PhonemeDF: A Synthetic Speech Dataset for Audio Deepfake Detection and Naturalness Evaluation

Vamshi Nallaguntla[1,*], **Aishwarya Fursule**[1], **Shruti Kshirsagar**[1], **Anderson R. Avila**[2,3]

[1] Wichita State University, Wichita, Kansas, USA
[2] Institut national de la recherche scientifique (INRS-EMT), Université du Québec, Canada
[3] INRS-UQO Mixed Research Unit on Cybersecurity, Gatineau, Canada
{vxnallaguntla@shockers., axfursule@shockers., shruti.kshirsagar@}wichita.edu,
Anderson.Avila@inrs.ca

Abstract

The growing sophistication of speech generated by Artificial Intelligence (AI) has introduced new challenges in audio deepfake detection. Text-to-speech (TTS) and voice conversion (VC) technologies can create highly convincing synthetic speech with naturalness and intelligibility. This poses serious threats to voice biometric security and to systems designed to combat the spread of spoken misinformation, where synthetic voices may be used to disseminate false or malicious content. While interest in AI-generated speech has increased, resources for evaluating naturalness at the phoneme level remain limited. In this work, we address this gap by presenting the Phoneme-Level DeepFake dataset (PhonemeDF), comprising parallel real and synthetic speech segmented at the phoneme level. Real speech samples are derived from a subset of LibriSpeech, while synthetic samples are generated using four TTS and three VC systems. For each system, phoneme-aligned TextGrid files are obtained using the Montreal Forced Aligner (MFA). We compute the Kullback–Leibler divergence (KLD) between real and synthetic phoneme distributions to quantify fidelity and establish a ranking based on similarity to natural speech. Our findings show a clear correlation between the KLD of real and synthetic phoneme distributions and the performance of classifiers trained to distinguish them, suggesting that KLD can serve as an indicator of the most discriminative phonemes for deepfake detection.

Keywords: Deepfake detection, Phoneme alignment, LibriSpeech

1. Introduction

Early synthesizers generated speech signals that sounded unnatural and robotic, making them easily recognizable (Nusbaum et al., 1997). This remained the case until recent advancements in generative artificial intelligence (Ren et al., 2019; Prenger et al., 2019), which led to improvements in the quality of synthetic speech. In fact, the growing sophistication of Text-to-Speech (TTS) and Voice Conversion (VC) systems can benefit society in multiple ways. Besides their impact on customer experience, high-quality synthetic speech can help expand accessibility to many people (Naayini et al., 2025). TTS, specifically, enables individuals with visual impairments or reading disabilities to access written content, such as e-books, websites, and documents. VC, on the other hand, improves personalization and communication by adapting voices for specific characters, accents, or age groups in the entertainment domain.

Despite its benefits, the generation of high-quality synthetic speech has introduced new challenges in spoofing detection. Deepfakes, in particular, pose significant challenges not only for voice biometric security applications but also for systems designed to combat the spread of spoken misinformation, where false content is deliberately generated using the voice of a targeted individual for malicious purposes (Yamagishi et al., 2021). In such a context, the ASVspoof challenge has been the main force providing the research community with guidelines and resources to promote the robustness of ASV systems (Wu et al., 2015). More recently, the organizers, realizing the rapid development of deep learning techniques used for voice conversion and speech synthesis, added the audio deepfake detection task to the ASVspoof 2021 challenge. In the following year, the Audio Deep Synthesis Detection (ADD) challenge was created by another group to cover attacks performed in more realistic scenarios, such as those involving background noise, fake clips in real speech signals, and new speech synthesis and voice conversion algorithms (Yi et al., 2022).

The main objective of the participants in such challenges is to develop solutions that can surpass the performance of baseline systems, creating new benchmarks for the task at hand. The focus is placed on developing new approaches for front-end and back-end systems. Besides the numerous contributions throughout the years, such approaches often lack a deep investigation of the phenomenon. In this work, we aim to fill this gap by studying the main acoustic differences between real and synthetic speech. We focus our analysis at the phoneme level to minimize the effects

* Corresponding author.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5905–5915
11–16 May 2026.

of linguistic variability. Our main objective is to investigate speech parameters associated with naturalness (e.g., speech rate, intensity range, articulation rate, average syllable duration) at the phoneme level and to compare the presence of these parameters in real and synthetic speech. Once these parameters are well understood, a model with an internal extractor of these parameters can be developed to compute a new naturalness score. In this work, we propose a comprehensive analysis framework to assess speech naturalness through phoneme-level alignment. We construct a parallel dataset comprising real speech from the LibriSpeech corpus and synthetic speech generated using four text-to-speech (TTS) models: MeloTTS (Zhao et al., 2023), XTTS v2 (Casanova et al., 2024), Chatterbox TTS (Resemble AI, 2025), and VITS TTS (Kim et al., 2021), and three voice conversion (VC) models: Chatterbox VC (Resemble AI, 2025), FreeVC (Li et al., 2023), and StarGAN VC (Li et al., 2021). Our findings highlight specific phonemes and acoustic patterns that contribute most to perceived unnaturalness. Aligned with that, recent phoneme-aware studies (Suthokumar et al., 2019; Dhamyal et al., 2021; Sivaraman et al., 2025) highlight that certain phoneme categories exhibit stronger separability between real and fake speech. Yet, existing datasets offer only utterance- or frame-level labels, lacking aligned phoneme boundaries or timestamps, which makes phoneme-based analysis and reproducibility challenging. To address this gap, we present a new PhonemeDF dataset. In particular, in this paper, the following contributions are made:

1. We introduce PhonemeDF, a new dataset containing synthetic audio and time-aligned phoneme boundaries extracted from TextGrids for both real and synthetic recordings.
2. We develop a phoneme-level detection framework that evaluates discriminability across phonemes using Kullback–Leibler divergence (KLD), and ML-based classifiers such as Logistic Regression (LR) and Support Vector Machine (SVM).
3. Lastly, we conduct a comparative study across handcrafted features and self-supervised learning (SSL) embeddings to assess their robustness and generalization in phoneme-level deepfake detection.

2. Related Work

2.1. Naturalness Assessment Based on Phonemes

The concept of naturalness is not clearly defined in the literature (Nussbaum et al., 2025). Our ability to perceive robotic characteristics in synthetic speech can be quite subjective (Nusbaum et al., 1997), and has decreased in recent years, given the advances in generative AI, which is now capable of generating high-quality speech. Studies addressing speech naturalness date back to the early nineties. In (Nusbaum et al., 1997), for instance, the authors proposed a new methodology for measuring the naturalness of specific aspects of synthesized speech, independent of its intelligibility. While naturalness is a multidimensional and subjective quality of speech, this methodology enables the assessment of the distinct contributions of prosodic, segmental, and source characteristics within an utterance. They conducted experiments showing that glottal source features and prosodic cues (like pitch and rhythm) help listeners distinguish real from synthetic speech. This provides a foundation for more focused evaluations of speech quality. Although a handful of studies have explored speech naturalness (Dall et al., 2014; Sellam et al., 2023; Vojtech et al., 2019), to the best of our knowledge, most of them are not focused on phoneme analysis, which is addressed in our study.

2.2. Deepfake detection datasets

While large-scale deepfake detection datasets such as ASVspoof (Todisco et al., 2019) and MLAAD (Li et al., 2024) provide extensive collections of real and synthetic speech, they lack phoneme-level annotations. Recent efforts have begun to address this limitation. Baser et al. (Baser et al., 2025) introduced PhonemeFake, a dataset focused on segmental deepfake manipulations, and proposed an adaptive bilevel detection model that achieved lower equal error rates (EERs) with 90% faster inference compared to prior baselines. Temmar et al. (Temmar et al., 2025) created a phoneme-annotated dataset. They developed a phoneme-to-word HuBERT-based framework to classify real vs. synthetic speech. Their analysis showed that diphthongs and fricatives exhibit the strongest deviations and provide interpretable phoneme-level cues for detection. Yang et al. (Yang et al., 2025) compared phonetic features with global audio-level representations to evaluate which feature types best discriminate genuine from synthetic speech.

Beyond dataset creation, several studies have explored phoneme-driven detection strategies. For example, Salvi et al. (Salvi et al., 2025) investigated speaker-specific phoneme profiles and evaluated test audio against these profiles to accurately identify synthetic artifacts. For replay attack detection, Suthokumar et al. (Suthokumar et al., 2019) demonstrated that phoneme-based models improve detection accuracy by capturing phoneme-dependent acoustic patterns. Dhamyal et al. (Dhamyal et al., 2021) in turn employed self-

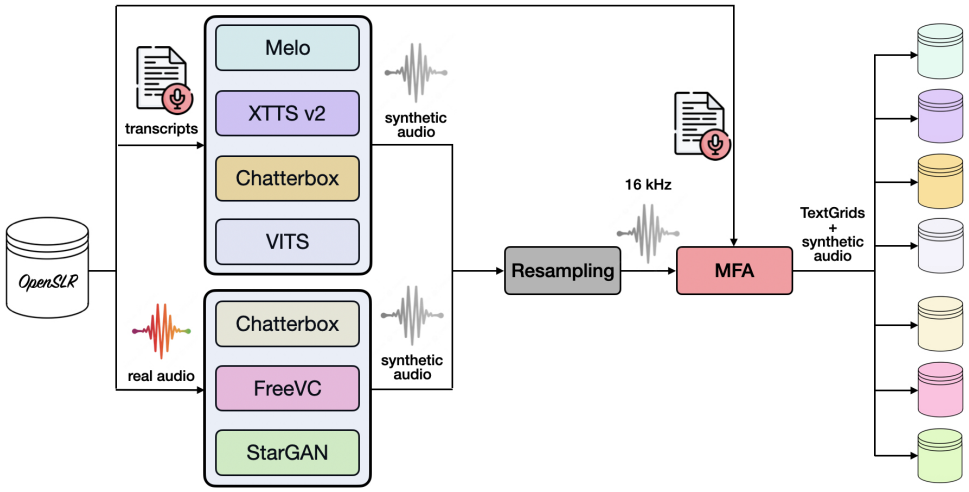


Figure 1: Overview of the dataset creation pipeline.

attention mechanisms to identify which phoneme-level spectral features most effectively distinguish real from synthetic speech. Sivaraman et al. (Sivaraman et al., 2025) adapted the AASIST detector (Jung et al., 2022) by partitioning speech into voiced and unvoiced segments, finding that voiced segments provide stronger discriminative cues. Zhang et al. (Zhang et al., 2025) proposed a phoneme-based detector. They utilized adaptive phoneme pooling and graph attention networks to capture inter-phoneme inconsistencies introduced during synthesis.

Research on discriminative acoustic cues has identified specific phonetic categories and features that reveal synthesis artifacts. The work in (Temmar et al., 2025; Sivaraman et al., 2025), for example, found that some phonemes, such as fricatives, diphthongs, and voiced segments, show more distinct cues from synthetic speech. Furthermore, work in (Zhu et al., 2024) showed that mismatches between linguistic text and speaking style can provide complementary cues for detecting synthetic speech. These results suggest that phoneme-level analysis could enhance detection accuracy and interpretability by providing greater granularity than utterance-level approaches.

3. PhonemeDF Dataset

The goal of PhonemeDF is to provide phoneme segmentation of parallel real and synthetic speech, enabling the investigation of differences between real and AI-generated signals at the phoneme level. We hypothesize that generative models are optimal at generating certain phonemes and less efficient at producing others. This can be used to leverage the optimization of audio deepfake classifiers. The procedure to construct the dataset is divided into two stages. The first stage consists of generating synthetic utterances from seven distinct synthesizers, followed by the segmentation of the speech into phoneme-level segments. These two stages are detailed in the next two sections and an overview of our data construction is shown in Figure 1.

3.1. Synthetic Speech Generation

We adopted four TTS systems, namely MeloTTS (Zhao et al., 2023), XTTS v2 (Casanova et al., 2024), Chatterbox TTS (Resemble AI, 2025), VITS TTS (Kim et al., 2021), and three VC models, referred to as Chatterbox VC (Resemble AI, 2025), FreeVC (Li et al., 2023), and StarGAN VC (Li et al., 2021) to generate synthetic speech. These speech generation systems were selected based on their open availability, coverage of modern paradigms, and diversity of synthesis mechanisms. LibriSpeech (Panayotov et al., 2015) was used as a reference corpus to generate parallel synthetic data.

All LibriSpeech files within the 100-hour subset were converted to WAV format, and their associated text transcripts were separated to ensure consistent filename alignment between text and audio. The TTS models synthesized speech directly from the transcripts, whereas the VC models used the original real audio as input and generated converted versions, as illustrated in Figure 1. When built-in reference speakers were available in a model, we utilized reference speakers directly. Ten speakers (five males and five females) from the VCTK corpus (Veaux et al., 2017) were selected as reference speakers to ensure consistent speaker identity and acoustic diversity across the generated data. As MeloTTS and VITS systems provide built-in voices, all available voices were used directly. For systems without built-in voices, the selected VCTK reference speaker recordings were used to condition the generated speech. We did not attempt to clone the

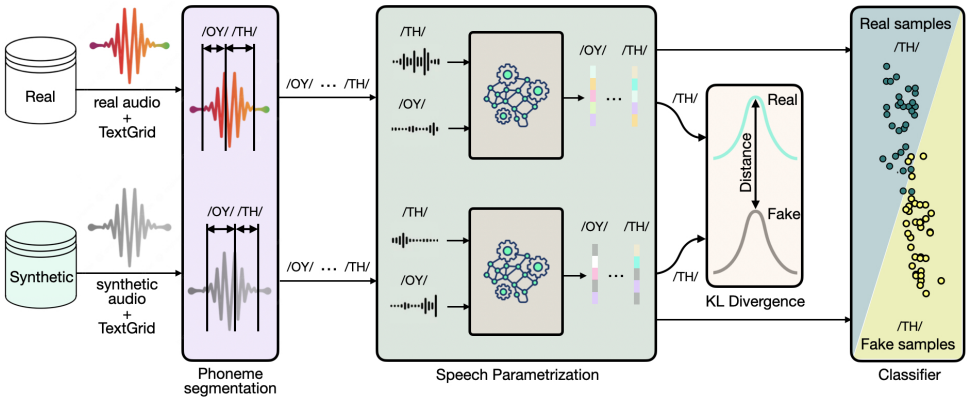


Figure 2: Per-phoneme evaluation pipeline for real vs. synthetic discrimination.

original LibriSpeech speakers. All synthesized files were resampled to 16 kHz to maintain a uniform sampling rate across the systems.

3.2. Phoneme-level Segmentation

We relied on the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to generate phoneme boundaries aligned with transcripts. Alignment was performed using the pretrained American English ARPAbet acoustic model (english_us_arpa) and its pronunciation dictionary. ARPAbet was chosen as it is the representation adopted for LibriSpeech and MFA. Thus, it ensures compatibility with alignment boundaries and the CMU pronunciation lexicon, while avoiding additional phoneme mapping or Grapheme-to-Phoneme (G2P) inference. To reduce sparsity and focus on phoneme-level acoustic characteristics, stress markers were removed (AA0/AA1/AA2 → AA). MFA then applies Viterbi forced alignment to produce TextGrid files containing phoneme labels and precise temporal boundaries. Table 1 summarizes the overall statistics of the PhonemeDF dataset. In total, it comprises approximately 730 hours of speech, equivalent to 199,773 synthetic speech samples, along with their respective TextGrid files, derived from 28,539 (100 hours) real utterances from the LibriSpeech corpus. Selected reference audios and their synthetic counterparts are available at the following link[1]. Note that a handful of TextGrids were inspected to verify alignment quality across both real and synthetic recordings.

4. Experimental Setup

In our experiments, we assess the capability of distinct synthesizers to generate synthetic phonemes. Our hypothesis is that certain phonemes are more challenging to generate, and therefore, classifying those phonemes will lead to higher accuracies compared to phonemes that are easier to generate and therefore more similar to their real counterparts. In the following section, we describe how we use PhonemeDF for the purpose of this study.

[1]https://github.com/Vamshi-Nallaguntla/PhonemeDF

Dataset	# Files	Avg. (s)	Total (h)
LibriSpeech	28,539	12.69	100
MeloTTS	28,539	10.15	80
XTTS v2	28,539	10.47	82
Chatterbox TTS	28,539	10.15	80
VITS TTS	28,539	10.77	85
Chatterbox VC	28,539	12.70	100
FreeVC	28,539	12.68	100
StarGAN VC	28,539	12.68	100
PhonemeDF	199,773	11.37	≈ 730

Table 1: Statistics showing number of files, average duration of an audio file (in seconds), and total duration (in hours) for real and synthetic datasets.

4.1. Phoneme-Level Evaluation Setup

As shown in Figure 2, we conduct our evaluation using parallel samples comprising real and synthetic speech. PhonemeDF includes the full 100 hours of real speech from the LibriSpeech subset together with their corresponding TextGrid alignments. Synthetic versions of these recordings are generated for each synthesis system to form parallel real–synthetic pairs. For computational efficiency, the experiments were conducted on a controlled subset of the dataset. Specifically, we selected 1000 real utterances from the LibriSpeech portion of the corpus and used their corresponding synthetic counterparts from each system. The subset was constructed in a balanced manner to ensure that all 251 speakers from the LibriSpeech subset are represented. Using phoneme boundaries extracted from the TextGrid files, speech signals are segmented into phoneme-level audio units. Figure 3 shows a sample TextGrid with phoneme and word boundaries. Each phoneme

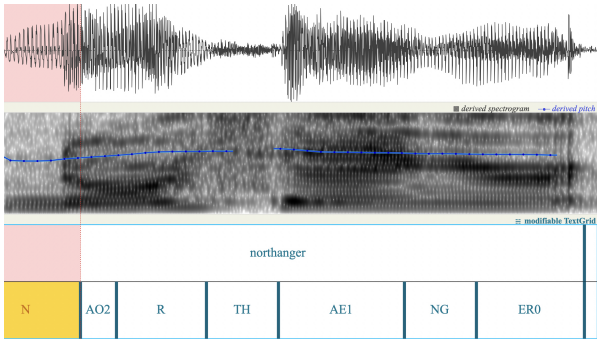


Figure 3: TextGrid showing phoneme boundaries aligned using Montreal Forced Aligner (MFA).

segment is processed separately and undergoes speech parametrization. We adopt both handcrafted and deep learning speech representations with the aim of assessing how synthetic phonemes behave within these categories. Thus, for handcrafted features, we rely on log-Mel Spectrograms (LogSpec) and Linear-Frequency Cepstral Coefficients (LFCC), and for SSL embeddings we considered WavLM (Chen et al., 2022) and wav2vec 2.0 (Baevski et al., 2020).

For the handcrafted features, LogSpec was computed using 80 Mel bins, a 512-point Fast Fourier Transform (FFT), and a 10 ms hop size, while LFCCs used 20 coefficients with identical parameters. Frame-level features were then mean-pooled across each phoneme segment using the boundary timestamps from MFA TextGrids. Similarly, for the SSL representations, phoneme-level embeddings were extracted based on frame-level hidden states, and mean pooling was applied to each phoneme segment. Both models process audio at a 16 kHz sampling rate with a 25 ms window size, with a 20 ms hop length, producing 1024-dimensional embeddings per frame.

4.2. Similarity of Synthetic and Real Phonemes

To measure the similarity between real and synthetic phoneme representations, we compute the KLD between their distributions. For each phoneme in PhonemeDF, the embeddings extracted from real and synthetic speech are modeled as multivariate Gaussian distributions. Let the distributions of real and synthetic phoneme embeddings be denoted as $\mathcal{N}_R(\mu_R, \Sigma_R)$ and $\mathcal{N}_S(\mu_S, \Sigma_S)$, where μ represents the mean vector and Σ the covariance matrix.

The KLD between two multivariate Gaussian distributions is defined as

$$D_{\mathrm{KL}}(\mathcal{N}_R \| \mathcal{N}_S) \quad (1)$$

Since KLD is asymmetric, we compute the symmetric KLD between real and synthetic phoneme distributions:

$$D_{\mathrm{sym}} = \frac{1}{2}\left[D_{\mathrm{KL}}(\mathcal{N}_R \| \mathcal{N}_S) + D_{\mathrm{KL}}(\mathcal{N}_S \| \mathcal{N}_R)\right] \quad (2)$$

Higher D_{sym} values indicate larger distributional differences between real and synthetic phoneme embeddings, while lower values suggest higher similarity between the two distributions. We hypothesize that phonemes with higher divergence values will be easier to distinguish from their synthetic counterparts and, therefore, will yield higher classification accuracies.

4.3. Classification Model

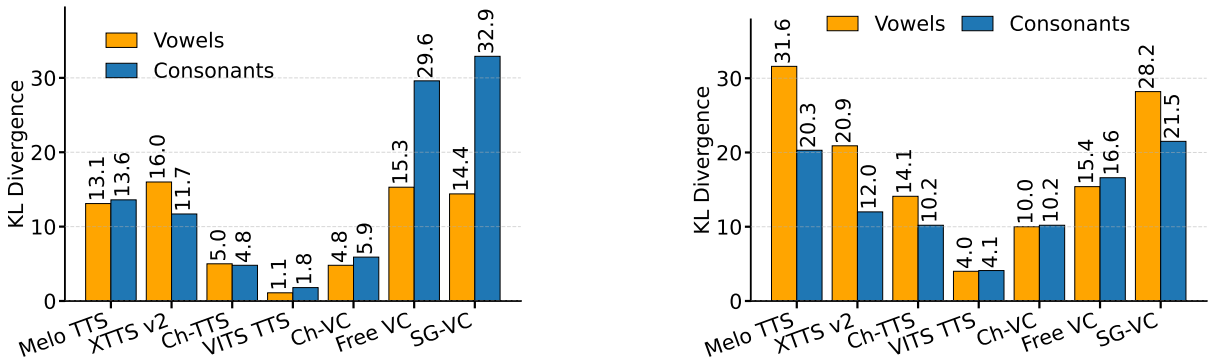
Two binary classifiers are adopted to evaluate the impact of phonemes on deepfake detection. We rely on LR (Hosmer et al., 2013) and SVM (Cortes and Vapnik, 1995). This yields 78 independent classifiers per feature type and synthesis system (39 phonemes $\times$ 2 classifiers). Classification accuracy quantifies the practical detectability of each phoneme type, complementing the statistical separability measured by KLD.

As evaluation metrics, we adopted KLD, accuracy, and Pearson correlation. Pearson correlation is used to assess the relationship between phoneme-level KLD values and classification accuracy across phonemes. The null hypothesis assumes no linear relationship between the two variables, $H_0 : r = 0$, where r denotes the Pearson correlation coefficient.

5. Results and Discussion

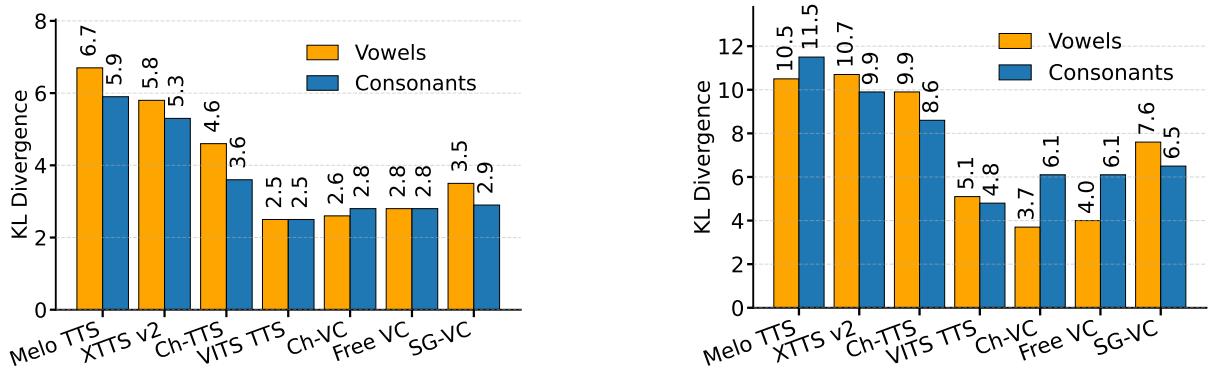
5.1. KLD for Vowels and Consonants

Figure 4 and Figure 5 show the KLD scores averaged across vowels and consonants for handcrafted features and SSL embeddings. The results are presented for each synthesizer. For handcrafted representations, the divergence patterns are mostly influenced by the spectral envelope and energy differences. LFCC and LogSpec representations rank voice-conversion systems as providing the highest mismatch between synthetic and real phonemes. In particular, SG-VC produces large divergences for both vowels and consonants in LFCC, reaching KLD values of 14.4 and 32.9, respectively, while FreeVC also shows a large consonant divergence of 29.6. Similarly, LogSpec reveals substantial mismatches for MeloTTS vowels, 31.6, and SG-VC vowels, 28.2, with consonant divergence remaining high for SG-VC, 21.5, and MeloTTS, 20.3. In contrast, VITS TTS produces synthetic phonemes that are very close to real ones in the handcrafted feature space, with very low KLD



(a) LFCC (b) LogSpec

Figure 4: KLD between synthesized and real speech using handcrafted features.



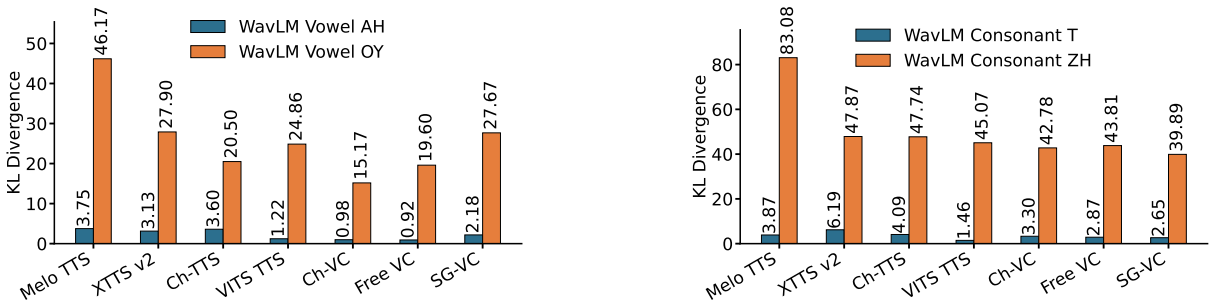
(a) Wav2vec2 (b) WavLM

Figure 5: KLD between synthesized and real speech using self-supervised features.

values, i.e., 1.1 and 1.8 in LFCC, and 4.0 and 4.1 for LogSpec, respectively, for vowels and consonants. Overall, consonants tend to diverge more than vowels for the VC systems (e.g., FreeVC and SG-VC), reflecting the sensitivity of spectral features to transient and fricative energy patterns. For TTS systems, the behavior is more mixed. XTTS v2 and MeloTTS show larger vowel mismatches, whereas VITS exhibits slightly higher divergence for consonants. In general, LogSpec produces higher KLD scores than LFCC, suggesting that the handcrafted feature representation adopted has a significant impact on the way real and synthetic phonemes are represented. The TTS and VC systems play an equally important role.

For the SSL representations, the average KLD scores decrease considerably compared to handcrafted features. Wav2Vec2 produces moderate divergences, with MeloTTS reaching the highest values, 6.7 for vowels and 5.9 for consonants, while VITS remains closest to real speech, i.e., 2.5 for both vowels and consonants. Compared to handcrafted features, WavLM representations produce lower KLD values overall, with most systems falling between 3 and 11. In this representation, Ch-VC and FreeVC produce relatively small vowel divergences, 3.7 and 4.0, but slightly larger consonant divergences, around 6.1. These results suggest that SSL embeddings capture higher-level phonetic structures and are less sensitive to raw spectral artifacts compared to handcrafted features.

Figure 6 presents examples of individual phonemes illustrating the lowest and highest divergences observed in the WavLM representation. For vowels, the monophthong AH consistently shows very small divergence across all systems, with values between 0.92 and 3.75, whereas the diphthong OY exhibits substantially larger divergence, reaching 46.17 for MeloTTS and remaining above 15 for all systems. This indicates that dynamic vowels with complex articulatory transitions are more difficult for synthesis models to reproduce accurately. A similar pattern is observed for consonants. The stop consonant /T/ shows relatively small divergence across systems, ranging from 1.46 to 6.19, indicating that simple closure–release patterns are reproduced reliably. In contrast, the fricative /ZH/ exhibits very large divergence, reaching 83.08 for MeloTTS and remaining above 39 for all systems, suggesting that high-frequency fricative noise is difficult for both TTS and VC models to replicate. Overall, these results highlight phoneme-level vari-



(a) WavLM (Vowel)

(b) WavLM (Consonant)

Figure 6: Lowest and highest KLD scores for (a) WavLM vowel representation and (b) WavLM consonant representation.

ability in synthetic speech. While phonemes such as /AH/ and /T/ are reproduced with high fidelity, others such as /OY/ and /ZH/ show substantial divergence, revealing limitations of current synthesizers.

5.2. Correlation Between KLD and ACC

In this experiment, we present the correlations between KLD scores and the accuracies for audio deepfake detection under different configurations. Within each table, results are presented by synthesizer system, type of classifier, speech representation, and the category of phonemes: vowels or consonants. In Table 2, we present performance for vowels with LFCC based on the LR classifier, providing moderate results for all systems, including Ch-VC, SG-VC, and Ch-TTS, where correlations range from 0.68, 0.74, and 0.79, respectively. Higher correlations are attained for consonants, as shown in Table 3, suggesting that for the LFCC representation, classifiers should rely more on consonants to distinguish between real and synthetic audio. LogSpec features show a similar pattern, with lower correlations for specific synthesizers. Table 6 provides the performance for vowels, with the highest correlations ranging from 0.84, 0.87, and 0.89, respectively, for Free VC, Melo TTS, and SG-VC. These results are based on LR, but a similar trend is found for the SVM. Low performance, however, was observed with the VITS synthesizer, with correlations as low as 0.07 and 0.04 for LR and SVM for the experiment with vowels, and a similar trend was found for consonants. These results suggest that for handcrafted features, larger KLD leads to higher classification accuracies and, therefore, phonemes that deviate more from real speech yield better detection rates.

For SSL representations, correlations are generally lower and more inconsistent across systems. Tables 4 and 5 present the results using Wav2Vec2 representations. In contrast to the handcrafted features, correlations vary substantially across synthesizers and phoneme categories. For vowels, in Table 4, several systems exhibit weak or even negative correlations, such as XTTS v2 and Ch-TTS, while VITS TTS shows a relatively high positive correlation (i.e., 0.75 for both LR and SVM). Other systems, including Ch-VC and Free VC, display moderate negative correlations, indicating that larger KLD does not necessarily translate into better classification performance when using this representation. For consonants, in Table 5, correlations tend to be negative for most systems, particularly for Melo TTS, VITS TTS, and SG-VC, suggesting that higher divergence is often associated with lower detection accuracy. A similar pattern is observed with WavLM representations, as shown in Tables 8 and 9. For vowels (see Table 8), correlations are generally weak and fluctuate around zero for most synthesizers, with only Ch-VC and Free VC showing moderate negative correlations. For consonants (Table 9), however, consistently strong negative correlations are observed across nearly all systems, with coefficients reaching as low as -0.83 for SG-VC and -0.79 for Melo TTS using LR. This indicates that, unlike handcrafted features, larger KLD in SSL representations often corresponds to lower classification accuracy. Overall, these results suggest that the relationship between phoneme-level divergence and detection performance differs substantially between handcrafted and SSL representations, with the latter exhibiting weaker and frequently inverse correlations.

5.3. Phoneme Rankings Across Systems

We identify the most discriminative phoneme categories by analyzing per-phoneme KLD, LR accuracy, and SVM accuracy across all feature representations. Across systems and features, consonants generally exhibit stronger discriminability than vowels. Handcrafted spectral representations highlight this difference clearly: LogSpec produces the highest divergence values (average KLD of 31.6 for vowels and 20.3 for consonants), followed by LFCC (13.1 for vowels and 13.6 for consonants),

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	0.68	4.99E-03	0.63	1.16E-02
XTTS v2	0.67	6.11E-03	0.69	4.07E-03
Ch-TTS	0.79	4.94E-04	0.80	3.89E-04
VITS TTS	0.49	0.065569	0.51	0.054600
Ch-VC	0.68	5.38E-03	0.69	4.05E-03
Free VC	0.71	3.06E-03	0.71	2.89E-03
SG-VC	0.74	1.73E-03	0.73	1.97E-03

Table 2: Results for vowels based on LFCC representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	0.79	4.28E-06	0.78	5.99E-06
XTTS v2	0.76	1.47E-05	0.76	1.69E-05
Ch-TTS	0.85	1.11E-07	0.85	1.38E-07
VITS TTS	0.65	5.79E-04	0.66	5.08E-04
Ch-VC	0.88	1.22E-08	0.84	2.92E-07
Free VC	0.65	6.37E-04	0.66	5.12E-04
SG-VC	0.82	1.16E-06	0.82	1.21E-06

Table 3: Results for consonants based on LFCC representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	0.29	0.298568	0.27	0.338104
XTTS v2	-0.43	0.109768	-0.39	0.155521
Ch-TTS	-0.36	0.187258	-0.50	0.059611
VITS TTS	0.75	1.30E-03	0.75	1.39E-03
Ch-VC	-0.60	0.017691	-0.61	0.016767
Free VC	-0.73	2.13E-03	-0.79	5.05E-04
SG-VC	0.33	0.227209	0.21	0.456768

Table 4: Results for vowels based on Wav2vec 2.0 representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	-0.50	0.013894	-0.59	2.19E-03
XTTS v2	0.31	0.134920	0.02	0.908695
Ch-TTS	-0.43	0.036489	-0.25	0.232014
VITS TTS	-0.56	4.25E-03	-0.16	0.443928
Ch-VC	-0.38	0.069623	-0.35	0.090622
Free VC	-0.29	0.173695	-0.01	0.963936
SG-VC	-0.59	2.61E-03	-0.43	0.038125

Table 5: Results for consonants based on Wav2vec 2.0 representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	0.87	2.48E-05	0.81	2.41E-04
XTTS v2	0.70	3.55E-03	0.76	9.30E-04
Ch-TTS	0.65	9.37E-03	0.74	1.59E-03
VITS TTS	0.07	0.803638	0.04	0.898342
Ch-VC	0.52	0.046131	0.61	0.014760
Free VC	0.84	7.92E-05	0.88	1.69E-05
SG-VC	0.89	8.82E-06	0.91	3.19E-06

Table 6: Results for vowels based on LogSpec representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	0.41	0.046255	0.56	4.82E-03
XTTS v2	0.62	1.29E-03	0.55	4.94E-03
Ch-TTS	0.57	3.70E-03	0.54	6.44E-03
VITS TTS	0.00	0.997485	-0.17	0.428621
Ch-VC	0.74	3.57E-05	0.69	1.86E-04
Free VC	0.61	1.72E-03	0.63	9.98E-04
SG-VC	0.63	8.65E-04	0.71	1.04E-04

Table 7: Results for consonants based on LogSpec representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	-0.41	0.132823	-0.31	0.264627
XTTS v2	0.18	0.527040	0.17	0.537221
Ch-TTS	-0.04	0.882777	0.01	0.965355
VITS TTS	0.40	0.145025	0.47	0.080362
Ch-VC	-0.65	8.13E-03	-0.67	6.15E-03
Free VC	-0.78	6.69E-04	-0.78	6.43E-04
SG-VC	-0.10	0.718714	-0.21	0.453530

Table 8: Results for vowels based on WavLM representation.

	LR		SVM	
System	r	p-value	r	p-value
Melo TTS	-0.79	4.18E-06	-0.74	3.43E-05
XTTS v2	-0.60	1.93E-03	-0.60	1.89E-03
Ch-TTS	-0.56	4.76E-03	-0.61	1.44E-03
VITS TTS	-0.78	7.42E-06	-0.73	5.40E-05
Ch-VC	-0.58	3.19E-03	-0.53	8.13E-03
Free VC	-0.33	0.115046	-0.44	0.032071
SG-VC	-0.80	2.82E-06	-0.75	2.55E-05

Table 9: Results for consonants based on WavLM representation.

while SSL embeddings yield lower divergence (e.g., WavLM averages of 10.5 for vowels and 11.5 for consonants). These results indicate that handcrafted spectral features emphasize acoustic mismatches between real and synthetic speech more strongly, whereas SSL embeddings capture more subtle phonetic variations. Among vowels, diphthongs consistently appear as the most discriminative category across feature types. In particular, /OY/ and /EY/ frequently exhibit the highest divergence values across synthesis systems, followed by /AW/ and /AY/. These phonemes involve dynamic formant trajectories that are difficult for generative models to reproduce accurately. In contrast, simpler monophthongs such as /AH/ and /UH/ consistently show lower divergence values, indicating that they are easier for synthesis systems to approximate. For consonants, fricatives and plosives dominate the top ranks. Fricatives such as /SH/, /S/, and /ZH/ exhibit large divergence values in spectral features due to their broadband noise characteristics, while plosives including /P/, /B/, and /T/ frequently appear among the most discriminative phonemes in SSL embeddings, reflecting the importance of transient temporal cues captured by these models. When considering vowels and consonants jointly, three phoneme groups consistently emerge as discriminative across systems and feature types: diphthongs (e.g., /OY/), fricatives (e.g., /SH/ and /S/), and plosives (e.g., /P/ and /B/). These categories capture different synthesis artifacts, including complex formant transitions, sustained spectral turbulence, and rapid articulatory bursts.

Our findings align with and extend recent phoneme-level deepfake analyses. Temmar et al. (Temmar et al., 2025) also showed that diphthongs and fricatives are highly discriminative, consistent with our observation that /OY/, /SH/, and /F/ frequently appear among the most informative phonemes. However, we extend prior work by evaluating 39 phoneme categories across seven synthesis systems and four feature representations. Different feature types reveal complementary phonetic sensitivities. SSL embeddings (WavLM and wav2vec 2.0) emphasize transient consonants such as plosives, while handcrafted spectral features (LogSpec and LFCC) highlight fricatives and diphthongs that manifest as spectral irregularities. At the system level, StarGAN-VC and MeloTTS consistently exhibit the largest divergence from real speech, while VITS remains closest to natural speech. Overall, handcrafted features highlight broader spectral mismatches, whereas SSL embeddings reveal finer phonetic inconsistencies, providing complementary insights into synthesis realism and phoneme-level variability.

6. Conclusion

This work introduced PhonemeDF, a dataset for audio deepfake detection with phoneme-level annotations. It comprises nearly 200k synthetic utterances and about two million aligned phoneme segments generated using seven TTS and VC systems. We used the data to analyze the discriminability of phonemes through KLD and supervised classification using both handcrafted spectral features and SSL speech representations. Our results show that certain phoneme categories—particularly diphthongs, fricatives, and plosives—consistently provide strong cues for distinguishing synthetic from real speech. Handcrafted spectral representations emphasize large acoustic mismatches between real and synthetic speech, while SSL embeddings capture more subtle phonetic inconsistencies. Additionally, we observe systematic differences across synthesis models, with voice conversion systems generally producing larger phoneme-level deviations from natural speech than modern TTS models. Overall, our findings highlight the value of phoneme-level analysis for understanding synthesis artifacts and suggest that combining complementary feature representations may improve the robustness of future deepfake detection systems.

7. Ethics Statement and Limitations

While this work aims to improve the detection of synthetic speech, the dataset and analysis may also indirectly facilitate the study of synthesis artifacts that could be exploited to improve generation systems. The dataset is restricted to English speech and a limited set of synthesis models, which may limit the generalization of the findings to other languages or emerging speech generation technologies. Furthermore, the reference speech is derived from a specific corpus and recording condition, which may introduce biases in speaker characteristics, recording environments, and speaking styles. In addition, phoneme boundaries are obtained using forced alignment, which may introduce small segmentation errors that affect phoneme-level analysis. Our experiments rely on a limited set of feature representations and relatively simple classifiers, and the evaluation focuses primarily on statistical divergence and classification accuracy without extensive perceptual validation. Future work will expand the dataset to additional languages and synthesis systems and incorporate perceptual studies to better relate phoneme-level differences to human judgments. We will also investigate phoneme transitions, as artifacts at phoneme boundaries may help detect partially manipulated synthetic speech.

8. Bibliographical References

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

Oguzhan Baser, Ahmet Ege Tanriverdi, Sriram Vishwanath, and Sandeep P. Chinchali. 2025. Phonemefake: Redefining deepfake realism with language-driven segmental manipulation and adaptive bilevel detection. *arXiv preprint arXiv:2506.22783*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: A massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

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

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Rasmus Dall, Junichi Yamagishi, and Simon King. 2014. Rating naturalness in speech synthesis: The effect of style and expectation. In *Speech Prosody 2014*.

Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, and Agha Ali Raza. 2021. Using self attention dnns to discover phonemic features for audio deep fake detection. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1178–1184. IEEE.

David W. Hosmer, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression*, 3rd edition. John Wiley & Sons, Hoboken, NJ.

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

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Jingyi Li, Weiping Tu, and Li Xiao. 2023. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. 2024. Audio anti-spoofing detection: A survey. *arXiv preprint arXiv:2404.13914*.

Yinghao Aaron Li, Ali Zare, and Nima Mesgarani. 2021. Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. *arXiv preprint arXiv:2107.10394*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

Prudhvi Naayini, Praveen Kumar Myakala, Chiranjeevi Bura, Anil Kumar Jonnalagadda, and Srikanth Kamatala. 2025. Ai-powered assistive technologies for visual impairment. *arXiv preprint arXiv:2503.15494*.

Howard C. Nusbaum, Alexander L. Francis, and Anne S. Henly. 1997. Measuring the naturalness of synthetic speech. *International Journal of Speech Technology*, 2:7–19.

Christine Nussbaum, Sascha Frühholz, and Stefan R. Schweinberger. 2025. Understanding voice naturalness. *Trends in Cognitive Sciences*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32.

Resemble AI. 2025. Chatterbox-TTS. GitHub repository. https://github.com/resemble-ai/chatterbox.

Davide Salvi, Viola Negroni, Sara Mandelli, Paolo Bestagini, and Stefano Tubaro. 2025. Phoneme-level analysis for person-of-interest speech deepfake detection. *arXiv preprint arXiv:2507.08626*.

Thibault Sellam, Ankur Bapna, Joshua Camp, Diana Mackinnon, Ankur P. Parikh, and Jason Riesa. 2023. Squid: Measuring speech naturalness in many languages. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ganesh Sivaraman, Hemlata Tak, and Elie Khoury. 2025. Investigating voiced and unvoiced regions of speech for audio deepfake detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Gajan Suthokumar, Kaavya Sriskandaraja, Vidhyasaharan Sethu, Chamith Wijenayake, and Eliathamby Ambikairajah. 2019. Phoneme specific modelling and scoring techniques for anti spoofing system. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6106–6110. IEEE.

Dia Elhak Temmar, Assia Hamadene, Vamshi Nallaguntla, Aishwarya Fursule, Mohand Saïd Allili, Shruti Kshirsagar, and Anderson R. Avila. 2025. Phonetic analysis of real and synthetic speech using hubert embeddings: Perspectives for deepfake detection. In *2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 86–91. IEEE.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. 2019. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2017. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. Available at: http://dx.doi.org/10.7488/ds/1994.

Jennifer M. Vojtech, Jacob P. Noordzij Jr, Gabriel J. Cler, and Cara E. Stepp. 2019. The effects of modulating fundamental frequency and speech rate on the intelligibility, communication efficiency, and perceived naturalness of synthetic speech. *American Journal of Speech-Language Pathology*, 28(2S):875–886.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. 2015. Asvspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015, Automatic Speaker Verification Spoofing and Countermeasures Challenge, colocated with INTERSPEECH 2015*, pages 2037–2041. ISCA.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. Asvspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

Tianle Yang, Chengzhe Sun, Siwei Lyu, and Phil Rose. 2025. Forensic deepfake audio detection using segmental speech features. *Forensic Science International*, page 112768.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. Add 2022: The first audio deep synthesis detection challenge. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE.

Kuiyuan Zhang, Zhongyun Hua, Rushi Lan, Yushu Zhang, and Yifang Guo. 2025. Phoneme-level feature discrepancies: A key to detecting sophisticated speech deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1066–1074.

Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. Melotts: High-quality multi-lingual multi-accent text-to-speech. GitHub repository. https://github.com/myshell-ai/MeloTTS.

Yi Zhu, Surya Koppisetti, Trang Tran, and Gaurav Bharaj. 2024. Slim: Style-linguistics mismatch model for generalized audio deepfake detection. *Advances in Neural Information Processing Systems*, 37:67901–67928.