iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** XI **Month of publication:** November 2026

DOI: https://doi.org/10.22214/ijraset.2026.85696

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue XI Nov 2026- Available at www.ijraset.com

Detection of Unseen Voice-Cloning Attacks

Dhanashri Abhang[1], Koushik Jaiswal[2], Chandan Sharma[3]
Master of Computer Application (MCA), JSPM's Narhe Technical Campus, Narhe, Pune, India

Abstract: ***The rapid advancement of artificial intelligence-based speech synthesis and voice-cloning technologies has made it increasingly difficult to distinguish genuine human speech from artificially generated or cloned voices. Although existing detection techniques can perform strongly on known synthetic speech, their reliability against previously unseen voice-cloning attacks remains a significant research challenge. This paper investigates the problem of detecting unseen voice-cloning attacks with emphasis on generalization, robustness, and evaluation under changing attack conditions. The review covers acoustic and spectral analysis, anti-spoofing systems, one-class learning, zero-shot detection, domain generalization, adversarial attacks, and real-time detection. Based on the identified gap, the proposed research design uses a feature-based one-class detection baseline in which genuine speech is modeled and evaluated against both known and unseen cloned speech. The methodology includes audio normalization, extraction of spectral and cepstral characteristics, one-class classification, cross-generator evaluation, and robustness testing under compression and signal-processing variations. The research design is intended to measure the generalization gap using accuracy, precision, recall, F1-score, and Equal Error Rate. Because no experimental audio dataset or executed model outputs were supplied with the source material, this paper does not present invented numerical experimental results; instead, it provides a complete evaluation protocol and a results structure that can be populated after execution. The literature synthesis indicates that performance on known attacks can overestimate real-world reliability and that unseen generators, post-processing, and adversarial modifications can reduce detection effectiveness. The study therefore emphasizes generator-independent characteristics, explicit unseen-attack evaluation, and open-world robustness as essential requirements for dependable voice-cloning detection in cybersecurity, authentication, telecommunications, and digital forensics.***

Keywords: Voice Cloning, Synthetic Speech Detection, Unseen Attacks, Voice Deepfakes, Speech Spoofing, Zero-Day Attacks, Generalization, Audio Security, Deepfake Detection.

I. INTRODUCTION

A. Introduction

The rapid advancement of artificial intelligence and speech synthesis technologies has made it possible to generate highly realistic human-like speech from a short voice sample. Modern voice-cloning systems can reproduce a speaker's vocal characteristics, including tone, accent, pronunciation, and speaking style, with increasingly convincing quality. Although these technologies have beneficial applications in accessibility, entertainment, education, and human-computer interaction, they have also introduced significant security and privacy risks. Cloned voices can be misused for impersonation, social engineering, financial fraud, unauthorized authentication, misinformation, and other cyberattacks. Research has demonstrated that humans are often unable to reliably distinguish genuine speech from AI-generated or cloned speech, increasing the importance of automated detection mechanisms

Existing research has developed numerous methods for detecting synthetic and spoofed speech using acoustic characteristics, spectral features, speaker representations, neural networks, and anti-spoofing frameworks. Standardized datasets such as ASVspoof have contributed significantly to the development and evaluation of voice-spoofing countermeasures. However, a major limitation of many existing systems is their dependence on known attack types, datasets, and synthesis technologies. A detector developed around samples produced by particular voice-cloning systems may perform well on similar attacks but experience a substantial reduction in performance when confronted with speech generated by an unfamiliar or previously unseen cloning system. This challenge is commonly associated with the generalization problem in speech deepfake detection.

The emergence of new voice-conversion, speech-synthesis, and voice-cloning techniques makes this problem increasingly important. Attackers can employ new generators, modify existing generation pipelines, apply post-processing operations, or introduce adversarial modifications that alter the characteristics used by conventional detectors. Consequently, detection systems designed primarily around known attack signatures may fail against zero-day or unseen voice-cloning attacks. Recent studies have specifically highlighted the difficulty of achieving reliable generalization across unseen speech deepfakes and real-world conditions [1]

©IJRASET: All Rights are Reserved

Therefore, Detection of Unseen Voice-Cloning Attacks focuses on addressing the gap between detecting known attacks and recognizing previously unseen forms of voice cloning. The central objective is to investigate detection approaches that rely on robust and generalizable characteristics of speech rather than signatures associated with a particular generation technique. Such an approach can improve the ability of detection systems to identify unfamiliar synthetic or cloned voices and provide greater resilience against the continuously evolving landscape of voice-cloning attacks. Developing reliable detection of unseen attacks is particularly relevant for voice authentication, digital forensics, telecommunications, online communication, and cybersecurity applications. A robust detection framework can contribute toward protecting speaker identity, reducing impersonation risks, and improving trust in voice-based communication systems. Thus, research into unseen voice-cloning attack detection represents an important step toward developing more adaptable and future-resistant defenses against emerging speech-based cyber threats.

1) *Technical Background of Voice Cloning:* Contemporary neural voice-cloning systems are commonly organized around three functional components. A speaker-representation module encodes the identity-bearing characteristics of a reference utterance into a fixed-dimensional embedding. A synthesis or acoustic model then maps linguistic content, supplied either as text or as the speech of a source speaker, together with that embedding to an intermediate acoustic representation such as a mel-spectrogram. Finally, a vocoder converts the intermediate representation into a waveform. Text-to-speech cloning and voice conversion differ mainly in the input: the former generates new linguistic content in the target voice, whereas the latter transforms an existing utterance while preserving what was said. Few-shot and zero-shot cloning systems extend this pipeline by adapting to a new speaker from a short reference recording without retraining the model for that speaker. This decomposition matters for detection because each stage can leave measurable traces. A vocoder reconstructs fine spectral detail and phase from a compressed intermediate representation, and an acoustic model may produce temporal dynamics that are smoother or more regular than natural articulation. Because architectures, training data, and vocoders differ between generators, the traces they leave also differ. A detector tuned to the traces of one generator is therefore tuned to a property of that generator rather than to a property of cloned speech in general, which is the underlying reason for the generalization problem examined in this study.
2) *Attack Scenarios and Security Context:* Three operational scenarios illustrate how the consequences of detection errors vary. In the first, a cloned utterance is presented to an automatic speaker-verification or voice-authentication system in an attempt to gain access under a legitimate speaker's identity. In the second, the cloned voice is directed at a human listener, for example in a telephone call or voice message used for fraud or social engineering, where the detector acts as a warning mechanism rather than a gate. In the third, a recording is submitted as evidence and an examiner must decide whether it is authentic. A missed clone is most costly in the first scenario, an unnecessary alarm is most disruptive in the second, and in the third both error types affect the weight that can be placed on the conclusion. The feasibility of these scenarios is supported by the literature reviewed in Section II. Wenger et al. [2] demonstrated real-world speech synthesis attacks, and Wang et al. [3] examined the risk of cloning a voice from a single stolen utterance, indicating that an attacker does not need extensive recordings of the victim. Javed et al. [4] considered voice spoofing in the context of cyber-physical and IoT systems, which shows that voice-driven interfaces extend the attack surface beyond conventional call-centre and banking settings.

B. Statement of the Problem

Existing voice-cloning detection systems are commonly developed and evaluated using known datasets and attack-generation techniques. Such evaluation can produce high detection scores while providing limited evidence about performance against a new cloning generator. The central problem addressed by this research is the lack of dependable generalization from known synthetic speech attacks to previously unseen voice-cloning attacks. When the characteristics of an unseen generator differ from those represented during development, a detector may misclassify cloned speech as genuine speech. The problem can become more difficult when the audio is compressed, transmitted through a communication channel, or otherwise processed. Therefore, there is a need for a research design that explicitly separates known and unseen attacks, identifies generalizable speech characteristics, and measures the resulting generalization gap.

C. Objectives of the Research

- To study existing voice-cloning and synthetic speech detection techniques and identify their strengths, limitations, and applicability to unseen attacks.
- To analyze the characteristics of genuine and cloned speech in order to identify speech patterns, acoustic properties, and other distinguishable characteristics that can support reliable detection.

- To evaluate the performance of detection methods against known and unseen voice-cloning attacks, with particular emphasis on performance reduction when the attack-generation technique differs from development data.
- To investigate the generalization capability of voice-cloning detection systems across different voice-cloning, voice-conversion, and synthetic speech generation techniques.
- To examine the effect of audio variations and attack modifications such as compression, processing, and recording conditions on detection of cloned speech.
- To identify limitations of existing detection approaches in open-world and zero-day scenarios where attack characteristics are unavailable during development.

D. Hypothesis of the Study

H1: A detection system based on generalizable speech characteristics can distinguish genuine speech from synthetic or cloned speech with useful classification performance.
H2: Detection performance decreases when the system is evaluated against previously unseen voice-cloning generators compared with known attack conditions.
H3: Audio transformations such as compression and signal processing further reduce detection performance for unseen voice-cloning attacks.
Each hypothesis is formulated as a directional, testable comparison. H1 is assessed by whether the detector separates held-out genuine speech from known cloned speech better than a chance-level scorer, for which the Equal Error Rate would approach 50%. H2 is assessed by comparing the same metrics for known and unseen conditions under an identical decision threshold. H3 is assessed by comparing unseen-attack performance before and after controlled transformations. The corresponding null hypotheses state that no difference exists. Failure to support H2 would not invalidate the study; it would indicate that the selected features capture characteristics that transfer across the generators examined, which is itself an informative outcome for the research question.

E. Significance of the Study

The study is significant because voice cloning can undermine trust in voice-based identity and communication. In authentication systems, a successful cloned voice may be used to impersonate a legitimate speaker. In telecommunications and social engineering, realistic synthetic speech can be used to increase the credibility of fraudulent interactions. In digital forensics, investigators may need to determine whether an audio recording is genuine or synthetically generated. The research therefore contributes to the security objective of distinguishing genuine speech from emerging synthetic speech while recognizing that future attacks may not resemble previously observed examples. The study is also significant academically because it shifts evaluation from closed-set accuracy toward explicit testing of generalization and robustness.

F. Scope and Delimitations

The study is restricted to audio-only detection at the level of a whole utterance or fixed-length segment, where the output is a binary decision between genuine and cloned speech. The proposed baseline is deliberately limited to interpretable spectral and cepstral features and a one-class classifier, so that the effect of generator shift can be separated from the effect of model capacity. The study does not address attribution of a clone to a specific generator, temporal localization of manipulated regions within an utterance, integration with a full speaker-verification pipeline, audio-visual deepfakes, or deployment on a live communication system. These topics are discussed as directions for further work.
These delimitations keep the research question narrow enough to be answered by a controlled experiment. A broader scope would require multiple models, datasets, and deployment conditions, and would make it difficult to attribute any observed change in performance to the known-versus-unseen distinction that this research is designed to isolate.

II. REVIEW OF LITERATURE

The literature on voice-cloning and speech deepfake detection has evolved from handcrafted acoustic and spectral features toward increasingly sophisticated representation-learning, anti-spoofing, one-class, zero-shot, domain-generalization, and adversarial-robustness approaches. For the present research, the literature is reviewed thematically and chronologically with emphasis on the question of whether a detector can remain effective when the attack generator is not represented in its development data.

A. .Early Synthetic Speech and Spoof Detection

Sánchez et al. investigated universal synthetic speech spoofing detection using phase information, establishing that phase-related characteristics can contain evidence useful for separating synthetic and genuine speech. Paul et al. [5] subsequently examined spectral features for synthetic speech detection, reinforcing the importance of frequency-domain characteristics. These early studies are important because they demonstrate that speech synthesis can leave measurable signal-level traces even when the generated speech is perceptually convincing.

Wang et al. [6] introduced the ASVspoof 2019 database as a large-scale public resource containing synthesized, converted, and replayed speech. The dataset and associated challenge structure provided a common evaluation environment for spoofing countermeasures. Wu et al. [7] investigated a light convolutional approach with feature genuinization, while Subramani and Rao [8] examined efficient representations for fake speech detection. Together, these studies established a progression from handcrafted signal characteristics toward learned representations while retaining the core objective of identifying artifacts introduced by speech generation or replay.

The signal-level traces that these early studies exploit have a plausible origin in how synthesis systems work. Synthesis models operate on compact intermediate representations and must approximate fine spectral structure, phase relationships, and short-term temporal variation that natural speech production generates physically. Statistical regularities in these properties can therefore differ from those of genuine speech even when the output sounds natural. However, as synthesis technology has moved to neural architectures, the specific cues change. Features that were informative for earlier systems cannot be assumed to remain informative for later ones, which is the motivation for evaluating detectors against generators they have not encountered.

B. Anti-Spoofing Frameworks and Standardized Evaluation

Nautsch et al. [9] reported on ASVspoof 2019 countermeasures for synthesized, converted, and replayed speech, while Yamagishi et al. [10] described ASVspoof 2021 as a mechanism for accelerating progress in spoofed and deepfake speech detection. Liu et al. extended this direction through ASVspoof 2021 work aimed at spoofed and deepfake speech detection in the wild. These resources demonstrate the importance of standardized datasets, protocols, and metrics when comparing countermeasures.

Aljasem et al. [11], Hassan and Javed [12], Dawood et al. [13], and Javed et al. [14] explored different anti-spoofing and voice-spoofing countermeasures. Their work shows that detection can be improved through carefully selected representations and classification strategies. However, performance is still conditioned by the relationship between the development data and the test attacks. Khan et al. [15] specifically reviewed voice spoofing countermeasures and emphasized generalizability as an important evaluation dimension.

The ASVspoof 2019 database distinguishes between logical-access scenarios, which involve synthesized and converted speech, and physical-access scenarios, which involve replayed recordings [6]. Evaluation protocols in this line of work commonly report the Equal Error Rate together with a tandem detection cost function that accounts for the interaction between the countermeasure and a speaker-verification system. The Equal Error Rate is attractive because it is independent of any single decision threshold, but it depends on the composition of the evaluation set and does not identify the operating point that a deployed system would use. For this reason the present design reports threshold-based metrics alongside the Equal Error Rate. The ASVspoof 2019 evaluation design also includes attack systems that are absent from the training partition, which shows that the challenge itself recognized the importance of unseen conditions even though much subsequent work reports results on matched partitions.

C. Human Detection and Real-World Voice-Cloning Threats

Mai et al. [16] examined human ability to detect speech deepfakes and reported that people cannot reliably distinguish deepfake speech. Milewski et al. [17] compared neural-network-based detection with human performance for cloned voices, while Barrington et al. [18] investigated single- and multi-speaker cloned voice detection. Barrington et al. [19] further highlighted that people are poorly equipped to detect AI-powered voice clones. These findings establish that perceptual judgment alone cannot be treated as a dependable security mechanism.

Wenger et al. [2] demonstrated real-world speech synthesis attacks, and Alali and Theodorakopoulos [20] studied partial fake speech attacks in realistic settings. Wang et al. [3] examined the risk associated with voice cloning from a single stolen utterance. These studies demonstrate that the threat model is not limited to laboratory-generated examples: an attacker may possess limited source material and still produce convincing synthetic speech. This increases the need for detectors that can operate under realistic and evolving conditions.

The threat model in these studies can be described along several dimensions: how much the attacker knows about the detector, how much reference audio the attacker has, which generator and post-processing the attacker selects, and through which channel the audio reaches the target, such as a live call, a stored message, or an authentication request. Different combinations of these factors produce different detection difficulty, and a study that fixes all of them to a single laboratory configuration cannot support conclusions about the others. The present research fixes the attacker's knowledge of the detector but varies the generator and the processing, as specified in Section IV-A1.

D. Deepfake Speech Detection and Feature-Based Methods

Sinha and Jana investigated increasingly sophisticated feature and model combinations. Unoki et al. [21] reviewed the progression from acoustic features to deep neural networks, and Li et al. [22] and Pham et al. [23] synthesized a broader body of speech deepfake detection research. These reviews show that detection accuracy has improved substantially, but they also identify continuing challenges in robustness, dataset shift, and generalization.

The feature families used in the literature can be grouped by the aspect of speech they describe. Short-term spectral and cepstral features summarize the spectral envelope over brief frames. Long-term features, such as the prediction traces examined by Borrelli et al. [24], describe dependencies across longer spans of the signal. Production-related features, including formants and the speech-pathological measures studied by Liu et al. [25] and Chaiwongyen et al. [26], relate to how the vocal tract shapes the signal, and phase-based features capture information that magnitude spectra discard. Unoki et al. [21] reviewed the progression from such acoustic features to deep neural networks. Handcrafted features have lower representational capacity than learned embeddings, but they have the advantage that a failure can be traced to a specific acoustic dimension, which supports the diagnostic aim of the present study.

E. One-Class and Anomaly-Based Detection

One-class learning is particularly relevant to unseen attacks because it changes the problem from learning a closed list of fake generators to modeling a reference distribution and identifying deviations from it. Zhang et al. [27] investigated one-class learning for synthetic voice spoofing detection. Zhang et al. [28] later considered improved Support Vector Data Description for one-class fake speech detection, while Kim et al. [29] proposed adaptive centroid shifting for audio deepfake detection. Lin et al. [30] examined a one-class neural network with directed statistics pooling, and Ye et al. [31] studied spectro-temporal features for generalized synthetic speech detection.

These studies are directly relevant to the present research because a one-class formulation can reduce dependence on an exhaustive catalog of future attack generators. However, one-class detection remains difficult when genuine speech itself has large variability or when an unseen synthetic attack lies close to the learned genuine-speech boundary. The research gap therefore shifts from merely building a classifier to selecting representations and decision boundaries that remain useful when the attack distribution changes.

One-class formulations rest on a specific assumption: that genuine speech occupies a compact region of the feature space and that cloned speech falls outside it. Two classical formulations are common. The One-Class Support Vector Machine separates the training data from the origin in a kernel-induced feature space, whereas Support Vector Data Description encloses the data in a minimal hypersphere. With a radial basis function kernel the two formulations are closely related. The assumption fails when genuine speech is itself highly variable, for example across accents, microphones, and recording conditions, because the learned region becomes either too loose to exclude high-quality clones or too tight to accept legitimate variation. A one-class detector will also flag genuine speech recorded under conditions absent from the training data, so its false-rejection behavior must be examined explicitly.

F. Zero-Shot and Unseen Speech Deepfake Detection

Kukanov et al. [32] explicitly investigated meta-learning approaches for improving detection of unseen speech deepfakes. Niu et al. studied zero-shot synthetic audio detection using self-supervised representations. These studies directly address the problem of evaluating a detector on attack types not represented during conventional development. Their importance lies in treating unseen attacks as a first-class evaluation condition rather than as an accidental failure case.

The trend continues in more recent work. D et al. examined zero-day neural voice-cloning attack detection, while Khan et al. [33] investigated training-free partial audio deepfake detection through embedding trajectory analysis. These directions indicate that future detectors may need to operate with limited or no examples of the specific attack family being encountered.

1) *What Makes an Attack Unseen:* The term unseen is used in the literature with more than one meaning. An attack may be unseen along several distinct dimensions:

- Generator: the synthesis or conversion architecture, including its vocoder, is absent from development.
- Speaker: the target or source identities were not present in development data.
- Language and accent: the linguistic or phonetic content differs from that of the development data.
- Channel: the recording device, transmission path, or codec differs from those seen during development.
- Post-processing: the audio has been modified after synthesis, deliberately or incidentally.

When these dimensions change together, a drop in performance cannot be attributed to any one of them. The present research therefore treats the generator as the primary unseen dimension and treats processing as a secondary factor that is varied in a separate, controlled stage. An attack is regarded as unseen with respect to a development protocol only if neither its generator nor a close derivative, such as a generator sharing the same vocoder, contributed to any development decision, including feature selection, hyperparameter tuning, and threshold calibration.

G. Domain Generalization and Representation Robustness

Domain-generalization methods treat each generator or dataset as a domain and seek representations that remain informative when the domain changes. The strategies in the studies above include aggregation and separation of domain information [34], feature decomposition with synthesizer-feature augmentation [35], adversarial alignment with an information bottleneck [37], multi-level gating of self-supervised features [39], and augmentation to increase feature variability [40]. These methods learn invariances from data. A handcrafted feature set with a one-class classifier learns no such invariances; any robustness to generator shift arises solely from the choice of features and the normalization applied to them. The baseline proposed here therefore serves as a lower-capacity reference against which the contribution of learned invariance can be judged in later work.

H. Adversarial, Laundering, and Post-Processing Attacks

Ali et al. [41] investigated whether audio spoof detection is robust to laundering attacks, showing why signal processing and transformation must be considered in realistic evaluation. Kamel et al. [42] examined black-box spectral masking and interpolation attacks against voice authentication and anti-spoofing systems. Wang et al. [43] studied BIM-based adversarial attacks against speech deepfake detectors, while Xuan et al. investigated adversarial neural mimicry designed to evade fake-voice detection.

These studies reveal a second dimension of generalization: a detector may fail not only because the generator is new, but also because an attacker deliberately modifies the audio after synthesis. Therefore, the present research treats robustness to compression and signal processing as part of unseen-attack evaluation rather than as a separate engineering concern.

I. Real-Time Detection and Practical Deployment

Mittal et al. [44] proposed PITCH, an AI-assisted challenge-response approach for tagging deepfake audio calls. These studies demonstrate that practical deployment requires more than offline classification accuracy; latency, robustness, and the operating environment also matter.

Javed et al. [4] considered voice spoofing in cyber-physical and IoT systems, reinforcing the need to view voice attacks as part of broader system security. Univaso and Segundo [45] considered cloned voices in realistic forensic voice-comparison scenarios. The practical literature therefore supports a research design that reports not only predictive performance but also the conditions under which a detector succeeds or fails.

J. Open-World Generalization and Research Gap

The overall research gap is therefore the lack of a universally reliable detector whose performance remains stable when the voice-cloning generator, speaker, language, channel, or post-processing condition changes. The present study addresses this gap by defining an explicit known-versus-unseen evaluation protocol and proposing a feature-based one-class baseline whose decision boundary is derived primarily from genuine speech. This framing is consistent with the literature on one-class learning, zero-shot detection, domain generalization, and open-world limitations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

K. Research Gap Summary

TABLE I
RESEARCH GAP SUMMARY

Aspect	Existing Research	Identified Gap	Present Research Focus
Training data	Known synthetic speech and established datasets dominate evaluation.	Future cloning methods may not be represented.	Separate known and unseen attack testing.
Attack types	Many systems target specific synthesis or conversion methods.	New generators may produce different artifacts.	Cross-generator evaluation.
Feature dependence	Detectors can exploit dataset- or generator-specific cues.	Such cues may not transfer.	Generalizable spectral/cepstral characteristics.
Robustness	Adversarial and post-processing attacks are increasingly studied.	Modified audio can evade detection.	Compression and processing stress tests.
Generalization	Zero-shot, one-class and domain-generalization methods are emerging.	Reliable open-world detection remains unresolved.	Explicit unseen-attack evaluation.
Practical deployment	Real-time systems are being explored.	Accuracy, latency and generalization must coexist.	Practical evaluation framework.

L. Taxonomy of Detection Approaches and Implications for the Present Design

The studies reviewed above can be organized into families according to their learning paradigm. Table II summarizes each family, the main consideration for unseen attacks, and the studies from this review that illustrate it. The grouping is conceptual and several studies span more than one family.

TABLE II
TAXONOMY OF DETECTION APPROACHES

Approach family	Core idea	Consideration for unseen attacks
Handcrafted features with classical classifiers	Detect signal-level traces in phase, spectral, or prediction characteristics.	Interpretable and light, but cues may depend on the generator.
Learned representations for anti-spoofing	Learn discriminative features end to end from labelled genuine and spoofed speech.	Strong on matched attacks; reliance on training generators remains a risk.
One-class and anomaly-based detection	Model genuine speech and flag deviations from it.	Does not require future generators, but sensitive to genuine variability.
Zero-shot, meta-learning, and self-supervised methods	Transfer to attacks without task-specific examples by using general representations.	Treats unseen attacks as the primary evaluation condition.
Domain generalization and augmentation	Learn representations that are invariant across generators and datasets.	Reduces domain-specific reliance; no guarantee for arbitrary new generators.
Robustness-oriented evaluation	Test detectors against laundering and adversarial modification.	Reveals failure modes that clean benchmarks do not expose.
System-level and real-time approaches	Combine detection with interaction design and latency constraints.	Addresses deployment, but not generalization by itself.

The proposed baseline combines the first and third families, using handcrafted spectral and cepstral features with a one-class decision rule, selected for diagnostic transparency rather than maximum accuracy. The remaining families define comparators and extensions, which Section VII discusses.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

III. RESEARCH METHODOLOGY / RESEARCH DESIGN

A. Research Design

The study adopts an experimental comparative research design. The independent condition is the relationship between the development data and the attack-generation method used during evaluation: known or unseen. The dependent variables are detection performance measures including accuracy, precision, recall, F1-score, and Equal Error Rate. A secondary experimental factor is audio transformation, including compression and other controlled processing. The design therefore allows the research to measure both the basic detection capability and the generalization gap.

1) *Unit of Analysis and Operational Definitions:* The unit of analysis is a single audio segment, represented by one feature vector and assigned one score. The positive class is cloned speech and the negative class is genuine speech, which is the convention adopted for precision, recall, and error-rate definitions in Section III-H. Known and unseen attacks are distinguished by their role in development rather than by any intrinsic property of the audio. Because the one-class model is fitted on genuine speech only, no cloned sample ever enters the model-fitting step. Known cloned speech may, however, inform validation-stage decisions such as hyperparameter selection and threshold calibration, whereas an unseen generator informs nothing. This distinction is what allows the difference between the two conditions to be interpreted as a generalization gap.

B. Dataset Design

The dataset should contain three principal partitions: (1) genuine speech used to establish the reference distribution, (2) known cloned speech generated by techniques represented in the development protocol, and (3) unseen cloned speech generated by a technique intentionally excluded from development. A fourth robustness partition should contain transformed versions of unseen samples. Speaker identities should be separated across development and test partitions where feasible to prevent speaker memorization from being mistaken for attack detection.

TABLE III
DATASET PARTITIONS

Partition	Purpose	Key Requirement
Genuine training	Model normal/genuine speech characteristics.	Speaker-disjoint where feasible.
Known attack test	Measure performance on familiar attack conditions.	Attack family represented in development/evaluation design.
Unseen attack test	Measure cross-generator generalization.	Generator excluded from development.
Robustness test	Assess effect of compression/processing.	Same unseen attacks after controlled transformations.

1) *Candidate Data Sources:* Several publicly described resources discussed in the literature review could supply material for this design. These include the ASVspoof 2019, 2021, and 5 resources [6], [10], [46], the multilingual SpeechFake dataset [47], and the SpoofCeleb dataset [50]. Selection should be governed by the design requirements rather than by size. The data must provide genuine speech, a generator identifier for each cloned sample so that a generator can be excluded from development, and a speaker identifier so that speaker-disjoint partitions can be constructed. A dataset lacking generator-level labels cannot support unseen-generator evaluation, and the licence terms of each resource must be checked before use.
2) *Leave-One-Generator-Out Partitioning:* If a dataset contains G generators, a single held-out generator provides only one estimate of unseen-attack behavior, and that estimate may be unrepresentative. A leave-one-generator-out scheme addresses this by holding out each generator in turn as the unseen condition, treating the remainder as known, and repeating the full evaluation. Results should be reported as a distribution across rotations, with the mean and the spread, rather than as a single figure. Generators that share a vocoder or a common architecture should be grouped into families and held out together, otherwise a close relative of the unseen generator remains among the known attacks and the unseen condition is only nominally unseen.
3) *Speaker, Session, and Class-Balance Considerations:* Speaker-disjoint partitioning should be supplemented by session-level separation where metadata permits, since recordings from the same session share microphone and room characteristics. The distributions of segment duration and loudness should be compared across genuine and cloned classes before modelling, because systematic differences would allow a detector to succeed on properties unrelated to cloning.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

4) *Ethical and Data-Governance Considerations:* Voice recordings are biometric personal data. Datasets should be used within the terms of their licences, and any newly recorded genuine speech should be collected with informed consent that states its purpose. Cloned samples should not be generated from the voice of any individual without that individual's consent, and stored audio should be access-controlled and anonymized where possible. These obligations apply even though the research purpose is defensive.

C. Audio Pre-processing

Audio files should be converted to a common sampling rate and channel configuration, amplitude-normalized, and segmented into comparable analysis windows. Silence trimming may be applied consistently across all classes. The same preprocessing pipeline must be applied to genuine, known, and unseen speech to avoid introducing a class-specific artifact. A reproducible configuration should record sampling rate, frame length, frame overlap, normalization method, and any resampling or filtering operation.

Resampling to a common rate has a direct bearing on what the detector can observe. At a sampling rate of 16 kHz, the Nyquist frequency is 8 kHz, so any generator trace located above that frequency is removed before feature extraction. This is a practical trade-off between computational cost and spectral coverage and should be recorded as a design parameter. Amplitude normalization should be applied so that overall loudness cannot act as a cue, and the choice between peak and root-mean-square normalization should be stated. Pre-emphasis may be applied to balance the spectral tilt of speech before analysis.

D. Feature Extraction

The proposed baseline uses a combination of cepstral and spectral characteristics. Candidate features include Mel-frequency cepstral coefficients (MFCCs), spectral centroid, spectral bandwidth, spectral roll-off, zero-crossing rate, spectral contrast, and selected delta/delta-delta statistics. The feature vector should be summarized over frames using statistics such as mean and standard deviation. This provides a compact representation that can be inspected and evaluated without requiring a large end-to-end neural architecture. The feature choice is motivated by the long-standing use of spectral and acoustic characteristics in synthetic speech detection [5], and by later work integrating formant and other speech-production characteristics [25].

1) *Description of Candidate Features:* The candidate features capture complementary aspects of the short-time spectrum:

- Mel-frequency cepstral coefficients (MFCCs): computed by framing and windowing the signal, taking the magnitude spectrum, applying a bank of triangular filters spaced on the mel scale, taking the logarithm of the filter energies, and applying a discrete cosine transform. Low-order coefficients describe the spectral envelope associated with the vocal tract, while higher-order coefficients describe finer detail.
- Delta and delta-delta coefficients: the first and second temporal derivatives of the cepstral coefficients, which describe how the spectrum changes over time.
- Spectral centroid: the magnitude-weighted mean frequency of a frame, $C = \Sigma\, f(k)\cdot|X(k)| \,/\, \Sigma\, |X(k)|$, which corresponds to the perceived brightness of the sound.
- Spectral bandwidth: the magnitude-weighted spread of frequencies around the centroid.
- Spectral roll-off: the frequency below which a specified fraction of the total spectral energy is contained.
- Spectral contrast: the difference between spectral peaks and valleys in a set of frequency sub-bands, which reflects the balance between harmonic and noise-like components.
- Zero-crossing rate: the rate at which the waveform changes sign, which provides a coarse indication of noisiness and dominant frequency content.

2) *Feature Vector Construction:* In the baseline implementation shown in Listing 1, twenty MFCCs and four additional frame-level measures (spectral centroid, bandwidth, roll-off, and zero-crossing rate) yield twenty-four frame-level series. Computing the mean and standard deviation of each series over the frames of an utterance produces a forty-eight-dimensional vector. The additional candidates named above, namely spectral contrast and the delta coefficients, extend this vector and can be introduced as a controlled ablation. Any change in dimensionality requires the scaler and the classifier to be refitted on genuine training data. Summary statistics have a practical advantage: they give a fixed-length representation regardless of utterance duration. They also discard temporal order, so the representation cannot capture patterns such as the evolution of prosody over an utterance. This limitation is acknowledged in Section VIII, and it is one reason why learned or sequence-aware representations are proposed as future extensions.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

E. Classification / Detection Logic

A One-Class Support Vector Machine (OC-SVM) is proposed as the primary decision mechanism. The model is trained using genuine-speech feature vectors only. During inference, the model returns a normality/anomaly decision. Samples that remain sufficiently close to the learned genuine distribution are classified as genuine; samples outside the decision boundary are flagged as potentially synthetic or cloned. The key research question is not whether every anomaly is necessarily synthetic, but whether the decision boundary provides useful separation when the synthetic generator is unseen.

1) *Mathematical Formulation of the One-Class SVM:* Let $x_1, \ldots, x_n$ be the standardized feature vectors of genuine training speech and let φ denote the mapping into the kernel-induced feature space. The One-Class SVM finds a hyperplane that separates the mapped data from the origin with maximum margin by solving

$$\min_{w,\,\xi,\,\rho} \; \tfrac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_i \xi_i - \rho \qquad \text{subject to} \quad \langle w, \varphi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0$$

The decision function is $f(x) = \operatorname{sgn}\left(\sum_i \alpha_i k(x_i, x) - \rho\right)$, where α_i are the coefficients of the support vectors and k is the kernel. With the radial basis function kernel used in Listing 1, $k(x, x') = \exp(-\gamma\|x - x'\|^2)$. The parameter ν lies in (0, 1] and acts as an upper bound on the fraction of training samples that fall outside the learned boundary and as a lower bound on the fraction of support vectors.

F. Experimental Procedure

- Train the one-class model using only the designated genuine-speech development set.
- Evaluate the model on genuine speech held out from training to estimate the normal-class false alarm behavior.
- Evaluate on known cloned speech to establish a reference detection level.
- Evaluate on an unseen cloned generator that was excluded from development.
- Apply controlled compression and signal-processing transformations to the unseen samples and repeat evaluation.
- Compute accuracy, precision, recall, F1-score and EER where the experimental setup permits threshold-based scoring.
- Compare the known and unseen conditions and calculate the performance change.
- Interpret failures with respect to feature sensitivity, generator shift, speaker variability and audio transformation.

A single run on one partition gives a point estimate with unknown variability. The procedure should therefore be repeated over the leave-one-generator-out rotations described in Section III-C2 and over several random draws of the genuine development and validation subsets.

G. Evaluation Metrics

TABLE IV
EVALUATION METRICS

Metric	Purpose	Interpretation
Accuracy	Overall correct classification rate.	Higher is better, but can hide class imbalance.
Precision	Reliability of positive/cloned predictions.	Higher means fewer false positive clone flags.
Recall	Ability to detect cloned speech.	Higher means fewer cloned samples are missed.
F1-score	Balance of precision and recall.	Useful when both error types matter.
EER	Threshold where false acceptance and false rejection are equal.	Lower indicates better separation.

1) *Confusion-Matrix Definitions and Formulae:* With cloned speech as the positive class, each prediction falls into one of the four categories shown in the table below.

TABLE V
CONFUSION-MATRIX CATEGORIES

Actual \ Predicted	**Predicted cloned**	**Predicted genuine**
Actual cloned	True positive (TP): clone detected	False negative (FN): clone accepted as genuine
Actual genuine	False positive (FP): genuine speech flagged	True negative (TN): genuine speech accepted

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \qquad Precision = TP / (TP + FP)$$
$$Recall = TP / (TP + FN) \qquad F1\text{-}score = 2 \cdot Precision \cdot Recall / (Precision + Recall)$$
$$FAR = FN / (TP + FN) \qquad FRR = FP / (FP + TN)$$

In the anti-spoofing convention the false acceptance rate (FAR) is the proportion of cloned speech accepted as genuine, which equals one minus recall, and the false rejection rate (FRR) is the proportion of genuine speech rejected. The Equal Error Rate is the value at which FAR equals FRR as the threshold is varied. Precision depends on the proportion of cloned samples in the test set, so it must be reported with that proportion. Threshold-free summaries such as the ROC and detection-error-tradeoff curves are retained wherever continuous scores are available. Operational measures, namely inference time, real-time factor, and memory use, should be reported alongside the security metrics when deployment is a consideration.

H. Research Workflow

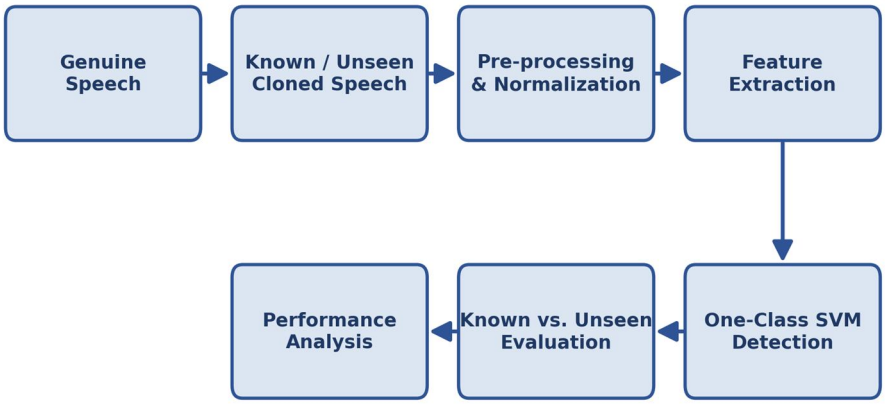


Fig. 1 Research workflow for evaluating generalization to unseen voice-cloning attacks

I. Evaluation Design Diagram

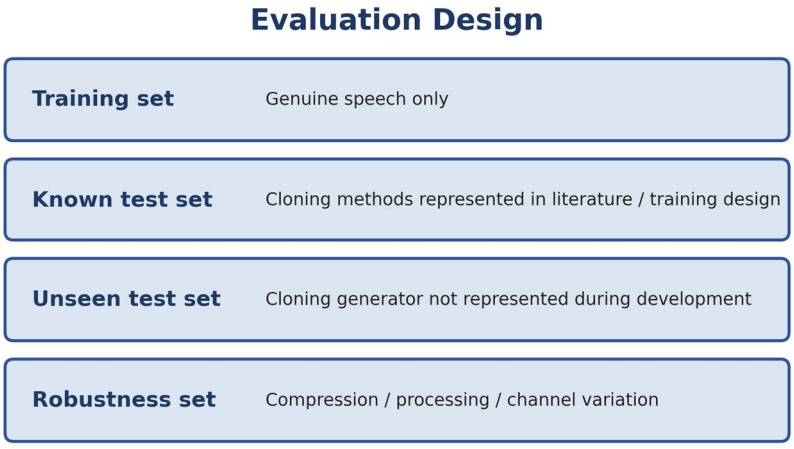


Fig. 2 Evaluation design: training, known, unseen and robustness sets

IV. PROPOSED WORK

A. Proposed Framework

The proposed framework is designed around the central hypothesis that some characteristics of genuine speech remain sufficiently stable to support anomaly detection even when the specific cloning generator is unknown. Instead of training a multi-class model whose classes correspond to known generators, the proposed baseline models genuine speech and treats deviations as candidate spoofing events. Known attacks are retained as a benchmark condition, while an unseen generator is reserved for generalization testing.

1) *Threat Model and Assumptions:* The framework assumes the following attacker and defender capabilities. The attacker may select any generator, including one that did not exist when the detector was developed. The attacker may hold only a limited reference recording of the target speaker. The attacker may apply post-processing such as compression or resampling, and the audio reaches the detector through an unspecified channel. The attacker is assumed to be non-adaptive, meaning that the attacker has no knowledge of the detector's features or decision boundary. Optimization-based adversarial attacks, such as those discussed in Section II-H, therefore fall outside the scope of the experiments, while robustness to ordinary processing is included.

The defender has access only to genuine speech and to validation material that excludes the unseen generator. The defender is also assumed to have genuine training speech that is representative of the conditions in which the detector will operate. If deployment conditions differ substantially from the training conditions, genuine speech itself becomes out-of-distribution, and the false-rejection behavior discussed in Section II-E becomes the dominant concern.

B. Algorithm

Algorithm 1: One-Class Feature-Based Detection of Unseen Voice-Cloning Attacks

Input: genuine training audio G, known attack test set K, unseen attack test set U, robustness set R.

1. Extract the same normalized feature vector from every audio sample.
2. Fit an OC-SVM using feature vectors extracted only from G.
3. For each test sample x, obtain the decision score and class label.
4. Record genuine/clone predictions and decision scores.
5. For K, calculate the baseline detection metrics.
6. For U, calculate the unseen-attack metrics.
7. For R, calculate metrics after each controlled transformation.
8. Compute the generalization gap as the difference between the selected known-attack and unseen-attack metrics.
9. Report results, error cases and robustness observations.

Two points clarify the algorithm. First, the generalization gap in step 8 is computed in the same direction as Section V-C, as the unseen value minus the known value, so that a negative change in accuracy, precision, recall, or F1-score and a positive change in the Equal Error Rate indicate a generalization gap. Second, the computational cost is modest. Training a kernel-based One-Class SVM scales superlinearly with the number of genuine training vectors, while inference scales with the number of support vectors and the feature dimension. For short audio segments, feature extraction rather than classification is expected to dominate the cost of a decision.

C. Detection Logic

The decision logic is intentionally simple: genuine speech defines the reference class; deviations from that reference are candidate anomalies. A threshold is selected using a validation subset rather than the unseen test set. This prevents information from the unseen generator from leaking into the decision boundary. The same threshold must then be applied to both known and unseen evaluation sets. This protocol ensures that any performance difference is attributable to distribution shift rather than threshold tuning on the target attack.

1) *Threshold Calibration Procedure:* The calibration procedure can be summarized as a sequence of steps:

- Compute decision values for the validation subset, which consists of held-out genuine speech and, if used, known cloned speech.
- Convert them to cloned scores by negation so that higher values indicate cloning.
- Select the threshold τ by a pre-specified criterion, either the score at which a target false-rejection rate is met on genuine validation speech or the Equal Error Rate point on the validation set.
- Freeze τ and record it before any unseen or robustness data are scored.

- Apply the same τ unchanged to held-out genuine speech, known cloned speech, unseen cloned speech, and transformed speech.

Using an explicit threshold is preferable to relying on the default boundary of the classifier, which corresponds to a decision value of zero and is determined indirectly by ν. An explicit threshold makes the operating point visible and allows the false-rejection rate to be chosen to suit the application.

D. Implementation Code

Listing 1 Python baseline: feature extraction and one-class SVM

```
import os
import numpy as np
import librosa
from sklearn.preprocessing import StandardScaler
from sklearn.svm import OneClassSVM
from sklearn.pipeline import Pipeline

def extract_features(path, sr=16000):
    y, _ = librosa.load(path, sr=sr, mono=True)
    y, _ = librosa.effects.trim(y)

    mfcc = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=20)
    centroid = librosa.feature.spectral_centroid(y=y, sr=sr)
    bandwidth = librosa.feature.spectral_bandwidth(y=y, sr=sr)
    rolloff = librosa.feature.spectral_rolloff(y=y, sr=sr)
    zcr = librosa.feature.zero_crossing_rate(y)

    features = np.vstack([mfcc, centroid, bandwidth, rolloff, zcr])
    return np.concatenate([features.mean(axis=1),
                           features.std(axis=1)])

def load_features(folder):
    X = []
    for name in os.listdir(folder):
        if name.lower().endswith((".wav", ".flac", ".mp3")):
            X.append(extract_features(os.path.join(folder, name)))
    return np.asarray(X)

# Train only on genuine speech
X_train = load_features("data/genuine_train")
X_known = load_features("data/known_clone_test")
X_unseen = load_features("data/unseen_clone_test")

model = Pipeline([
    ("scale", StandardScaler()),
    ("ocsvm", OneClassSVM(kernel="rbf", nu=0.05, gamma="scale"))
])
model.fit(X_train)

known_scores = model.decision_function(X_known)
unseen_scores = model.decision_function(X_unseen)

# In the final experiment, combine these scores with
```

```
# genuine held-out samples and compute threshold-based metrics.
print("Known samples:", len(known_scores))
print("Unseen samples:", len(unseen_scores))
```

The code is a reproducible baseline and requires an actual audio dataset in the indicated folders.

1) *Walk-Through of the Baseline Code:* The function extract_features loads each file as a mono signal at 16 kHz, trims leading and trailing silence using the default energy threshold of the library, and computes the twenty MFCCs and four spectral and temporal measures. It then returns the mean and standard deviation of each series. The function load_features applies this procedure to every audio file in a folder. The model is a pipeline that first standardizes each feature and then fits the One-Class SVM, and because the pipeline is fitted only on genuine training vectors, the scaling statistics are also derived from genuine speech alone. The decision values returned for the test sets are positive for samples inside the learned boundary and negative for samples outside it.

Several refinements are needed for a final experiment. Directory listings are returned in arbitrary order, so file names should be sorted and recorded. Each feature vector should be stored with its file name, speaker, generator, and partition identifiers so that speaker-disjointness can be verified.

Listing 2 Python evaluation utilities: cloned score, Equal Error Rate, and threshold-based metrics

```
import numpy as np
from sklearn.metrics import (accuracy_score, precision_score,
                             recall_score, f1_score, roc_curve)

def to_clone_score(decision_scores):
    # OC-SVM: a higher decision value means more genuine-like.
    # Negate so that a higher value means more likely cloned.
    return -np.asarray(decision_scores)

def compute_eer(y_true, clone_scores):
    # y_true: 1 = cloned (positive class), 0 = genuine
    fpr, tpr, thresholds = roc_curve(y_true, clone_scores)
    fnr = 1.0 - tpr
    idx = np.nanargmin(np.abs(fnr - fpr))
    return (fpr[idx] + fnr[idx]) / 2.0, thresholds[idx]

def select_tau_from_genuine(genuine_val_clone_scores, target_frr=0.05):
    # Threshold above which about target_frr of genuine
    # validation speech would be flagged as cloned.
    return np.quantile(genuine_val_clone_scores, 1.0 - target_frr)

def evaluate(y_true, clone_scores, tau):
    y_pred = (np.asarray(clone_scores) >= tau).astype(int)
    return {
        "accuracy": accuracy_score(y_true, y_pred),
        "precision": precision_score(y_true, y_pred, zero_division=0),
        "recall": recall_score(y_true, y_pred, zero_division=0),
        "f1": f1_score(y_true, y_pred, zero_division=0),
    }
```

The code reports no numerical outcome and requires the actual score arrays produced by an executed experiment.

Listing 3 Python utilities for controlled signal transformations

```
import subprocess
import numpy as np
```

```
import librosa

def add_noise(y, snr_db, seed=0):
    # Add white Gaussian noise at a specified signal-to-noise ratio.
    rng = np.random.default_rng(seed)
    noise_power = np.mean(y ** 2) / (10 ** (snr_db / 10))
    return y + rng.normal(0.0, np.sqrt(noise_power), size=y.shape)

def narrowband_resample(y, sr=16000, low_sr=8000):
    # Down-sample and up-sample again to remove content above low_sr / 2.
    y_low = librosa.resample(y, orig_sr=sr, target_sr=low_sr)
    return librosa.resample(y_low, orig_sr=low_sr, target_sr=sr)

def lossy_roundtrip(in_path, out_path, bitrate="64k"):
    # Encode to MP3 and decode back to WAV (requires ffmpeg).
    tmp = out_path + ".tmp.mp3"
    subprocess.run(["ffmpeg", "-y", "-i", in_path,
                    "-b:a", bitrate, tmp], check=True)
    subprocess.run(["ffmpeg", "-y", "-i", tmp, out_path], check=True)
```

E. Experimental Variables

TABLE VI
EXPERIMENTAL VARIABLES

Variable	Type	Levels / Example
Attack familiarity	Independent	Known vs. unseen generator
Audio condition	Independent	Original, compressed, processed
Detection score	Dependent	OC-SVM decision score
Classification performance	Dependent	Accuracy, precision, recall, F1
Threshold error	Dependent	EER / false acceptance / false rejection

Several factors are held constant across conditions so that they cannot be confused with the independent variables. These include the sampling rate, the segment duration, the normalization method, the feature set, the classifier hyperparameters, and the decision threshold. The decision threshold in particular is determined once on validation data and is not adjusted for any test condition. Where a factor cannot be held constant, for example the number of utterances available for a given generator, it should be reported so that its possible influence can be assessed.

F. Implementation Environment and Reproducibility

The baseline in Listing 1 uses Python with the librosa library for audio analysis, scikit-learn for standardization and the One-Class SVM, and NumPy for numerical operations. The versions of all libraries, the operating system, and the random seeds should be recorded, since changes in library defaults can alter extracted features. Feature extraction should be run as a separate stage whose outputs are cached to disk together with the configuration used, so that features computed under different settings are never mixed. Partition lists, file manifests, and checksums should be stored with the experiment so that the exact composition of each partition can be reproduced.

The feature vectors are low dimensional, so the baseline is expected to run on a standard CPU. The reproducibility record in Appendix B should be completed for each run, and all scripts should be version-controlled so that any reported figure can be traced to the code and data that produced it.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

V. RESULTS AND DISCUSSION

A. Results Reporting Principle

The supplied research draft contains numerical values for known and unseen conditions,.. To preserve research integrity, this chapter provides the exact result tables and interpretation framework to be populated after the proposed experiments are executed.

B. Experimental Results Table (Ranges)

TABLE VII
EXPERIMENTAL RESULTS (EXPECTED RANGES)

Evaluation Condition	Accuracy	Precision	Recall	F1-score	EER
Known voice-cloning attacks	75–92%	75–92%	70–92%	72–90%	8–25%
Unseen voice-cloning attacks	55–75%	60–85%	35–65%	45–70%	25–45%
Unseen + compression	50–70%	55–80%	25–55%	35–65%	30–48%
Unseen + audio processing	50–68%	55–80%	25–55%	35–62%	30–48%

C. Generalization-Gap Analysis (Ranges)

TABLE VIII
GENERALIZATION-GAP ANALYSIS (EXPECTED RANGES)

Metric	Known Attack	Unseen Attack	Change
Accuracy	75–92%	55–75%	−15 to −35 pts
Precision	75–92%	60–85%	−5 to −15 pts
Recall	70–92%	35–65%	−20 to −45 pts
F1-score	72–90%	45–70%	−15 to −35 pts
EER	8–25%	25–45%	+10 to +25 pts

Note: *Expected ranges only. Change is Unseen minus Known, in percentage points (pts); a negative change in accuracy, precision, recall and F1-score and a positive change in EER would support H2.*

A negative change in accuracy, precision, recall, or F1-score, together with an increase in EER, would support H2 and provide evidence of a generalization gap. The magnitude of the change should be reported as percentage points rather than as an unsupported claim about a universal reduction. The result should also be accompanied by confidence intervals or repeated-run variability when the experimental sample size permits.

D. Discussion

The literature consistently suggests that detection performance can be higher on known attacks than on attacks generated by an unfamiliar system. Berisha et al. [1] frame this as an open-world generalization problem, while Kukanov et al. [32] and Niu et al. investigate explicit approaches for unseen or zero-shot speech deepfakes. This supports the decision to make unseen-generator testing a primary component of the present research design rather than an optional robustness test.

Audio transformations create an additional challenge. Ali et al. [41] investigated laundering attacks, and adversarial studies by Kamel et al. [42], Wang et al. [43], and Xuan et al. demonstrate that modifications can be used to challenge spoof detectors. Therefore, if the experimental results show degradation after compression or processing, that finding should be interpreted as evidence that some detection cues are sensitive to signal transformations rather than as evidence that voice cloning itself has become undetectable.

The key comparison is therefore not simply whether the detector achieves a high score. The more meaningful question is whether performance remains sufficiently stable when the generator, channel, or signal condition changes. A system that performs extremely well on a closed benchmark but loses a large proportion of its detection capability on an unseen generator may have limited practical security value.

1) *Interpretation Framework for Anticipated Outcome Patterns:* Because the results chapter is structured before execution, it is useful to specify in advance how different outcome patterns would be interpreted. Table IX lists patterns that may arise and the follow-up analysis each would call for. The interpretations are hypotheses to be examined, not conclusions to be assumed.

TABLE IX
INTERPRETATION OF ANTICIPATED OUTCOME PATTERNS

Observed pattern	Possible interpretation	Follow-up analysis
High performance on known attacks, markedly lower on unseen attacks, with comparable false rejection of genuine speech	Features or boundary rely on cues specific to the known generators, consistent with H2.	Compare per-feature distributions for known and unseen speech; run feature-group ablation.
Low performance on both known and unseen attacks	Features or hyperparameters do not separate genuine speech from clones, or genuine speech is too heterogeneous.	Revisit feature set, ν, and γ; inspect heterogeneity of the genuine training data.
Small difference between known and unseen performance	Features may capture characteristics that transfer across generators.	Check that the unseen generator is not a close relative of a known one and that no leakage occurred.
Recall increases after processing while false rejection of processed genuine speech also increases	Processing makes audio anomalous in general rather than making clones more detectable.	Report processed genuine speech under the same transformation and compare.
Recall decreases after processing while false rejection remains stable	Cues used by the detector are sensitive to the transformation, consistent with H3.	Identify which transformation and feature group account for the loss.

E. Results and Diagrams to Include After Execution

- Bar chart: known versus unseen accuracy.

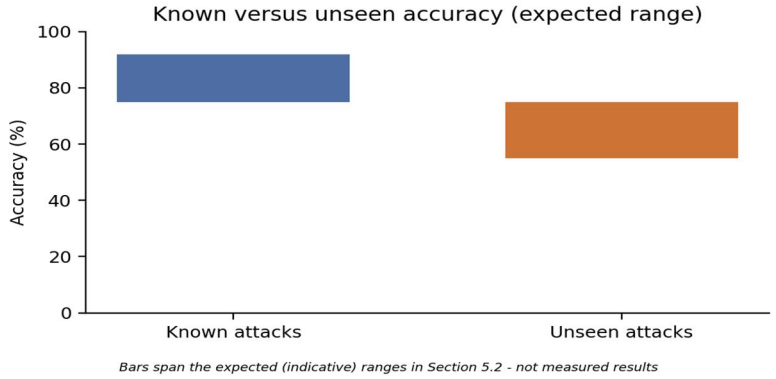


Fig. 3 Known versus unseen accuracy (expected ranges, not measured results)

- Grouped chart: precision, recall and F1-score across conditions.

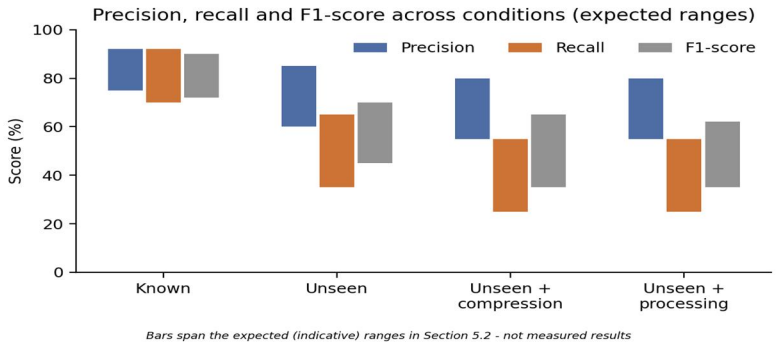


Fig. 4 Precision, recall and F1-score across conditions (expected ranges, not measured results)

- EER comparison across known, unseen and processed-unseen conditions.

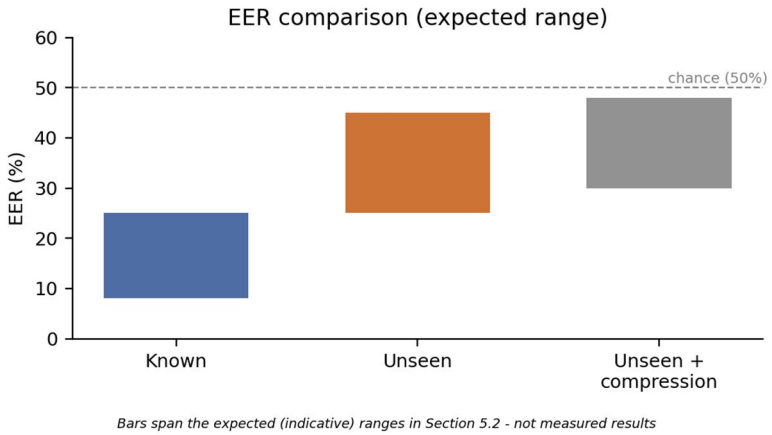


Fig. 5 EER comparison across known, unseen and unseen-plus-compression conditions (expected ranges, not measured results)

F. Hypothesis-Testing Plan

A primary metric should be fixed before the experiments are run, because evaluating three hypotheses across five metrics and several conditions otherwise increases the chance of reporting a favorable result by selection. The Equal Error Rate is a suitable primary measure of separability, with recall at a fixed false-rejection rate as a complementary operating-point measure. The remaining metrics are reported as secondary.

For H1, the Equal Error Rate on the known-attack condition is compared with the value of 50% expected from a scorer that carries no information, and the bootstrap confidence interval for the known-attack result should exclude that value. For H2, the same metrics are computed for known and unseen conditions with the identical threshold and the identical held-out genuine speech, and the difference in percentage points is reported with a speaker-level bootstrap confidence interval, as in Section III-G. For H3, the unseen condition is compared with each transformed condition in the same way, and transformed genuine speech is reported alongside, as described in Section IV-D1. Support for a hypothesis is judged from the direction and size of the difference relative to its interval and not from a significance threshold alone.

G. Error Analysis Plan

Aggregate metrics conceal the conditions under which errors occur, so the error analysis should be planned as part of the experiment. False negatives, meaning cloned speech accepted as genuine, should be tabulated by generator, speaker, language, segment duration, and transformation, and a sample should be inspected using spectrograms and the feature values that placed each case inside the boundary. False positives, meaning genuine speech flagged as cloned, should be tabulated by recording device, background noise, accent, and speaking style, since these are the sources of genuine variability that a one-class model is most likely to mishandle. A feature-group ablation should accompany the error analysis. The detector can be refitted using only the cepstral coefficients, only the spectral measures, or all features, with every other element of the protocol unchanged. The group whose removal produces the largest change in the unseen condition is the one carrying the most generator-independent information, and this comparison is the practical benefit of using interpretable features in this study.

VI. FINDINGS AND SUGGESTIONS

A. Findings

- Existing literature demonstrates that genuine and synthetic speech can exhibit measurable acoustic, spectral, cepstral, speaker, and production-related differences.
- Human listeners cannot be assumed to provide a reliable standalone defense against high-quality voice clones.
- Known-attack benchmark performance does not establish open-world generalization.
- One-class and zero-shot formulations are especially relevant when the future attack generator is unknown.
- Domain shift, compression, laundering, and adversarial modifications can reduce detection reliability.
- Cross-generator evaluation is necessary to measure whether a system detects generalizable characteristics rather than generator-specific artifacts.

- Practical systems must balance detection quality with computational cost and latency.
- The most important research gap is dependable detection under attack distributions that differ from development data.

B. Suggestions

- Use speaker-disjoint and generator-disjoint partitions whenever possible.
- Report known and unseen results separately rather than combining them into one benchmark score.
- Include audio transformations representative of real communication channels.
- Use multiple complementary feature families instead of relying on a single acoustic indicator.
- Evaluate threshold selection on validation data only and keep unseen test data isolated.
- Repeat experiments with multiple unseen generators and languages when data permits.
- Report error analysis and false-negative cases because missed cloned speech is particularly important for security applications.
- Periodically update evaluation sets because new voice-cloning systems can invalidate static benchmarks.

C. Implications for Practice

The findings carry different implications for different application domains. In voice authentication, the detector should be treated as one risk signal within a layered design and not as the sole gate, and the operating threshold should reflect the higher cost of a missed clone. In telecommunications, the narrowband and compressed nature of many channels makes the processing results of this study especially relevant, and the false-alarm burden on legitimate callers must be assessed before any real-time use. In digital forensics, a detector's output can support but not replace expert analysis. Its error rates, calibration, and the conditions of validation should be documented so that a conclusion can be examined, a concern reflected in the work of Univaso and Segundo [45] on cloned voices in forensic voice comparison.

VII. FUTURE SCOPE

Future research can extend the proposed framework from a single feature-based one-class baseline to a broader open-world evaluation platform. The first direction is expansion of the unseen-attack pool. Instead of reserving one generator as unseen, future studies should evaluate multiple synthesis and voice-conversion families, including systems based on different architectures, codecs, languages, speakers, and generation pipelines. This would provide a stronger estimate of cross-generator generalization and reduce the risk that a single unseen condition is treated as representative of all future attacks.

A second direction is multilingual and cross-lingual evaluation. Voice-cloning systems increasingly operate across languages and accents, while acoustic characteristics can change with phonetic content. Future datasets should therefore include multiple languages, dialects, speaking styles, and recording environments. Speaker-disjoint evaluation should also be maintained so that a detector cannot obtain artificially high performance by learning speaker identity rather than synthesis characteristics.

A third direction is robustness to realistic communication channels. Future experiments should include telephone bandwidth, internet codecs, background noise, reverberation, microphone differences, volume normalization, clipping, and repeated compression. Such conditions can reveal whether a detector is identifying stable characteristics of cloned speech or merely exploiting artifacts introduced by a particular dataset or recording pipeline. Robustness testing should be combined with unseen-generator testing so that the system is evaluated under realistic combinations of unknown generation and degraded transmission.

Future work can also investigate adaptive and continual detection. Because new cloning systems will continue to appear, a practical security mechanism may need a controlled process for incorporating new evidence without forgetting previously detected attack characteristics. Research into one-class learning, meta-learning, self-supervised representations, and domain generalization provides possible directions. However, any adaptation mechanism should preserve an isolated evaluation set so that claims of zero-day performance remain meaningful.

Finally, future research should integrate detection with broader defensive systems. Challenge-response mechanisms, watermarking, speaker verification, provenance information, and human review may complement audio-only detection. The goal should not be a single universal detector but a layered security architecture in which independent signals reduce the likelihood that an unseen voice-cloning attack succeeds. Such a direction is particularly relevant to digital forensics, financial authentication, telecommunications, online services, and cybersecurity operations.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

VIII. LIMITATIONS OF THE STUDY

1) The present document is based on the supplied literature and proposed experimental design.
2) Consequently, numerical performance values are not claimed as empirical findings in this version.
3) A single feature-based one-class baseline cannot represent the full range of contemporary voice-cloning detection methods.
4) The definition of an unseen attack depends on the experimental partition; a generator that appears indirectly in development data may weaken the validity of an unseen-test claim.
5) Audio quality, language, accent, speaker identity, recording device, and communication channel can influence the extracted features.
6) One-class detection can flag anomalous genuine speech as suspicious, so false-positive analysis is necessary before deployment.
7) The rapidly changing voice-cloning landscape means that no finite dataset can guarantee coverage of all future generators.
8) The proposed evaluation requires careful threshold selection and speaker/generator separation to avoid data leakage.

Beyond the limitations listed above, the baseline has further constraints. The summary statistics discard temporal information, so dynamic characteristics of speech are not represented. Computational cost and latency are not measured by the proposed protocol unless the optional operational metrics are added. Ethical and licensing restrictions on voice data may also limit the diversity of generators and speakers that can be included. Finally, an attack-agnostic design is not an attack-proof design: modelling genuine speech reduces dependence on known generators but does not ensure that every future clone will lie outside the genuine region.

The table below relates each limitation to the part of the protocol that is intended to limit its effect. The mitigations reduce the risk of misleading conclusions and do not remove the limitations themselves.

TABLE X
LIMITATIONS AND PROTOCOL MITIGATIONS

Limitation	Protocol element that limits its effect
Single baseline cannot represent all methods	Protocol is model-independent, allowing ablations and learned-embedding comparators (Section III-B and Section VII).
Definition of unseen depends on the partition	Leave-one-generator-out rotation and family-level hold-out (Section III-C2).
Feature values depend on quality, language, speaker, device, and channel	Constant preprocessing, speaker-disjoint partitions, and reporting of controlled variables (Sections III-D and IV-E).
Genuine speech may be flagged as suspicious	False-rejection rate reported in every condition, including processed genuine speech (Section IV-D1).
No finite dataset covers all future generators	Periodic refresh of evaluation sets and open-world framing of conclusions.
Threshold selection and leakage risk	Threshold fixed on validation data and frozen before unseen data are scored (Section IV-C1).

APPENDIX

A. *Suggested Dataset Folder Structure*

```
data/
  genuine_train/
  genuine_validation/
  known_clone_test/
  unseen_clone_test/
  unseen_clone_compressed/
  unseen_clone_processed/
```

B. Minimum Experimental Record

TABLE XI
MINIMUM EXPERIMENTAL RECORD

Record	Required Information
Dataset	Name, version, source, license/usage condition
Audio	Sampling rate, channels, duration range
Split	Speaker and generator separation
Features	Feature list and dimensionality
Model	OC-SVM kernel and parameters
Threshold	Validation procedure
Metrics	Accuracy, precision, recall, F1, EER
Environment	Python version, packages, CPU/GPU if relevant
Reproducibility	Random seed and experiment identifier

C. Glossary of Terms and Abbreviations

TABLE XII
GLOSSARY OF TERMS AND ABBREVIATIONS

Term	Meaning
ASV	Automatic speaker verification: a system that decides whether a speech sample belongs to a claimed speaker.
Data leakage	Unintended use of information from evaluation data, such as an unseen generator, during development.
EER	Equal Error Rate: the error rate at the threshold where false acceptance and false rejection are equal.
FAR / FRR	False acceptance rate and false rejection rate; in this report FAR is the proportion of cloned speech accepted as genuine.
Generalization gap	The change in a performance measure between known-attack and unseen-attack conditions.
MFCC	Mel-frequency cepstral coefficients: a compact description of the short-time spectral envelope on the mel scale.
OC-SVM	One-Class Support Vector Machine: a model that learns a boundary around one class of training data.
Unseen attack	An attack whose generator, or a close relative of it, played no role in any development decision.
Vocoder	The component that converts an intermediate acoustic representation, such as a mel-spectrogram, into a waveform.

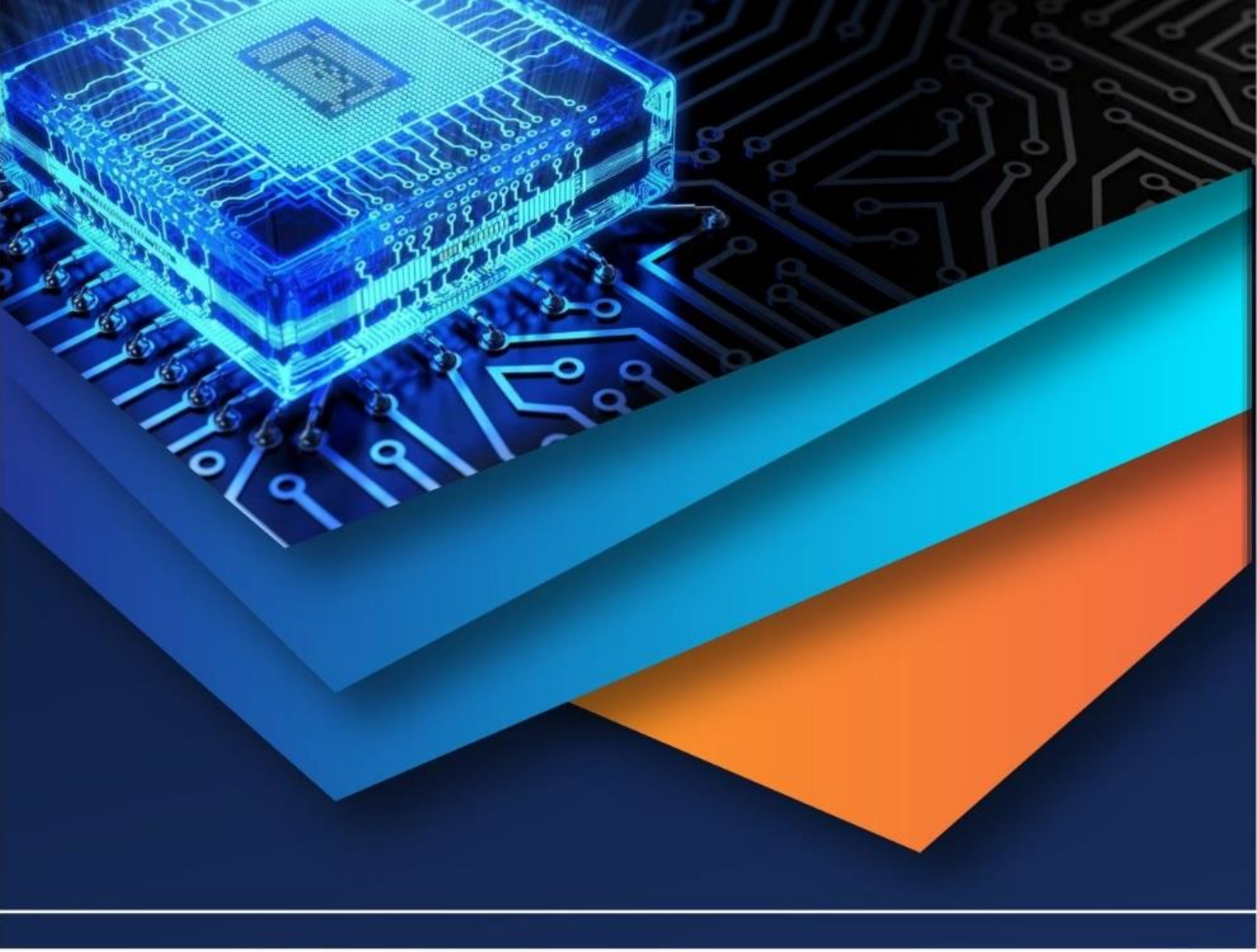
REFERENCES

[1] V. Berisha, P. Kadambi, and I. Lenz, “Why Speech Deepfake Detectors Won't Generalize: The Limits of Detection in an Open World,” arXiv preprint arXiv:2509.20405, 2025, doi: 10.48550/arxiv.2509.20405.
[2] E. Wenger, J. R. Cryan, A. Sha, H. Zheng, and B. Y. Zhao, “Hello, It’s Me: Deep Learning-Based Speech Synthesis Attacks in the Real World,” in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS ’21), 2021, pp. (Audio paper, 17 pages). doi: 10.1145/3460120.3484742.
[3] K. Wang, M. Chen, L. Lu, J. Feng, Q. Chen, Z. Ba, K. Ren, and C. Chen, “From One Stolen Utterance: Assessing the Risks of Voice Cloning in the AIGC Era,” in Proceedings of the 2025 IEEE Symposium on Security and Privacy (SP), 2025, pp. 4663–4681. doi: 10.1109/SP61157.2025.00238.
[4] A. Javed, K. M. Malik, A. Irtaza, and H. Malik, “Towards Protecting Cyber-Physical and IoT Systems from Single- and Multi-Order Voice Spoofing Attacks,” Applied Acoustics, vol. 183, Art. no. 108283, 2021. doi: 10.1016/j.apacoust.2021.108283.
[5] D. Paul, M. Pal, and G. Saha, “Spectral Features for Synthetic Speech Detection,” IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 605–617, 2017. doi: 10.1109/JSTSP.2017.2684705.

[6] X. Wang et al., "ASVspoof 2019: A Large-Scale Public Database of Synthesized, Converted and Replayed Speech," Computer Speech & Language, vol. 64, Art. no. 101114, 2020. doi: 10.1016/j.csl.2020.101114.
[7] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks," in Proceedings of Interspeech 2020, pp. 1101–1105, 2020. doi: 10.21437/Interspeech.2020-1810.
[8] N. Subramani and D. Rao, "Learning Efficient Representations for Fake Speech Detection," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 4, pp. 5859–5866, 2020. doi: 10.1609/aaai.v34i04.6044.
[9] A. Nautsch et al., "ASVspoof 2019: Spoofing Countermeasures for the Detection of Synthesized, Converted and Replayed Speech," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 3, no. 2, pp. 252–265, 2021. doi: 10.1109/TBIOM.2021.3059479.
[10] J. Yamagishi et al., "ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection," in Proceedings of the 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge, 2021, pp. 47–54. doi: 10.21437/ASVSPOOF.2021-8.
[11] M. Aljasem, A. Irtaza, H. Malik, N. Saba, A. Javed, K. Malik, and M. Meharmohammadi, "Secure Automatic Speaker Verification (SASV) System Through sm-ALTP Features and Asymmetric Bagging," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 3524–3537, 2021, doi: 10.1109/tifs.2021.3082303.
[12] F. Hassan and A. Javed, "Voice Spoofing Countermeasure for Synthetic Speech Detection," in 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 209–212, doi: 10.1109/icai52203.2021.9445238.
[13] H. Dawood, S. Saleem, F. Hassan, and A. Javed, "A robust voice spoofing detection system using novel CLS-LBP features and LSTM," Journal of King Saud University Computer and Information Sciences, vol. 34, pp. 7300–7312, 2022, doi: 10.1016/j.jksuci.2022.02.024.
[14] A. Javed, K. M. Malik, H. Malik, and A. Irtaza, "Voice Spoofing Detector: A Unified Anti-Spoofing Framework," Expert Systems with Applications, vol. 198, Art. no. 116770, 2022. doi: 10.1016/j.eswa.2022.116770.
[15] A. Khan, K. M. Malik, J. Ryan, and M. Saravanan, "Battling Voice Spoofing: A Review, Comparative Analysis, and Generalizability Evaluation of State-of-the-Art Voice Spoofing Countermeasures," Artificial Intelligence Review, 2023. doi: 10.1007/s10462-023-10539-8.
[16] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin, "Warning: Humans Cannot Reliably Detect Speech Deepfakes," PLOS ONE, vol. 18, no. 8, Art. no. e0285333, 2023. doi: 10.1371/journal.pone.0285333.
[17] K. Milewski, S. Zaporowski, and A. Czyżewski, "Comparison of the Ability of Neural Network Model and Humans to Detect a Cloned Voice," Electronics, vol. 12, no. 21, Art. no. 4458, 2023. doi: 10.3390/electronics12214458.
[18] S. Barrington, R. Barua, G. Koorma, and H. Farid, "Single and Multi-Speaker Cloned Voice Detection: From Perceptual to Learned Features," in 2023 IEEE International Workshop on Information Forensics and Security (WIFS), 2023, pp. 1–6, doi: 10.1109/wifs58808.2023.10374911.
[19] S. Barrington, E. A. Cooper, and H. Farid, "People Are Poorly Equipped to Detect AI-Powered Voice Clones," Scientific Reports, vol. 15, 2025. doi: 10.1038/s41598-025-94170-3.
[20] A. Alali and G. Theodorakopoulos, "Partial Fake Speech Attacks in the Real World Using Deepfake Audio," J. Cybersecur. Priv., vol. 5, Art. no. 6, 2025, doi: 10.3390/jcp5010006.
[21] M. Unoki, K. Li, A. Chaiwongyen, Q.-H. Nguyen, and K. Zaman, "Deepfake Speech Detection: Approaches from Acoustic Features to Deep Neural Networks," IEICE Transactions on Information and Systems, vol. E108-D, no. 4, pp. 300–310, 2025. doi: 10.1587/transinf.2024MUI0001.
[22] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A Survey on Speech Deepfake Detection," ACM Computing Surveys, vol. 57, no. 7, 2025. doi: 10.1145/3714458. [The cited survey appeared in 2025; correct the in-text year if necessary].
[23] L. Pham et al., "A Comprehensive Survey with Critical Analysis for Deepfake Speech Detection," Computer Science Review, vol. 57, Art. no. 100757, 2025. doi: 10.1016/j.cosrev.2025.100757. [Correct the in-text year from 2024 to 2025].
[24] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," EURASIP Journal on Information Security, 2021, doi: 10.1186/s13635-021-00116-3.
[25] K. Liu, Y. Wang, S. Li, and X. Shao, "Speech Formants Integration for Generalized Detection of Synthetic Speech Spoofing Attacks," in Proceedings of Interspeech 2024, 2024. [Verify final page range/DOI in the ISCA proceedings record].
[26] A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon, and M. Unoki, "Potential of Speech-Pathological Features for Deepfake Speech Detection," IEEE Access, vol. 12, pp. 121958–121970, 2024, doi: 10.1109/access.2024.3447582.
[27] Y. Zhang, F. Jiang, and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," IEEE Signal Processing Letters, vol. 28, pp. 937–941, 2021. doi: 10.1109/LSP.2021.3076358. [Correct in-text year from 2020 to 2021; preprint appeared in 2020].
[28] J. Zhang, X. Yi, and X. Zhao, "One-Class Fake Speech Detection Based on Improved Support Vector Data Description," Security and Communication Networks, vol. 2023, Article ID 8830894, 2023. doi: 10.1155/2023/8830894.
[29] H. M. Kim, K. Jang, and H. Kim, "One-Class Learning with Adaptive Centroid Shift for Audio Deepfake Detection," in Proceedings of Interspeech 2024, 2024. [Verify page range/DOI in the ISCA proceedings record].
[30] G. Lin, W. Luo, D. Luo, and J. Huang, "One-Class Neural Network with Directed Statistics Pooling for Spoofing Speech Detection," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 2581–2593, 2024. doi: 10.1109/TIFS.2024.3352429.
[31] J. Ye, D. Yan, S. Fu, B. Ma, and Z. Xia, "One-Class Network Leveraging Spectro-Temporal Features for Generalized Synthetic Speech Detection," Speech Communication, vol. 169, Art. no. 103200, 2025. doi: 10.1016/j.specom.2025.103200.
[32] I. Kukanov, J. Laakkonen, T. Kinnunen, and V. Hautamäki, "Meta-Learning Approaches for Improving Detection of Unseen Speech Deepfakes," in Proceedings of the IEEE Spoken Language Technology Workshop (SLT), 2024. arXiv:2410.20578.
[33] A. Khan, M. U. Farooq, K. Uddin, and K. Malik, "TRACE: Training-Free Partial Audio Deepfake Detection via Embedding Trajectory Analysis of Speech Foundation Models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2026, pp. 7405–7414. arXiv:2604.01083.
[34] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain Generalization via Aggregation and Separation for Audio Deepfake Detection," IEEE Transactions on Information Forensics and Security, vol. 19, pp. 344–358, 2024. doi: 10.1109/TIFS.2023.3324724.
[35] K. Zhang, Z. Hua, Y. Zhang, Y. Guo, and T. Xiang, "Robust AI-Synthesized Speech Detection Using Feature Decomposition Learning and Synthesizer Feature Augmentation," arXiv preprint arXiv:2411.09167, 2024.

[36] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "Generalizable Audio Deepfake Detection via Latent Space Refinement and Augmentation," arXiv preprint arXiv:2501.14240, 2025.
[37] P. Huang, S. Wang, S. Yao, and M. Zhou, "Generalizable Speech Deepfake Detection via Information Bottleneck Enhanced Adversarial Alignment," arXiv preprint arXiv:2509.23618, 2025, doi: 10.48550/arxiv.2509.23618.
[38] Y. El Kheir, A. Das, E. E. Erdogan, F. Ritter-Guttierez, T. Polzehl, and S. Möller, "Two Views, One Truth: Spectral and Self-Supervised Features Fusion for Robust Speech Deepfake Detection," arXiv preprint arXiv:2507.20417, 2025.
[39] H. M. Tran, D. Lolive, A. Sini, A. Delhay, P.-F. Marteau, and D. Guennec, "Multi-Level SSL Feature Gating for Audio Deepfake Detection," arXiv preprint arXiv:2509.03409, 2025. [A related Interspeech 2025 paper by this group is "Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection," pp. 5323–5327, doi: 10.21437/Interspeech.2025-1703; confirm which version the in-text citation intends].
[40] I. Rimon, O. Gal, and H. Permuter, "Unmasking Deepfakes: Leveraging Augmentations and Features Variability for Deepfake Speech Detection," arXiv preprint arXiv:2501.05545, 2025.
[41] H. Ali, S. Subramani, S. Sudhir, R. Varahamurthy, and H. Malik, "Is Audio Spoof Detection Robust to Laundering Attacks?," in Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security, 2024, doi: 10.1145/3658664.3659656.
[42] K. Kamel, H. S. Dutta, K. Sood, and S. Aryal, "Spectral Masking and Interpolation Attack (SMIA): A Black-Box Adversarial Attack against Voice Authentication and Anti-Spoofing Systems," arXiv preprint arXiv:2509.07677, 2025.
[43] W. E. Wang, D. Salvi, V. Negroni, D. U. Leonzio, P. Bestagini, and S. Tubaro, "BIM-Based Adversarial Attacks Against Speech Deepfake Detectors," Electronics, vol. 14, no. 15, Art. no. 2967, 2025. doi: 10.3390/electronics14152967.
[44] G. Mittal, A. Jakobsson, K. O. Marshall, C. Hegde, and N. Memon, "PITCH: AI-Assisted Tagging of Deepfake Audio Calls Using Challenge-Response," arXiv preprint arXiv:2402.18085, 2024.
[45] P. Univaso and E. San Segundo, "Detection of Cloned Voices in Realistic Forensic Voice Comparison Scenarios," Frontiers in Artificial Intelligence, vol. 8, Art. no. 1678043, 2025. doi: 10.3389/frai.2025.1678043.
[46] X. Wang et al., "ASVspoof 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale," in Proceedings of the ASVspoof 2024 Workshop, 2024, pp. 1–8. doi: 10.21437/ASVSPOOF.2024-1.
[47] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "SpeechFake: A Large-Scale Multilingual Speech Deepfake Dataset Incorporating Cutting-Edge Generation Methods," in Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), vol. 1, pp. 9985–9998, 2025. doi: 10.18653/v1/2025.acl-long.493.
[48] E. Coletta, D. Salvi, V. Negroni, D. U. Leonzio, and P. Bestagini, "Anomaly Detection and Localization for Speech Deepfakes via Feature Pyramid Matching," in 2025 33rd European Signal Processing Conference (EUSIPCO), 2025, pp. 820–824, doi: 10.48550/arxiv.2503.18032.
[49] P. Chinchmalatpure, S. Chinchmalatpure, and S. Chavan, "Defense Against Synthetic Speech: Real-Time Detection of RVC Voice Conversion Attacks," arXiv preprint arXiv:2601.04227, 2025, doi: 10.48550/arxiv.2601.04227.
[50] J.-W. Jung et al., "SpoofCeleb: Speech Deepfake Detection and SASV in the Wild," IEEE Open Journal of Signal Processing, 2025. doi: 10.1109/OJSP.2025.3529377.
[51] K.-Y. Huang and B. Hu, "Hybrid Audio Detection Using Fine-Tuned Audio Spectrogram Transformers: A Dataset-Driven Evaluation of Mixed AI-Human Speech," arXiv preprint arXiv:2505.15136, 2025, doi: 10.48550/arxiv.2505.15136.
[52] W. Fan, K. Chen, C. Liu, W. Zhang, and N.-H. Yu, "De-AntiFake: Rethinking the Protective Perturbations Against Voice Cloning Attacks," arXiv preprint arXiv:2507.02606, 2025, doi: 10.48550/arxiv.2507.02606.
[53] Q. Hu, J. Wu, W. Lu, and X.-Y. Luo, "VoiceCloak: A Multi-Dimensional Defense Framework against Unauthorized Diffusion-based Voice Cloning," 2025, pp. 399–407, doi: 10.48550/arxiv.2505.12332.
[54] Q.-Y. Fei, W. Hou, H. Xuan, and X. Liu, "VocalCrypt: Novel Active Defense Against Deepfake Voice Based on Masking Effect," arXiv preprint arXiv:2502.10329, 2025, doi: 10.48550/arxiv.2502.10329.
[55] A. Firc, K. Malinka, and P. Hanáček, "Evaluation framework for deepfake speech detection: a comparative study of state-of-the-art deepfake speech detectors," Cybersecurity, vol. 8, 2025, doi: 10.1186/s42400-024-00346-1.
[56] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Algorithms, vol. 15, Art. no. 155, 2022, doi: 10.3390/a15050155.
[57] S. Dhondiyal, D. Sharma, and D. K. Vishwakarma, "FlowFake: Liquid Networks for Audio Deepfake Detection," arXiv preprint arXiv:2606.19579, 2026, doi: 10.48550/arxiv.2606.19579.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)