IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

"AI-BASED SIGN LANGUAGE TO TEXT AND SPEECH CONVERTER"

[1]Supriya Shinde ,[2]Saloni Vernekar, [3]Gayatri Lohar, [4]Ashlesha Powar,[5]Mr.Sagar Chavan

[1]Student, [2]Student, [3]Student, [4]Student, [5]HOD

[1] Department Of Computer Science And Engineering ,

[1]Sanjay Ghodawat Institute (Atigre), Kolhapur, India

Abstract: The aim of this project is to build a smart system that can turn sign language into spoken words and written text. It is designed to help people who are deaf or cannot speak communicate more easily with those who do not understand sign language. Most sign language translator devices today are expensive and can only show text. To solve this, the system uses modern technologies such as Artificial Intelligence (AI) and Machine Learning (ML)[2]. These allow it to quickly recognize hand gestures and convert them into clear speech and readable text. The system uses computer vision to track and understand hand movements, along with a speech module that produces natural-sounding voice. It is simple, affordable, and easy to carry, making it useful in everyday situations like schools, hospitals, and public places. Another important feature is that the system can improve over time. As it collects more data, it can learn new signs and support different sign languages, becoming more accurate. We extract optical flow features based on human pose estimation and, using a linear classifier, show these features are meaningful with an accuracy of 80%.[1]

I. INTRODUCTION

Communication plays a very important role in our everyday life because it allows people to share ideas, feelings, and information with one another. For people who have difficulty hearing or speaking, expressing themselves can be hard. They often use sign language to communicate, but most people do not understand it[5]. This makes it difficult for them to interact with others easily.Today, technology, especially Artificial Intelligence (AI) and Machine Learning (ML), can help reduce this problem. We already use tools like Google Translate to understand different spoken languages, but these tools cannot understand sign language. Relying on a human translator is also not practical or affordable for daily use.This project is designed to solve that problem by creating a system that can automatically recognize and understand sign language. The system can quickly and accurately convert sign language gestures into text or speech in real time. The Realtime Sign Language Detection Using LSTM Model is a deep learning-based project that aims to recognize and interpret sign language gestures in real-time.[3]The main goal of the "Sign Language Detection" project is to make communication easier for people with hearing or speech difficulties and help them connect with those who do not know sign language.

II. LITERATURE REVIEW

The paper **"**Fingerspelling Recognition in the Wild with Iterative Visual Attention**"** by Bowen Shi and colleagues explains how recognizing sign language, especially fingerspelling, is a complex task. This is because signers move their hands very quickly, and the signs often blend together. The researchers worked on recognizing fingerspelling in American Sign Language (ASL) using videos taken from real-world sources like YouTube and social media, instead of controlled studio environments. This makes their study more realistic since it deals with different lighting conditions, camera angles, and signer variations that occur naturally.On the other hand, the paper **"**Real-time Sign Language Fingerspelling Recognition using Convolutional Neural Networks from Depth Map" by Byeongkeun Kang and his team focuses on using Convolutional Neural Networks (CNNs) to recognize fingerspelling through depth maps (which capture how far the hands are from

the camera). Their system could recognize 31 different letters and numbers, achieving very high accuracy—99.99% for signers it had already seen and around 83–85% for new signers. The study also found that the system's performance improved as more data from different people was used for training.Building on these ideas, the proposed project—AI-based Sign Language to Speech and Text Converter—goes a step further. Instead of only recognizing gestures, it also translates them into both text and spoken words. This means that when a person uses sign language, the system will instantly display and say what they mean, helping them communicate more easily with people who do not know sign language. By combining computer vision, machine learning, and speech synthesis, this system creates a real-time communication tool that is practical, inclusive, and affordable for everyday use. It aims to break down the communication barrier between hearing-impaired individuals and the rest of society.

III. OBJECTIVES

This project aims to create a simple and practical tool that makes it easier for people who use sign language to communicate with those who don't understand it. The first step is to gather a large collection of sign language gestures, including letters, numbers, and everyday words or phrases. This collection will be used to train an AI system to accurately recognize different hand movements and positions[2].When a sign is detected, the system will instantly convert it into natural and clear speech. This means that as soon as a person makes a sign, the system will speak the message out loud, allowing both sides to communicate quickly and easily.A key feature of the project is its ability to work in multiple languages. After recognizing a sign, the system can translate it into different languages and display or speak the translated message. This makes the tool useful for people from different countries and language backgrounds.The system is designed to be lightweight and run smoothly on everyday devices such as smartphones and laptops, without needing costly or special equipment[3]. Its portable and user-friendly design ensures it can be used comfortably anywhere.In short, the main purpose of this project is to bridge the communication gap between people who are deaf or hard of hearing and others, by creating a real-time AI system that can understand and translate sign language using a standard camera.

IV. Technical Constraints

4.1 Hardware And Device Limitations

The performance and usability of the AI-Based Sign Language to Speech and Text Converter depend a lot on the device it runs on. Since the system uses computer vision and deep learning to recognize hand gestures, it needs enough processing power, good camera quality, and sufficient memory to work smoothly in real time.The system works on devices with cameras, like laptops, smartphones, or webcams[4]. The quality of the camera and its frame rate are very important for accurate gesture detection. Cameras with low resolution or low frame rates can make gestures appear blurry or incomplete, which lowers recognition accuracy.Although the system is designed to be portable and work on different devices, its speed, accuracy, and reliability still depend on the device's hardware. To overcome these limitations, techniques like model compression, quantization, and edge AI acceleration can be used. These methods help the system run efficiently on a wider range of devices without slowing down.

4.2 Data And Model Limitations

The success of the AI-Based Sign Language to Speech and Text Converter depends a lot on the quality, variety, and size of the datasets used to train its AI models. However, there are some challenges that can affect how accurate and reliable the system is in real-world use[3].One major challenge is data diversity. Many sign language datasets are recorded in controlled environments, with consistent lighting, plain backgrounds, and only a few people signing. When the system is used in real-life situations—where lighting, camera angles, skin tones, and backgrounds vary—its accuracy can drop. While techniques like data augmentation and transfer learning can help, they cannot completely replace the variety found in real-world data.Another challenge is data labeling and annotation. Creating and marking these datasets is time-consuming and sometimes inconsistent. Differences in signing speed, hand positions, and motion blur can introduce errors into the data, making the AI model less reliable.To improve the system, creating standardized datasets and encouraging collaborative, open-source contributions could help provide more diverse, high quality data and make the AI model more accurate and dependable.

4.3 Network Constraints

The performance and usability of the AI-Based Sign Language to Speech and Text Converter can be affected by internet connection, especially when it uses online processing or cloud-based speech generation. While the core gesture recognition can run on the device itself, features like cloud AI processing, updating datasets, or multilingual text-to-speech need a reliable internet connection[5].In everyday use, slow or unstable internet and low bandwidth can interrupt the smooth conversion of signs into speech or text. Real-time gesture detection often uses high-quality video, which requires a steady flow of data between the camera and the system. If the connection is weak, users may experience delays, missing frames, or mistakes in recognizing gestures.This problem is even more noticeable in areas with poor internet access, such as rural or remote locations. Online text-to-speech services, like Google Text-to-Speech, need a continuous internet connection. Without it, speech output may fail or switch to an offline mode, which delivers lower-quality audio.

V.Operational Design

5.1 Recognition accuracy

Recognition accuracy means how correctly a system can identify the hand gestures shown to it. It helps to check how well the gesture recognition system is working. If the accuracy is high, it shows that the system can easily recognize different hand shapes, movements, and directions, even when lighting or background conditions are not perfect[1]. The accuracy depends on things like the quality of the data used, how well the system has been trained, the clarity of the camera, and the environment where it is used. For smooth and reliable communication, the system should have an accuracy level of about 90–95% or higher.

5.2 Response Latency

Response latency is the time a system takes to respond after someone shows a gesture. In simple words, it is the short pause between making a gesture and the system turning it into words or text. When the system reacts quickly, communication feels natural and smooth, making it easy to talk and understand each other[6]. If the system is slow, it can interrupt the conversation and make it harder to follow. To make the system faster, it should recognize gestures quickly, work efficiently, and send data without delays. Ideally, the system should respond in less than 200 milliseconds to keep communication clear and in real time.

5.3 Speech Quality

Speech quality is about how clear, natural, and easy to understand the voice is when it is generated by a Text-to-Speech (TTS) system. It makes sure that the spoken output correctly matches the intended words, tone, and pronunciation. If the audio is unclear or sounds unnatural, it can lead to confusion or make communication difficult.To measure speech quality, a method called the Mean Opinion Score (MOS) is often used[7]. In this method, people listen to the audio and rate it based on how natural, clear, and smooth it sounds.Modern TTS systems use advanced AI models like Tacotron 2 and FastSpeech 2, which help make the generated voice sound more realistic, expressive, and closer to a real human voice.

5.4 User Experience

User experience refers to how easy and comfortable it is for people to use the system. It looks at how simple the interface is, how quickly users can understand it, and how easily they can perform tasks such as changing the language, adjusting speech speed, or switching display modes.A good user experience means the system should have a clean and simple design, respond quickly, and provide clear visual feedback in real time[6]. It should also be easy to use for people of all ages and abilities, even for those who are not familiar with technology.When users find the system smooth, accurate, and convenient to use, it shows that the design successfully balances performance, reliability, and simplicity—leading to higher user satisfaction.

VI. METHODOLOGY

6.1 Collecting Data

The first step is to gather a large number of videos featuring people using sign language. These videos can be sourced from publicly available collections or created specifically for the project[2]. Having recordings of different individuals is essential so the system can learn to recognize various styles and ways of signing.

6.2 Preparing the Videos

Once the videos are collected, they need to be prepared for further processing[4]. This involves breaking the videos down into single frames—like snapshots—and ensuring all images are resized consistently[1]. Any unnecessary background distractions are removed to help the system concentrate on the hands and face, which carry the meaning in sign language.

6.3 Extracting Important Details

A specialized program known as a Convolutional Neural Network (CNN), such as ResNet50, is used to analyze each frame[7]. This tool identifies important features like the shape of the hands, finger positions, and facial expressions—key elements that convey meaning in sign language.

6.4 Tracking Motion Over Time

Sign language involves a sequence of movements, so another type of model—either an LSTM or a Transformer—is used to understand how these frames connect over time. This allows the system to grasp the flow and timing of gestures, which is crucial for accurate interpretation.[5]

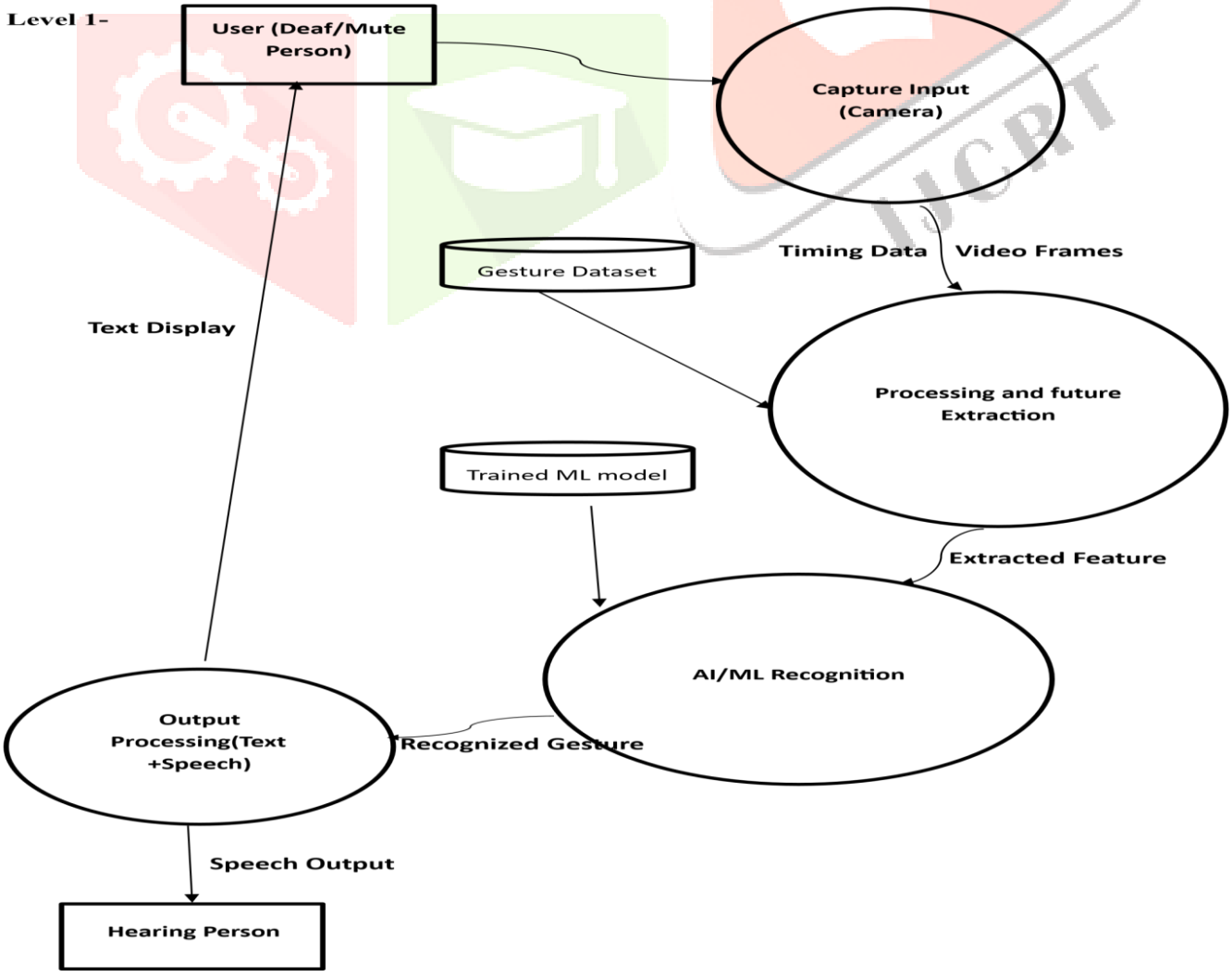
6.5 Interpreting the Signs

After the system learns the movements and features, it translates the gestures into corresponding words or letters. This step effectively converts the visual signs into written language that can be understood.

6.6 Converting Text to Speech

Finally, the written words are transformed into spoken language using a Text-to-Speech (TTS) engine. The spoken output can be played aloud or displayed as text, helping others to understand the signed message easily.

6.7 System Architecture



VII. Benefits

7.1 Instant Communication Without Using Hands Constantly

This system helps people with hearing challenges talk right away, without always needing to rely on hand gestures[5]. It makes conversations smoother and more natural, especially in places where sign language is common.

7.2 Works with Normal Devices

The system can operate using regular cameras on phones and computers. Since it doesn't require any special or costly tools, it can be used easily at home, work, or anywhere else. The majority of the methods use one or two CNN models integrated to their applications.[7]

7.3 Low Cost and Easy to Expand

It doesn't cost much to set up or use, which makes it a practical choice for individuals, communities, and organizations. Plus, it can be quickly expanded to help more people as needed.

7.4 Promotes Fair Access for Everyone

This technology helps hearing-impaired people in important areas like:[5]
Customer Support: They can get help and information just like everyone else.
Healthcare: It improves communication between doctors and patients, ensuring patients receive accurate care without confusion.

VIII. LIMITATIONS

Despite progress in recent years, sign language recognition systems still face many challenges. One of the first issues is that the system can have trouble when hand movements aren't clear or overlap with each other. If the gestures aren't shown properly, the system might misinterpret them or miss the meaning altogether. The clarity of the video plays a huge role too—poor lighting or low-resolution footage can make it hard for the system to work accurately[2].Another challenge is picking up facial expressions and emotional cues, which are essential parts of signing. Many systems simply can't read faces well enough to understand these non-verbal elements. Also, in order to teach the system to recognize full sentences, developers need very large datasets that are already labeled and organized—which can take a lot of time and effort to collect.Some signs are more complex and involve fast or multi-step motions. These types of gestures are harder for the system to recognize correctly, especially in conversations where signs flow quickly and change with the context. The technology often struggles to follow along in real-time and misses the meaning behind the signs.On top of that, not everyone signs the same way. Some people sign quickly, while others move more slowly or add personal flair. These differences in signing styles can confuse the system if it hasn't been trained to handle that variety. Most systems are also built around just one version or dialect of sign language, so they might not recognize signs used in other regions.Vocabulary is another limitation. If someone uses signs that the system hasn't seen before, it won't know how to interpret them. Accurate recognition also depends on how well the system can follow hand and finger movements. Small errors here can change the meaning of a sign entirely. Nuances in how a sign is made—like speed, direction, or shape—can be subtle but important, and the system might miss them.There's also the issue of recognizing things beyond hand signs, such as head movements, eyebrow raises, or posture—things known as non-manual markers[6]. The whole machine learning process has some limitations to it like Poor quality of data, Lack of training data, Slow implementation, etc.[4] Many systems either ignore these cues or don't detect them properly. A noisy or cluttered background can also throw the system off, making it harder to focus on the person signing.Finally, these systems need strong hardware to run well. Real-time translation, in particular, demands a lot of computing power[3]. Even with all the right tools, mistakes still happen during live interpretation, which can affect the flow of communication when speed and accuracy are crucial.

IX. Future Scope –

9.1 Understanding Complete Sentences

Most current systems can only catch individual signs, like a single word or letter. A big step forward would be teaching these tools to understand full sentences and the meaning behind them. This would make conversations flow more naturally, just like regular speech.

9.2 Reading Facial Expressions and Body Language

Sign language includes much more than just hand movements. Facial expressions, lip shapes, and even how someone moves their body are key to understanding what they mean. A smarter system should be able to recognize these parts too, so it captures the full message.[2]

9.3 Translating into Different Spoken Languages

After recognizing the signs, the system could be designed to speak the message aloud in different languages. This way, people from various countries could understand it, making the tool useful on a global level.

9.4 Making the System Easy to Carry and Use

If the system could run on a smartphone or a small device, people would be able to take it with them wherever they go. Whether they're at work, on the street, or at home, they could use it without needing any special equipment.

9.5 Real-Time Conversations Across Distances

Using cloud technology, the system could help people communicate even if they're far apart[3]. Someone could sign in one place, and the system would instantly speak their message to someone else, miles away. This would be great for remote conversations.

9.6 Adjusting to Personal Signing Styles

Just like everyone speaks differently, everyone also signs in their own way[6]. A more advanced system could learn how a specific person signs and get better at understanding them over time. This would make it much more accurate for each individual user.

9.7 Using It in Education, Healthcare, and the Workplace

This kind of technology would be a huge help in schools, hospitals, and offices[7]. It would make it easier for Deaf and hard-of-hearing people to join discussions, ask questions, or get important information—leading to more inclusive and accessible environments for everyone.

REFERENCES

[1] A. Moryossef, I. Tsochantaridis, and R. Aharoni, "Real-time sign language detection using human pose estimation," in Computer Vision – ECCV 2020 Workshops, E. Ricci, S. R. Bulo, C. Snoek, O. Lanz, S. Messelodi, and N. Sebe, Eds. Cham: Springer, 2020, pp. 1–17. [Online].

[2] GitHub, "Sign-language-recognition-system · GitHub topics," GitHub, 2024. [Online].

[3] A. Adhikary, "Realtime sign language detection using LSTM model," GitHub repository, 2023. [Online].

[4] M. Rafay, "Sign language recognition using Python & MediaPipe," Medium, Jul. 18, 2023. [Online].

[5] M. Garimella, "Sign language recognition with advanced computer vision," *Towards Data Science*, Mar. 28, 2022. [Online].

[6] Roboflow 100, "Sign language object detection model," Roboflow Universe, 2023. [Online].

[7] Artificial intelligence technologies for sign language," Frontiers in Robotics and AI, vol. 8, Art. no. 697336, 2021. [Online].