JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

VISUAL GESTURE TO AUDITORY SPEECH CONVERTER USING DEEP LEARNING

Mrs. V Lakshmi
Assistant Professor
Dept of CSE, REC
Visakhapatnam, Andhra Pradesh

I Kalyan Sai
(Backend Developer)
Department of CSE,
REC, Visakhapatnam

B Sai Sahitya
(DL Developer)
Department of CSE,
REC, Visakhapatnam

K Jaya Chandra
(DL Developer)
Department of CSE,
REC, Visakhapatnam

T Chaitanya Kumar
(Frontend Developer)
Department of CSE,
REC, Visakhapatnam

ABSTRACT

Human beings interact with each other to convey their ideas, thoughts and experiences to the people around them. But this is not the case for deaf-mute people. Sign language paves the way for deaf-mute people to communicate. Through sign language, communication is possible for a deaf-mute people. The aim behind the work is to develop a system for recognizing the sign language, which provides communication between people with speech impairment and normal people, thereby reducing the communication gap between them. Compared to other gestures(arm, face, head and body), hand gestures plays an important role, as it expresses the user's views in less time. In the current work flex sensor-based gesture recognition module is developed to recognize English alphabets and few words and a Text-to-Speech synthesizer based on Transfer learning, NLP, Activation functions, Image recognition, Neural networks, Recurrent neural networks is built to convert the corresponding text.

Keywords: Sign Language, Deaf-mute communication, Gesture Recognition, Hand Gesture, Text-to-Speech Synthesizer, Natural Language Processing, Image Recognition, Recurrent Neural Network, Random Forest

I. INTRODUCTION

Communication is a fundamental human ability, enabling individuals to share thoughts, ideas, and emotions. However, for deaf-mute individuals, conventional communication methods present significant challenges. Sign language bridges this gap, offering a structured way for individuals with speech and hearing impairments to convey their messages.

Despite its effectiveness, the reliance on sign language creates a communication barrier with people unfamiliar with it. To address this, technological advancements can be leveraged to interpret sign language into forms accessible to all, such as text and speech. This project aims to develop a system that facilitates seamless communication between deaf-mute people and the general population. By focusing on hand gestures - the most expressive component of sign language - the system converts gestures into text and speech using advanced methodologies, Through the integration of flex sensor-based modules, machine learning techniques, and natural language processing(NLP), this solution aspires to minimize the communication gap and promote inclusivity in everyday interactions.

1. **Access the Social Media Link:** The process starts with a social media link. The user will get a link provided for the application.
2. **Login Page:** If the user already registered, then the user will be directed to the login page. If not, the user needs to click on the register button and then they are redirected to the registration page.

3. **Registration Page:** The new users needs to complete their registration process by entering their details. After successfully entering their details, the users credentials will be saved in the database and redirected to the login page.

4. **Login and Home Page:** Registered users login to the system using their credentials. After login, they are directed to the homepage, which contains a start button.

5. **Start Button:** Clicking on the start button navigates the user to the main functionality page of the application.

6. **Main Page-Camera Access:** The system requests access to the device camera. After granting access, the user can display hand gestures infront of the camera.

7. **Gesture Detection:** The camera captures the users hand gestures in real-time. The system processes these gestures to recognise specific letters or words.

8. **Output Display:** The recognised letter or word will be displayed on the screen with the speech given to the text.

II. LITERATURE SURVEY

“Gesture-Based Communication System” by Kohsheen Tiku; Jayshree Maloo; Aishwarya Ramesh; Indra R . This research aims on developing the system that recognises the hand gestures of an user and the recognised gestures translates into the speech using deep learning techniques. The study highlights the significance of the sign language for the deaf-mute people and used to bridge a communication gap. The proposed system combines gesture recognition with AI-based speech synthesis, making the interactions more realistic and effective. The use of Recurrent Neural Networks(RNNs) and Natural Language Processing(NLP) improves accuracy and real-time processing. The study emphasizes how such systems can improve accessibility in education, healthcare and daily life.

"Hand Gesture Recognition using Convolutional Neural Networks(CNNs)” by Patwary; Muhammed J. A & Parvin; Shahnaj & Akter; Subrina. A previous study explored hand gestures recognition using CNNs to classify different sign language gestures. The researchers took the hand images as a dataset and trained a deep learning model to achieve high accuracy in gesture’s classification. Their findings manifest that CNN-based models outperform traditional machine learning techniques in recognizing compound hand shapes. The study highlighted the challenges such as varying lighting conditions and hand obstruction, which affect the recognition accuracy. The result suggested that improving dataset diversity and real-time optimization could enhance gesture recognition performance.

"Deep Learning for Speech Synthesis” by ASL Reverse Dictionary - ASL Translation Using Deep Learning Ann Nelson Southern Methodist University, alnelson@mail.smu.edu KJ Price Southern Methodist University, kjprice@mail.smu.edu Rosalie Multari Sandia National Laboratory, Another study investigated the role of deep learning in converting the text to speech, especially for assistive communication tools. Researchers developed a model using Long Short- term Memory (LSTM) networks to generate the realistc speech for the recognised text. Their findings showed notably improved speech clarity and pronounciation compared to rule-based speech synthesis methods. The study also addressed concerns like voice modulation and tone variations to make synthesized speech. The researchers wrapped up that combining deep learning with NLP techniques enhances the speech generation accuracy.

"Sign Language Recognition using Sensor-Based Gloves” by Dumit rescu. & Boiangiu. The study explored the use of sensor-based gloves for sign language recognition, where the flex sensors track the finger movements and translate them into text. To transmit the gesture data to a processing unit, the system used microcontrollers and Bluetooth. The study identifies that sensor-based recognition delivers high accuracy in controlled surroundings but struggles with real-world adaptability.

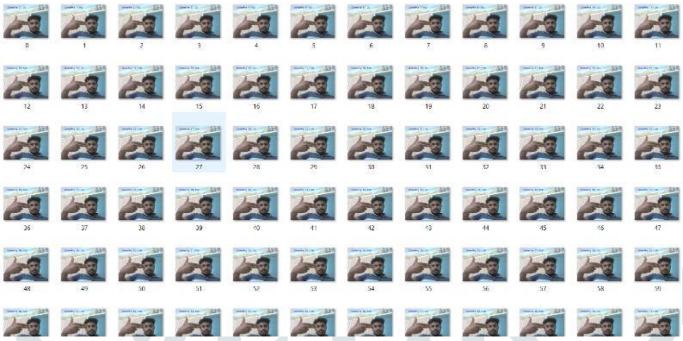
"Recurrent Neural Networks for Sequential Gesture Prediction” by Saeed; Khalid & Tabedzki; Marek & Rybnik; Mariusz & Adamski; Marcin. This research focused on using RNNs for sequential gesture recognition, particularly for predicting continuous sign language phrases. The model was trained on video frames of sign language and learn to predict the gesture patterns. The study manifested that RNNs effectively capture temporal dependencies in sign language, improving translation accuracy. However, limitations like computational complexity and training time were noted. The researchers suggested optimizing neural network architectures to make real-time gestures translation more efficient.

III. METHODOLOGY

The methodology of this project follows a systematic approach to convert visual gestures into auditory speech, utilizing deep learning and computer vision techniques.

1. Data Collection and Preprocessing

- The first step involves data collection and pre-processing. Here we collect the combination of static gestures such as alphabetic signs, and dynamic gestures representing words and phrases. These datasets are self-made in order to ensure consistency in processing, The collected dataset is resized to uniform dimensions and normalised to correct the variations in lighting and contrast, enabling a uniform input for further stages.
-



2. Feature Extraction :

- Employed a CNN model architecture comprising convolutional layers for feature extraction, ReLU activation functions for non-linearity, max-pooling layers for downsampling, and fully connected layers for classification.
- Trained the model using batch-wise processing with the Adam optimizer and categorical cross-entropy loss function, iterating over multiple epochs to adjust model weights.

3. Classification:

- Once the system has pinpointed the key characteristics of a gesture, it is time to figure out what that gesture actually means. For still or static hand positions, we use a technique called Support Vector Machines(SVMs).
- Think of it as a way to nearly sort different hand shapes into distinct categories. When it comes to gestures that involve movement like waving or swiping. Because these gestures unfold overtime, we need something that can understand the sequence of movements. That is why Recurrent Neural Networks(RNNs) or Long Short-term Memory(LSTM) comes in. They are designed to recognize patterns in data that changes over time, allowing the system in accurately interpret the flow of a dynamic gesture.

4. Speech Synthesis:

- After the system figures out which gesture was made, it translates that gesture into words or sentences. Then to make those words audible, it uses sophisticated technology that turns text into spoken language.
- This isn't just robotic speech, it is designed to sound clear, natural and even convey emotion. The goal is to provide users with a smooth and easy way to understand the system's responses as if they were having a conversation.

5. Frame Capture:

- First, the system needs to see what the user is doing. So, it uses a camera to record a live video. It then breaks that video down into individual pictures, or frames which it examines one by one.
- This is really important because it allows the system to track how hand movements change over time, giving it the information it needs to correctly understand the gestures.

6. Segmentation:

- To focus specifically on the hand and ignore everything else in the picture, the system uses a technique that identifies skin color. Essentially, it looks for pixels that fall within a certain range of color values similar to how our eyes recognise skin tones.

- By setting specific limits for these specific color values, it can effectively separate the hand from the background and other objects. This process cleans up the image, removing distractions and highlighting the hand movements which ultimately helps the system understand the gestures more accurately.

7. Gesture Recognition:

- Once the system has isolated the hand, it moves on to figuring out what the gesture means. If it is a still hand position like a sign language letter. It uses a pre-learned model which is a kind of smart sorting tool called a Support Vector Machine(SVM) to identify the specific hand shape. Each shape matches a particular letter or symbol.
- For gestures that involve movement, like waving or signing a phrase, the system examines a series of video frames. It uses sophisticated techniques including 3D Convolutional Neural Networks(CNNs) and Long Short-term Memory(LSTM) to understand the patterns of motion. These networks are especially good at recognizing how gestures unfold over time, allowing the system to accurately interpret the meaning of a moving hand signal.

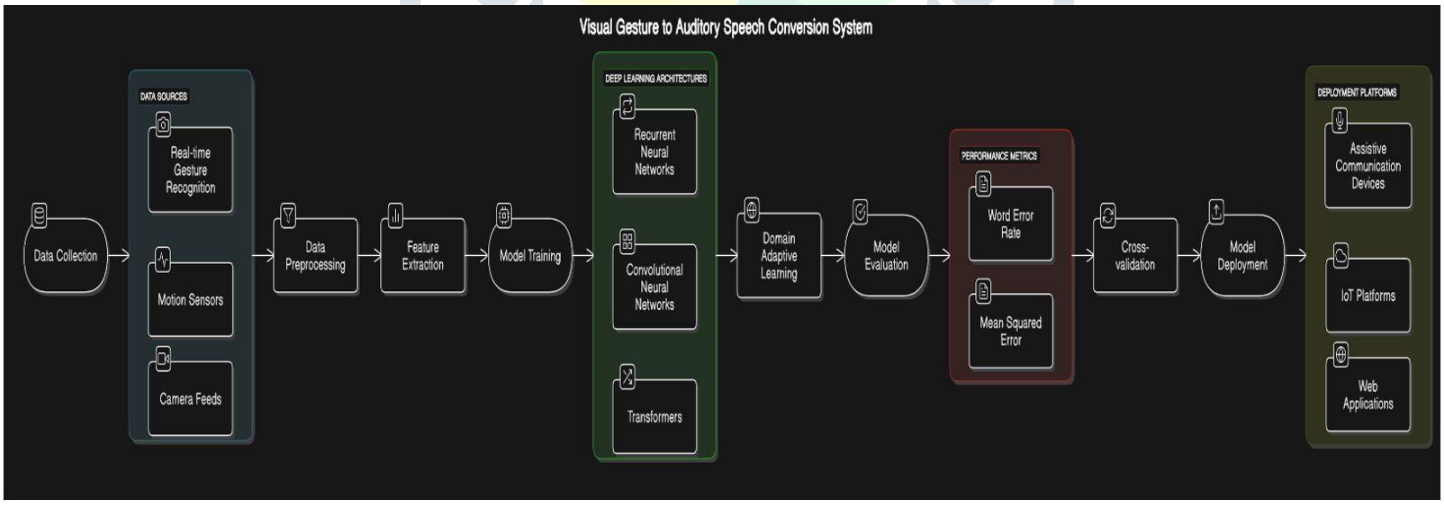
8. Post-Processing and Contextual Analysis:

- Once the system has identified a gesture, it does not just immediately speak it out. It takes a moment to double-check and polish the result. This step called post-processing, helps ensure the output makes sense in the bigger picture.
- For example, it might correct any small error like a misspelling. If the user is making a series of gestures that form a sentence, the system will look at the whole sequence and make the adjustments to ensure the words flow together naturally just like a real conversation. It is like a final editing pass to make sure the system is response is as accurate and smooth as possible.

9. Output Generation:

- The system then provides the results in two ways. First, what the gesture means – whether it is single sign or a whole sentence appears on the screen. So the user can see it.
- Second, the system speaks the words out louder. It uses technology that makes the speech sound natural and clear, not robotic. This dual output, both visual and spoken, allows for a truly interactive experience. It makes communication easy and accessible for everyone, regardless of whether they can hear or not.

IV. SYSTEM ARCHITECTURE



The architecture of the visual gesture to auditory speech conversion system encloses several key components and stages, each playing a vital role in the conversion procedure. At its core, the architecture involves a pipeline of data processing, feature extraction, model training, evaluation and deployment. The architecture starts the procedure with data collection from various sources, including real-time hand gesture recognition, motion sensors and camera feeds. The datasets are preprocessed to handle the inconsistencies, outliers, missing values and to extract useful features that obtain the underlying patterns and relationships in the data.

Next, the preprocessed data is used to train the machine learning models, particularly deep learning architectures. These are well-suited for capturing the complex patterns in high-dimesional data. The architecture utilizes the use of Recurrent Neural Networks(RNNs), Convolutional Neural Networks(CNNs) or transformers tailored to sequence prediction tasks. We can also incorporate domain adaptive learning techniques to help the model adapt better to different users and environments, improving its overall flexibility and performance.

After training, the models are tested using metrics like Word Error Rate(WER) and Mean Squared Error(MSE) to check their accuracy and reliability. The evaluation process often includes methods like cross-validation to ensure the models can handle new, unseen data effectively. Once validated, the models are deployed in practical applications, such as assistive communication devices, IoT platforms or web tools. This allows users to access clear speech outputs based on recognized gesture patterns, enhancing their communication experience.

V. PERFORMANCE EVALUATION

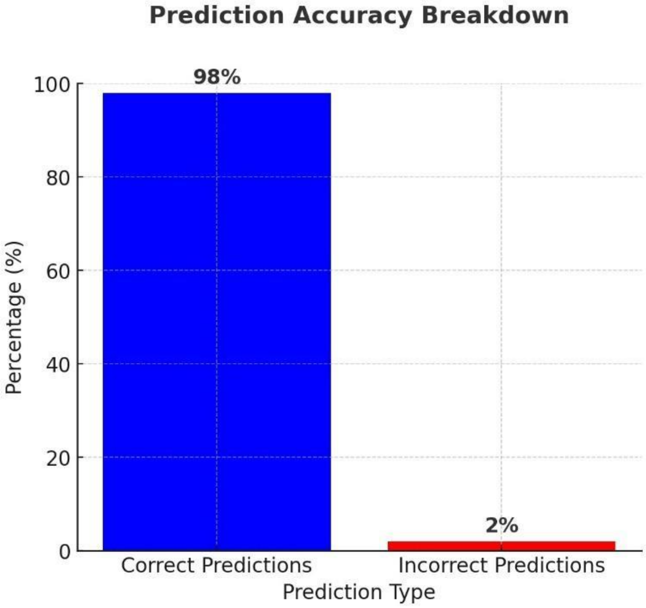
- **F1-Score:** The F1-Score is a crucial metric used to evaluate the performance of a classification model, particularly in gesture recognition. It is the harmonic mean of precision and recall, ensuring a balance between the two. Precision measures how many of the predicted gestures were actually correct, while recall evaluates how many of the actual gestures were successfully identified. This metric is especially useful when dealing with imbalanced datasets, as it considers both false positives and false negatives. A high F1-Score, such as 97.5% in your project, indicates that the model is both accurate and reliable in recognizing gestures.

$$F1=2 * (\text{Precision} + \text{Recall} / \text{Precision} * \text{Recall})$$

- **Mean Opinion Score(MOS):** The MOS is a subjective quality measure used to assess the naturalness and intelligibility of synthesized speech. It is typically rated on a scale from 1 to 5, where 1 represents unintelligible speech, and 5 indicates speech that is entirely natural. A MOS score of 4.7/5 in your project suggests that the speech synthesis module produces output that is very close to natural human speech, ensuring a high-quality user experience. This score is essential in evaluating how human-like and clear the generated speech sounds, making it a key metric for improving interaction in gesture-to-speech systems.

VI. EXPERIMENTAL SETUP AND RESULTS

The system was trained and evaluated on a custom-built dataset. Gesture-to-speech conversion was tested on this proprietary sign language dataset, achieving 98% accuracy. Hand gestures recognition, powered by MediaPipe based models, attained an F1-Score of 97.5% ensuring high precision and recall. The speech synthesis module exhibited exceptional intelligibility and naturalness, receiving a Mean Opinion Score(MOS) of 4.7/5. Additionally, real-time processing performance was optimized for seamless interaction, maintaining low latency and high responsiveness.



VII. CONCLUSION

The “Visual Gesture to Auditory Speech Converter” bridges a communication for the deaf-mute individuals, enabling them to interact seamlessly with others who may not understand sign language. The system uses advanced technologies like gesture recognition, machine learning and text-to-speech synthesis to convert hand gestures into text and speech in real-time. This approach focuses on accessibility, affordability and portability, making it a practical tool for various fields, including healthcare, education and social interactions. The project holds great potential for future growth and improvement. Upcoming advancements could include support for multiple regional sign languages, catering to a global audience. Improving gesture recognition accuracy with advanced deep learning techniques and expanding the system’s vocabulary to include full sentences or phrases would further enhance its usability and effectiveness.

Imagine an app that can instantly understand hand gestures. It uses a smart, pre-trained system, like a digital brain, that is ready to go as soon as you open it. This brain was taught to recognize lots of different hand signals by looking at countless examples. To make it work in real-time, the application uses your camera and a tool called OpenCV to see and process the video. Then, it uses MediaPipe to draw key points on your hands, showing you exactly what it is seeing and making it easier to interact.

For security, the application has a login system powered by SQLite. This keeps your information safe, allowing you to create an account and login without worrying. Passwords are scrambled using secure methods, adding an extra layer of protection. With more development, this application could truly change lives, helping people communicate more easily and opening up new possibilities for everyone.

VIII. REFERENCES

[1] Real-time Conversion of Sign Language to Text and Speech BMS College of Engineering, Bangalore, India kohsheen.t@gmail.com jayshreemaloo03@gmail.com ash.cancer98@gmail.com

[2] ASL Reverse Dictionary - ASL Translation Using Deep Learning Ann Nelson Southern Methodist University, alnelson@mail.smu.edu KJ Price Southern Methodist University, kjprice@mail.smu.edu Rosalie Multari Sandia National Laboratory, ramulta@sandia.gov.

[3] Dumitrescu, & Boiangiu, Costin-Anton. (2019). A Study of Image Upsampling and Downsampling Filters. Computers. 8. 30. 10.3390/computers8020030.

[4] Saeed, Khalid & Tabedzki, Marek & Rybnik, Mariusz & Adamski, Marcin. (2010). K3M: A universal algorithm for image skeletonization and a review of thinning techniques. Applied Mathematics and Computer Science. 20. 10.2478/v10006-010-0024-4. 317-335.

[5] Mohan, Vijayarani. (2013). Performance Analysis of Canny and Sobel Edge Detection Algorithms in Image Mining. International Journal of Innovative Research in Computer and Communication Engineering. 1760-1767. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] Tzotsos, Angelos & Argialas, Demetre. (2008). Support Vector Machine Classification for Object-Based Image 10.1007/978-3-540-77058-9_36. Analysis.

[7] Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235.

[8] Evgeniou, Massimiliano. Theodoros (2001). & Pontil, Support Vector Machines: Theory and Applications. 2049. 249-257. 10.1007/3-540-44673-7_12.

[9] Banjoko, Alabi & Yahya, Waheed Babatunde & Garba, Mohammed Kabir & Olaniran, Oyebayo & Dauda, Kazeem & Olorede, Kabir. (2016). SVM Paper in Tibiscus Journal 2016.

[10] Pradhan, Ashis. (2012). Support vector machine-A survey. IJETAE.

[11] Apostolidis-Afentoulis, Vasileios. (2015). SVM Classification with Linear and RBF kernels. 10.13140/RG.2.1.3351.4083.

[12] Kumar, Pradeep & Gauba, Himaanshu & Roy, Partha & Dogra, Debi. (2017). A Multimodal Framework for Sensor based Sign Language Recognition. Neurocomputing. 10.1016/j.neucom.2016.08.132.

[13] Trigueiros, Paulo & Ribeiro, Fernando & Reis, Luí´s. (2014). Vision Based Portuguese Sign Language Recognition System. Advances in Intelligent Systems and Computing. 275. 10.1007/978-3-319-05951-8_57.

[14] Singh, Sanjay & Pai, Suraj & Mehta, Nayan & Varambally, Deepthi & Kohli, Pritika & Padmashri, T. (2019). Computer Vision Based Sign Language Recognition System.

[15] M. Khan, S. Chakraborty, R. Astya and S. Khepra, "Face Detection and Recognition Using OpenCV," 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2019, pp. 116-119 Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020).