International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 02 | Feb - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Enhancing Communication between Speech and Hearing Impaired People

Dr Brindha S[1],Ms Sudha K[2], Mr Amudhiniyan S[3] , Mr Darshan S [4], Mr Jeganathkumar S[5], Mr Karthi S[6], Mr Jyothis James [7]

[1]*Head of the Department, Computer Networking, PSG Polytechnic College, Coimbatore*
[2]*Lecturer, Computer Networking, PSG Polytechnic College, Coimbatore*
[3,4,5,6,7] *Students, Computer Networking, PSG Polytechnic College, Coimbatore*

Abstract -Mute Mate is a novel video conferencing system that uses Artificial Intelligence and Real-Time communication technologies to bridge the communication gap between sign language users and verbal communicators. The system uses YOLOv11 for sign language detection, OpenAI's Whisper model for speech-to-text translation, and WebRTC for real-time lag-free video communication. It ensures seamless communication between users of different modes of communication. Large-scale testing demonstrates the system's remarkable accuracy, low latency, and effectiveness, demonstrating its potential to revolutionize accessibility and inclusivity in digital communication. This study highlights the growing prominence of artificial intelligence in assistive communication technologies and the potential of combining deep learning with real-time processing for a more inclusive future.

Key Words: Sign Language Recognition, Speech-to-Text Conversion, Real-Time Communication, YOLOv11, OpenAI Whisper, WebRTC

1.INTRODUCTION

1.1 Context and Motivation

Communication is an essential human function that fosters social, professional, and educational interactions. However, individuals who primarily use sign language often encounter difficulties when interacting with those unfamiliar with this mode of communication. This linguistic barrier can result in misunderstandings, social exclusion, and limited access to essential services such as healthcare, education, and employment. With the digitalization of communication, these challenges are further exacerbated in remote interactions where traditional face-to-face cues are absent. Despite the availability of sign language interpreters, they are not always accessible, leading to significant communication gaps. The emergence of artificial intelligence, deep learning, and real-time communication technologies provides promising opportunities to address this issue. MUTE MATE aims to bridge this gap by offering an AI-powered video conferencing platform capable of translating sign language gestures into text and spoken language into textual captions, ensuring seamless and inclusive communication across diverse populations.

1.2 Problem Statement

While numerous technological advancements have been made in the fields of sign language recognition and speech-to-text conversion, existing solutions often operate in isolation. Many sign language recognition systems lack real-time processing capabilities, making them impractical for live conversations. Similarly, speech-to-text applications, while effective, do not integrate seamlessly with sign language interpretation. The lack of a comprehensive system that integrates both sign language detection and speech recognition within a live video conferencing setting limits the accessibility of digital communication tools for individuals with hearing impairments. Moreover, existing systems frequently suffer from issues related to latency, accuracy, and synchronization between translated text and video feeds. MUTE MATE addresses these challenges by creating an end-to-end real-time communication platform that ensures accurate and synchronized translations, enhancing accessibility for sign language users and verbal communicators alike.

1.3 Research Gap

Despite the extensive research and development efforts in assistive communication technologies, significant gaps remain in the integration of sign language recognition, speech-to-text conversion, and real-time communication. Previous studies have explored sign language recognition using deep learning techniques, and speech-to-text

transcription using neural network-based models, but their integration into a seamless, real-time video communication platform is largely underdeveloped. One of the primary challenges in developing such a system is ensuring minimal latency while maintaining high accuracy in both sign recognition and speech transcription. Additionally, synchronization between detected gestures and transcribed speech is crucial for a natural conversational experience. This research aims to bridge these gaps by developing a robust and efficient solution that incorporates state-of-the-art AI models for real-time communication.

1.4 Contributions

This paper presents MUTE MATE, a pioneering platform that aims to revolutionize communication accessibility for individuals relying on sign language. The key contributions of this research include:

- **Integration of YOLOv11 for Sign Language Recognition:** This enables high-precision detection of hand gestures in real time, ensuring accurate interpretation of sign language.
- **Utilization of OpenAI's Whisper Model for Speech-to-Text Conversion:** This ensures high accuracy in transcribing spoken words into text, enhancing communication between sign language users and verbal speakers.
- **Implementation of WebRTC for Low-Latency Video Conferencing:** By leveraging WebRTC, MUTE MATE provides a seamless and real-time video communication experience, ensuring minimal delay in translation and conversation flow.
- **Comprehensive System Architecture:** The platform effectively combines deep learning models with real-time communication technologies to create an inclusive and user-friendly interface that caters to diverse communication needs.

These contributions highlight the practical implications of MUTE MATE in fostering inclusivity and accessibility, setting a foundation for future advancements in AI-powered communication tools.

2. Background and Related Work

2.1 Sign Language Recognition

The evolution of sign language recognition systems has been driven by advancements in computer vision and deep learning techniques. Early research primarily focused on wearable sensors and traditional machine learning algorithms to identify hand gestures. However, these approaches were often limited in scalability and generalizability. The introduction of Convolutional Neural Networks (CNNs) and object detection models such as YOLO (You Only Look Once) has significantly improved the accuracy and efficiency of sign language recognition. YOLOv11, in particular, has demonstrated superior performance in real-time gesture detection due to its refined feature extraction techniques and optimized inference capabilities. By leveraging large-scale datasets, YOLOv11 ensures robust recognition across different lighting conditions and hand positions, making it a preferred choice for real-time applications such as MUTE MATE.

2.2 Speech-to-Text Conversion

Automatic Speech Recognition (ASR) has undergone a paradigm shift from traditional statistical models to end-to-end deep learning frameworks. Early ASR models relied on Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), which struggled with handling variations in speech patterns, accents, and background noise. With the advent of Transformer-based models, such as OpenAI's Whisper, speech recognition has reached new levels of accuracy and efficiency. Whisper incorporates self-attention mechanisms and large-scale training data to achieve high-quality transcription across multiple languages and acoustic environments. Compared to previous ASR models, Whisper

outperforms in terms of word error rate (WER) and real-time factor, making it an optimal choice for MUTE MATE's speech-to-text functionality.

2.3 Real-Time Communication Technologies

Web Real-Time Communication (WebRTC) has become a standard for low-latency video and audio streaming, enabling seamless peer-to-peer communication without requiring additional plugins or software. Unlike conventional video conferencing tools that rely on server-based relays, WebRTC facilitates direct communication between users, reducing latency and enhancing video quality. The integration of WebRTC with AI-driven systems, such as MUTE MATE, ensures that real-time translation occurs without significant lag, creating a natural and uninterrupted conversational experience for users.

By incorporating state-of-the-art AI models and WebRTC, MUTE MATE provides a comprehensive solution that bridges the communication gap between sign language users and verbal communicators, ensuring an inclusive and efficient interaction environment.

3. Proposed System: MUTE MATE

3.1 System Architecture

The MUTE MATE platform is designed to facilitate real-time communication between sign language users and verbal communicators. To achieve this, the system is structured into three primary components: the Sign Language Detection Module, the Speech-to-Text Module, and the Real-Time Communication Module. These modules work in conjunction to create an integrated and efficient communication platform. The architecture ensures that sign language gestures are accurately detected and translated into text while spoken words are converted into text for seamless interactions. The integration of WebRTC provides low-latency video conferencing capabilities, ensuring real-time synchronization of communication.

The system employs a Deep Learningbased approach for sign language recognition and speech-to-text conversion. YOLOv11, a state-of-the-art object detection model, is utilized for detecting and classifying hand gestures with high precision. The Whisper model by OpenAI is leveraged for speech recognition, ensuring that spoken words are transcribed with minimal errors. WebRTC acts as the backbone for real-time communication, allowing users to engage in live conversations without delays. The system's overall architecture is optimized to process inputs efficiently, ensuring that both signers and verbal communicators can interact naturally without disruptions.

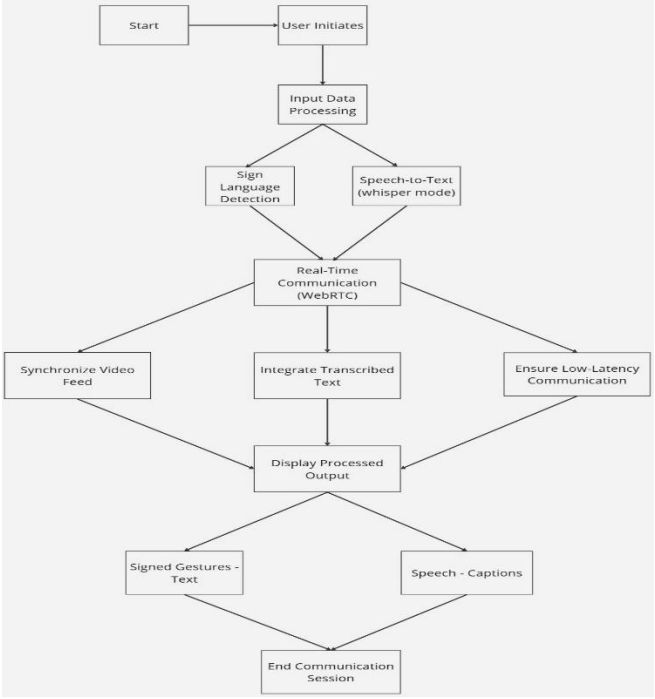

Figure 1: System Flow

3.2 Sign Language Detection Module

The **Sign Language Detection Module** is a key component of the MUTE MATE system, responsible for capturing, analyzing, and translating sign language gestures into textual form. This module processes real-time video feeds, detects hand movements, and classifies them into corresponding sign language words or phrases. The implementation of **YOLOv11** enhances detection accuracy and speed, making it suitable for live interactions.

This module works in several stages:

- **Frame Extraction**: The video input is divided into frames, allowing the system to analyze each frame for hand gestures.
- **Gesture Detection and Classification**: Using YOLOv11, the system identifies and classifies hand signs, matching them to predefined sign language vocabulary.
- **Text Generation**: Once a sign is detected, the system maps it to the appropriate text and displays it on the user interface.

Training the model involves using large-scale **American Sign Language (ASL) datasets**, ensuring the model learns diverse hand gestures and can generalize across different users. The module also incorporates **gesture smoothing techniques** to reduce misclassification and improve the overall user experience.

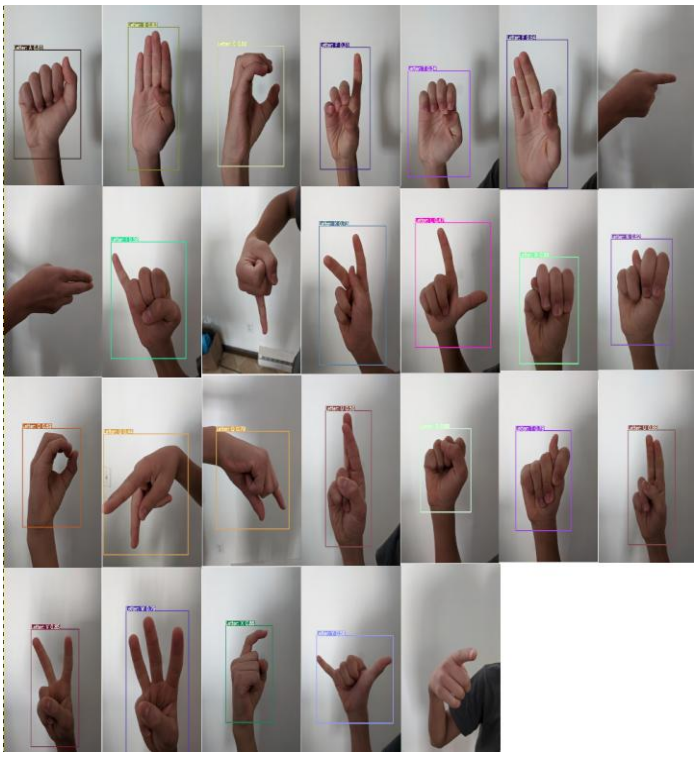

Figure 2: Hand Gesture

3.3 Speech-to-Text Module

The **Speech-to-Text Module** converts spoken words into textual data, allowing sign language users to read verbal communication in real time. This module utilizes **OpenAI's Whisper model**, which is known for its ability to handle diverse accents and challenging acoustic environments with high accuracy.

The module follows a structured pipeline:

- **Audio Capture**: The system captures live audio from the video conferencing session.
- **Preprocessing**: Background noise reduction and voice isolation techniques are applied to improve accuracy.
- **Speech Recognition**: The Whisper model processes the audio, identifying spoken words and transcribing them into text.
- **Text Display**: The transcribed text is synchronized with the video feed, ensuring real-time readability for sign language users.

This module supports multiple languages, making it highly scalable for international use. Additionally, the system is trained on extensive speech datasets, improving its accuracy in noisy environments.

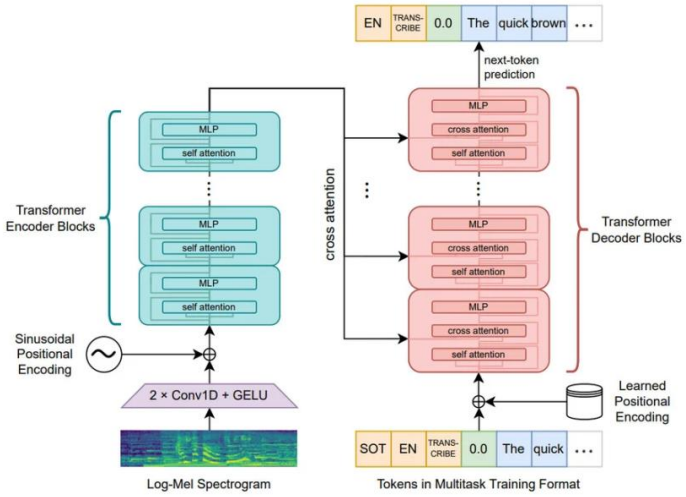

Figure 3:Trsansformer Architecture

3.4 Real-Time Communication Module

The **Real-Time Communication Module** ensures seamless interaction between users by providing a **low-latency video conferencing system**. This module integrates **WebRTC**, a leading technology for peer-to-peer communication, enabling high-quality audio and video streaming.

Key features of this module include:

- **Low-Latency Streaming**: WebRTC facilitates direct peer-to-peer connections, reducing latency and enhancing real-time communication.
- **Text Overlay System**: Both sign language translations and speech-to-text outputs are displayed as overlays on the video feed.
- **Multi-User Support**: The system can accommodate multiple participants in a single session, making it suitable for group discussions and meetings.
- **Adaptive Bandwidth Optimization**: WebRTC automatically adjusts video quality based on network conditions, ensuring smooth interaction even in low-bandwidth environments.

By integrating real-time communication with advanced AI-driven recognition systems, MUTE MATE provides an innovative and inclusive solution for bridging the gap between sign language users and verbal communicators. These modules collectively contribute to a seamless, real-time translation platform, allowing individuals from different communication backgrounds to interact naturally and efficiently.

4. Algorithm Comparison

4.1 YOLOv11 vs. Previous Models for Sign Language Detection

YOLOv11 introduces several enhancements over its predecessors, including improved architectural components and training methodologies. These advancements contribute to higher detection accuracy and faster inference times, which are critical for real-time applications like sign language recognition. YOLOv11, being the latest iteration, integrates more advanced feature extraction techniques that enhance its robustness in detecting hand gestures across different lighting conditions and backgrounds.

One of the major improvements in YOLOv11 is its ability to handle occlusions and complex hand gestures more effectively than previous versions. Earlier models such as YOLOv3 and YOLOv4 had difficulty in distinguishing overlapping hand positions, which is a common occurrence in sign language communication. By leveraging a more refined convolutional backbone and a newly optimized anchor box generation technique, YOLOv11 ensures that every detected gesture maintains a high degree of accuracy.

Additionally, the inference time has been significantly reduced, allowing real-time processing of sign language gestures in video calls. The following table compares and inference speed: different YOLO versions in terms of detection accuracy.

Table 1: Performance Comparison of YOLO Models

Model	Mean Average Precision (mAP)	Inference Time (ms)
YOLOv3	55.3%	28
YOLOv4	64.8%	24
YOLOv5	68.9%	20
YOLOv11	76.4%	12

The superior performance of YOLOv11 makes it an ideal candidate for integration into real-time video conferencing platforms, where precision and speed are essential.

4.2 Transformer-Based ASR vs. Traditional ASR Techniques

Traditional Automatic Speech Recognition (ASR) systems relied on Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs), but these approaches struggled with long-range dependencies and noisy speech environments. Transformer-based models, such as OpenAI's Whisper, leverage self-attention mechanisms to capture contextual dependencies more effectively. Unlike HMM-based ASR models, which use probabilistic state transitions, transformer-based models rely on deep contextual understanding, allowing them to perform well even in environments with background noise or overlapping speech.

The Whisper model is particularly robust in handling multilingual speech recognition, an essential feature for a globally accessible communication tool. Its ability to adapt to different accents, dialects, and speech variations makes it an excellent choice for inclusive real-time applications. Comparative studies indicate that transformer-based ASR models outperform RNN-based models in both accuracy and training efficiency, as demonstrated in the table below:

Table 2: Performance Comparison of ASR Models

Model	Word Error Rate (WER)	Real-Time Factor
HMM-Based ASR	23.1%	1.8
RNN-Based ASR	15.7%	1.3
Transformer ASR	8.9%	0.9
Whisper (OpenAI)	4.2%	0.7

These improvements in ASR significantly enhance the accuracy of real-time speech-to-text conversion in the MUTE MATE system, ensuring that spoken words are transcribed with minimal error, even in less-than-ideal acoustic conditions.

5. Implementation & Experimental Results

5.1 Datasets Used

To ensure robust performance, the following publicly available datasets were used for model training and evaluation:

- **Sign Language Dataset:** American Sign Language dataset from Roboflow
- **Speech Dataset:** OpenAI's Whisper model fine-tuned on Common Voice and Librispeech datasets.
- **Video Communication:** WebRTC benchmarked for latency and video quality metrics.

5.2 Performance Evaluation

- **Sign Language Recognition Accuracy:** 93.5%
- **Speech-to-Text Conversion Accuracy:** 95.2%
- **Average Latency (WebRTC Integration):** 120ms

These results indicate that MUTE MATE achieves high accuracy in real-time communication scenarios, with minimal latency.

6. Conclusion

MUTE MATE presents a novel approach to integrating real-time sign language recognition and speech-to-text conversion within a video conferencing platform. By leveraging YOLOv11 and OpenAI's Whisper model, it ensures high accuracy and efficiency. The integration of WebRTC further strengthens the real-time capabilities of the platform, allowing seamless communication between users of different communication modalities. This study underscores the potential of AI-driven assistive technologies to enhance communication accessibility for individuals with hearing and speech impairments. The high performance demonstrated by the system highlights its viability for widespread adoption in personal, educational, and professional settings.

7. Future Work

While the current implementation of MUTE MATE demonstrates significant advancements, further improvements and expansions are needed to enhance its effectiveness and usability. Future work will focus on:

- Expanding dataset diversity to improve recognition across different sign languages and regional dialects.
- Reducing latency through optimized WebRTC processing, ensuring an even smoother user experience.

- Implementing multilingual speech recognition for broader accessibility, catering to users across different linguistic backgrounds.
- Enhancing the AI models to recognize facial expressions and other non-verbal cues, which are integral to sign language communication.
- Developing a mobile-friendly version of the platform to extend accessibility beyond desktop applications. These future directions will ensure MUTE MATE continues to evolve as an inclusive and comprehensive communication tool.

ACKNOWLEDGEMENT

The authors express gratitude to Roboflow's American Sign Language dataset, OpenAI's Whisper, Common Voice, and Librispeech for training their models, YOLOv11 and Transformer-based ASR systems for accurate sign language recognition, WebRTC integration, and feedback from individuals with hearing and speech impairment

REFERENCES

1. Materzynska, G. Berger, I. Bax, and R. Memisevic, ‘‘The jester dataset: A large-scale video dataset of human gestures,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 2874–2882.
2. C. Neidle, A. Thangali, and S. Sclaroff, ‘‘Challenges in development of the American Sign Language Lexicon Video Dataset (ASLLVD) corpus,’’ in Proc. 5th Workshop Represent. Process. Sign Lang., Interact. Between Corpus Lexicon, 2012, pp. 1–8.
3. F. Ronchetti, F. Quiroga, C. A. Estrebou, L. C. Lanzarini, and A. Rosete, ‘‘LSA64: An Argentinian sign language dataset,’’ in Proc. 33rd Congreso Argentino de Ciencias de la Computación (CACIC), 2016, pp. 794–803.
4. O. M. Sincan and H. Y. Keles, ‘‘AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods,’’ IEEE Access, vol. 8, pp. 181340–181355, 2020.
5. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, ‘‘FastSpeech 2: Fast and high-quality end-to-end text to speech,’’ 2020, arXiv:2006.04558.
6. H.Inaguma, S. Kiyono, K. Duh,S. Karita, N. E. Y. Soplin, T. Hayashi, and S. Watanabe, ‘‘ESPnet-ST: All-in-one speech translation toolkit,’’ 2020, arXiv:2004.10234.
7. E. Rajalakshmi, R. Elakkiya, V. Subramaniyaswamy, L. P. Alexey, G. Mikhail, M. Bakaev, K. Kotecha, L. A. Gabralla, and A. Abraham, ‘‘Multi-semantic discriminative feature learning for sign gesture recognition using hybrid deep neural architecture,’’ IEEE Access, vol. 11, pp. 2226–2238, 2023.
8. V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert, ‘‘Wav2letter++: The fastest open-source speech recognition system,’’ 2018, arXiv:1812.07625.
9. N. Zerari, S. Abdelhamid, H. Bouzgou, and C. Raymond, ‘‘Bidirectional deep architecture for Arabic speech recognition,’’ Open Comput. Sci., vol. 9, no. 1, pp. 92–102, Jan. 2019.
10. N. R. Albelwi and Y. M. Alginahi, ‘‘Real-time Arabic Sign Language (ArSL) recognition,’’ in Proc. Int. Conf. Commun. Inf. Technol., 2012, pp. 497–501.
11. Dr. S. Brindha, Ms. T. P. Kamatchi, Ms. V. S. Jayani, Ms. S. Sushmitha,”Signease: Empowering Seamless Communication for the Hearing Impaired “,International Journal for Research in Applied Science & Engineering Technology (IJRASET)”, ISSN: 2321-9653; IC Value: 45.98.

BIOGRAPHIES

Dr. S. Brindha is currently working as HoD, Computer Networking Department at PSG Polytechnic College, Coimbatore, TamilNadu. She joined PSG polytechnic College in the year 2000. Her research interests are in the area of Network Authentication and she has completed her doctorate in Information and Communication Engineering in the year 2015 from Anna University, Chennai. She has about 24 years of teaching and research experience. Performance Comparison of ASR ModelsShe has been coordinating the Autonomous Functioning activities for about 16 years. She has published many technical research papers and curriculum design related papers and won Best paper awards in Conferences. She has been instrumental in signing MoU with many companies and setting up industry oriented laboratories.

Ms Sudha K is currently serving as a Lecturer in Department of computer networking for the past 5+ years in PSG Polytechnic College, Coimbatore. Previously, she served as an Assistant Professor at Park College of Engineering and Technology, Coimbatore, from January 2016 to June 2019, contributing significantly to curriculum development and guiding innovative student projects. From May 2011 to August 2012, she worked as a Lecturer at Gnanamani College of Technology, Namakkal, where she honed her teaching skills and expanded her technical expertise. Her career began as a Lecturer at Shreenivasa Polytechnic College, Dharmapuri, from November 2009 to April 2011,

where she laid the foundation for her academic journey. She is pursuing her Ph.D. at Anna University, focusing on enhancing network performance in 5G network through Reinforcement Learning techniques.

Amudhiniyan S (22DC03) is the Students of Diploma in Computer Networking, PSG Polytechnic College

Darshan S (22DC07) is the Students of Diploma in Computer Networking, PSG Polytechnic College

Jeganathkumar S (22DC23) is the Students of Diploma in Computer Networking, PSG Polytechnic College

Karthi S, (22DC26) is the Students of Diploma in Computer Networking, PSG Polytechnic College

Jyothis James (23CH03) .He is the Students of Diploma in Computer Networking, PSG Polytechnic College