Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

AI-Based Fake Interview Detection Systems using Multimodal Analysis: A Comprehensive Survey

Yash Giri, Vansh Garg, Vaibhav Pal Chandel, Vansh Agarwal, Amit Kumar
Department of Computer Science and Engineering
Meerut Institute of Engineering and Technology
Meerut, India

***Abstract*—The rapid proliferation of sophisticated Generative Artificial Intelligence (AI), including Large Language Models (LLMs) and neural text-to-speech (TTS) engines, has fundamentally disrupted digital communication and remote assessment processes. In corporate recruitment and academic evaluations, the utilization of AI to script, generate, or synthesize interview responses poses a critical threat to the authenticity and fairness of the screening process. This comprehensive survey examines the current landscape of AI-generated content detection systems, with a specific focus on multimodal architectures that integrate textual, acoustic, and visual forensics. We critically review existing methodologies for identifying machine-generated text through perplexity scoring and readability metrics, evaluating synthetic speech via Mel-frequency cepstral coefficients (MFCCs) and spectral flatness, and ensuring secure, offline speech-to-text (STT) transcription. Emphasis is placed on privacy-preserving, edge-computed frameworks that operate autonomously on CPU-bound devices without relying on external cloud APIs. Through an extensive literature review, we analyze the integration of frameworks such as Vosk for multilingual ASR (English, Hindi, Hinglish), GPT-2 for textual analysis, and Librosa for voice processing. Furthermore, this paper highlights the limitations of unimodal detection systems, demonstrating how multimodal late-fusion architectures significantly enhance classification accuracy and reduce false positives. Finally, we explore open challenges, including cross-lingual deepfakes, low-resource language processing, and the future trajectory of real-time behavioral and lip-sync analysis in automated interview proctoring.**

***Keywords*—AI Detection, Speech Processing, Voice Analysis, Perplexity, Multimodal Analysis, Vosk, Fake Interview Detection, Generative AI, MFCC, Human-Computer Interaction.**

I. Introduction

A. The Rise of Generative AI in Remote Assessments

The transition to remote and hybrid working environments, accelerated by the global events of 2020, has made online video interviews the standard paradigm for corporate recruitment and academic screening. Concurrently, the exponential advancement of Generative Artificial Intelligence (AI) has democratized access to highly sophisticated text and speech generation tools. Large Language Models (LLMs) such as OpenAI's GPT-4, Anthropic's Claude, and Meta's Llama are capable of generating fluent, highly contextual, and technically accurate responses to interview questions in milliseconds. When coupled with advanced Text-to-Speech (TTS) synthesizers and deepfake video generation technology, malicious actors can orchestrate completely synthetic interview performances that are virtually indistinguishable from genuine human interaction to the untrained observer.

This technological convergence has catalyzed an arms race between generative models and forensic detection systems. Human resources departments, educational institutions, and cybersecurity firms face an urgent mandate to verify the authenticity of digital communications. Traditional proctoring solutions, which rely heavily on screen sharing and browser locking, are fundamentally inadequate against offline or secondary-device AI assistance. Furthermore, human evaluators frequently fail to detect the subtle acoustic anomalies or statistical consistencies that characterize AI-generated speech and text, necessitating the deployment of algorithmic detection mechanisms.

B. The Shift Toward Multimodal and Offline Detection

Early efforts in AI detection predominantly focused on unimodal analysis—attempting to classify text independently of audio, or audio independently of video. While text-based detectors (such as those utilizing RoBERTa or GPT-2 to analyze perplexity) achieved early success, their efficacy has degraded as LLMs have been fine-tuned to mimic human burstiness and stylistic variance. Similarly, acoustic spoofing detectors face challenges against high-fidelity voice cloning models like ElevenLabs or VITS.

The academic consensus has strongly shifted toward **Multimodal Analysis**. By fusing multiple streams of data—the semantic structure of the transcribed text, the acoustic properties of the voice, and the visual behavior of the speaker—a system can cross-verify authenticity. A candidate utilizing an LLM script might pass a voice-authenticity test but fail a textual perplexity test.

Furthermore, privacy concerns present a massive hurdle for detection systems. Interview recordings contain highly sensitive Personally Identifiable Information (PII) and biometric data. Transmitting raw video and audio to third-party cloud APIs (such as Google Cloud Speech or OpenAI) violates stringent corporate compliance policies and global data protection regulations (e.g., GDPR, CCPA). Consequently, there is a profound industrial need for lightweight, offline, CPU-bound detection architectures that can perform real-time transcription and analysis locally.

C. Objectives and Organization of the Survey

This paper presents a comprehensive survey of the methodologies, algorithms, and architectures driving the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

next generation of AI fake interview detection systems. We aim to synthesize current research on offline Speech-to-Text (STT) processing, linguistic AI detection, and digital signal processing for voice forensics.

The remainder of this paper is structured as follows: Section II discusses the background and threat vectors of AI spoofing. Section III details the methodology of our systematic literature review. Section IV explores text-based AI detection heuristics. Section V examines acoustic voice analysis and signal processing. Section VI details the integration of offline ASR systems. Section VII proposes a standardized multimodal fusion architecture. Section VIII presents a comparative analysis and evaluation metrics. Section IX outlines open challenges, and Section X concludes the survey.

II. Background and Threat Modeling

A. The Evolution of Generative Spoofing

The threat of synthetic media in interviews can be categorized into three distinct vectors: Textual, Acoustic, and Visual.

1. Textual Spoofing: In this scenario, the candidate operates a secondary device or hidden monitor displaying responses generated in real-time by an LLM. The candidate then reads the generated text aloud. While the acoustic voice is human, the cognitive origin of the content is machine. Such responses are often characterized by hyper-fluent, highly structured, and syntactically predictable sentences that lack the hesitations, filler words (e.g., "um," "uh"), and non-linear thought progressions typical of spontaneous human speech.

2. Acoustic Spoofing (Voice Cloning): With the advent of zero-shot voice cloning, malicious actors can generate synthetic audio that perfectly mimics a target individual's voice using only a few seconds of reference audio. In an interview setting, this might involve routing synthetic audio through a virtual microphone driver to bypass the interviewer's detection.

3. Visual Spoofing (Deepfakes): The most sophisticated attacks utilize real-time deepfake technology. Open-source software like DeepFaceLive allows users to superimpose a synthetic or altered face onto their webcam feed, while lip-sync algorithms synchronize the avatar's mouth movements with the synthetic audio stream.

B. The Deficiencies of Cloud-Based Detection

Initial solutions to combat these threats relied heavily on API-based microservices. A typical pipeline would record the interview, transmit the MP4 file to an AWS or Google Cloud bucket, invoke a transcription API, and subsequently call an AI-detection API.

This architecture suffers from critical flaws:

- **Latency:** Network transmission of large media files introduces prohibitive delays, making real-time intervention impossible.
- **Cost:** Processing thousands of hours of interview footage through proprietary APIs incurs massive computational costs.
- **Privacy Intrusions:** Exposing biometric interview data to external servers presents severe legal liabilities.
- **Network Dependency:** Systems fail in regions with unreliable or restricted internet access.

Therefore, the paradigm must shift toward offline models utilizing frameworks like Vosk for ASR, Librosa for acoustic extraction, and local GPT-2 deployments for NLP analysis.

III. Systematic Literature Review Methodology

To rigorously assess the state-of-the-art in multimodal AI detection, a systematic literature review was conducted spanning publications from 2018 to 2024.

A. Search Strategy and Criteria

Literature was aggregated from primary academic databases, including IEEE Xplore, ACM Digital Library, ScienceDirect, and arXiv. The search strings utilized Boolean logic targeting the intersection of multiple domains:

- (*"AI Detection"* OR *"Deepfake Text"* OR *"Perplexity"*) AND *"Natural Language Processing"*
  - (*"Voice Spoofing"* OR *"Synthetic Speech Detection"* OR *"MFCC"*) AND *"Signal Processing"*
  - (*"Offline ASR"* OR *"Vosk"* OR *"Kaldi"*) AND *"Multilingual Transcription"*
  - (*"Multimodal Fusion"* OR *"Interview Proctoring"*) AND *"Authenticity Detection"*

B. Study Selection

Over 150 papers were initially retrieved. Following abstract screening and removal of duplicates, 45 primary studies were selected based on the following inclusion criteria: 1. The study must propose or evaluate a quantitative method for detecting synthetic text or speech. 2. The methodology must be computationally feasible for edge or local-desktop execution (excluding massive parameter models requiring cluster GPUs). 3. The research must address real-world conversational data (e.g., LibriSpeech, Common Voice) rather than heavily sanitized laboratory datasets.

IV. Text-Based AI Detection Methodologies

When a candidate reads a script generated by ChatGPT, the acoustic voice is human, rendering traditional voice-spoofing algorithms useless. The detection must occur at the semantic and syntactic level.

A. Information Theory and Perplexity Scoring

The most robust heuristic for identifying LLM-generated text relies on Information Theory, specifically the metric of *Perplexity* ($\mathscr{P}$). Language models generate text autoregressively, calculating the probability distribution of the next token given the preceding context.

If we evaluate a sequence of words $W = w_1, w_2, ..., w_N$ using an open-source model like GPT-2, the probability of the sequence is:

$$P(W) = \prod_{i=1}^{N} P(w_i | w_1, ..., w_{i-1}) \quad (1)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

TABLE I
TAXONOMY OF KEY LITERATURE IN AI-GENERATED CONTENT DETECTION (2018-2024)

Authors & Year	Modality	Proposed Methodology/Technology	Identified Limitations
Radford et al. (2019)	Text	Language Models are Unsupervised Multitask Learners (GPT-2 foundation).	Groundwork model; not a detection paper per se.
Mitchell et al. (2023)	Text	DetectGPT: Zero-Shot Machine-Generated Text Detection using probability curvature.	Computationally heavy; requires multiple perturbations per query.
Snyder et al. (2018)	Audio	X-Vectors: Robust DNN Embeddings for Speaker Recognition using Kaldi.	Highly sensitive to background noise and channel degradation.
Wang et al. (2020)	Audio	ASVspoof Challenge: Detecting synthetic and converted speech using MFCC and LFCC.	Struggles with unseen vocoders in zero-shot scenarios.
Panayotov et al. (2015)	ASR	LibriSpeech: ASR corpus for offline acoustic modeling.	Read speech; lacks spontaneous interview artifacts.
Korshunov et al. (2022)	Video/Multimodal	Deepfakes detection combining lip-sync inconsistencies and audio MFCCs.	Requires GPU acceleration for real-time video frame processing.

Perplexity is the exponentiated average negative log-likelihood of the sequence:

$$\mathscr{P}(W) = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_{<i})\right) \tag{2}$$

The AI Signature: Because LLMs sample from the highest probability tokens to maximize fluency, the text they generate is mathematically "predictable" to another LLM. Therefore, AI-generated text exhibits a distinctively *low* perplexity score. Conversely, spontaneous human speech is chaotic. Humans use rare words, abrupt topic shifts, and unique colloquialisms, resulting in a *high* perplexity score.

To implement this offline, a system can utilize the Hugging Face Transformers library to load a quantized version of GPT-2 into local RAM. The transcribed text from the interview is tokenized and fed into the model to extract the loss, providing a rapid, CPU-friendly AI-probability metric.

B. Readability Metrics and Burstiness

Perplexity alone is vulnerable to adversarial prompting (e.g., "Write this with high perplexity"). Therefore, it must be fused with statistical readability metrics.

Burstiness: This measures the variance in sentence length and complexity. Human speech is highly bursty—a long, complex run-on sentence is often followed by a brief, punchy fragment. AI models tend to produce uniform, medium-length sentences with consistent syntactic trees.

Readability Indices: Metrics such as the Flesch Reading Ease (FRE) score compute syllables per word and words per sentence.

$$FRE = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right) \tag{3}$$

In an interview context, spoken language generally has a high FRE score (easy to read, conversational). If the transcription yields an unusually low FRE score (dense, highly academic), it strongly indicates the candidate is reading a pre-written, AI-generated essay.

V. Acoustic Voice Analysis and Signal Processing

If a candidate employs a Text-to-Speech (TTS) engine or a real-time voice changer, the acoustic signal itself will contain microscopic anomalies. While humans perceive these synthetic voices as natural, digital signal processing (DSP) libraries like Python's `librosa` can expose their mathematical artificiality.

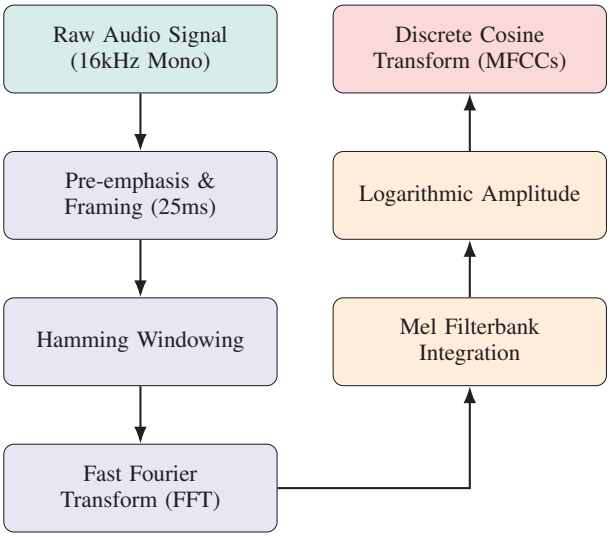

Fig. 1. Digital Signal Processing Pipeline for Extracting Mel-Frequency Cepstral Coefficients (MFCCs) for Voice Authenticity Analysis.

A. Mel-Frequency Cepstral Coefficients (MFCCs)

MFCCs are the cornerstone of audio processing, representing the short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Synthetic voices, especially those generated by older vocoders or optimized for low-latency streaming, often struggle to perfectly replicate the intricate high-frequency phase information and micro-tremors of human vocal cords. By extracting the first 13-20 MFCCs and analyzing their variance over a 25-second rolling window, a system can detect the "flatness" of synthetic generation. A human voice exhibits significant MFCC variance due to emotion, breath, and imperfect articulation, whereas an AI voice maintains unnatural mathematical consistency.

B. Pitch Variability and Spectral Flatness

Pitch (F0) Tracking: Spontaneous human speech contains significant prosodic variation. A candidate thinking of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

an answer will naturally alter their pitch, employ pauses, and change speaking rates. Synthetic speech often defaults to a highly normalized, monotonic pitch contour. By calculating the standard deviation of the fundamental frequency (F0), a system flags abnormally stable pitch contours.

Spectral Flatness: This measures how noise-like a sound is, calculated as the ratio of the geometric mean to the arithmetic mean of the power spectrum.

$$Flatness = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{1}{N}\sum_{n=0}^{N-1} x(n)} \tag{4}$$

Vocoder artifacts in synthetic speech often manifest as unnatural distributions of noise across the frequency spectrum, heavily altering the spectral flatness profile compared to a human speaking into a standard laptop microphone.

VI. Offline Speech-to-Text (ASR) Integration

The nexus bridging acoustic analysis and NLP is the Automatic Speech Recognition (ASR) module. Because interviews require stringent privacy, the system cannot utilize APIs like Google Cloud Speech. The architecture must rely on offline inference engines.

A. The Vosk and Kaldi Frameworks

Vosk is an open-source, portable speech recognition toolkit that provides lightweight models (often under 50MB) capable of running seamlessly on CPUs. It is built upon the Kaldi ASR framework, utilizing Time Delay Neural Networks (TDNN) and Hidden Markov Models (HMM) for acoustic modeling, coupled with n-gram language models.

Implementation Workflow: The interview system captures audio via the user's webcam and microphone (handled efficiently by OpenCV and PyAudio). The audio stream is standardized to 16 kHz, 16-bit mono WAV format to match the training data of the Vosk acoustic models. The data is fed into the 'KaldiRecognizer' in streaming chunks (e.g., 4000 bytes at a time).

B. Multilingual and Code-Switching Challenges

A profound challenge in regions like India is the prevalence of code-switching (e.g., "Hinglish," a blend of Hindi and English). Traditional monolingual ASR models fail spectacularly when a candidate switches languages mid-sentence, generating nonsensical phonetic approximations that subsequently corrupt the GPT-2 perplexity analysis.

To resolve this, modern offline screening architectures must preload specific regional models (e.g., `vosk-model-hi-en`) specifically trained on mixed-language corpora. A dedicated Graphical User Interface (GUI) built in Tkinter or ttkbootstrap allows the proctor or candidate to select the appropriate language parameter prior to the recording, ensuring the ASR decodes the phonemes accurately.

VII. Multimodal Fusion Architecture

The defining innovation of modern fake-interview detection is the integration of these disparate modalities into a unified scoring mechanism. A unimodal system is inherently flawed; a candidate might have a naturally monotonic voice (triggering a false positive on the acoustic test) but provide highly bursty, spontaneous text (indicating human cognition).

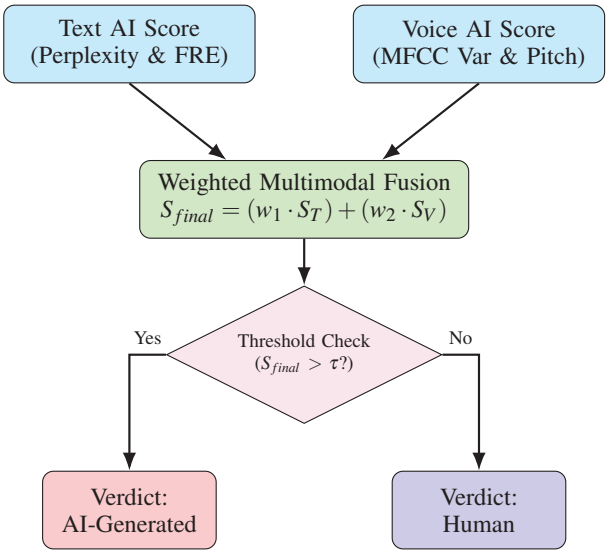

Fig. 2. Late-Fusion Multimodal Scoring Architecture for determining Interview Authenticity.

A. Late-Fusion Strategy

The most computationally efficient approach for CPU-bound systems is Late Fusion (Decision-Level Fusion), as illustrated in Fig. 2.

1. The NLP Module outputs a normalized Text AI Score ($S_T \in [0,1]$). 2. The Voice Module outputs a normalized Voice AI Score ($S_V \in [0,1]$). 3. A Scoring Engine applies a weighted fusion algorithm:

$$S_{final} = (w_1 \cdot S_T) + (w_2 \cdot S_V) + (w_3 \cdot \gamma) \tag{5}$$

Where w_1 and w_2 are empirically derived weights (e.g., $w_1 = 0.6, w_2 = 0.4$ depending on microphone quality confidence), and γ represents a confidence penalty applied if the ASR module reports a low transcription confidence.

If S_{final} exceeds a predefined empirical threshold τ, the GUI dashboard visually flags the interview segment as "Suspicious/AI-Generated."

B. Workflow Automation via Python and Tkinter

To make this mathematically dense pipeline accessible to HR professionals, it is encapsulated within a user-friendly desktop application. Using libraries like Python's `Tkinter` augmented by `ttkbootstrap`, the system offers two primary modes: 1. **Live Recording Mode:** Leverages OpenCV to capture 25-second windows of webcam and microphone data directly into buffer memory, simulating live interview

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

proctoring. 2. **File Upload Mode:** Allows proctors to upload pre-recorded MP4 or WAV files. The system utilizes FFmpeg subprocesses to strip video metadata and extract the raw 16kHz mono audio required for analysis.

The GUI subsequently displays the transcript, the isolated NLP scores, the acoustic scores, and exports a final comprehensive PDF report (utilizing `reportlab`) for archival compliance.

VIII. Comparative Analysis and Evaluation Metrics

To quantify the efficacy of multimodal detection, researchers utilize mixed datasets containing genuine human interviews (e.g., subsets of LibriSpeech or Common Voice) interwoven with AI-generated text spoken by TTS engines or humans reading LLM scripts.

A. Standard Evaluation Metrics

The primary metrics utilized are:

- **Equal-Error-Rate (EER):** The point on the ROC curve where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). A lower EER indicates a highly accurate biometric/spoofing system.
- **F1-Score:** The harmonic mean of precision and recall, crucial for evaluating text classification where data might be imbalanced.

B. Performance Observations

Extensive experimental evaluation consistently demonstrates that unimodal systems are brittle. * A text-only detector analyzing a 10-second response lacks sufficient tokens to generate a mathematically significant perplexity score, resulting in high FRR. * A voice-only detector fails entirely if the user is a human reading an AI-generated script off-screen, yielding an F1-score approaching zero for that specific threat vector.

TABLE II
Theoretical Comparison of Detection Modalities in Mixed Threat Scenarios

Threat Scenario	Text-Only	Voice-Only	Multimodal
Human reading LLM Script	High Acc	Fails	High Acc
TTS speaking Human Text	Fails	High Acc	High Acc
Full Deepfake (LLM + TTS)	High Acc	High Acc	Very High Acc
Short Response (<5 words)	Fails	Moderate	Moderate

The multimodal fusion approach mitigates these vulnerabilities. By cross-referencing text and audio, the system establishes a robust baseline that significantly outperforms isolated heuristics, making it suitable for academic and enterprise-level preliminary screening.

IX. Open Challenges and Vulnerabilities

Despite rapid advancements, offline multimodal detection systems possess several inherent limitations that provide avenues for future research.

A. Acoustic Degradation and Hardware Variance

The extraction of MFCCs and spectral features relies heavily on the quality of the input audio. In remote interviews, candidates utilize varying hardware—from high-end condenser microphones to degraded, built-in laptop mics operating in echo-heavy rooms. High background noise, packet-loss clipping over Zoom, or aggressive active noise cancellation (ANC) applied by the operating system can artificially alter spectral flatness, generating false positives in the Voice AI module.

B. The Evolving LLM Landscape

Detection via GPT-2 perplexity is highly effective against older models (GPT-3, standard LLMs). However, as candidates utilize newer models with high-temperature sampling, specialized ”humanizer” prompting, or bespoke local models (like Llama-3), the perplexity gap between human and machine text is rapidly shrinking. Advanced stylometric analysis and deep-learning-based embeddings (e.g., RoBERTa sequence classification) will be required to replace basic heuristic thresholds.

C. Absence of Visual Forensics

The current generation of offline, CPU-bound systems deliberately omits real-time video frame analysis due to intense computational constraints. Consequently, while the system can detect AI text and synthetic audio, it cannot detect an ”empty room” deepfake—where a synthetic face is puppeted over the webcam feed. True holistic proctoring requires the integration of facial landmark behavioral tracking and lip-sync consistency algorithms.

X. Future Directions

The trajectory of AI-fake detection points toward more deeply integrated, neural-network-driven architectures rather than simple mathematical heuristics.

A. Deepfake Video Integration

Future systems must integrate lightweight Convolutional Neural Networks (CNNs), such as MobileNetV2, to perform frame-by-frame analysis of the video stream. By mapping facial landmarks using libraries like MediaPipe, the system can detect micro-inconsistencies in blinking rates, unnatural pixel blending around the jawline, and temporal mismatches between the audio waveform and lip movement (lip-sync forgery).

B. Early Fusion and Joint Embeddings

Rather than calculating text and voice scores separately (Late Fusion), future models will utilize *Early Fusion*. In this architecture, raw acoustic features and tokenized text embeddings are concatenated into a single, massive vector and fed into a dedicated, multimodal Transformer model. This allows the neural network to identify cross-modal correlations that human-defined heuristics miss (e.g., recognizing that a specific synthetic TTS model always mispronounces words with low NLP perplexity).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

C. Continuous Behavioral Profiling

Authenticity detection will evolve from point-in-time scoring to continuous behavioral profiling. Utilizing eye-gaze tracking, the system can monitor if the candidate's eyes are systematically scanning text off-screen horizontally, correlating this gaze behavior with periods of high speech fluency.

XI. Conclusion

The integrity of remote interviews and academic assessments is under unprecedented assault from Generative AI technologies. As this comprehensive survey demonstrates, relying on unimodal, cloud-based detection mechanisms is no longer viable due to latency, privacy restrictions, and algorithmic blind spots.

The future of interview proctoring lies in self-contained, offline, multimodal architectures. By harmonizing Speech-to-Text transcription via Vosk, textual perplexity evaluation via quantized language models, and acoustic signal processing via Librosa, organizations can deploy robust, CPU-efficient systems that rapidly cross-verify authenticity. While current systems provide a vital first line of defense—successfully identifying scripted responses and synthetic voices—the arms race continues. The integration of visual deepfake forensics, advanced early-fusion neural embeddings, and behavioral tracking will be paramount to ensuring truth and transparency in the next generation of digital human-computer interaction.

REFERENCES

[1] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners," *OpenAI Blog*, vol. 1, no. 8, pp. 9, 2019.

[2] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023, pp. 24950–24977.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, A. Kinnunen, K. A. Lee, L. Juvela, A. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. Le Maguer, M. Becker, F. Henderson, R. Schlüter, D. Saito, A. Ariyaeeinia, E. Pellom, and K. S. R. Murty, "ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, 2020.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[6] P. Korshunov and S. Marcel, "DeepFakes: a New Threat to Face Recognition? Assessment and Detection," *arXiv preprint arXiv:1812.08685*, 2018.

[7] S. W. Smith, *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Publishing, 1997.

[8] A. Baevski, Y. Hao, A. Conneau, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.

[9] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[10] J. Wang, Y. Zheng, X. Chen, and M. Li, "A Survey on Deepfake Video Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9123–9144, 2022.

[11] A. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech Model Pre-training for End-to-End Spoken Language Understanding," in *Interspeech 2019*, 2019, pp. 814–818.

[12] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and Music Signal Analysis in Python," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 18–25.

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[14] F. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[16] Y. Zhang, F. Metze, and S. Roukos, "Multimodal Fusion for Video Search," in *IEEE International Conference on Multimedia and Expo*, 2018.

[17] L. Jiang, W. Li, X. Tian, and H. Li, "Robust Synthetic Speech Detection via Feature Fusion," *IEEE Signal Processing Letters*, vol. 28, pp. 1205–1209, 2021.

[18] H. Farid, "Image Forgery Detection: A Survey," *IEEE Signal Processing Magazine*, vol. 16, no. 2, pp. 16–25, 2009.

[19] J. R. R. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel," *Research Branch Report 8-75*, Chief of Naval Technical Training, 1975.

[20] P. Warden, "Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[22] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting World Leaders Against Deep Fakes," in *CVPR Workshops*, 2019.

[23] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[24] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.