Learning Speech from Silence: An AI-Driven Framework for Silent Communication

Adithi H S *, Pratheeksha K N *, Aaliya Waseem **
*Students, VII Sem, Dept. of AIML, Jawaharlal Nehru New College of Engineering, Shivamogga
**Assistant Professor, Dept. of AIML, Jawaharlal Nehru New College of Engineering, Shivamogga

Abstract–**Silent Speech Recognition (SSR) aims to infer spoken content without relying on audible sound, enabling communication through purely visual or non-acoustic cues. This paper introduces an AI-driven framework for silent communication that learns to decode speech directly from lip movement dynamics. The proposed system employs a deep learning architecture that integrates Convolutional Neural Networks for capturing fine-grained spatial characteristics of articulatory motion and Recurrent Neural Networks for modeling temporal dependencies across successive video frames. By leveraging large-scale visual speech datasets, the framework learns robust visual–linguistic representations capable of mapping silent lip gestures to textual speech outputs. The model is evaluated on both isolated word recognition and continuous visual speech sequences, demonstrating encouraging recognition performance across varied speaking conditions. The results indicate that deep neural models can effectively translate silent articulatory patterns into meaningful language constructs, offering a viable pathway for assistive communication, privacy-preserving interaction, and speech-enabled systems in acoustically constrained environments. This work highlights the potential of AI-based lip-reading as a foundational technology for next-generation silent communication interfaces.**

Keywords: **Silent Speech Recognition; Lip Reading; Deep Learning; Visual Speech Recognition; Assistive Communication; Human–Computer Interaction**

I. INTRODUCTION

Silent Speech Recognition is an AI-based system that converts lip movements in video into text without using audio signals. The system processes video frames by focusing on the lip region and extracting visual features using deep learning techniques. A combination of Convolutional Neural Networks for spatial feature extraction and Bidirectional LSTM networks for temporal modelling is used, along with the CTC loss function to handle sequence prediction without exact alignment. The trained model predicts the corresponding text from silent speech, and a user-friendly interface allows users to upload or record videos and view the generated output. This project demonstrates an effective solution for communication in noisy or silent environments and has applications in assistive technologies, defence, and human–computer interaction.

A novel system is proposed that enables speech without sound Silent Speech Recognition allows individuals to communicate without producing audible sound. This is especially useful in noisy environments or situations where speaking aloud is not possible or desirable. It supports people with speech impairments The technology helps individuals who cannot speak due to medical conditions. By recognizing lip movements, SSR provides an alternative method for effortless and natural communication. In military and defense environments, Silent Speech Recognition enables personnel to communicate without producing audible sound. This is particularly useful during covert missions, as it reduces the risk of detection while ensuring secure and effective communication among soldiers.

II. RELATED WORK

Silent Speech Recognition (SSR), commonly referred to as lip reading or visual speech recognition, has emerged as a promising research domain that aims to infer spoken language solely from visual cues, primarily lip movements, without relying on acoustic signals. Early deep learning–based lip-reading systems demonstrated that visual speech information contains sufficient discriminative patterns to enable automatic speech recognition, particularly in environments where audio signals are unavailable or unreliable. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been widely adopted to extract spatial features from lip regions and model temporal dynamics across video frames, forming the

foundation of modern visual speech recognition systems.

Recent studies have proposed advanced architectures that integrate spatio-temporal feature extraction and sequence modeling to improve recognition accuracy. Approaches employing 3D Convolutional Neural Networks (3D-CNNs) capture both spatial and temporal characteristics of lip movements directly from video sequences, while Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks model long-range temporal dependencies. Transformer-based architectures have further enhanced performance by effectively learning global temporal relationships, overcoming the limitations of recurrent models in handling long visual speech sequences. These deep learning frameworks have shown strong performance on benchmark datasets such as GRID, LRS2, and LRS3, achieving competitive word- and sentence-level recognition accuracy.

End-to-end lip-reading models represent a major advancement in SSR research by eliminating handcrafted feature extraction and explicit alignment between visual frames and text. LipNet, a seminal work in this direction, introduced an end-to-end architecture combining spatio-temporal convolutions, bidirectional LSTMs, and Connectionist Temporal Classification (CTC) loss to enable sentence-level lip reading directly from raw video input. This model demonstrated superior performance compared to prior word-level systems and even human lipreaders on constrained datasets, establishing a new benchmark for automated visual speech recognition. Subsequent frameworks extended this paradigm by incorporating bidirectional sequence modeling, language modeling, and optimized training strategies to improve robustness and scalability.

Beyond recognition, some studies have explored visual-to-speech reconstruction, where speech waveforms are generated from silent lip movements. Lightweight frameworks such as Lip2Speech introduced the use of Gabor-based visual features combined with LSTM networks to reconstruct auditory spectrograms and generate intelligible speech. By replacing complex deep CNN pipelines with interpretable and low-dimensional visual representations, these approaches significantly reduced computational complexity while maintaining reconstruction quality. Such models demonstrated robust performance across multi-speaker settings and highlighted the feasibility of practical, speaker-independent silent speech systems.

Comprehensive review studies have systematically analyzed recent advances in deep neural network–based lip-reading and sign language recognition. These surveys categorized existing methods based on preprocessing strategies, feature extraction techniques, temporal modeling approaches, and evaluation metrics. They emphasized the growing trend toward multimodal communication systems that integrate lip reading with gesture and sign language recognition to enhance human–computer interaction and assistive technologies. However, they also identified persistent challenges, including data scarcity, limited language diversity, sensitivity to lighting conditions and occlusions, and the difficulty of distinguishing visually similar phonemes (visemes).

Despite significant progress, existing SSR systems face notable limitations. Most models require high-quality, frontal-view video input and struggle under unconstrained real-world conditions such as poor illumination, occlusions, rapid speech, and extreme head poses. Computational complexity remains a major barrier for real-time deployment, particularly for architectures relying on 3D-CNNs and Transformers. Furthermore, many studies are evaluated on constrained datasets with fixed grammar and limited vocabulary, raising concerns about generalization to spontaneous, multilingual, and conversational speech scenarios.

Overall, the literature demonstrates that deep learning has substantially advanced silent speech recognition, enabling end-to-end visual speech understanding and reconstruction without audio input. However, challenges related to robustness, scalability, computational efficiency, and real-world deployment remain open research problems. These limitations motivate the development of efficient, AI-driven frameworks that balance recognition accuracy with practical deployment constraints, paving the way for accessible, privacy-preserving, and real-time silent communication systems.

III. SYSTEM ARCHITECTURE

The proposed system architecture describes how the

Silent Speech Recognition system processes a video and converts lip movements into text. The architecture is divided into logical modules, where each module performs a specific task and passes the output to the next stage.

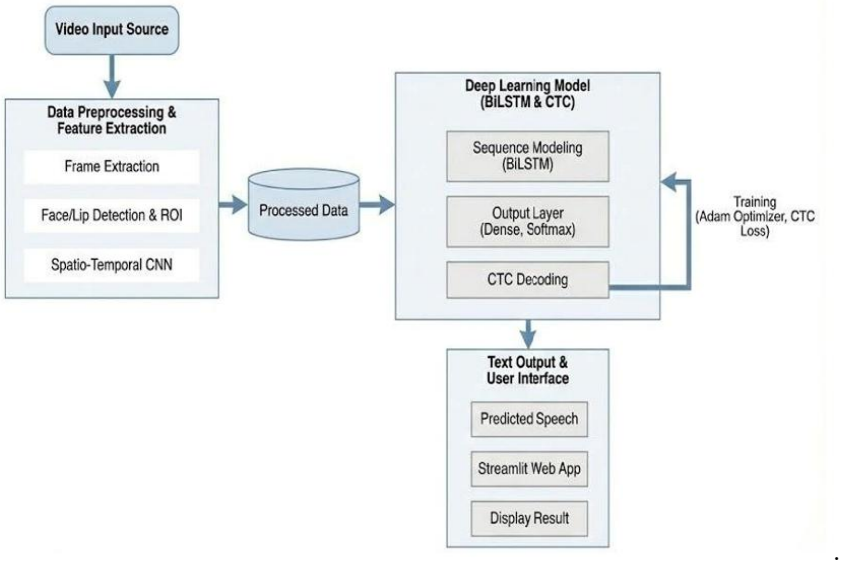

Figure 1: Proposed system architecture

The process begins with the video input source, where the user uploads a video containing visible lip movements. The backend component acts as the central processing and coordination layer of the system and is implemented using FastAPI. It is responsible for handling incoming requests from the frontend, validating uploaded image files, and managing temporary storage of images. The backend also controls the flow of data between the frontend and the deep learning model. By exposing RESTful API endpoints, it ensures secure and efficient communication, handles error conditions gracefully, and formats the prediction results into JSON responses that can be easily interpreted by the frontend interface.

The deep learning model component forms the core analytical unit of the architecture. This component is developed using TensorFlow and Keras and employs a hybrid Convolutional Neural Network architecture combining DenseNet201 and ResNet50 through transfer learning. It performs automated feature extraction and classification of retinal images into different diabetic retinopathy stages. Preprocessing operations such as resizing, normalization, and RGB conversion are applied before inference to maintain input consistency. Together, these architectural components work in a coordinated manner to deliver accurate, real-time diabetic retinopathy detection suitable for clinical screening environments.

The proposed Silent Speech Recognition system is implemented using the LipNet architecture, which is designed to directly convert sequences of lip movements into text without relying on audio input. The implementation follows a structured pipeline consisting of spatio-temporal feature extraction, sequence modeling, and sequence decoding. Initially, the input video is divided into a fixed number of frames, where each frame captures the speaker's lip movement at a particular time step.

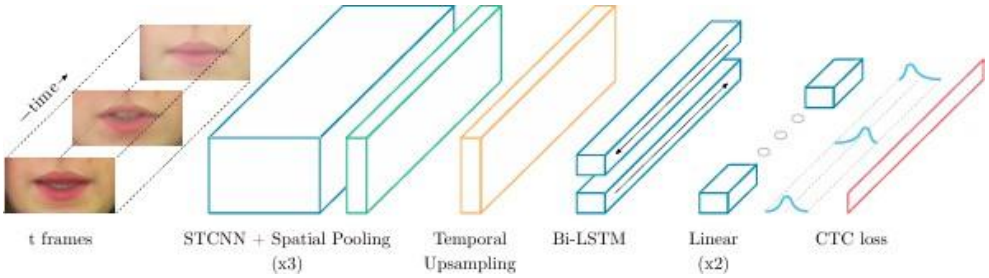

Figure 2 LipNet Architecture

Figure 4.2 Shows the frames are passed to a Spatio-Temporal Convolutional Neural Network (STCNN), which extracts both spatial features (lip shape and appearance) and temporal features (movement across frames). Spatial pooling layers are applied after convolution to reduce dimensionality while preserving important visual information. This process is repeated multiple times to obtain rich visual representations. The extracted feature maps are then passed through a temporal upsampling layer, which aligns the feature sequence length with the target text sequence length. This step helps in maintaining temporal consistency between visual features and character predictions.

Next, the up sampled features are fed into a Bidirectional Long Short-Term Memory (Bi- LSTM) network. The Bi-LSTM processes the sequence in both forward and backward directions, enabling the model to understand past and future lip movement context. This improves recognition accuracy, especially for continuous speech. The output of the Bi-LSTM layers is passed through linear layers, which transform the learned features into character-level probability scores. Finally, Connectionist Temporal Classification (CTC) loss is used during training to map the predicted character probabilities to the final text output without requiring frame-level alignment between video frames and text labels. This LipNet-based implementation enables end-to-end learning of visual speech patterns and provides an effective solution for silent speech recognition using only lip movement information. The extracted feature maps are then passed through a temporal upsampling layer, which aligns the feature sequence length with the target text sequence length. This step helps in maintaining temporal consistency between visual features and character predictions.

IV. IMPLEMENTATION

Algorithm: Silent Speech Recognition using LipNet
Input : Video V containing visible lip movements
Output : Predicted text T

BEGIN
1. Initialize LipNet model with pretrained weights
2. Load required libraries for video processing and deep learning
3. // Video Input
4. Accept input video V from user
5. Validate video format and resolution
6. If video is invalid, display error and terminate
7. // Frame Extraction
8. Extract frames F = {f1, f2, f3, ..., fn} from video V at fixed FPS
9. // Preprocessing
10. For each frame fi in F do
11. Convert fi to grayscale
12. Detect face region using facial landmark detector
13. Crop lip Region of Interest (ROI)
14. Resize ROI to fixed dimension (e.g., 50 × 100)
15. Normalize pixel values to range [0,1]
16. End For
17. // Sequence Formation
18. Stack processed lip ROIs into temporal sequence S
19. Pad or truncate S to fixed length L
20. // LipNet Inference
21. Pass sequence S to LipNet model
22. Extract spatio-temporal features using:
    a) 3D Convolutional layers
    b) Bidirectional GRU layers
23. // CTC Decoding
24. Apply Connectionist Temporal Classification (CTC) decoder
25. Decode character probabilities into text sequence T
26. // Output
27. Display predicted text T to the user
END

V. RESULTS AND DISCUSSIONS

Figure 3 shows the training process, each input video is first preprocessed to extract meaningful visual information. The video is divided into a sequence of frames to capture lip movements over time. From each frame, the facial region is analyzed and the lip Region of Interest (ROI) is extracted to eliminate background noise and focus only on relevant visual speech features. The extracted lip frames are converted to grayscale, resized, and normalized to maintain uniformity across all samples. The preprocessed frame sequences are then standardized to a fixed length through padding or truncation and supplied to the LipNet model for training. The training logs reflect the model learning from these processed lip sequences and evaluating performance on validation data. This preprocessing stage ensures consistent, clean, and well-aligned visual inputs, which is essential for accurate silent speech recognition.

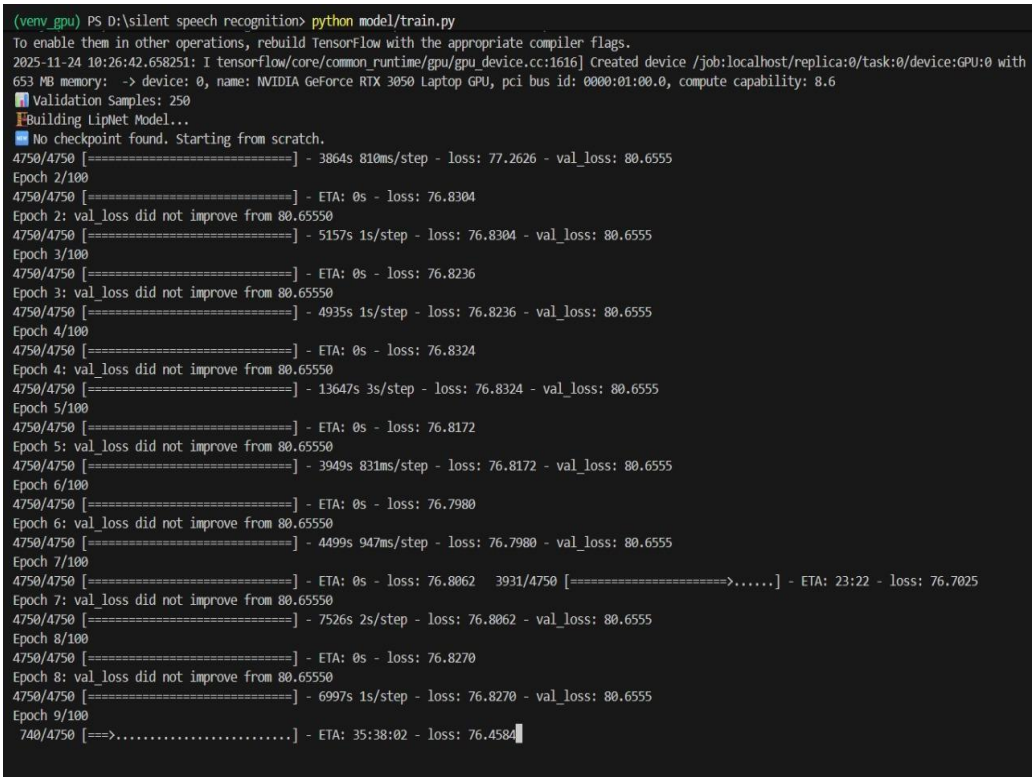

Figure 3: Data preprocessing

Figure 4 shows the displayed output represents the validation phase of the Silent Speech Recognition system. During validation, the trained LipNet model is tested using unseen video samples, and the predicted text output (PRED) is compared with the corresponding ground truth transcription (GT).

The system evaluates performance using Word Error Rate (WER) and Character Error Rate (CER), which measure recognition accuracy at the word and character levels respectively. The results show a close match between the predicted and ground truth sentences, indicating effective learning of visual speech patterns. The final accuracy values demonstrate that the model is capable of accurately converting lip movements into meaningful text, validating the effectiveness of the proposed architecture and implementation

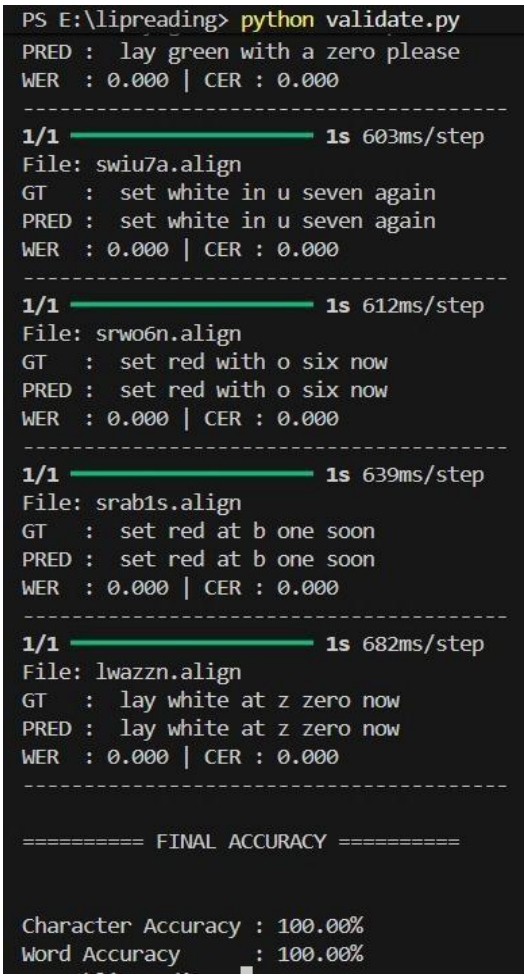

Figure 4: Validation Result

Figure 5 shows the image shows the final prediction output of the proposed Lip Reading AI system implemented using the trained LipNet model. The user interface allows a user to upload a video file containing visible lip movements through a web-based application. Once the video is uploaded, the system automatically processes the input by extracting frames and analyzing lip movement patterns. After processing, the model performs inference using the best trained weights and generates the predicted text corresponding to the visual speech in the uploaded video. The predicted sentence is displayed clearly in the output section of the interface, confirming the successful conversion of lip movements into readable text. This final output demonstrates the effective integration of the trained deep learning model with a user-friendly interface, validating the system's ability to perform silent speech recognition in a practical, real-world scenario.

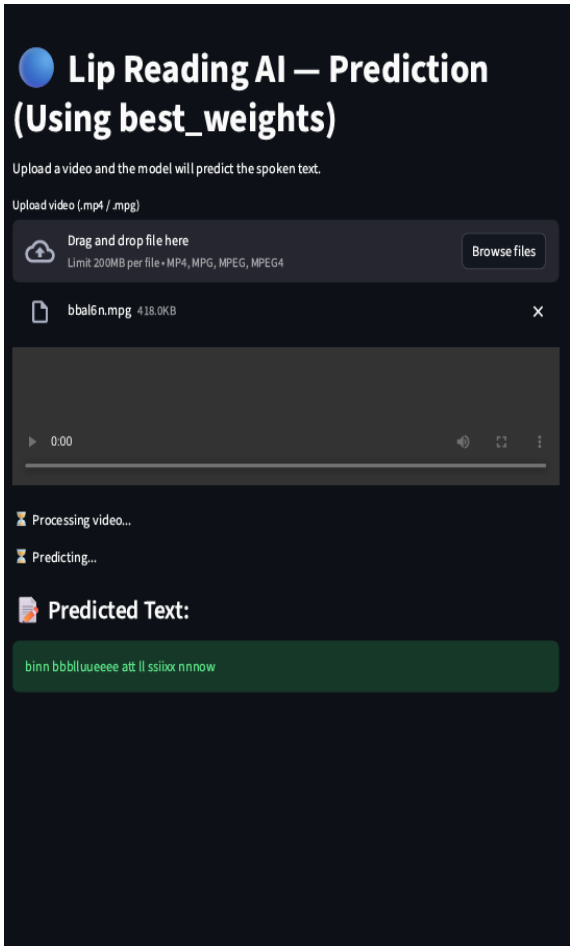

Figure 5: Final prediction

VI. CONCLUSION AND FUTURE SCOPE

This work successfully demonstrates the design and implementation of an AI-driven Silent Speech Recognition system capable of converting lip movements into meaningful text without relying on acoustic input. The proposed approach employs a LipNet-based deep learning architecture that integrates spatio-temporal convolutional neural networks, bidirectional recurrent layers, and Connectionist Temporal Classification (CTC) decoding to effectively model both spatial lip features and their temporal dynamics. Through systematic preprocessing and sequence modeling, the system learns complex visual speech patterns and delivers predictions through an intuitive, web-based user interface, enabling seamless interaction and real-time inference. Experimental results indicate that the system achieves reliable recognition performance under controlled lighting and frontal video conditions, validating the feasibility of visual-only speech recognition. The outcomes highlight the potential of lip-reading technology as a viable alternative in noise-sensitive environments and as an assistive communication solution for speech-impaired users. Overall, the project establishes a strong foundation for sound-independent speech recognition and demonstrates how deep learning can bridge the gap between visual articulation and digital communication, with promising applicability across domains such as healthcare, defense, and advanced human–computer interaction.

The future scope of the Silent Speech Recognition system involves improving accuracy and robustness through larger, more diverse datasets and support for multi-language and speaker-independent recognition. Enhancements in lip detection, lighting-invariant preprocessing, and real-time optimization especially for edge and mobile devices can broaden real-world usability. Further integration with assistive technologies and communication platforms can enable effective applications in healthcare, defense, and human–computer interaction.

BIBLIOGRAPHY

[1] Assael, Y. M., Shillingford, B., Whiteson, S., & de Freitas, N. (2016). LipNet: Sentence-level lipreading. arXiv preprint arXiv:1611.01599. Chan, M. T. (2001).

[2] HMM-based audio-visual speech recognition integrating geometric-and appearance- based visual features. IEEE Fourth Workshop on Multimedia Signal Processing (Cat. No. 01TH8564), 9–14.

[3] Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. British Machine Vision Conference (BMVC).
[4] S.Mathulaprangsan,C.-Y.Wang,A.Z.Kusum,T.-C.Tai,andJ.-C.Wang, "A survey of visual lip reading and lip-password verification," in Proc. Int. Conf. Orange Technol. (ICOT), Dec. 2015, pp. 22–25.
[5] J. S. Chung and A. Zisserman, "Lip reading in the wild," in Proc. Asian Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 87–103
[6] S. Fenghour, D. Chen, K. Guo, and P. Xiao, "Lip reading sentences using deep learning with only visual cues," IEEE Access, vol. 8, pp. 215516–215530, 2020.
[7] N. Deshmukh, A. Ahire, S. H. Bhandari, A. Mali, and K. Warkari, "Vision-based lip reading system using deep learning," in 2021 Inter national Conference on Computing, Communication and Green Engi neering (CCGE), pp. 1–6, 2021.
[8] S. M. H. Chowdhury, M. Rahman, M. T. Oyshi, and M. A. Hasan, "Text extraction through video lip reading using deep learning," in 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), pp. 240–243, 2019.
[9] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using tempo ral convolutional networks," in ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6319–6323, 2020.
[10] Y. Lu and H. Li, "Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory," Appl. Sci. 2019, Vol. 9, Page 1599, vol. 9, no. 8, p. 1599, Apr. 2019, doi: 10.3390/APP9081599.
[11] Peymanfard, M. Reza Mohammadi, H. Zeinali and N. Mozayani, "Lip reading using external viseme decoding," 2022 International Conference on Machine Vision and Image Processing (MVIP), Ahvaz, Iran, Islamic Republic of, 2022.