Fraud Detection in Online Interview Using Audio Visual Synchronization

1.Mr. B. Ramana Babu, Assistant Professor, 2.Y. Sireesha, 3.R. Suresh
4.V. Sai Srinivas , 5.V. Hemanth,6.S. Sumendra Kumar
Department of Computer Science Lendi Institute of Engineering and technology
JNTU GV, Vizianagaram,Andhra Pradesh,India
ramanababu.b@lendi.edu.in, ysireesha8985@gmail.com,sureshrouthu098@gmail.com,
vallepusaisrinivas@gmail.com,vabbilisettyhemanth@gmail.com,sahusumendrakumar46642@gmail.com

***Abstract*—The increasing reliance on remote communication platforms for processes like job interviews necessitates robust methods to verify participant authenticity. This project proposes an automated system to detect potentially fraudulent or spoofed interviews by analyzing audio-visual synchronization cues. Manual verification methods lack scalability and struggle against advanced manipulations like deepfakes or pre-recorded responses. To address this, the system segments interview recordings into manageable clips and extracts text independently from lip movements (via lip-sync-to-text models) and spoken audio (using speech recognition). These parallel text streams are aligned using timestamped subtitles, then transformed into semantic vector embeddings. The similarity between these embeddings is measured using metrics like Cosine Similarity. A low similarity score indicates a potential manipulation. This system offers a scalable, objective, and datadriven solution for authenticating remote interviews and enhancing recruitment integrity.**

Keywords—Audio-Visual Synchronization, Lip Reading, Speech Recognition, Semantic Embeddings,

I. INTRODUCTION

The global shift to remote work and digital communication has revolutionized hiring, making video interviews a standard, convenient practice. However, this transition has also increased security risks—particularly the rise of interview fraud through proxy participants who impersonate actual candidates. Such deception can lead to poor hiring decisions, financial losses, legal risks, and reputational harm. Traditional verification methods like ID checks, facial recogn ition, and manual supervision are often ineffective against advanced spoofing tactics such as deepfakes or pre-recorded responses. These approaches lack the real-time accuracy needed to detect subtle inconsistencies during interviews.

To address this, the project proposes an automated system that verifies candidate authenticity by analysing the synchronization between lip movements and spoken audio. Interview videos are segmented into clips, and text is independently extracted from audio and visual streams. These are converted into semantic embeddings and compared using a customized similarity algorithm. Low similarity scores may indicate impersonation. Leveraging machine learning, speech processing, and computer vision, this scalable system provides a reliable, real-time solution to enhance security in remote hiring.

The integrity of video interviews can be compromised through a variety of deceptive techniques, ranging from basic impersonation to highly sophisticated manipulations. Common forms include the use of proxy candidates, where someone else attends the interview on behalf of the actual applicant, and pre-recorded responses played back during live sessions. More advanced threats are emerging with the rise of AI-driven technologies,.

Various machine learning and signal processing techniques are increasingly applied to detect manipulation in remote video interviews by analyzing audio-visual synchronization. This project introduces a novel approach grounded in the principle that natural human speech requires tightly coordinated movements of the lips, tongue, and jaw with the produced audio. The system independently extracts text from visual lip movements using lip-reading models and from audio using speech-to-text conversion. These two streams are then semantically analyzed and compared using natural language processing techniques. The underlying hypothesis is that genuine, unmanipulated interviews will show high semantic similarity between the visual and audio-derived text, while manipulated segments—such as those involving dubbing, inaccurate lip-syncing, or synthetic speech—will exhibit noticeable discrepancies. This method provides a robust foundation for identifying spoofed interviews using multimodal data analysis

II. LITERATURE REVIEW

LipForensics leverages VSR techniques, as detailed by Oghbaie et al. (2025), by pretraining on lip-reading datasets like LRW and LRS2, using 3D CNNs and Transformers to model lip motion. Its spatio-temporal network mirrors VSR models like AV-HuBERT, extracting mouth region features and capturing temporal dynamics. Pretraining learns natural lip patterns, enabling detection of forged movements during fine-tuning on FakeAVCeleb. Unlike VSR's text output, LipForensics performs binary classification (real vs. fake), adapting VSR pipelines for security. The survey highlights self-supervised learning, which could enhance LipForensics' robustness to limited forgery data. Both rely on datasets with diverse speech, though LRW's vocabulary limits generalization. LipForensics' high ROC AUC reflects VSR's progress, like 26.9% WER on LRS2. Its lightweight design aligns with VSR's real-time goals. Multimodal trends in VSR suggest LipForensics could integrate audio cues. This alignment underscores VSR's versatility in biometric applications.

RVTALL's multimodal dataset advances VSR by integrating 7.5 GHz UWB and 77 GHz mmWave radar data, laser, and depth camera visuals, capturing lip and vocal cord movements. Unlike LRW/LRS3, used in VSR models like AV-HuBERT, RVTALL's RF modalities detect physical vibrations, enabling contactless lip reading. Its 400 minutes of data from 20 participants, covering vowels, words, and sentences, support tasks like vowel classification and speech enhancement. Oghbaie et al. (2025) highlight VSR's need for diverse data, which RVTALL addresses through novel modalities. The dataset's open-access scripts facilitate model training, aligning with VSR's push for reproducibility. Microsoft Kinect V2 ensures high-fidelity audiovisual data, complementing radar inputs. RVTALL's focus on silent speech aligns with VSR applications for accessibility. Its validation via machine learning underscores reliability for research. The dataset could enhance VSR models by adding physical speech cues. RVTALL sets a new standard for multimodal VSR datasets.

The survey by Oghbaie et al. (2025) reviews deep learning's role in advancing automatic lip reading, transforming VSR into a standalone technology. It outlines the VSR pipeline: lip region extraction, feature extraction with 3D CNNs, and classification using Transformers or LSTMs. 7 Datasets like LRW and LRS2 drive progress, reducing WER to 26.9% on LRS2. Applications include silent speech interfaces, biometrics, and accessibility. Pretraining, as in AV-HuBERT, enhances robustness via audiovisual data. Self-supervised learning addresses data scarcity, a key innovation. The survey emphasizes lightweight models for real-time use, aligning with LipForensics' needs. Multimodal integration, like RVTALL's RF data, is a future trend. Evaluation metrics like WER guide progress. This work maps VSR's current state and potential.

The survey by Oghbaie et al. (2025) details technical advancements in VSR architectures, emphasizing deep learning's role in achieving low WERs, such as 26.9% on LRS2. Early models like LipNet used 3D CNNs and LSTMs with CTC loss, while recent ones leverage Transformers for superior temporal modeling. Conformer architectures combine convolution and self-attention, capturing both local and global lip movement patterns. Self-supervised models like AV-HuBERT pretrain on audiovisual data, enhancing robustness to noise and speaker variability. Feature extraction now integrates facial landmarks and optical flow, improving lip region precision. Lightweight models, such as those based on MobileNet, enable real-time VSR on mobile devices. Datasets like LRW and LRS3 provide diverse training data, though vocabulary limitations persist. The survey highlights the shift to end-to-end training, reducing reliance on separate phoneme recognition. Multimodal inputs, as in RVTALL's RF data, are emerging to bolster accuracy. These advancements align with LipForensics' need for robust lip motion analysis.

Lina and Latha (2023) introduce a facial spoofing detection method using weighted deep ensemble learning, combining DenseNet201 and MiniVGG architectures to enhance biometric security. A comparative study of DenseNet201, DenseNet169, VGG16, MiniVGG, and ResNet50 showed DenseNet201's superior recall and MiniVGG's high precision, justifying their selection. The ensemble model weights predictions from both architectures to improve detection accuracy for spoofing attacks like print, replay, and 3D mask attacks. The method processes facial images to classify them as genuine or spoofed, achieving robust performance across diverse conditions. Unlike LipForensics, which focuses on lip motion, this approach analyzes entire facial features, leveraging deep learning's representational power. The study emphasizes generalization, addressing unseen attack types, a key concern in face anti-spoofing. Its use of pre-trained models aligns with VSR's pretraining strategies, as noted in Oghbaei et al. (2025). The method's computational efficiency supports real-time applications, similar to LipForensics' goals.

Performance Evaluation

LipNet, a pioneering end-to-end model for sentence-level lip-reading, and subsequent visual speech recognition (VSR) systems are typically evaluated using metrics borrowed from automatic speech recognition (ASR). The most common and important performance metrics for LipNet.

Word Error Rate (WER): This is the most widely used metric for evaluating the overall accuracy of a lip-reading system at the word level. It measures the number of errors (substitutions, deletions, and insertions) required to transform the predicted sequence of words into the ground truth sequence, divided by the total number of words in the ground truth. A lower WER indicates better performance.

The formula for WER i

Precision, a key metric, measures how many of the predicted positive cases were actually correct. It is calculated using the formula:

A high precision rate ensures that only the intended individuals are recognized, reducing misclassification errors.

$$\mathrm{WER} = \frac{S+D+I}{N}$$

Where: ● S is the number of substitutions. ● D is the number of deletions. ● I is the number of insertions. ● N is the total number of words in the reference (ground truth) defined as:

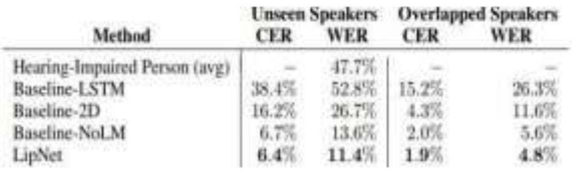
Character Error Rate (CER): Similar to WER, CER measures the accuracy at the character level. It calculates the number of errors (substitutions, deletions, and insertions of characters) needed to change the predicted character sequence into the ground truth sequence, divided by the total number of characters in the ground truth. A lower CER indicates better performance. The formula for CER is:

$$\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c}$$

Where: ● Sc is the number of character substitutions. ● Dc is the number of character deletions. ● Ic is the number of character insertions. ● Nc is the total number of characters in the reference (ground truth

Comparison

Table Lipnet model:



	Unseen Speakers		Overlapped Speakers	
Method	CER	WER	CER	WER
Hearing-Impaired Person (avg)	–	47.7%	–	–
Baseline-LSTM	38.4%	52.8%	15.2%	26.3%
Baseline-2D	16.2%	26.7%	4.3%	11.6%
Baseline-NoLM	6.7%	13.6%	2.0%	5.6%
LipNet	**6.4%**	**11.4%**	**1.9%**	**4.8%**

Performance of LipNet on the GRID dataset compared to the baselines, measured on two splits: (a) evaluating on only unseen speakers, and (b) evaluating on a 255 video subset of each speakers' sentences. The table 5.1 clearly shows that LipNet significantly outperforms the baseline models (Baseline-LSTM, Baseline-2D, Baseline-NoLM) in both CER and WER for both unseen and overlapped speaker scenarios. For instance, on overlapped speakers, LipNet achieves a WER of 4.8%, much lower than the baselines (e.g., Baseline-LSTM at 26.3%). It also shows that the average hearing-impaired person's lip-reading performance on the GRID dataset has a much higher WER (47.7%) than LipNet, highlighting the superior capability of the LipNet model on this specific task and dataset. The distinction between unseen and overlapped speakers is important, showing how well models generalize to new individuals. LipNet's performance is better on overlapped speakers, as expected, but it still maintains a relatively low error rate on unseen speakers compared to the baselines.

MACHINE LEARNING

Various machine learning and signal processing techniques are increasingly applied to detect manipulation in remote video interviews by analyzing audio-visual synchronization. This project introduces a novel approach grounded in the principle that natural human speech requires tightly coordinated movements of the lips, tongue, and jaw with the produced audio. The system independently extracts text from visual lip movements using lip-reading models and from audio using speech-to-text conversion. These two streams are then semantically analyzed and compared using natural language processing techniques. The underlying hypothesis is that genuine, unmanipulated interviews will show high semantic similarity between the visual and audio-derived text, while manipulated segments—such as those involving dubbing, inaccurate lip-syncing, or synthetic speech—will exhibit noticeable discrepancies. This method provides a robust foundation for identifying spoofed interviews using multimodal data analysis Analyzing video, extracting meaningful information from subtle lip movements, transcribing audio accurately, and comparing semantic content necessitates the use of Machine Learning (ML). Human analysis is impractical for large volumes and prone to subjective bias and fatigue. ML offers several key advantages: PatternRecognition: ML, especially deep learning models, excels at learning complex, non-linear patterns in high-dimensional data, such as video frames and audio signals. Automation&Scalability: ML pipelines can process and analyze numerous interview recordings automatically, offering a scalable solution for large organizations. Objectivity: ML reduces subjectivity in assessments by relying on quantitative similarity scores derived from data, ensuring consistent decision-making. Adaptability: ML models can be fine-tuned or retrained as new spoofing techniques emerge

Learns from Audio-Visual Data: ML models in this project learn from both audio and visual data of video interviews, such as lip movements and spoken words. By analyzing these features, the model identifies patterns and inconsistencies that often indicate fraudulent behavior like lip-sync manipulation or pre-recorded responses. Improves Fraud Detection Accuracy: Using machine learning algorithms like Random Forest, SVM, and deep learning models, the system compares and selects the best-performing models for detecting interview fraud. This improves prediction accuracy, making the system more reliable than traditional, manual verification methods. Gets Better with More Data: As more interview data is collected or new manipulation techniques emerge, ML models can be retrained with updated datasets, continuously improving the system's ability to detect new types of interview fraud. Handles Multi-Modal Data: The system processes both audio and visual features simultaneously, allowing it to understand complex interactions between spoken words and lip movements. This holistic approach is difficult to achieve manually, especially when dealing with large datasets.

Instant Fraud Detection: The integration of ML models into a user-friendly interface allows real-time analysis of interview recordings. HR professionals can quickly assess interview authenticity and flag suspicious candidates, making the hiring process more efficient. Personalized Analysis: The system analyzes each interview individually, using the unique data from both the audio and video streams to provide specific, accurate results. This personalized approach helps to detect manipulation tailored to each interview, rather than relying on generic detection methods.

System Analysis Definition

System analysis is a problem-solving technique that decomposes a system into its component pieces for the purpose of studying how well those component parts work and interact to accomplish their purpose. System analysis is the process of studying a procedure in order to identify its goals and purposes and create systems and procedures that will achieve them in an efficient way. The development of a computer-based information system includes a systems analysis phase which produces or enhances the data-model which itself is a precursor to creating or enhancing a database. There are a number of different approaches to system analysis.

Existing System:

Interviewers rely on observing eye movements, facial expressions, and audio/video glitches to gauge candidate engagement. However, this approach is subjective, prone to error, and easily fooled by subtle manipulations like deep fakes or pre-recorded answers. It can also be mentally taxing for interviewers, leading to inconsistent judgments. Traditional methods, such as asking for a photo ID, confirm identity but don't assess interview authenticity. They ensure the person is who they claim to be but fail to verify if the candidate is engaged or providing spontaneous responses.Proctoring tools monitor eye movements and background noise, but they do not address the core issue of interview authenticity—audio-visual synchronization. These systems cannot effectively detect manipulated interviews that involve synchronized lip movements and speech.

Lack of Objectivity: Many traditional methods rely on subjective human judgment or basic data inputs. Human observation, for example, is highly susceptible to individual biases, fatigue, or distractions, which can affect the accuracy of fraud detection. Scalability Issues: The existing systems are not easily scalable, particularly for large organizations or high volumes of interviews. Manual methods, such as human observation, are time-consuming, and identity verification may not be efficient when dealing with numerous candidates.Intrusiveness: Some identity verification and proctoring methods can feel intrusive to candidates, potentially affecting the candidate experience and possibly deterring applicants. This can also create privacy concerns, especially if video or audio is constantly being monitored.Ineffectiveness Against Advanced Spoofing Techniques: Traditional methods are often ineffective against sophisticated AI-driven spoofing techniques, such as deepfakes or seamlessly injected pre-recorded answers. These manipulations can be nearly impossible to detect through human observation or even basic proctoring software. Failure to Analyze Audio-Visual Synchronization: The core issue of detecting manipulations—whether the audio matches the lip movements—remains unsolved by current systems. Human observation cannot reliably assess whether the spoken words align with the lip movements, and most proctoring software fails to detect this kind of subtle inconsistency.

Challenges

Data Quality and Availability: Poor video or audio quality can hinder the model's accuracy, affecting lip-sync and speech-to-text algorithms, leading to misclassifications. Imbalanced Data: An imbalance between genuine and manipulated interviews can result in biased models, making the system more effective at detecting genuine interviews than manipulations. Feature Selection: Identifying key features in video and audio data that indicate manipulation is complex but crucial for model performance. Overfitting: Models may perform well on training data but struggle with new interview recordings, requiring proper validation and data augmentation to avoid overfitting. Interpretability: Deep learning models can be "black boxes," making it difficult to explain predictions to HR professionals or stakeholders. Data Privacy and Security: Sensitive interview data must be handled according to privacy regulations (e.g., GDPR, HIPAA), ensuring data security. Generalization: Models trained on one group may not work well for different demographics or technological setups, requiring broad adaptability. Integration with Hiring Platforms: Integrating the system with real-world hiring platforms may face technical, operational, and regulatory challenges, requiring compatibility and legal compliance.

Benefits of the System:

The proposed system brings a significant improvement in ensuring the integrity of remote interviews by leveraging advanced machine learning, computer vision, and natural language processing techniques. One of the primary benefits is objectivity. Unlike traditional methods that depend heavily on human observation, this system uses quantifiable similarity metrics between audio and visual streams to flag potentially spoofed content. This eliminates bias, reduces human error, and ensures consistency in decision-making. Another major advantage is automation and scalability. The system can process a large number of interview videos quickly and without human supervision, making it highly suitable for large-scale recruitment processes. It can be configured for real-time alerts or used for post-interview batch processing, offering flexibility in deployment according to the organization's needs. The system also enhances security and authenticity by identifying subtle manipulations that typically escape human detection, such as lip-sync discrepancies, dubbed audio, or deepfake-based 14 impersonation. Unlike intrusive proctoring tools that may affect candidate comfort or privacy, this system offers a non-intrusive solution that analyzes naturally occurring data during the interview. Moreover, the results produced by the system are transparent and explainable. By outputting similarity scores and highlighting mismatched segments, the tool helps HR professionals understand the rationale behind the flagging, making it easier to take informed actions. It is also designed to be modular and integration-friendly, enabling seamless connection with existing hiring platforms or applicant tracking systems. Furthermore, the system is built with future adaptability in mind. As manipulation tactics become more sophisticated, the system's architecture allows for enhancements through additional detection layers or updated AI models. Additionally, it is privacy-conscious, adhering to regulations such as GDPR and HIPAA to ensure candidate data is handled responsibly and securely. Lastly, the system offers substantial cost and time efficiency. By reducing the need for repeated interviews, minimizing human workload, and increasing the accuracy of fraud detection, it helps organizations improve hiring decisions while saving time and resources

A. Proposed System

In our interview spoof detection project, we propose an advanced system that uses machine learning techniques to overcome the limitations of traditional manual observation and basic proctoring software. Instead of relying on subjective human judgment or surface-level monitoring, the proposed system learns from audio-visual data to identify subtle inconsistencies between lip movements and spoken audio, offering accurate, scalable, and objective fraud detection.

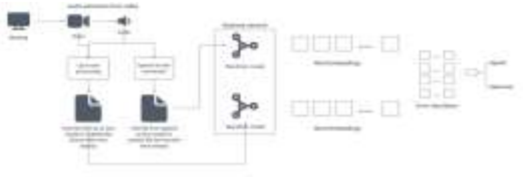


Fig. 1. System Architecture

B. Pre-Processing

A recorded meeting video is the initial input to your system. This video contains two main components: 1. Visual information (video frames) → contains the person's facial and lip movements. 2. Audio signal → contains the actual speech. To process them separately for analysis, the system splits the video into two independent streams:

Video Stream (for Lip Reading):

The video stream refers to the sequence of image frames in the video. This stream is used to analyse lip movements of the speaker, frame by frame. Each frame is processed to focus on the region of interest (ROI)—usually the mouth/lips area. Later, these visual cues are fed into a lip-to-text model that predicts spoken words based purely on lip motion.

Audio Stream (for Speech Recognition):

The audio track is extracted from the video file using audio processing tools (e.g., ffmpeg, Libros).This audio contains the spoken words as sound waves. It is passed into a speech-to-text engine (like Google Speech API, Deep Speech, Whisper) to generate a transcript of spoken content.

Lip to Text Processing Lip to Text Processing refers to the process of converting silent lip movements captured from a video into textual data, predicting what words were spoken based solely on visual input. This module is crucial in scenarios where audio may be missing, tampered with, or unreliable, and it serves as the foundation for comparing visual and auditory speech in deepfake detection. It follows these main steps: 1. Lip Region Extraction Lip region extraction is the process of isolating the mouth area from a video frame for further analysis. This involves:

Detecting the face in each frame using face detection algorithms (e.g., D lib, Media pipe, OpenCV Haar cascades).

Identifying facial landmarks to precisely locate the lip contours or bounding box. 20 Cropping the lip area while maintaining consistency across frames (accounting for head movement).

Feature Extraction from Lip Movements:

After isolating the lip region, the next step is to convert the video frames into a feature representation that captures the visual patterns of speech. This step uses: ● Convolutional Neural Networks (CNNs) to extract spatial features from each frame (e.g., shape, texture of lips). ● Optionally, optical flow techniques to capture motion information between frames.

In models like Mu Se Talk, the visual encoder maps these lip features into a latent embedding space that aligns with speech features. Temporal Modeling: Lip reading is a sequence problem — words are formed over multiple frames, not in a single snapshot. Therefore, after extracting per-frame features, they are fed into models that understand temporal patterns:

Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks to capture dependencies across time. Or, more modern alternatives like Transformers that use attention mechanisms to model long-range temporal context. In Mu Se Talk architecture, a temporal module integrates frame-wise features into a global representation of the spoken utterance.

Mapping to Phonemes/Words: Once a temporal feature sequence is created, the system decodes it into a sequence of phonemes (basic sound units) or directly into words. Two common approaches: CTC (Connectionist Temporal Classification) decoding: Allows the model to output sequences of varying lengths without strict alignment between input frames and output letters. ● Sequence-to-sequence models (with attention): Maps the input lip features to output text tokens using encoder-decoder frameworks. In Mu Se Talk (which is designed to generate lips from speech), the embedding space connecting lips ↔ speech can be used in reverse to estimate likely speech content from lip motion. 5. Timestamp Alignment and Subtitle Generation: 21 As words are predicted, the system keeps track of when each word starts and ends based on frame indices or internal timing. Mu Se Talk: Although MuSeTalk is mainly a speech-driven talking head synthesis model (generating realistic lip-sync from audio), it includes key modules useful for lip-to-text tasks:

A cross-modal embedding space connecting visual and audio modalities.A visual encoder that maps lip images to embeddings similar to speech embeddings. Ability to fine-tune or invert mappings for lip→ speech inference. ● By leveraging Mu Se Talk: We use its pretrained ability to associate lip shapes with speech information. Instead of generating lips from speech, we extract latent speech representations from lips, then decode these to phonemes/words.

Speech to Text Conversion (Audio → Speech Recognition → Text) :

This step takes the extracted audio stream from the meeting recording and converts it into a written text transcript using an Automatic Speech Recognition (ASR) system. Here's how it works: 1. Input: Audio Stream The audio track is separated from the video during the earlier extraction phase. This input could be in formats like .wav, .mp3, etc.

WORKING PRINCIPLE

INPUT: Audio-visual recordings of remote interviews.

OUTPUT:Detection of fraudulent behavior (e.g., deepfakes,lip-sync mismatches) in real-time or post-interview analysis.

Step-by-Step Workflow:

1. User Access & Authentication:
 The candidate enters the system through a secure login portal on the Webinaar Meeting Application, ensuring authorized access.

2. Interview Recording:
 The system captures synchronized audio and video data from the interview, including the candidate's

speech and facial movements.

3. Visual Feature Extraction (LipNet):
 A deep learning model called LipNet processes the video frames of the candidate's lip movements. It uses spatiotemporal convolutions (STCNN) and Bi-GRU layers to generate sentence-level textual transcriptions from lip movements.

4. Audio Transcription (Whisper):
 Simultaneously, the audio is transcribed using Whisper, a robust speech recognition model that converts spoken words into text.

5. Word Embedding Generation (Word2Vec):
 The transcribed texts (from both LipNet and Whisper) are converted into numerical vectors (embeddings) using a pretrained Word2Vec model trained on the Common Crawl corpus.

6. Semantic Comparison (Siamese Network with Cosine Similarity):
 Both vectors are passed through a Siamese network to compute the Cosine Similarity. A high similarity indicates that the spoken words match the lip movements, whereas a low score may indicate a mismatch or manipulation.

7. Decision Output:
 Based on a similarity threshold, the system flags possible fraud

III. Proposed Methodologies

The tools and methodology to implement and evaluate face detection and tracking are listed below.

A. OpenCV

The proposed interview spoof detection system introduces several key advantages over traditional interview monitoring approaches. First and foremost, it offers improved accuracy by analyzing fine-grained audio-visual cues, such as synchronization between lip movements and spoken content, which are often imperceptible to human reviewers or basic proctoring software. This makes it highly effective in identifying both overt and subtle forms of manipulation. Another major advantage is faster verification, as the system automates the evaluation process, significantly reducing the time needed to assess the authenticity of interviews. This allows HR 18 professionals to make quicker and more confident hiring decisions. The system is also capable of detecting sophisticated manipulations, including deepfakes, audio dubbing, and pre-recorded responses, which are increasingly used to bypass traditional identity checks. The solution provides personalized assessments by evaluating each interview based on unique characteristics extracted from both video and audio inputs. This context-aware approach increases the reliability of the detection process. Its scalability ensures that it can be used effectively by organizations handling large volumes of interviews, such as during campus placements or global remote hiring drives. Furthermore, the system serves as a support tool for HR professionals, offering objective insights and helping reduce cognitive load and fatigue that can result from repeated manual evaluations. Finally, the system is designed for continuous improvement, with the ability to learn from new data and adapt to emerging spoofing techniques over time, thereby ensuring sustained performance and resilience in real-world applications

B. Data for lipnet

GRID is a large multitalker audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Sentences are of the form "put red at G9 now". The corpus, together with transcriptions, is freely available for research use.

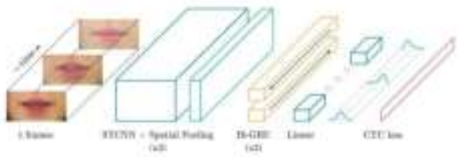
C. Face Detection

This system can recognize faces from HD video collected images for the purpose of studying and detecting the face. Face from Section IV above, face detection detects where a face is in a picture, and it is accomplished by scanning the various picture scales and detecting the face by extracting the precise patterns. The A Haar-Like Feature function is used to build the prototype. Haar classifier facial detection is used in OpenCV to build a search window that scrolls over images and checks if a certain section of a picture resembles a face or not.

D. Feature Extraction

Face detection feature extraction involves locating the features of face components in an image. This process is done using a series of mathematical operations. First, the image is converted into grayscale. Then, the pixels are divided into blocks. Each block contains a small area of the image. Next, the image is examined for changes in color intensity. These changes indicate the presence of a face. Finally, the location of the face is determined by comparing the size and shape of the face to a pre-defined template. This process is used in many applications, including face detection, facial expression recognition, and human activity recognition. This process is divided into identification and verification. This solution focuses on two terms: identification to detect the face in real-time video and verification application for facial recognition.

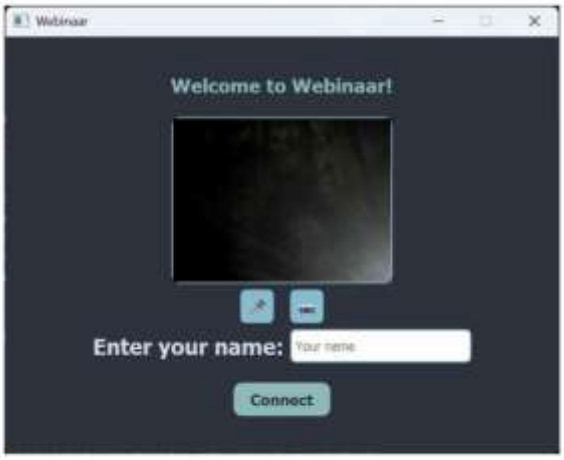
Fig. 2. Feature Selection for Face Detection



The greatest matching score obtained in the previous stage is declared in the final phase of face detection. The configuration will define how the application should act..

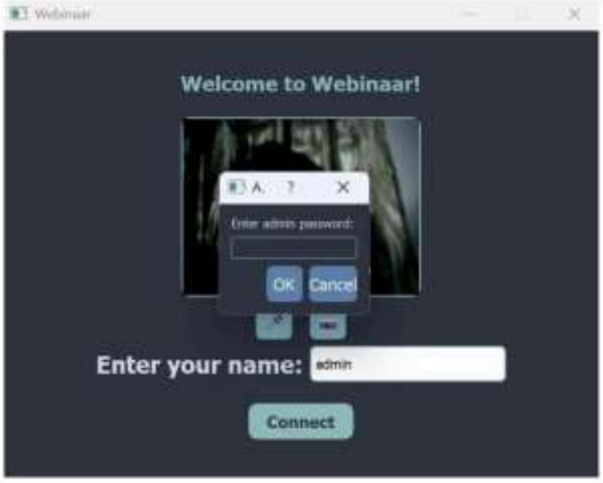
IV. Result

The management system using facial recognition is very easy to use and works smartly in less time. This is an automatic system. Once an administrator has created a student profile in the database, it is automatically used in the facial recognition and recognition process. To initialize this system, the administrator first creates all student profiles.GRID is a large multitalker audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Sentences are of the form "put red at G9 now". The corpus, together with transcriptions, is freely available for research use



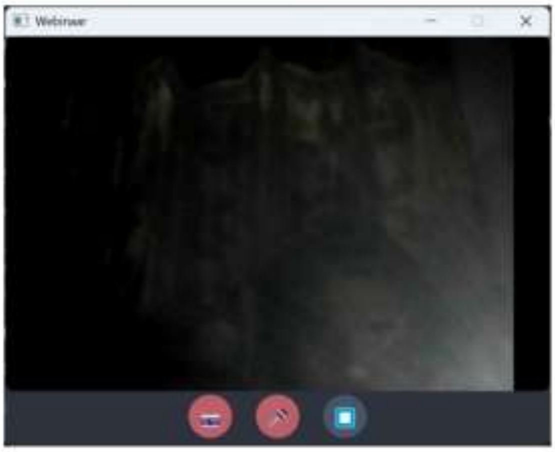
This figure depicts the initial screen a user would encounter when accessing the "Webinaar meeting application." Based on typical web application design, this screen likely serves as the entry point to the system. It would typically include elements such as: Branding: The application's name ("Webinaar meeting application") and possibly a logo. Call to Action: Buttons or links to initiate key actions, such as starting a new meeting, joining an existing meeting, or potentially logging in or signing up. Brief Description: A short explanation of what the application does, likely emphasizing its purpose for online meetings. Navigation: Links to other relevant pages like "About Us," "Features," or "Contact." In the context of your fraud detection project, this welcome screen is the gateway to the interview environment where recordings are made, which are then processed by your system. It sets the stage for the user's interaction with the meeting platform itself

User Authentication:



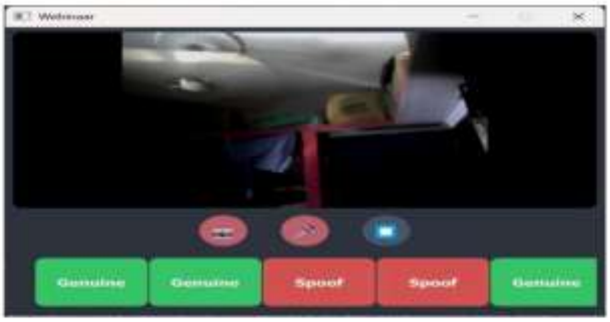
This figure illustrates the process or interface related to user authentication. This is a critical step for security, ensuring that only authorized individuals can access and use the meeting application and potentially the fraud detection features. The image likely shows a login or sign-up interface, which would include: Input Fields: Text fields for entering credentials such as username/email and password. Login/Sign-up Buttons: Buttons to submit the entered information. Password Recovery Option: A link for users who have forgotten their password. Registration Link: An option for new users to create an account. Security Indicators: Visual cues indicating a secure connection (e.g., HTTPS). The "Authentication Protocol" title suggests that this figure might also visually represent the underlying security measures in place during the login process, although the image itself would primarily show the user-facing interface. Secure authentication is vital to protect the integrity of the interview process and the sensitive data involved

Meeting screen:



This figure shows the main interface where the online interview or meeting actually takes place. This is the environment where the audio-visual data is generated that your fraud detection system will analyze. The meeting screen typically includes a variety of elements for facilitating communication and interaction: Video Feeds: Display areas showing the video streams of the participants (the interviewer and the candidate). Audio Controls: Buttons to mute/unmute the microphone. Video Controls: Buttons to turn the camera on/off. Meeting Timer: An indicator showing the duration of the meeting. End Call Button: A prominent button to leave or end the meeting. Recording Indicator: Crucially for your project, there would likely be a visual indicator showing that the meeting is being recorded.

Output:



This meeting screen is the source of the raw video and audio data that your system processes. The quality and nature of the interaction captured on this screen directly impact the performance of your audio-visual synchronization analysis. These GUI images provide a visual context for the environment in which the data for your fraud detection system is generated and collected. They represent the user-facing side of the online interview process that your backend system is designed to analyze for signs of manipulation

V .CONCLUSION

Online interviews have become common in modern recruitment, but they are increasingly vulnerable to sophisticated fraud techniques such as deepfakes, audio dubbing, and pre-recorded answers. Traditional verification tools and manual observation are often insufficient to detect these manipulations, highlighting the need for more robust and automated solutions .The proposed system addresses this challenge by analyzing the synchronization between a candidate's lip movements and spoken audio. Using advanced machine learning techniques, it extracts and compares semantic content from both visual and audio sources. Models like LipNet, Whisper, Word2Vec, and Sentence-BERT are used to generate embeddings, and similarity is measured with Cosine Similarity to flag inconsistencies. This automated approach offers several advantages: improved accuracy in detecting subtle mismatches, the ability to scale across large volumes of interviews, and reduced reliance on subjective human judgment. It also supports continuous updates, allowing the system to adapt to emerging spoofing methods and strengthen the security of remote hiring processes.

REFERENCES

[1] Assael, Y., Shillingford, B., Whiteson, S., & Freitas, N. (2016). LipNet: End-to-End Sentence-level Lipreading. arXiv preprint arXiv:1611.01599. [Referenced in]Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D & Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585(7825), 357-362. (Reference for NumPy) [2] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781. (Reference for Word2Vec) [3] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830. (Reference for Scikit-learn/Cosine Similarity) [4] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 484–492). [Referenced in] [5] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. arXiv preprint arXiv:2212.04356. (Reference for Whisper) [6] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. [7] Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. (Reference for Gensim/Word2Vec) [8] Rudrabha. (n.d.). Wav2Lip GitHub Repository. Retrieved from https://github.com/Rudrabha/Wav2Lip [Referenced in] [9] S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In Workshop on Multi-view Lip-reading, ACCV, 2016b. [10] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555, 2014. [11] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. [12] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for largevocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1): 30–42, 2012. [13] F. DeLand. The story of lip-reading, its genesis and development. 1931. [14] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior. Recent advances in the automatic recognition of audiovisual speech. Proceedings of the IEEE, 91(9):1306–1326, 2003. [15] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first faciallandmark localization challenge. In IEEE International Conference on Computer Vision Workshops, pp. 397–403, 2013. [16] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European Conference on Computer Vision, pp. 818–833, 2014. [17] G. Zhao, M. Barnard, and M. Pietikainen. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 11(7):1254–1265, 2009. [18] Eric H. Huang, R. Socher, C. D. Manning and Andrew Y. Ng. Improving Word Representations via Global Context and Multiple Word Prototypes. In: Proc. Association for Computational Linguistics, 2012. [19] T. Mikolov. Language Modeling for Speech Recognition in Czech, Masters thesis, Brno University of Technology, 2007. [20] T. Mikolov, S. Kombrink, L. Burget, J. Cernock ˇ y, S. Khudanpur. Extensions of recurrent neural ´ network language model, In: Proceedings of ICASSP 2011. [21] D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning internal representations by backpropagating errors. Nature, 323:533.536, 1986. [22] P. D. Turney. Measuring Semantic Similarity by Latent Relational Analysis. In: Proc. International Joint Conference on Artificial Intelligence, 2005. [23] R. Socher, E.H. Huang, J. Pennington, A.Y. Ng, and C.D. Manning. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In NIPS, 2011. [24] T. Mikolov, W.T. Yih, G. Zweig. Linguistic Regularities in Continuous Space Word Representations. NAACL HLT 2013. [25] Y. Bengio, R. Ducharme, P. Vincent. A neural probabilistic language model. Journal of Machine Learning Research, 3:1137-1155, 200