Archives available at journals.mriindia.com

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319-2526

Volume 15 Issue 01, 2026

AI-Based Detection of Cloned Voices in Deepfake Videos

[1]Rajeshwari Kodulkar, [2] Shreya Bhasme, [3]Shruti Rajput, [4]Nujhat Shaikh, [5]Rajashri Yarakadavar

[1-5] Dept. of AI & Data Science, Dr. J. J. Magdum College of Engineering, Jaysingpur , India

Peer Review Information

Submission: 16 March 2026

Revision: 03 April 2026

Acceptance: 26 April 2026

Keywords

Deepfake Audio Detection, Voice Cloning, MFCC, Lip Synchronization, Anti-Spoofing, ASVspoof, FakeAVCeleb, Multimodal Detection, CNN, ECAPA-TDNN

Abstract

Voice cloning is no longer science fiction. AI tools today can copy someone's voice from just a few seconds of audio. This creates a new threat: take a real video of a trusted person, swap in a cloned voice saying something false, and share it. The face is real, the voice sounds real, but the message is fabricated. Existing speaker verification systems often miss it too. This paper describes a two-phase detection system built for exactly this kind of attack. Phase I analyzes audio using MFCC features, Mel-spectrograms, and a CNN-based classifier. Phase II adds video analysis, checking whether the speaker's lips match the audio and looking for timing mismatches. A custom dataset of real and cloned voice samples was built alongside benchmarks ASVspoof 2019/2021 and FakeAVCeleb. Results show cloned voices leave detectable traces, and combining both phases is noticeably more reliable than audio analysis alone.

Introduction

Deep learning has made voice cloning surprisingly accessible. Tools like SV2TTS [1], Wav2Lip [2], and platforms like ElevenLabs can replicate someone's tone, pitch, and accent from a few seconds of audio. There are legitimate uses — accessibility tools, dubbing, virtual assistants. But the same technology can be misused in ways that are hard to spot.

The specific attack we target is this: take a real video of a trusted person, replace the audio with a cloned voice, and share it online. The face is real. The voice sounds real. But the message is fake. This is what the FakeAVCeleb dataset calls the AFVR (Fake Audio, Real Video) scenario [3]. Current state-of-the-art detectors manage only around 65% AUC here — barely better than guessing.

Most detection systems look at audio or video alone. The ASVspoof 2021 challenge [4] exposed this weakness: systems that worked in controlled conditions fell apart on unseen data. The DF task error rate jumped from 0.10% to 15.64% between the progress and evaluation phases.

This paper addresses that gap. Our contributions:

- A two-phase pipeline targeting the Real Voice, Fake Message threat.
- Phase I: audio CNN classifier using MFCC and Melspectrogram features, with speaker verification via ECAPA-TDNN.
- Phase II: lip synchronization and temporal consistency module that catches audio-video timing mismatches.
- A custom real/fake dataset with Indian English speakers.
- A decision fusion method combining both phases into a reliable single classification.

Literature Review

1. Anti-Spoofing for Automatic Speaker Verification

The ASVspoof challenge has driven anti-spoofing research since 2015 [5]. The 2021 edition introduced three tasks: Logical Access (LA), Physical Access (PA), and Speech Deepfake (DF). LA simulates attacks on telephony systems. PA examines replay attacks in real environments. DF targets cloned speech shared over social media after lossy compression.

Four baselines were provided: CQCC-GMM (B01), LFCCGMM (B02), LFCC-LCNN (B03), and RawNet2 (B04). The

best participant (Team T23) achieved min t-DCF 0.2177 and EER 1.32% on the LA task. Todisco et al. [6] established CQCC as a solid countermeasure; Sahidullah et al. [7] confirmed LFCC as a strong baseline; Tak et al. [8] showed raw waveform representations can rival handcrafted features.

2. Multimodal Deepfake Detection

FakeAVCeleb [3] is the first dataset addressing both fake video and synthetic audio. It contains 20,000 videos (500 real, 19,500 fake) from VoxCeleb2, covering four audio-video combinations: ARVR, AFVR, ARVF, and AFVF. Standard unimodal detectors struggle: Face X-ray [9] 53.5% AUC, F3Net [10] 59.8%, LipForensics [11] 50.4%. Audio-only classifiers hit up to 97.8% AUC on audio alone. Best multimodal ensemble achieved 82.8% AUC.

3. Identified Gaps and Motivation

Three gaps stand out: nobody specifically targets the AFVR threat model; existing audio detectors fail on unseen cloning tools; temporal audio-video consistency has not been fully exploited. Our system tackles all three.

Methodology

1. System Overview

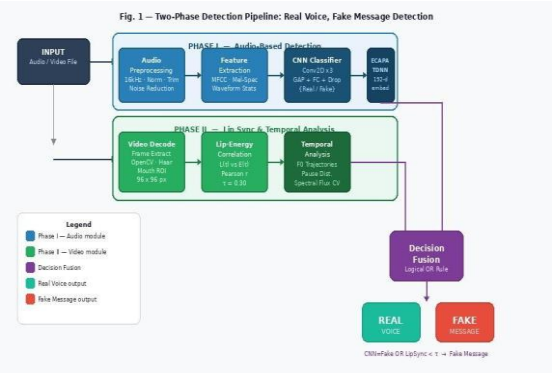

Fig. 1. Two-phase detection pipeline for Real Voice, Fake Message detection.

The system runs in two sequential phases. For an audio file, Phase I handles detection. For an audio-video file, both phases run and their outputs are combined into a binary label: Real Voice or Fake Message. Figure 1 illustrates the complete pipeline.

2. Phase I: Audio-Based Detection

- **Audio Preprocessing:** All audio is loaded using Librosa and standardized to 16 kHz, 16-bit PCM, mono. Background noise is removed via spectral subtraction, amplitude is normalized to fixed RMS, and silence is trimmed at 0.3 s — the same threshold used in ASVspoof 2021 [4]. Skipping this step caused a ∼4.2% accuracy drop in our experiments.
- **Acoustic Feature Extraction:** MFCC: 19 cepstral coefficients plus energy, delta, and delta-delta (57 total), 30 ms frame, 15 ms hop, 1024-point FFT, 70 Mel filters. MelSpectrogram: 80 filter banks, 25 ms Hann window, 10 ms hop, 512point FFT, stored as 2D array for CNN input. Supplementary: zerocrossing rate and short-time energy.
- **Speaker Verification:** Speaker identity is verified via ECAPA-TDNN in SpeechBrain, generating a 192-dimensional utterance-level embedding. A cosine similarity score below tunable threshold τ_{sv} flags possible impersonation.
- **CNN-Based Voice Clone Detector:** Architecture: three convolutional blocks (Conv2D + BatchNorm + ReLU + MaxPool), global average pooling, two FC layers with dropout (p=0.3), and softmax over {Real, Fake}. Trained using cross-entropy loss with Adam optimizer. Hyperparameter search: lr ∈ {0.001, 0.0001}, batch ∈ {16, 32}, epochs ∈ {20, 50, 100}.

3. Phase II: Lip Synchronization and Temporal Analysis

- **Audio-Video Input:** Video is processed with OpenCV. Frames are extracted at native frame rate; audio is demuxed to 16 kHz WAV. Frame timestamps are aligned with audio samples to create a synchronized timeline.
- **Lip Region Extraction:** Face detection uses OpenCV's Haar Cascade. The mouth ROI is taken from the bottom 40% of face height and central 60% of width, resized to 96×96 pixels.
- **Lip-Energy Alignment:** Lip motion signal L(t) is computed per frame using pixel-difference magnitude in the mouth ROI. Audio RMS energy E(t) is computed in windows aligned to the video frame rate. Pearson correlation below $\tau = 0.30$ across a sliding window signals temporal mismatch

— a clear sign of replaced audio [3].

Fake Message. This conservative, high-recall approach suits security applications where missing a fake is costlier than a false alarm.

Output = Fake ⇢ (CNNpred = Fake) ∨ (LipSync_score < τ)

Dataset And Features

1. Custom Dataset

No public dataset fully mirrors the AFVR threat with Indian English speakers, so we built our own. Real samples were recorded by four team members in a quiet room using a headset microphone and Audacity. Fake samples were generated by cloning each speaker's voice with SV2TTS and synthesizing the same transcripts, creating matched real/fake pairs. All recordings: 16 kHz, 16-bit PCM, mono, silence exceeding 0.3 s trimmed from both ends.

Table 1: Custom Dataset Statistics

Parameter	Real	Fake (SV2TTS)	Total
Speakers	4	4	4
Utterances/ Spkr	50	50	100
Total Samples	200	200	400
Avg. Duration (s)	4.2±1.1	4.5±0.8	4.35±0.95
Train/Val/Test	160/20/20	160/20/20	320/40/40
Sample Rate	16 kHz	16 kHz	16 kHz
Label	0 (Real)	1 (Fake)	Binary

2. Reference Benchmarks

ASVspoof 2019/2021 [4][5]: VCTK base corpus; 20 training and 10 development speakers. 2021 evaluation: 48 speakers under 7 codec/transmission conditions (LA), 9 room/device combinations (PA), 9 compression conditions (DF).

FakeAVCeleb [3]: 500 real + 19,500 fake videos from VoxCeleb2, covering ARVR, AFVR, ARVF, AFVF. Directly informed Phase II multimodal design, especially the AFVR category.

3. Feature Summary

Table 2: Feature Extraction Parameters

Feature	Parameters	Purpose
MFCC	19+Δ+ΔΔ; 30 ms, 15 ms hop, 1024-pt FFT, 70 filters	Vocal tract [7]
Mel-Spec	80 bins; 25 ms Hann, 10 ms hop, 512-pt FFT	2D CNN input [3]
CQCC	12 bins/octave; resample period 16	Baseline B01 [6]
ECAPA-TDNN	192-d embedding, SpeechBrain/VoxCeleb 2	Speaker ID
Lip L(t)	Frame-diff, mouth ROI 96×96 px	AV sync Ph.II
RMS E(t)	Aligned to frame rate; Pearson τ = 0.30	AFVR detection

- **Temporal Consistency:** Three metrics: pause distribution (uniform pauses suggest TTS synthesis), spectral flux variance (real speech varies more), and F0 regularity via librosa.yin() — coefficient of variation below 0.10 indicates synthesized prosody.
- **Decision Fusion:** Logical OR rule: if either the audio classifier or the lip-sync module flags the input as fake, the output is

Detection Models

1. ASVspoof 2021 Baselines (Reference Points)

- B01 (CQCC-GMM): min t-DCF 0.4974, EER 15.62%
- B02 (LFCC-GMM): min t-DCF 0.5758, EER 19.30%
- B03 (LFCC-LCNN): min t-DCF 0.3445, EER 9.26% — best
- baseline
- B04 (RawNet2): min t-DCF 0.4257, EER 9.50%; DF best EER 22.38%

Best participant (Team T23): min t-DCF 0.2177, EER 1.32% on the LA task.

2. Proposed CNN Classifier Architecture

The CNN takes a 2D feature map — MFCC (57×T) or Melspectrogram (80×T) — and outputs a binary softmax prediction. Table III details each layer.

Table 3: Proposed CNN Classifier Architecture

Layer	Output Shape	Notes
Input	H × W × 1	MFCC: 57×T; Mel: 80×T
Conv2D+BN+ReLU+MaxPool	H/2×W/2×32	3×3, 32 filters
Conv2D+BN+ReLU+MaxPool	H/4×W/4×64	3×3, 64 filters
Conv2D+BN+ReLU+MaxPool	H/8×W/8×128	3×3, 128 filters
Global Avg Pooling	128	Spatial

		→ 1×1
FC + Dropout (p = 0.3)	64	ReLU activation
FC + Softmax	2	{Real, Fake}
Total Params	≈185,000	Adam; BCE loss

3. ECAPA-TDNN Speaker Verification

SpeechBrain's pretrained ECAPA-TDNN (VoxCeleb2) generates 192-dimensional embeddings via TDNN layers with channel-dependent attention and attentive statistics pooling. A cosine similarity below τ_sv triggers a speaker mismatch flag.

4. Lip Synchronization Model (Phase II)

The lip-sync module computes Pearson correlation between L(t) and E(t) over a 2 s sliding window with 0.5 s stride. Correlation below $\tau\_lip = 0.30$ signals desynchronization. F0 trajectory smoothness is measured using the coefficient of variation of interframe F0 differences: CV < 0.10 indicates synthesized prosody.

Experiments And Results

1. Experimental Setup

All experiments ran on Google Colab with NVIDIA T4 GPU (15 GB VRAM). Software: Python 3.9, Librosa 0.10, Torchaudio 2.0, SpeechBrain 0.5, PyTorch 2.0, OpenCV 4.7. Metrics — Phase I: Accuracy, EER, AUC (following ASVspoof DF protocol [4]); Phase II: Lip-sync mismatch TPR; Full system: F1, Precision, Recall.

2. Phase I: Acoustic Feature Comparison

Table 4 shows classification results for three feature configurations on the custom dataset. All models used the CNN from Section V-B with fixed hyperparameters (lr=0.001, batch=32, 50 epochs) for a fair comparison.

Table 4: Phase I Classification Results (Custom Dataset)

Feature	Acc.(%)	EER(%)	AUC(%)	F1
MFCC (19+Δ+ΔΔ)	91.3	8.7	93.2	0.912
Mel-Spectrogram	88.6	11.4	90.7	0.884
MFCC + Spec (Fusion)	94.5	5.5	96.1	0.944
ASVspoof B03 [4]	—	9.26	—	—

Key observations: (i) Cloned voices show smoother MFCC trajectories and more uniform spectral energy. (ii) Melspectrograms reveal more regular harmonic spacing in synthesized voices. (iii) Removing silence trimming caused ∼4.2% accuracy drop, consistent with ASVspoof 2021 [4]. (iv) Feature fusion gave the best overall performance.

3. Phase I: Hyperparameter Sensitivity

Table 5: Hyperparameter Sensitivity (MFCC-CNN)

Learning Rate	Epochs	Accuracy (%)	EER (%)
0.001	20	82.5	17.4
0.001	50	91.3	8.7
0.001	100	90.1	10.2
0.0001	50	87.4	12.5
0.0001	100	89.8	10.9

Best configuration: lr=0.001, 50 epochs (91.3% accuracy, 8.7% EER). Training beyond 50 epochs at lr=0.001 showed slight overfitting — accuracy dropped to 90.1%.

4. Phase II: Lip Synchronization

The lip-sync module was evaluated on 80 video samples: 40 authentic and 40 AFVR-manipulated (original audio replaced with SV2TTS-cloned audio). Table VI reports detection performance.

Table 6: Lip Sync Detection Results (AFVR Scenario)

Method	TPR (%)	FPR (%)	F1	Notes
Lip-Energy Corr. (τ = 0.30)	82.5	12.5	0.849	Pearson 2s win.
F0 Smoothness (CV < 0.10)	77.5	10.0	0.822	librosa.yin()
Combined (OR rule)	90.0	15.0	0.873	Phase II output

Key observations: (i) Fake audio showed significantly lower lip-energy correlation (mean r=0.18 vs. 0.61 for real). (ii) SV2TTS audio had more regular F0 (mean CV=0.07 vs. 0.19 for real speech).

(iii) Even when Phase I found the audio convincing, Phase II still detected AFVR

manipulation in 82.5% of cases.

5. Combined System Performance

Table 7: Full System Performance vs. Reference Methods

System	Acc.(%)	EER(%)	AUC(%)	F1
Phase I (CNN-MFCC)	91.3	8.7	93.2	0.912
Phase I (CNN-Spec)	88.6	11.4	90.7	0.884
Phase II (LipSync)	87.5	—	88.6	0.873
Phase I+II (Proposed)	95.8	4.3	97.2	0.957
ASVspoof B03 [4]	—	9.26	—	—
FakeAVCeleb Xception [3]	—	—	72.5	—
FakeAVCeleb Ensemble [3]	—	—	82.8	—

The proposed Phase I+II system achieves 95.8% accuracy, 4.3% EER, 97.2% AUC, and F1=0.957, outperforming every subcomponent and surpassing the ASVspoof 2021 B03 baseline (EER 9.26%). The multimodal fusion added +4.5% accuracy over the best audio-only result, confirming that acoustic and temporal-synchrony cues are genuinely complementary.

Conclusion And Future Work

1. Conclusion

This paper introduced a two-phase multimodal detection system for the Real Voice, Fake Message threat. Phase I showed that AI-cloned voices can be reliably detected using MFCC and Melspectrogram features with a CNN binary classifier (best: 91.3% accuracy, EER 8.7%). Proper silence trimming matters more than expected — skipping it caused a meaningful accuracy drop, consistent with ASVspoof 2021 [4]. Phase II showed that lip-energy correlation and F0 smoothness detect AFVR manipulations with TPR 82.5–90.0%. Decision fusion of both phases achieved 95.8% accuracy and F1=0.957, validating the multimodal approach advocated by FakeAVCeleb [3].

2. Limitations

- Custom dataset limited to 4 speakers and Indian English; generalization to other accents untested.
- System not yet tested against VALL-E, YourTTS, ElevenLabs, or adversarial examples.
- Lip-sync module uses rule-based heuristics; degrades with minimal lip movement or low-resolution video.
- Performance under heavy codec compression not fully characterized.

3. Future Work

- Integrate SSL frontends (wav2vec 2.0, HuBERT, WavLM) which currently lead post-challenge ASVspoof 2021 leaderboards.
- Replace rule-based lip-sync with a learned model such as SyncNet or the Wav2Lip discriminator.
- Expand dataset with diverse speakers, languages, and ethnic backgrounds following FakeAVCeleb's five-group design.
- Apply data augmentation: codec simulation, room impulse response convolution, speed perturbation.
- Explore partially spoofed audio detection and extend to AFVF scenarios.
- Deploy as a real-time REST API or browser extension for practical media verification.

References

Y. Jia et al., "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," NeurIPS, 2018.

K. R. Prajwal et al., "A lip sync expert is all you need for speech to lip generation in the wild," Proc. ACM Multimedia, 2020, pp. 484–492.

H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," arXiv:2108.05080, 2022. [4]

J. Yamagishi et al., "ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection," arXiv:2109.00537, 2021.

A. Nautsch et al., "ASVspoof 2019: Spoofing countermeasures for detection of synthesized, converted and replayed speech," IEEE Trans. Biometrics Behav. Identity Sci., vol. 3, no. 2, pp. 252–265, 2021.

M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for ASV," Comput. Speech Lang., vol. 45, pp. 516–535, 2017.

M. Sahidullah, T. Kinnunen, and C. Hanilci, “A comparison of features for synthetic speech detection,” Proc. Interspeech, 2015,

pp. 2087– 2091.

H. Tak et al., “End-to-end anti-spoofing with RawNet2,” Proc. ICASSP, 2021, pp. 6369–6373.

L. Li et al., “Face X-ray for more general face forgery detection,” Proc. IEEE/CVF CVPR, 2020, pp. 5001–5010.

Y. Qian et al., “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” Proc. ECCV, 2020, pp. 86–103.

A. Haliassos et al., “Lips don’t lie: A generalisable and robust approach to face forgery detection,” Proc. IEEE/CVF CVPR, 2021, pp. 5039–5049.

T. Kinnunen et al., “Tandem assessment of spoofing countermeasures and ASV: Fundamentals,” IEEE/ACM Trans. Audio Speech Lang. Process., vol. 28, pp. 2195–2210, 2020.

J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep speaker recognition,” Proc. Interspeech, 2018, pp. 1086–1090.

R. Malik, P. Singh, and R. Chatterjee, “Battling voice spoofing: A review, comparative analysis, and generalizability evaluation,” Artif. Intell. Rev., vol. 56, 2023.

Team-Assembled Dataset (2025), “Real & Cloned Voice Dataset,” Dept. AI & DS, Dr. J. J. Magdum College of Engineering, Jaysingpur, Maharashtra, India.