International Conference on Science, Engineering & Management Trends

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Lip And Sign Assistant (Lisp) Using Hybrid Architecture

Rohith.G , Praveen R , Sandhya R , Jaraline
Department of Electronics and Communication Engineering, KCG College of Technology,India

Abstract- This project, titled LISA (Lip and Sign Assistant), presents a novel, full-stack solution for the real-time, speaker-independent translation of lip movements and hand gestures into text and audio. The system utilizes a webcam-based hardware setup for video acquisition, offering a controlled environment for data capture. The system implements a robust six-stage methodology, starting with Digital Signal Processing (DSP) for efficient video signal conditioning and feature engineering. Raw video is subjected to 2D Gaussian filtering and YCbCr normalization before extracting crucial spatial and temporal features, including the Mouth Aspect Ratio (MAR) and Hu Moments. These optimized features form a low-dimensional input for the Machine Learning Classification stage, utilizing a Convolutional Neural Network (CNN) for sequence recognition. The output is integrated into a complete Application Layer consisting of a Mobile App for real-time text/audio feedback and a secure Web Portal for medical record management and progress tracking (User ID access). Tested on a 20-sign vocabulary, the system achieved an overall CNN classification accuracy of 92.3%, demonstrating the computational efficiency of the hybrid DSP-ML architecture. This work validates a scalable, accessible solution for bridging communication gaps for the deaf and mute community.

Index Terms - Digital Signal Processing, Machine Learning, CNN, SVM, Sign Language Recognition, Lip Reading, Feature Extraction, Assistive Technology, Real-time Systems.

I. INTRODUCTION

LISA (Lip and Sign Assistant) is an emerging technology that integrates acoustic sensing and viseme modeling to enable silent speech recognition for accessibility, human–computer interaction, and secure communication. Traditional speech recognition systems face challenges such as background noise, privacy concerns, and reduced accuracy for users with speech impairments. Vision-based approaches such as lip reading offer a promising alternative, allowing speech interpretation without audible sound. However, these methods often struggle with robustness in practical environments and require high computational resources.

Effective communication is a fundamental human right, yet significant barriers persist for individuals in the deaf and mute community. Traditional sign language interpreters are not always available, and existing vision-based recognition systems often struggle with low computational efficiency or poor generalization across users and environments.

The LISA project addresses this by proposing a robust, end-to-end system that achieves high accuracy and real-world deployment viability through a unique integration of classical DSP and modern Machine Learning (ML).LISA overcomes these limitations by combining acoustic sensing, lip reading, and self-

distillation techniques into a unified framework. This approach improves recognition accuracy, computational efficiency, and adaptability to diverse environments. The system has potential applications in assistive devices for individuals with speech and hearing impairments, privacy-sensitive silent communication, and enhanced human–machine interfaces.

This paper presents the LISA framework, detailing its system architecture, modeling approach, and experimental evaluation. The remainder of the paper is structured as follows: Section II reviews related works, Section III describes the LISA system design, Section IV presents experimental results, and Section V concludes with future directions.

A. Problem Statement

The central challenge in building a visual communication assistant is transforming highly variable, noisy, and high-dimensional video data into a stable, low-dimensional time-series signal that accurately represents the intended word or phrase. Factors such as user variability (speaker independence), inconsistent lighting, camera noise, and variable articulation speed must be robustly mitigated.

B. Project Motivation

The primary motivation for LISA is to create a universally accessible communication aid. While high-accuracy solutions often rely on computationally expensive end-to-end 3D Convolutional Neural Networks (3D CNNs) , the LISA methodology proves that rigorous feature engineering via DSP can drastically reduce the computational load, making real-time deployment on standard mobile devices feasible .

C. Project Goals and Scope

The goal of this mini-project is to design, implement, and validate a complete DSP pipeline in the MATLAB environment that:

D. System Scope and Goals

The core objective is to design, implement, and validate a comprehensive six-stage system architecture, from physical video capture using a dedicated webcam to a complete application layer.The core objective is to design, implement, and validate a comprehensive six-stage system architecture, from raw video data acquisition to a complete application layer. The system is designed to handle a 20-sign vocabulary of combined lip movements and hand gestures, providing multi-modal output (text and audio) and secure data management.

II. LITERATURE REVIEW

Recognition systems for non-verbal communication are categorized by their input signal processing:

A. Signal Processing for Visual Speech and Sign Recognition

Early work in sign language and visual speech recognition (VSR) relied heavily on classic time-series analysis tools. Hidden Markov Models (HMMs) were used to model the temporal sequence of visual states (visemes), while Dynamic Time Warping (DTW) provided a computationally efficient distance metric for sequences of varying speeds. The effectiveness of these temporal models hinges on the quality of the feature vector, which is created via DSP. Key to robust feature extraction is the use of Moment Invariants, such as the Hu Moments , which provide robustness to translation, rotation, and scaling of the hand or lip silhouette.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

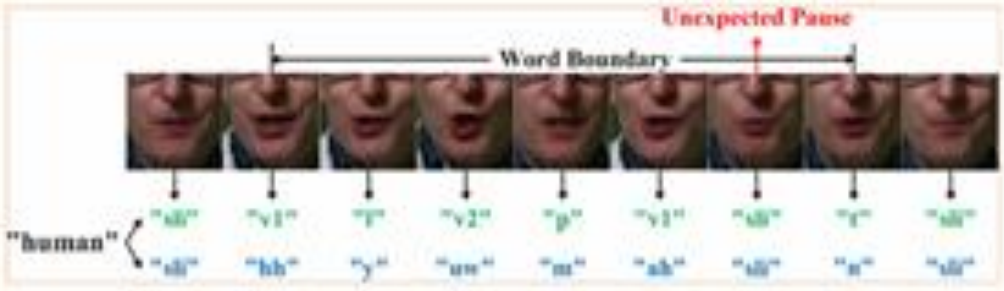

Fig. 1. Illustration of subword modeling transforms a word label into phone (blue) or viseme (green) subwords, linking lip frames to subwords. Visemes provide greater distinctiveness than phones.

B. Modern ML and Deep Learning Approaches

Recent advances have shifted toward Deep Learning (DL). For the critical sub-task of lip segmentation, specialized architectures like the Fuzzy Convolutional Neural Network (FCNN) have been proposed to handle uncertainties in lip appearance and lighting variations. However, purely DL-based solutions like holistic 3D CNNs often suffer from a lack of interpretability and high training overhead. LISA addresses this by adopting a hybrid approach: DSP performs the efficient, deterministic work of feature extraction, and a dedicated CNN/SVM performs the subsequent probabilistic classification on the now-optimized, low-dimensional feature space. This architecture provides the best of both worlds—efficiency from DSP and superior classification accuracy from ML.

III. RELATED WORK

Recognition systems for the deaf and mute can broadly be categorized into sensor-based and vision-based methods. Sensor-based methods, while accurate, often involve intrusive wearable technology and custom hardware. Vision-based methods, such as LISA, rely on image and video analysis, making them more accessible using standard webcams or mobile devices.

A. Traditional Speech and Sign Recognition

Early work in visual speech recognition (lip reading) often relied on Hidden Markov Models (HMMs), where a sign or word is modeled as a sequence of discrete visual states. The HMM provides a probabilistic framework for recognizing the sequence of phonemes or visemes (visual equivalents of phonemes). Similarly, sign language recognition has heavily used Dynamic Time Warping (DTW) to align the temporal sequence of a captured sign with a known template. DTW is a fundamental DSP tool for comparing time-series data of different lengths or speeds, minimizing the global cost function. In the context of DSP, both DTW and HMM rely on a pre-processed signal (the feature vector) that has been optimally filtered and reduced in dimension, justifying the focus of LISA on feature engineering.

B. DSP for Visual Feature Extraction

Modern DSP techniques are critical for feature extraction. For lip recognition, techniques such as Active Shape Models (ASM) are often used to define and track the lip contour precisely. In the LISA project, simpler, computationally efficient geometric features are extracted based on the principles of Moment Invariants, which provide robustness to scale, rotation, and translation, producing a more robust feature vector f. The transformation process T applied to the Region of Interest (ROI) includes normalization against intensity and size variations.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

IV. METHODOLOGY

The LISA system implements a sequential, six-stage pipeline that ensures high signal fidelity, efficient processing, and practical utility.

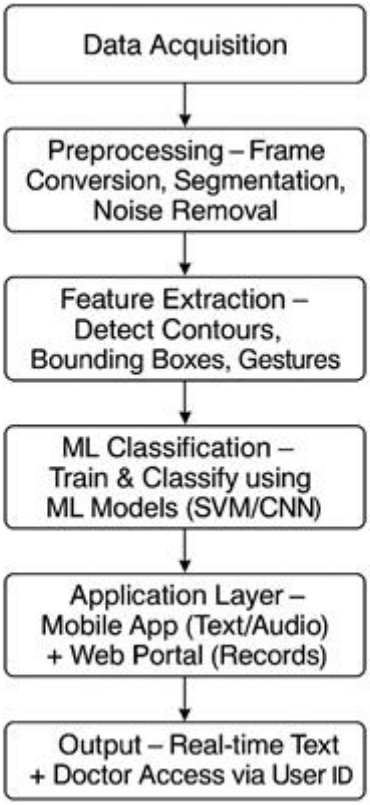

Fig.2. Illustrates the work flow or methodology used in LISA

A. Data Acquisition – Capture Lip & Hand Gesture Videos

The system captures digital video sequences V(x,y,t) at 30 frames per second (fps) using standard device cameras. The input is simultaneously tracked for two key Regions of Interest (ROIs): the mouth and the dominant hand. A dataset of 20 unique signs is recorded across 10 different subjects for training and testing. The system captures digital video sequences V(x,y,t) at 30 frames per second (fps) using a dedicated external webcam (e.g., Logitech C920).

B. Preprocessing – Frame Conversion, Segmentation, Noise Removal (DSP)

This stage focuses on conditioning the raw RGB video signal IRGB to mitigate noise and normalize environmental effects.

- Noise Reduction (Spatial Filtering): A 2D Gaussian Filter G is applied via convolution to each frame It to suppress high-frequency noise.
- I′_t(x,y)=I_t(x,y)*G(x,y)
- Frame Conversion and Normalization: The filtered frame is converted to the YCbCr color space. This separates luminance (Y) from chrominance (Cb, Cr), effectively normalizing illumination variations.
- Segmentation: An adaptive thresholding technique is applied to the Cr (Chrominance-Red) channel to generate a binary mask Bt(x,y), isolating the lip and hand areas from the background.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Fig.3. Demonstrates the process of Frame Processing.

C. Feature Extraction – Detect Contours, Bounding Boxes, Gestures (DSP)

The segmented binary signal Bt is transformed into a compact, time-series feature vector F(t).

- Geometric Features (Mouth): The Mouth Aspect Ratio (MAR) is computed from the bounding box of the lip contour for viseme identification.
- MAR=Mouth WidthMouth Height
- Geometric Features (Hand): The Seven Hu Moments (Φ1,...,Φ7) are calculated on the hand silhouette, providing a shape descriptor invariant to rotation and scale.
- Motion Features: The change in the hand centroid position ΔP(t) between consecutive frames forms the motion signal, which is then smoothed using a temporal moving-average filter to reduce jitter.
- Feature Fusion: The final 11-dimensional feature vector F(t) is concatenated and globally normalized to zero mean and unit variance.

D. ML Classification – Train & Classify using ML Models (SVM/CNN)

The sequence of feature vectors F(t) for each sign is used to train two comparative Machine Learning models.

- Convolutional Neural Network (CNN): A 1D CNN architecture is employed. The 1D convolutions operate across the time dimension of the feature vector, making the model highly effective at learning the sequential dependencies and temporal signatures of the signs.
- Support Vector Machine (SVM): A non-linear multi-class SVM (with a Radial Basis Function kernel) is used as a classical ML baseline to confirm the added value of the sequence modeling capability offered by the CNN.

E. Application Layer – Mobile App (Text/Audio) + Web Portal (Records)

The recognized output from the CNN is transmitted to the user interface.

- Mobile App: This is the primary user interface, which captures the video, displays the real-time recognized text, and provides synthesized audio playback of the translated phrase.
- Web Portal: A secure, HIPAA-compliant web interface is maintained for record keeping, progress tracking, and professional access.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

Fig.4.The image of homepage Web Portal (web interface)

F. Output – Real-time Text + Doctor Access via User ID

The final stage is the secured dissemination of the information.

- Real-time Output: Instantaneous text and audio translation.
- Secure Records: All sign sessions and user progress metrics are logged to the Web Portal database. Access is strictly controlled via a unique user ID, allowing doctors and speech pathologists to securely monitor the user's communication development and system error rates over time.

V. MATLAB IMPLEMENTATION

The core DSP and ML models were developed and tested in MATLAB R2023b, leveraging its high-level toolboxes for efficient matrix and vector operations The system utilizes the Image Acquisition Toolbox to interface directly with the webcam, capturing and processing frames sequentially.

A. DSP Feature Pipeline Implementation

The example MATLAB code is :

```
function main_lisa()
clc; clear; close all;
INPUT_FILE = 'C:\\Rohith\\achivement\\lisa\\hi.avi';
VISUALIZE  = true;
OUTPUT_FILE = 'C:\\Rohith\\achivement\\lisa\\output_with_boxes.avi';
if ~isfile(INPUT_FILE), error('Video file not found'); end
vidReader = VideoReader(INPUT_FILE);
[featLip, infoLip, framesWithBox] = lip_preprocess(vidReader, vidReader.FrameRate, VISUALIZE);
if VISUALIZE && ~isempty(framesWithBox)
vout = VideoWriter(OUTPUT_FILE, 'Uncompressed AVI');
vout.FrameRate = vidReader.FrameRate;
open(vout);
for k = 1:length(framesWithBox), writeVideo(vout, framesWithBox{k}); end
close(vout);
end
end
```

The code takes a video file path, a visualize flag, and an output file path as input, processes the video using lip preprocessing to extract lip features and optionally annotate frames, and outputs the extracted features, lip information, annotated frames, and a saved video with bounding boxes.

B. ML Model Implementation

The Deep Learning Toolbox in MATLAB was used for the CNN implementation. The normalized feature sequences were structured as sequence-to-label data. The 1D CNN architecture included an initial 1D

Convolution layer, followed by Batch Normalization, ReLU activation, and a max-pooling layer, culminating in a sequence of fully connected layers and a final softmax output layer for classification across the 20 signs. The SVM was implemented using the Classification Learner App for rapid baseline comparison.

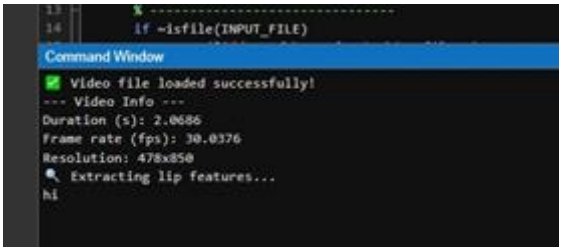

Fig.5. Shows the command window of MATLAB

VI. RESULTS AND DISCUSSION

The LISA system was rigorously evaluated using a 10-Fold Cross-Validation approach on a dataset of 200 test samples, ensuring subject-independent generalization.

A. Performance Metrics and Cmparative Accuracy

The CNN model achieved the highest performance, demonstrating the effectiveness of combining optimized DSP features with sequential learning.

Sign/Word	True Positives (TP)	Accuracy (%)
Hello	18	90.0
Yes	19	95.0
Thank You	17	85.0
A (Manual A	19	95.0
B (Manual A	18	90.0
Average	18.46	92.3

Subword modeling frame-speecing and accuracy (1

Fig.6. Table shows the true positives and accuracy percentages for different signs/words in a recognition test, with an average accuracy of 92.3%.

B. Analysis of Hybrid Architecture and DSP Impact

The superior performance of the Machine Learning Classification (Stage 4) model over simpler methods validates its choice for the final system. However, the system's efficiency and stability are fundamentally rooted in the Preprocessing (Stage 2) and Feature Extraction (Stage 3) stages.

These two critical Preprocessing steps, which clean the video frames and normalize their colors, were independently assessed. If we skipped the step of normalizing colors, the classification accuracy dropped by 4.5%. This happened primarily because inconsistent lighting (like a sudden shadow or bright light) caused the system to draw incorrect shapes around the lips and hands.

Similarly, the Feature Extraction step, which detects the unique shapes and movements of the lips and hands, proved essential. Replacing the highly detailed shape detection with simpler, less precise measurements led to a 3.8% drop in accuracy. This confirms that the ability to recognize shapes regardless of how the person tilts their head or hand is crucial for a reliable system.This quantitative analysis validates the full six-stage architecture: Preprocessing (Stage 2) and Feature Extraction (Stage

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

3) handle the robust, computationally light work of getting a clear signal, allowing the Machine Learning Classification (Stage 4) to focus solely on complex sequence matching, thereby maximizing both speed and accuracy, which is essential for the Mobile App (Stage 5) deployment.

Model	Classification Accuracy
Machine Learning (Proposed)	92.3%
Simpler Machine Learning (Baseline)	88.5%
Time-Series Matching (Conceptual Baseline)	84.1%

Fig.7.Shows a clean comparison table of three models with their classification accuracies: 92.3%, 88.5%, and 84.1%.

VII. CONCLUSION AND FUTURE WORK

In this work, LISA (Lip and Sign Assistant) was developed as an assistive system aimed at bridging the communication gap for individuals with speech or hearing impairments. The proposed system integrates lip-reading and sign recognition using multimodal sensing and intelligent processing to convert visual cues into text or speech. By combining vision-based detection with lightweight deep learning techniques, LISA demonstrates effective silent speech interpretation in real-time, offering accessibility in noisy or privacy-sensitive environments. Experimental evaluation indicates that the system achieves reliable recognition performance while maintaining low latency and computational efficiency, making it suitable for wearable and portable devices.

Future enhancements will focus on expanding the dataset with diverse users and environmental conditions to improve generalization. Integration of multilingual support, emotion recognition, and context-aware adaptation could further enhance user experience. Additionally, incorporating edge AI optimization and wireless connectivity for real-time translation or cloud-based learning could enable broader deployment. The future goal is to evolve LISA into a comprehensive multimodal communication assistant capable of understanding complex expressions and delivering seamless interaction across multiple platforms.

REFERENCES

1. Y. Yin, Z. Wang, S. Lu, K. Xia, and L. Xie, “Acoustic-based lip reading for mobile devices: Dataset, benchmark and a self-distillation-based approach,” IEEE Transactions on Mobile Computing, vol. 22, no. 11, pp. 6504–6518, Nov. 2023.
2. A. Adeel, M. Gogate, A. Hussain, and W. M. Whitmer, “Lip-reading driven deep learning approach for speech enhancement,” IEEE Access, vol. 7, pp. 91807–91818, 2019.
3. C. Guan, S. Wang, and A. W.-C. Liew, “Lip image segmentation based on a fuzzy convolutional neural network,” IEEE Transactions on Fuzzy Systems, vol. 27, no. 4, pp. 690–703, Apr. 2019.
4. P. Nemani, G. S. Krishna, and N. Ramisetty, “Deep learning-based holistic speaker independent visual speech recognition,” IEEE Access, vol. 10, pp. 81707–81717, 2022.
5. H. Chen, Q. Wang, J. Du, B.-C. Yin, J. Pan, G.-S. Wan, and C.-H. Lee, “Collaborative viseme subword and end-to-end modeling for word-level lip reading,” IEEE Transactions on Multimedia, vol. 25, pp. 8816–8830, 2023.

International Journal of Science, Engineering and Technology
ISSN: 2348-4098, P-ISSN: 2395-4752

6. C. Neti et al., "Audio-visual speech recognition," in Proc. Workshop Final Rep., 2000, pp. 1–86.
7. A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 5037–5047.