IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 February 2026, accepted 16 March 2026, date of publication 19 March 2026, date of current version 25 March 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3675783

RESEARCH ARTICLE

Enhancing Synthesized Speech Detection Using Modified Group Delay and Self-Supervised Compact Convolutional Transformer

SOULAIMAN MOUALLA[1], OUMAYMA AL DAKKAK[1], OMAR HAMDOUN[1], AND ASSEF JAFAR[2]
[1]Higher Institute for Applied Sciences and Technology (HIAST), Damascus, Syria
[2]Syrian Private University (SPU), Damascus, Syria

Corresponding author: Soulaiman Moualla (sulieman.muallah@hiast.edu.sy)

This work was supported by the Higher Institute for Applied Sciences and Technology (HIAST).

ABSTRACT The escalating sophistication of synthesized speech necessitates detection systems that target fundamental physiological constraints rather than surface artifacts. Our approach leverages Modified Group Delay (MGD) to capture the vocal tract characteristics that remain challenging for synthesizers to replicate. We then use this robust representation to train a Compact Convolutional Transformer (CCT) model via selfsupervised learning. This training is based on masked prediction, inspired by models like Wav2Vec2, using an Information Noise-Contrastive Estimation (InfoNCE) loss to force the model to learn the local, deep, and contextual structure of real human speech. The trained CCT thus becomes a powerful encoder, producing discriminative embeddings. Finally, a simple supervised binary classifier is trained on top of these embeddings to distinguish between real and synthesized speech, creating a hybrid detection system resilient to novel and evolving synthetic threats. This method achieved state-of-the-art performance when evaluated on the ASVspoof 2019 LA dataset, with an Equal Error Rate (EER) of 0.0093%, thus performing a new benchmark without augmentation. Furthermore, it demonstrated superior generalization, with an EER of 0.61% on the more challenging ASVspoof 2021 LA dataset and excellent performance with an EER of 6.13% on the In-The-Wild dataset.

INDEX TERMS Synthesized speech detection, modified group delay, compact convolutional transformer, self- supervised learning, information noise-contrastive estimation loss.

I. INTRODUCTION

The proliferation of synthesized speech technologies, particularly advanced Text-to-Speech (TTS) and Voice Conversion (VC) systems, poses significant security risks through potential misuse in fraud and disinformation. To counter these threats, robust detection systems are needed that combine advanced signal processing and machine learning to identify subtle artifacts in synthesized speech, ensuring reliable discrimination from human speech across diverse conditions [1]. The rapid advancement of voice spoofing has driven the evolution of Automatic Speaker Verification (ASV) systems, with the ASVspoof challenges providing crucial benchmarks through increasingly complex attack scenarios. These datasets enable the development and evaluation of countermeasures, with performance measured primarily using the Equal Error Rate (EER) metric, which balances false acceptance and rejection rates at an optimal threshold to measure detection performance across diverse spoofing conditions [2].

The efficacy of spoofed speech detection systems is fundamentally dependent on the choice of feature representations, which are broadly classified into hand-crafted and learning-based methodologies [3].

Conventional spectral features, particularly those derived from Fourier transform magnitude, prove inadequate for capturing essential temporal speech dynamics, resulting in a critical performance limitation. Although phase information

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

© 2026 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

possesses theoretical potential to address this shortcoming, its inherent instability prevents direct implementation [4]. Modern neural vocoders such as WaveNet, HiFi-GAN, and diffusion-based speech synthesizers excel at generating highfidelity magnitude spectra [5]; however, their phase reconstruction remains an unresolved problem. Most systems either employ heuristic phaserecovery algorithms (e.g., GriffinLim) or learn implicit phase relationships that do not fully capture the causal, physically governed groupdelay structure of the human vocal tract [6].

This challenge has catalyzed the development of enhanced post-processed features, notably the Modified Group Delay (MGD), which improves the phase modeling of contemporary vocoders and enables robust vocal tract characterization [7]. The superior discriminative capability of MGD originates from its ability to capture physiological speech attributes that current speech synthesis techniques cannot accurately reproduce [8]. In contrast to fixed feature engineering, learning-based models dynamically extract relevant representations. Modern self-supervised models like Wav2Vec2 exemplify this approach but incur substantial computational costs due to their large size [9].

The Compact Convolutional Transformer (CCT) addresses this by employing architectural efficiencies like sequence pooling and reduced layers to achieve a compact footprint while maintaining the performance [10]. Fundamentally, both Wav2Vec2 and CCT share a hybrid design, leveraging convolutions for local patterns and transformers for global context.

In the present work, we adopt a hybrid detection framework that combines self-supervised representation learning with supervised binary classification. We first employ a lightweight pre-training stage to train a CCT model, utilizing the strong representation offered by MGD features. As in Wav2Vec2, the training paradigm uses a self-supervised objective, specifically a masked prediction task. The CCT is compelled to internalize the complicated contextual patterns present in authentic human speech by learning to reconstruct randomly masked segments of the MGD representations without ever being exposed to artificial examples.

Contrastive learning is a key technique in self-supervised learning and highly depends on the quality of its positive pairs, which are assumed to be augmented from real speech [11]. Among contrastive learning approaches, the Information Noise-Contrastive Estimation (InfoNCE) loss function and its updated versions have consistently demonstrated the best performance across diverse modalities [12], [13], [14], [15].

So, the masked prediction task in our work is especially well-suited for the InfoNCE loss. It trains the model to minimize similarity with distracting negative samples and maximize the similarity between the predicted representation and the corresponding ground-truth representation in the masked portions. The negative samples in this implementation are purposefully selected from other real speech samples in the same training batch rather than being augmented or artificially created.

The resulting CCT model acts as a powerful encoder, generating highly discriminative embeddings. A simple supervised binary classifier is subsequently trained on these embeddings to separate bona fide from synthesized speech, yielding a robust hybrid system capable of generalizing to novel and evolving synthetic threats.

In summary, the main contributions of this work are as follows:

- This is, to the best of our knowledge, the first work to propose a hybrid detection framework for synthesized speech that integrates MGD features with a self-supervised CCT model followed by supervised binary classification. It provides a computationally efficient and lightweight alternative to massive models like Wav2Vec2, creating a highly effective detector of synthesized speech.
- The use of InfoNCE loss is a key strength of our approach, which powerfully shapes the embedding space. Crucially, we define negative samples not through artificial augmentation but by strategically leveraging other predicted representations within the same batch. This creates a structured and compact latent space for genuine speech. Synthesized speech, with its subtle artifacts, produces embeddings that lie far from this cluster.
- Our method sets a new state-of-the-art on the ASVspoof 2019 LA dataset and exhibits strong generalization capabilities on the more challenging ASVspoof 2021 LA and the In-The-Wild benchmark, validating its practical robustness.

The remainder of this paper is structured as follows: Section II surveys related works on synthesized speech detection. Section III introduces the proposed system. Section IV describes the experimental setup, followed by results discussion in Section V. Section VI provides concluding remarks.

II. LITERATURE REVIEW

The initial wave of research, largely driven by challenges like ASVspoof, focused on identifying spoofing artifacts through digital signal processing. Researchers first analyzed short-term spectral characteristics like Linear Frequency Cepstral Coefficients (LFCC), Mel-Frequency Cepstral Coefficients (MFCC), and Inverted MFCC (IMFCC), operating on the assumption that vocoders left identifiable traces in the spectral envelope [16]. This approach evolved to include techniques like the Linear Frequency Bank (LFB) to address computational issues [17] and Log Power Spectrum (LPS) features with per-band analysis and multi-band fusion to capture the importance of low-frequency information [18]. To overcome the limitation of ignoring temporal information, the field shifted to phase-based features. The MGD function, which effectively captures synthetic artifacts, represents a key development in this category [19]. As attacks grew more sophisticated, feature design evolved to

incorporate long-term spectral analysis like Constant-Q Cepstral Coefficients (CQCC) [20] and prosodic features such as fundamental frequency F0 and energy [21]. The fundamental limitation of these hand-crafted methods has now led to a paradigm shift toward deep learning, where self-supervised models like Wav2Vec2 and WavLM learn discriminative features directly from raw audio. Studies showed the critical importance of early network layers for detection [22], [23].

To detect synthesized speech, researchers have adapted computer vision models by treating speech features as 2D data. Early approaches used models like Light Convolutional Neural Network (LCNN), a base model in ASVspoof challenges, to learn local features [24]. Subsequently, Graph Attention Networks (GAT) [25] and Audio Anti-Spoofing using Integrated Spectro-Temporal graph attention networks (AASIST) [26], captured complex spectral and temporal relationships. The field then progressed to hybrid models like the CCT, which combined the efficiency of Convolutional Neural Networks (CNNs) with the global context of transformers [27]. Recognizing that CNNs struggle with spatial variations, researchers also explored Capsule Networks (CapsNets), which are inherently better at handling changes in input perspective. Architectures like Attention-Based Cascaded Capsule Network (ABC-CapsNet) leveraged this by using attention layers to refine focus and improve detection accuracy [28].

The loss function concerns the model's optimization objective, which directly shapes its learned features. Research evolved through successive refinements in one-class learning to handle unknown attacks. It began with the One-Class Softmax loss (OC-Softmax), which created a compact representation of genuine speech using angular margin separation [29]. To address OC-Softmax's oversight of speaker diversity, the Speaker Attractor Multi-Center One-Class learning (SAMO) method was developed, clustering genuine samples around multiple speaker-specific centers [30]. Later, to leverage synthesized spoofed data, researchers employed a contrastive loss with a specialized mini-batch strategy to optimize the model on training pairs [31]. Unlike the original AASIST, which utilized weighted cross-entropy, AASIST2 implemented the Additive Margin Softmax (AM-Softmax) for better class separation and robustness [32]. Recently, a novel framework integrated Latent Space Refinement (LSR) to model spoofing variability and Latent Space Augmentation (LSA) to generate diverse samples and further enhance generalization [33].

Recent progress in selfsupervised learning highlights the growing adoption of masking strategies to enhance feature learning. In parallel, imagebased methods underscore the need for architectures that capture both local detail and global context. For instance, global featureinjected blindspot networks [34] leverage masking to integrate global information without compromising local reconstruction. This evolution has shifted from CNNdominated models, which focus primarily on local patterns, toward architectures that explicitly model longrange dependencies. Hybrid designs such as nonlocallocal featurecoupled networks [35] address this by combining dilated convolutions for local features with lightweight selfattention for global context. Beyond architectural considerations, self-supervised methods must also contend with two inherent challenges: the scarcity of anomaly samples and the semantic ambiguity of anomalous events. Approaches like temporalspatial decoupled multitask learning [36] respond by constructing multiple proxy subtasks that separately handle spatial and temporal dimensions, thereby extracting more discriminative representations than singletask baselines.

Building on the preceding review, we propose a system that benefits from phase-based feature extraction and self-supervised representation learning within a computationally efficient lightweight architecture. This system then combines the representation-learning advantages of one-class learning with the discriminative precision of supervised classification.

III. METHODOLOGY

Our methodology follows a three-stage pipeline (see **Fig. 1**): (1) Feature extraction using MGD to exploit its phase-sensitive advantages; (2) Self-supervised representation learning, where a CCT model is trained on temporally sampled MGD representations from only real speech using masking and InfoNCE loss to create a structured embedding space; and (3) Supervised binary classification, where a lightweight classifier uses embeddings from the frozen CCT to distinguish real from synthesized speech. This hybrid pipeline enables effective deployment of the complete detection system.

A. PHASE 1: FEATURE EXTRACTION

Most modern vocoders act as the final stage in TTS systems, translating a mel-spectrogram back into a playable audio waveform. A mel-spectrogram is a highly compressed, lossy representation as it reduces the dimensionality of audio by focusing on human perception (the mel scale). The process of mapping a raw audio waveform to a mel-spectrogram involves a magnitude-only representation (often logarithmic) that discards phase information entirely [37]. Reconstructing the phase from these representations is mathematically ill-posed because many different phase structures can map to the same magnitude spectrogram [38].

Unlike the magnitude spectrum, the phase spectrum holds vital temporal details. However, its raw form is unstable due to phase wrapping (bounded between $-\pi$ and π).

This limitation introduces subtle but detectable artifacts in the reconstructed waveform, making it difficult for vocoders to recover the exact phase required for perfect signal reconstruction. The Group Delay (GD) function addresses this by calculating the negative derivative of the phase, which effectively unwraps the phase and extracts the rate of change [39]. To mitigate the unpredictable behavior of the GD, two

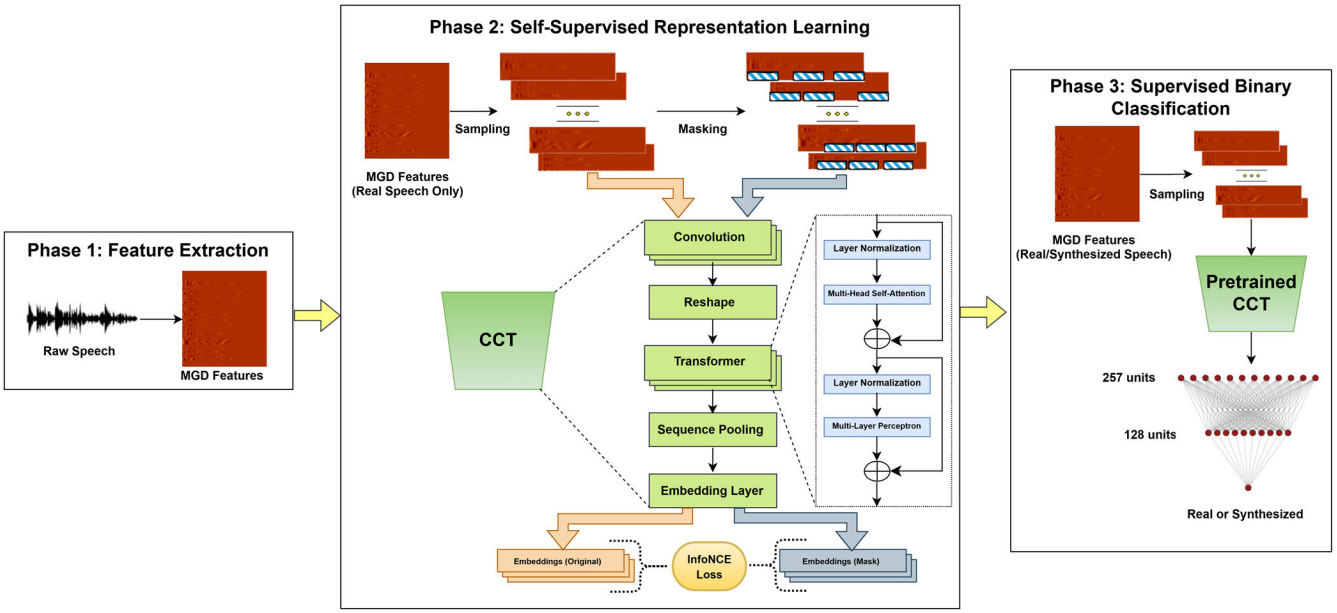


FIGURE 1. The proposed methodology.

parameters, α and γ, are introduced to attenuate spurious spikes and restore the dynamic range of the speech spectrum. The MGD function is defined as [40]:

$$\tau_m(w) = \frac{\tau(w)}{|\tau(w)|} |\tau(w)|^{\alpha} \tag{1}$$

By supposing:

$$\tau(w) = \frac{X_R(w).Y_R(w) + X_I(w).Y_I(w)}{S(w)^{2\gamma}} \tag{2}$$

where $X_R(w)$ and $X_I(w)$ are the real and imaginary parts of the Fourier transform of a signal $x[n]$, respectively. $Y_R(w)$ and $Y_I(w)$ are the real and imaginary parts of the Fourier transform of $nx[n]$, respectively. The parameters α and γ range from 0 to 1. $S(w)$ represents the smoothed version of the magnitude spectrum [39]. The resulting smoothed spectrum mitigates formant peaks, allowing the MGD function to clearly highlight subtle artifacts that expose synthetic speech.

Homomorphic processing is a standard approach for converting speech spectra into effective features. It generates linearly decorrelated features that facilitate diagonal covariance modeling in distributions of speech feature vectors. Thus, the discrete cosine transform (DCT II) is employed to transform the modified group delay function τ_m into cepstral coefficients. Let N_f be the discrete Fourier transform order. The transformation is then defined as [40]:

$$c(n) = \sum_{k=0}^{k=N_f} \tau_m(k) \cos\left(\frac{n(2k+1)\pi}{N_f}\right) \tag{3}$$

In the following example, we analyze a noisy synthesized signal with three resonant frequencies (100 Hz, 250 Hz, 400 Hz), as demonstrated in **Fig. 2**. In this analysis, the log-magnitude spectrum successfully identifies the three resonant frequencies as distinct peaks, accurately representing the spectral envelope. Conversely, the group delay function is rendered ineffective by large spurious spikes and suffers from a corrupted dynamic range due to noise and windowing effects, which together obscure all discriminative features. The MGD rectifies this by employing parametric smoothing and compression to attenuate the formant spikes. It is highly effective in mitigating the effects of noise, yielding a robust representation that retains the important characteristics lost by the other methods.

In addition, examination of MGD cepstral coefficients ($\alpha = 0.4$, $\gamma = 0.4$), computed using a 512-point FFT with 20 ms frames and a 10 ms shift, for real and synthesized speech of the same utterance, as shown in **Fig. 3**, reveals three critical patterns. First, these coefficients are concentrated in the lower FFT points, consistent with the energy compaction property of the DCT [41]. Second, the coefficient magnitudes for synthesized speech consistently exhibit higher values than the genuine ones, indicating greater spectral distortion and phase discontinuity. Third, while the overall spectral trajectory of synthesized speech mirrors the genuine pattern, it lacks fine-grained structural detail. This absence is particularly evident in the higher-order coefficients, revealing the synthesis model's inability to perfectly replicate the complex source-filter interactions of human speech production.

B. PHASE 2: SELF-SUPERVISED REPRESENTATION LEARNING

Motivated by the illustrative examples in **Fig. 2** and **Fig. 3**, which demonstrate the MGD function's unique ability

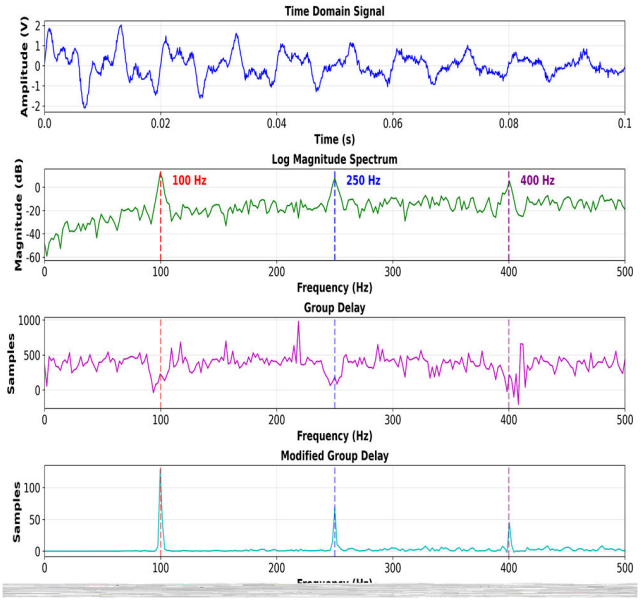


FIGURE 2. A comparative analysis of a noisy synthesized signal featuring three resonances (100 Hz, 250 Hz, 400 Hz), with its corresponding log-magnitude spectrum, GD and MGD (Top to Bottom).

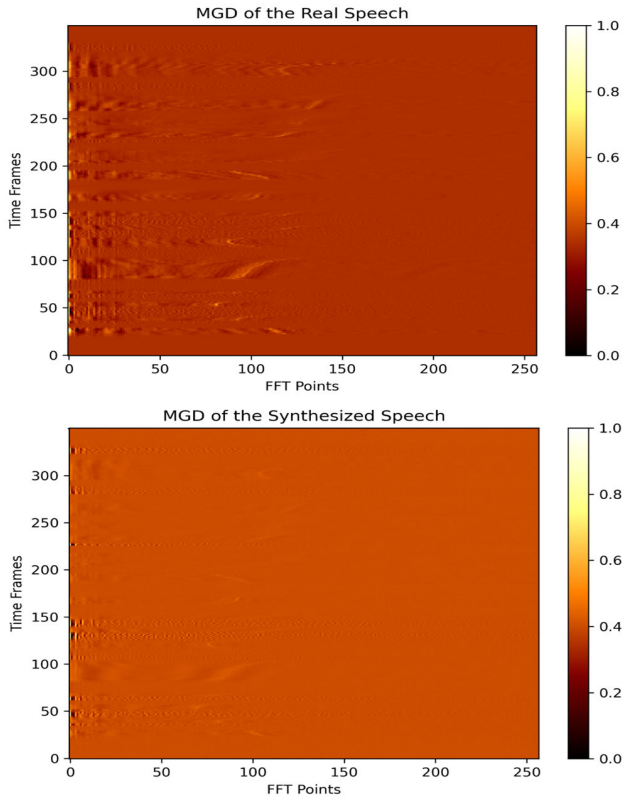


FIGURE 3. MGD cepstral coefficients for real (Top) and synthesized (Down)speech of the same utterance across time-frequency analysis.

to attenuate noise and capture subtle synthesis artifacts, we leverage this robust MGD representation to train a CCT model. Employing a self-supervised, masked prediction objective inspired by Wav2Vec2, the CCT model is trained with an InfoNCE loss to learn the complex structure of authentic human speech.

1) SAMPLING OF MGD FEATURES FROM REAL SPEECH

MGD features are extracted from the real speech utterances of the ASVspoof 2019 LA training partition after removing silence from the beginning and end of utterances. These features are computed on 20 ms windows with a 10 ms shift, using 512-point FFT, forming a 2D representation of dimensions (T, 257), where T is the number of frames for each utterance, and 257 corresponds to the positive half of the 512-point FFT spectrum. To standardize the input dimensionality for the CCT model, a fixed-length segment of 16 consecutive frames (with a 50% overlap) is extracted from each audio's MGD representation (any excess frames being discarded). This uniform sampling strategy is applied across the entire training partition to obtain a substantial dataset of fixed-size 2D feature maps (16×257), optimally structured for training the CCT model.

Each 16-frame segment captures a unique authentic acoustic context from a different part of a long utterance, encompassing variations in phonemes, prosody, and co-articulation.

Fig. 4 depicts the transformation of MGD features from a 3-second real utterance into fixed-size (16×257) samples using 16-frame overlapping segments. The original 300-frame sequence is converted into standardized samples, discarding the final 4 frames to maintain dimensional uniformity.

2) MASKING STRATEGY

Inspired by the masking technique originally introduced in BERT [42] and later implemented in Wav2Vec2 for speech [9], we apply a combined masking strategy with the following distribution for masked patches:

- 80% are replaced with a constant value.
- 10% are replaced with random Gaussian noise ($\mu = 0$, $\sigma = 0.1$) to learn invariance to small, meaningless distortions
- 10% are left unchanged.

We choose a constant mask value of 0.9, which is far from the values mean ($\sim$0.48), to create a harder and more unambiguous pretext task. To force the model to develop a robust understanding of genuine speech, we mask three random segments in the target (16th) frame.

The model must then reconstruct them using context from the previous 15 frames and the unmasked parts of the target itself. This prevents over-reliance on specific cues and enhances generalization for effective detection in the MGD domain.

Fig. 5 demonstrates the masking strategy on MGD features of size 16×257, where the mask is applied only to the 16th frame. The visualization key is as follows: constant masks (0.9…0.9), regions with added noise (0.3…0.6 and 0.4…0.5), and unchanged segments (empty).

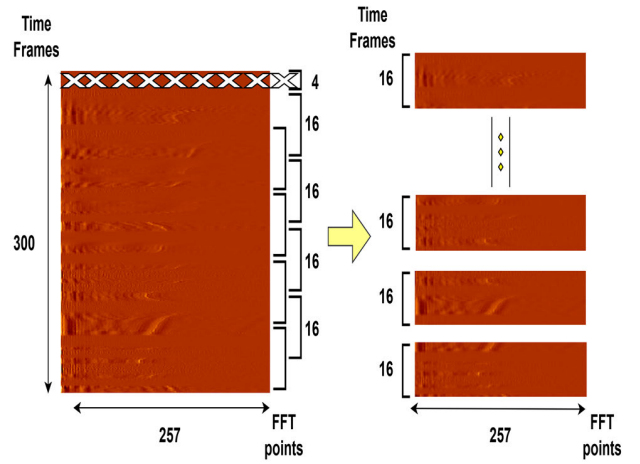


FIGURE 4. Sampling of MGD features extracted from a single real speech.

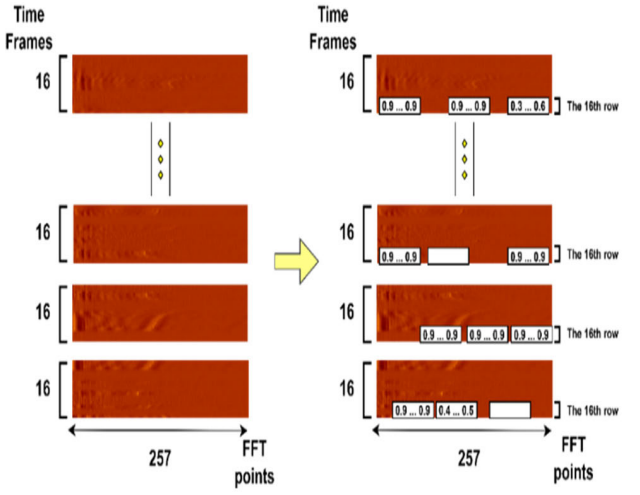


FIGURE 5. Masking strategy.

3) COMPACT CONVOLUTIONAL TRANSFORMER (CCT)

While inspired by self-supervised models like Wav2Vec2 [9], the CCT innovates by prioritizing architectural efficiency for resource-constrained tasks. It retains the proven hybrid CNN-Transformer design, where CNNs capture local acoustic patterns and the transformers model global context. However, CCT introduces two key modifications for efficiency: Sequence Pooling, which summarizes the token sequence to reduce computational cost, and Stochastic Depth, which randomly skips layers during training to improve generalization [10]. To tailor the model for our specific task, the standard classification head is replaced by a linear projection embedding layer, whose dimensionality is configured to match the feature size of MGD representations per time frame. The pseudo-code for the CCT model is presented in **Algorithm 1**.

The CCT function transforms an input batch X into an output y through three sequential stages. First, in the

Algorithm 1 Function CCT

Input: X // Shape: $128 \times 16 \times 257 \times 1$
Output: y // Shape: 128×257

$Z^{(0)} = X$
// — 1. Convolutional Embedding —
for $l = 1$ to 2 **do** // Two convolutional layers
  $Z^{(l)}_{\text{conv}} = \text{ReLU}\left(\text{Conv2d}\left(Z^{(l-1)}\right)\right)$
  $Z^{(l)} = \text{MaxPool}\left(\text{Pad}\left(Z^{(l)}_{\text{conv}}, 1\right); \text{kernel} = 3, \text{stride} = 2\right)$
end for
$Z_0 = \text{reshape_to_tokens}\ (Z^{(2)})$
$Z_0 = Z_0 + E_{\text{pos}}$ // Add learnable positional embeddings
// — 2. Transformer Encoding —
for i = 1 to 3 **do** // Three transformer layers
  $Z'_{i-1} = \text{LayerNorm}\left(Z_{i-1}\right)$
  $A_i = \text{MSA}(Z'_{i-1})$ // MSA with 3 heads
  $Z'_i = \text{StochasticDepth}\ (A_i) + Z_{i-1}$ // First residual connection
  $Z''_i = \text{LayerNorm}\left(Z'_i\right)$
  $F_i = \text{MLP}(Z''_i)$ // two-layer MLP with GELU activation
  $Z_i = \text{StochasticDepth}\ (F_i) + Z'_i$ // Second residual connection
end for
//— 3. Sequence Pooling and Output —
$Z_L = \text{LayerNorm}\ (Z_3)$
$\alpha = \text{softmax}\ (Z_L)$ // Compute attention weights
$z = \alpha Z_L$ // Compute weighted global representation
$y = \text{EmbeddingLayer(z)}$ // Final linear projection
End Function

Convolutional Embedding stage, the input passes through two convolutional layers. Each layer performs a 2D convolution followed by a ReLU activation, zero-padding, and max-pooling to extract hierarchical local features and reduce spatial dimensionality. The resulting feature maps are then reshaped into a sequence of tokens, and learnable positional embeddings are added to retain spatial information.

Second, in the Transformer Encoding stage, this sequence is processed by three transformer layers. Each layer applies layer normalization followed by multi-head self-attention (MSA) to model global dependencies. This is followed by another layer normalization and a two-layer Multilayer Perceptron (MLP) with GELU activation, whose output is also integrated via a second stochastic depth residual connection.

In the last stage, the Sequence Pooling and Output stage, the processed sequence undergoes a final layer normalization. An attention mechanism then computes a weighted sum to aggregate the entire sequence into a single, global representation vector z, which is finally projected to the target output dimensionality through the linear embedding layer to produce y.

4) TRAINING BY THE INFORMATION NOISE-CONTRASTIVE ESTIMATION (INFONCE) LOSS

The self-supervised framework of models, such as Wav2Vec2, is built upon a contrastive learning paradigm. This paradigm involves training a model to solve a pretext task where it must identify the correct latent speech representation for a masked audio segment from a set of negative distractors, thereby acquiring robust, context-aware embeddings without labeled data [9]. Contrastive learning

is particularly valuable for high-dimensional data where traditional losses, like mean-squared error, prove inadequate.

The standard practice in contemporary contrastive learning involves a progression from Noise Contrastive Estimation (NCE) to its more powerful descendant, InfoNCE. The latter variant is expressly recommended for its efficacy in leveraging numerous negative samples to learn more discriminative representations [11]. The masked prediction objective is therefore optimized using an InfoNCE loss [12], which structures the latent space by maximizing the similarity between the predicted embeddings of masked MGD features and their corresponding ground-truth for the same masked positions. It also minimizes similarity with negative distractors. A critical aspect of our implementation involves the negative samples, which are drawn from other predicted embeddings of MGD features within the same batch. The training process operates on batches of size B, where each batch contains paired sets of original and masked MGD features of size 16×257, as illustrated in **Fig. 6**. The CCT model processes both sets to generate corresponding original and masked embeddings. The InfoNCE loss is then computed by contrasting the embeddings at identical masked positions (the hatched regions in **Fig. 6**) across the batch, thereby forcing the model to align original and masked pairs while discriminating against non-matching pairs.

Given original embeddings $E_{original} \in \mathrm{R}^{\mathrm{B}\times\mathrm{L}}$ and masked embeddings $E_{mask} \in \mathrm{R}^{\mathrm{B}\times\mathrm{L}}$, with the indices of masked positions, where B is the batch size and L is the length of inferred embeddings from the CCT model, the procedure first gathers only the masked positions to form $E^m_{original}, E^m_{mask} \in \mathrm{R}^{\mathrm{B}\times\mathrm{N}_m}$, where N_m is the number of masked positions for each sample in the batch.

Rows from the matrices $E^m_{original}$ and E^m_{mask}, each denoted $E_i \in \mathrm{R}^{1\times\mathrm{N}_m}$, are L2-normalized as:

$$\hat{E}_i = \frac{E_i}{\|E_i\|_2} \tag{4}$$

The model then computes a similarity matrix $S \in \mathrm{R}^{\mathrm{B}\times\mathrm{B}}$ using a scaled cosine similarity. Each element in the matrix is defined as:

$$S_{ij} = \frac{\hat{E}_{original,i}\hat{E}^T_{mask,j}}{\tau} \tag{5}$$

where S_{ij} is the similarity at the same masked positions between the i-th sample of the original embedding and the j-th sample of the masked embedding. The temperature parameter τ is a crucial hyperparameter in the InfoNCE loss function. It is used to sharpen the distribution, which helps the model focus on hard negative samples. The loss treats the diagonal elements S_{ii} as positive pairs and all other elements S_{ij} (where $i \neq j$) as negative pairs, using a cross-entropy formulation as follows:

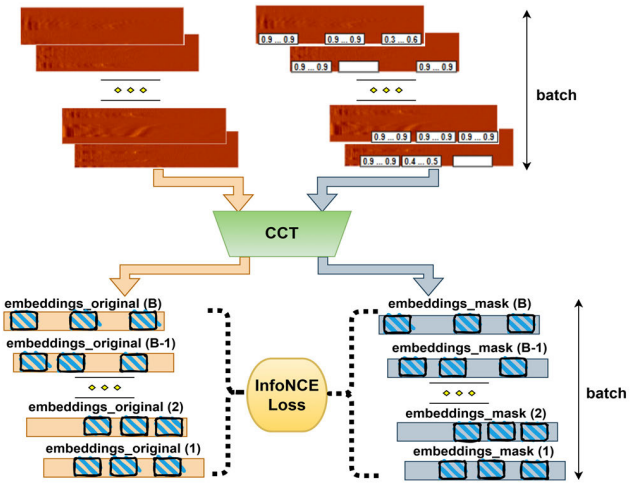


FIGURE 6. Training CCT by InfoNCE loss.

Algorithm 2 Self-Supervised Pre-training of CCT

Input: MGD features from only real speech X_{MGD} (Shape: samples × 16 × 257 × 1)
Output: Trained CCT Model
Preprocess the input X_{MGD}
// — Apply Min-Max Normalization —
X_{norm} = (X_{MGD} - min(X_{MGD})) / (max(X_{MGD}) - min(X_{MGD})) 4:
for e = 1 to Epochs **do**
  X_{norm} = shuffle (X_{norm}) // Shuffle data at the start of each epoch
  for each batch X_B in X_{norm} **do** // X_B shape: 128 × 16 × 257 × 1 for B=128
  // — Apply Masking —
  X_{mask} = apply_masking_strategy (X_B) // Mask 3 random intervals on the 16th frame
  masked_positions = get_current_masked_positions()
  // For each sample in batch
  y_{mask} = CCT (X_{mask}) // Model prediction; y_{mask} shape: 128 × 257
  $\mathrm{y}_{original}$ = CCT (X_B) // Model prediction; $\mathrm{y}_{original}$ shape: 128 × 257
  // — Loss Calculation & Optimization —
  Loss = InfoNCE ($\mathrm{y}_{original}$, y_{mask}, masked_positions)
  // Contrastive loss at masked positions
  optimize (CCT parameters based on Loss)
  // using AdamW optimizer
  end for
16:**end for**
End Algorithm

$$\mathrm{L} = -\frac{1}{B}\sum_{i=1}^{B} log\left(\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp\left(S_{ij}\right)}\right) \tag{6}$$

Thus, the self-supervised training procedure yields a CCT model that learns robust representations of genuine speech. This process is outlined in **Algorithm 2**.

After normalizing the input MGD features, each training batch is processed in two ways: one in its original form and one with randomly masked patches. The model's objective, enforced by an InfoNCE loss, is to make the masked input representation match the original (unmasked) representation at the same masked positions. Consequently, the CCT model

learns context-aware and robust representations suitable for downstream fine-tuning.

C. PHASE 3: SUPERVISED BINARY CLASSIFICATION

After self-supervised training on real speech MGD features, the CCT model serves as a robust encoder. In this phase, a supervised binary classifier is trained on discriminative embeddings generated by the frozen CCT encoder.

Normalized MGD representations are extracted from the audio segments in the ASVspoof 2019 LA training set, which contains both real and synthesized samples. These MGD features are passed through the frozen CCT encoder to produce a 257-dimensional embedding vector for each segment.

The binary classifier is implemented as a lightweight MLP with three dense layers. It takes the 257-dimensional CCT embedding as input to a first hidden layer of 257 units with ReLU activation, followed by a second hidden layer of 128 units with ReLU activation. A dropout layer (rate = 0.25) is applied after the second layer for regularization. The final output layer consists of a single unit with a sigmoid activation function, producing the segment-level probability $s_i \in [0, 1]$ that the i-th segment is synthesized.

During this stage, the parameters of the CCT encoder remain frozen. The binary classifier is trained exclusively using binary cross-entropy loss on labeled embeddings (0 for real and 1 for synthesized).

At the utterance level, the normalized MGD features are first computed for the entire input utterance. This full representation is then partitioned into consecutive, overlapping segments with a 50% overlap to reduce temporal noise sensitivity. Each segment is processed independently to obtain a segment-level detection score s_i. The final utterance-level score U is computed as the mean of all segment scores as follows:

$$U = \frac{1}{N} \sum_{i=1}^{N} s_i \tag{7}$$

where N is the total number of segments. The utterance is then classified as synthesized if $U > \theta$, where $\theta = 0.5$ is the decision threshold; otherwise, it is classified as real. This simple yet effective aggregation strategy leverages the calibrated probabilities from all segments to produce a stable, utterance-level decision.

Therefore, this hybrid architecture synergizes a self-supervised CCT encoder with a lightweight supervised classifier, yielding a detection system that is both accurate and adaptable to novel synthetic threats.

The efficacy of our methodology is rigorously validated across multiple benchmarks, including the ASVspoof 2019 LA evaluation partition, the more challenging ASVspoof 2021 LA dataset, and the In-The-Wild benchmark, demonstrating consistent performance and practical robustness against diverse spoofing attacks.

TABLE 1. Summary of the employed datasets.

Dataset	Subset	#Real	#Spoofed
ASVspoof 2019 LA	Training	2580	22800
	Development	2548	22296
	Evaluation	7355	63882
ASVspoof 2021 LA	Progress	1676	14788
	Evaluation	14816	133360
	Hidden	1960	14966
In-The-Wild		19963	11816

IV. EXPERIMENTAL SETUP

A. DATASET

All systems were trained utilizing the ASVspoof 2019 LA training partition, which contains 25,380 utterances encompassing 6 distinct VC/TTS spoofing attacks [43]. The CCT model is pretrained using a self-supervised learning objective exclusively on the real speech utterances from this partition. During the development of the detection system, the ASVspoof 2019 LA development set, which comprises 24,844 utterances with the same 6 VC/TTS attack variants as the training partition [43], was utilized for model validation throughout the training process.

To assess generalization capability, we conducted a comprehensive evaluation across multiple challenging benchmarks. The systems were tested on: (1) the ASVspoof 2019 LA evaluation set (19LA) featuring 71,237 utterances with 13 diverse spoofing attacks [43]; (2) the ASVspoof 2021 LA dataset (21LA) comprising 181,566 bona fide and spoofed utterances, transmitted through diverse telephony systems, including Voice-over-IP (VoIP) and Public Switched Telephone Network (PSTN) infrastructures, divided into progress, evaluation, and hidden subsets [2]; and (3) the In-The-Wild (ITW) benchmark containing 31,779 utterances collected under realistic and uncontrolled conditions representing the more challenging evaluation scenario [44]. A summary of the employed datasets and their partitions is presented in **Table 1**.

B. EVALUATION METRIC

System performance is evaluated using the Equal Error Rate (EER), which is the standard metric in ASVspoof challenges. The EER is defined as the point on the Detection Error Tradeoff (DET) curve where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. Formally, this occurs at a specific threshold θ on the inference score [45], where:

$$\mathrm{FAR}(\theta) = \frac{\#\{\textit{spoofed trials with score} > \theta\}}{\#\textit{Total spoofed utterances}} \tag{8}$$

$$\mathrm{FRR}(\theta) = \frac{\#\{\textit{genuine trials with score} \leq \theta\}}{\#\textit{Total genuine utterances}} \tag{9}$$

The EER is then calculated as:

$$\mathrm{EER} = \mathrm{FAR}\left(\theta^*\right) = \mathrm{FRR}\left(\theta^*\right) \tag{10}$$

where θ^* is the threshold that satisfies the equality. A lower EER indicates a more robust and accurate detection system, as it signifies a better balance between correctly identifying spoofs and genuine speech.

C. TRAINING DETAILS

For feature extraction, MGD representations were derived, as mentioned earlier, using a 512-point FFT applied to 20 ms frames with a 10 ms hop size and configured with parameters $\alpha = 0.4$ and $\gamma = 0.4$. This framing strategy ensures signal stationarity within each analysis window, fulfilling the fundamental requirement for effective short-term spectral feature extraction.

For each utterance, fixed length 16frame MGD segments were extracted with 50% overlap as inputs to the CCT. Each segment received an independent score; the final utterance-level detection score was computed as the mean of these segment scores. Classification uses a threshold $\theta = 0.5$: utterances with scores strictly above θ are classified as synthesized, while others are classified as real.

The temperature parameter τ in the InfoNCE loss was chosen to be 0.01, and its role is subtle but has a profound impact on the quality of the learned representations.

The models were trained for 100 epochs with a batch size of 128 using the AdamW optimizer, which was configured with a learning rate of 0.001 and a weight decay of 0.0001 for both the CCT model and the final binary classifier.

We employed binary cross-entropy loss because of its proven efficacy in binary classification problems, particularly for discriminating between bona fide and synthesized speech. These hyperparameters were determined via a greedy optimization algorithm. All experiments were executed on an NVIDIA GeForce RTX 4090 GPU.

V. RESULTS

A. COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART SYSTEMS

To assess the overall efficacy of the proposed framework, we conducted comprehensive evaluations across three distinct datasets. All systems, including our proposed method and recent state-of-the-art approaches, were trained exclusively on the ASVspoof 2019 LA training set to ensure fair comparison. We performed comparative analysis against existing methods from the literature that utilized identical training data, as illustrated in **Table 2**.

As evidenced in **Table 2**, contemporary research frequently leverages Wav2Vec2 as a front-end feature extractor, capitalizing on its rich, contextual speech representations to enhance generalization in spoofing detection.

Notable implementations include Wav2Vec2+ Linear [31], which employs multiple linear layers, and Wav2Vec2+AASIST [46] as well as its enhanced variant Wav2Vec2+AASIST2 [32] that integrates separate spectral and temporal graphs via self-attention. In contrast, Wav2Vec2+LogReg [47] shows that even a simple logistic regression classifier can achieve remarkably strong generalization.

TABLE 2. Performance comparison in terms of (EER %) ↓ across multiple datasets. All methods were trained on the ASVspoof 2019 LA training set. Best performing results are in bold, while second-best results are underlined.

System	19LA	21LA	ITW
Wav2Vec2+AASIST [46]	-	0.82	-
Wav2Vec2+Linear [31]	0.22	3.63	16.17
Wav2Vec2+LogReg [47]	0.50	-	7.20
Wav2Vec2+AASIST2 [32]	0.15	1.61	-
Wav2Vec2+STJ-GAT+BLDL [22]	0.06	**0.56**	-
WavLM+AttM [23]	0.65	3.50	-
FTDKD [48]	0.30	2.96	13.01
OCKD [49]	0.39	0.90	7.68
ABC-CapsNet [28]	<u>0.06</u>	-	-
Dense MoE-LoRA [50]	0.29	4.24	9.77
LSR+LSA [33]	0.12	1.05	<u>5.54</u>
Poin-HierNet [51]	0.11	0.94	**4.91**
Our proposed model	**0.0093**	<u>0.61</u>	6.13

Further advancing this paradigm, Wav2Vec2+STJ-GAT +BLDL [22] incorporates a Spectro-Temporal Joint Graph Attention Network (STJ-GAT) with BiLSTM-based Long-term Dependency Learning (BLDL) to simultaneously model temporal and spectral discrepancies, achieving state-of-the-art performance with an EER of 0.56% on 21LA.

Furthermore, the WavLM +AttM framework [23] introduces an attentive merging mechanism to strategically combine embeddings from multiple transformer encoders of WavLM, effectively condensing speech features for enhanced discrimination.

Parallel methodological advances leverage knowledge distillation to create efficient detection models; notably, Frequency-Time Domain Knowledge Distillation (FTDKD) [48] conducts dual-domain distillation to maintain detection efficacy even with low-quality audio inputs, while One-Class Knowledge Distillation (OCKD) [49] learns the distributional characteristics of bona fide speech to improve anomaly detection capabilities.

Recent advancements demonstrate several effective strategies for adapting pre-trained models and refining loss functions. Parameter-efficient methods like Low-Rank Adaptation (LoRA) integrated into Wav2Vec2's attention mechanisms, combined with Mixture-of-Experts (MoE) architectures (Dense MoE-LoRA) [50], show significant promise for domain adaptation with limited data.

Concurrently, novel loss functions have substantially improved discrimination between authentic and spoofed embeddings. The margin loss in ABC-CapsNet [28] guided the model to achieve a notable 0.06% EER on 19LA. In contrast, (LSR +LSA) [33] reduces intra-class variance through prototype refinement, attaining 5.54% EER on ITW. Most impressively, Poin-HierNet [51] constructs domain-invariant hierarchical representations in the Poincaré

space, achieving state-of-the-art performance with 4.91% EER on the challenging ITW. Our proposed methodology establishes new state-of-the-art performance on 19LA with an exceptional EER of 0.0093%, while simultaneously achieving top-tier results on the more challenging 21LA and ITW with EERs of 0.61% and 6.13%, respectively. This sustained effectiveness, across diverse spoofing scenarios and dataset conditions, underscores our system's strong generalization capabilities and reliability.

Beyond quantitative benchmarks, we posit that the system's effectiveness stems from the synergistic interplay between MGD's phase-sensitive representation and the CCT's hybrid local-global processing. MGD amplifies micro-phase inconsistencies inherent to synthetic speech, providing a low-level discriminative signal. The CCT encoder then integrates these local cues through convolutional layers and models their global context via transformers, capturing both fine-grained artifacts and broader contextual anomalies. This complementary interaction yields a latent space that is semantically rich and structurally aligned with natural speech, thereby enhancing generalization across diverse spoofing techniques and acoustic conditions.

B. REPRESENTATION AND ERROR ANALYSIS

To qualitatively evaluate the discriminative capability of our learned representations, we provide a visual analysis. **Fig. 7** presents a 2D Principal Component Analysis (PCA) visualization, demonstrating the discriminative power of the speech representations learned by our method. The plot reveals clear separation between genuine and spoofed classes for the 19LA and 21LA datasets. In contrast, the ITW dataset exhibits significant overlap between classes, indicating a more challenging and less structured feature distribution.

To better understand the specific challenges within ITW, we analyze some individual cases in **Fig. 8**. This figure illustrates representative MGD features for four distinct classification scenarios, providing insight into the model's decision boundaries for ITW samples.

a) True Negative (Real → Real): The MGD features exhibit smooth, continuous transitions between successive frames, particularly evident around the 40th frame. This temporal coherence reflects the natural phase evolution of genuine human speech, which the model correctly identifies.
b) True Positive (Spoofed → Spoofed): The features display pronounced discontinuities and erratic jumps in value across frames. This artifact is a characteristic signature of the underlying vocoder, whose synthesis process introduces phase inconsistencies that are captured by the MGD function.
c) False Positive (Real → Spoofed): This misclassification involves an utterance with short, intermittent bursts of speech. The frequent and abrupt transitions between speech and silence appear to generate phase patterns that the model conflates with synthetic artifacts, leading to the erroneous rejection of a bona fide sample.
d) False Negative (Spoofed → Real): This sample represents a high-quality spoofing attack with minimal audible artifacts. The corresponding MGD features show a convergence of values and relatively smooth frame-to-frame transitions, closely mimicking the temporal structure of real speech. This successful deception underscores the challenge posed by advanced synthesis systems that minimize spectral distortion and phase discontinuity.

C. COMPUTATIONAL EFFICIENCY ANALYSIS

To conduct a rigorous efficiency comparison, we benchmark our CCT model against large-scale pre-trained speech models, specifically HuBERT [52] and multilingual Wav2Vec2-XLS-R [53] variants. These models share convolution-transformer hybrid architectures similar to our CCT design and are known to learn robust, generalizable speech representations that generalize across diverse tasks. Notably, Wav2Vec2-based approaches have demonstrated particular effectiveness in synthesized speech detection [22], [32], [46], [31], [47].

For fair comparison, all models were modified to accept 16×257 MGD spectrograms as input and output 257-dimensional embeddings, which then serve as input to the proposed supervised binary classifier for final detection.

We computed both Floating Point Operations (FLOPs), reported in GigaFLOPs (GFLOPs), and average inference times for the entire pipeline by substituting the CCT model with each alternative approach. FLOPs were calculated using an approximately 4second utterance from the ASVspoof 2019 LA evaluation set, while inference times were averaged across 100 randomly selected utterances from the same set on the same NVIDIA GeForce RTX 4090 GPU.

As shown in **Table 3**, our proposed model demonstrates dramatic efficiency advantages over large pretrained speech models. With only 0.91M parameters, our model achieved drastic reduction in scale which was reflected in a memory footprint of just 0.00347 GB, which was 100 to 2500 times smaller than the baselines, a critical advantage for deployment in memory-constrained environments.

In terms of computational demand, our model required only 7.02 GFLOPs to process a 4-second utterance, representing 4 to 62 times fewer operations than the pre-trained models. This efficiency stems directly from our architectural choice to process fixedlength MGD spectrogram segments, thereby avoiding the quadratic complexity of full-sequence attention over raw audio.

In contrast, our model achieved an average inference time of 45.2 ms, operating 3 to 26 times faster than the benchmarked models. The notably low standard deviation ($\pm$3.1 ms) indicates consistent and predictable latency, a key requirement for real-time systems.

This demonstrates that for synthesized speech detection, our lightweight architecture provides substantially better efficiency than adapting large general-purpose models, even

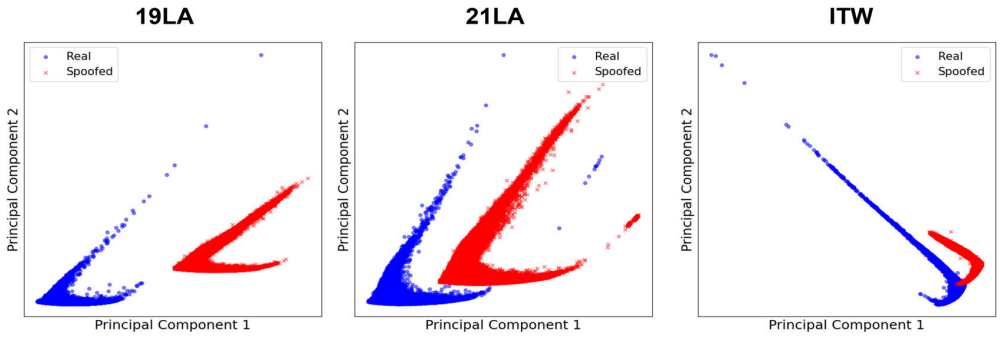


FIGURE 7. 2D PCA plot of real (blue) vs. spoofed (red) speech for the 19LA, 21LA, and ITW datasets.

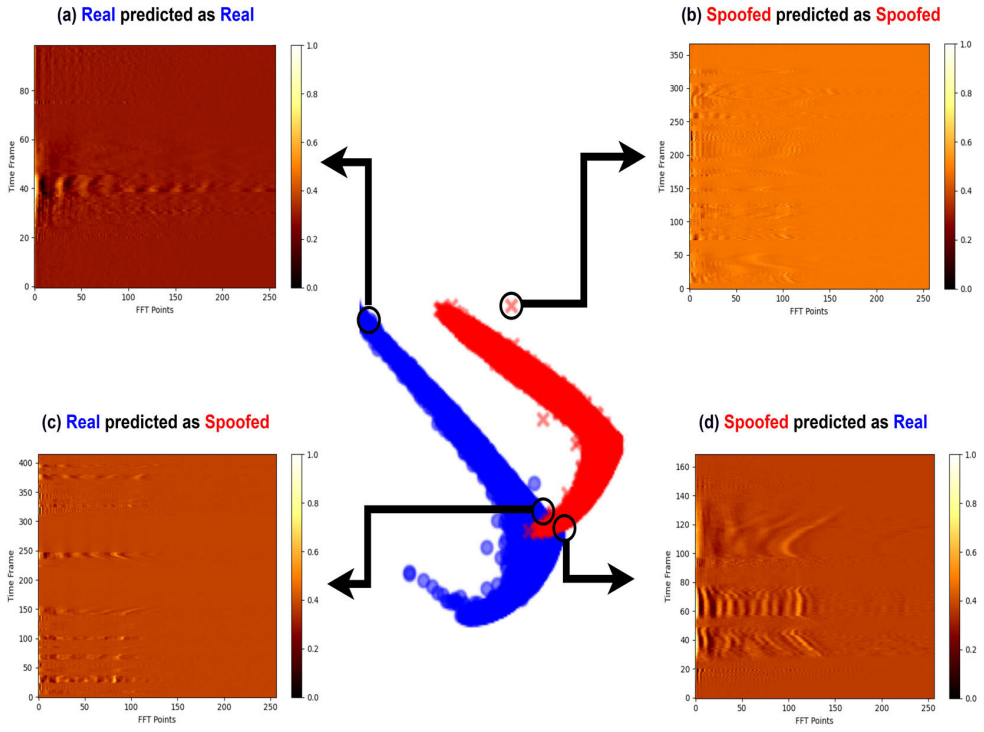


FIGURE 8. MGD features for some classification scenarios on ITW: (a) True Negative, (b) True Positive, (c) False Positive.

when those models offer broad speech representation capabilities.

D. ROBUSTNESS ANALYSIS

To evaluate the robustness of the proposed system against channel degradation, we conducted a series of experiments with Additive White Gaussian Noise (AWGN). The Signal-to-Noise (SNR) was varied from 0 dB to 12 dB in 3 dB increments, covering conditions from severe to moderate degradation. These experiments were performed across 19LA, 21LA, and ITW to assess generalization capability under noisy conditions as shown in **Fig. 9**.

The robustness evaluation under additive noise reveals a consistent performance hierarchy across all SNR levels. The 19LA dataset demonstrates exceptional noise resilience, with its EER decreasing sharply from 32.97% at 0 dB to 1.12% at 12 dB, indicating that the system can recover near-perfect performance with only moderate improvements in signal quality. The 21LA dataset, while more affected by noise, also shows strong recovery, particularly between 6 dB and 12 dB.

The ITW dataset demonstrated the highest sensitivity to noise degradation, with its EER remaining critically high (15.76%) even at 12 dB SNR.

This behavior highlights that environmental noise disproportionately affects the detection of in-the-wild spoofing attacks, presenting a distinct challenge for real-world deployment; nevertheless, it confirms that the proposed methodology remains fundamentally effective in noisy conditions, with performance systematically approaching clean-audio baselines as SNR increases.

The proposed system exhibits strong generalization largely because MGD features supply a noise-robust, phase-based

TABLE 3. Computational efficiency comparison: Parameter count, model size, GFLOPs for a 4second utterance, and average inference time (± Standard deviation) across 100 random samples.

Model	#Params	Size (GB)	GFLOPs	Inference Time (ms)
HuBERT (Base)*	95M	0.378	27.79	152.7 ± 8.5
HuBERT (Large)*	317M	1.26	71.81	293.4 ± 12.3
HuBERT (X-Large)*	964M	3.85	200.55	878.6 ± 24.7
Wav2Vec2-XLS-R (0.3B)*	317M	1.27	71.81	297.2 ± 11.9
Wav2Vec2-XLS-R (1B)*	956M	3.86	200.55	887.4 ± 26.1
Wav2Vec2-XLS-R (2B)*	2162M	8.65	438.71	1185.8 ± 34.2
Our proposed model	0.91M	0.00347	7.02	45.2 ± 3.1

*The models are available at https://huggingface.co/facebook.

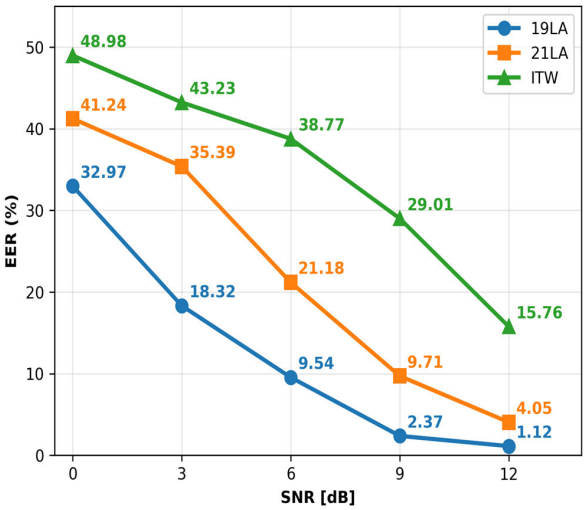


FIGURE 9. EER vs SNR across 19LA, 21LA, and ITW datasets.

foundation that retains discriminative information even at low SNR levels, while the CCT's hybrid local-global architecture actively explores and models the distinct patterns of genuine versus synthesized speech that remain discernible despite additive noise interference. However, the persistently elevated EER on the ITW dataset indicates that real-world noise encompasses convolutional distortions, reverberation, and non-stationary interference that can fundamentally alter phase relationships. Consequently, while the system can reliably reject noise-mimicking artifacts in controlled laboratory data, it encounters greater difficulty when noise and spoofing cues interact in complex, unpredictable ways.

E. ABLATION STUDIES

To validate design choices, we performed ablation studies across features (MGD alternatives), self-supervised models (CCT alternatives), and losses (InfoNCE alternatives). We then optimized key hyperparameters that strongly influence performance, specifically the MGD parameters (α, γ) and the temperature τ in the InfoNCE loss function.

1) FEATURE ABLATION

We evaluated the proposed MGD features against the most common acoustic features in synthesized speech detection: MFCC, IMFCC, CQCC, and LFCC. The input dimensionality was standardized to 80 coefficients per time frame for all feature types, where both the CCT model and the final classifier were adjusted to accept this 80-dimensional input. This configuration is aligned with established practices in audio deepfake detection, where 80-dimensional LFCC features have been demonstrated as a suitable choice for domain generalization [54].

The comparative analysis of acoustic features reveals a clear performance hierarchy, as illustrated in **Fig. 10**, empirically validating the theoretical strengths of the MGD features. As anticipated, traditional magnitude-based features like MFCC yielded the highest EER, as they are fundamentally aligned with the human perception model that modern neural vocoders are designed to mimic. Similarly, IMFCC, while offering enhanced sensitivity to high-frequency artifacts, demonstrated only a marginal improvement over standard MFCC. CQCC provided substantially better performance by adaptively combining high frequency resolution at low frequencies with high temporal resolution at high frequencies. LFCC also proved highly effective, with its uniform linear scaling adeptly capturing specific wide-band vocoder artifacts, showing good generalization.

Ultimately, the superior performance of MGD features stems from their unique focus on phase discontinuities, a vulnerability inherent in vocoders optimized solely for magnitude spectrum fidelity.

This finding confirms that advancing spoofing detection requires fundamentally new, physically-grounded representations, not merely refinements to conventional magnitude-based features.

2) ARCHITECTURAL COMPARISON

We evaluated the CCT model against state-of-the-art self-supervised audio models. To ensure a fair comparison, we adapted all baselines to process MGD features and kept the final binary classifier fixed. Each model, including HuBERT [52] and multilingual Wav2Vec2-XLS-R [53] variants, was modified with an input projection layer to accept 16×257 MGD spectrograms and an output embedding layer to produce 257-dimensional representations, aligning with our CCT's input/output dimensions. During training,

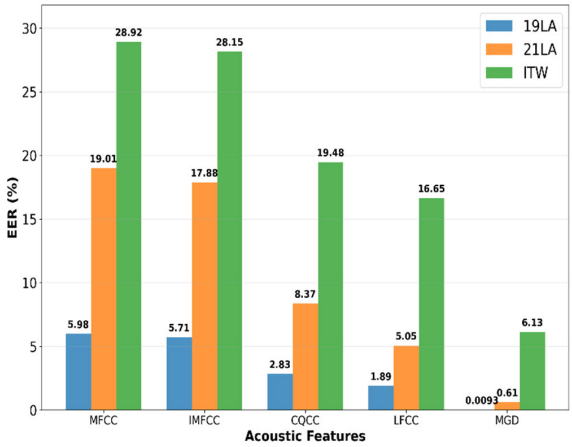


FIGURE 10. EER comparison of different acoustic features across 19LA, 21LA, and ITW datasets.

TABLE 4. Performance comparison in terms of (EER %) ↓ across multiple datasets for systems trained with different self-supervised methods. Best performing results are in **bold**, while second-best results are underlined.

Model	#Params	19LA	21LA	ITW
HuBERT (Base)*	95M	0.39	4.98	16.14
HuBERT (Large)*	317M	0.28	4.63	15.97
HuBERT (X-Large)*	964M	0.21	3.37	13.50
Wav2Vec2-XLS-R (0.3B)*	317M	0.35	4.52	13.88
Wav2Vec2-XLS-R (1B)*	956M	0.18	3.20	11.17
Wav2Vec2-XLS-R (2B)*	2162M	<u>0.12</u>	<u>2.76</u>	<u>9.85</u>
Proposed (CCT)	0.91M	**0.0093**	**0.61**	**6.13**

*The models are available at https://huggingface.co/facebook.

we froze the pre-trained layers of these large models and trained only the added projection and embedding layers.

As shown in **Table 4**, increasing model scale yields negligible gain across all baselines. Scaling HuBERT from 95M to 964M parameters reduces EER on ITW by only 2.64% (16.14% → 13.50%), while scaling Wav2Vec2-XLS-R from 317M to 2,162M parameters improves by just 4.03% (13.88% → 9.85%). Consequently, even the largest baseline (Wav2Vec2-XLS-R2B) underperforms our 0.91M-parameter CCT by a factor of 13× on 19LA (0.12% vs. 0.0093%) and 1.6× on ITW (9.85% vs. 6.13%).

This performance gap stems from a fundamental misalignment in training objectives. HuBERT and Wav2Vec2-XLS-R are optimized to reconstruct masked latent features derived from raw audio, whereas the CCT is trained exclusively within the MGD domain to detect the phase-reconstruction failures that characterize neural vocoders.

This task-specific optimization enables the CCT to learn a separable manifold for spoofing artifacts, avoiding the representation misalignment that limits adapted large models and demonstrating that architectural specialization outperforms scaling of generic pre-trained models.

3) LOSS FUNCTION ABLATION

To rigorously evaluate our choice of contrastive objective, we conducted an ablation study comparing widely-used

TABLE 5. Performance comparison in terms of (EER %) ↓ across multiple datasets for systems trained with different losses. Best performing results are in **bold**, while second-best results are underlined.

Loss	Parameters	19LA	21LA	ITW
Triplet Loss	Margin = 0.5	0.45	2.77	19.26
NT-Xent	$\tau = 0.01$	0.32	2.63	15.97
SupCon	$\tau = 0.01$	0.018	1.31	11.94
RINCE	$\tau = 0.01$	**0.0091**	<u>0.93</u>	<u>9.28</u>
Proposed (InfoNCE)	$\tau = 0.01$	<u>0.0093</u>	**0.61**	**6.13**

alternative losses against the proposed InfoNCE formulation, with results detailed in **Table 5**. The Triplet Loss (margin = 0.5), which relies on relative distances between anchor, positive, and negative samples [55], achieved an EER of 0.45% on 19LA and exhibited poor generalization to unseen datasets, indicating its learned representations lack robustness across diverse conditions.

The NT-Xent loss ($\tau = 0.01$), a symmetric contrastive objective popularized by SimCLR [56], improved upon Triplet Loss (0.32% on 19LA) but still fell short, suggesting that its uniform treatment of negative samples is less effective than InfoNCE's discriminative weighting.

The Supervised Contrastive Loss (SupCon) ($\tau = 0.01$), which leverages speaker labels to cluster same-class samples [57], performed remarkably well on the in-domain 19LA set (0.018%), closely approaching our result; however, its reliance on labeled data limits applicability in self-supervised settings, and its performance degraded on more challenging benchmarks (1.31% on 21LA, 11.94% on ITW).

Interestingly, RINCE ($\tau = 0.01$), a robust variant of InfoNCE designed to handle noisy negative samples [58], closely matched our method on 19LA (0.0091%) and 21LA (0.93%), but exhibited a noticeable drop on the realistic ITW dataset (9.28%), implying that its robustness mechanisms offer little advantage in our clean, curated training data. In contrast, our InfoNCE implementation ($\tau = 0.01$) delivered balanced performance across all three evaluation datasets.

These results demonstrate that InfoNCE's ability to contrast each masked embedding against multiple in-batch negative samples produces a more generalizable and discriminative latent space, optimally suited for detecting synthetic artifacts across both controlled and in-the-wild conditions.

4) MGD PARAMETER ABLATION

The feature extraction employs MGD representations computed using a 512-point FFT with 20 ms frames and a 10 ms hop size, parameterized by coefficients α and γ. We systematically evaluate the impact of these parameters on detection performance, with comprehensive results detailed in **Table 6**.

At low parameter values ($\alpha = 0.2$, $\gamma = 0.2$), inadequate spike suppression and spectral smoothing yield noisy representations that challenge the CCT model, forcing it to prioritize noise filtering over discriminative feature learning. This combination yields poor performance on 19LA,

TABLE 6. Performance comparison in terms of (EER %) ↓ across multiple datasets for systems trained with different mgd parameters. Best performing results are in bold, while second-best results are underlined.

α	γ	19LA	21LA	ITW
0.2	0.2	2.1538	4.65	38.57
0.2	0.4	1.0879	2.58	37.12
0.2	0.6	<u>0.0105</u>	1.12	28.07
0.2	0.8	4.2502	8.87	40.59
0.4	0.2	0.0567	0.95	9.98
0.4	0.4	**0.0093**	<u>0.61</u>	<u>6.13</u>
0.4	0.6	0.0174	**0.58**	7.57
0.4	0.8	2.1783	9.40	29.82
0.6	0.2	2.4512	11.15	18.74
0.6	0.4	0.0121	1.07	**6.08**
0.6	0.6	1.6815	3.20	18.15
0.6	0.8	4.7322	19.65	31.02
0.8	0.2	3.2096	20.60	40.09
0.8	0.4	1.7896	10.35	19.81
0.8	0.6	4.1129	17.55	39.48
0.8	0.8	6.1283	35.10	52.65

TABLE 7. Performance comparison in terms of (EER %) ↓ across multiple datasets for systems trained with different values of the temperature parameter in the infonce loss. Best performing results are in bold, while second-best results are underlined.

τ	19LA	21LA	ITW
0.001	**0.0019**	8.36	44.20
0.005	<u>0.0062</u>	**0.41**	35.57
0.01	0.0093	<u>0.61</u>	**6.13**
0.05	0.0097	5.67	<u>17.46</u>
0.1	3.3568	13.58	31.23

21LA, and ITW with EERs of 2.15%, 4.65%, and 38.57%, respectively.

Conversely, high parameter values ($\alpha = 0.8$, $\gamma = 0.8$) produce excessive smoothing that eliminates crucial discriminative artifacts, resulting in oversimplified features devoid of synthetic speech fingerprints. This leads to substantially degraded performance on 19LA, 21LA, and ITW with EERs of 6.13%, 35.10%, and 52.65%, respectively.

Asymmetric parameter combinations reveal distinct trade-offs. Low α with high γ preserves temporal spikes while oversmoothing spectral information, whereas high α with low γ excessively suppresses spikes while retaining spectral details.

The balanced configuration ($\alpha = 0.4$, $\gamma = 0.4$) provides optimal trade-offs between spike suppression and spectral smoothing, delivering clean, discriminative features that preserve phasebased synthesis artifacts while suppressing formant-dominated spikes. This configuration achieves optimal balanced performance on 19LA, 21LA, and ITW with EERs of 0.0093%, 0.61%, and 6.13%, respectively.

This analysis confirms that careful parameter selection is critical for MGD-based synthetic speech detection, with balanced values providing optimal performance across diverse evaluation conditions.

5) TEMPERATURE PARAMETER ABLATION

The analysis of the temperature parameter τ in the InfoNCE loss reveals a crucial trade-off between precision and generalization in our detection system, as illustrated in **Table 7**.

At low temperatures ($\tau = 0.001$), we observe exceptional performance on 19LA with an EER of 0.0019%. However, the model yields the poorest performance on ITW with an EER of 44.20%, indicating severe overfitting to laboratory conditions.

Conversely, higher temperatures ($\tau = 0.1$) lead to progressive performance degradation across all datasets as similarity distributions become excessively diffuse.

The optimal balance emerges at ($\tau = 0.01$), which maintains strong detection on 19LA with an EER of 0.0093%, while achieving respectable performance on 21LA and ITW with EERs of 0.61% and 6.13%, respectively.

This demonstrates that temperature selection fundamentally controls the bias-variance trade-off, with ($\tau = 0.01$) providing optimal gradient weighting that balances sensitivity to clear spoofing artifacts with robustness to subtle real-world variations.

VI. CONCLUSION

This work has established a comprehensive hybrid detection framework for synthesized speech detection, combining self-supervised representation learning with supervised binary classification. Our approach first employs a lightweight pre-training stage that synergistically integrates MGD feature extraction with CCT architecture, followed by supervised classification. The phase-based MGD features proved instrumental in capturing subtle artifacts that are notoriously difficult for speech synthesizers to replicate, while the self-supervised CCT model learned robust representations of genuine speech through masked contrastive learning. Our methodology demonstrated exceptional performance across multiple benchmarks, achieving a remarkable EER of 0.0093% on 19LA, 0.61% on the more challenging 21LA, and 6.13% on the diverse ITW.

These results not only validate the efficacy of our approach but also set a new standard for generalized audio deepfake detection. Our framework effectively balances high precision on controlled laboratory data with strong robustness against real-world variations and unseen spoofing attacks.

To improve performance in noisy conditions, future efforts could involve integrating speech enhancement front-ends as a preprocessing module, alongside data augmentation with a more diverse set of noise profiles and room impulse responses. Furthermore, exploring self-supervised learning paradigms that learn invariant representations across clean, enhanced, and noisy domains could improve generalization.

DATA AVAILABILITY

The ASVspoof 2019 LA dataset is available at https://datashare.ed.ac.uk/handle/10283/3336. The ASVspoof 2021 LA dataset is available at https://zenodo.org/records/4837263. The In-The-Wild dataset is available at https://deepfake-total.com/in_the_wild

REFERENCES

[1] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "A survey on speech deepfake detection," *ACM Comput. Surv.*, vol. 57, no. 7, pp. 1–38, Feb. 2025, doi: 10.1145/3714458.

[2] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2507–2522, 2023, doi: 10.1109/TASLP.2023.3285283.

[3] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, "Audio deepfake detection: What has been achieved and what lies ahead," *Sensors*, vol. 25, no. 7, p. 1989, Mar. 2025, doi: 10.3390/s25071989.

[4] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. Interspeech*, Sep. 2015, pp. 2052–2056, doi: 10.21437/interspeech.2015-465.

[5] S. Bhushan, V. Prakash Mishra, V. Rishiwal, S. Arunkumar, and U. Agarwal, "Advancing text-to-speech systems for low-resource languages: Challenges, innovations, and future directions," *IEEE Access*, vol. 13, pp. 155729–155758, 2025, doi: 10.1109/ACCESS.2025.3605236.

[6] Y. Ai and Z.-H. Ling, "Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096553.

[7] H. A. Murthy and V. Gadde, "The modified group delay function and its application to phoneme recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Proceedings. (ICASSP)*, Apr. 2003, pp. 68–71, doi: 10.1109/ICASSP.2003.1198718.

[8] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its applications in speech technology," *Sadhana*, vol. 36, no. 5, pp. 745–782, Oct. 2011, doi: 10.1007/s12046-011-0045-1.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), Dec.*, 2020, pp. 12449–12460.

[10] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[11] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020, doi: 10.3390/technologies9010002.

[12] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[13] S. Zare and H. Van Nguyen, "Evaluating and improving domain invariance in contrastive self-supervised learning by extrapolating the loss function," *IEEE Access*, vol. 11, pp. 137758–137768, 2023, doi: 10.1109/ACCESS.2023.3339775.

[14] J. Abbas, C. Zhang, and B. Luo, "EnsCL-CatBoost: A strategic framework for software requirements classification," *IEEE Access*, vol. 12, pp. 127614–127628, 2024, doi: 10.1109/ACCESS.2024.3452011.

[15] H. Zou, M. Lu, and Q. Zhang, "Efficient unsupervised semantic similarity through dual-masked prompting," *IEEE Access*, vol. 13, pp. 163999–164007, 2025, doi: 10.1109/ACCESS.2025.3611338.

[16] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," in *Proc. Interspeech*, Sep. 2015, pp. 2087–2091, doi: 10.21437/interspeech.2015-472.

[17] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of audio deepfake detection," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Nov. 2020, pp. 132–137, doi: 10.21437/odyssey.2020-19.

[18] Y. Zhang, W. Wang, and P. Zhang, "The effect of silence and dual-band fusion in anti-spoofing system," in *Proc. Interspeech*, Aug. 2021, pp. 4279–4283, doi: 10.21437/interspeech.2021-1281.

[19] S. Singh, K. Khoria, and H. A. Patil, "Modified group delay cepstral coefficients for voice liveness detection," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 146–150, doi: 10.23919/EUSIPCO54536.2021.9616343.

[20] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Jun. 2016, pp. 283–290, doi: 10.21437/odyssey.2016-41.

[21] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation features," in *Proc. INTERSPEECH*, Aug. 2023, pp. 3844–3848, doi: 10.21437/INTERSPEECH.2023-1254.

[22] H. Wu, W. Guo, Z. Zhang, W. Zhao, S. Peng, and J. Zhang, "Spoofing speech detection by modeling local spectro-temporal and long-term dependency," in *Proc. Interspeech*, Sep. 2024, pp. 507–511, doi: 10.21437/INTERSPEECH.2024-251.

[23] Z. Pan, T. Liu, H. B. Sailor, and Q. Wang, "Attentive merging of hidden embeddings from pre-trained speech model for anti-spoofing detection," in *Proc. Interspeech*, Sep. 2024, pp. 2090–2094, doi: 10.21437/interspeech.2024-1472.

[24] G. Lavrentyeva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlov, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. Interspeech*, Sep. 2019, pp. 1033–1037, doi: 10.21437/interspeech.2019-1768.

[25] H. Tak, J.-W. Jung, J. Patino, M. Todisco, and N. Evans, "Graph attention networks for anti-spoofing," in *Proc. Interspeech*, Aug. 2021, pp. 2356–2360, doi: 10.21437/interspeech.2021-993.

[26] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6367–6371, doi: 10.1109/ICASSP43922.2022.9747766.

[27] E. R. Bartusiak and E. J. Delp, "Synthesized speech detection using convolutional transformer-based spectrogram analysis," in *Proc. 55th Asilomar Conf. Signals, Syst., Comput.*, Oct. 2021, pp. 1426–1430, doi: 10.1109/IEEECONF53345.2021.9723142.

[28] T. M. Wani, R. Gulzar, and I. Amerini, "ABC-CapsNet: Attention based cascaded capsule network for audio deepfake detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 2464–2472, doi: 10.1109/CVPRW63382.2024.00253.

[29] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Process. Lett.*, vol. 28, pp. 937–941, 2021, doi: 10.1109/LSP.2021.3076358.

[30] S. Ding, Y. Zhang, and Z. Duan, "SAMO: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094704.

[31] X. Wang and J. Yamagishi, "Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10094779.

[32] Y. Zhang, J. Lu, Z. Shang, W. Wang, and P. Zhang, "Improving short utterance anti-spoofing with Aasist2," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 11636–11640, doi: 10.1109/ICASSP48485.2024.10448049.

[33] W. Huang, Y. Gu, Z. Wang, H. Zhu, and Y. Qian, "Generalizable audio deepfake detection via latent space refinement and augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5, doi: 10.1109/ICASSP49660.2025.10888328.

[34] D. Wang, L. Zhuang, L. Gao, X. Sun, and X. Zhao, "Global feature-injected blind-spot network for hyperspectral anomaly detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024, doi: 10.1109/LGRS.2024.3449635.

[35] D. Wang, L. Ren, X. Sun, L. Gao, and J. Chanussot, "Nonlocal and local feature-coupled self-supervised network for hyperspectral anomaly detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 18, pp. 6981–6993, 2025, doi: 10.1109/JSTARS.2025.3542457.

[36] D. Wang, Q. Wang, Q. Hu, and K. Wu, "Temporal-spatial decoupled self-supervised multi-task learning for video anomaly detection and localization in intelligent transportation surveillance systems," *IEEE Trans. Intell. Transp. Syst.*, early access, Apr. 21, 2025, doi: 10.1109/TITS.2025.3559166.

[37] G. Hu, H. Tan, and R. Li, "A mel spectrogram enhancement paradigm based on CWT in speech synthesis," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Aug. 2024, pp. 401–405, doi: 10.1109/ialp63756.2024.10661192.

[38] M. Allain, S. Aslan, W. Coene, S. Dirksen, J. Dong, J. Flamant, M. Iwen, F. Krahmer, T. V. Leeuwen, O. Melnyk, A. Menzel, A. P. Mosk, V. Nikitin, P. Salanevich, G. Plonka, and M. Wellershoff, "Phasebook: A survey of selected open problems in phase retrieval," *Sampling Theory, Signal Process., Data Anal.*, vol. 23, no. 2, pp. 6–8, Oct. 2025, doi: 10.1007/s43670-025-00117-0.

[39] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Process.*, vol. 40, no. 9, pp. 2281–2289, Sep. 1992, doi: 10.1109/78.157227.

[40] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 1, pp. 190–202, Jan. 2007, doi: 10.1109/TASL.2006.876858.

[41] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE Trans. Comput.*, vol. C-23, no. 1, pp. 90–93, Jan. 1974, doi: 10.1109/T-C.1974.223784.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[43] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. H. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. Interspeech*, Sep. 2019, pp. 1008–1012, doi: 10.21437/interspeech.2019-2249.

[44] N. Müller, P. Czempin, F. Diekmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Proc. Interspeech*, Sep. 2022, pp. 2783–2787, doi: 10.21437/interspeech.2022-108.

[45] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Interspeech*, Aug. 2017, pp. 2–6, doi: 10.21437/interspeech.2017-1111.

[46] H. Tak, M. Todisco, X. Wang, J.-W. Jung, J. Yamagishi, and N. Evans, "Automatic speaker verification spoofing and deepfake detection using Wav2vec 2.0 and data augmentation," in *Proc. Speaker Lang. Recognit. Workshop (Odyssey)*, Jun. 2022, pp. 112–119, doi: 10.21437/odyssey.2022-16.

[47] O. Pascu, A. Stan, D. Oneata, E. Oneata, and H. Cucu, "Towards generalisable and calibrated audio deepfake detection with self-supervised representations," in *Proc. Interspeech*, Sep. 2024, pp. 4828–4832, doi: 10.21437/interspeech.2024-1302.

[48] B. Wang, Y. Tang, F. Wei, Z. Ba, and K. Ren, "FTDKD: Frequency-time domain knowledge distillation for low-quality compressed audio deepfake detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 4905–4918, 2024, doi: 10.1109/TASLP.2024.3492796.

[49] J. Lu, Y. Zhang, W. Wang, Z. Shang, and P. Zhang, "One-class knowledge distillation for spoofing speech detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 11251–11255, doi: 10.1109/ICASSP48485.2024.10446270.

[50] J. Laakkonen, I. Kukanov, and V. Hautamäki, "Mixture of low-rank adapter experts in generalizable audio deepfake detection," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Oct. 2025, pp. 2211–2216, doi: 10.1109/APSIPAASC65261.2025.11249190.

[51] M. Yang, Y. Gu, Q. He, Y. Li, P. Zhang, Y. Chen, Z. Wang, H. Zhu, J. Liu, and W. Wang, "Generalizable audio deepfake detection via hierarchical structure learning and feature whitening in Poincaré sphere," in *Proc. Interspeech*, Aug. 2025, pp. 2255–2259, doi: 10.21437/interspeech.2025-1594.

[52] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021, doi: 10.1109/TASLP.2021.3122291.

[53] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, Sep. 2022, pp. 2278–2282, doi: 10.21437/interspeech.2022-143.

[54] Y. Xie, H. Cheng, Y. Wang, and L. Ye, "Domain generalization via aggregation and separation for audio deepfake detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 344–358, 2024, doi: 10.1109/TIFS.2023.3324724.

[55] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823, doi: 10.1109/CVPR.2015.7298682.

[56] W. ågren, "The NT-xent loss upper bound," 2022, *arXiv:2205.03169*.

[57] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18661–18673.

[58] D. T. Hoffmann, N. Behrmann, J. Gall, T. Brox, and M. Noroozi, "Ranking info noise contrastive estimation: Boosting contrastive learning via ranked positives," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2022, pp. 897–905, doi: 10.1609/aaai.v36i1.19972.



SOULAIMAN MOUALLA received the B.S. degree in telecommunication engineering and the M.S. degree in telecommunication systems from the Higher Institute for Applied Sciences and Technology (HIAST), Syria, in 2016 and 2021, respectively. He is currently pursuing the Ph.D. degree in signal processing. Since 2016, he has been an Assistant Teacher with HIAST, for a number of signal processing and artificial intelligence courses. His research interests include signal processing, speech processing, artificial intelligence, analog communications, digital communications, and intrusion detection systems.



OUMAYMA AL DAKKAK received the combined Engineering and D.E.A. (Advanced Studies Diploma) degrees in electronic systems from École Nationale Supérieure d'Électronique et de Radioelectricité Grenoble (ENSERG), in 1985, and the Ph.D. degree in electronic system from Institute of Technology (Institut polytechnique de Grenoble INPG-France), done in the Institute of speech communication ICP (now GIPSA), in 1988. She has been a Professor and the Research Director with the Telecommunication Department, Higher Institute for Applied Science and Technology (HIAST), Damascus, Syria, since 2002. Her research interests include signal processing, speech processing: speech synthesis and recognition, machine learning, and digital communication.



OMAR HAMDOUN received the Engineering degree in automatic systems and electronics from the Higher Institute for Applied Sciences and Technology (HIAST), Syria, in 2001, the master's degree in intelligent systems and communication from the University of Cergy-Pontoise, France, in 2007, and the Ph.D. degree in computer vision from École des Mines de Paris (ParisTech), France, in 2010. He is currently an Associate Researcher and a Lecturer with HIAST. His research interests include computer vision, deep learning, multimodal sensor fusion, person re-identification, semantic SLAM, vision–language models, and assistive artificial intelligence.



ASSEF JAFAR received the B.Eng. degree in electronics engineering from Damascus University, in 1991, and the M.S. and Ph.D. degrees in pattern recognition from Bauman Moscow State Technical University, in 1997 and 2004, respectively.

He is currently an Assistant Professor with the Faculty of Artificial Intelligence, Syrian Private University. His research interests include deep learning, machine learning, speech processing, and image processing.