

Profiling the Voice: Speaker-Specific Phoneme Fingerprinting for Speech Deepfake Detection

Jun Xue^{1,2}, Tong Zhang^{1,2}, Zhuolin Yi^{1,2}, Yihuan Huang^{1,2}, Yi Chai^{1,2}, Yiyang Zhang^{1,2} and Yanzhen Ren^{1,2*}

¹Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education

²School of Cyber Science and Engineering, Wuhan University

{junxue, tongzhang, renyz}@whu.edu.com,

Abstract

The rapid advancement of generative AI has made audio deepfakes increasingly indistinguishable from authentic human vocals, posing significant threats to persons-of-interest (POI) such as public figures. Current detection systems primarily rely on generic, black-box models that fail to capture speaker-specific idiosyncratic traits and lack interpretability. In this paper, we propose Phoneme-based Voice Profiling (PVP), a novel personalized defense framework. By shifting the detection paradigm from macro-utterance analysis to micro-phonetic modeling, PVP captures the unique acoustic distributions underlying a POI’s habitual articulatory patterns. Specifically, our framework models speaker-specific phonetic realizations using lightweight Gaussian Mixture Models (GMMs) estimated solely from bona fide reference speech. This design enables data-efficient profiling and robust generalization to previously unseen spoofing attacks without requiring heavy spoof-specific training. Furthermore, we introduce the first large-scale Chinese POI deepfake dataset to benchmark speaker-specific detection. Experimental results demonstrate that PVP significantly outperforms state-of-the-art generic detectors in POI spoofing scenarios, achieving substantial EER reductions while providing fine-grained, phoneme-level interpretability for forensic analysis. Code and data are available at: <https://github.com/JunXue-tech/PVP>

1 Introduction

With the rapid development of artificial intelligence generated content (AIGC), speech synthesis and voice conversion technologies have achieved unprecedented realism, making speech deepfakes increasingly accessible and convincing. As a result, speech deepfake detection has become a critical research topic, especially in high-stakes scenarios such as judicial forensics, governmental communications, and public statements by celebrities or other well-known figures. In

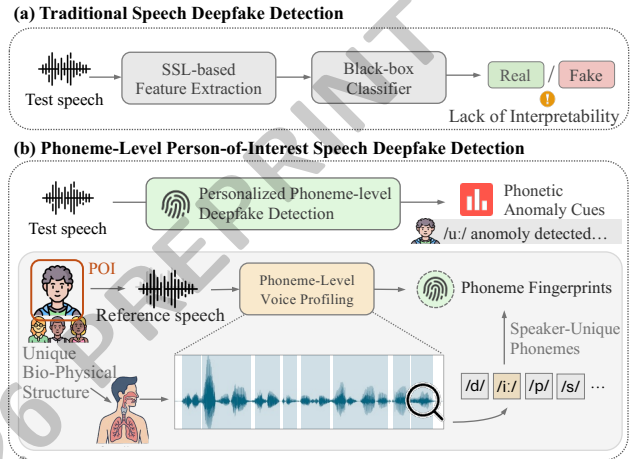
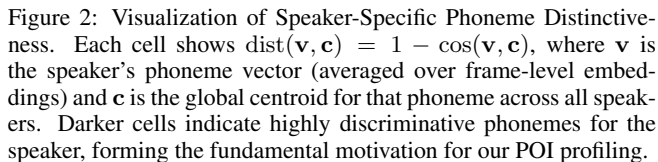


Figure 1: Illustration of our personalized and interpretable detection paradigm. (a) Traditional speech deepfake detection relies on black-box classifiers and lacks interpretability. (b) Our approach performs phoneme-level voice profiling based on the fact that each speaker’s phoneme realizations are shaped by individual vocal characteristics and habitual articulation. By extracting speaker-unique phonemes and constructing “Phoneme Fingerprints” from reference speech, the system conducts granular verification on test speech and uncover “Phonetic Anomaly Cues”, pinpointing specific suspicious phonetic segments for enhanced forensic transparency.

these contexts, a single forged utterance attributed to a specific speaker may lead to severe social, legal, or economic consequences. Therefore, beyond generic detection systems, there is an urgent need for personalized, speaker-specific, and interpretable deepfake speech protection mechanisms.

Most existing studies [Zhang *et al.*, 2025a; Xue *et al.*, 2026; Xue *et al.*, 2023; Xue *et al.*, 2024] formulate speech deepfake detection as a straightforward classification problem, where an entire speech segment is processed by an end-to-end black-box model to output utterance-level decision. A common paradigm is to extract front-end representations using self-supervised learning (SSL) models, followed by a complex back-end classifier to perform discrimination (Fig 1a). Representative works [Tak *et al.*, 2022; Zhang *et al.*, 2024; Tran *et al.*, 2025] adopt convolutional or transformer-based architectures to model global acoustic patterns and have achieved impressive performance on widely used benchmark datasets [Todisco *et al.*, 2019; Müller *et al.*,

*Corresponding author.



However, such evaluations are typically conducted under speaker-agnostic settings, where training and test speakers are loosely constrained. As a result, the learned decision boundaries tend to emphasize dataset-level artifacts—such as synthesis-model-specific spectral patterns, vocoder traces, or corpus-dependent biases—that are consistent across speakers but largely unrelated to individual articulation habits. In contrast, in POI scenarios, attackers often explicitly tailor the synthesis process to a single target speaker by fine-tuning generative models with bona fide speech from that individual [Ali *et al.*, 2025]. Under such conditions, generic detectors often struggle to capture subtle, speaker-dependent inconsistencies and provide limited interpretability regarding *why* a given utterance is detected as fake.

To address this limitation, we turn our attention to the phoneme. Compared to frame-level artifacts that are often transient and model-dependent, phonemes offer a linguistically grounded unit whose acoustic realizations exhibit stable, speaker-specific patterns. Although phoneme categories are shared across speakers, their realizations are shaped by individual vocal characteristics and habitual articulation behaviors, leading to consistent inter-speaker differences. These fine-grained regularities are difficult for current speech synthesis and voice cloning systems to reproduce faithfully, often giving rise to systematic phoneme-level inconsistencies. While recent studies [Zhang *et al.*, 2025b; Baser *et al.*, 2025] have begun to exploit phoneme-related information to guide deepfake detection models, phonemes are still treated as auxiliary signals within speaker-agnostic, utterance-level classifiers. Such designs overlook the fact that phoneme realizations are inherently speaker-dependent and physiologically constrained, limiting their ability to support personalized analysis and interpretable detection in person-of-interest (POI) scenarios.

sis on the *FamousFigures* dataset. We sample genuine speech from ten speakers and extract phoneme-related features using the MMS model [Ge *et al.*, 2025], a multilingual speech deepfake detection system. These features are then visualized as heatmaps. As shown in Fig 2, distinct phoneme pronunciation patterns can be observed across different speakers, indicating that phoneme-level representations capture stable and discriminative speaker-specific characteristics. These observations suggest that phonemes provide a natural and interpretable basis for personalized modeling in speech deepfake detection.

Building on these insights, we propose a novel, plug-and-play phoneme-level modeling framework for speaker-specific deepfake detection (Fig. 1b). Given a small amount of bona fide reference speech from a POI, our method first performs phoneme-level alignment and extracts phoneme-conditioned acoustic representations. For each phoneme, a lightweight Gaussian Mixture Model (GMM) is constructed to characterize the speaker’s habitual articulatory distribution. At inference time, a test utterance is decomposed into phonetic units and evaluated against the corresponding phoneme models, producing fine-grained consistency scores that reflect how well each phoneme realization matches the target speaker’s profile. These phoneme-level consistency scores expose localized speaker-dependent deviations, offering fine-grained and interpretable evidence for deepfake analysis.

Distinct from existing heavy-training paradigms, our method is agnostic to synthesis algorithms, thus reduces overfitting to spoof-specific artifacts, and generalizes naturally to unseen spoofing attacks. Moreover, its additive nature allows it to function as an interpretable “plugin” that can be seamlessly combined with existing detection backbones. Our contributions can be summarized as:

- **Speaker-Specific Phoneme Profiling:** We propose a personalized deepfake detection framework that explicitly captures speaker-specific articulatory patterns using lightweight statistical modeling, enabling effective POI protection with minimal bona fide reference data and without relying on spoof-specific training.
- **Fine-Grained and Interpretable Phonetic Evidence:** We introduce an interpretable phoneme-level scoring and tiered decision mechanism that reveals explicit phonetic anomaly cues, facilitating forensic analysis and transparent reasoning beyond black-box utterance-level classifiers.
- **Chinese POI Deepfake Dataset:** To fill the gap in speaker-centric benchmarks for Mandarin, we present the first large-scale Chinese POI deepfake dataset. It covers diverse public figures and modern speech generation techniques, enabling systematic evaluation of personalized spoofing attacks.

2.1 Self-Supervised Learning for Speech Deepfake Detection

The paradigm of speech deepfake detection has shifted significantly from handcrafted features [Xue *et al.*, 2022;

Lavrentyeva *et al.*, 2019] (e.g., LFCC, CQCC) to representation learning, driven by the success of Self-Supervised Learning (SSL) foundation models. Models such as wav2vec 2.0 [Baevski *et al.*, 2020], HuBERT [Hsu *et al.*, 2021], and WavLM [Chen *et al.*, 2022] leverage massive unlabeled corpora to encode rich acoustic and phonetic information in a latent and entangled manner, proving highly effective at capturing subtle artifacts introduced by neural vocoders and up-sampling layers. While Wav2Vec 2.0 utilizes contrastive predictive coding to retain phase information crucial for detecting synthesis discontinuities, WavLM explicitly incorporates a denoising objective, offering superior robustness against channel variation and background noise. Recent studies have further explored cross-lingual generalization using models like XLS-R [Babu *et al.*, 2021] to address the scarcity of non-English benchmarks, particularly for tonal languages like Mandarin. However, most SSL-based detectors utilize these embeddings as global, utterance-level representations, often discarding the fine-grained, localized phonetic discrepancies that are critical for identifying high-fidelity, targeted spoofing attacks.

2.2 Speech Deepfake Detection: From Agnostic to Speaker-Specific

Existing detection frameworks can be broadly categorized into speaker-agnostic and speaker-specific approaches. Speaker-agnostic methods aim to learn universal forensic features applicable to any voice. Common architectures in this domain, such as RawNet2 [Tak *et al.*, 2021b] and AA-SIST [Jung *et al.*, 2022a], employ sinc-convolutions and graph attention networks, respectively, to model spectro-temporal artifacts. Despite their success on standard benchmarks (e.g., ASVspoof), these “black-box” models often struggle with generalization against unseen generation algorithms and lack interpretability.

Conversely, Speaker-Specific or POI detection reframes the task as a verification problem, leveraging reference data to protect specific identities. Early works [Jung *et al.*, 2022b] integrate automatic speaker verification with anti-spoofing to detect personalized clones, while recent advancements utilize speaker profiles to detect speaker-specific impersonation. Notably, emerging research has begun to investigate leveraging phoneme-level features for enhancing detection performance [Zhang *et al.*, 2025b], or performing phoneme-level analysis on POI [Salvi *et al.*, 2025]. Our work builds upon this interpretable direction by establishing a comprehensive Chinese POI benchmark and introducing an adaptive profiling mechanism that explicitly models these idiosyncratic articulatory patterns.

3 Dataset Construction

To facilitate the study of POI deepfake detection under realistic conditions, we construct a large-scale speaker-centric Mandarin dataset containing both genuine and synthesized speech from public figures. The dataset is designed to benchmark personalized impersonation attacks, enabling fine-grained analysis of speaker-dependent artifacts.

Item	Min	Max	Avg.	Total
Utt. per Speaker	1,848	71,080	20,494	204,944
Real Utt. per Speaker	308	11,848	3,401	34,014
Fake Utt. per Speaker	1,540	59,232	17,093	170,930
Duration per Utt. (s)	1.25	50.00	7.25	—
Total Duration (h)	—	—	—	412.48

Table 1: Dataset statistics of the constructed speaker-specific spoofing detection dataset.

Data Collection and Preprocessing. We collect approximately 400 hours of real speech from 10 target speakers on major Chinese online streaming platforms. To ensure acoustic consistency, all recordings are processed using a voice activity detection (VAD) tool to extract speech-only segments. Non-speech regions and excessively short or long segments are discarded.

To further guarantee speaker purity, we leverage pretrained automatic speaker verification (ASV) embeddings to verify speaker identity. Segments whose embeddings deviate significantly from the target speaker profile are removed, effectively filtering out background speakers, interviewers, and cross-talk commonly present in in-the-wild recordings.

Speech Synthesis. For each target speaker, we generate spoofed speech using five representative zero-shot text-to-speech (TTS) systems: F5-TTS, IndexTTS, LLASA, OpenAudio-S1-mini, and VOXCPM. The selected TTS and VC systems are chosen to cover a wide range of modern zero-shot speech synthesis paradigms, including diffusion-based, autoregressive, and large language model (LLM) based architectures. All systems support reference-based voice cloning, enabling speaker-specific synthesis without fine-tuning. For each synthesis, a single genuine utterance from the target speaker is used as the reference signal, and the synthesis text is kept identical to the corresponding real speech.

Post-processing. All real and synthesized audio samples are resampled to a unified sampling rate and undergo identical post-processing procedures to eliminate confounding factors introduced by format or codec differences. The resulting dataset provides a clean and balanced benchmark for evaluating speaker-specific deepfake detection methods.

Dataset Statistics and Composition. Table 1 summarizes the overall statistics of the constructed dataset. The dataset covers ten target speakers with balanced real and synthesized utterances. For synthesized data, samples are evenly generated by multiple zero-shot TTS and VC systems to ensure diversity in generation mechanisms and artifacts.

4 Methodology

We propose a speaker-specific and phoneme-aware speech deepfake detection framework designed for POI scenarios. Given a small amount of reference speech from a target speaker, our method constructs a personalized phoneme-level acoustic profile and performs interpretable detection by measuring phonetic consistency at inference time.

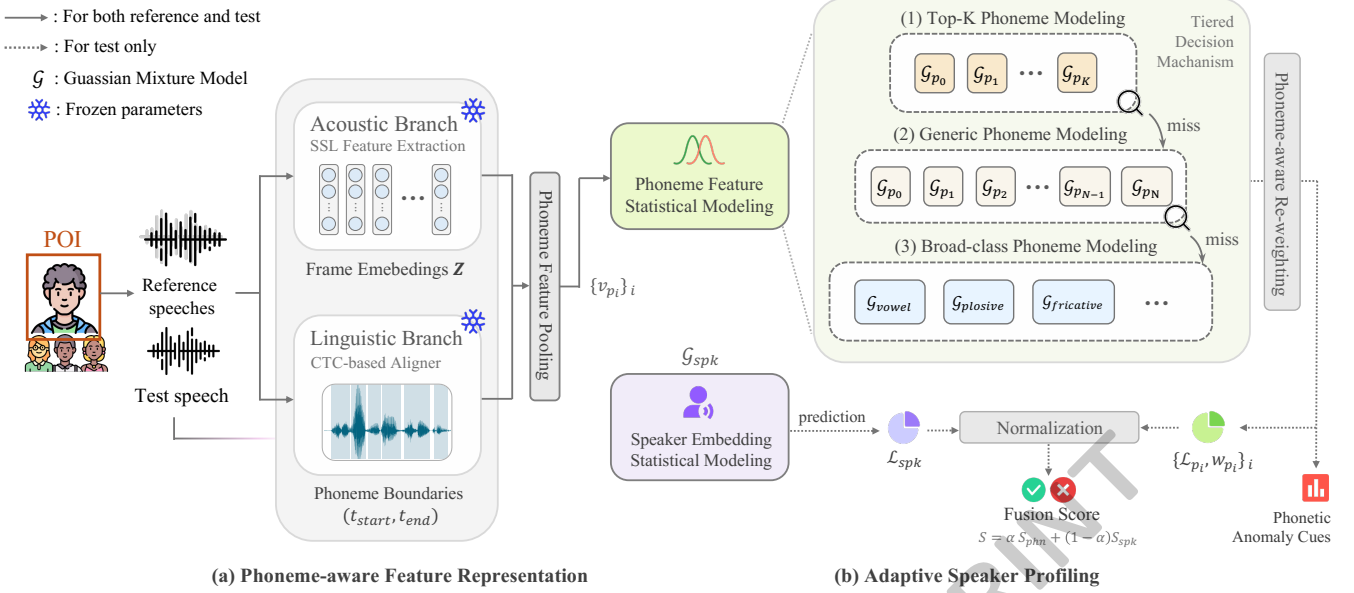


Figure 3: The detailed architecture of our proposed personalized speech deepfake detection framework. The pipeline consists of two primary stages: (a) Phoneme-aware Feature Representation, which utilizes a dual-branch extractor to generate pooled phoneme vectors $\{v_{p_i}\}_i$; and (b) Adaptive Speaker Profiling, which models both fine-grained phonetic distributions and holistic speaker embeddings. A tiered decision mechanism (Top-K, Generic, and Broad-class) is employed to handle linguistic sparsity. Finally, a hybrid score S is derived by fusing the phoneme-level likelihood S_{phn} and the global identity score S_{spk} , enabling the identification of interpretable Phonetic Anomaly Cues.

4.1 Phoneme-aware Feature Representation

To achieve a comprehensive defense, we extract features at two distinct granularities: micro-level phoneme representations and macro-level speaker embeddings.

Phoneme Feature Extraction: We employ SSL model fine-tuned on speech deepfake datasets as the backbone $\mathcal{F}$. For a given input x , the model extracts a sequence of high-dimensional acoustic embeddings:

$$\mathbf{Z} = \mathcal{F}(x) = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T], \quad \mathbf{z}_t \in \mathbb{R}^D \quad (1)$$

where T is the number of frames, and D is the embedding dimension. These features encompass rich prosodic and phonetic information learned from diverse linguistic contexts.

Simultaneously, a Connectionist Temporal Classification (CTC) based aligner $\mathcal{A}$ is utilized to determine the phonetic boundaries. For each phoneme p appearing in the utterance, the aligner provides its temporal interval $[t_{start}, t_{end}]$. To obtain a fixed-length representation for each phoneme instance, we apply a mean-pooling operation over the corresponding frame indices:

$$\mathbf{v}_p = \frac{1}{t_{end} - t_{start} + 1} \sum_{t=t_{start}}^{t_{end}} \mathbf{z}_t \quad (2)$$

The resulting vector $\mathbf{v}_p$ serves as a micro-level descriptor of the speaker’s idiosyncratic pronunciation for phoneme p .

Global Speaker Embedding: To capture the holistic vocal identity, we utilize a pre-trained ASV model, denoted as $\mathcal{F}_{spk}$. For the input utterance x , we extract a fixed-dimensional global speaker embedding:

$$\mathbf{e} = \mathcal{F}_{spk}(x), \quad \mathbf{e} \in \mathbb{R}^{D_{spk}} \quad (3)$$

This embedding encapsulates long-term spectral and prosodic characteristics, providing a robust identity anchor complementary to the transient phonetic features.

4.2 Adaptive Speaker Profiling

A single phoneme may exhibit multiple articulation modes for the same speaker due to coarticulation, prosodic variation, and speaking style. To explicitly model such intra-speaker phonetic variability, we employ GMMs to model the distribution of each phoneme in the speaker’s acoustic space. This formulation enables flexible modeling of intra-speaker variability while retaining a compact and interpretable statistical representation.

Phoneme Statistical Modeling: Let $\mathcal{E}_S = \{x_i\}_{i=1}^N$ denote the reference utterances of the target speaker. For each unique phoneme p in the reference set $\mathcal{E}_S$, we fit a GMM with K_p components. The probability density function for a phoneme vector $\mathbf{v}$ is defined as:

$$P(\mathbf{v}|\lambda_p) = \sum_{k=1}^{K_p} \pi_k \mathcal{N}(\mathbf{v}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (4)$$

where $\lambda_p = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K_p}$ denotes the model parameters, including mixture weights π_k , mean vectors $\boldsymbol{\mu}_k$, and covariance matrices $\boldsymbol{\Sigma}_k$. To prevent over-fitting on sparse data, we restrict $\boldsymbol{\Sigma}_k$ to be a diagonal covariance matrix and adaptively adjust K based on the sample size N_p of the phoneme.

Profile Reliability Weighting: Not all phonemes are equally discriminative or stable for a given speaker. To quantify the reliability of a phoneme profile, we compute a confidence weight w_p based on the average log-likelihood ($\bar{\mathcal{L}}_p$) of

the reference samples:

$$\bar{\mathcal{L}}_p = \frac{1}{N_p} \sum_{i=1}^{N_p} \log P(\mathbf{v}_i | \lambda_p) \quad (5)$$

The weight w_p is then derived using an exponential scaling factor α :

$$w_p = \exp(\bar{\mathcal{L}}_p / \alpha) \quad (6)$$

A higher w_p indicates that the speaker’s pronunciation of phoneme p is highly consistent (compact in the feature space), making it a more reliable “fingerprint” for subsequent deepfake detection.

Global Identity Modeling: To model the speaker’s holistic variance, we fit a separate GMM, denoted as λ_{spk} , to the set of global embeddings $\{\mathbf{e}_i\}_{i=1}^N$ extracted from the reference utterances $\mathcal{E}_S$. Similar to the phonetic branch, we employ a diagonal covariance matrix to prevent overfitting on limited data:

$$P(\mathbf{e} | \lambda_{spk}) = \sum_{k=1}^{K_{spk}} \pi_k \mathcal{N}(\mathbf{e} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (7)$$

where K_{spk} is number of mixture components. This global model serves as a coarse-grained verifier to ensure the test utterance falls within the target speaker’s general acoustic manifold.

4.3 Identification of Salient Phoneme Fingerprints

Phonemes differ in their stability and speaker-discriminative capacity. Rather than treating all phonetic units equally, we identify a speaker-specific subset of *salient phoneme fingerprints*, denoted as $\mathcal{P}_{salient}$, which capture the most consistent articulatory patterns of the POI.

Let $\mathcal{P}_{all}$ denote the set of phoneme types observed in the reference utterances of the target speaker. As described in Algorithm 1, the selection of salient phonemes is driven by the phoneme reliability weights w_p introduced in Section 4.2. Specifically, phonemes in $\mathcal{P}_{all}$ are ranked according to w_p , and only the top- K most reliable units are retained to form $\mathcal{P}_{salient}$. By focusing on phonemes that exhibit high intra-speaker consistency, this selection yields a compact, speaker-specific, and interpretable phonetic fingerprint.

4.4 Tiered Decision Mechanism

In practical POI scenarios, test utterances are often short or linguistically constrained, leading to incomplete phoneme coverage. Let $\mathcal{P}_{test}$ denote the set of phonemes detected in a test utterance. Relying solely on a fixed set of salient phonemes may therefore result in unreliable or missing decisions. To address this issue, we adopt a tiered decision mechanism that adaptively selects the most informative level of phonetic evidence available at inference time.

As illustrated in Algorithm 1, the decision process proceeds in a coarse-to-fine manner. If $\mathcal{P}_{test}$ contains any speaker-specific salient phonemes, the final score is computed as a reliability-weighted average of their normalized phoneme-level scores s_p , yielding a high-precision and interpretable decision. When no salient phonemes are observed

Algorithm 1 Speaker-specific Salient Phoneme Modeling and Tiered Decision

Input: Enrollment set $\mathcal{P}_{all}$, weights $\{w_p\}$, test utterance X , detected phonemes $\mathcal{P}_{test}$, parameter K .

Output: Detection score $S(X)$.

// Salient Phoneme Fingerprint Construction

1: $\mathcal{P}_{sorted} \leftarrow \text{Sort}(\mathcal{P}_{all}, \{w_p\}, \text{desc})$

2: $\mathcal{P}_{salient} \leftarrow \text{SelectTopK}(\mathcal{P}_{sorted}, K)$

// Tiered Decision Mechanism

3: **if** $\mathcal{P}_{test} \cap \mathcal{P}_{salient} \neq \emptyset$ **then**

4: $S(X) \leftarrow \text{WeightedAvg}(\{s_p\}_{p \in \mathcal{P}_{test} \cap \mathcal{P}_{salient}}, \{w_p\})$

5: **else if** $\mathcal{P}_{test} \cap \mathcal{P}_{all} \neq \emptyset$ **then**

6: $S(X) \leftarrow \text{Avg}(\{s_p\}_{p \in \mathcal{P}_{test} \cap \mathcal{P}_{all}})$

7: **else**

8: $\mathcal{C}_{test} \leftarrow \{C(p) \mid p \in \mathcal{P}_{test}\}$

9: $S(X) \leftarrow \text{Avg}(\{s_c\}_{c \in \mathcal{C}_{test}})$

10: **end if**

11: **return** $S(X)$

but other reference phonemes are present, the system falls back to a weighted average over all available phoneme-level scores. In the extreme case where phoneme overlap is sparse, phonemes are mapped to broader phonetic categories via a predefined function $C(\cdot)$ (e.g., vowels, plosives or fricatives), and category-level scores s_c are used to ensure coverage.

This tiered approach prioritizes discriminative, speaker-consistent phonemes whenever possible, while maintaining robustness against linguistic sparsity and content mismatch.

4.5 Hybrid Scoring and Fusion Strategy

Before combining phoneme-level evidence, we normalize raw likelihood scores to ensure comparability across phonemes. For each phoneme p , we compute its log-likelihood $\mathcal{L}_p = \log P(\mathbf{v}_p | \lambda_p)$ under the corresponding phoneme GMM, and apply a Sigmoid-based normalization:

$$s_p = \sigma(\mathcal{L}_p; \beta_{spk}, \gamma_{spk}) = \frac{1}{1 + \exp(-(\mathcal{L}_p - \beta_{spk})/\gamma_{spk})} \quad (8)$$

where β and γ are the centering and scaling hyper-parameters determined empirically.

Finally, we fuse phoneme-level evidence with global speaker identity to produce the final detection score. We employ the tiered decision mechanism to compute the phoneme-level score $S_{phn}(X)$. Parallely, we compute the log-likelihood of the test utterance’s global embedding $\mathbf{e}_{test}$ given the global profile λ_{spk} , followed by Sigmoid normalization:

$$S_{spk}(X) = \sigma(\log P(\mathbf{e}_{test} | \lambda_{spk}); \beta_{spk}, \gamma_{spk}) \quad (9)$$

where β_{spk} and γ_{spk} are normalization parameters centered on the impostor distribution.

To balance the sensitivity to local artifacts and the robustness of global identity, the final detection score S_{final} is obtained via linear interpolation:

$$S_{final}(X) = \alpha \cdot S_{phn}(X) + (1 - \alpha) \cdot S_{spk}(X) \quad (10)$$

where $\alpha \in [0, 1]$ controls the trade-off between phoneme-level and speaker-level evidence.

Backbone	ZH-Famous				En-Famous				Pooled			
	Baseline		with PVP		Baseline		with PVP		Baseline		with PVP	
	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)
hubert-xlarge	63.70	39.80	89.36	19.21	86.88	19.76	93.81	11.96	75.29	29.78	91.58	15.58
wav2vec2-small	74.39	33.08	88.84	19.37	84.58	24.16	96.17	9.93	79.48	28.62	92.50	14.65
wav2vec2-large	78.81	28.19	91.81	15.95	93.93	12.66	98.22	4.46	86.37	20.42	95.01	10.20
wav2vec2-xlsr-1b	88.43	19.17	92.93	14.91	95.65	11.30	98.21	6.66	92.04	15.23	95.57	10.78
mms-300m	84.11	21.13	94.58	11.37	93.99	13.97	96.61	7.24	89.05	17.55	95.59	9.305

Table 2: Plug-and-play evaluation of the proposed Phoneme-based Voice Profiling (PVP) module with different backbone feature extractors on two POI datasets, reported in terms of AUC (↑) / EER (↓) (%).

5 Experiments

5.1 Implementation Details

Datasets. We evaluate our method on two POI deepfake datasets to assess its effectiveness and robustness under diverse linguistic and attack conditions. In addition to our proposed Chinese POI dataset, we adopt the *Famous Figures* dataset [Ali *et al.*, 2025], an English POI deepfake benchmark featuring diverse speech synthesis attacks under zero-shot, few-shot, and fine-tuned settings. Both our dataset (denoted as **ZH-Famous**) and *Famous Figures* (denoted as **EN-Famous**) are designed for POI scenario and include previously unseen synthesis algorithms, making them well suited for assessing speaker-centric and attack-agnostic defenses.

Model Configuration. For phoneme-level alignment, we employ a wav2vec2-large-xlsr-53 model¹ to extract phoneme boundary timesteps. For speaker-level representation, we adopt a pretrained ECAPA-TDNN model² to extract fixed-dimensional global speaker embeddings.

Each phoneme is modeled using a GMM with diagonal covariance matrices. Unless otherwise specified, we utilize 1% of the bona fide speech from each speaker’s available data as reference data for all experiments. We fix the number of mixture components to $K_p = 5$ for phoneme-level models and $K_{spk} = 5$ for the global speaker model, together with a covariance regularization term of 10^{-3} for numerical stability. The number of salient phonemes is set to $K = 12$ across all experiments. For likelihood normalization, we apply a sigmoid mapping with $\beta = -2000$ and $\gamma = 200$. For final score fusion, we set $\alpha = 0.8$.

Evaluation Metrics. We report Equal Error Rate (EER, %) and Area Under the ROC Curve (AUC, %) as the primary evaluation metrics. Lower EER and higher AUC indicate better detection performance.

5.2 Plug-and-Play Evaluation

A core design goal of the proposed Phoneme-based Voice Profiling (PVP) module is plug-and-play compatibility with existing deepfake detection backbones. To validate this property, we integrate PVP with a diverse set of SSL-based encoders [Ge *et al.*, 2025] and evaluate performance on POI and cross-lingual benchmarks.

¹<https://huggingface.co/facebook/wav2vec2-xlsr-53-espeak-cv-ft>

²<https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb>

Model	ZH-Famous		EN-Famous	
	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)
LCNN [Hessam <i>et al.</i> , 2017]	41.73	53.83	43.91	53.81
RawNet2 [Tak <i>et al.</i> , 2021b]	46.51	50.84	35.09	59.01
RawGAT-ST [Tak <i>et al.</i> , 2021a]	38.17	58.94	27.67	68.67
LibriSeVoc [Sun <i>et al.</i> , 2023]	55.12	46.85	50.48	50.40
AASIST [Jung <i>et al.</i> , 2022a]	33.27	60.90	34.04	62.83
XLSR+AASIST [Tak <i>et al.</i> , 2022]	44.54	50.95	45.72	53.06
XLSR+SLS [Zhang <i>et al.</i> , 2024]	54.55	42.65	50.14	45.62
ML-SSLFG [Tran <i>et al.</i> , 2025]	45.24	50.41	64.49	39.50
PLFD-ADD [Zhang <i>et al.</i> , 2025b]	61.63	41.74	47.42	51.85
PVP (Ours)	94.58	11.37	96.61	7.24

Table 3: Comparison with state-of-the-art methods on POI datasets. Results are reported as AUC (↑) / EER (↓) (%).

As summarized in Table 2, incorporating PVP consistently improves detection performance across all backbones on both datasets. On the EN-Famous benchmark, PVP integration yields an average EER reduction of approximately 8.3% and an average AUC improvement of about 5.6% across different backbones. On the ZH-Famous dataset, PVP brings more substantial gains, achieving an average EER reduction of around 12.1% and an average AUC increase of approximately 13.6%, demonstrating strong effectiveness under cross-lingual and unseen generation conditions.

5.3 Comparison with State-of-the-Art Methods

We compare our method with state-of-the-art deepfake detection approaches under a unified speaker-centric evaluation protocol. As shown in Table 3, the proposed method consistently achieves the best performance across both benchmarks, with substantially higher AUC and lower EER than all competing approaches.

Compared to general-purpose detectors, our framework exhibits a clear advantage in capturing fine-grained, speaker-dependent articulatory patterns that are critical for detecting POI speech synthesis. In particular, the large performance margins indicate that modeling phoneme-level consistency from bona fide speech provides strong robustness against unseen spoofing attacks, even without exposure to attack-specific data during modeling. Despite being designed for personalized defense, the method also maintains strong performance across both English and Chinese benchmarks, demonstrating favorable cross-lingual generalization.

Method Variant	ZH-Famous		EN-Famous	
	AUC (↑)	EER (↓)	AUC (↑)	EER (↓)
w/o Phoneme Modeling	92.80	14.95	94.98	9.62
w/o GMM	92.70	14.49	98.16	7.20
w/o Speaker Embedding	92.78	13.44	98.42	7.43
Full PVP (Ours)	94.58	11.37	96.61	7.24

Table 4: Ablation study on the POI benchmarks. Results are reported as AUC (↑) / EER (↓) (%).

5.4 Ablation Study

We conduct ablation studies to examine the impact of key design choices in the proposed PVP framework under the same POI evaluation protocol. Specifically, we replace phoneme-level GMM likelihood modeling with cosine similarity between phoneme embeddings, remove the global speaker embedding branch, and discard phoneme-level profiling in favor of utterance-level modeling, respectively.

Table 4 shows that each component of PVP contributes positively to detection performance under the POI benchmarks. Removing phoneme-level profiling and reverting to utterance-level modeling causes the most consistent degradation, confirming the importance of fine-grained phonetic analysis for capturing speaker-specific articulatory regularities. While certain components may exhibit slightly better performance under specific conditions, their effectiveness is not consistently preserved across datasets and metrics. In contrast, the full PVP framework, which integrates phoneme-level modeling, probabilistic characterization, and global speaker cues, delivers the most stable and averaged-best performance. These results suggest that the strength of PVP lies not in any single module, but in the complementary interplay among its components.

5.5 Interpretability and Case Study

Beyond detection accuracy, our proposed PVP framework provides inherent interpretability by explicitly decomposing the final decision into phoneme-level, speaker-conditioned evidence. Unlike conventional end-to-end detectors that output utterance-level black-box scores, PVP allows each phonetic unit to be independently evaluated against the POI’s articulatory profile learned from bona fide reference speech, making the detection process transparent and traceable.

This focus on phoneme-level speaker inconsistency provides actionable forensic cues while preserving robustness and generalization to previously unseen attacks. Such interpretability arises naturally from the design of PVP rather than post-hoc analysis: by enforcing phoneme-level alignment and probabilistic modeling, detection decisions are grounded in deviations from a speaker’s own phonetic distribution instead of attack-specific patterns. As illustrated in Fig 4, bona fide speech maintains consistently high confidence across segments, whereas the synthetic clone exhibits localized “red flags” at specific phonemes where the generative model fails to replicate the POI’s unique articulatory habits. Moreover, such structured, phoneme-level evidence offers a promising interface for future reasoning-based or large-model-driven

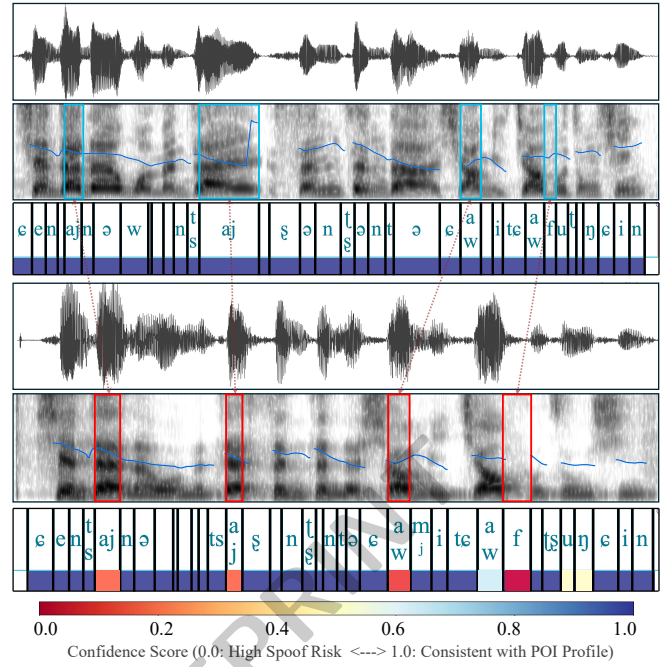


Figure 4: Visualization of Phonetic Interpretability and Anomaly Detection. This figure compares a bona fide utterance (top) with a synthetic clone (bottom) of the same linguistic content. The segments are time-aligned with their respective waveforms, spectrograms, and phoneme labels. The confidence heatmap at the bottom of each panel reflects the consistency of each phoneme with the POI’s articulatory profile: Blue (1.0) indicates high consistency, while Red (0.0) denotes a high spoofing risk.

forensic systems, where explicit articulatory consistency can serve as interpretable intermediate signals.

6 Conclusion

This paper introduces Phoneme-based Voice Profiling (PVP), a plug-and-play and lightweight framework for POI speech deepfake detection. By shifting detection from black-box utterance-level classification to phonetic consistency verification, PVP models speaker-specific articulatory patterns as personalized “fingerprints” using lightweight GMM-based statistics. This design enables high data efficiency and strong generalization to unseen synthesis algorithms, without relying on heavy spoof-specific training.

Extensive evaluations on both our newly curated ZH-Famous dataset and *Famous Figures* dataset demonstrate that PVP consistently outperforms state-of-the-art methods across diverse SSL backbones, achieving substantial reductions in EER. Beyond detection performance, the proposed framework provides transparent phonetic anomaly cues by exposing speaker-centric inconsistencies at the phoneme level, thereby bridging automated deepfake detection with interpretable POI forensic analysis. These findings suggest that phoneme-level voice profiling offers a principled and extensible direction for personalized and interpretable defense against POI speech synthesis.

Acknowledgements

This work is supported by the Natural Science Foundation of China (NSFC) under the grant NO.62572358, 62372334

References

- [Ali *et al.*, 2025] Hashim Ali, Surya Subramani, Raksha Varahamurthy, Nithin Adupa, Lekha Bollinani, and Hafiz Malik. Collecting, Curating, and Annotating Good Quality Speech deepfake dataset for Famous Figures: Process and Challenges. In *Interspeech 2025*, pages 3928–3932, 2025.
- [Babu *et al.*, 2021] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.
- [Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.
- [Baser *et al.*, 2025] Oguzhan Baser, Ahmet Ege Tanriverdi, Sriram Vishwanath, and Sandeep Chinchali. PhoneFake: Redefining Deepfake Realism with Language-Driven Segmental Manipulation and Adaptive Bilevel Detection. In *Interspeech 2025*, pages 5333–5337, 2025.
- [Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.
- [Dawalatabad *et al.*, 2021] Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na. Ecapa-tdnn embeddings for speaker diarization. In *Interspeech 2021*, pages 3560–3564, 2021.
- [Ge *et al.*, 2025] Wanying Ge, Xin Wang, Xuechen Liu, and Junichi Yamagishi. Post-training for deepfake speech detection. *arXiv preprint arXiv:2506.21090*, 2025.
- [Hessam *et al.*, 2017] Hessam, Mohammad Rastegari, and Ali Farhadi. Lcnn: Lookup-based convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7120–7129, 2017.
- [Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.
- [Jung *et al.*, 2022a] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. 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. IEEE, 2022.
- [Jung *et al.*, 2022b] Jee-weon Jung, Hemlata Tak, Hye-jin Shim, Hee-Soo Heo, Bong-Jin Lee, Soo-Whan Chung, Ha-Jin Yu, Nicholas Evans, and Tomi Kinnunen. Sasv 2022: The first spoofing-aware speaker verification challenge. *arXiv preprint arXiv:2203.14732*, 2022.
- [Lavrentyeva *et al.*, 2019] Galina Lavrentyeva, Sergey Novoselov, Andzhukaev Tseren, Marina Volkova, Artem Gorlanov, and Alexandr Kozlov. Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576*, 2019.
- [Müller *et al.*, 2022] Nicolas M. Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *Interspeech*, 2022.
- [Reynolds *et al.*, 2000] Douglas A Reynolds, Thomas F Quatieri, and Robert B Dunn. Speaker verification using adapted gaussian mixture models. *Digital signal processing*, 10(1-3):19–41, 2000.
- [Salvi *et al.*, 2025] Davide Salvi, Viola Negroni, Sara Mandelli, Paolo Bestagini, and Stefano Tubaro. Phoneme-level analysis for person-of-interest speech deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1586–1595, 2025.
- [Sun *et al.*, 2023] Chengzhe Sun, Shan Jia, Shuwei Hou, and Siwei Lyu. Ai-synthesized voice detection using neural vocoder artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 904–912, 2023.
- [Tak *et al.*, 2021a] Hemlata Tak, Jee-weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv preprint arXiv:2107.12710*, 2021.
- [Tak *et al.*, 2021b] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373. IEEE, 2021.
- [Tak *et al.*, 2022] Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.
- [Todisco *et al.*, 2019] Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech 2019*, pages 1008–1012, 2019.

- [Tran *et al.*, 2025] Hoan My Tran, Damien Lolive, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau, and David Guennec. Multi-level ssl feature gating for audio deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 11766–11775, 2025.
- [Xue *et al.*, 2022] Jun Xue, Cunhang Fan, Zhao Lv, Jianhua Tao, Jiangyan Yi, Chengshi Zheng, Zhengqi Wen, Minmin Yuan, and Shegang Shao. Audio deepfake detection based on a combination of f0 information and real plus imaginary spectrogram features. In *Proceedings of the 1st international workshop on deepfake detection for audio multimedia*, pages 19–26, 2022.
- [Xue *et al.*, 2023] Jun Xue, Cunhang Fan, Jiangyan Yi, Chenglong Wang, Zhengqi Wen, Dan Zhang, and Zhao Lv. Learning from yourself: A self-distillation method for fake speech detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [Xue *et al.*, 2024] Jun Xue, Cunhang Fan, Jiangyan Yi, Jian Zhou, and Zhao Lv. Dynamic ensemble teacher-student distillation framework for light-weight fake audio detection. *IEEE Signal Processing Letters*, 31:2305–2309, 2024.
- [Xue *et al.*, 2026] Jun Xue, Zhuolin Yi, Yihuan Huang, Yanzhen Ren, Yujie Chen, Cunhang Fan, Zicheng Su, Yonghong Zhang, and Bo Cai. Rtcfake: Speech deepfake detection in real-time communication. *arXiv preprint arXiv:2604.23742*, 2026.
- [Zhang *et al.*, 2024] Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6765–6773, 2024.
- [Zhang *et al.*, 2025a] Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yifang Guo, Yushu Zhang, and Guoai Xu. Multi-view collaborative learning network for speech deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1075–1083, 2025.
- [Zhang *et al.*, 2025b] Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yushu Zhang, and Yifang Guo. Phoneme-level feature discrepancies: A key to detecting sophisticated speech deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1066–1074, 2025.