MIND YOUR [m]S, CROSS YOUR [t]S: A LARGE-SCALE PHONETIC ANALYSIS OF SPEECH REPRODUCTION IN MODERN SPEECH GENERATORS

Boo Fullwood, Fabian Monrose

Georgia Institute of Technology
ECE & School of Cybersecurity and Privacy, USA

ABSTRACT

Despite rapid advances in synthetic speech generation, the failure modes of modern high-quality generators remain poorly characterized. Our study of phonetic error (i.e., the failure of generators to accurately reproduce specific characteristics of certain types of speech) offers a window into architectural limitations. Specifically, empirical evaluations on 23 contemporary generator configurations on single-speaker and multi-speaker datasets (110 speakers), reveal that all models reproduce pitch accurately but struggle with the spectral features of certain phone classes, notably nasals (like [m] in summer) and obstruents (like [t] in night). Architectural analysis shows that voice-cloning models consistently surpass text-to-speech (TTS) in phonetic accuracy, TTS errors stem primarily from the text-to-spectrogram stage, and vocoder training contributes little to differences in phone reproduction. Additionally, correlating phonetic error with model attention in modern detectors reveals that even state-of-the-art spectrogram-based detectors fail to exploit all relevant phone categories.

Index Terms— Deepfake Detection, Phonetic Analysis

1. INTRODUCTION

The proliferation of speech synthesis technology has led to an unprecedented diversity in speech synthesis techniques and a rapid rise in the quantity of synthetic audio. Detection systems have likewise advanced, drawing on high-level audio representations to achieve strong accuracy. Yet such detectors often reveal little about which speech characteristics generators reproduce poorly. Indeed, prior research [1, 2] on defeating modern detectors showed that many detectors rely on non-speech artifacts of the generation process—features that separate real and synthetic speech well in the lab but degrade under common manipulations like transcoding.

To remain robust, detectors must instead target speech characteristics essential for intelligibility. However, efforts to pinpoint where generators diverge from natural speech remain limited, constraining opportunities to guide detector retraining despite evidence of large potential gains [3, 4]. Recently, phonetic analysis was proposed [5] as a way to quantify the divergence of synthetic speech from natural speech by measuring the distribution of selected audio features across phone classes and comparing distributions between the synthetic speech and the baseline corpus. Distributional shift in these audio features may not always manifest as human-perceptible error, but expose fundamental errors in speech production.

Unfortunately, previous evaluations [5, 6] of phonetic error examined only small groups of generators and the characterization of outdated architectures. We provide the first large-scale phonetic analysis of modern speech generators, showing that earlier problem classes (fricatives, stops) persist, uncovering new ones (nasals, select vowels), and linking error patterns to key architectural factors. We further find that current detectors exploit only a subset of these phones, highlighting significant untapped potential for improved detection.

2. PHONETIC ANALYSIS

A central challenge in large-scale synthetic speech analysis is comparing samples with different underlying texts. Earlier work paired real and generated utterances with identical scripts, making error detection easy for training but revealing little about whether the generated phones exhibit the natural feature distributions of human speech. Phonetic analysis overcomes this by decomposing audio into phones [7], ensuring each unit shares a common vocal production mechanism and can be meaningfully compared [8]. The process begins by identifying and time-aligning phones within an utterance. In this work, we adopt the Montreal Forced Aligner (MFA) for time-aligning phones because of its strong performance against competing forced-alignment systems [9, 10]. MFA outputs a sequence of phones with precise start–end timestamps, including silence and non-speech regions. In our use, both overall alignment likelihood and phone duration deviation were consistent between real and synthetic speech.

Importantly, a synthetic phone need not be identical to a natural one to sound correct, but its features should fall within the natural phone's feature distribution. To build these distributions, we collect large sets of genuine and synthetic speech, identify phone boundaries via forced alignment, and compute multiple speech features. Each feature value is assigned to its corresponding phone and aggregated to form distributions for real speech and for each generator. Comparing these distributions reveals how well generators match natural phones and exposes shared failure modes across generator families. The chosen features are detailed in Section 2.1.

2.1. Feature Selection

Guided by prior work [5, 6], we select a broad set of features to isolate key characteristics of each phone class. These features (given in Table 1) ensure comprehensive coverage across classes. While measures requiring explicit frequency components ($F0$ and R_1, R_2) are ill-defined for unvoiced phones lacking glottal activation, the energy distribution in their noise-like spectra still reflects articulatory and airflow properties [8]. The *Spectral* features we use capture most point attributes of speech signals and emphasize traits often associated with natural-sounding speech.

Feature	Category	Target Phones
RMS Energy [5, 6]	General	All
$F0$ [5, 6]	Pitch	Voiced
R_1, R_2 [5, 6, 3]	Pitch	Voiced
Zero-Crossing Rate	General	All
Spectral Centroid [5, 6]	Spectral	All
Spectral Bandwidth [5, 6]	Spectral	All
Spectral Tilt	Spectral	All
Harmonicity-to-Noise	Spectral	All

Table 1: Phonetic Features included in analysis

2.2. Statistical Analysis

To measure the divergence between two phonetic distributions, we use a two-part statistical analysis. First, a two-sample Kolmogorov-Smirnov (KS) test determines whether there is *any* significant distributional shift between the two distributions. We additionally apply a Bonferroni multiple-test correction, as the large number of phones and features in a single cross-comparison significantly increases the risk of false null hypothesis rejection. We target a rejection threshold of $\alpha = 0.01$ after correction. For distributions that are significantly diverged by the above metric, we convert each distribution to a probability distribution function and compute the Wasserstein distance [11]. For distributions that were not statistically dissimilar, we report a distance of 0. This two-step approach avoids issues where insignificant differences between distributions may be exaggerated if other differences in the same comparison group are also numerically small.

3. DATASET GENERATION

We generate two groups of synthetic speech for this analysis: a single-speaker group to maximize the visibility of detector variation and a multi-speaker group to better represent the true distribution of English speech. Details are given in Table 2.

The single-speaker group is drawn from the LJSpeech corpus, which consists of a single female speaker reading nonfiction. We select 16 architectures, split into three subgroups to capture modern designs and isolate architectural components for later evaluation. 1000 samples are generated per architecture. *Subgroup A* evaluates 7 state-of-the-art vocoders [12, 13, 14] that synthesize speech directly from target Mel-spectrograms, avoiding confounds from TTS design. Because the vocoder performs the final waveform synthesis, its role is important [15]. From *subgroup A*, we select HiFiGAN [14] as the most broadly represented in modern work as the baseline vocoder used in the other subgroups. *Subgroup B* consists of 6 TTS systems, each of which generates Mel-spectrograms of the target speech, which are then vocoded by HiFiGAN. This separation allows us to compare the impact of TTS choice independent of vocoder differences. Finally, *subgroup C* uses the VITS[16] TTS from *subgroup B* and HiFiGAN from *subgroup A*, varying only the training corpus to distinguish architectural from training effects.

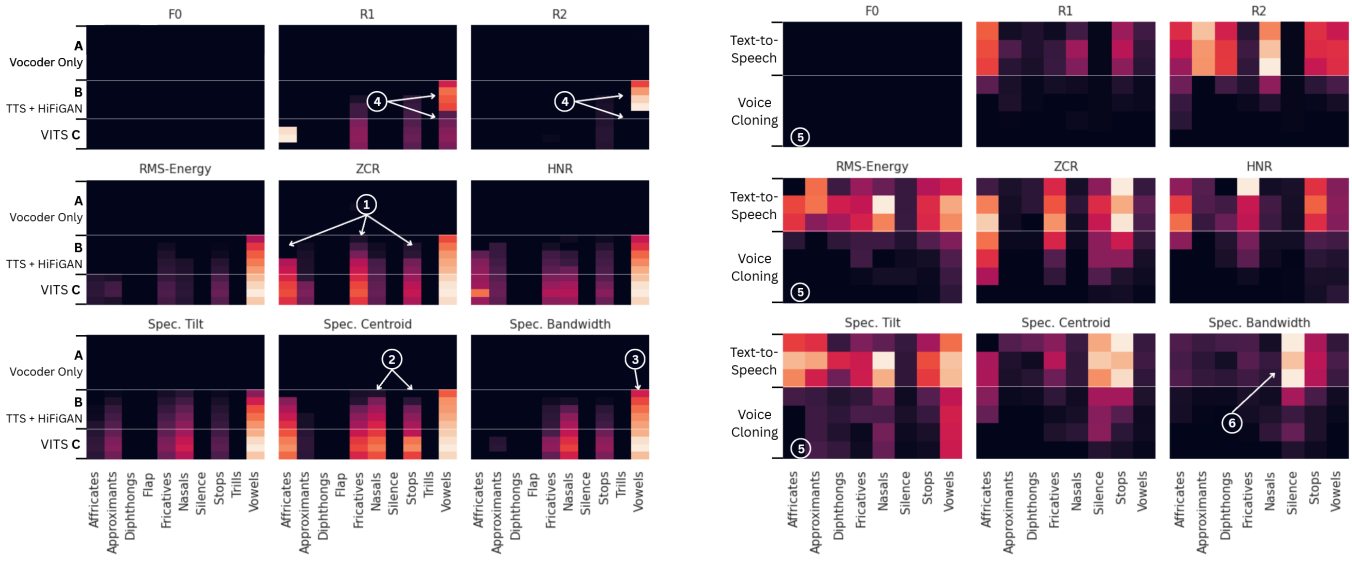
Group	Subgroup	Models
Single-Speaker	Vocoder Evaluation	MelGAN, MelGAN Large, FB MelGAN, MB MelGAN, WaveGlow, Parallel WaveGAN, HiFiGAN
Single-Speaker	TTS Evaluation	Tacotron2, NeuralHMM, Overflow, SpeedySpeech, VITSNeon
Single-Speaker	Training Evaluation	VITS-LJSpeech, VITS-Blizzard, VITS-SAM, VITS-VCTK
Multi-Speaker	TTS	YourTTS, XTTSv1.1, XTTSv2
Multi-Speaker	Voice Cloning	Openvoice V1, Openvoice V2, kNNVC, FreeVC

Table 2: Selected architectures by analysis groups

While the single-speaker group is comprehensive in its coverage of architectural variance, it is limited in its real-world application by the single-speaker nature of the source corpus. Therefore, we generate a multispeaker group based on the VCTK [17] corpus, which includes data from 110 English speakers. Given that the models used above would require retraining for each speaker, we opt to use models specific to multispeaker generation. We select three end-to-end TTS models [18, 19] (combined TTS and vocoder pipelines) and three voice cloning models [20, 18, 21], including multiple versions of models where available. Since the voice cloning models require an input voice, we randomly select speaker p311 in VCTK as the reference speaker. Audio samples from this speaker are used to prompt the voice cloning models, and transcripts of the original p311 samples are used as input for the TTS models.

4. RESULTS

In the single-speaker group (Fig. 1a), the difference in performance between Vocoder Only (*subgroup A*) systems and TTS systems (*subgroup B, C*) is striking. *Subgroup A* vocoders, fed Mel-spectrograms directly from source speech, showed minimal and low-magnitude distributional errors. In contrast, TTS systems using the same vocoder exhibited substantial phone- and feature-level errors, indicating that most deficiencies stem



(a) Single-speaker

(b) Multi-speaker

Fig. 1: Distributional differences between single generator samples and natural speech, for single-speaker and multi-speaker sets. Brighter colors indicate greater Wasserstein distance for that model-feature-phone class distribution.

from Mel-spectrogram generation during the TTS stage rather than from vocoding.

Obstruent phones ① dominated *subgroup B, C* errors, comprising 3 of the 5 most affected classes. Nasals were misproduced at rates comparable to stops ②, a pattern absent in prior analyses [5] but consistent with their shared articulatory traits, differing mainly in whether airflow is blocked orally or nasally. Vowels ③ also show high distributional shifts, likely reflecting the LJSpeech speaker's consistent vowel production.

The impact of vocoder retraining (*subgroup C*) is limited, with mostly consistent performance across all instances of the VITS-HiFiGAN architecture, indicating that the observed failures are primarily architecture-based rather than training dependent. Notably, however, the end-to-end training used by VITS largely reduced or eliminated R_1, R_2 dispersion error ④ compared to other *subgroup B* architectures.

> *Takeaway*: Most phonetic errors originate in the text-to-speech (TTS) stage that converts text to Mel-spectrograms. This stage consistently misrenders obstruent phones and also struggles with nasals and tightly constrained vowels.

In the multi-speaker group (Fig. 1b), the VC architectures achieve better reproduction of our selected features across all phone classes. In particular, FreeVC ⑤ achieves near-zero error across all features except *spectral tilt*, where it is on par with the other VC models. This is a noticeable improvement over the original VITS architecture it is derived from. Distributional errors for the end-to-end TTS models are more broadly distributed across phone classes compared to the single-speaker *subgroup B, C* models for most features, although ZCR and HNR errors are still strongly associated with obstruents. The TTS models also showed increased error in non-speech regions (silence) ⑥ associated with poor reproduction of the background noise floor or speaker breaths. These errors are notably absent from the single-speaker group.

Contrary to previous work [5], all models in both groups accurately reproduced *F0* characteristics for all classes, and error in formant dispersion (R_0, R_1) was limited. This reflects the improved consistency of modern architectures, highlighting the increasing difficulty they pose for deepfake detectors.

> *Takeaway*: Modern end-to-end TTS systems still reproduce obstruent phones poorly, but pitch-related errors have largely been eliminated, and voice-cloning (VC) models perform exceptionally well across all features and phone classes.

5. ALIGNMENT OF DISTRIBUTIONAL FEATURES WITH DETECTOR ATTENTION

These results naturally raise the question of whether modern detectors exploit these phone-level distributional differences or rely on other cues. To assess that, we train an Audio Spectrogram Transformer (AST) for audio deepfake detection on each of our groups. We choose AST because it delivers state-of-the-art accuracy on common detection tasks [22, 23] and offers strong interpretability [24]. Because AST uses a spectrogram representation of the speech signal as its input, it is well-suited to capture the phone-related features analyzed here.

To determine what phones are important to detector classification, we perform multiscale occlusion. We begin with

an unaltered spectrogram input and compute the detector's corresponding classification confidence for the true class of the target sample. We then systematically remove information from the spectrogram, setting progressively larger regions of the input to a baseline value. The occluded regions span the entire frequency axis of the spectrogram. If the affected region is important to classification, we expect to see a decrease in the detector's confidence. This change in confidence can then be mapped back onto the original input, allowing for the localization of important regions. These regions can then be mapped back to the underlying phones in the same way as the features used in the previous analysis, allowing us to compute an 'average importance' for each phone. Given that we expect larger changes in confidence when larger regions are masked, we scale the change in attention by the width of the occluded area before aggregating regions for each sample.

The output of this occlusion process is a single 'attention' value for each column of the input spectrogram, which represents the average loss of confidence when the information in that column is removed. The attention for a given phone can then be computed as the average attention for all spectrogram columns that fall within that phone's time range.

5.1. Results

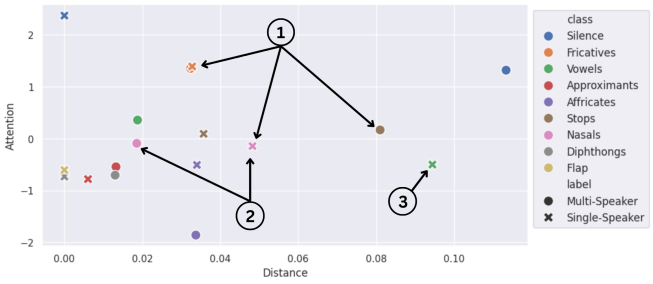


Fig. 2: Alignment between detector attention and distributional distance. Attention reflects the average reduction in model confidence when a phone of the given class is removed from the target sample.

Overall, the detector attention shown in Fig. 2 across the different phone classes aligns well with the results of our phonetic analysis, with the most fricatives, stops, and nasals as the most relevant phonetic classes (①). However, affricate attention is only 42% that of the fricative rate, despite similar phonetic error rates in both groups. Detector attention is largely consistent between groups for each phone class, independent of differences in phonetic error (e.g. attention for nasals is similar for both groups despite significantly higher error in the multi-speaker group ②). This indicates that architectural features of the detector constrain the separation and identification of useful phonetic units.

This analysis also reveals that detector reliance on non-speech regions continues even with access to discriminatory speech characteristics. In the single-speaker group, silence regions had an average attention 23% greater than fricatives, the second most attended to class, and 61% greater than the average attention across all classes. This is despite exceptionally low distributional error in non-speech regions for this group. In the multi-speaker group, silence attention is high, similar to the average fricatives, aligning with high spectral differences in that group. In the single-speaker group, vowels received minimal attention despite significant errors (③).

> *Takeaway*: Detection attention could be greatly improved by emphasizing or enforcing focus on key phone classes. Because relative attention across classes appears stable across training, achieving this may require alternative architectures or an ensemble of phone-specific detectors.

6. RELATED WORK

Several works [1, 2] show that mainstream countermeasures rely on artifacts of speech synthesis (clicks, discontinuities, and unnatural silence regions), rather than intrinsic characteristics of speech. These artifacts are often generator-specific [1, 2] and fragile to even slight signal changes [1]. This has motivated data augmentation methods and training methods to push detectors toward direct speech cues. Notably, detectors trained on isolated unvoiced speech regions (e.g., individual obstruents) outperform those trained on whole audio samples [4], despite the latter having access to both voiced and unvoiced regions. Performance improved when fused with the whole-audio model, indicating strong complementarity between highly focused phonetic models and more general detectors.

Direct natural–synthetic comparisons have revealed systematic generation errors, including a bias toward higher-frequency regions where poor synthesis is less perceptible [3]. Phonetic analyses [6, 5] confirmed that reproduction errors vary by phone class: obstruents with strong turbulent airflow are less faithfully generated than more tonal phones, with later work focusing specifically on fricatives [6]. Yet these studies were constrained by outdated architectures, limited generator variety, and narrow speaker diversity [5, 6]—limitations we overcome in this paper.

7. CONCLUSION

Applying phonetic analysis to a large set of diverse, modern speech generators reveals that reproduction of obstruent phones is still a challenge, despite significantly improved perceptual quality. The rate and significance of errors vary greatly between generator classes, with the raw performance of modern vocoders alone outstripping text-to-speech generation. Similarly, end-to-end voice conversion significantly outperforms similar multispeaker TTS systems. Finally, despite apparent focus on relevant phonetic regions, detectors still rely heavily on non-speech signals for classification, increasing the risk of detector failure in the presence of signal degradation.

A. REFERENCES

[1] Andre Kassis and Urs Hengartner, "Breaking security-critical voice authentication," in *IEEE Symposium on Security and Privacy (SP)*. IEEE, 2023, pp. 951–968.

[2] Tsu-Hsien Shih, Chin-Yuan Yeh, and Ming-Syan Chen, "Does audio deepfake detection rely on artifacts?," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[3] Harry Maltby, Julie Wall, Cornelius Glackin, Mansour Moniri, Nigel Cannings, and Iwa Salami, "A frequency bin analysis of distinctive ranges between human and deepfake generated voices," in *International Joint Conference on Neural Networks Models*, 2024.

[4] Ganesh Sivaraman, Hemlata Tak, and Elie Khoury, "Investigating voiced and unvoiced regions of speech for audio deepfake detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[5] Ayushi Pandey, Sébastien Le Maguer, Julie Carson-Berndsen, and Naomi Harte, "Production characteristics of obstruents in WaveNet and older TTS systems.," in *Interspeech*, 2022, pp. 2373–2377.

[6] Sriyugesh Bhyravajjula, Ayushi Pandey, and Arun Baby, "Fricatives in modern text-to-speech synthesizers," in *Proc. Speech Synthesis Workshop*, 2025, pp. 222–227.

[7] Roger K. Moore and Lucy Skidmore, "On the use/misuse of the term 'phoneme'," in *Interspeech 2019*, 2019, pp. 2340–2344.

[8] Henning Reetz and Allard Jongman, *Phonetics: Transcription, production, acoustics, and perception*, John Wiley & Sons, 2020.

[9] Rotem Rousso, Eyal Cohen, Joseph Keshet, and Eleanor Chodroff, "Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment," in *Interspeech*, 2024, pp. 1525–1529.

[10] Jian Zhu, Cong Zhang, and David Jurgens, "Phone-to-audio alignment without text: A semi-supervised approach," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8167–8171.

[11] Cédric Villani, "The wasserstein distances," in *Optimal transport: old and new*, pp. 93–111. Springer, 2009.

[12] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[13] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[14] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17022–17033, 2020.

[15] Joel Frank and Lea Schönherr, "WaveFake: A dataset to facilitate audio DeepFake detection," Nov. 2021.

[16] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*, 2021, pp. 5530–5540.

[17] Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al., "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, pp. 271–350, 2019.

[18] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti, "YourTTS: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*, 2022, pp. 2709–2720.

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

[20] Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun, "Openvoice: Versatile instant voice cloning," 2024.

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

[22] Tuan Duy Nguyen Le, Kah Kuan Teh, and Huy Dat Tran, "Continuous learning of transformer-based audio deepfake detection," 2024.

[23] Chirag Goel, Surya Koppisetti, Ben Colman, Ali Shahriyari, and Gaurav Bharaj, "Towards attention-based contrastive learning for audio spoof detection," in *Interspeech*, Aug. 2023, pp. 2758–2762.

[24] Boo Fullwood and Fabian Monrose, "Seeing is believing: Interpreting behavioral changes in audio deepfake detectors arising from data augmentation," in *Proceedings of the Workshop on Artificial Intelligence and Security (AISec)*, 2025.