ISSN No:-2456-2165

SpeakVision: A Comprehensive Survey on End-to-End Sentence Level Lipreading

Ashwini M Rayannavar
Assistant Professor, Information Science and Engineering
RNS Institute of Technology
Bengaluru, India

Rakshit Chouhan
Student, Information Science and Engineering
RNS Institute of Technology
Bengaluru, India

Aman Ali Gazi
Student, Information Science and Engineering
RNS Institute of Technology
Bengaluru, India

Maitree Rajesh Patel
Student, Information Science and Engineering
RNS Institute of Technology
Bengaluru, India

Abstract:- SpeakVision is a speech reading framework capable of extracting speech from audio-video inputs using an AI-based model. A new integrated approach, using both sight and sound, is needed for situations when a voice signal is obscured, or when seeing the apparatus is much easier than hearing it. SpeakVision leverages AI technologies, such as, 3D convolutional layers for extracting spatial features, Bidirectional LSTMs for temporal information and CTC decoding for generating text. Video preprocessing techniques were applied to optimize model performance, and the results were developed into an easy-to-use Streamlit interface for interactive visualization.

I. INTRODUCTION

In today's world, combining visual and audio data is becoming important for building smarter technology. Traditional systems that rely only on visual information often struggle to understand complex situations where sound could provide valuable context. Speak Vision aims to bridge this gap by using both visual and audio data to better interpret user needs, making it useful for communication aid.

SpeakVision uses artificial intelligence, which combines different types of data. Systems designed for real-time adjustments, SpeakVision focuses on analysing pre-recorded or controlled datasets, making it ideal for research and offline use.

This paper explores the AI techniques behind Speak Vision, such as machine learning and deep learning, to process and understand visual and audio information.

II. LITRATURE SURVEY

There has been noticeable progress in combining visual and audio data to create smarter systems. These systems, which analyze multiple types of data simultaneously, are more effective in understanding and responding to real-world situations.

[1] Learning from Multiple Sources: A study by Ngiam et al. (2011) explored how combining different types of data, like video and audio, can make AI systems more accurate and reliable. They found that fusing visual and audio data allows the system to better understand its environment, especially when one type of data might not be enough. This is a crucial idea for Speak Vision, where we aim to use both visual and audio cues to improve overall understanding.

[2] Lip-Reading and Audio for Speech Recognition: Ephrata et al. (2018) worked on models that combine lip-reading with audio for better speech recognition.

[3] Extracting Meaningful Features from Visual and Audio Data: Simonyan and Zisserman (2014) showed that CNNs are very effective at extracting features from images. Similarly, Cho et al. (2014) found that RNNs are great for processing audio data.

III. OBJECTIVES

- *Develop an Automated Lip-Reading Model:*
  The goal is to build a deep learning-based system that can interpret speech by reading lip movements from video inputs. This involves training a model to identify and map facial movements to corresponding phonemes or words.

- *Enhance Speech Recognition with Visual Input:*
  We aim to create a model that combines both video and audio signals to ensure accuracy.

- *Provide Text Output for Lip-Read Speech:*

One of the main objectives is to convert lip-read speech into text. The system will analyze facial expressions and lip movements from video frames to generate an accurate transcription of the spoken words.

IV. AI MODELS AND ALGORITHMS USED IN LIP READING

Our lip reading and speech-to-text system uses a combination of advanced AI models and algorithms to convert lip movements and speech into text. The models and algorithms used are:

- *Convolutional Neural Networks (CNNs):*

CNNs are used to analyze lip movements in videos. They are good at identifying patterns and changes in images, which helps the system understand the subtle movements of lips as someone speaks. By processing each video frame, CNNs help the system figure out what sounds or words are being spoken, even without the actual sound.

- *Connectionist Temporal Classification (CTC):*

CTC helps the system match lip movements or sounds to the correct text. This method helps the system to improve the accuracy of its predictions.

- *Multimodal Fusion Network:*

Our system does not just rely on lip movements but also considers the audio. To combine both the visual and auditory data, we use a fusion network. This approach improves accuracy, especially when one type of data is unclear, helping the system make better predictions.

- *Support Vector Machines (SVMs):*

SVMs help classify specific speech sounds, like phonemes or words. These models fine-tune the system's predictions by distinguishing between different speech sounds based on the visual features of lip movements.

V. DATA SOURCES AND SOFTWARE INTEGRATION

A. Data Sources

- *Lip Movement Data (Video Frames):*

The main source of data for the SpeakVision comes from videos that capture the user's lip movements. These videos are analyzed frame by frame to observe how the lips change shape while speaking.

- *Audio Data (Speech Sounds):*

While lip reading can work without sound, combining it with audio makes the system much more accurate. The system also processes the sounds that the user is speaking, which helps confirm or clarify the lip movements. When lip movements alone might not be enough to determine a word, the audio helps the system fill in the gaps, improving overall accuracy.

- *Environmental Data:*

External factors can affect the quality of the video and the audio. The system observes these factors and adjusts to enhance accuracy.

B. Software Integration

- *Application Programming Interfaces:*

APIs are the connectors that allow different parts of the system to work together smoothly. APIs help the system pull in video and audio data, process it, and then produce the resulting text.

- *Data Analytics and Machine Learning Tools:*

The system's ability to interpret lip movements and audio is through advanced data analytics and machine learning. These tools analyze the incoming data, recognizing patterns in the lip movements and sounds to predict what is being said. As the system is used more, it learns from feedback, becoming more accurate in its predictions over time.

- *Multimodal Fusion:*

SpeakVision combines both visual and audio data to make the system more accurate. When one of these data sources might be unclear, the other helps fill in the gaps.

VI. SYSTEM ARCHITECTURE

The SpeakVision system architecture is designed to efficiently process and interpret lip movements and audio data in real-time.

A. Data Collection Layer

- *Video Capture:*

The video feed is processed frame by frame to identify lip shapes and movements.

- *Audio Capture:*

The audio helps to provide additional context for the lip movements, enhancing the system's accuracy.

- *Environment Sensing:*

This data helps adjust video and audio quality to improve lip reading accuracy and speech clarity

B. Preprocessing Layer

- *Video Enhancement:*

The video is preprocessed i.e., it is adjusted for better light conditions and stabilizing the video feed.

- *Audio Enhancement:*

The audio is cleaned up by reducing background noise and enhancing the speech signal to make it clearer. Noise cancellation techniques and volume normalization may be applied.

- *Feature Extraction:*

Both video and audio data are processed to fetch features. For video, movement of the lip is fetched from each frame. For audio, phonetic features are fetched.

C. Ai Models And Processing Layer

- *Convolutional Neural Networks (CNNs):*

CNNs are used to process and analyze the lip movements in the video. These networks recognize patterns in images, which helps them understand the changes in lip shapes as the speech proceeds.

- *Connectionist Temporal Classification (CTC):*

CTC is used to align the speech and lip movements, which may not perfectly correspond due to factors like speech speed. CTC helps the system make the predictions more reliable.

- *Multimodal Fusion Network:*

A multimodal fusion network combines the visual and auditory data to make predictions more accurate.

D. Output Layer:

The system generates the text corresponding to the detected speech. This text can be displayed on a screen or used for further processing.

VII. CHALLENGES, LIMITATIONS AND FUTURE DIRECTIONS

- *Accuracy with Multiple Speakers and Unclear Speech:*

One of the biggest challenges for SpeakVision is understanding conversations where multiple people are speaking at once or when the speech is unclear. When multiple people are talking simultaneously or speaking quickly, the system can struggle make accurate transcriptions. To improve, the system needs to get better at identifying different speakers and understanding fast or unclear speech, making it more reliable in real-world conversations.

- *Support for More Languages and Accents:*

SpeakVision is limited in the variety of languages and accents it can understand. Since people speak differently in various regions, the system can struggle with dialects or less common accents. To make it more versatile, expanding its support to include a wider range of languages and regional accents will help it reach a larger audience and improve accuracy in diverse settings.

- *Real-Time Video Challenges*:

Real-time videos can be challenging, because training the data with real-time facial and speech recognition is difficult. Future upgrades could focus on making the system more adaptable to real-time videos.

- *Understanding Context and Ambiguous Speech:*

While lip reading can provide a lot of useful information, SpeakVision still faces difficulties in understanding the full context of what is being said. Words or phrases can be ambiguous, and without understanding the broader context, the system might misinterpret what is being communicated. To improve, the system could be enhanced to better grasp the flow of conversation and the subtle cues that shape meaning.

VIII. CONCLUSION

SpeakVision is a project that brings together lip reading and speech-to-text technologies to offer a more inclusive and efficient way for people to communicate.

While the system is promising, it is still a work in progress, especially when it comes to handling challenges like background noise, or diverse accents. The potential for improvement is huge, with future advancements likely to make it even more accurate, adaptable, and responsive.

Looking ahead, we expect SpeakVision to keep evolving, incorporating new AI technologies and optimization techniques that will make it an even more powerful tool.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to all the faculty of Information Science and Engineering Department at RNS Institute of Technology. We are particularly thankful to our Guide, Ms. Ashwini M Rayannavar, for her invaluable guidance and support in shaping our ideas and refining our research methodology.

REFERENCES

[1]. Yannis M. Assael1, Brendan Shillingford1, Shimon Whiteson1 & Nando de Freitas123 Department of Computer Science, University of Oxford, Oxford, UK 1 Google DeepMind, London, UK 2 CIFAR, Canada 3, "LipNet"

[2]. Fu, S. Yan, and T. S. Huang. Classification and feature extraction by simplification. IEEE Transactions on Information Forensics and Security, 3(1):91–100, 2008.K. Eves and J. Valasek, "Adaptive control for singularly perturbed systems examples," Code Ocean, Aug. 2023. [Online]. Available: https://codeocean.com/capsule/4989235/tree

[3]. Garg, J. Noyola, and S. Bagadia. Lip reading using CNN and LSTM. Technical report, Stanford University, CS231n project report, 2016.

[4]. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5):602–610, 2005.

[5]. McGurk and J. MacDonald. Hearing lips and seeing voices. Nature, 264:746–748, 1976.

[6]. Noda, Y. Yamaguchi, K. Nakadai, H. G. Okuno, and T. Ogata. Lipreading using convolutional neural network. In INTERSPEECH, pp. 1149–1153, 2014.

[7]. F. Woodward and C. G. Barber. Phoneme perception in lipreading. Journal of Speech, Language, and Hearing Research, 3(3):212–222, 1960.