1. INTRODUCTION

This chapter introduces automatic lip-reading processes, and the rise of technology for humans. It talks about the motivation behind this work, objectives and research contributions and possible applications. Finally, it ends with the outline of the thesis.

1.1 BACKGROUND

Since humans are social creatures, interacting with other people is essential to surviving. A person's character, cognitive abilities, way of life, comprehension, and many other aspects of human existence are shaped by this interaction. A person's communication style and volume with others can have a big impact on these things. In addition to aiding in the development of behavioral skills, effective communication is crucial for psychological growth.

God created mankind with the three fundamental senses that are necessary for communication: hearing, speech, and vision. Humans have developed a variety of languages to improve this capacity, which has increased the meaning, interactivity, and accessibility of communication.

Human vision, speech, and hearing must all be coordinated for effective communication. The brain combines auditory and visual information from the ears and eyes to form a coherent input. Vocal signals are produced by the brain once it has processed this input. In this sense, the generation of communication signals depends on the information that is perceived by the eyes and heard by the ears.

As a result, the sensory organs—the eyes and ears—provide the brain with most of its messages. The processed vocal or speech signal is produced by the mouth. Under the assumption that the sense organs are functioning normally, this coordination enables interaction with the external environment through quick and efficient processing.

For efficient communication, vision and hearing alone are usually adequate if the language is already learned. Hearing, however, is crucial to language acquisition, particularly for young children and infants. A baby or young child picks up bits of information from speech, arranges it, and learns a language [1]. Hearing is therefore essential in the early phases of language acquisition.

Effective communication might be hampered by any sensory impairment. Humans learn the act by watching, and using their sense of sight to acquire the abilities needed for daily tasks. The development of speaking abilities—which are necessary for interaction with the outside world—depends heavily on the speech signal. However, proper synchronization with hearing is crucial for the development of speaking abilities. Accurate hearing promotes language development, which improves speaking skills.

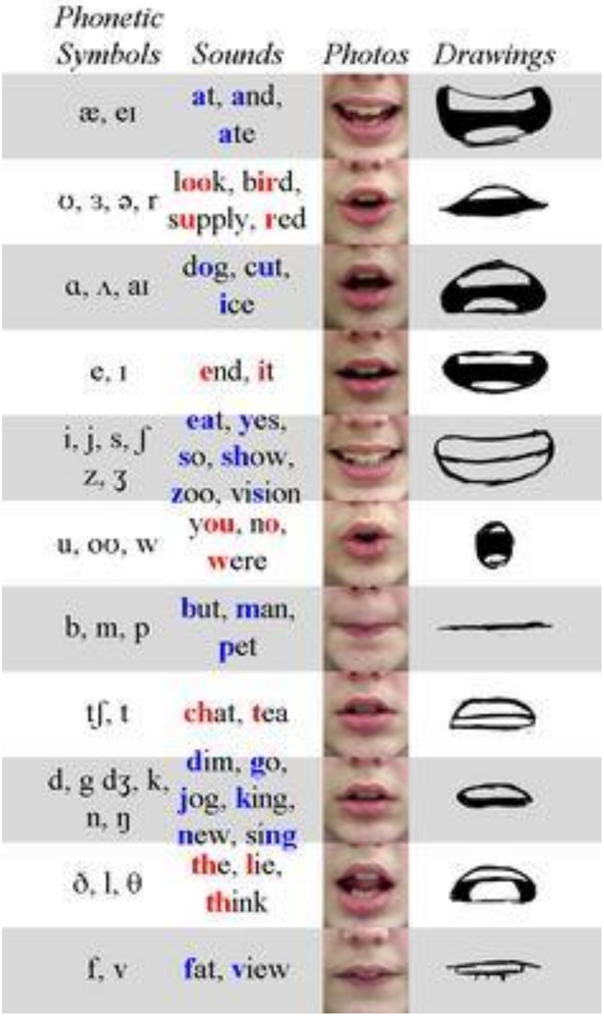
It becomes difficult to communicate if any of these skills are lacking. Communication might be nearly impossible in cases of severe hearing impairment as hearing is essential for proper speech and language development. Communication difficulties are highest in frequency for children who are born deaf or who become deaf before learning to talk [2]. They typically don't develop speech and communicate with other deaf people primarily through sign language, even though they might not have any innate speech issues.

Sign language is a visual language commonly used by deaf individuals for communication. It conveys meaning through a combination of body language, facial expressions, and hand gestures. Sign language was developed organically within deaf communities as a natural means of communication. Abbe de l'Épée, a French educator, established one of the first public schools for the deaf in Paris and developed a new method of teaching using sign language, known as Old French Sign Language (OFSL) [3]. Together with European deaf instructor Laurent Clerc, American educator Thomas Hopkins Gallaudet founded what is now known as the American School for the Deaf, the nation's first permanent school for the deaf, in the second decade of the 1800s. American Sign Language (ASL), one of the most extensively used sign languages in North America today, was created by adapting parts of French Sign Language, notably Old French Sign Language [4].

Lip reading is another way to compensate for the lack of auditory information, understand speech, and integrate with the hearing world [5]. Lip reading, also known as speechreading, is a skill which is developed and refined over time by deaf and hard-of-hearing people.

Although sign language is a valuable technique for communication within the deaf community, it may not be as effective when interacting with non-deaf individuals. While lip reading has its limitations, it can help bridge the communication gap between deaf and hearing people [6], [7], [8]. When deaf persons need to interpret signals from others who don't know sign language, miscommunications might happen [1], which can cause major problems, particularly in emergency situations.

Training deaf children to learn their mother tongue through lip reading and additional resources can improve their ability to communicate with hearing individuals [8]. In this approach, deaf or hard-of-hearing children focus on the lip movements of a speaker, creating mental images of the patterns and articulation of speech. Even though they cannot hear the sound, children can still perceive a significant amount of information from these visual cues, enabling them to gradually learn the language [9]. Lip reading in preschool can help deaf or hard-of-hearing children acquire their mother tongue, providing a solid foundation for future learning [10], [11]. Figure 1.1 shows lip movements and drawings for some English words. Along with these human efforts, technology has joined hands to assist in this process and benefit humanity.



Phonetic Symbols	Sounds	Photos	Drawings
æ, eɪ	at, and, ate		
ʊ, ɜ, ə, r	look, bird, supply, red		
ɑ, ʌ, aɪ	dog, cut, ice		
e, ɪ	end, it		
i, j, s, ʃ z, ʒ	eat, yes, so, show, zoo, vision		
u, oʊ, w	you, no, were		
b, m, p	but, man, pet		
tʃ, t	chat, tea		
d, g dʒ, k, n, ŋ	dim, go, jog, king, new, sing		
ð, l, θ	the, lie, think		
f, v	fat, view		

Figure 1 phonetic symbols, lip movements, and drawings for some English words (image courtesy by https://www.deviantart.com/)

1.2 RISING OF TECHNOLOGY IN HELP OF HUMANKIND

Although the concept of the computer was first conceived by Charles Babbage in the 19th century, significant advancements and research accelerated in the mid-20th century with the development of electronic computers. The ENIAC (Electronic Numerical Integrator and Computer) [12], built in 1940 was the first fully functional Electronic Computer. ENIAC was designed to solve high-speed calculations during World War-II. EDVAC (Electronic Discrete Variable Automatic Computer) [12] designed in 1949 used the stored program concept. UNIVAC (Universal Automatic Computer) [12] was designed for business applications and data processing tasks such as statistical analysis. It became famous for correctly forecasting the results of the 1952 U.S. presidential election [13]. The field of computer technology has progressed through significant stages: from transistors and second-generation computers to high-level programming languages and integrated circuits; from third-generation computers and time-sharing minicomputers, to personal computers, operating systems, and graphical user interfaces (GUIs), culminating in the modern digital age. Advancements in hardware, software, networking, and other related fields have significantly shaped the evolution of computers, leading to the versatile and powerful systems we use today [14], [15].

The ambition to design machines that mimic human behaviour, combined with significant advances in technology, algorithms, and computing power, gave rise to a new concept in Automatic Speech Recognition (ASR) after the mid-20th century. Early attempts at Automatic Speech Recognition (ASR) design were based on discrete sound patterns or acoustic features. The first ASR system designed in 1952, Davis, Biddulph, and Balashek of Bell Laboratories built a system named ‘Digit recognizer’ [16]. The second model of digit recognizer based on more frequency bands was designed by Dudley and Balashek in 1958 [17]. In 1954, Sumby and Pollak [18] tested the impact of visual observation of a speaker’s facial and lip movements on oral speech intelligibility and found that a speaker’s facial and lip movements can significantly improve oral speech intelligibility. While this research opened the doors for considering visual cues in speech recognition, image processing techniques were not well-defined or widely developed at the time. As a result, researchers primarily focused on ASR [19], [20], [21], [22], and [23] which involves analysing and interpreting speech signals captured by microphones without considering visual information such as lip movements or facial expressions [24]. Meanwhile field of image processing continued to expand from fundamental image processing algorithms to digital image processing techniques. In 1984, Petajan's work on lip reading is cited as one of the earliest efforts to systematically explore the use of visual information from lip movements to improve video-based automatic speech recognition (Video-ASR) [25]. This has been discussed in detail in chapter 2 of Literature Study.

Work in Video-ASR started in the second last decade of the 20th century, but in the last three decades, it accelerated due to proliferation in Digital Imaging techniques, advancement in Computational power and emerging approaches like machine learning and deep learning.

1.3 MOTIVATION, PROBLEM STATEMENT, OBJECTIVES, SCOPE, CONTRIBUTIONS, AND APPLICATIONS

1.3.1 Motivation

The motivation for this research stems from several key factors:

1. **Language Preservation and Accessibility**: Enhancing the accessibility of the Gujarati language through advanced technological solutions can support language preservation and promote its use in digital communication.
2. **Advancement in Speech Recognition**: Focusing on visual information alone, particularly 2D images with depth, can significantly improve speech recognition systems by accurately interpreting lip movements. This approach is especially beneficial in noisy environments or for individuals with hearing impairments, where reliance on auditory information alone may be insufficient.
3. **Technological Innovation**: Utilizing advanced machine learning techniques, including CNN-LSTM models and MobileNet [26], to advance the capabilities and effectiveness of lip reading technology.
4. **Personal and Educational Applications**: Developing tools that can aid in language learning and assistive technologies, particularly for children with hearing impairments. By learning their mother tongue through lip reading at an early age, deaf children can build a strong foundation that enables them to learn other languages through lip reading as well. This can greatly enhance their communication abilities and educational opportunities.

1.3.2 Problem Statement

1.3.3 Research Objectives

To achieve the noble goal, the following objectives are set:

1. To design and develop a Gujarati Alphabet Dataset.
2. To design, develop and test lip extraction algorithm(s).
3. To design and develop algorithms in such a way so that it/they remain(s) generic and can be used for lip detection and extraction in general.
4. To develop a test program for the proposed algorithm(s) to ensure that the objectives are met.

1.3.4 Research Scope and Contribution

1. **GJVarna Dataset Creation**: The focus of this research is to create a complete dataset for the Gujarati alphabet as no dataset is available for the Gujarati language. Videos are recorded for speakers of different age groups for 34 consonants of Gujarati Alphabets ensuring data quality. Following the 1st Objective, a dataset named GJVarna has been created with 51000 images.
2. **ViLiDEx Algorithm**: Implementing a pre-processing pipeline using Dlib [27] for facial landmark detection and frame selection to ensure that only relevant frames are used for training and testing. Alphabet utterances of different speakers may vary in time. For a long utterance, the total number of frames is more compare to the short utterance, and if the first odd/even frames are kept, key frames may be discarded. So, fulfilling the 2nd objective is to design an algorithm for extra frame removal and dataset creation. The algorithm has been named - ViLiDEX. ViLiDEx algorithm has been used to remove extra frames and store key frames, ensuring consistent and relevant data for model training.
3. **Model Utilization, Validation and Evolution**: MobileNet, a pre-trained model, has been used for alphabet classification and recognition based on five classes: Guttural, Palatal, Retroflex, Dental, and Labial. The model is applied to the GJVarna dataset, which is divided into 66% for training and 33% for testing. Cross-validation techniques have been to ensure the accuracy and robustness of the

MobileNet model in recognizing the Gujarati alphabets. Metrics such as accuracy, precision, recall, and computational efficiency have been used to assess the model.

1.3.5 Applications

A few possible applications of this work could be:

1. The GJVarna can be used as a base for Lip Reading work in the Gujarati language as it is the first ever dataset created for alphabets of the Gujarati Language.
2. ViLiDEx algorithm can be used to remove extra frames and store keyframes, ensuring consistent and relevant data for model training for any other language alphabets too with few customizations.
3. Pre-trained MobileNet model with GJVarna can be used as a base for further research in the Gujarati language.

1.4 OUTLINE OF THESIS

Chapter 1 is about the introduction of lip reading. It includes how deaf children are trained without and with lip reading. It tells the history of how technology has evolved and implemented lip reading. This chapter also covers motivation, problem statement, objectives, scope, research contributions and applications. Chapter 2 carries out the literature study for Automatic Lip Recognition (ALR), which includes the foundation of ALR, steps, methods, and various approaches of ALR. It also includes various datasets and languages for which ALR has been developed. It also discusses ALR work carried out for Indian languages and Gujarati language. Chapter 3 explains the proposed ViLiDEx algorithm and dataset created for the Gujarati language. Chapter 4 carries out and explains a detailed analysis of the testing of the algorithm. In Chapter 5, the conclusion, limitations and future roadmap have been discussed. The thesis ends with chapter 6 listing out the publications and chapter 7 containing the references that have been used for this work.