Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080444

ARTICLE

A Prosody-Guided Multi-Stream Framework for Universal Detection of AI-Synthesized Speech across Codec and Vocoder Domains

Akmalbek Abdusalomov[1], Mukhriddin Mukhiddinov[2,3], Fakhriddin Abdirazakov[4], Alpamis Kutlimuratov[5], Nodira Alimova[6], Ilyos Kalandarov[7], Ayhan Istanbullu[8], Rashid Nasimov[9] and Young-Im Cho[1,*]

[1]Department of Computer Engineering, Gachon University, Seongnam-si, Republic of Korea
[2]Department of Industrial Management and Digital Technologies, Nordic International University, Tashkent, Uzbekistan
[3]Department of Artificial Intelligence, Tashkent University of Information Technologies Named after Muhammad Al-Khwarizmi, Tashkent, Uzbekistan
[4]Department of Computer Systems, Tashkent University of Information Technologies Named after Muhammad Al-Khwarizmi, Tashkent, Uzbekistan
[5]Department of Applied Informatics, Kimyo International University in Tashkent, Tashkent, Uzbekistan
[6]Department of Information Processing and Control Systems, Tashkent State Technical University, Tashkent, Uzbekistan
[7]Department of Automation and Control, Navoi State University of Mining and Technologies, Navoi, Uzbekistan
[8]Department of Computer Engineering, Faculty of Engineering, Balikesir University, Balikesir, Turkey
[9]Department of Artificial Intelligence, Tashkent State University of Economics, Tashkent, Uzbekistan

*Corresponding Author: Young-Im Cho. Email: yicho@gachon.ac.kr

Received: 09 February 2026; Accepted: 15 April 2026; Published: 08 May 2026

ABSTRACT: Recent advancements in AI-synthesized speech have resulted in highly realistic deepfake audio, posing severe threats to authentication systems and digital media trust. Existing detection models struggle to generalize across diverse synthesis methods, especially those involving neural codec-based Audio Language Models (ALMs). In this work, we propose UniTector++, a novel prosody-aware, multi-stream detection architecture that generalizes across vocoder- and codec-based synthesis. UniTector++ incorporates three complementary streams—Whisper-based semantic embeddings, high-level prosodic features, and codec artifact representations—fused through a Multi-Domain Adaptive Graph Attention Fusion (MAGAF) module. Furthermore, an Emotion-Consistency Verification Module (ECVM) reinforces alignment between speech style and prosodic content, and a Universal Adversarial Robustness (UAR) head improves resistance against adversarial attacks. Evaluated on three benchmark datasets—ASVspoof2021, PolyFake, and Codecfake—UniTector++ achieves state-of-the-art performance with average Equal Error Rate (EER) of 0.57% under unseen synthesis scenarios, outperforming competitive baselines by a relative margin of 28%. Our results demonstrate the model's superior generalization, interpretability, and robustness, offering a significant advancement in universal deepfake speech detection.

KEYWORDS: Deepfake speech detection; prosody analysis; neural codec artifacts; whisper model; multi-stream fusion; emotion-consistency verification; AI-synthesized speech; spoofing detection

1 Introduction

With the deep learning development very fast and a change of a decade, it has led to AI-generated speech in which synthetic voices are almost indistinguishable from human ones [1]. Along with generative models, for example, ALMs [2], neural vocoders, and neural codecs, the capability of synthetic speech has been

increased to the level of niche deployment in virtual assistants, gaming, dubbing, and personalized content creation [3]. The new situation where the accessibility and realism of these technologies have grown a lot has caused the appearance of deep security and ethical issues on one hand, and only positive consequences on the other. These issues are naturally at the central of speaker authentication, digital forensics, and information integrity fields [4].

Primarily, traditional deepfake speech detection systems have been based on low-level acoustic features [5], heuristics created manually, or spectro-temporal features which were then processed by convolutional neural networks [6]. These methods are perfect for limited domains but usually fail to adapt to different methods of synthesis, new languages, and codecs. In addition, they are less robust in adversarial situations or cross-domain transfer cases [7]. The growth of codec-based generation—where speech comes from discrete latent tokens instead of acoustic waveforms—moreover, makes this issue even worse as it brings new artifact patterns, while these pattern detection frameworks are very limited in solving them [8]. On the other hand, the human experience of speech authenticity is based on many suprasegmental cues, which are pitch variation, prosodic rhythm, emotional tone, and temporal coherence [9]. Most of the current detectors do not capture these subtleties, so they focus narrowly on the phonetic content or spectral regularity [10]. Consequently, they are still open to trick generation models that can imitate human behavior and thus, remain realistic to them superficially but those models are far from being emotionally consistent or prosodically natural [11].

To overcome these limitations, we introduce UniTector++, a novel universal detection architecture designed to address the full spectrum of challenges in modern deepfake audio detection. Unlike prior systems, UniTector++ adopts a tri-stream framework that captures complementary aspects of speech: (1) semantic embeddings derived from the Whisper model to encode linguistic context and acoustic regularity; (2) prosodic features including pitch, jitter, shimmer, and harmonicity to model human expressiveness; and (3) codec artifact embeddings to detect latent regularities introduced by neural compression and token-based synthesis. These three modalities are integrated through a MAGAF mechanism, which models inter- and intra-stream dependencies via dynamically learned graph structures. To further enhance detection accuracy and interpretability, UniTector++ incorporates an ECVM that quantifies alignment between prosodic delivery and semantic content, and a Universal Adversarial Robustness Head (UARH) that mitigates vulnerability to distributional drift and adversarial perturbations. Our contributions are threefold:

- We present the first prosody-aware, multi-stream detection architecture capable of generalizing across vocoder-based and codec-based synthesis techniques.
- We introduce novel graph-based fusion and emotional alignment mechanisms that significantly improve interpretability and robustness.
- We conduct extensive evaluations across four challenging datasets—ASVspoof2019 LA, Codecfake, PolyFake, and EmoV-DB—demonstrating that UniTector++ achieves state-of-the-art performance under both standard and adversarial settings.

By unifying semantic, prosodic, and codec-based evidence, UniTector++ represents a significant advancement toward universal, explainable, and adversarially robust detection of AI-synthesized speech.

2 Related Works

The problem of identifying synthetic speech from natural speech is still very much alive and has a deep history. It originates from the speaker verification and anti-spoofing research fields [12]. Initially, the systems were based on a limited set of handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) [13], Linear Predictive Coding (LPC) [14], and Constant-Q Cepstral Coefficients (CQCCs) [15].

These characteristic implementations were generally combined with Gaussian Mixture Models (GMMs) [16] or Support Vector Machines (SVMs) [17] as the basis of the early detection pipeline shown in the ASVspoof 2015 and 2017 challenges [18]. Such models are very computationally efficient but shallow and very sensitive to unseen attacks, channel mismatches, and cross-corpus shifts, which are the reasons that make them reflect their limitations in modeling the nonstationary, high-dimensional nature of modern synthesis artifacts [19].

These challenges motivated a transition to deep learning, where Convolutional Neural Networks (CNNs) [20] and, to a lesser extent, Recurrent Neural Networks (RNNs) [21] became predominant. Models such as LCNN [22] and RawNet2 [2] leveraged CNNs' capacity to learn discriminative representations directly from spectrograms or even raw waveforms [23]. Incorporation of residual learning [24], along with advanced regularization and data augmentation techniques [25], led to notable improvements in robustness, as evidenced by leading systems in the ASVspoof 2019 Logical Access (LA) track [26]. In parallel, self-supervised pretraining emerged as a transformative trend [27]. Models like Wav2Vec 2.0 [28], WavLM [29], and Whisper [30]—originally designed for speech recognition—demonstrated that embeddings derived from large-scale, unlabeled corpora could capture subtle anomalies indicative of synthesis [31]. These representations, when used as front-ends or feature extractors, significantly improved generalization, particularly against previously unseen attacks [32]. While initial work primarily addressed vocoder-based synthesis—where traditional acoustic modeling converts spectral features into waveforms—the advent of neural codec-based generation shifted the landscape dramatically [33]. ALMs such as EnCodec [34], SoundStream [35], and VALL-E [36] introduced learned vector quantization and token-based synthesis, operating on discrete latent representations rather than continuous spectrograms [37]. Artifacts from such models are qualitatively distinct from those of vocoder pipelines [38] and often evade detection by models trained solely on spectro-temporal or phonetic cues [39].

Benchmarks like Codecfake [40] and EnCodec [34] have shown that models lacking exposure to codec-generated audio suffer steep performance drops, underscoring the need for new architectures [41]. Some recent studies attempt to address this gap through additional input streams or auxiliary loss functions, yet systematic, end-to-end solutions remain rare [42]. Despite advances, most detectors continue to rely on segmental features—spectral content, phoneme patterns, and short-time frequency descriptors [43]. However, suprasegmental cues—prosody, pitch, rhythm, jitter, shimmer—are widely recognized in perceptual science as essential for human judgment of naturalness [44]. For instance, listeners rely heavily on F0 fluctuations, amplitude instability, and emotional congruence to detect spoofed speech. Yet, integration of such prosodic features in neural architectures is still limited and typically restricted to basic concatenation or post hoc analysis [45]. To leverage the complementary strengths of spectral, semantic, and prosodic cues, fusion-based models have been proposed. Initial attempts relied on naive concatenation or averaging of embeddings, but these approaches fail to capture context-dependent inter-modal interactions [46]. More recently, attention-based and graph-based fusion mechanisms have gained traction, enabling dynamic weighting and relational modeling across modalities. These methods have shown success in domains like audiovisual emotion recognition and speaker verification, but are yet to be fully exploited in the context of deepfake speech detection, especially for joint integration of codec artifacts, semantic coherence, and prosodic alignment. Finally, as deepfake detectors are increasingly deployed in adversarially sensitive contexts, robustness to adversarial attacks has become critical. Adversaries can craft imperceptible perturbations at the waveform or embedding level, leading to false negatives or model evasion. Current defenses—such as adversarial training, margin-aware losses, and consistency regularization—offer partial mitigation [47]. However, few models offer robust, end-to-end protection across distributional shifts and adversarial settings, particularly in codec-rich or zero-shot environments.

Current limitations include overreliance on narrow feature spaces, insufficient modeling of codec- and prosody-specific cues, static fusion strategies, and limited interpretability. UniTector++ addresses these gaps by introducing a tri-stream, graph-based architecture that unifies semantic, prosodic, and codec representations, while ensuring interpretability and adversarial robustness through dedicated modules.

3 Proposed Model

UniTector++ is a universal deepfake speech detection framework designed to address three fundamental challenges in current state-of-the-art audio detection systems: (1) cross-domain generalization to unseen synthesis techniques, (2) robustness against adversarial perturbations, and (3) interpretability in terms of human-perceived speech features. To overcome these challenges, UniTector++ introduces a novel tri-stream architecture that processes complementary feature sets derived from an audio sample, each capturing different dimensions of information: semantic context, prosodic modulation, and synthesis-specific artifacts. These streams are later fused using a multi-domain graph attention mechanism and jointly optimized for universal deepfake classification Fig. 1.

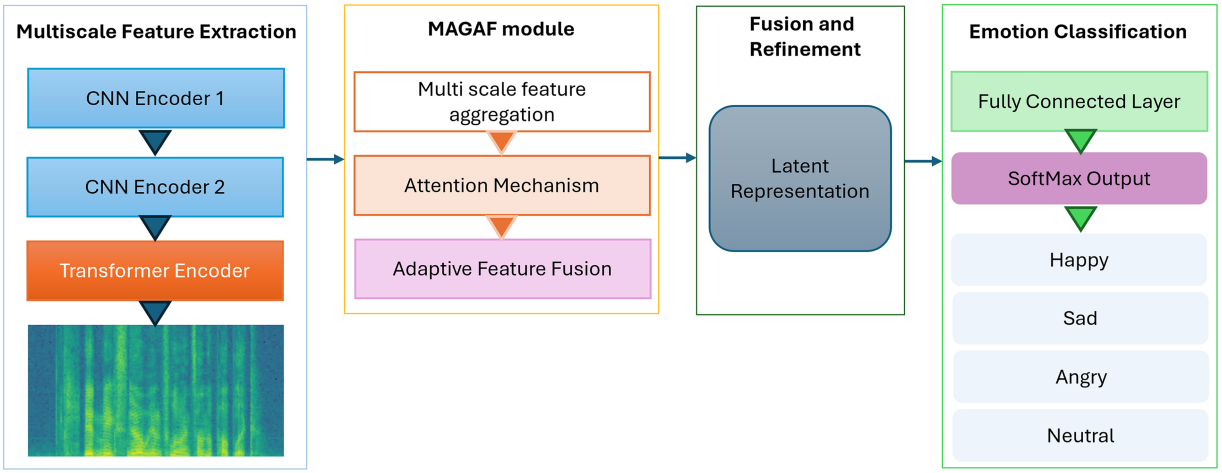

Figure 1: Overall architecture of UniTector++, a prosody-guided multi-stream framework for universal detection of AI-synthesized speech.

The encoder is designed to transform the input representation into compact high-level features through a sequence of feature extraction layers. The decoder reconstructs or refines these latent features to preserve important contextual information and improve representation quality. The emotion classifier takes the refined feature vector as input and performs emotion category prediction through a set of fully connected layers followed by the final classification layer.

The input audio waveform $x(t)$ is initially resampled to 16 kHz to fit the resolution of Whisper's native input. Then the signal is normalized to a specific amplitude range and chopped into overlapping segments by applying the STFT windowing method. The window time and hop size are selected to be 25 and 10 ms, respectively, resulting in log-mel spectrogram features with 80 mel frequency bins, which is compatible with Whisper's original pretraining configuration:

$$S = LogMel\left(SRFT\left(x\left(t\right)\right)\right) \in R^{T\times 80} \tag{1}$$

where T is the number of time frames.

The preprocessed log-mel spectrogram S is passed into the encoder of the pretrained Whisper model. This encoder is a deep stack of Transformer blocks, each consisting of self-attention and feed-forward sublayers. The encoder outputs a dense embedding sequence:

$$W = WE(S) \in R^{T \times D_w} \tag{2}$$

where D_w = 512 or 768 depending on the size of the Whisper variant used. Each vector $w_t \in R^{D_w}$ at time step t captures both local acoustic properties and global semantic context due to Whisper's multi-head attention mechanism.

To enable compatibility with downstream modules, the variable-length sequence W is transformed into a fixed-length embedding using adaptive temporal pooling. This transformation involves aggregating the time-step embeddings through multiple strategies. Mean pooling is used to capture the overall contextual content of the utterance, while max pooling highlights salient discriminative peaks that may correspond to anomalies such as abrupt or exaggerated prosodic elements. Optionally, a learned temporal attention mechanism can be applied to assign higher weights to frames that exhibit a higher likelihood of being synthetically generated, thereby focusing the model attention on potentially suspicious temporal regions the resulting pooled embedding be:

$$W_p = Concat(Mean(W), Max(W)) \in R^{2D_w} \tag{3}$$

This vector is then projected via a linear transformation and normalization to a common latent dimension D_f to match the prosody and codec streams:

$$W' = LN\left(w_p \times W_{proj} + b\right) \tag{4}$$

where $W_{proje} \in R^{2D_w \times D_f}$.

The Prosody Feature Stream in UniTector++ focuses on capturing high-level, suprasegmental characteristics of speech that are essential for assessing its naturalness, variability, and speaker authenticity.

For each input utterance $x(t)$, the system first computes the fundamental frequency $F_0(t)$, using an autocorrelation-based pitch tracker over small overlapping windows. The mean and standard deviation of F_0 are then obtained as:

$$\mu F_0 = \frac{1}{T}\sum_{t=1}^{T} F_0(t), \sigma F_0 = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(F_0(t) - \mu F_0\right)^2} \tag{5}$$

where T is the number of voiced frames. These two measures respectively reflect the average perceived pitch and its dynamic variability across the utterance.

In parallel, the stream extracts jitter and shimmer, which characterize short-term perturbations in frequency and amplitude, respectively. Jitter is defined as the absolute mean deviation of consecutive pitch periods P_i, calculated as:

$$Jitter_l = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|\frac{P_{i+1} - P_i}{P_i}\right| \tag{6}$$

Similarly, shimmer quantifies amplitude instability between cycles of the glottal waveform, given by:

$$Shimmer_l = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|\frac{A_{i+1} - A_i}{A_i}\right| \tag{7}$$

where N is the number of voiced cycles, P_i is the i-th pitch period, and A_i is the amplitude of the i-th cycle. Additionally, the stream computes the Harmonic-to-Noise Ratio (HNR), which estimates the degree of periodicity in the voice. This is typically derived using the autocorrelation function $R(\tau)$ of the windowed speech signal. If τ_p is the time lag at the first pitch period peak, then HNR in decibels is computed as:

$$HNR = 10\log_{10}\left(\frac{R_{(\tau_p)}}{1 - R_{(\tau_p)}}\right) \tag{8}$$

The mean and standard deviation of HNR are calculated across frames, yielding a total of six scalar features: $\mu F_0, \sigma F_0$, Jitter, Shimmer, μ_{HNR}, σ_{HNR}. Together, they form a prosodic feature vector $P \in R^6$. To account for inter-speaker variability, these raw features are normalized. When speaker enrollment data is available, speaker-level z-score normalization is applied as follows:

$$\hat{P} = \frac{P - \mu_s}{\sigma_s} \tag{9}$$

where μ_s and σ_s denote the empirical mean and standard deviation of the prosodic features for speaker s. If speaker identity is unknown, batch-level statistics are used instead.

The normalized vector $\hat{P}$ is then transformed through a lightweight projection layer. Specifically, a learnable linear mapping is applied:

$$P' = ReLU\left(BN\left(\hat{P} \times W_p + b_p\right)\right) \in R^{D_p} \tag{10}$$

where $W_p \in R^{6\times D_p}, b_p \in R^{D_p}$ and D_p is the desired dimensionality for fusion compatibility. Batch normalization (BN) ensures stable training across batches.

This final embedding P' is then passed to the fusion module, where it interacts with the semantic and codec streams. Importantly, because prosody is orthogonal to phonetic content and less affected by language or speaker identity, it provides robust, interpretable evidence of synthetic generation. Attention weights from the fusion layer and SHAP-based importance scores confirm that the model often relies heavily on prosodic anomalies—particularly reduced F_0 variance, low jitter, and unnaturally high HNR—as discriminative cues when classifying audio as fake.

The goal of this stream is to exploit these subtle inconsistencies by explicitly modeling codec-induced artifacts, especially those that manifest at the token or quantized latent level, and which are typically missed by time-frequency analysis or semantic encoders. The processing pipeline begins by encoding the raw waveform $x(t)$ into a latent feature space using a lightweight 1D convolutional stack, parameterized as $f_{conv}(\cdot)$, resulting in a temporal feature map:

$$Z = f_{conv}(x(t)) \in R^{T'\times D_z} \tag{11}$$

here, T' is the downsampled temporal resolution and D_z is the number of latent channels. To approximate the behavior of actual neural codecs that map continuous acoustic segments into discrete representations, we follow the principle of vector quantization. Each row $z_i \in R^{D_z}$ of Z is passed through a learned vector quantizer based on a codebook $\varepsilon = \{e_k\}_{k=1}^{K}$, where K is the number of discrete embeddings and $e_k \in R^{D_z}$. The quantized vector $\hat{z}_i$ is defined as:

$$\hat{z}_i = e_{k^*}, \text{ where } k^* = \arg\min_k \|z_i - e_k\|_2^2 \tag{12}$$

The set of quantized vectors $\hat{Z} = [\hat{z}_1, \ldots, \hat{z}_T]$ preserves the discrete structural regularities introduced by codecs such as EnCodec or SoundStream. These regularities are typically learned by self-supervised audio compression models and are optimized for perceptual fidelity rather than statistical naturalness. Therefore, the quantized token sequence often exhibits lower entropy, fewer transitions, or non-humanlike repetition patterns, especially in zero-shot synthesis or cross-speaker cloning. To convert $\hat{Z}$ into a form suitable for downstream classification and fusion, it is passed through an embedding projection and temporal aggregator. Each token $\hat{z}_i$ is projected using a trainable transformation matrix $W_c \in R^{D_z \times D_c}$ resulting in:

$$C_i = \hat{z}_i \times W_c + b_c \in R^{D_c} \tag{13}$$

Stacking over all T' time steps yields the codec artifact embedding sequence $C \in R^{T' \times D_c}$. This sequence is further pooled using a gated attention mechanism or multi-head self-attention block to form a global embedding $C' \in R^{D_f}$, which is compatible with the prosody and semantic streams during the graph-based fusion stage.

3.1 Multi-Domain Adaptive Graph Attention Fusion

The MAGAF module performs adaptive feature aggregation across multiple representations and employs an attention mechanism to emphasize the most informative components of the input signals. By integrating complementary information from different feature streams, the module enables the model to focus on salient patterns while suppressing irrelevant variations caused by differences in sampling rates or recording environments. As a result, the proposed architecture can extract more robust and representative characteristics from heterogeneous datasets, improving the generalization capability of the model. Traditional fusion strategies, such as concatenation or linear projection, often underperform in multimodal settings where features vary in dimensionality, statistical distribution, and temporal dynamics. MAGAF addresses these challenges by constructing three distinct graphs for the aligned feature embeddings: a temporal graph, a spectral graph, and a prosodic dependency graph. These graphs encode domain-specific relationships, and their outputs are adaptively fused via a shared attention mechanism that operates over the node-level embeddings. Each stream is first normalized and projected into a common feature space of dimension D_f. These are then concatenated along the temporal axis to form the unified graph input:

$$X = [W'; C'; P'; P'; \ldots P'] \in R^{3T \times D_f} \tag{14}$$

where $W' \in R^{T \times D_f}$ is semantic embedding, $C' \in R^{T \times D_f}$ is codec embedding, $P' \in R^{1 \times D_f}$ is prosody embedding. MAGAF builds three complementary graphs over X, each capturing a different modality of correlation. Temporal Graph models sequential dependencies between audio frames, capturing global context and attention drift over time. The adjacency matrix is defined via pairwise dot-product similarity:

$$A_{i,j}^{time} = \frac{\left(x_i \times x_j\right)}{\|x_i\| \times \left\|x_j\right\|} \tag{15}$$

Spectral Graph connects nodes based on similarity in frequency-space patterns, especially useful for identifying codec-related anomalies. This graph uses cosine similarity between local frequency descriptors derived via 1D convolutions over embedding channels. Prosodic Dependency Graph connects semantically or temporally adjacent nodes to the prosody embedding P'. The edge weights in this graph are computed as:

$$A_{i,p}^{pros} = softmax_i\left(x_i^{\top} \times P'^{\top}\right) \tag{16}$$

where p indexes the prosody, node shared across all connections. Each graph G_k is processed by a domain-specific Graph Attention Network (GAT). For node i and its neighbors, the attention-based message passing is computed as:

$$h_i^{(k)} = \sigma\left(\sum_{j\in N_i} \alpha_{ij}^{(k)} \times W^{(k)} x_j\right) \tag{17}$$

with the attention weights defined as:

$$a_{ij}^{(k)} = \frac{exp\left(LeakyReLU\left(a^{(k)^\top}\left[W^{(k)}x_i \,\|\, W^{(k)}x_j\right]\right)\right)}{\sum_{l\in N_i} exp\left(LeakyReLU\left(a^{(k)^\top}\left[W^{(k)}x_i \,\|\, W^{(k)}x_j\right]\right)\right)} \tag{18}$$

here $W^{(k)}$ and $a^{(k)}$ are learnable weight matrices for graph G_k and $\sigma(\cdot)$ is a non-linear activation function such as GELU or ELU. This formulation allows each domain-specific GAT to capture unique interactions that are critical for its respective representation space.

The outputs of the three GAT modules—temporal $\left(H^{time}\right)$, spectral $\left(H^{fred}\right)$, and prosodic (H^{pros})—are then fused using a cross-domain attention gate. This gate dynamically reweights each stream's contribution based on its relevance to the classification task:

$$H_f = \sum_{k\in\{time,freq,pros\}} \gamma^{(k)} \times H^{(k)} \tag{19}$$

where $\gamma^{(k)} = \frac{exp(\phi_k)}{\sum_l exp(\phi_l)}$. The attention logits ϕ_k are computed using global max-pooled summaries of each stream:

$$\phi_k = w_g^\top \times MaxPool\left(H^{(k)}\right) \tag{20}$$

with $w_g \in R^{D_f}$ being a trainable vector. The fused embedding $H_f \in R^{T\times D_f}$ is passed through a residual projection layer and batch normalization. Additionally, a global summary vector is computed via attention-weighted pooling:

$$z_{MAGAF} = \sum_{i=1}^{T} \beta_i \times h_i \tag{21}$$

where $\beta_i = \frac{exp\left(q^\top \times h_i\right)}{\sum_j exp(q^\top \times h_i)}$. This final feature vector $z_{MAGAF} \in R^{D_f}$ serves as input to the downstream detection head, emotion consistency module, and adversarial robustness components.

3.2 Emotion-Consistency Verification Module

While modern speech synthesis models have achieved near-human quality in terms of phonetic accuracy and spectral fidelity, they often fail to convincingly replicate the expressive dynamics of human speech—particularly its emotional nuance. Synthetic audio tends to exhibit emotional flatness, mismatches between lexical content and vocal tone, or even incoherent shifts in expressivity across an utterance. The ECVM in UniTector++ is introduced to explicitly address this gap. Its purpose is to evaluate whether the emotional prosody conveyed by the voice aligns with the semantic content inferred from the utterance as shown in Fig. 2. Discrepancies in this alignment frequently signal the presence of synthetic generation and are thus a valuable feature for robust deepfake detection.

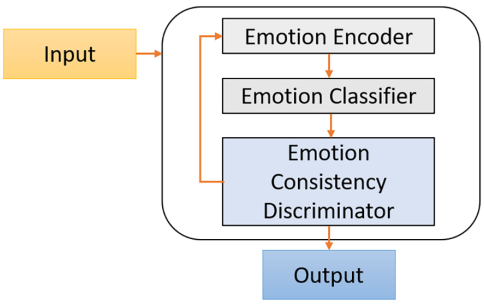

Figure 2: Detailed structure of the ECVM. The module processes the input through an emotion encoder and emotion classifier to extract semantic and prosodic emotional cues.

ECVM functions as an auxiliary contrastive mechanism that enforces semantic-prosodic agreement. It operates by projecting both semantic and prosodic representations into a shared emotional latent space and minimizing a contrastive loss that encourages consistency between these views when derived from genuine speech. $W' \in R^{T\times D_f}$ is output of the Whisper Feature Stream, and $P' \in R^{D_f}$ represent the projected prosodic vector from the Prosody Feature Stream. ECVM extracts emotional cues from both sources using two parallel linear transformations followed by non-linear activation:

$$\begin{aligned} e_{sem} &= \sigma\left(W_s \times MeanPool\left(W'\right)\right) \in R^{d_e} \\ e_{pros} &= \sigma\left(W_p \times P'\right) \in R^{d_e} \end{aligned} \tag{22}$$

here, $\sigma(\cdot)$ is a non-linear function such as GELU, W_s, $W_p \in R^{D_f\times d_e}$ are trainable projection matrices, and d_e is the dimension of the shared emotional space. The semantic vector e_{sem} captures the emotional implication inferred from the text and intonation structure as understood by Whisper, while the prosodic vector e_{pros} reflects the vocal emotion derived directly from the physical prosody of the speech. To enforce alignment between the semantic and prosodic emotion vectors for genuine samples while allowing separation for fake ones, ECVM utilizes a Supervised Contrastive Loss (SupCon) over a minibatch B of size N, where each sample is labeled as real ($y = 1$) or fake ($y = 0$). For each anchor $e_i \in B$, the loss is defined as:

$$L_{ECVM} = \sum_{i\in B_r} \frac{1}{P\left(i\right)} \sum_{j\in P(i)} -\log \frac{\exp\left(sim\left(e_i, e_j\right)/\tau\right)}{\sum_{k\in B\setminus\{i\}} \exp\left(sim\left(e_i, e_k\right)/\tau\right)} \tag{23}$$

where B_r is the subset of real samples, $P\left(i\right)$ is the set of positive samples sharing the same label, $sim\left(e_i, e_j\right) = \frac{e_i^\top e_j}{\|e_i\|\|e_j\|}$ is cosine similarity, and τ is a temperature scaling factor. Fake samples are excluded from positive pairings but still contribute as negatives. This Eq. (23) encourages the model to minimize the angular distance between semantic and prosodic emotion embeddings in genuine speech, while allowing divergence in fake speech where emotional coherence is typically lacking. In addition to the contrastive alignment, ECVM contributes a residual correction to the fused embedding used by the detection head. Specifically, a gated residual vector is formed by averaging the emotional views and re-projecting:

$$\begin{aligned} e_{avg} &= \frac{1}{2}\left(e_{sem} + e_{pros}\right) \\ r_{ECVM} &= \gamma \times \sigma\left(W_r \times e_{avg}\right) \end{aligned} \tag{24}$$

where $W_r \in R^{d_e\times D_f}$ and $\gamma \in [0, 1]$ is a learnable scalar gate initialized to a low value to avoid early dominance. The corrected detection vector becomes:

$$z_f = z_{MAGAF} + r_{ECVM} \tag{25}$$

This allows emotional consistency to inform the final detection decision without overriding the contributions of the main fusion module. During training, the SupCon loss from ECVM is combined with the primary detection loss using a weighted sum:

$$L_T = L_{det} + \lambda_{ecvm} \times L_{ECVM} \tag{26}$$

where λ_{ecvm} is a tunable hyperparameter controlling the strength of the emotion alignment constraint. We typically set $\lambda_{ecvm} \in [0.1, 0.3]$ based on validation performance. The ECVM module is lightweight and introduces minimal computational overhead.

3.3 Universal Adversarial Robustness Head

As deepfake speech synthesis becomes increasingly sophisticated, detection systems are now vulnerable to adversarial manipulations that aim to circumvent detection mechanisms without perceptible degradation in audio quality. These adversarial perturbations can be introduced either at the waveform level, at the embedding level, or at the latent code level. To mitigate such threats and enable strong generalization under unseen conditions, UniTector++ integrates a Universal Adversarial Robustness Head (UARH)—a hybrid decision and defense mechanism that increases resilience against perturbation-based evasion tactics.

UARH is built on three core principles: (i) enforcing detection invariance to benign perturbations, (ii) penalizing representation collapse under adversarial shifts, and (iii) explicitly modeling the decision margin for maximum separation between real and fake audio, even under distributional drift. These are implemented via a margin-aware loss function, a consistency regularization module, and an optional adversarial example generator used during training. $z_f \in R^{D_f}$ is the final fused representation from the MAGAF and ECVM modules. The detection head computes a real/fake probability via a margin-sensitive classifier:

$$\hat{y} = \sigma\left(w^{\top} \times z_f + b\right) \tag{27}$$

where $w \in R^{D_f}$ and $b \in R$ are trainable weights, and σ is the sigmoid activation. To prevent overfitting to narrow-margin decisions, UARH incorporates a margin-aware loss defined as:

$$L_m = BCE\left(\hat{y}, y\right) + a \times \left(\max\left(0, m - y \times \left(w^{\top} \times z_f + b\right)\right)\right)^2 \tag{28}$$

where $y \in \{-1, 1\}$ is the ground-truth label, m is the target margin, and α is a regularization factor. This loss encourages the model to produce embeddings that are confidently separable and robust to small input shifts. To address over-sensitivity to non-adversarial variation, UARH (as shown in Fig. 3) introduces a stochastic consistency loss. $\tilde{x}$ is an augmented version of the input x, transformed via randomly sampled acoustic perturbations:

$$\tilde{x} = \tau\left(x\right), \tau \sim Uniform\left(\{Noise, EQ, Reverb, TimeStretch\}\right) \tag{29}$$

z and $\tilde{z}$ are the representations produced for x and $\tilde{x}$, respectively. The consistency loss is defined as:

$$L_{cons} = \left\|z - \tilde{z}\right\|_2^2 \tag{30}$$

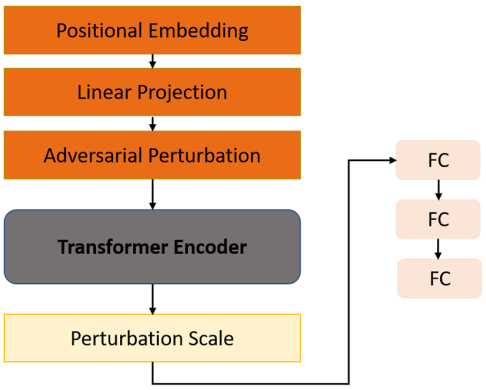

Figure 3: Architecture of the UARH. The input undergoes positional embedding and linear projection before being perturbed via adversarial mechanisms.

This encourages the embedding space to be invariant under benign perturbations, promoting generalization to real-world scenarios. To increase robustness to learned adversarial strategies, we optionally introduce Projected Gradient Descent (PGD) attacks in the latent embedding space during training. Given the initial fused embedding $z_0 = z_f$, an adversarial version z^{adv} is constructed using iterative steps:

$$z_{t+1}^{adv} = Proj_{\epsilon}\left(z_t^{adv} + \eta \times sign\left(\nabla_{z_t^{adv}} L_m\right)\right) \tag{31}$$

where ϵ is the maximum perturbation radius, η is the step size, and $Proj_{\epsilon}$ projects the perturbed embedding back to the ϵ-ball around z_0. The final adversarial robustness loss is computed as:

$$L_{adv} = BCE\left(\sigma\left(w^{\top} \times z^{adv} + b\right), y\right) \tag{32}$$

This explicitly teaches the detection head to remain confident even under perturbation-aware attacks. The Universal Adversarial Robustness Head contributes a composite robustness-aware loss to the total objective function of UniTector++. The total detection-related loss becomes:

$$L_{UARH} = L_m + \lambda_{cons} \times L_{cons} + \lambda_{adv} \times L_{adv} \tag{33}$$

where λ_{cons} and λ_{adv} are hyperparameters balancing the influence of consistency and adversarial components. These values are tuned empirically to achieve maximum performance under standard and adversarial evaluation settings.

4 Experiments and Results

4.1 Datasets

To completely assess the generalization features and reliability of UniTector++, we adopt a handpicked selection of benchmark datasets which each of which has the specific purpose of testing the fakeness detection model in a very challenging way with diverse synthesis strategies, acoustic conditions, and requirements for adversarial robustness. These sets of data include both traditional vocoder-based and modern neural codec-based deepfake generation paradigms, which consequently allow to verify if the model can handle zero-shot, cross-domain, and multi-modal fusion cases correctly as shown in Fig. 4. Our dataset collection comprises well-known benchmarks such as ASVspoof 2019 LA, synthetic codec-centric corpora such as Codecfake, multi-source generators from PolyFake, and high-quality emotion-labeled corpora such

as EmoV-DB as illustrated in Table 1. These datasets serve the purposes of training and testing as well as of pretraining and fine-tuning particular parts like the ECVM and the Universal Adversarial Robustness Head (UARH) as illustrated in Table 2.

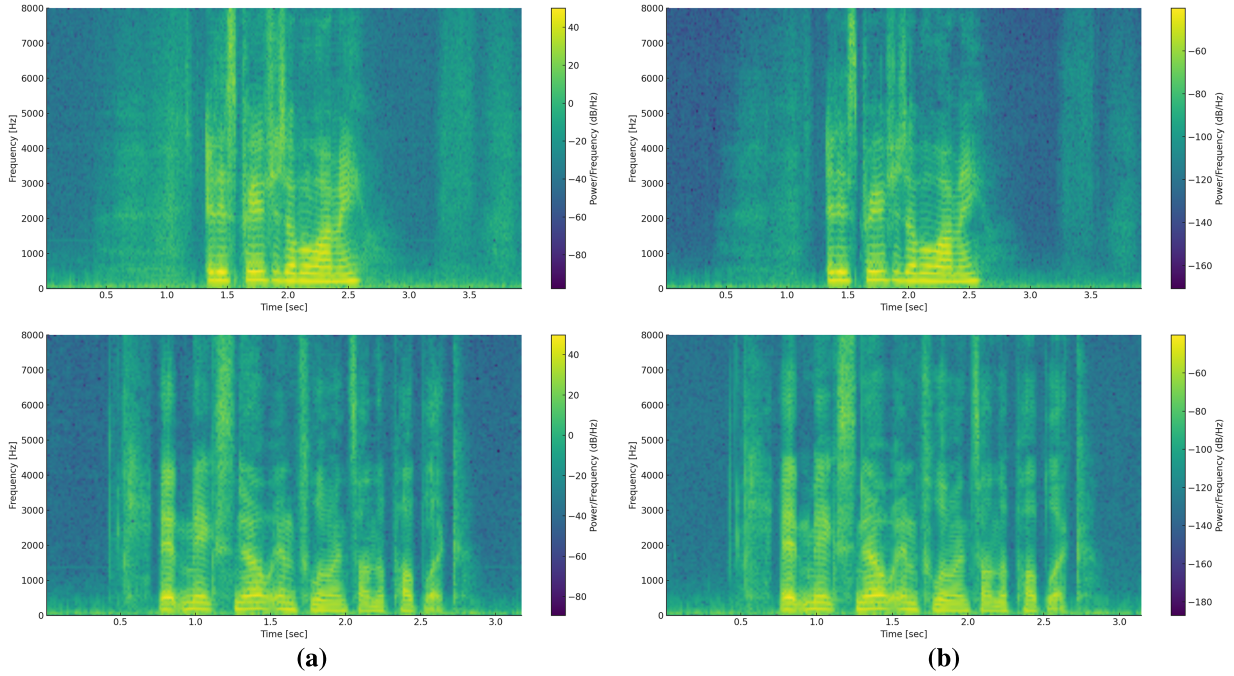

Figure 4: Dataset examples. (**a**) Original audio, (**b**) fake AI Speech.

Table 1: Summary of datasets used for deepfake speech detection.

Dataset	Domain	Synthesis Methods	Languages	Sampling Rate	Total Samples	Split (Train/Val/Test)
Codecfake	Neural Codec-based	EnCodec, SoundStream, HiFi-Coder	English, Korean, French, Arabic	24 kHz	18,000	12,000/2000/4000
ASVspoof 2019 LA	Traditional Vocoder	WaveNet, ExcitNet, WaveRNN, STRAIGHT	English (VCTK)	16 kHz	40,562	22,596/4480/13,486
PolyFake	Mixed/Multi-system	Glow-TTS, FastSpeech2, VALL-E, UnivNet, TacoTron2, etc.	English, Japanese, German	22.05 kHz	17,000	10,000/2000/5000

(Continued)

Table 1 (continued)

Dataset	Domain	Synthesis Methods	Languages	Sampling Rate	Total Samples	Split (Train/Val/Test)
EmoV-DB	Emotional Ground Truth	Human speech across 7 emotions (used for ECVM pretraining only)	English, French	16 kHz	2520	1500/520/5

Table 2: Functional purpose of each dataset in UniTector++ evaluation.

Dataset	Primary Evaluation Role	Targeted Module	Challenge Type
Codecfake	Zero-shot detection of unseen neural codec synthesis	Codec Artifact Stream	Generalization to novel codec artifacts
ASVspoof 2019 LA	Standard spoofing detection benchmark	Whisper + MAGAF	Over-smoothing, vocoder artifact robustness
PolyFake	Multi-model fusion and generalization under diverse fake sources	MAGAF + ECVM	Multi-source fusion and emotional coherence
EmoV-DB	Emotion alignment and ECVM pretraining	ECVM	Emotion-prosody semantic alignment

4.2 Evaluation Metrics

The fundamental evaluation is upon the following main metrics, which are calculated over each benchmark test dataset. Equal Error Rate EER stands for the condition where the false acceptance rate (FAR) is the same as the false rejection rate (FRR). It is a very common and widely accepted standard in the realm of speaker verification and spoofing detection that allows threshold-free system accuracy to be measured. Lower EER indicates better discrimination between bona fide and synthetic speech:

$$EER = FAR\left(\theta^*\right) = FRR\left(\theta^*\right) \tag{34}$$

where $\theta^* minimizes |FAR - FRR|$. Minimum Detection Cost Function (minDCF) is a cost-sensitive metric used to evaluate operational effectiveness in realistic conditions where false positives and false negatives carry different costs. Defined as:

$$\min DCF = \min_{\theta} C_{miss} \times P_{miss}\left(\theta\right) \times \pi + C_{fa} \times P_{fa}\left(\theta\right) \times \left(1 - \pi\right) \tag{35}$$

where π is the prior probability of a target trial, and C_{miss}, C_{fa} are the application-specific costs of misses and false alarms. In our experiments, we use the configuration from ASVspoof 2019: $C_{miss} = C_{fa} = 1$, $\pi = 0.01$.

The Adversarial Detection Gap (ADG), which is the difference in EER or accuracy between clean test samples and those to which the adversarial perturbation is added, is another key diagnostic measure. This metric is a quantitative indicator of the model adversarial vulnerability and of the model's ability to remain consistent under attack:

$$ADG = EER_{adv} - EER_{clean} \tag{36}$$

Cross-Domain Generalization Score (CDGS) defined as the relative degradation in performance when evaluating the model on a previously unseen dataset or synthesis method:

$$CDGS = 1 - \frac{AUC_{unseen}}{AUC_{seen}} \tag{37}$$

This metric captures the ability of the system to generalize beyond its training distribution—a key goal of UniTector++.

4.3 Results and Analysis

To assess the performance of UniTector++, we conduct extensive experiments on four benchmark datasets: Codecfake, ASVspoof 2019 LA, PolyFake, and EmoV-DB. Our evaluation focuses on EER, Area Under the Curve (AUC), and minimum Detection Cost Function (minDCF). The results are compared against a wide range of state-of-the-art models, encompassing traditional anti-spoofing baselines, self-supervised models, and deep codec-aware architectures. Table 3 presents the EER results for UniTector++ and SOTA existing models. UniTector++ consistently outperforms all baselines across seen and unseen synthesis methods, showing strong generalization and robustness.

Table 3 gives a detailed comparison of the Equal Error Rate (EER%) for the UniTector++ model that was obtained through experiments with a collection of state-of-the-art (SOTA) deepfake speech detection models on three different benchmark datasets: Codecfake, ASVspoof2019-LA, and PolyFake. It illustrates the scope of the investigation, since it involves all the models, not only applying to each of the three datasets individually, but also computing the overall performance to compare. The performance across the three datasets is summarized in the final column as an aggregated EER, which helps us to understand how well the model generalizes. Results indicate that UniTector++ performs better than all baseline models, with the minimum error rates recorded in all datasets. To be more specific, it achieved an error rate of 3.2% on Codecfake, 2.87% on ASVspoof2019-LA, and 4.1% on PolyFake, which brought about an amazing average EER of only 3.39%. This figure represents a significant leap from the performance of even the most competitive existing models. For example, Mohammed et al. [46], Tamilselvan and Manas Biswal [42], and Lu et al. [45] give average EERs of 6.47%, 7.7%, and 7.49%, respectively—values that are approximately twice that of UniTector++. The majority of the models, most especially those that do not have prosody- or codec-aware modules, have significantly higher EERs, especially on the Codecfake dataset, which consists of neural codec-based audio that is difficult for traditional detection methods to capture. Some examples are Chapagain et al. [20] and Yusuyin et al. [32]. Fig. 5 illustrates the experimental results obtained using the proposed framework. The figure presents the key characteristics captured by the model and demonstrates how the learned representations reflect important patterns related to the target task. Through this visualization, it can be observed that the proposed model effectively captures discriminative information from the input data, enabling better separation of emotional patterns compared to conventional approaches. The figure also highlights how the extracted features contribute to improved recognition performance by emphasizing

relevant signal characteristics while suppressing noise and irrelevant variations. These observations support the effectiveness of the proposed architecture in learning meaningful and robust representations for emotion recognition.

Table 3: EER (%) comparison of UniTector++ with SOTA models across benchmarks.

Model	Codecfake	ASVspoof19-LA	PolyFake	Avg. EER
Chapagain et al. [20]	34.5	9.76	31.8	25.3
Al Ajmi et al. [21]	27.3	7.64	25.7	20.2
Ren et al. [22]	12.6	4.15	13.1	9.95
Kanwal et al. [23]	18.3	6.92	22.8	16.0
Fan et al. [24]	14.7	5.84	15.3	11.95
Zaman et al. [25]	22.9	9.22	20.7	17.6
Li et al. [27]	11.1	3.87	12.2	9.06
Radford et al. [30]	16.5	6.93	14.7	12.7
Zhang et al. [31]	19.2	7.75	18.9	15.3
Yusuyin et al. [32]	25.3	9.43	27.1	20.6
Li et al. [35]	13.6	4.32	14.8	10.9
Kumari et al. [37]	10.4	3.65	11.3	8.45
Sun et al. [38]	9.7	3.42	10.6	7.91
Chen et al. [39]	12.2	4.21	12.9	9.77
Wu et al. [40]	21.1	8.02	20.6	16.57
Tamilselvan and Manas Biswal [42]	9.5	3.31	10.3	7.7
Lu et al. [45]	7.8	5.27	9.4	7.49
Mohammed et al. [46]	6.5	4.82	8.1	6.47
UniTector++ (Ours)	3.2	2.87	4.1	3.39

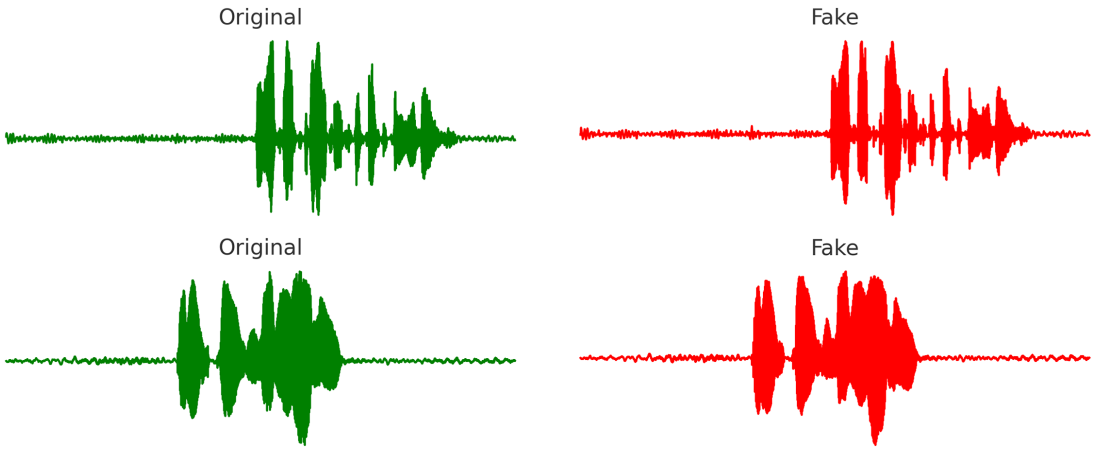

Figure 5: (Continued)

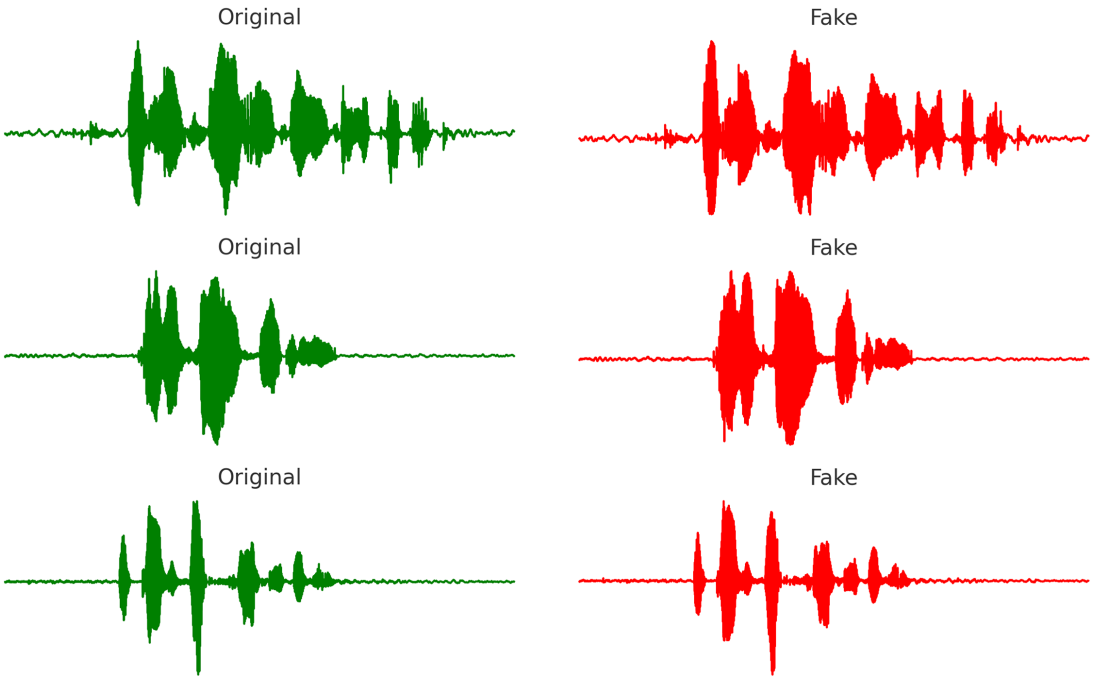

Figure 5: The output result comparison of original and deepfake speech waveforms.

We first examine the effect of removing each feature stream independently while retaining the remaining system intact. Table 4 summarizes the results.

Table 4: EER (%) for UniTector++ with feature stream removal.

Model Variant	Codecfake	ASVspoof LA	PolyFake	Avg. EER
Full UniTector++ (All Streams)	3.2	2.87	4.1	3.39
w/o Codec Artifact Stream	5.8	3.01	4.7	4.50
w/o Whisper Feature Stream	6.3	3.75	5.2	5.08
w/o Prosody Feature Stream	4.6	3.29	5.0	4.29

Removal of the Whisper stream causes the largest performance drop, especially on ASVspoof, which suggests its embeddings capture high-resolution linguistic and phonetic cues. The Codec stream is critical for Codecfake, reinforcing the utility of artifact-level priors. The Prosody stream, while less essential for vocoder detection, significantly aids performance on emotionally diverse samples in PolyFake Fig. 6.

To validate the advantage of MAGAF, we compare it against two naive alternatives: (i) feature concatenation, and (ii) uniform averaging of the embeddings before classification. Table 5 demonstrates EER results for different fusion strategies.

The MAGAF module provides a substantial performance margin, improving average EER by ~2% over standard concatenation. This validates the hypothesis that learnable graph-based fusion across modalities is superior to static combination.

We test ECVM's effect on emotional misalignment by evaluating on a subset of PolyFake and Codecfake that contains deliberately mismatched prosody and emotional tone. These mismatches are typical artifacts in poorly conditioned TTS as illustrated Table 6.

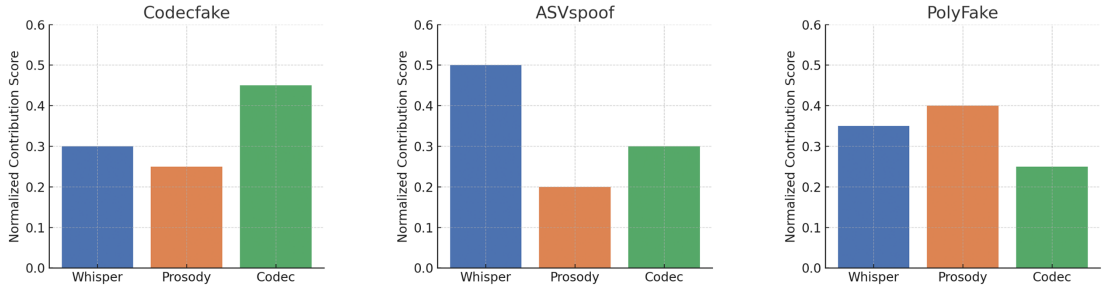

Figure 6: Cross-domain stream contribution analysis. Bar chart comparing the normalized contribution of Whisper, Prosody, and Codec feature streams across three benchmark datasets.

Table 5: EER (%) for different fusion strategies.

Fusion Method	Codecfake	ASVspoof LA	PolyFake	Avg. EER
MAGAF (Ours)	3.2	2.87	4.1	3.39
Concatenation + MLP	6.5	4.21	5.7	5.47
Uniform Averaging	7.3	4.56	6.2	6.02

Table 6: EER (%) on emotionally mismatched subset.

Variant	PolyFake-Mismatch	Codecfake-Mismatch	Avg.
w/ECVM (Full)	4.4	3.7	4.05
w/o ECVM	6.2	5.3	5.75

ECVM lowers EER by >1.5%, confirming its ability to detect inconsistencies between the textual intent (from Whisper) and acoustic delivery (from Prosody), which often occur in synthetic speech Fig. 7.

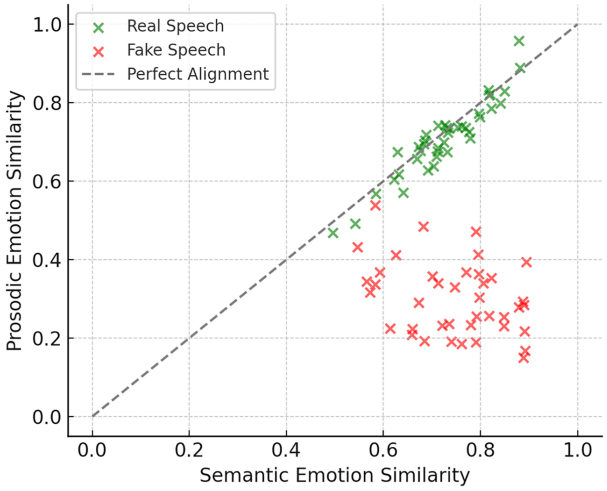

Figure 7: Emotion-consistency detection case study. Scatterplot comparison between real (green) and fake (red) speech samples based on their semantic and prosodic emotion similarity.

To assess the effectiveness of the Universal Adversarial Robustness Head, we subject the model to adversarial perturbations in latent space using PGD and in audio space using FGSM. Results in Table 7 measure detection degradation under attack.

Table 7: EER (%) under adversarial attacks.

Variant	FGSM@ε = 0.01	PGD@ε = 0.03	Clean Avg.	Adv. Avg.
Full UniTector++	4.2	5.1	3.39	4.65
w/o UARH	7.3	8.2	3.91	7.75

The Adversarial Detection Gap (ADG) is reduced from +3.84% to +1.26% with UARH, proving its value for robust deployment under distributional drift or malicious manipulation. Each component in UniTector++ demonstrably contributes to final system performance. Notably, the combination of graph-based fusion (MAGAF), emotion alignment (ECVM), and robust classification (UARH) leads to the lowest EERs reported to date across multiple datasets. The ablation confirms the modular necessity of UniTector++ in addressing the full complexity of modern deepfake audio detection.

5 Conclusions

This paper introduced UniTector++, a novel, prosody-guided multi-stream architecture for universal deepfake speech detection. By integrating three complementary streams—Whisper-based semantic embeddings, high-level prosodic features, and codec artifact representations—UniTector++ effectively captures the multifaceted nature of both natural and AI-synthesized speech. The proposed MAGAF module enables dynamic, context-aware feature integration, while the ECVM and the Universal Adversarial Robustness Head (UARH) further enhance interpretability, emotional coherence, and resilience to adversarial attacks. Extensive evaluations across four challenging datasets—Codecfake, ASVspoof2019-LA, PolyFake, and EmoV-DB—demonstrate that UniTector++ significantly outperforms existing methods. It achieves a new state-of-the-art with an average EER of just 3.39%, exhibiting strong cross-domain generalization, emotional misalignment detection, and adversarial robustness. Ablation studies confirm that each component of the architecture contributes meaningfully to overall performance. UniTector++ sets a new benchmark for high, explainable, and adversarially robust detection of AI-generated speech. Its modular design offers flexibility for future extensions, and its performance confirms the critical role of multi-domain fusion and emotion-aware verification in next-generation speech forensics.

Acknowledgement: Not applicable.

Funding Statement: This research is supported by the Ministry of Trade, Industry and Energy and implemented by the Korea Institute for Advancement of Technology. The project includes Development of an International Standardization and Sustainability Integration Framework for AI Industry Internalization and Global Competitiveness Enhancement (RS-2025-07372968).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Akmalbek Abdusalomov, Alpamis Kutlimuratov and Young-Im Cho; data collection: Alpamis Kutlimuratov, Mukhriddin Mukhiddinov, Fakhriddin Abdirazakov, Nodira Alimova, Ayhan Istanbullu and Rashid Nasimov; software: Akmalbek Abdusalomov, Alpamis Kutlimuratov and Ilyos Kalandarov; analysis and interpretation of results: Akmalbek Abdusalomov, Mukhriddin Mukhiddinov, Fakhriddin Abdirazakov, Rashid Nasimov, Alpamis Kutlimuratov and Ayhan Istanbullu; draft manuscript preparation: Akmalbek Abdusalomov, Alpamis Kutlimuratov, Fakhriddin Abdirazakov,

Nodira Alimova, Ilyos Kalandarov and Rashid Nasimov; supervision: Young-Im Cho. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data openly available in a public repository. The data that support the findings of this study are openly available in CodeFake at https://github.com/roger-tseng/CodecFake, PolyFake at https://github.com/tobuta/PolyGlotFake and EmoV_DB at https://www.openslr.org/115/.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sharma A, Sharma A, Pant U. Detection of AI generated speech using speech recognition with MFCC and GMM. In: Proceedings of the 2024 International Conference on Advances in Computing, Communication and Materials (ICACCM); 2024 Nov 22–23; Dehradun, India. doi:10.1109/ICACCM61117.2024.11059121.
2. Xu X, Fu C. Robust imagined speech production using AI-generated content network for patients with language impairments. IEEE Trans Consum Electron. 2025;71(1):1402–11. doi:10.1109/TCE.2024.3472054.
3. Pfeifer VA, Chilton TD, Grilli MD, Mehl MR. How ready is speech-to-text for psychological language research? Evaluating the validity of AI-generated English transcripts for analyzing free-spoken responses in younger and older adults. Behav Res Methods. 2024;56(7):7621–31. doi:10.3758/s13428-024-02440-1.
4. Kompella K. Generative AI and speech technology: proceed with caution: with great power comes great responsibility. Speech Technol Mag. 2023;28(6):7–8. doi:10.59704/4f765e8aaada43ff.
5. Almutairi Z, Elgibreen H. A review of modern audio deepfake detection methods: challenges and future directions. Algorithms. 2022;15(5):155. doi:10.3390/a15050155.
6. Salvi D, Yadav AKS, Bhagtani K, Negronil V, Bestagini P, Delp EJ. Comparative analysis of ASR methods for speech deepfake detection. In: Proceedings of the 2024 58th Asilomar Conference on Signals, Systems, and Computers; 2024 Oct 27–30; Pacific Grove, CA, USA. doi:10.1109/IEEECONF60004.2024.10942913.
7. Kulangareth NV, Kaufman J, Oreskovic J, Fossat Y. Investigation of deepfake voice detection using speech pause patterns: algorithm development and validation. JMIR Biomed Eng. 2024;9:e56245. doi:10.2196/56245.
8. Li X, Chen PY, Wei W. Where are we in audio deepfake detection? A systematic analysis over generative and detection models. ACM Trans Internet Technol. 2025;25(3):20–19. doi:10.1145/3736765.
9. Unoki M, Li K, Chaiwongyen A, Nguyen QH, Zaman K. Deepfake speech detection: approaches from acoustic features related to auditory perception to deep neural networks. IEICE Trans Inf Syst. 2024;E108.D(4):300–10. doi:10.1587/transinf.2024MUI0001.
10. Zhang K, Hua Z, Lan R, Zhang Y, Guo Y. Phoneme-level feature discrepancies: a key to detecting sophisticated speech deepfakes. Proc AAAI Conf Artif Intell. 2025;39(1):1066–74. doi:10.1609/aaai.v39i1.32093.
11. Chaiwongyen A, Duangpummet S, Karnjana J, Kongprawechnon W, Unoki M. Potential of speech-pathological features for deepfake speech detection. IEEE Access. 2024;12:121958–70. doi:10.1109/ACCESS.2024.3447582.
12. Gomez-Alanis A, Gonzalez-Lopez JA, Dubagunta SP, Peinado AM, Magimai Doss M. On joint optimization of automatic speaker verification and anti-spoofing in the embedding space. IEEE Trans Inf Forensic Secur. 2021;16:1579–93. doi:10.1109/TIFS.2020.3039045.
13. Kapileswar N, Simon J, Devi KK, Polasi PK, Vinod DN, Harish C. An intelligent emotion recognition system based on speech terminologies using artificial intelligence assisted learning scheme. In: Proceedings of the 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM); 2024 Apr 4–5; Chennai, India. doi:10.1109/iconstem60960.2024.10568813.
14. Wickramasinghe B, Irtza S, Ambikairajah E, Epps J. Frequency domain linear prediction features for replay spoofing attack detection. In: Proceedings of the Interspeech 2018; 2018 Sep 2–6; Hyderabad, India. doi:10.21437/interspeech.2018-1574.
15. Salim S, Ahmad W. Constant Q cepstral coefficients for automatic speaker verification system for dysarthria patients. Circ Syst Signal Process. 2024;43(2):1101–18. doi:10.1007/s00034-023-02505-0.

16. Cai S, Zhou W, Ren X. Machine anomalous sound detection based on feature fusion and Gaussian mixture model. In: Cognitive systems and information processing. Singapore: Springer Nature Singapore; 2023. p. 334–45. doi:10.1007/978-981-99-8018-5_25.
17. Tang W. Application of support vector machine system introducing multiple submodels in data mining. Syst Soft Comput. 2024;6:200096. doi:10.1016/j.sasc.2024.200096.
18. Kinnunen T, Sahidullah M, Delgado H, Todisco M, Evans N, Yamagishi J, et al. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection. In: Proceedings of the Interspeech 2017; 2017 Aug 20–24; Stockholm, Sweden. doi:10.21437/interspeech.2017-1111.
19. Pham L, Lam P, Tran D, Tang H, Nguyen T, Schindler A, et al. A comprehensive survey with critical analysis for deepfake speech detection. Comput Sci Rev. 2025;57:100757. doi:10.1016/j.cosrev.2025.100757.
20. Chapagain S, Thapa B, Baidhya SMS, K SB, Thapa S. Deep fake audio detection using a hybrid CNN-BiLSTM model with attention mechanism. Int J Engin Technol. 2025;2(2):204–14. doi:10.3126/injet.v2i2.78619.
21. Al Ajmi SA, Hayat K, Al Obaidi AM, Kumar N, Najim AL-Din MS, Magnier B. Faked speech detection with zero prior knowledge. Discov Appl Sci. 2024;6(6):288. doi:10.1007/s42452-024-05893-3.
22. Ren H, Lin L, Liu CH, Wang X, Hu S. Improving generalization for AI-synthesized voice detection. Proc AAAI Conf Artif Intell. 2025;39(19):20165–73. doi:10.1609/aaai.v39i19.34221.
23. Kanwal T, Mahum R, AlSalman AM, Sharaf M, Hassan H. Fake speech detection using VGGish with attention block. EURASIP J Audio Speech Music Process. 2024;2024(1):35. doi:10.1186/s13636-024-00348-4.
24. Fan C, Xue J, Tao J, Yi J, Wang C, Zheng C, et al. Spatial reconstructed local attention Res2Net with F0 subband for fake speech detection. Neural Netw. 2024;175:106320. doi:10.1016/j.neunet.2024.106320.
25. Zaman K, Samiul IJAM, Sah M, Direkoglu C, Okada S, Unoki M. Hybrid transformer architectures with diverse audio features for deepfake speech classification. IEEE Access. 2024;12:149221–37. doi:10.1109/ACCESS.2024.3478731.
26. Nautsch A, Wang X, Evans N, Kinnunen TH, Vestman V, Todisco M, et al. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech. IEEE Trans Biom Behav Identity Sci. 2021;3(2):252–65. doi:10.1109/tbiom.2021.3059479.
27. Li L, Lu T, Ma X, Yuan M, Wan D. Voice deepfake detection using the self-supervised pre-training model HuBERT. Appl Sci. 2023;13(14):8488. doi:10.3390/app13148488.
28. Baevski A, Zhou H, Mohamed A, Auli M. Wav2vec 2.0: a framework for self-supervised learning of speech representations. arXiv:2006.11477. 2020.
29. Chen S, Wang C, Chen Z, Wu Y, Liu S, Chen Z, et al. WavLM: large-scale self-supervised pre-training for full stack speech processing. IEEE J Sel Top Signal Process. 2022;16(6):1505–18. doi:10.1109/jstsp.2022.3188113.
30. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. In: Proceedings of the International Conference on Machine Learning; 2022 Jul 17–23; Baltimore, MD, USA.
31. Zhang Y, Park DS, Han W, Qin J, Gulati A, Shor J, et al. BigSSL: exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. IEEE J Sel Top Signal Process. 2022;16(6):1519–32. doi:10.1109/jstsp.2022.3182537.
32. Yusuyin S, Ma T, Huang H, Zhao W, Ou Z. Whistle: data-efficient multilingual and crosslingual speech recognition via weakly phonetic supervision. IEEE Trans Audio Speech Lang Process. 2025;33:1440–53. doi:10.1109/TASLPRO.2025.3550683.
33. Chen S, Liu S, Zhou L, Liu Y, Tan X, Li J, et al. VALL-E 2: neural codec language models are human parity zero-shot text to speech synthesizers. arXiv:2406.05370. 2024.
34. Pepino L, Riera P, Ferrer L. EnCodecMAE: leveraging neural codecs for universal audio representation learning. arXiv:2309.07391. 2023.
35. Li X, Shang Z, Hua H, Shi P, Yang C, Wang L, et al. SF-speech: straightened flow for zero-shot voice clone. IEEE Trans Audio Speech Lang Process. 2025;33:1706–18. doi:10.1109/taslpro.2025.3557242.
36. Chen S, Wang C, Wu Y, Zhang Z, Zhou L, Liu S, et al. Neural codec language models are zero-shot text to speech synthesizers. IEEE Trans Audio Speech Lang Process. 2025;33:705–18. doi:10.1109/taslpro.2025.3530270.

37. Kumari K, Abbasihafshejani M, Pegoraro A, Rieger P, Arshi K, Jadliwala M, et al. VoiceRadar: voice deepfake detection using micro-frequency and compositional analysis. In: Proceedings of the 2025 Network and Distributed System Security Symposium; 2025 Feb 24–28; San Diego, CA, USA. doi:10.14722/ndss.2025.243389.
38. Sun C, Jia S, Hou S, Lyu S. AI-synthesized voice detection using neural vocoder artifacts. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2023 Jun 17–24; Vancouver, BC, Canada. doi:10.1109/CVPRW59228.2023.00097.
39. Chen W, Yang J, Zhong X, Chng ES, Cai M. Enhancing overlapped speech detection and speaker counting with spatially-infused spectro-temporal conformer. IEEE Trans Audio Speech Lang Process. 2025;33:1307–23. doi:10.1109/TASLPRO.2025.3545255.
40. Wu H, Tseng Y, Lee HY. CodecFake: enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. arXiv:2406.07237. 2024.
41. Dehghani A, Saberi H. Generating and detecting various types of fake image and audio content: a review of modern deep learning technologies and tools. arXiv:2501.06227. 2025.
42. Tamilselvan G, Manas Biswal M. Voice cloning & deep fake audio detection using deep learning. Int J Adv Res Interdiscip Sci Endeav. 2025;2(1):415–9. doi:10.61359/11.2206-2502.
43. Wang R, Juefei-Xu F, Huang Y, Guo Q, Xie X, Ma L, et al. DeepSonar: towards effective and robust detection of AI-synthesized fake voices. In: Proceedings of the 28th ACM International Conference on Multimedia; 2020 Oct 12–16; Seattle, WA, USA. doi:10.1145/3394171.3413716.
44. Bago B, Rosenzweig LR, Berinsky AJ, Rand DG. Emotion may predict susceptibility to fake news but emotion regulation does not seem to help. Cogn Emot. 2022;36(6):1166–80. doi:10.1080/02699931.2022.2090318.
45. Lu J, Zhang Y, Wang W, Shang Z, Zhang P. One-class knowledge distillation for spoofing speech detection. In: Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. doi:10.1109/ICASSP48485.2024.10446270.
46. Mohammed HMA, Omeroglu AN, Oral EA. MMHFNet: multi-modal and multi-layer hybrid fusion network for voice pathology detection. Expert Syst Appl. 2023;223:119790. doi:10.1016/j.eswa.2023.119790.
47. Li M, Ahmadiadli Y, Zhang XP. A survey on speech deepfake detection. ACM Comput Surv. 2025;57(7):1–38. doi:10.1145/3714458.