HCFD: A Benchmark for Audio Deepfake Detection in Healthcare

Mohd Mujtaba Akhtar[1]* **Girish**[2]* **Muskaan Singh**[3]†
[1]Veer Bahadur Singh Purvanchal University, India [2]UPES, India [3]Ulster University, UK
{mmakhtar.research, girish.research.pr}@gmail.com, m.singh@ulster.ac.uk

Abstract

In this study, we present Healthcare Codec-Fake Detection (HCFD), a new task for detecting codec-fakes under pathological speech conditions. We intentionally focus on codec-based synthetic speech in this work, since neural codec decoding forms a core building block in modern speech generation pipelines. First, we release Healthcare CodecFake, the first pathology-aware dataset containing paired real and NAC-synthesized speech across multiple clinical conditions and codec families. Our evaluations show that SOTA codec-fake detectors trained primarily on healthy speech perform poorly on Healthcare CodecFake, highlighting the need for HCFD-specific models. Second, we demonstrate that PaSST outperforms existing speech-based models for HCFD, benefiting from its patch-based spectro-temporal representation. Finally, we propose **PHOENIX-Mamba**, a geometry-aware framework that models codec-fakes as multiple self-discovered modes in hyperbolic space and achieves the strongest performance on HCFD across clinical conditions and codecs. Experiments on HCFK show that **PHOENIX-Mamba** (PaSST) achieves the best overall performance, reaching 97.04 Acc on E-Dep, 96.73 on E-Alz, and 96.57 on E-Dys, while maintaining strong results on Chinese with 94.41 (Dep), 94.40 (Alz), and 93.20 (Dys). This geometry-aware formulation enables self-discovered clustering of heterogeneous codec-fake modes in hyperbolic space, facilitating robust discrimination under pathological speech variability. **PHOENIX-Mamba** achieves topmost performance on the HCFD task across clinical conditions and codecs.

1 Introduction

Imagine a clinical voice sample collected to track disease progression being replaced by a codec-generated replica that sounds natural to clinicians and passes automated checks. This scenario is increasingly plausible as neural audio codecs and modern speech generation pipelines enable high-fidelity synthetic speech at scale (Zeghidour et al., 2022; Défossez et al., 2023; Borsos et al., 2023b; Wang et al., 2023). While audio deepfake detection has progressed rapidly, its robustness under pathology-driven speech variability—a defining property of healthcare audio—remains underexplored. Pathological speech systematically alters prosody and articulation, which can obscure codec-induced cues and cause detectors trained on healthy speech to fail in clinical settings. Recent benchmarks further reveal robustness gaps under realistic conditions, including telephony-based deepfakes (Yamagishi et al., 2021) and in-the-wild conversational speech (Wang et al., 2024), which reflect realistic recording conditions and spontaneous speech patterns. Moving to clinical speech, a key barrier is the lack of pathology-aware healthcare deepfake data. While recent work has advanced codec-fake detection, the datasets used for training and evaluation are largely built on healthy, clean speech, with little pathology-aware healthcare deepfake data. As a result, detectors are rarely exposed to disease-related acoustic variation, even though pathological speech differs markedly in prosody, articulation, and phonation. These shifts can change how codec traces appear, limiting transfer to clinical settings and leaving healthcare speech under-served in the current codec-fake detection ecosystem. Healthcare speech is increasingly captured beyond controlled clinic visits—via telehealth consultations, contact-center triage, and remote screening—then transmitted and stored through real-world audio pipelines. In such settings, voice serves a dual role: it is both a clinical signal for monitoring neurocognitive and motor disorders and, in many deployments, an identity signal for patient or staff verification. This combination makes healthcare audio a practical target for modern voice-based attacks.

*Equal contribution as a first author.
†Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 34829–34843

Practitioner and industry analyses warn that rapidly advancing AI voice synthesis and cloning can enable social-engineering and account-takeover attempts against healthcare services, underscoring the need for pathology-aware safeguards that remain reliable under clinically realistic variability [1]. Motivated by this real-world risk surface, we study whether current codec-fake detectors remain reliable when speech is shaped by clinically realistic variability, and we introduce a pathology-aware benchmark and framework to address this gap.

As a remedy for this gap, in this work, we present Healthcare CodecFake (HCFK) and define Healthcare CodecFake Detection (HCFD) as a task for codec-fake detection under pathological speech conditions. Healthcare CodecFake provides paired bona fide and Neural Audio Codecs (NACs)-synthesized speech across multiple clinical conditions and codec families. We benchmark state-of-the-art codec-fake (CF) detectors, including PaSST Koutini et. al (2022), on HCFK and observe a pronounced performance drop, revealing that current CF detectors are not designed for pathology-driven clinical audio. Healthcare speech samples inherently reflect condition-dependent changes in prosody, articulation, and phonation (e.g., altered voice quality, reduced intelligibility, and atypical temporal patterns), which can distort or conceal the subtle traces introduced by NACs, thereby highlighting the limitations of existing CF detection approaches and underscoring the necessity for pathology-aware, healthcare-specific CF detection frameworks. Next, *we hypothesize that the use of pretrained audio representations mitigates the impact of pathology-driven variability by separating condition-specific speech characteristics from codec-related cues, resulting in improved generalization across clinical conditions.* To validate this hypothesis, we conduct a comprehensive comparative study across a diverse set of pretrained audio models, including both speech-focused and audio–language (multimodal) PTMs, under the HCFD setting. Through extensive experiments on HCFK spanning multiple clinical conditions and codec families, we show that PTM-based representations offer more reliable performance than conventional CF detectors, supporting our hypothesis. Despite the gains offered by PTMs, residual performance gaps under clinical variability indicate that representation choice alone is insufficient, prompting us to develop a dedicated pathology-aware detection framework. To this end, we propose **PHOENIX-Mamba:** **P**rototypical **H**yperbolic **O**rganization for **E**vidence **N**ormalization and **I**nference using e**X**ponential-map, a framework tailored to the challenges of codec-fake detection in pathological speech. **PHOENIX-Mamba** integrates long-context temporal modeling with geometry-aware, prototype-based clustering to capture the heterogeneous structure of codec artifacts in clinical speech. By organizing learned evidence representations into multiple fake modes within a hyperbolic space, the framework separates codec-induced cues from disease-related acoustic effects in a principled manner. This design enables consistent inference across clinical conditions and codec families, addressing failure modes that arise in pathology-agnostic detection pipelines. Extensive experiments on HCFK demonstrate that **PHOENIX-Mamba** consistently delivers stronger performance than state-of-the-art codec-fake detectors and strong PTM baselines on the HCFD task. **Our Key contributions are as follows:**

(i) We present Healthcare CodecFake (HCFK) and formally define the novel task of Healthcare Codec-Fake Detection (HCFD), addressing a critical gap in pathology-aware healthcare deepfake detection.; (ii) We benchmark state-of-the-art codec-fake (CF) detectors on HCFK and demonstrate substantial performance degradation, highlighting the need for pathology-aware, healthcare-specific detection frameworks.; (iii) We hypothesize and validate that pretrained audio representations (speech-focused and audio–language PTMs), due to their large-scale pretraining, are better suited for HCFD; a comprehensive comparison across diverse PTMs supports this claim.; (iv) We propose **PHOENIX-Mamba**—a pathology-aware detection framework that integrates long-context temporal modeling with hyperbolic, prototype-based clustering to capture heterogeneous codec artifacts, achieving consistently stronger performance than state-of-the-art CF detectors and strong PTM baselines on HCFD.

2 Related Work

Wu et al. (2024) and Lu et al. (2024) showed that vocoder-trained deepfake detectors generalize poorly to codec-synthesized speech. Wu

[1] Relevant practitioner/industry discussions: Pindrop, HIT Consultant, VoiceBiometrics.ai, OpenAccessGovernment.

Dataset access, code, and evaluation resources are provided at https://helixometry.github.io/HCFD/.

et al. (2024) introduced a VCTK-based codec-synthesized benchmark evaluated with AASIST, while Lu et al. (2024) extended benchmark construction to VCTK and AISHELL3 and compared AASIST and LCNN using mel-spectrogram and Wav2vec2 features. Subsequent studies extended CodecFake benchmarks by increasing codec diversity (Chen et al., 2025a) and explored unified semantic–acoustic representations for detection, e.g., via SASTNet (Chen et al., 2025b). Xie et al. (2025) considered unified detection across codec- and vocoder-synthesized speech, reporting improved cross-mechanism generalization with sharpness-aware optimization. Empirical studies have documented that audio deepfake detectors exhibit non-trivial performance degradation when evaluated under acoustic mismatches relative to their training and validation conditions (Li et al., 2025).

3 Healthcare Codecfake Dataset

In this section, we describe the datasets, neural audio codecs, and synthesis pipeline used to construct HCFK. All recordings are sourced from established corpora accessed under their respective data-use agreements; we collect no new data, include no personally identifiable information, and generate synthetic audio solely for research and evaluation. Our intent is defensive: to support detection and risk mitigation for malicious use of AI-generated healthcare speech; we explicitly discourage misuse of the methods or generated samples.

Access to the underlying corpora remains subject to their original licenses, consent provisions, and access agreements. To support reproducibility, we will provide the exact split files together with the full codec-generation pipeline, including preprocessing details, codec configurations, and reconstruction steps, so that HCFK can be recreated deterministically given approved access to the source datasets.

3.1 Healthcare Speech Datasets

We select benchmark healthcare speech datasets spanning multiple clinical conditions to construct a realistic evaluation setting for healthcare-oriented codec-fake detection. The benchmark includes corpora in English and Chinese, enabling cross-lingual evaluation in healthcare speech.

Depression: we use DAIC-WOZ (Gratch et al., 2014) as the English corpus, consisting of 189 semi-structured interviews with the virtual interviewer Ellie. For Chinese, we use the EATD-Corpus (Shen et al., 2022), which contains interview-style responses from 162 volunteers with SDS-based depression annotations.

Alzheimer: we use ADReSS/ADReSSo (Luz et al., 2020) as the English corpus, a widely used benchmark for Alzheimer's disease detection that provides standardized and balanced audio samples derived from the DementiaBank Pitt "Cookie Theft" picture-description task. For Chinese, we use NCMMSC (Shakeri et al., 2025), a Mandarin dementia benchmark with clinically annotated recordings for cognitive impairment assessment.

Dysarthria: we use TORGO (Rudzicz et al., 2012) as the English corpus, which contains speech from individuals with dysarthria. For Chinese, we use the Chinese Dysarthria Speech Database (CDSD) (Wan et al., 2024), a large-scale Mandarin corpus comprising approximately 133 hours of recordings. Across all conditions, the original recordings are treated as bona fide speech, and paired codec-generated samples are synthesized from the same utterances using our codec pipeline.

3.2 Neural Audio Codecs

Following Lu et al. (2024) and Wu et al. (2024), we use the same family of neural audio codecs, focusing on state-of-the-art, publicly released models that are easy to reproduce and widely used.

Speechtokenizer (Zhang et al., 2024) [1]: It is a unified speech tokenizer built on an RVQ-GAN style neural codec. The model uses an EnCodec-based convolutional encoder–decoder backbone with Residual Vector Quantization (RVQ).

Descript Audio Codec (Kumar et al., 2024) [2]: It is a VQ-GAN based neural audio codec targeting high-fidelity reconstruction. The approach discretizes encoder features with RVQ and trains the generator using adversarial learning alongside multi-scale frequency-domain criteria to suppress codec artifacts.

Encodec (Défossez et al., 2022) [3]: It is a streaming, high-quality neural codec that couples a convolutional encoder–decoder with RVQ discretization. Training combines time-domain and frequency-domain reconstruction criteria with a spectrogram adversary for improved perceptual quality.

[1] https://github.com/ZhangXInFD/SpeechTokenizer.git

[2] https://huggingface.co/descript/dac_16khz

[3] https://huggingface.co/facebook/encodec_24khz

Soundstream (Zeghidour et al., 2021) [4]: It is an end-to-end neural codec tailored for low-bitrate speech compression. The model combines an encoder–decoder backbone with Residual Vector Quantization (RVQ) and multi-scale STFT discriminators, enabling high perceptual quality under aggressive compression (3–18 kbps).
Funcodec (Du et al., 2024) [5]: It is an open-source neural speech codec toolkit built to make modern codec models easy to train, reproduce, and integrate into downstream pipelines. It extends the FunASR ecosystem and provides unified training recipes and inference scripts.
Audiodec (Wu et al., 2023) [6]: It is a high-quality neural codec formulated as an end-to-end autoencoder. Training proceeds in two phases: it first learns the encoder–decoder with metric-based objectives, and then applies an adversarial refinement stage that updates only the decoder.
SNAC (Siuzdak et al., 2024) [7]: It is a multi-scale NAC that extends standard RVQ by allowing different quantizers. Concretely, it forms a hierarchical token representation by quantizing coarse-to-fine structure at multiple frame rates.
Detailed codec configurations and checkpoints used in our experiments are summarized in Appendix A. For reproducibility, we provide a consolidated repository of the NAC resources used in data generation: [8].

3.3 Health Care Codecfake Data Generation Pipeline

We construct HCFK using a controlled resynthesis protocol, building on established CodecFake-style dataset generation practices (Wu et al., 2024). Specifically, each bona fide pathological utterance is passed through a diverse set of NACs to produce paired codec-synthesized counterparts, reflecting the codec front-ends commonly adopted in modern audio language modeling pipelines (Borsos et al., 2023a; Lu et al., 2024). We begin with the healthcare speech corpora described in Section 3.1, spanning multiple pathological conditions and two languages (English and Chinese). Each segmented utterance is treated as a bona fide reference sample. To generate codec-synthesized spoofed counterparts, we employ a codec synthesis–resynthesis loop: for each waveform, we first pass the signal through the pre-trained encoder of a NAC to obtain a discrete latent representation, and then decode it back to the waveform domain using the corresponding decoder. We then decode the discrete representation using the corresponding codec decoder to reconstruct the signal, producing a codec-generated counterpart of the original clinical utterance. This reconstruction preserves most of the semantic content and speaker/subject traits, while introducing subtle artifacts arising from quantization and bandwidth constraints in the codec. As a result, HCFK contains high-fidelity yet spoofed pathological speech that reflects realistic codec-mediated generation pathways. We repeat this resynthesis procedure across a suite of NAC models, producing multiple codec-specific variants of HCFK. Concretely, for each bona fide utterance, we generate one paired codec-synthesized counterpart per codec, yielding a consistent one-to-one mapping between the source sample and each codec condition. For DAIC-WOZ (Gratch et al., 2014), which follows an interview-based evaluation protocol, we generate codec-synthesized samples separately within each predefined subset to keep train/dev/test strictly isolated. For ADReSS/ADReSSo (Luz et al., 2020) and NCMMSC (Shakeri et al., 2025), where standard partitions are released with the benchmarks, we retain these official splits and synthesize the corresponding codec-generated audio within each split. We keep the provided splits unchanged and synthesize the codec-generated audio within each split. Across all settings, train/dev/test splits are strictly speaker-disjoint. Bona fide test utterances come from speakers unseen during training, and codec-generated samples are synthesized only from utterances within the same split. This preserves speaker disjointness for both bona fide and fake classes and ensures that evaluation reflects generalization to unseen speakers rather than speaker-specific memorization.

4 Methodology

In this section, we provide the methodological details of our study. We begin by outlining the PTMs considered. Next, we define the downstream baselines trained on individual PTM representations. Finally, we present **PHOENIX-Mamba**, a geometry-aware detector that retains multiple localized evidences and models diverse fake modes using hy-

[4] https://github.com/haydenshively/SoundStream
[5] https://github.com/modelscope/FunCodec
[6] https://github.com/facebookresearch/AudioDec
[7] https://huggingface.co/hubertsiuzdak/snac_44khz
[8] https://github.com/CodeVault-girish/Neural-Codecs.git

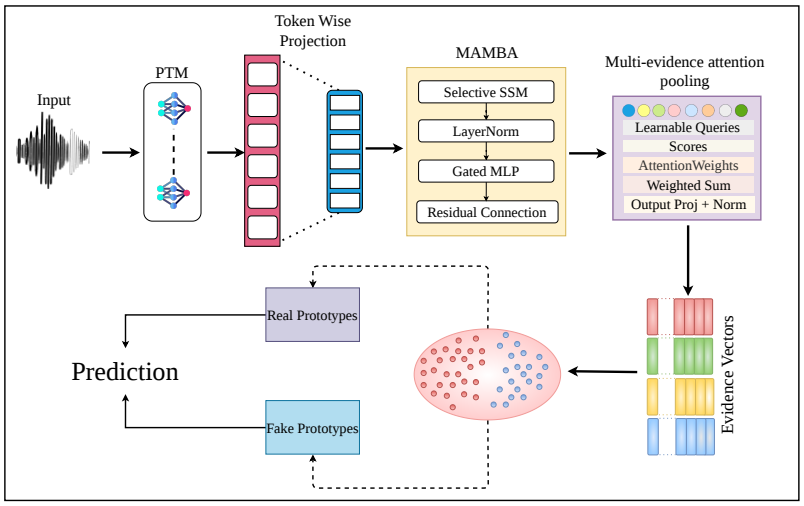

Figure 1: Proposed Framework: **PHOENIX-Mamba**

perbolic prototypes.

Method	Dep		Alz		Dys	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
	English					
AASIST (Tr. on CF)	48.62	44.03	34.19	32.51	36.71	34.39
AASIST (Tr. in-domain)	60.84	57.92	52.14	49.93	56.07	54.49
AASIST (wav2vec2.0)	63.55	51.29	57.76	54.98	59.35	57.16
RawNet2	60.46	50.27	58.96	53.87	57.24	54.68
LCNN	62.79	53.75	60.59	55.27	59.79	55.62
SAMO	64.37	54.78	61.89	58.41	60.27	56.73
Method	**Chinese**					
AASIST (Tr. on CF)	45.81	42.38	30.79	29.23	33.96	32.54
AASIST (Tr. in-domain)	58.06	54.79	48.51	46.12	54.84	51.03
AASIST (wav2vec2.0)	60.82	49.31	53.90	52.29	56.11	54.62
RawNet2	58.33	47.83	54.51	49.77	56.71	52.64
LCNN	60.32	46.98	56.29	50.08	58.64	54.22
SAMO	62.18	49.82	59.52	53.79	61.45	54.78

Table 1: Performance of prior detector baselines on three healthcare speech tasks: Depression (Dep), Alzheimer's (Alz), and Dysarthria (Dys), evaluated separately for English and Chinese. We report Accuracy and macro-F1. "Tr. on CF" denotes training on the standard CodecFake benchmark and evaluation on HCFK. "Tr. in-domain" denotes the within-dataset setting, where for each language–condition subset the model is trained on the training split, selected on the validation split, and evaluated on the held-out test split. In addition to AASIST, we include representative waveform-based (RawNet2), spectrogram-based (LCNN), and robust one-class/generalization-oriented (SAMO) anti-spoofing baselines.

Pre-Trained Models: We evaluate a diverse set of strong pre-trained encoders that have shown competitive performance across speech/audio benchmarks. Specifically, we consider self-supervised speech encoders WavLM[9] (Chen et al., 2022) and Wav2vec 2.0[10] (Baevski et al., 2020), the Whisper model[11] (Radford et al., 2023) (using its encoder representations), a supervised speaker-embedding extractor X-vector[12] (Snyder et al., 2018), and the spectrogram-based audio transformer PaSST[13] (Koutini et.al, 2022). All audio inputs are resampled to 16 kHz before being passed to the respective encoders. For WavLM, Wav2vec 2.0, and Whisper, we keep the PTMs frozen, extract the last hidden-state sequence (from the encoder in the case of Whisper), and obtain an utterance-level representation by average pooling over time; this pooled embedding is then fed to the downstream classifier. For X-vector, we use the VoxCeleb-trained model as a frozen feature extractor and directly feed the resulting fixed-dimensional utterance embedding to the classifier. For PaSST, we compute spectrogram inputs as required by the model and use its pooled representation as the utterance-level embedding for classification. The resulting feature vector sizes are: 768 for WavLM, Wav2vec 2.0, and PaSST; and 512 for Whisper and X-vector.

4.1 Individual Representation Modeling

We establish strong baselines by training standard downstream classifiers on individual pre-trained speech representations, using a lightweight 1D-CNN (two Conv1D–BN–activation–max-pooling

[9] https://huggingface.co/microsoft/wavlm-base
[10] https://huggingface.co/facebook/wav2vec2-base
[11] https://huggingface.co/openai/whisper-base
[12] https://huggingface.co/speechbrain/spkrec-xvect-voxceleb
[13] https://github.com/kkoutini/PaSST

blocks followed by flattening and a small dense predictor) and an FCN that removes the convolutional front-end while keeping the same dense predictor. Detailed hyperparameter settings and model configurations are provided in the Appendix B.

4.2 Proposed Framework: PHOENIX-Mamba

We propose **PHOENIX-Mamba**, a geometry-aware approach for detecting codec-generated speech in healthcare data. The complete pipeline is depicted in Figure 1. Given an input utterance x, we first extract a sequence of latent features $X = [x_1, \ldots, x_T] \in \mathbb{R}^{T \times D}$ using an upstream encoder. We then apply a token-wise alignment (adapter) map $\phi : \mathbb{R}^D \to \mathbb{R}^d$ to obtain $U = \phi(X) \in \mathbb{R}^{T \times d}$. The aligned sequence is passed through a Mamba-style selective state-space backbone f_θ to build context-enriched representations $Z = f_\theta(U) = [z_1, \ldots, z_T] \in \mathbb{R}^{T \times d}$. Instead of collapsing the full sequence into a single pooled vector, **PHOENIX-Mamba** compresses Z into a small set of M evidence vectors $E = [e_1, \ldots, e_M] \in \mathbb{R}^{M \times d}$. This evidence construction is implemented with a learnable pooling operator: $e_m = \sum_{t=1}^{T} a_{m,t} z_t$, where $a_{m,t} \geq 0$ and $\sum_{t=1}^{T} a_{m,t} = 1$. The weights $a_{m,t}$ are produced by a differentiable scoring mechanism. This design allows the model to retain multiple localized cues that may be unevenly distributed across the utterance. To model heterogeneity in codec artifacts, we embed each evidence vector into a metric manifold. We use the Poincaré ball $\mathcal{M} = \mathbb{B}_c^h = \{v \in \mathbb{R}^h : c\|v\|^2 < 1\}$ with curvature $-c$. Each evidence vector is mapped to the manifold using a differentiable projection $\psi : \mathbb{R}^d \to \mathcal{M}$, giving $h_m = \psi(e_m) \in \mathcal{M}$ and $H = [h_1, \ldots, h_M] \in \mathcal{M}^M$. In practice, we project to the tangent space and apply the exponential map at the origin: $h_m = \mathrm{Exp}_0^c(We_m)$, where $\mathrm{Exp}_0^c(y) = \tanh(\sqrt{c}\|y\|)\,\frac{y}{\sqrt{c}\|y\|}$. Distances are computed using the hyperbolic geodesic distance $d_c(\cdot, \cdot)$. We perform classification using prototype-based reasoning in $\mathcal{M}$. We parameterize a single negative prototype $p_- \in \mathcal{M}$ for the real class and K positive prototypes $\{p_{+,1}, \ldots, p_{+,K}\} \subset \mathcal{M}$ to capture diverse fake modes. For each evidence point h_m, we compute soft responsibilities over the positive prototypes using a temperature-controlled distance softmax:

$$q_{m,k} = \frac{\exp(-d_{\mathcal{M}}(h_m, p_{+,k})/\tau)}{\sum_{j=1}^{K} \exp(-d_{\mathcal{M}}(h_m, p_{+,j})/\tau)} \quad (1)$$

This allows **PHOENIX-Mamba** to self-discover multiple clusters within the fake class using only binary supervision. We compute evidence-level scores using manifold distances. The negative score is defined as $s_-(h_m) = -d_{\mathcal{M}}(h_m, p_-)$. The positive score is computed as a smooth softmin over the K positive modes: $s_+(h_m) = \log \sum_{k=1}^{K} \exp(-d_{\mathcal{M}}(h_m, p_{+,k})/\tau)$. We aggregate these scores across the M evidence vectors to obtain instance-level logits $S_- = \frac{1}{M} \sum_{m=1}^{M} s_-(h_m)$ and $S_+ = \frac{1}{M} \sum_{m=1}^{M} s_+(h_m)$. The final probability is computed via a softmax: $P(y = + \mid x) = \mathrm{softmax}([S_-, S_+])_+$.

We train **PHOENIX-Mamba** end-to-end using only real/fake labels. The base objective is the cross-entropy classification loss $\mathcal{L}_{\mathrm{cls}}$ on logits $[S_-, S_+]$. To encourage compact and meaningful positive clusters, we introduce a geometry-aware clustering loss:

$$\mathcal{L}_{\mathrm{cluster}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} q_{m,k}\, d_{\mathcal{M}}(h_m, p_{+,k}) + \gamma \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} q_{m,k} \log q_{m,k} \quad (2)$$

The first term pulls evidence embeddings toward their assigned positive prototypes, while the entropy term controls assignment sharpness. To avoid prototype collapse and maintain separation between modes, we add a repulsion loss:

$$\mathcal{L}_{\mathrm{sep}} = \sum_{1 \leq i < j \leq K} \exp\big(-d_{\mathcal{M}}(p_{+,i}, p_{+,j})\big) + \sum_{k=1}^{K} \exp\big(-d_{\mathcal{M}}(p_{+,k}, p_-)\big) \quad (3)$$

The total loss is given by: $\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda\, \mathcal{L}_{\mathrm{cluster}} + \beta\, \mathcal{L}_{\mathrm{sep}}$, where $\lambda, \beta, \gamma \geq 0$ control the contribution of geometry-aware regularization. Under this objective, the backbone learns evidence representations that support robust discrimination, while the positive prototypes self-organize into multiple modes that capture heterogeneous codec artifacts. The trainable parameters for **PHOENIX-Mamba** range from 2 M to 5 M, depending on the input representation dimension.

4.3 Training Details and Hyperparameters

We train the model with AdamW for 20 epochs, using a batch size of 32, weight decay of 0.01,

PTMs	Dep				Alz				Dys			
	FCN		CNN		FCN		CNN		FCN		CNN	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
ENGLISH												
WLM	69.51	66.83	71.42	69.13	57.21	55.62	59.80	58.17	63.83	60.47	65.77	63.21
WV2	73.29	71.95	76.09	73.61	60.78	59.42	63.46	61.53	66.78	64.89	69.35	66.92
WHI	71.58	68.46	73.38	71.56	63.84	59.72	65.98	62.81	67.15	65.91	69.48	68.26
XVE	67.05	66.34	69.84	68.29	56.26	54.98	58.32	57.64	63.52	62.39	66.80	64.69
PST	76.71	73.89	78.98	76.62	66.09	62.57	67.94	65.27	69.27	67.86	71.03	70.54
CHINESE												
WLM	66.04	63.79	69.28	67.59	55.03	53.81	56.34	53.67	60.41	58.03	63.19	60.47
WV2	71.68	69.84	72.94	69.20	58.33	55.89	61.27	60.02	62.97	59.38	65.88	64.23
WHI	67.93	64.02	70.51	66.82	59.91	56.24	63.53	61.39	65.34	64.01	66.41	64.60
XVE	64.52	62.49	67.08	64.96	54.68	50.91	54.67	51.83	59.96	56.73	64.27	61.39
PST	74.16	70.63	75.69	72.19	63.49	61.05	65.71	64.24	67.52	66.10	67.36	65.02

Table 2: Performance on HCFK. Abbreviations: Dep = Depression; Alz = Alzheimer's; Dys = Dysarthria; WLM = WavLM; WV2 = wav2vec 2.0; WHI = Whisper; XVE = x-vector; PST = PaSST. The same abbreviations are used consistently in Table 3.

and gradient clipping at 1.0. For evaluation, Accuracy and macro-F1 are computed with a validation-selected decision threshold, while EER is computed from the same score distributions. All main experiments use a single shared default configuration. The final hyperparameter values are summarized in Appendix B.

5 Results & Discussion

Generalization from prior codec-deepfake detectors: We first examine whether a detector trained under standard codec-deepfake benchmarks (Jung et al., 2022) can generalize to pathological healthcare speech, a setting that couples codec artifacts with clinically-driven acoustic deviations. Table 1 reports results for prior detector baselines on the English and Chinese subsets across the three clinical tasks. Training AASIST on the original Codec-Fake distribution yields near-chance performance across all three tasks; on the English subset the scores are 48.62/34.19/36.71, while on the Chinese subset they are 45.81/30.79/33.96 for Dep/Alz/Dys, respectively. This behavior indicates a pronounced distribution shift: codec-induced cues in pathological speech are confounded by condition-specific acoustics and recording variability, limiting the reliability of models transferred from healthy-speech codec benchmarks. When trained separately on each healthcare dataset, performance improves but remains limited, suggesting that more informative representations and mechanisms for capturing heterogeneous cues are still needed. To broaden this comparison, we additionally evaluate RawNet2 (Tak et al., 2021), LCNN (Wu et al., 2020), and SAMO (Ding et al., 2023), representing waveform-based, spectrogram-based, and robust generalization-oriented detector families. Although these baselines provide modest improvements over the weakest transfer setting, they still remain clearly below the stronger PTM-based approaches reported next. The AASIST variant equipped with a wav2vec 2.0 backbone is also consistently stronger than the standard AASIST settings, but still leaves a substantial gap to the best-performing methods introduced later. Taken together, these results show that the challenge is not specific to a single detector architecture, but reflects a broader difficulty of transferring existing codec-fake detectors to pathological speech.

PTM comparison with FCN and CNN as downstream networks: We next benchmark a diverse set of pre-trained encoders to assess which representations are most effective for healthcare codec-deepfake detection. Using the same train/test protocol, we train two lightweight downstream heads (FCN and 1D-CNN) on top of each frozen PTM embedding, and report results in Table 2 for both English and Chinese subsets. Two consistent observations emerge. First, the CNN head—despite its simplicity—yields stronger performance than the FCN in most settings, indicating that local temporal structure remains informative and can be exploited with a shallow convolutional frontend. Second, among all upstream encoders, PaSST provides the strongest single-representation baseline across tasks and languages, achieving the best overall per-

formance in both English and Chinese. We further observe that Alzheimer's is consistently more challenging than the other clinical tasks for all PTMs, suggesting that condition-induced variability and broader acoustic differences make codec artifacts harder to isolate in this setting. Finally, performance is generally lower on the Chinese subset than on English for the same configuration, highlighting an additional cross-lingual shift that compounds the healthcare-domain mismatch. Notably, the comparatively strong results of Whisper across tasks align with prior findings that multilingual speech foundation models often provide more robust features for deepfake-related tasks than monolingual SSL encoders, likely due to broader linguistic diversity during pre-training (Phukan et al., 2025).

PTMs	Dep		Alz		Dys	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
ENGLISH						
WLM	94.27	91.52	92.84	91.56	95.07	92.67
WV2	95.68	93.14	95.39	92.78	94.82	93.19
WHI	94.12	92.38	93.11	91.34	95.43	92.82
XVE	94.46	91.59	91.65	90.07	93.24	91.54
PST	97.04	96.81	96.73	95.20	96.57	94.28
CHINESE						
WLM	91.56	89.74	89.25	86.62	91.48	90.29
WV2	93.04	90.42	93.08	91.54	92.03	89.47
WHI	90.29	88.19	90.71	89.03	93.87	91.86
XVE	92.17	89.87	87.93	84.72	89.61	86.95
PST	94.41	92.10	94.40	92.18	93.20	91.42

Table 3: **PHOENIX-Mamba** results. Abbreviations follow Table 2.

Results of `PHOENIX-Mamba`: To understand where the improvements come from, we compare **`PHOENIX-Mamba`** against the strongest single-representation baselines reported in Table 2 under the same train/test protocol. The results in Table 3 show that replacing standard single-vector classification with our evidence-driven, prototype-based reasoning consistently improves performance across all upstream encoders, tasks, and languages, indicating that the gains are primarily driven by the proposed modeling strategy rather than any single PTM. With PaSST, **`PHOENIX-Mamba`** achieves the best overall performance, reaching 97.04/96.81 on Dep, 96.73/95.20 on Alz, and 96.57/94.28 on Dys for the English subset (Acc/F1), and maintaining strong results on the Chinese subset with 94.41/92.10 (Dep), 94.40/92.18 (Alz), and 93.20/91.42 (Dys). Notably, the largest gains appear on Alzheimer's, where variability is highest, reinforcing the importance of explicitly handling heterogeneity in healthcare speech. The consistent gains suggest that **`PHOENIX-Mamba`** is effective because it aligns the decision process with the structure of the problem—localized cues and heterogeneous artifact modes—rather than relying on a single pooled representation.

To complement Accuracy and macro-F1, we also report EER, a standard threshold-independent metric in synthetic speech detection (Sheth et al., 2025). Using the same PaSST upstream representation, the CNN baseline attains EERs of 14.01/16.52/15.93 on English Dep/Alz/Dys and 18.42/20.04/21.78 on Chinese, whereas **`PHOENIX-Mamba`** achieves markedly lower values of 5.17/6.29/6.23 and 6.54/5.42/6.79, respectively. This further confirms the advantage of the proposed method under the same evaluation setting.

Generalization Beyond the In-Domain Setting: We also study the generalization behavior of **`PHOENIX-Mamba`** under a held-out codec-family protocol. For this evaluation, the seven codec families in HCFK are partitioned into seen and unseen subsets by randomly assigning five families to training and holding out the remaining two exclusively for testing. As shown in Table 4, the relative behavior across upstream encoders remains largely stable in this more challenging setting, with PaSST continuing to achieve the strongest overall performance. This suggests that the proposed framework captures evidence that remains informative even for codec families not encountered during training.

To further examine transfer beyond codec variation, we also evaluated **`PHOENIX-Mamba`** under a cross-pathology setting in which training and testing were performed on different clinical conditions. When trained on Depression and evaluated on Alzheimer's, **`PHOENIX-Mamba`** achieved 95.88 Acc / 93.41 F1 / 6.57 EER in English and 91.79 Acc / 89.05 F1 / 6.82 EER in Chinese. Under a broader transfer setting, where training was performed on Depression and Dysarthria and evaluation was conducted on Alzheimer's, performance further improved to 98.53 Acc / 97.21 F1 / 3.66 EER in English and 97.84 Acc / 95.10 F1 / 3.79 EER in Chinese. These results indicate that **`PHOENIX-Mamba`** retains transferable codec-fake evidence even when the target condition differs from those observed during training, and that broader pathology coverage during training further strengthens generalization.

PTMs	English						Chinese					
	Dep		Alz		Dys		Dep		Alz		Dys	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
WavLM	91.43	89.14	87.65	85.98	91.32	89.42	89.62	88.74	85.45	83.73	88.21	86.92
Wav2Vec2	91.67	88.34	92.82	90.38	91.49	87.04	88.45	87.64	90.92	88.69	89.82	87.54
Whisper	89.51	87.93	91.42	88.61	92.74	89.74	90.04	87.61	89.54	87.94	90.05	88.59
X-vector	90.37	88.52	85.27	83.74	88.49	87.29	89.75	86.05	87.72	85.82	90.92	89.34
PaSST	**95.59**	**93.29**	**94.66**	**92.04**	**95.17**	**93.28**	**94.17**	**93.74**	**95.09**	**92.75**	**93.99**	**93.02**

Table 4: Seen–unseen codec evaluation results.

Methods	Dep		Alz		Dys	
	Acc	F1	Acc	F1	Acc	F1
ENGLISH						
CNN Head	82.26	80.73	75.52	72.13	79.37	77.91
BiGRU Head	87.69	84.91	82.86	80.49	86.61	83.73
Single evidence: set $M=1$	73.51	72.02	55.03	52.67	67.94	65.02
PHOENIX-Euc	83.62	81.24	79.48	77.16	84.72	83.67
PHOENIX-Mamba (Full)	97.04	94.81	96.73	94.20	96.05	93.28
CHINESE						
CNN Head	79.81	77.26	73.12	70.46	76.89	74.28
BiGRU Head	83.62	80.91	78.95	77.87	84.03	81.65
Single evidence: set $M=1$	71.04	69.38	52.46	50.24	65.26	63.77
PHOENIX-Euc	80.59	77.14	77.08	74.63	80.91	77.98
PHOENIX-Mamba (Full)	94.41	92.10	94.40	92.18	93.20	91.42

Table 5: Ablation study of **PHOENIX-Mamba**

5.1 Ablation Study

To quantify the contribution of key design choices in **PHOENIX-Mamba**, we perform an ablation study along three axes, keeping the training protocol fixed and varying a single component at a time (Table 5).

Role of Temporal Modeling: We first examine the impact of the sequence modeling head used for downstream reasoning. Starting from a lightweight CNN head, we replace it with a stronger recurrent alternative (BiGRU head) while keeping the remaining pipeline unchanged. This comparison isolates the benefit of richer temporal dependencies over shallow local modeling.

Role of Multi-Evidence Pooling A central design choice in **PHOENIX-Mamba** is to retain multiple localized cues via M evidence vectors rather than collapsing an utterance into a single summary. To isolate this effect, we reduce the evidence set to a single vector by setting $M=1$ (single-evidence variant), while keeping the backbone and classifier identical. This ablation tests whether preserving multiple evidences is necessary for healthcare speech, where codec artifacts may appear intermittently and non-uniformly.

Role of Geometry-Aware Multi-Mode Reasoning Finally, we evaluate the importance of the geometry-aware prototype reasoning used to model heterogeneous fake modes. We compare the full **PHOENIX-Mamba** to a Euclidean counterpart (PHOENIX-Euc) that removes the hyperbolic embedding/clustering component while retaining the same overall architecture and optimization setup. This variant isolates the contribution of the manifold-based multi-mode structure beyond standard Euclidean classification.

6 Conclusion

In this work, we initiate a focused study of Healthcare CodecFake Detection (HCFD), targeting codec-generated audio deepfakes in pathological speech—a setting that remains underexplored despite its practical relevance to clinical and telehealth communication. To support systematic evaluation, we introduce Healthcare CodecFake (HCFK), a benchmark constructed by resynthesizing pathological speech across multiple clinical conditions and two languages using a diverse set of neural audio codecs. Our findings show that detectors trained under standard codec-deepfake benchmarks exhibit limited transfer to healthcare speech, highlighting a substantial domain shift and motivating dedicated solutions for this problem. Building on this insight, we benchmark a range of pre-trained audio/speech encoders and observe that representation quality and downstream modeling both play an important role. To move beyond single-vector classification, we propose **PHOENIX-Mamba**, a geometry-aware framework that retains multiple localized evidences and models the fake class through multiple prototype modes in hyperbolic space. Across clinical tasks and languages, **PHOENIX-Mamba** consistently improves over strong PTM baselines, and ablations confirm the importance of multi-evidence pooling and geometry-aware multi-mode reasoning. We expect HCFK and **PHOENIX-Mamba** to provide a foundation for more reliable evaluation and continued progress on codec-deepfake detection in healthcare-oriented speech.

Limitations & Future Work

Limitations HCFK provides a first pathology-aware benchmark for healthcare codec-fake detection, but currently covers a limited set of conditions and languages. We focus on codec-based resynthesis with a fixed set of NACs under a controlled protocol; broader real-world channel effects and other attack families (e.g., TTS/VC/diffusion, replay/recapture, adversarial post-processing) are not included. We evaluate detection only and do not address generator/codec attribution or open-set uncertainty for unseen attacks.

Future work We will expand coverage to more conditions, languages, and recording scenarios, including telehealth-style settings, and extend the benchmark to additional attack families beyond NAC resynthesis. We also plan to study open-set detection, uncertainty estimation, and attribution, as well as privacy-preserving evaluation and more interpretable artifact analysis.

Ethical considerations

This work is motivated by the need to safeguard clinical and healthcare communication against codec-generated audio manipulation. We do not collect any new human-subject recordings. The benchmark is constructed by applying NACs to existing speech datasets that are available for research use under their respective licenses and access conditions. As an additional quality-control measure, a certified speech therapist with clinical experience in pathological speech assessment qualitatively reviewed a subset of bona fide and codec-synthesized samples. The proposed models and benchmarks are not intended for medical diagnosis, treatment decisions, or deployment as standalone security mechanisms in clinical workflows.

Acknowledgments

The authors gratefully acknowledge the support of the United States–Ireland–Northern Ireland R&D Partnership Programme (USI-207), and access to the Tier 2 High-Performance Computing resources provided by the Northern Ireland High Performance Computing (NIHPC) facility, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant No. EP/T022175/1.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and 1 others. 2023a. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023b. Audiolm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Xuanjun Chen, Jiawei Du, Haibin Wu, Lin Zhang, I Lin, I Chiu, Wenze Ren, Yuan Tseng, Yu Tsao, Jyh-Shing Roger Jang, and 1 others. 2025a. Codecfake+: A large-scale neural audio codec-based deepfake speech dataset. *arXiv preprint arXiv:2501.08238*.

Xuanjun Chen, I Lin, Lin Zhang, Haibin Wu, Hung-yi Lee, Jyh-Shing Roger Jang, and 1 others. 2025b. Towards generalized source tracing for codec-based deepfake speech. *arXiv preprint arXiv:2506.07294*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Trans. Mach. Learn. Res.*, 2023.

Siwen Ding, You Zhang, and Zhiyao Duan. 2023. Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhihao Du, Shiliang Zhang, Kai Hu, and Siqi Zheng. 2024. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 591–595. IEEE.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood,

Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371.

Koutini et. al. 2022. Efficient Training of Audio Transformers with Patchout. In *Interspeech 2022*, pages 2753–2757.

Koutini et.al. 2022. Efficient Training of Audio Transformers with Patchout. In *Interspeech 2022*, pages 2753–2757.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2024. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36.

Xiang Li, Pin-Yu Chen, and Wenqi Wei. 2025. Measuring the robustness of audio deepfake detectors. *Preprint*, arXiv:2503.17577.

Yi Lu, Yuankun Xie, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Zhiyong Wang, Xin Qi, Xuefei Liu, Yongwei Li, Yukun Liu, Xiaopeng Wang, and Shuchen Shi. 2024. Codecfake: An initial dataset for detecting llm-based deepfake audio. In *Interspeech 2024*, pages 1390–1394.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: The adress challenge. In *Interspeech 2020*, pages 2172–2176.

Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Priyabrata Mallick, Pailla Balakrishna Reddy, Arun Balaji Buduru, and Rajesh Sharma. 2025. Towards source attribution of singing voice deepfake with multimodal foundation models. In *Proc. INTERSPEECH*, pages 1673–1677.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff. 2012. The torgo database of acoustic and articulatory speech from speakers with dysarthria. *Lang. Resour. Eval.*, 46(4):523–541.

Arezo Shakeri, Mina Farmanbar, and Krisztian Balog. 2025. Multiconad: A unified multilingual conversational dataset for early alzheimer's detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 3627–3637, New York, NY, USA. Association for Computing Machinery.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251.

Farhan Sheth, Girish, Mohd Mujtaba Akhtar, and Muskaan Singh. 2025. Curved worlds, clear boundaries: Generalizing speech deepfake detection using hyperbolic and spherical geometry spaces. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1923–1932, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Hubert Siuzdak, Florian Grötschla, and Luca A Lanzendörfer. 2024. Snac: Multi-scale neural audio codec. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373.

Yan Wan, Mengyi Sun, Xinchen Kang, Jingting Li, Pengfei Guo, Ming Gao, and Su-Jing Wang. 2024. CDSD: Chinese Dysarthria Speech Database. In *Interspeech 2024*, pages 4109–4113.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers. *Preprint*, arXiv:2301.02111.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi H. Kinnunen, Nicholas Evans, Kong Aik Lee, and Junichi Yamagishi. 2024. Asvspoof 5: crowdsourced speech

data, deepfakes, and adversarial attacks at scale. In *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, pages 1–8.

Haibin Wu, Yuan Tseng, and Hung yi Lee. 2024. Codecfake: Enhancing anti-spoofing models against deepfake audios from codec-based speech synthesis systems. In *Interspeech 2024*, pages 1770–1774.

Yi-Chiao Wu, Israel D Gebru, Dejan Marković, and Alexander Richard. 2023. Audiodec: An open-source streaming high-fidelity neural audio codec. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. 2020. Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks. In *Interspeech 2020*, pages 1101–1105.

Yuankun Xie, Yi Lu, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Jianhua Tao, Xin Qi, Xiaopeng Wang, Yukun Liu, Haonan Cheng, and 1 others. 2025. The codecfake dataset and countermeasures for the universally detection of deepfake audio. *IEEE Transactions on Audio, Speech and Language Processing*.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, and Héctor Delgado. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2022. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:495–507.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. 2024. Speechtokenizer: Unified speech tokenizer for speech language models. In *The Twelfth International Conference on Learning Representations*.

Appendix

In the appendix, we provide:

- **Section A: Neural Audio Codecs.** Details of the neural audio codec families and the specific publicly released checkpoints used in our experiments.
- **Section B: Hyperparameters and System Configurations.** A summary of key training hyperparameters and geometry-related settings (Table 6).
- **Section C: Visualization Analysis.** Additional qualitative analyses including confusion matrices (Figure 2) and t-SNE plots (Figure 3) for representative configurations.

A Neural Audio Codecs

Following Lu et al. (2024) and Wu et al. (2024), we use the same family of neural audio codecs, focusing on state-of-the-art, publicly released models that are easy to reproduce and widely used.

Speechtokenizer (Zhang et al., 2024)[14]: It is a unified speech tokenizer built on an RVQ-GAN style neural codec. The model uses an EnCodec-based convolutional encoder–decoder backbone with Residual Vector Quantization (RVQ). We use the 16 kHz SpeechTokenizer setting in this work.

Descript Audio Codec (Kumar et al., 2024)[15]: It is a VQ-GAN-based neural audio codec targeting high-fidelity reconstruction. The approach discretizes encoder features with RVQ and trains the generator using adversarial learning alongside multi-scale frequency-domain criteria to suppress codec artifacts. We evaluate its checkpoints at 16 kHz, 24 kHz, and 44 kHz sampling rates.

Encodec (Défossez et al., 2022)[16]: It is a streaming, high-quality neural codec that couples a convolutional encoder–decoder with RVQ discretization. Training combines time-domain and frequency-domain reconstruction criteria with a spectrogram adversary for improved perceptual quality. We evaluate the 24 kHz and 48 kHz models in our study.

Soundstream (Zeghidour et al., 2021)[17]: It is an end-to-end neural codec tailored for low-bitrate speech compression. The model combines an encoder–decoder backbone with Residual Vector Quantization (RVQ) and multi-scale STFT discriminators, enabling high perceptual quality under aggressive compression (3–18 kbps). We adopt the 16 kHz variant in our experiments (`soundstream_16khz`).

Funcodec (Du et al., 2024)[18]: It is an open-source neural speech codec toolkit built to make modern

[14] https://github.com/ZhangXInFD/SpeechTokenizer.git

[15] https://huggingface.co/descript/dac_16khz

[16] https://huggingface.co/facebook/encodec_24khz

[17] https://github.com/haydenshively/SoundStream

[18] https://github.com/modelscope/FunCodec

codec models easy to train, reproduce, and integrate into downstream pipelines. It extends the FunASR ecosystem and provides unified training recipes and inference scripts for widely used neural codec families.

Audiodec (Wu et al., 2023)[19]: It is a high-quality neural codec formulated as an end-to-end autoencoder. Training proceeds in two phases: it first learns the encoder–decoder with metric-based objectives to ensure stable optimization, and then applies an adversarial refinement stage that updates only the decoder to enhance waveform realism. In our experiments, we use the 28 kHz and 48 kHz AudioDec variants.

SNAC (Siuzdak et al., 2024)[20]: It is a multi-scale neural audio codec that extends standard Residual Vector Quantization by allowing different quantizers to operate at different temporal resolutions. Concretely, it forms a hierarchical token representation by quantizing coarse-to-fine structure at multiple frame rates. In this study, we employ various sampling rates, specifically 24 kHz, 32 kHz, and 44 kHz.

B Hyperparameters and System Configurations

Table 6 summarizes the key hyperparameters used in all experiments, including the Poincaré-ball geometry settings, VQ and HEL configurations, OT/Sinkhorn parameters, and optimization details.

C Visualization Analysis

C.1 Confusion Matrices

Figure 2 provides confusion matrices for representative PHOENIX-Mamba configurations across tasks, PTM backbones, and languages. These plots offer a class-wise view of where the model is most reliable and where errors concentrate, complementing the aggregate Acc/F1 trends reported in the main results.

C.2 t-SNE Plots

Figure 3 visualizes learned utterance representations from selected PHOENIX-Mamba configurations. These projections provide an intuitive view of class separability and complement the class-wise error patterns observed in the main results.

Hyperparameter	**Value**
Geometry / Manifold (Poincaré ball)	
Hyperbolic curvature	$\kappa = -1.0$
Hyperbolic embedding dim	$h = 128$
Sequence & Evidence Pooling	
Adapter output dim	$d = 256$
# evidence vectors (glimpses)	$M = 4$
Binary Prototypes (self-discovered positive modes)	
Negative prototypes	$\lvert p_- \rvert = 1$
# positive prototypes (modes)	$K = 4$
Temperature (assignments / soft-min)	$\tau = 0.1$
Loss Weights	
Cluster loss weight	$\lambda = 1.0$
Separation (repulsion) weight	$\beta = 0.1$
Entropy regularizer weight	$\gamma = 0.05$
Optimization	
Optimizer	AdamW
AdamW betas	$(0.9, 0.999)$
AdamW epsilon	10^{-8}
Weight decay	0.01
LR (encoder, if finetuned)	3×10^{-5}
LR (new layers: ϕ, f_θ, W, prototypes)	1×10^{-4}
Gradient clipping	1.0
Epochs	20
Batch size	$\lvert \mathcal{B} \rvert = 32$

Table 6: Hyperparameters for geometry-aware sequence classification with self-discovered hyperbolic clusters. Shared default configuration used across all main experiments.

[19] https://github.com/facebookresearch/AudioDec

[20] https://huggingface.co/hubertsiuzdak/snac_44khz

Figure 2: Confusion matrices for selected PHOENIX-Mamba configurations: (a) Depression with PaSST on Chinese; (b) Depression with Wav2vec 2.0 on English; (c) Dysarthria with PaSST on Chinese; (d) Alzheimer's with PaSST on Chinese; (e) Alzheimer's with Whisper on Chinese; and (f) Alzheimer's with WavLM on Chinese. These plots summarize prediction stability and highlight the dominant error modes for each configuration.

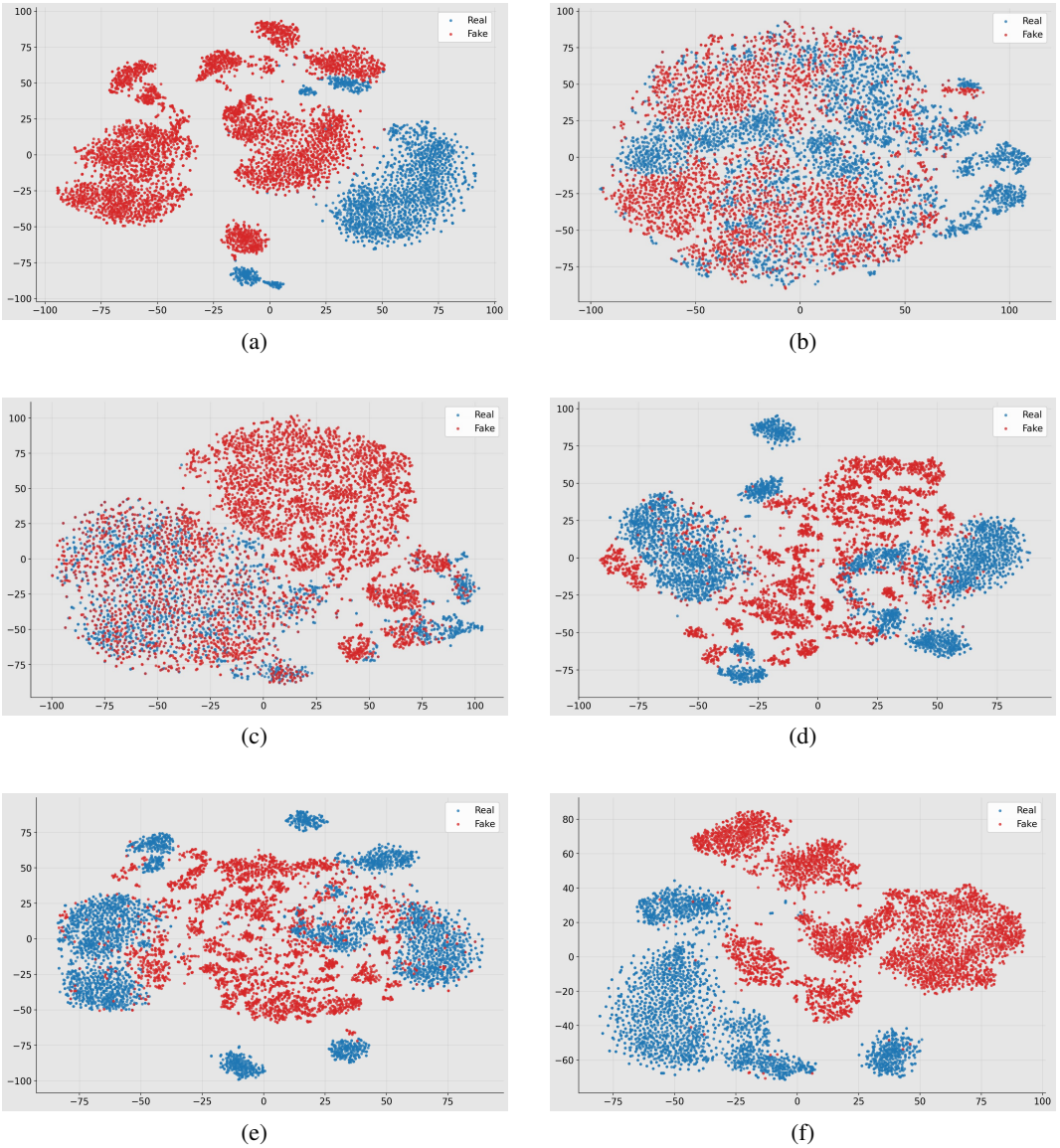

Figure 3: t-SNE visualizations: (a) Depression with PaSST on English PHOENIX-Mamba; (b) Alzheimer's with WavLM on Chinese; (c) Alzheimer's with WavLM on English; (d) Alzheimer's with PaSST on English PHOENIX-Mamba; (e) Dysarthria with PaSST on English PHOENIX-Mamba; and (f) Dysarthria with PaSST on Chinese PHOENIX-Mamba. These plots provide an intuitive view of class separability.