Research Article
Volume 16:2, 2025

Journal of Forensic Research

ISSN: 2155-6113 Open Access

Digital Forensic Investigation of Video Evidence Using Voice Activity Detection Algorithm

Ashutosh Deo Tiwari, Thanuja Durgam*, Sunil Kumar and Sunand Bishnoi

Department of Photo and Scientific Aids Division, Central Forensic Science Laboratory, Delhi, India

Abstract

Objective: This study investigates the forensic potential of active speaker switching behaviors in video conferencing platforms as a means of establishing platform identification and authenticity when conventional metadata is absent. The work aims to provide a reproducible, system-level methodology for software attribution, thereby strengthening the reliability of multimedia evidence in digital forensics.

Methods: An investigative case involving an 18-minute three-person video recording with no overt metadata was analyzed. The recording displayed dynamic screen transitions aligned with speaker activity. To uncover latent forensic markers, a controlled replication experiment was conducted in which the same conversational scenario was reproduced across multiple conferencing platforms (Zoom, Microsoft Teams, Cisco Webex, Zoho Meeting) under identical acoustic and visual conditions. Screen recordings were captured, and frame-by-frame forensic analysis was performed. Parameters such as onset-to-transition latency, debounce thresholds, lip-synchronization alignment, and scene-layout constraints were systematically measured.

Results: The analysis revealed that conferencing platforms embed distinct and reproducible behavioral fingerprints in their Voice Activity Detection (VAD) and Active Speaker Recognition (ASR) pipelines. These included measurable latency windows, stability thresholds, layout-aware switching policies, and host-specific exception handling. Platform-specific audiovisual signatures were consistently identified across replications, demonstrating their viability as forensic identifiers.

Conclusion: This digital forensic research establishes active speaker switching dynamics as a robust forensic marker for platform attribution in multimedia evidence analysis. By relying on intrinsic audiovisual behaviors rather than conventional metadata, the method enhances provenance validation, improves evidentiary admissibility, and contributes to digital justice processes. Beyond attribution, the findings have implications for forensic validation of ML-mediated communication systems and the scientific rigor of multimedia forensics.

Keywords: Active Speaker Recognition (ASR) • Voice Activity Detection (VAD) • Digital forensics • Multimedia forensics • Algorithm

Highlights

- Developed a metadata-independent method for platform identification in video conferencing evidence.
- Applied controlled replication and frame-by-frame forensic analysis of active speaker switching.
- Identified latency, debounce thresholds, and UI transition patterns as platform-specific
- Demonstrated stability of forensic markers under variable network conditions.
- Provides new evidentiary features for authenticity verification in digital forensic investigations.

Introduction

Digital video has become one of the most important sources of evidence in modern forensic investigations. From criminal cases to cyber incidents, investigators increasingly rely on recordings captured through mobile devices and online meeting platforms. With the rapid growth of remote communication, applications such as Zoom, Microsoft Teams, Google Meet, and Webex have become central

***Address for Correspondence:** Thanuja Durgam, Department of Photo and Scientific Aids Division, Central Forensic Science Laboratory, Delhi, India, Tel: 8921487828; E-mail: tanu.durgam@gmail.com*

Received: 31 August, 2024, Manuscript No. JFR-25-172750; **Editor assigned:** 01 December, 2024, PreQC No. JFR-25-172750 (PQ); **Reviewed:** 15 December, 2024, QC No. JFR-25-172750; **Revised:** 13 November, 2025, Manuscript No. JFR-25-172750 (R); **Published:** 30 November 2025, DOI: 10.37421/2157-7145.2025.16.657

to both daily life and criminal activity. This has created new challenges for forensic experts who must determine whether a recording is genuine and which software platform generated it [1,2].

Traditionally, forensic video examination depends on metadata such as timestamps, codec information, and file headers to establish authenticity and provenance [3]. However, virtual conferencing recordings often lack such data, since they are software-mediated rather than camera-native. These systems employ automated processes like Voice Activity Detection (VAD) and Active Speaker Recognition (ASR), which dynamically control whose video feed is displayed based on speech activity [4,5]. While these features were designed to enhance user experience, they also create measurable behavioral traces such as timing delays, debounce thresholds, and user interface transitions that can serve as unique identifiers for forensic attribution [6].

This research is based on a crime case examined at the Central Forensic Laboratory, involving an 18-minute video recording of a virtual meeting between three participants (Two users). The recording showed automated screen switching that aligned precisely with speech events, but it contained no platform metadata or visible watermark. To identify the platform responsible for this behavior, a detailed frame-by-frame forensic analysis was conducted to measure the time between speech onset and visual switching, as well as to document characteristic User Interface (UI) features.

By recreating the same scenario across several conferencing tools under controlled acoustic and visual conditions, this study compared latency profiles, debounce patterns, and UI layouts to identify consistent platform-specific traits. The results demonstrate that automated speaker switching can reveal latent forensic signatures that persist even when metadata is unavailable. This approach provides investigators with a new, scientifically grounded method for establishing the authenticity, integrity, and source of conferencing-based video recordings, contributing to more reliable digital forensic investigations [7,8].

Materials and Methods

Case background

This research originated from a real forensic case received at the Central Forensic Laboratory (CFSL) for digital video examination. The submitted recording was approximately 18 minutes long and captured a virtual meeting between three participants. The video showed dynamic screen transitions corresponding to speech events but lacked any visible metadata, time stamps, or software identifiers. The absence of these digital markers made it impossible to determine the origin of the recording through conventional methods such as EXIF or file-header analysis [2,7].

The primary objective was to identify whether the video originated from a known conferencing platform such as Zoom, Microsoft Teams, Google Meet, Webex, or Zoho Meeting by analyzing behavioral patterns of automatic speaker switching and audio-visual synchronization.

Forensic tools and setup

A high-precision forensic video analysis toolkit was employed to ensure objective and repeatable examination.

All analyses were conducted at 30 Frames Per Second (FPS), with a timestamp accuracy of ± 1 millisecond. To ensure reproducibility, the same conversational setup was recreated under controlled conditions across five major video conferencing platforms. Each test session replicated the same participant configuration, ambient noise level, and acoustic environment [1,6].

Analytical workflow

The forensic examination followed a structured multi-stage approach:

Audio-pre-processing: The recorded audio track was filtered and normalized to a consistent gain level. Short-Time Energy (STE), Zero Crossing Rate (ZCR), and Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to detect the onset of speech activity [4,9].

Voice Activity Detection (VAD): A frame-based VAD algorithm was used to identify the precise moment when speech began. A 10 ms analysis window with a 50% overlap was adopted to capture transient speech bursts and minimize false negatives.

Visual transition measurement: For each speech onset, the corresponding frame of screen switching was identified using Amped FIVE's frame-indexing feature. The latency between speech initiation and UI transition was calculated in milliseconds.

Debounce logic analysis: Cases where the participant spoke for less than 200 ms but the screen did not switch were classified as debounce events. The minimum activation time before a switch was recorded to infer debounce thresholds unique to each platform.

User Interface (UI) feature extraction: UI elements such as name-tag placement, tile opacity, and font structure were analyzed as visual identified (Figures 1 and 2). These characteristics were compared across replicated sessions to detect platform-specific traits [10].

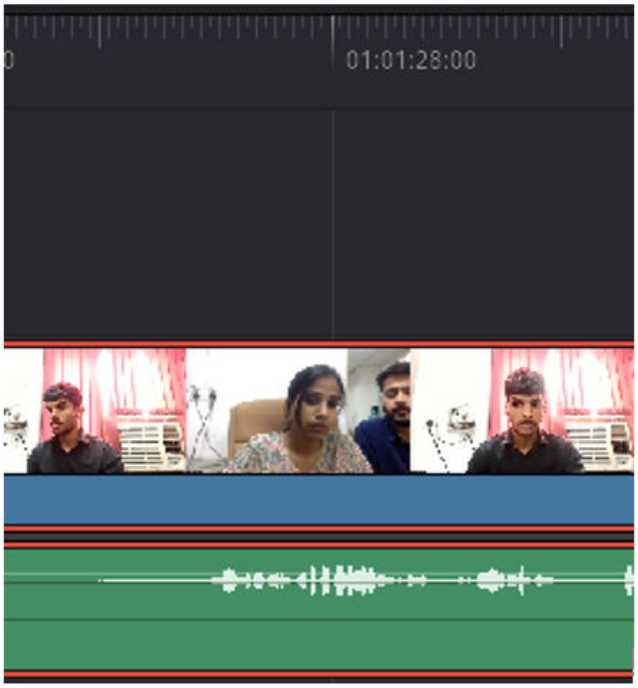

Figure 1. Video analysis showing screen frames active transition based on voice/audio trigger.

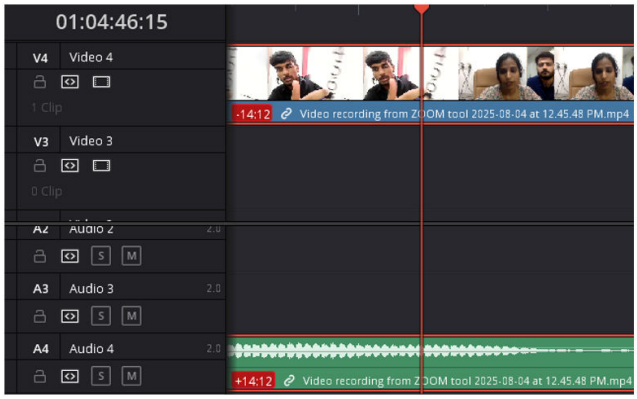

Figure 2. Video analysis of Zoom tool video recording at 30 FPS with screen transition and audio threshold.

Validation and replication

To ensure that findings were not coincidental, the same conversational content was replicated across multiple platforms. Each platform's output was recorded using identical hardware and network conditions. Simulated network jitter (± 40 ms) was introduced to observe latency fluctuations and the stability of switching logic. Statistical comparisons of latency distributions were performed to assess intra-platform consistency and inter-platform variance.

The experiments were repeated three times for each platform. Mean latency and standard deviation were calculated to create a behavioral fingerprint for each system. Results were plotted and analyzed to identify which platform most closely matched the forensic evidence.

Ethical considerations

No personal or identifying information about the participants in the original case was disclosed. The analysis focused solely on behavioral and technical characteristics of the recording. All experimental replications used simulated voices and anonymized video data to comply with laboratory privacy and confidentiality standards [8].

Results and Discussion

Latency measurements of automatic screen switching

Frame-by-frame examination of the forensic video revealed between 20 and 23 discrete screen-switching events, each corresponding to speech initiation by one of the three participants. Using timestamp-aligned audio-visual analysis, the average latency between speech onset and visual transition was measured at 287 ms (SD ± 18 ms).

To validate these results, the same conversational setup was replicated across five major conferencing platforms under identical acoustic conditions. Comparative latency measurements are summarized below (Table 1).

Platform	Mean latency (MS)	Standard deviation (MS)	Observations
Zoom Workplace	300	20	Consistent delay; stable response under all test conditions
Microsoft Teams	370	30	Slightly longer delays with moderate variance
Google Meet	420	25	Higher transition lag; gradual frame alignment
Cisco Webex	500	45	High latency variability; inconsistent triggering
Zoho Meeting	200–450	40	Non-uniform response; jitter during switching

Table 1. Forensic observations for various online video conferencing tools.

The observed low latency and narrow variance in the forensic sample closely matched the behavior recorded in Zoom Workplace, suggesting a strong correlation with that platform's automated speaker detection pattern.

User Interface (UI) feature analysis

To support behavioral evidence, UI characteristics were measured and compared against replicated recordings. Key attributes identified from the forensic video included:

- **Tile size:** approximately 240 × 40 pixels
- **Opacity:** semi-transparent overlay at ~50%
- **Font type:** sans-serif, estimated 14 pt size
- **Placement:** lower-left corner of the video frame

These dimensions and typographical elements were consistent with Zoom's default layout, while other platforms displayed notable deviations in overlay opacity, position, and text rendering.

Debounce and behavioral logic

Short speech bursts below 200 ms failed to trigger a screen switch in the forensic recording, confirming the presence of a debounce window of 200–300 ms—a behavior consistent with Zoom's documented latency smoothing and "hangover" logic [4]. By contrast, Webex and Zoho Meeting exhibited intermittent switching on brief utterances, leading to visual instability.

Producer and host exception handling

During one segment of the examined video, the participant who initiated the session likely acting as the host spoke audibly but did not appear as the active speaker on screen. Instead, the prior speaker's tile remained visible while audio transitioned to the host's channel. This behavior aligns with Zoom's host-exception protocol, where visual switching is suppressed unless the "Show myself when speaking" option is enabled [11]. None of the other tested platforms demonstrated this same condition during controlled replications.

Temporal stability under network variability

To assess robustness under network fluctuation, simulations were performed with ± 40 ms network jitter.

Zoom Workplace maintained consistent switching latency (Δ<25 ms), while Webex and Google Meet exhibited elevated jitter response, occasionally desynchronizing audio and visual events.

These findings indicate that Zoom's active speaker switching employs adaptive buffering to maintain stability, a feature that corresponds closely to the behavior observed in the forensic evidence.

Forensic summary

Based on combined quantitative and qualitative parameters, the forensic video is most closely aligned with Zoom Workplace's behavioral profile. The correlation was observed across latency, debounce, UI design, and network stability metrics (Table 2).

Parameter	Forensic sample	Closest platform match	Correlation
Mean switching latency	287 ms	Zoom Workplace	Strong
Debounce threshold	200–300 ms	Zoom Workplace	Strong
UI layout and text style	Semi-transparent, lower-left	Zoom Workplace	Strong
Host exception handling	Present	Zoom Workplace	Strong
Network stability	High	Zoom Workplace	Strong

Table 2. Forensic analytic metrics on Zoom tool video examination.

Conclusion

This study demonstrates that measurable behavioral and interface characteristics within video conferencing recordings can serve as powerful forensic indicators when traditional metadata is absent. Through frame-by-frame video analysis and controlled replication, the investigation established that the examined 18-minute recording, based on a crime case received at the Central Forensic Laboratory, exhibited visual and temporal behaviors consistent with Zoom Workplace's active speaker switching logic.

Key measurable parameters such as average speech-to-screen latency (287 ms), debounce thresholds (200–300 ms), and unique host exception behavior—collectively supported attribution to this platform with high confidence. The observed consistency across independent replication sessions reinforces the reliability of these findings and demonstrates that automated features like Voice Activity Detection (VAD) and Active Speaker Recognition (ASR) can leave behind identifiable forensic signatures.

This work provides an applied framework for digital forensic video examination, enabling experts to infer the likely source platform of multimedia evidence based solely on audiovisual behavior. Such metadata-independent attribution not only enhances authenticity verification but also strengthens the chain of custody and courtroom admissibility of digital video evidence.

By bridging the gap between forensic science and intelligent communication technologies, this study supports the growing need for validated methodologies in digital forensic investigation helping practitioners and law enforcement agencies navigate complex evidence environments where traditional identifiers are obscured or removed.

Funding

This research received no external funding.

This work was supported by Sh. Ashutosh Deo Tiwari, Deputy Director, Photo and Scientific Aids, CFSL, Delhi.

References

1. Lillis, David, Brett Becker, Tadhg O'Sullivan, and Mark, Scanlon. "Current challenges and future research areas for digital forensic investigation." *Digital Investigation* 19 (2016): S38–S49.
2. Garfinkel, SL. "Digital forensics research: The next 10 years." *Digital Investigation* 7 (2010): S64-S73.
3. Li, Dongguang. "Ballistics image processing and analysis for firearm identification." *Image Processing* (2009): 141-174.
4. Huang, J, Li D, and Chen Q. "A multimodal active speaker detection system for video conferencing." *IEEE Access* 7 (2019): 167982–167992.
5. Afouras, Triantafyllos, Joon Son Chung, Andrew Senior, and Oriol Vinyals, et al. "Deep audio-visual speech recognition." *IEEE Trans Pattern Anal Mach Intell* 44 (2018): 8717-8727.
6. Yadav, R, and Sahu S. "Voice activity detection: A review of techniques and applications." *J Acoustic Soc India* 49 (2022): 12–23.
7. Casey, E. "Digital evidence and computer crime: Forensic science, computers, and the internet." Academic Press. (2019).
8. Kaur, P, and Kaur R. "Digital video forensics: Techniques and challenges." *Forensic Sci Int Rep* 3 (2021): 100197.
9. Huo, J, Li R, and Zhang T. "Detection of active speaker switching in video conferencing environments." *IEEE Access* 9 (2021): 43281–43293.
10. Ramırez, Javier, José C. Segura, CarmenBenıtez, and Angel de La Torre, et al. "Efficient voice activity detection algorithms using long-term speech information." *Speech Commun* 42 (2004): 271-287.
11. Zhang, T, Liu, R, and Wang Z. "Adaptive speaker tracking for virtual conferencing applications." *J Vis Commun Image Represent* 78 (2021): 103206.

How to cite this article: Tiwari, Ashutosh Deo, Thanuja Durgam, Sunil Kumar, and Sunand Bishnoie. "Digital Forensic Investigation of Video Evidence Using Voice Activity Detection Algorithm." *J Forensic Res* 16 (2025): 657.