Phoneme-Level Analysis for Person-of-Interest Speech Deepfake Detection

Davide Salvi, Viola Negroni, Sara Mandelli, Paolo Bestagini, Stefano Tubaro

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Italy

{davide.salvi, viola.negroni, sara.mandelli, paolo.bestagini, stefano.tubaro}@polimi.it

Abstract

Recent advances in generative AI have made the creation of speech deepfakes widely accessible, posing serious challenges to digital trust. To counter this, various speech deepfake detection strategies have been proposed, including Person-of-Interest (POI) approaches, which focus on identifying impersonations of specific individuals by modeling and analyzing their unique vocal traits. Despite their excellent performance, the existing methods offer limited granularity and lack interpretability. In this work, we propose a POI-based speech deepfake detection method that operates at the phoneme level. Our approach decomposes reference audio into phonemes to construct a detailed speaker profile. In inference, phonemes from a test sample are individually compared against this profile, enabling fine-grained detection of synthetic artifacts. The proposed method achieves comparable accuracy to traditional approaches while offering superior robustness and interpretability, key aspects in multimedia forensics. By focusing on phoneme analysis, this work explores a novel direction for explainable, speaker-centric deepfake detection.

1. Introduction

Recent advancements in generative AI have made it possible to produce hyper-realistic synthetic content with unprecedented ease. While these technologies hold enormous potential for applications in entertainment, education, and accessibility, they also pose significant risks to security, privacy, and trust in digital communication [17]. One of the most concerning menaces is represented by deepfakes, synthetic multimedia content generated through Deep Learning (DL) techniques that depict individuals performing actions or making statements they never actually did [22, 40]. In the audio domain, speech deepfakes allow the generation of synthetic speech that mimics the voice of a target speaker with remarkable realism, enabling malicious applications such as impersonation, fraud, and disinformation.

To combat the growing risks associated with the misuse of synthetic speech, the multimedia forensics community
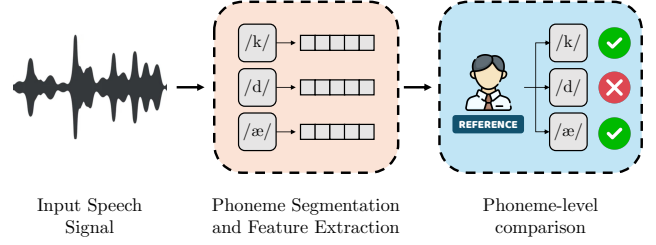

Figure 1. Proposed phoneme-based Person-of-Interest (POI) speech deepfake detection pipeline.

has developed a wide range of countermeasures, primarily aimed at protecting Automatic Speaker Verification (ASV) systems and enhancing deepfake detection by reliably distinguishing between real and fake speech signals [13, 31]. To this end, a variety of approaches have been explored, including transformer-based architectures [6, 14, 45], Mixture of Experts (MoE) models [26, 42], and others DL-based techniques [4, 19, 37, 44, 47].

More recently, POI-oriented methods have emerged as a promising strategy to enhance speaker-specific protection [1, 12]. These approaches create a speaker profile based on a set of reference recordings from a target individual requiring protection. During inference, the profile is used to compute the similarity between a given speech sample and the reference speaker's characteristics, determining the authenticity of the analyzed track [29]. Compared to conventional deepfake detection systems, POI-based methods offer two key advantages. First, their speaker-centric nature leads to higher accuracy in detecting attacks against specific individuals. Second, since they rely exclusively on real speech data rather than training on synthetic samples, they proved remarkable generalization capabilities and enhanced robustness against various deepfake generation techniques [48].

One of the most widely used approaches in POI-based methods involves representing each speech sample with a single feature vector, obtained using either supervised learning models or pre-trained foundation models [30]. While this approach has demonstrated strong detection performance, it is inherently limited in granularity. Representing an entire speech segment with a fixed-length embedding

may result in the loss of fine-grained cues that are critical for detecting synthetic speech. Additionally, this coarse representation reduces interpretability, as it does not provide information about which specific aspects of the speech signal contribute to the final decision. As a result, it may be more susceptible to relying on spurious artifacts or shortcuts rather than genuine indicators of synthetic speech [41].

In this work, we propose a phoneme-level approach to POI-based speech deepfake detection, shifting from full-signal analysis to a novel fine-grained phoneme-centric evaluation, as illustrated in Figure 1. Rather than processing the entire speech signal as a whole and extracting a single, global feature representation, our method decomposes the input audio into individual phonemes and then processes each phonetic unit independently to extract a dedicated feature vector from it. These phoneme-level embeddings, derived from reference speech, are aggregated to construct a speaker-specific phoneme profile. During inference, the test audio is similarly decomposed into phonemes, and each of these is individually compared to the speaker profile, enabling the detection of synthetic artifacts at a much finer temporal and linguistic resolution.

Our method is grounded on the idea that speech deepfake generators, despite their sophistication, struggle to replicate short, fundamental units of speech, such as phonemes, with perfect fidelity. By refining the analysis and focusing only on the most critical parts of the signal, our method identifies the specific phonemes that deviate significantly from the reference speaker's characteristics, resulting in a more robust and reliable analysis. Moreover, the proposed approach inherits the advantages of one-class POI-based methods, making it more resilient to different deepfake generators and diverse recording conditions.

This work explores phoneme-level analysis as a promising approach for POI-based speech deepfake detection. Our findings support the viability of this method and suggest future research directions, such as integrating phonetic knowledge into end-to-end detectors and developing more interpretable, speaker-centric solutions. Understanding which specific aspects of speech can be reliably used to discriminate between real and fake audio, and localizing generation artifacts within a signal in both time and frequency domains, could serve as valuable tools to combat voice spoofing in an era of increasingly realistic synthetic media.

The rest of the paper is organized as follows. Section 2 provides background on POI and phoneme-based speech deepfake detection methods. Section 3 formally defines the problem at hand and details the proposed detection pipeline. Section 4 outlines the experimental setup used to evaluate the method. Section 5 presents the results along with analysis and discussion. Finally, Section 6 summarizes the key findings and outlines directions for future work.

2. Background

POI-Based Speech Deepfake Detection. The rapid advancement of synthetic speech generation techniques has led to growing interest in developing reliable methods for detecting speech deepfakes. In recent years, the research community has proposed a wide range of techniques based on diverse detection paradigms [3, 20], framing the task as a binary classification problem and relying heavily on supervised learning.

While these supervised methods have shown promising results, they also have important limitations. For instance, accurately identifying speech generated by methods not seen during training remains a significant challenge and undermines the robustness of these systems in real-world conditions. To address this limitation, Person-of-Interest (POI)-based methods have emerged as an alternative. These approaches focus on speaker-specific analysis, training exclusively on real data from a given target speaker. This approach leads to improved accuracy and robustness for the enrolled speaker, as well as enhanced generalization across a wide range of speech generation techniques.

While several POI-based deepfake detection methods have been proposed in the video domain [1, 2, 11, 12], their application to audio remains relatively underexplored. Notable contributions in this space include the work by Pianese et al. [29, 30], who introduced a POI-based framework for speech deepfake detection that reframes the problem as a speaker verification task: assessing whether the voice in a test sample matches the identity it claims. This approach falls between the speaker verification and speech deepfake detection tasks, providing strong performance in both fields. The contamination between ASV and speech deepfake detection has also been highlighted in other recent studies [16, 43], which observed that ASV systems often exhibit an inherent capacity to reject spoofed audio without explicit training for that purpose.

A related and more recent POI-inspired approach is presented in [10], where the authors frame speech deepfake detection as an anomaly detection problem. By training solely on real data, the system learns to detect deviations characteristic of synthetic speech. The use of speaker-specific models in this context closely aligns with POI settings and provides compelling evidence for the efficacy of one-class learning in voice spoofing detection.

Phoneme-Based Speech Deepfake Detection. Phoneme-based speech deepfake detectors are a class of forensic classifiers that leverage the linguistic structure of speech by extracting information tied to phonetic content, which can be used to improve the detection capabilities of the model.

One of the earliest studies to investigate this aspect is [36]. There, the authors analyzed phonetic variations in

replay attacks, motivated by earlier works in emotion recognition [35] and speaker verification [21], where incorporating phonetic information proved effective. The underlying idea behind this work is that different phonemes are affected differently by the channel distortions introduced during replay attacks, due to their distinct spectral characteristics. Specifically, some phonemes emphasize high-frequency energy, while others concentrate on lower frequency bands. To validate their hypothesis, the authors use Rectangular Filter Cepstral Coefficients (RFCCs) and a phoneme recognizer [34] to train phoneme-specific pairs of Gaussian Mixture Models (GMMs). Their analysis identified fricatives, nasals, and silences as particularly informative for detecting replay attacks, aligning with prior findings that showed that replay attacks introduce artifacts primarily in high-frequency regions [25].

Building on this idea, Dhamyal et al. [15] proposed the first explainable DL-based approach to study phonetic differences between real and fake speech. Inspired by speech recognition techniques [9, 27], they introduced a self-attention mechanism within a SENet model trained on Constant-Q Cepstral Coefficients (CQCCs). Attention weights were then aligned with annotated phoneme boundaries using Dynamic Time Warping (DTW), enabling frame-level interpretability. Their findings reinforced the importance of fricatives and nasals but also showed that focusing solely on vowels yielded detection performance that surpassed all individual phoneme classes.

More recently, Zhang et al. [46] proposed a phoneme-sequence-based approach to deepfake detection. Unlike prior methods that analyze phonemes in isolation, their technique models the temporal sequence of phonemes across an entire utterance. They fine-tuned a pre-trained Large Language Model (LLM) for phoneme recognition and aligned frame-level acoustic features, extracted using a shared encoder, with corresponding phonemes. These sequences were then passed to a binary classifier to determine whether the utterance was real or fake. While this method sacrifices the fine-grained analysis of specific phoneme types, it achieves performance comparable or superior to many state-of-the-art detectors.

In contrast to these prior works, our method leverages phonemes as foundational units in a POI-based detection framework. It is based on the hypothesis that synthetic speech generators struggle to accurately replicate a speaker's unique phonetic patterns. In practical scenarios, when a reference speaker can provide a large number of real tracks to protect their identity, our method constructs a personalized phoneme dictionary using compact embeddings derived from these genuine samples. Test recordings can then be analyzed by comparing their phonemes against this dictionary to assess whether they match the target speaker's genuine speech characteristics or not.
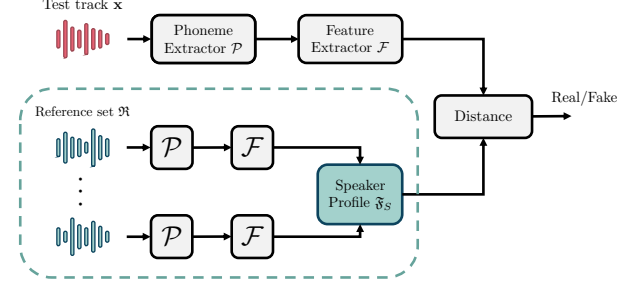

Figure 2. Pipeline of the proposed phoneme-level POI speech deepfake detection method.

3. Proposed Method

In this section, we formalize the POI speech deepfake detection task and introduce our phoneme-level detection framework.

3.1. Problem Formulation

The POI-based speech deepfake detection problem is formally defined as follows. Let $\mathfrak{R} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{I-1}\}$ denote a set of I speech tracks $\mathbf{x}$, all uttered by a target speaker S. These tracks are taken as references to construct a speaker profile, which captures the unique vocal characteristics of S. During inference, we are given a test speech signal $\mathbf{x}$, sampled at a frequency f_s and associated with a binary class label $y \in \{0, 1\}$, where $y = 0$ denotes that $\mathbf{x}$ is an authentic recording of S, while $y = 1$ indicates that $\mathbf{x}$ is a deepfake attempting to imitate S. The goal of the task is to predict the class label y of the test signal $\mathbf{x}$ by measuring its distance from the speaker profile.

3.2. Proposed System

In this paper, we propose a phoneme-based approach to POI speech deepfake detection. Our detection pipeline, illustrated in Figure 2, is composed of the following steps:

Phoneme Extraction: Each speech signal $\mathbf{x}$ (including both reference and test tracks) is processed by a *Phoneme Extractor* $\mathcal{P}$ to segment it into its constituent phonemes. For each phoneme a in the phonetic alphabet $\mathfrak{A}$, which includes all the International Phonetic Alphabet (IPA) phonemes, we define the set $\mathfrak{P}_{\mathbf{x}}^a = \{p_0^a, p_1^a, \ldots, p_{N-1}^a\}$ that comprises all the realizations of the phoneme a in $\mathbf{x}$. This segmentation enables fine-grained analysis of the speech signal.

Feature Extraction: Each phoneme instance p_i^a is processed by a *Feature Extractor* $\mathcal{F}$, generating a d-dimensional feature vector $\mathbf{f}_i^a \in \mathbb{R}^d$. This results in a set $\mathfrak{F}_{\mathbf{x}}^a$ which contains features corresponding to all phonemes in $\mathfrak{P}_{\mathbf{x}}^a$. Each embedding $\mathbf{f}_i^a$ encodes speaker-specific phoneme characteristics that are critical to distinguish between real and fake speech. To represent the entire track $\mathbf{x}$ we aggregate all the sets $\mathfrak{F}_{\mathbf{x}}^a$ as in $\mathfrak{F}_{\mathbf{x}} = \{\mathfrak{F}_{\mathbf{x}}^a\}_{a \in \mathfrak{A}}$.

Speaker Profile Construction: To build a target speaker profile, we repeat the phoneme and feature extraction steps across a set of reference tracks $\mathfrak{R}$ from the person of interest. This process produces the speaker profile set $\mathfrak{F}_S$, which captures the phoneme-level acoustic patterns unique to the target speaker S.

Inference: Given a test track $\mathbf{x}$, we compute its feature representation $\mathfrak{F}_{\mathbf{x}}$ as described above. For each phoneme a, the feature vector $\mathbf{f}_i^a$ is compared against the corresponding phoneme representation in $\mathfrak{F}_S$ using a distance function, measuring how closely each test phoneme matches the reference speaker's typical realization.

Classification: Phoneme-level distances are aggregated to compute a global similarity score between the test signal and the speaker profile. This score quantifies the overall deviation of $\mathbf{x}$ from the expected voice pattern of speaker S. This score is then used to classify the test sample as either authentic ($y = 0$) or synthetic ($y = 1$).

Our phoneme-level strategy offers several advantages over traditional speech deepfake detection methods. While conventional models provide only a binary classification, our framework can identify the specific phonemes and time intervals where synthetic artifacts occur. This fine-grained analysis enhances interpretability, a critical aspect in multimedia forensics. Additionally, by focusing only on the most relevant portions of the signal itself, i.e., phonemes, rather than analyzing the entire track, our method is inherently robust to post-processing operations that inject noise in the recording or apply lossy compression, as we will show in Section 5.2.

4. Experimental Setup

In this section, we detail the experimental setup used in our analyses. We begin by describing the speech processing models we used for phoneme extraction and feature computation. Next, we outline the methodology for constructing speaker profiles and measuring distances between test tracks and reference profiles, along with the baseline we used to validate our findings. Finally, we introduce the datasets used in our experiments.

4.1. Speech Processing Models

To process the input speech data, we employed two distinct pre-trained instances of Wav2Vec 2.0 [5], a state-of-the-art self-supervised learning model speech representation. The first model is the Phoneme Extractor $\mathcal{P}$ and consists of a fine-tuned version of Wav2Vec 2.0 on the LJSpeech Phonemes dataset [8], enabling it to predict phoneme sequences from raw audio waveforms. The model processes the speech signal by dividing it into $25\,\mathrm{ms}$ frames with $5\,\mathrm{ms}$ overlap and assigns a phoneme to each frame where applicable. The second model is the Feature Extractor $\mathcal{F}$ and consists of the *base* version of Wav2Vec 2.0. It operates following the same frame-based processing pipeline as $\mathcal{P}$ but outputs a dense feature vector of dimension $d = 768$ for each frame, encoding rich acoustic information.

The two models operate sequentially: the input speech track $\mathbf{x}$ is first processed by $\mathcal{P}$ to extract phonemes and then passed through $\mathcal{F}$ to compute feature vectors. Since the two models share the same architecture and processing pipeline, their outputs are inherently time-aligned. This allows us to directly associate each phoneme prediction from $\mathcal{P}$ with a corresponding feature vector from $\mathcal{F}$.

We retain only those feature vectors that are aligned with phoneme-labeled frames, discarding features corresponding to non-linguistic segments such as silence or unvoiced frames. In cases where a phoneme spans multiple consecutive frames, the associated feature vectors are averaged to produce a single representative embedding for that phoneme. Prior to model inference, all audio signals are normalized to have zero mean and unit variance. This standardization step reduces variability in amplitude and dynamic range across the tracks, enabling more robust phoneme prediction and feature extraction.

4.2. Speaker profile and distance computation

For each speaker S, we construct a profile $\mathfrak{F}_S$ in the form of a dictionary, where each key represents a phoneme a, and the associated value is the set of feature vectors $\mathfrak{F}_S^a$ for that phoneme across all the reference tracks, as in

$$\mathfrak{F}_S = \{a \mapsto \mathfrak{F}_S^a \mid a \in \mathfrak{A}_S\}, \tag{1}$$

where $\mathfrak{A}_S \subseteq \mathfrak{A}$ denotes the set of phonemes realized by the speaker S in the reference tracks.

The number of reference utterances we use to construct each speaker profile is $I = 100$. This value is selected as a trade-off based on empirical validation, satisfying two main conditions. The first one is phoneme coverage. With $I = 100$ utterances per speaker, we ensure that most, if not all, phonemes in the speaker's inventory are sufficiently represented in $\mathfrak{F}_S$, meaning that $\mathfrak{A}_S \approx \mathfrak{A}$. Using fewer utterances may lead to data sparsity, where some phonemes are underrepresented or missing altogether. The second condition is dataset balance, a practical consideration driven by experimental constraints. Limiting the number of reference tracks to $I = 100$ allows the retention of a sufficient number of remaining utterances for test purposes across all speakers. Using a larger value for I would disproportionately reduce the number of test samples available for some speakers, especially in datasets with limited total utterances per speaker.

During inference, the feature vector of each phoneme in a given test track $\mathbf{x}$ is compared to all corresponding phoneme entries in the speaker profile $\mathfrak{F}_S$. Phoneme-level similarity is computed using the minimum element-wise cosine distance, following the approach of Pianese et al. [30].

This process is repeated for all phonemes in the test track, and the final distance between the test track $\mathbf{x}$ and the speaker profile $\mathfrak{F}_S$ is computed as the average of the individual phoneme distances. Formally, for each phoneme occurrence $(p_i^a, \mathbf{f}_i^a)$, we compute the cosine distance to all vectors in the corresponding entry $\mathfrak{F}_S^a$ of the speaker profile, and retain the minimum distance, as in

$$d_i = \min_{\mathbf{f} \in \mathfrak{F}_S^a} \left(1 - \frac{\langle \mathbf{f}_i^a, \mathbf{f} \rangle}{\|\mathbf{f}_i^a\| \cdot \|\mathbf{f}\|} \right). \tag{2}$$

Then, the overall distance between $\mathbf{x}$ and $\mathfrak{F}_S$ is obtained as

$$D(\mathbf{x}, \mathfrak{F}_S) = \frac{1}{N} \sum_{n=0}^{N-1} d_n, \tag{3}$$

where N is the number of phonemes contained in $\mathbf{x}$. This final score $D(\mathbf{x}, \mathfrak{F}_S)$ is used as a decision metric for classifying $\mathbf{x}$ as either authentic or synthetic.

4.3. Baseline

To validate the effectiveness of the proposed pipeline, we compare it against a baseline method that follows the traditional POI speech deepfake detection framework outlined in [30]. In this approach, each speech track $\mathbf{x}$ is represented by a single global feature vector $\mathbf{f_x}$, computed as the average of the frame-level feature vectors generated by the feature extraction model $\mathcal{F}$, as in

$$\mathbf{f_x} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbf{f_x}^t, \tag{4}$$

where T is the number of frames contained in $\mathbf{x}$.

Unlike our phoneme-level approach, which constructs a structured speaker profile as a dictionary indexed by phonemes, this baseline represents each speaker as a flat set of utterance-level feature vectors, each representing an entire reference signal, as in

$$\mathfrak{F}_S = \left\{ \mathbf{f}_{\mathbf{x}_i} \in \mathbb{R}^d \mid i = 1, \ldots, I \right\}, \tag{5}$$

where $I = 100$ is the number of reference utterances per speaker, consistent with our proposed method to ensure a fair comparison.

During inference, the similarity between a test track and a speaker profile is determined by computing the minimum cosine distance between the test track's feature vector $\mathbf{f_x}$ and the feature vectors in $\mathfrak{F}_S$. This distance serves as the final decision score, used to assess the authenticity of $\mathbf{x}$.

4.4. Datasets

We evaluate the proposed framework on four publicly available English-language speech deepfake datasets to provide a comprehensive assessment of its generalization capabilities across diverse conditions and synthesis techniques.

All audio data are sampled to a uniform sampling rate of $f_s = 16\,\text{kHz}$.

ASVspoof 2019 [38]. Released for the homonymous challenge, this dataset was designed to develop effective ASV models. We use its Logical Access-*eval* partition, which includes real speech samples from VCTK [39] and synthetic speech generated by 13 different speech synthesis models. The dataset features 67 speakers, enabling speaker-specific analyses. Detection performance is reported as the average across all speakers.

In-the-Wild [24]. This dataset is designed to evaluate speech deepfake detectors in real-world conditions. It consists of approximately 38 hours of audio data (17 hours fake, 21 hours real) featuring 54 celebrities and politicians. The fake clips were created by segmenting publicly accessible video and audio files, while the real clips come from publicly available material featuring the same speakers.

Purdue speech dataset [7]. This corpus includes 25 000 synthetic speech tracks generated by 5 diffusion model-based voice cloning methods: ProDiff, DiffGAN-TTS, ElevenLabs, UnitSpeech, and XTTS. Real speech data are sourced from LJSpeech [18] and 10 speakers from LibriSpeech [28].

TIMIT-TTS [32]. This is a speech dataset that includes only fake audio samples, generated from 12 different Text-to-Speech (TTS) methods that reproduce the voice of Linda Johnson from LibriVox. It is created based on the Vid-TIMIT corpus [33] by generating a synthetic copy of its tracks using the considered synthetic speech generators. We consider the *single speaker/clean* partition of this set, pairing it with real speech data from LJSpeech for evaluation.

5. Results

In this section, we evaluate the performance of the proposed phoneme-based framework for speech deepfake detection. The analysis aims to validate our hypothesis regarding the benefits of phoneme-level information for robustness and interpretability in the task at hand.

5.1. Detection performance

In our first experiment, we compare the speech deepfake detection performance of the proposed method against the considered baseline. As described in Section 4, both frameworks rely on the same feature encoder, i.e., the *base* version of Wav2Vec 2.0, but differ in treating the input signal. The proposed model incorporates a phoneme-level pre-processing step, while the baseline processes the entire signal without any linguistic structuring.

Table 1 shows the results of this analysis in terms of Area Under the Curve (AUC) and Equal Error Rate (EER) across the four evaluated speech deepfake datasets. The overall performance is comparable between the two methods. The

Table 1. Performance metrics (AUC and EER) for the considered methods, presented as percentage values (%).

	ASVspoof 2019		In-the-Wild		Purdue		TIMIT-TTS	
	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓
Baseline	**98.12**	**6.84**	74.69	30.88	81.32	26.49	84.54	**21.33**
Ours (phoneme)	79.45	25.82	**78.19**	**27.48**	**89.49**	**17.55**	**84.96**	23.56

Figure 3. Comparison of average input signal duration (in seconds) processed by the baseline and proposed frameworks across the considered datasets.

baseline achieves superior results on the ASVspoof 2019 dataset, while the proposed phoneme-level approach outperforms it on both the In-the-Wild and Purdue datasets. For the TIMIT-TTS corpus, the two approaches yield nearly identical performance.

Although the improvement in absolute detection performance might appear incremental, a deeper analysis reveals that it represents a significant finding. Figure 3 shows that the proposed phoneme-based model processes, on average, over 65% less of the input signal during inference, while still achieving comparable or superior detection performance. This efficiency stems from the core idea of our framework, which selectively analyzes only the portions of the signal associated with phonemes while discarding non-informative regions for the task at hand, such as silence or acoustically unclear content (see Section 4.1).

This analysis also provides insight into the relatively poor performance of the proposed approach on the ASVspoof 2019 dataset with respect to the baseline (Table 1). In this specific corpus, the average amount of analyzed content per utterance is often shorter than one second, severely limiting the amount of information that can be used to perform the detection process. This limitation, combined with known issues in ASVspoof 2019 regarding silences [23] that may push the baseline performance, could be the reason behind the observed performance disparity. On the other hand, our performance improves significantly on datasets containing longer test utterances, such as Purdue. Notably, this dataset contains the longest utterances among those evaluated and is also where our approach achieves the greatest performance improvement over the baseline.

These results indicate that the phoneme-based approach effectively isolates the most informative segments of the signal, capturing the critical aspects necessary for determining its authenticity. The benefits are twofold: (i) a more compact and interpretable decision process and (ii) improved efficiency and potential scalability in practical deployment scenarios. We further investigate these advantages in the following experiments.

5.2. Robustness to Post-Processing Perturbations

We now evaluate the robustness of the proposed method to common post-processing perturbations, such as additive noise, lossy compression, and quantization artifacts, in comparison to the baseline system. Our underlying hy-
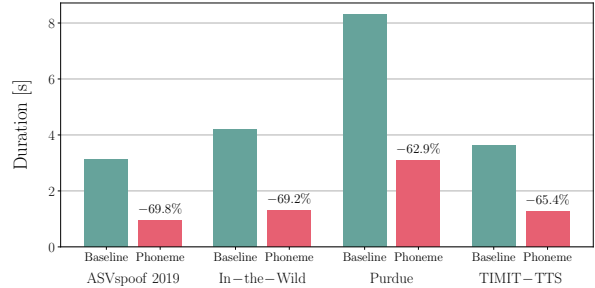
pothesis is that the phoneme-based framework, by focusing on the most informative and meaningful regions of the signal, is inherently more robust to degradations that predominantly affect less critical or non-phoneme regions. In contrast, the baseline system, which processes the entire waveform indiscriminately, may extract features from segments that are more vulnerable to such perturbations, potentially degrading its performance under adverse conditions.

To test this hypothesis, we subject the test tracks to three types of distortions: additive Gaussian noise at various Signal-to-Noise Ratio (SNR) levels ranging from 25 to 10 dB, MP3 compression at a bitrate of 128 kbps, and 8-bit μ-law quantization. Processing is applied only to the test tracks, while the reference set remains clean. This setup simulates realistic conditions for POI methods, where clean reference recordings of the target speaker are assumed to be available, but test samples may be subject to noise or other distortions.

Table 2 presents the results of this analysis in terms of EER variations (ΔEER) relative to the clean condition. The results confirm our hypothesis: the proposed method consistently outperforms the baseline in terms of robustness and exhibits lower sensitivity to noise. Notably, the EER increase of the phoneme-based method remains below 5% across almost all tested conditions, while the baseline exhibits an EER increase exceeding 10% on both ASVspoof 2019 and Purdue datasets.

In some cases, the proposed framework even yields negative ΔEER values, indicating not just robustness but actual performance improvements under distortion. This may be attributed to the fact that after post-processing, only the clearest phonemes of the input signal are detected, while ambiguous or low-quality ones are filtered out, resulting in cleaner feature representations. Also, we observe increasing performance degradation as the SNR decreases, suggesting that noise injection obscures information useful for distinguishing between real and fake samples. On the other hand, both MP3 compression and μ-law quantization introduce relatively minor performance drops. Nonetheless, the

Table 2. Change in EER (ΔEER, in %) relative to clean test conditions, different post-processing operations are applied. Lower variation indicates greater robustness.

	SNR = 25 dB		SNR = 20 dB		SNR = 15 dB		SNR = 10 dB		MP3 compression		μ-law quantization	
	Baseline	Ours (phoneme)	Baseline	Ours (phoneme)	Baseline	Ours (phoneme)	Baseline	Ours (phoneme)	Baseline	Ours (phoneme)	Baseline	Ours (phoneme)
ASVspoof 2019	13.69	**0.20**	14.66	**0.40**	15.46	**0.68**	12.97	**5.06**	-0.08	**-0.24**	0.08	**-0.56**
In-the-Wild	5.24	**-0.34**	5.34	**0.17**	5.18	**2.81**	4.54	6.77	2.50	**1.10**	0.63	**-0.08**
Purdue	7.95	**0.00**	11.59	**1.32**	10.60	**3.97**	13.91	**5.24**	7.42	**6.74**	4.95	**3.64**
TIMIT-TTS	0.00	**-1.66**	0.67	**-3.52**	3.34	**-4.80**	4.67	**-2.74**	0.91	**-1.04**	1.21	**-0.12**
Average	6.72	**-0.45**	8.07	**-0.41**	8.65	**0.67**	9.02	**3.58**	2.69	**1.64**	1.72	**0.72**

proposed method consistently outperforms the baseline under these conditions as well.

These findings suggest that phoneme-level modeling not only improves efficiency but also gives increased robustness against common distortions encountered in real-world scenarios.

5.3. Interpretability analysis

The final aspect we want to assess in the proposed method is its interpretability. Since our framework performs a fine-grained analysis of the input signal and operates at the phoneme level, it is capable of identifying the specific phonemes that deviate from the speaker's reference profile. This allows for temporal interpretability, as it highlights precisely when and where in the utterance anomalies occur. These insights are particularly valuable in several forensic scenarios, such as courtrooms and legal proceedings, where explainability is essential for justifying detection outcomes. Understanding why a speech segment is classified as fake (e.g., because a specific phoneme is articulated differently from how the speaker typically pronounces it) provides highly actionable forensic evidence.

To investigate this interpretability aspect, we categorized all phonemes into seven groups based on the standard IPA classification and evaluated the detection performance using only one category at a time. The goal of this experiment is to determine which phoneme classes are most challenging for deepfake generators to replicate, and therefore, which contribute most to effective detection. Table 3 summarizes the phoneme categories that we considered.

Table 4 presents the results of this analysis. Vowels and plosives emerge as the most discriminative phoneme categories, consistent with findings from [15], while affricates appear to be the least informative. Also, no single phoneme category achieves better performance than the combined set of all phonemes, indicating the benefit of comprehensive phonetic coverage. An important observation is that the occurrence of phonemes in the analyzed data significantly impacts their utility. For instance, vowels occur far more frequently than diphthongs or approximants, leading to more robust speaker profiles and more reliable classification. In future work, we plan to explore strategies to effectively utilize all phoneme categories, regardless of their frequency in the analyzed speech.

Category	Phonemes
Vowels	/i/, /æ/, /ɑ/, /ɪ/, /ʌ/, /u/, /ɔ/, /ə/, /ɜr/, /ɛ/, /ʊ/
Diphthongs	/aɪ/, /oʊ/, /ɔɪ/, /aʊ/, /eɪ/
Plosives	/k/, /p/, /t/, /b/, /d/, /g/
Fricatives	/ʃ/, /s/, /z/, /θ/, /ð/, /f/, /v/, /ʒ/, /h/
Affricates	/tʃ/, /dʒ/
Approximants	/l/, /j/, /w/, /r/
Nasals	/m/, /n/, /ŋ/

Table 3. Phoneme categories considered in the analysis

Table 4. Detection performance per phoneme category in terms of EER (↓). Best discriminative categories are highlighted in bold green, worst in italic red.

	ASVspoof 2019	In-the-Wild	Purdue	TIMIT-TTS	Average
All phonemes	25.82	27.48	17.55	23.56	23.60
Vowels	32.85	**29.54**	**28.59**	30.22	**30.30**
Diphthongs	38.36	36.99	39.42	34.25	37.26
Plosives	**32.29**	31.60	32.74	**25.80**	30.61
Fricatives	33.47	34.19	36.61	40.51	36.20
Affricates	*40.57*	*41.53*	*46.71*	*43.02*	*42.96*
Approximants	36.05	32.12	33.67	31.48	33.33
Nasals	36.05	35.78	42.97	33.53	37.08

5.4. Case Study: Interpretability in Action

To further illustrate the advantages of the proposed method in terms of interpretability, we conduct a focused case study involving two well-known speakers from the In-the-Wild dataset: Donald Trump and Queen Elizabeth II. For each speaker, we analyze both a real and a synthetic version of the same utterance, applying our phoneme-level framework to compute distances between the test signals and the corresponding speaker-specific profiles. The authentic recordings are sourced from the In-the-Wild dataset, as well as the tracks used to construct the speaker profiles. On the other hand, the synthetic counterparts are generated using the ElevenLabs voice cloning tool to ensure identical linguistic content across the real and fake samples.

The sentence analyzed for Donald Trump is *"Be constantly guarded and protected"* (/bi kɑnstəntli gɑrdɪd ænd prətɛktɪd/), and for Queen Elizabeth II it is *"Germany has reconciled with all her neighbors"* (/dʒɜrməni hæz rkənsaɪld wɪθ ɔl hɜr neɪbɜrz/). Figure 4 shows the phoneme-wise distance scores computed by our method for both the
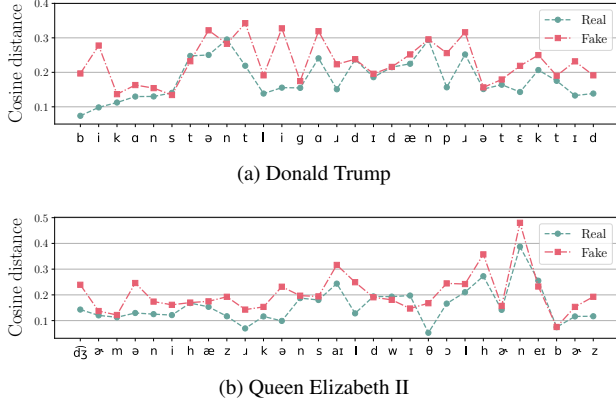
(a) Donald Trump

(b) Queen Elizabeth II

Figure 4. Case study: Phoneme-distance comparison between real and fake speech signals for two speakers.

real and fake utterances, highlighting its capacity for detailed, interpretable analysis.

In general, the synthetic speech samples exhibit higher phoneme-level deviations from the reference profiles compared to authentic speech. Additionally, the method identifies which specific phonemes contribute most to these deviations. In the case of Donald Trump, the phonemes /i/ and /ɑ/ show the largest mismatches, while for Queen Elizabeth II, the highest deviation is observed in the phoneme /ə/.

This case study highlights the practical forensic utility of our approach. By localizing deviations at the phoneme level, the method may not only enable more precise deepfake detection but also provide interpretable evidence that can support human verification or judicial review.

5.5. Discussion and Limitations

The proposed phoneme-level POI-based detection framework offers a novel perspective on speech deepfake detection by combining speaker-specific modeling with fine-grained phonetic analysis and produces promising results.

However, the method also presents some limitations that merit further investigation. First, the framework requires access to a sufficiently large set of reference recordings from the target speaker to build a reliable phoneme-level profile. Ideally, this profile should include consistent coverage of all phoneme classes to ensure robustness during inference.

Second, the current framework depends on accurate phoneme segmentation, which is achieved through an external Automatic Speech Recognition (ASR) system. Errors in alignment can propagate through the pipeline and negatively impact detection performance. As reliance on external components introduces an additional source of uncertainty and complexity, future work should explore ways to reduce or eliminate this dependency, such as through end-to-end alignment-free alternatives.

Related to this, the computational efficiency of the current pipeline is another area for improvement. At present, the method relies on the sequential inference of two distinct models, one for phoneme alignment and one for detection. While this design enables modularity and interpretability, it is not optimal in terms of runtime performance.

Finally, our evaluation was limited to datasets containing only English speech. The extension of our proposed method to other languages, especially those with different phoneme alphabets, may be nontrivial and is a subject of future research.

6. Conclusions

In this paper, we introduced a novel phoneme-level approach to POI-based speech deepfake detection. Our method decomposes speech signals into phonemes and constructs a speaker-specific phoneme profile using only real reference recordings. During inference, each phoneme in a test utterance is independently compared to this profile to determine its authenticity, enabling a fine-grained and interpretable analysis.

This phoneme-centric design offers multiple advantages over traditional POI approaches. First, it enhances interpretability by revealing which specific phonetic elements deviate from the speaker's speech patterns. Second, it improves robustness, as the method focuses on the most relevant segments of the signal and is less prone to performance degradation in the case of post-processing.

Our findings highlight the potential of phoneme-level analysis as a promising direction for future research in speech deepfake detection. Future work could explore integrating phonetic analysis into end-to-end pipelines, further improving the robustness and interpretability of detection systems, as well as extending the framework to multilingual scenarios. Additionally, the development of more sophisticated phoneme-level speaker profiles, potentially incorporating speaker-specific traits such as accent, intonation, or prosodic patterns, could further increase the accuracy and reliability of speaker-aware deepfake detection systems.

Acknowledgment

References

[1] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting world leaders against deep fakes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2019. 1, 2

[2] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020. 2

[3] Irene Amerini, Mauro Barni, Sebastiano Battiato, Paolo Bestagini, Giulia Boato, Vittoria Bruni, Roberto Caldelli, Francesco De Natale, Rocco De Nicola, Luca Guarnera, et al. Deepfake media forensics: Status and future challenges. *Journal of Imaging*, 11(3):73, 2025. 2

[4] Luigi Attorresi, Davide Salvi, Clara Borrelli, Paolo Bestagini, and Stefano Tubaro. Combining Automatic Speaker Verification and Prosody Analysis for Synthetic Speech Detection. In *International Conference on Pattern Recognition (ICPR)*, 2022. 1

[5] 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 (NeurIPS)*, 2020. 4

[6] Emily R Bartusiak, Kratika Bhagtani, Amit Kumar Singh Yadav, and Edward J Delp. Transformer ensemble for synthesized speech detection. In *Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2023. 1

[7] Kratika Bhagtani, Amit Kumar Singh Yadav, Paolo Bestagini, and Edward J Delp. Are Recent Deepfake Speech Generators Detectable? In *ACM Workshop on Information Hiding and Multimedia Security*, 2024. 5

[8] Bookbot. Ljspeech phonemes dataset. https://huggingface.co/datasets/bookbot/ljspeech_phonemes, 2023. Accessed: 2025-02-14. 4

[9] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. 3

[10] Emma Coletta, Davide Salvi, Viola Negroni, Daniele Ugo Leonzio, and Paolo Bestagini. Anomaly detection and localization for speech deepfakes via feature pyramid matching. *arXiv preprint arXiv:2503.18032*, 2025. 2

[11] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[13] Luca Cuccovillo, Christoforos Papastergiopoulos, Anastasios Vafeiadis, Artem Yaroshchuk, Patrick Aichroth, Konstantinos Votis, and Dimitrios Tzovaras. Open challenges in synthetic speech detection. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022. 1

[14] Luca Cuccovillo, Milica Gerhardt, and Patrick Aichroth. Audio transformer for synthetic speech detection via multi-formant analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 1

[15] Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, and Agha Ali Raza. Using self attention DNNs to discover phonemic features for audio deep fake detection. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021. 3, 7

[16] Wanying Ge, Hemlata Tak, Massimiliano Todisco, and Nicholas Evans. Can spoofing countermeasure and speaker verification systems be jointly optimised? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 2

[17] Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. Privacy and security concerns in generative ai: A comprehensive survey. *IEEE Access*, 2024. 1

[18] Keith Ito and Linda Johnson. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 5

[19] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1

[20] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. A survey on speech deepfake detection. *ACM Computing Surveys*, 2025. 2

[21] Jianbo Ma, Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Kong Aik Lee. Speaker-phonetic vector estimation for short duration speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 3

[22] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence*, 53(4): 3974–4026, 2023. 1

[23] Nicolas M Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is silver, silence is golden: What do ASVspoof-trained models really learn? In *Interspeech*, 2021. 6

[24] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? In *Interspeech*, 2022. 5

[25] Parav Nagarsheth, Elie Khoury, Kailash Patil, and Matt Garland. Replay attack detection using dnn for channel discrimination. In *Interspeech*, 2017. 3

[26] Viola Negroni, Davide Salvi, Alessandro Ilic Mezza, Paolo Bestagini, and Stefano Tubaro. Leveraging mixture of experts for improved speech deepfake detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025. 1

[27] Shruti Palaskar and Florian Metze. Acoustic-to-word recognition with sequence-to-sequence models. In *IEEE Spoken Language Technology Workshop (SLT)*, 2018. 3

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015. 5

[29] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Deepfake audio detection by speaker verification. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2022. 1, 2

[30] Alessandro Pianese, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. Training-free deepfake voice recognition by leveraging large-scale pre-trained models. In *ACM Workshop on Information Hiding and Multimedia Security*, 2024. 1, 2, 4, 5

[31] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE Access*, 2024. 1

[32] Davide Salvi, Brian Hosler, Paolo Bestagini, Matthew C Stamm, and Stefano Tubaro. TIMIT-TTS: a Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access*, 2023. 5

[33] Conrad Sanderson. The VidTIMIT database. Technical report, IDIAP, 2002. 5

[34] Petr Schwarz, Pavel Matejka, and Jan Cernocky. Hierarchical structures of neural networks for phoneme recognition. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2006. 3

[35] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Phonetic and speaker variations in automatic emotion classification. In *Interspeech*, 2008. 3

[36] Gajan Suthokumar, Kaavya Sriskandaraja, Vidhyasaharan Sethu, Chamith Wijenayake, and Eliathamby Ambikairajah. Phoneme specific modelling and scoring techniques for anti spoofing system. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 2

[37] 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. In *Interspeech*, 2022. 1

[38] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. ASVspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech*, 2019. 5

[39] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Superseded-CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2016. 5

[40] Luisa Verdoliva. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. 1

[41] Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, page 101825, 2025. 2

[42] Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Jianhua Tao, Xiaopeng Wang, Yuankun Xie, Xin Qi, Shuchen Shi, Yi Lu, Yukun Liu, et al. Mixture of experts fusion for fake audio detection using frozen wav2vec 2.0. *arXiv preprint arXiv:2409.11909*, 2024. 1

[43] Jee weon Jung, Xin Wang, Nicholas Evans, Shinji Watanabe, Hye jin Shim, Hemlata Tak, Siddhant Arora, Junichi Yamagishi, and Joon Son Chung. To what extent can ASV systems naturally defend against spoofing attacks? In *Interspeech*, 2024. 2

[44] Yujie Yang, Haochen Qin, Hang Zhou, Chengcheng Wang, Tianyu Guo, Kai Han, and Yunhe Wang. A robust audio deepfake detection system via multi-view feature. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. 1

[45] Khalid Zaman, Islam JAM Samiul, Melike Sah, Cem Direkoglu, Shogo Okada, and Masashi Unoki. Hybrid transformer architectures with diverse audio features for deepfake speech classification. *IEEE Access*, 2024. 1

[46] Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yushu Zhang, and Yifang Guo. Phoneme-level feature discrepancies: A key to detecting sophisticated speech deepfakes. In *AAAI Conference on Artificial Intelligence*, 2025. 3

[47] Qishan Zhang, Shuangbing Wen, and Tao Hu. Audio deepfake detection with self-supervised xls-r and sls classifier. In *ACM International Conference on Multimedia*, 2024. 1

[48] You Zhang, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021. 1