in the field.

REFERENCES

REFERENCES

[1] : Alsulami, N.H., Jamal, A.T., Elrefaei, L.A. (2022). Deep learning-based approach for arabic visual speech recognition. Computers, Materials & Continua, 71(1), 85-108. https://doi.org/10.32604/cmc.2022.019450

[2] : Alper, B., et al. "Lip Reading Using Convolutional Neural Networks with and without Pre-Trained Models." Balkan Journal of Electrical and Computer Engineering, vol. 7, no. 1, 2019, pp. 1-8. DOI: 10.17694/bajece.479891

[3] : Peng, C., et al. "Lip Reading Using Deformable 3D Convolution and Channel-Temporal Attention." Proceedings of the International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2022, pp. 707–718. DOI: 10.1109/ICANN.2022.9911445

[4] : Erbey, A., & Barışçı, N. (2022). A Survey on Lip-Reading with Deep Learning. Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi / International Journal of Engineering Research and Development, 14(2), 844-860

[5] : https://www.mtapractice.com/2016/12/14/lip-reading-obstacles/

[6] : Faisal, M., & Manzoor, S. (2018). Deep Learning for Lip Reading using Audio-Visual Information for Urdu Language. ArXiv, abs/1802.05521.

[7] : Pandey, Laxmi. "Lip Reading as an Active Mode of Interaction with Computer Systems." 2024-05-01. eScholarship.org

[8] : https://en.wikipedia.org/wiki/Charles-Michel_de_l%27%C3%89p%C3%A9e

[9] : https://deafhistory.eu/index.php/component/zoo/item/1755-samuel-heinicke

[10] : https://www.britannica.com/science/deaf-history/The-19th-century

[11] : https://deafwebsites.com/lipreading/

[12] : Zarzycki, K., & Ławryńczuk, M. (2021). LSTM and GRU Neural Networks as Models of Dynamical Processes Used in Predictive Control: A Comparison of Models Developed for Two Chemical Reactors. Sensors, 21(16), 5625. https://doi.org/10.3390/s21165625

[13] : Fraser, S., Gagné, J. P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues. *Journal of speech, language, and hearing research: JSLHR*, *53*(1), 18–33. https://doi.org/10.1044/1092-4388(2009/08-0140)

[14] : https://www.theverge.com/2016/11/7/13551210/ai-deep-learning-lip-reading-accuracy-oxford

[15] : https://www.121captions.com/captioning-services-what-we-do/professional-lip-reading services/

[16] : Lu, Yuanyao, and Hongbo Li. 2019. "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory" *Applied Sciences* 9, no. 8: 1599. https://doi.org/10.3390/app908159

[17] : Chong, Teak-Wei & Lee, Boon Giin. (2018). American Sign Language Recognition Using Leap Motion Controller with Machine Learning Approach. Sensors. 18. 3554. 10.3390/s18103554.

[18] : https://rnid.org.uk/information-and-support/hearing-loss/hearing-implants/cochlear-implants/

[19] : https://www.neuralconcept.com/post/3d-convolutional-neural-network-a-guide-for-engineers

ملخص

قراءة الشفاه تشير إلى عملية فهم الكلام من خلال تفسير حركات الشفاه والتعبيرات الوجهية وغيرها من الإشارات البصرية دون الاعتماد على الصوت. تُستخدم هذه التقنية غالباً من قبل الأفراد الذين يعانون من نقص في السمع كوسيلة لفهم اللغة المنطوقة. يمكن تعزيز قراءة الشفاه بوسائل مختلفة، بما في ذلك التكنولوجيا.

في هذا العمل، نهدف إلى تصميم جهاز أوتوماتيكي يستطيع تحويل الكلمات المنطوقة إلى نصوص من خلال قراءة شفاه الأشخاص المتحدثين. قمنا في هذه الدراسة باستخدام العديد من الخوارزميات والبرمجيات لتجسيد النظام الذي توصلنا إليه حقيق هذا النظام.

Abstract

Lip reading refers to the practice of understanding speech by visually interpreting the movements of the lips, the facial expressions, and other visual cues without relying on sound. This technique is often used by individuals with hearing deficiencies as a way to comprehend spoken language. Lip reading can be enhanced through various means, including technology.

In this work, we aim to design an automatic device capable of converting spoken words into text by reading the lips of speakers. In this study, we have used numerous algorithms and software to develop the system we have achieved.

Résumé

La lecture labiale fait référence au processus de compréhension de la parole en interprétant les mouvements des lèvres, les expressions faciales et autres indices visuels sans se fier au son. Cette technique est souvent utilisée par les personnes souffrant de déficience auditive comme moyen de comprendre la langue parlée. La lecture labiale peut être améliorée par divers moyens, y compris la technologie.

Dans ce travail, nous visons à concevoir un dispositif automatique capable de convertir les mots prononcés en textes en lisant sur les lèvres des locuteurs. Dans cette étude, nous avons utilisé de nombreux algorithmes et logiciels pour concrétiser le système que nous avons développé.