A Novel Dataset for Testing Anti-spoofing Models in a Telephony Environment

Zachary Nicholas Houghton
Department of Linguistics
University of California, Davis
Davis, USA
znhoughton@ucdavis.edu

Dan Pluth
Vail Systems
Chicago, USA
dpluth@vailsys.com

Jordan Hosier
Vail Systems
Chicago, USA
jhosier@vailsys.com

Vijay K. Gurbani
Vail Systems
Chicago, USA
vgurbani@vailsys.com

***Abstract*—In the last few years, synthetic voices have become incredibly realistic and more difficult to distinguish from authentic, human voices. Although impressive, these advances raise security concerns, increasing the need for models that can discriminate between human and synthetic voices under real-world conditions. While previous work has proposed datasets and models that provide convincing results for high-quality recordings, very few studies have examined the efficacy of models under diverse conditions - both speaker and channel variations. Thus, it is unclear how well these models generalize to novel, less pristine channel conditions. In this paper, we present a novel dataset for testing the performance of such models under noisy conditions associated with the cellular telephone network. We improve upon previous methods by including a variety of synthesizers as well as languages. Finally, we demonstrate that a model trained on this dataset can achieve high accuracy on novel telephony data without any degradation in accuracy on non-telephonic audio.**

***Index Terms*—speech recognition, speaker recognition, automatic speech recognition, telephony, biometric authentication**

I. Introduction

The ability to create synthetic voices that can imitate an individual's voice has dramatically improved such that many of these synthetic voices are difficult to distinguish from the human voice that they imitate. The inability to discriminate synthetic voices from human voices is of great concern for many reasons. For example, synthetic voice clones can be used for deception and to spread misinformation using the familiar voice of an authority figure or political leader. When conducting business through the telephone (phone banking, for instance), it can be used to steal one's identity, or access bank accounts by impersonating the real user's voice in an interactive voice response setting. For example, recently thieves imitated a company executive's voice in order to steal hundreds of thousands of dollars.[1]

Thus, in this paper we present our work on detecting such voice spoofs in telephony speech. Our specific contributions are as follows:

- We create a telephony dataset that captures diverse channel conditions associated with cellular networks. We will provide this dataset and the corresponding code to the research community.[2]
- We present a novel approach to create high-quality synthetic telephony data.
- We train a model that exhibits high accuracy in discriminating real human voices from synthetic voices in both clear and telephony speech, even when encountering out-of-distribution synthetic samples created by high end commercial synthetic voice generation tools.
- We apply transfer learning by re-purposing a high-performance speaker embedding model for liveness detection, showcasing the adaptability of speaker representations to spoof detection tasks.
- We demonstrate the viability of a lesser-known fine-tuning approach of freezing the output layer and fine-tuning once to mitigate catastrophic forgetting.

The rest of this paper is structured as follows: Section II positions our work in context with the current literature on anti-spoofing, Section III outlines our datasets and methodology including a comprehensive account of the process we used to introduce cellular network characteristics in the datasets. We discuss our proposed changes to the NVIDIA NeMo pre-trained TitaNet speaker recognition model [1] to allow it to discriminate synthetic voices from live speakers. We present the results of our work and subsequent discussions in Section V, and we conclude in Section VI.

II. Related Work

In the last few years, deep learning methods have advanced rapidly, enabling text-to-speech models to achieve incredible results [2]. Among these models, data-driven techniques have resulted in text-to-speech and voice conversion models that are extremely realistic. Data-driven voice synthesis models, as their name suggests, learn the structure of the waveforms from data. For example, Wavenet [3] uses a generative model that produces speech by estimating the probability of the raw waveform (conditioned on all the previous waveforms). This approach has achieved state-of-the-art performance.

Voice cloning in text-to-speech as well as voice-conversion both work by learning about not only the acoustics of the

[1] https://www.scmp.com/print/news/world/article/3025772/ai-first-voice-mimicking-software-used-major-heist.

[2] Available at: https://github.com/vail-systems/IEEE-RTC-2025.

language they're being trained on, but also by learning how to encode speaker-embeddings. These embeddings capture various characteristics of the speaker, such as their speech rate or accent [4]. These speaker-embeddings can be learned from only a few samples of audio [4], sometimes as little as a few seconds of speech. By combining the speaker embeddings with any given text, an audio can be created making that 'person' say anything. The naturalness of these models combined with how little audio they require has these models suitable for abuse, with applications in identity theft and misinformation campaigns.

As a consequence of these advancements, there has been an increasing concern of spoofing attacks. Spoofing attacks are attempts to trick a system into granting an unauthorized user access. These attacks can take the form of several different forms, such as impersonation, synthetic speech, voice conversion, or replay attacks [5]. In the present study we focus on a specific case of anti-spoofing, liveness detection. In liveness detection, unlike other voice biometric systems such as automatic speaker verification, the system is only concerned with detecting whether the speech is being generated from a live speaker or whether it has been synthesized. For example, in an ideal world, the liveness detection model would perfectly verify that the audio is human-generated, and would flag any synthetic speech.

There have been several attempts to create datasets to train and test liveness detection models [6, 7, 8] along with novel models to detect spoofed voices [9, 10, 11, 12]. For example, the Multi-Language Audio Anti-Spoof Dataset (MLAAD) is a diverse dataset that contains data from 59 different text-to-speech models in 23 languages. However, a crucial limitation is that these recordings are all clean and relatively noise free. For these datasets to be useful in a real-world setting, a model must be able to achieve high performance in a noisier environment and across diverse channel conditions.

Previous work has looked at how some properties of speech affect liveness detection models' ability to correctly identify the speech as being human or not. For example, [13] examined how silence in the audio files affects a model's performance on liveness detection. They found that equal error rates (EERs) increase significantly when silence is removed from the samples. In other words, by concatenating silence to the beginning of an audio sample, one could fool a liveness detection model. They argued that this is because TTS algorithms generate speech with a lower proportion of silence compared to human speech.

Additionally, [12] examined whether a classification model could achieve high accuracy in detecting a replay attack (previously recorded audio of a human being played back into a microphone). In order to test their classification model, they instructed participants to speak aloud various commands into a cellphone and recorded them. The recordings were made in an open lab environment such that there was an element of noise present in the recordings. They found that their classification model was able to perform well on this dataset. However, while their dataset was recorded using a cellphone microphone, there are key properties of telephonic speech outside of just the microphone. For example, the carrier, location, and amount of network traffic can all have an effect on the quality of the audio as its transmitted across a network.

Similarly, there have been previous attempts to create datasets comprised of telephony speech [8, 9]. For example, [8] introduced the telephony dataset, *Phonespoof*. Their approach consisted of replaying synthetic voices on a computer and transferring it to a mobile phone using a 3Jack-4Jack cable. While it was transferred, the mobile phone called into one of two software: Smart Logger II or Smart Caller and this call was recorded. Variability in the recording conditions was introduced by using two different mobile phones and two different telecommunications operators. However, this dataset no longer reflects the current ecosystem. Specifically, the data set is limited in language diversity, having examined only English and Russian, and contains recordings obtained by using products that no longer exist to capture calls. It is not clear what effect these capture methods have on the quality of the audio. Finally, rather than capturing the audio with the cellphone microphones, they use a phone jack to directly feed audio into the phone. This creates a distinct audio characteristic that is only one of many possible ways for an attacker to bypass security measures.

Additionally, [9] examined the vulnerability of speaker verification systems against voice conversion attacks. They examined the performance of models ranging from simple Gaussian mixture models (GMMs) to a joint factor analysis (JFA) recognizer. Their results suggested that these systems are vulnerable to spoofing attacks, especially in telephonic speech (speech transmitted through a telephone or cellular network). However, since the paper's publication, there have been breakthroughs in both speaker-recognition models as well as text-to-speech (TTS) models. Thus, it is unclear whether these vulnerabilities remain.

More recently, [14] also examined the ability of a model to identify whether a voice is authentic or not. While they did not explicitly examine telephony speech, they did examine the effects of channel conditions. Specifically, the generated voice clones of speech taken from the VCTK dataset, the Mozilla Common Voice dataset, and the AnonVox dataset. This study employed three different TTS engines to create synthesized counterparts to the authentic speech datasets. The TTS engines used were XTTS[3], StyleTTS2 [15], and YourTTS [16]. Importantly, each of the datasets vary in their channel conditions.

In order to examine the role of channel conditions, [14] trained a TitaNet SVM model to identify whether the voices are real or fake. While TitaNet is a speaker-recognition model, its output consists of speaker embeddings, in this case used as input to a Support Vector Machine (SVM)

[3]https://github.com/coqui-ai/TTS

model to classify the voice as either real or spoofed. They then trained this model on real and spoofed voices from one dataset, holding out data from the remaining datasets. Results revealed that a model trained using this approach struggles with identifying whether a voice is real or spoofed in the out-of-domain (held out) dataset. Their results demonstrate the need for an in-depth examination of liveness detection methods in noisy speech.

III. Dataset and Methodology

Our training and validation datasets contain data from 5 datasets: M-AILABS [17], Multi-Language Audio Anti-Spoof Dataset [6], cellularized MLAAD, Clipwise (explained below), and ASVspoof2019. The "cellularization" process is explained in depth below.

Our test dataset comprises the five previously mentioned datasets along with three additional datasets: the ASVspoof2019 evaluation set [18], the Call Home dataset [19], and the cellularized ElevenLabs dataset – a version of the LibriSpeech dataset [20] which we then converted to synthetic speech using ElevenLabs. We describe each of these datasets in depth below, and a breakdown is included in Table I.

The motivation for the training set was to provide the model with as much information as possible with respect to a variety of synthesizers as well as a variety of channel conditions. The test set is designed to test a model's performance on out-of-domain distribution of synthesized data as well as out-of-domain distribution of telephony data samples captured over a cellular telephone network. The key features of the test set are that it contains 14 novel (i.e., not seen in training) synthesizers and speech that is both clean and telephonic. As such, high performance on the test set indicates that the model is able to generalize well.

- **M-AILABS**: M-AILABS [17] is a speech dataset that contains nearly a thousand hours of audio book recordings in several different languages. The recordings were produced in clean, relatively noise-free environments.

- **Multi-Language Audio Anti-Spoof Dataset (MLAAD)**: MLAAD [6] is a speech dataset based on M-AILABS and contains 59 different text-to-speech models in 26 different architectures. The corpus contains a total of 175.0 hours of synthetic voice audio in 23 different languages. The speech in the corpus is taken from audio books or speeches and interviews of public figures.

- **Cellularized MLAAD**: To create a noisier dataset, we transmitted a subset of the MLAAD corpus through a pipeline to generate telephonic versions of this data. We describe the data generation process in Section III-A. This process is the same for both the cellularized MLAAD dataset and the cellularized ElevenLabs dataset. This dataset is intended to emulate synthetic speech in a telephonic environment.

- **Cellularized ElevenLabs**: Similar to the Cellularized MLAAD dataset, this dataset contains synthetic speech that has gone through our cellularization process. Specifically, a subset of LibriSpeech [20] was cloned using ElevenLabs' state-of-the-art text-to-speech program. In order to provide speaker-variability, we used 175 distinct voices to generate the speech. More accurately, for each file in the subset of the LibriSpeech corpus, a voice was sampled at randomly from the list of 175 distinct voices and used to generate speech for the transcript of the LibriSpeech audio file. This synthetic speech was then cellularized using the process described in Section III-A. The goal was to emulate realistic synthetic telephony data. Further, ElevenLabs is considered a state-of-the-art synthesizer. Performance on this data is critical to determining the model's robustness to a zero-day attack. The LibriSpeech dataset was used to maximize the differences between the test set and the training set. The LibriSpeech dataset, similar to M-AILABS, is a speech corpus comprised of audio book recordings.

- **Clipwise**: The Clipwise dataset comprises calls between individuals and a financial institution. The audios have two channels (caller-agent interaction), however only the caller channel was used. The duration of the calls range in length from a few seconds to tens of minutes.

- **ASVspoof2019**: We use the training and evaluation sets from the logical access subset of their dataset of the ASVspoof 2019 dataset [18]. More specifically, the training set consists of speech from 20 different speakers (8 male, 12 female) and 6 different spoofing systems: 2 voice conversion (VC) systems and 4 text-to-speech (TTS) systems. VC systems use a combination of neural-network and spectral-filtering approaches while TTS systems use conventional a source-filter vocoder or a WaveNet-based vocoder. The evaluation set contains 13 novel synthesizers (not present in the training dataset). Thus performance on the ASVspoof2019 evaluation set indicates how well the model can generalize to novel synthesizers.

- **Call Home Dataset**: The call home dataset [19] consists of 120 unscripted 30-minute telephone conversations. These took place in North America between native American English speakers.

Altogether, Our dataset comprises 132,000 audio samples in training, 16,500 audio samples in validation, and 84,377 samples in testing (Table II). A breakdown of the duration of audio in each dataset is included below, however we include a brief summary of the training, validation, and test sets here. The training data comprises about 220.36 hours of data, the validation data comprises about 27.47 hours of data, and the test data comprises 96.77 hours of data. A plot of

Dataset	Audio samples	Training	Validation	Test
MLAAD	Synthetic	36000	4500	4500
M-AILABS	Human	24000	3000	3000
Cellularized MLAAD	Synthetic	16000	2000	2000
Clipwise	Human	40000	5000	5000
ASVspoof2019 Training	Mix	16000	2000	2000
ASVspoof2019 Eval	Mix	–	–	54540
Cellularized Elevenlabs	Synthetic	–	–	1788
Call Home	Human	–	–	11549

TABLE I: Dataset description.

the distribution of the mean duration of audio files in each dataset along with the standard deviation of duration of audio files is included in Figure 1.

A. Cellularization Process

In digital cellular communications, channel characteristics play an important role in spoof detection. As the data packets are transported over the radio channel, they encounter a wide variety of channel conditions, including radio resource contention, signal attenuation, and mobile handoffs [21]. Besides the inherent channel noise, there is also ambient noise when a user makes a phone call from a noisy environment (train station, city street, etc.). All of these factors can influence the audio quality and potentially make it more difficult for a model to detect whether a voice is human or not. Thus, our interest is in creating — and evaluating — a dataset that captures both the inherent and ambient noises associated with cellular telecommunications.

To simulate ambient noise, we randomly sampled a file from the MLAAD dataset (and the ElevenLabs/LibriSpeech dataset), overlaying it with a randomly sampled noise file from the MUSAN noise corpus [22]. To approximate real-world noise conditions, we randomized the introduction of the noise across the playout duration time, and we varied the noise volume randomly. Specifically, after normalizing the volume of the audio file and the volume of the noise, a number was uniformly sampled from $\mathcal{N}(25, 7.5)$. This number was then subtracted from the normalized volume of the noise.

The end result of this was a dataset that consisted of audio files with ambient noise of varying intensities present in different playout positions. To simulate the inherent cellular communications channel characteristics, we used three phones from different manufacturers across two service providers (AT&T and Verizon). Location diversity was also introduced by using the cellular phones in a crowded city apartment, a suburban home, and a suburban apartment. The audio samples created using the technique described in the above paragraph were subsequently played through one of the three cellular phones and transmitted through the service provider's network to create a cellularized MLAAD and ElevenLabs dataset.

The play through process consisted of playing each file that had ambient noise introduced to it on a laptop speaker and positioning a cellular phone such that the audio was captured by the cellular mic and transmitted on the cellular network. The cellular phone was connected to a telephony server that accepted the incoming call and stored the received audio on disk (companies like Twilio, Vonage, RingCentral, and FreeClimb provide such platforms, APIs, and phone numbers). This process is depicted in Figure 2.[4]

IV. Model

The present study utilized transfer learning to adapt the TitaNet speaker recognition model to the liveness detection task [1]. We decided to use TitaNet because it is a speaker recognition model that may have learned characteristics of speech that could help it learn to differentiate between real and spoofed voices[5]. Specifically, TitaNet is a 1D time depth-wise channel separable convolutional model that combines a ContextNet-like architecture with channel attention pooling. This approach feeds the features extracted from the ContextNet model into the attentive pooling layer. A visualization of our modifications to the architecture from [1] is presented in Figure 3.

To test our dataset, we used the NVIDIA NeMo pre-trained TitaNet speaker recognition model version 1.0 [1] with a cross-entropy loss function instead of an additive angular margin loss function[6]. We chose to use a speaker recognition model such as TitaNet because it is likely that TitaNet has learned some characteristics of the speech that might facilitate performance in our liveness detection task.

To target the TitaNet model speaker recognition model towards liveness detection, we swapped the 192 dimension softmax output layer with a two dimension softmax layer. Since fine-tuning the model with a new output layer with randomized weights can lead to catastrophic forgetting of the prior layers [23], we first froze all the other layers and finetuned only the new output layer.

In deep neural networks, the last layer becomes specialized for the specific task throughout learning [24]. Despite this, the weights in the other layers still contain valuable knowledge, such as features of different speakers' voices. This

[4]Our code for this process can be found at https://github.com/FreeClimbAPI/telephonizer.

[5]While we use TitaNet, many other models may also perform well using our dataset and we leave further evaluation to future studies.

[6]We originally used an additive angular margin loss function, however we found that for our task our model did not learn well with this loss function, perhaps because our model has no need to optimize the cosine distance between speaker embeddings, which is the main advantage of the additive angular margin loss function.

Dataset	Audio samples	Training Duration	Validation Duration	Test Duration
MLAAD	Synthetic	62.45	7.76	7.81
M-AILABS	Human	41.09	5.12	5.16
Cellularized MLAAD	Synthetic	30.65	3.81	3.85
Clipwise	Human	54.67	6.82	8.70
ASVspoof2019 Training	Mix	16.39	2.02	2.05
ASVspoof2019 Eval	Mix	–	–	54.20
Cellularized Elevenlabs	Synthetic	–	–	3.03
Call Home	Human	–	–	11.97

TABLE II: Duration of audio for each dataset in hours.

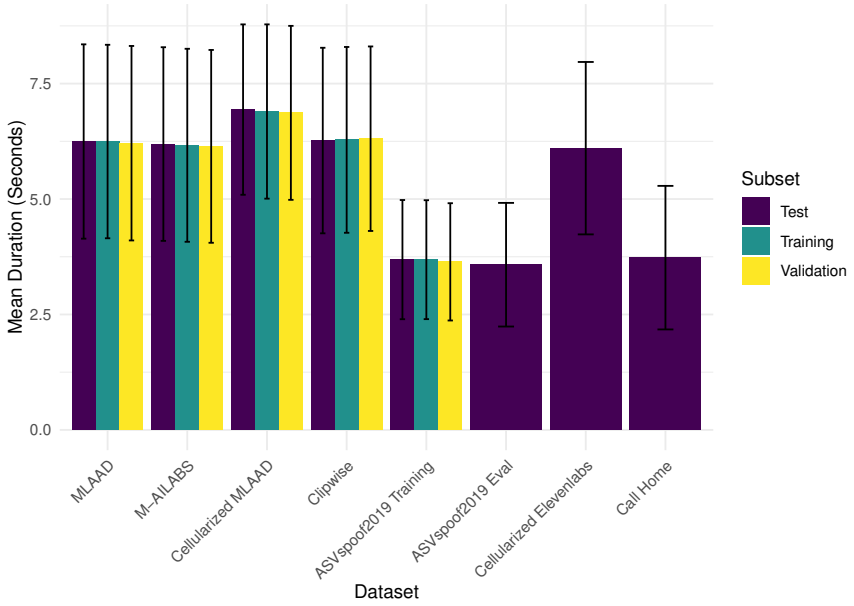

Fig. 1: Mean duration of audio files for each dataset. Lines indicate ± 1 standard deviation.

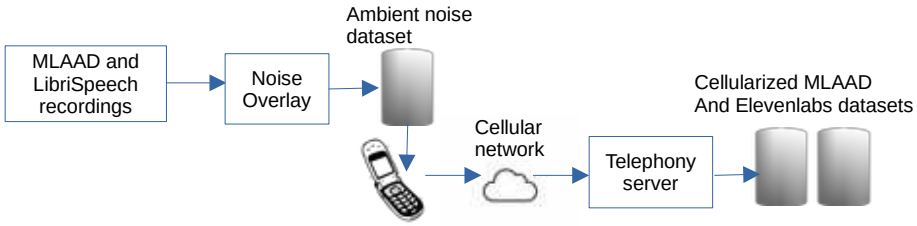

Fig. 2: A visualization of our cellularization process, a process by which we created noisy, telephony samples from the clean relatively noise-free MLAAD recordings.

general knowledge may be useful in categorizing whether a voice is human or synthetic. Through freezing these other layers and retraining the output layer, we train the model for the new task without forgetting all the knowledge it has learned.

Finally, once the new output layer was trained, we fine-tuned the entire model, without any layers frozen, to minimize the cross-entropy function.

V. Results and Discussion

Table III shows the confusion matrix on our full test set, while Table IV shows the results on the entire test set. Our overall accuracy is 0.959 with an EER of 0.041 (a low EER is preferred as the model minimizes the chances of false positives and false negatives). Additionally, a plot of EER as a function of different thresholds can be found in Figure 4. Interestingly, the optimal threshold is closer to 1, demonstrating that a threshold that is strict about acceptance performs best. With respect to the positive class being recognized as a synthetic voices, the model also exhibits high

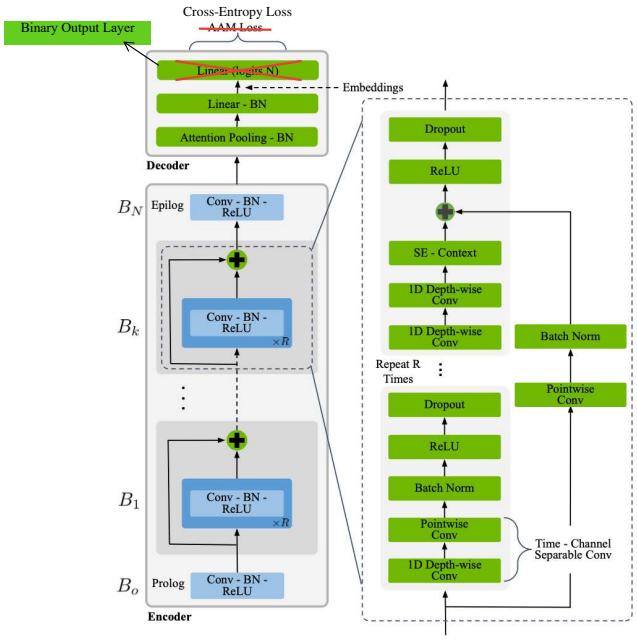

Fig. 3: TitaNet model with binary output layer and cross-entropy loss function. This image is a modified version of the one from [1].

precision and recall.

	Actual	
Predicted	**Synthetic**	**Human**
Synthetic	55176	1091
Human	2336	25774

TABLE III: Confusion matrix of our model results.

Statistic	
Precision	0.981
Recall	0.959
Accuracy	0.959
EER	0.041

TABLE IV: Model statistics.

However, it is important to stratify the results, as performance in- and out-of-domain will vary. Specifically, the model performs exceptionally well on datasets which have the same synthesizers or the same shared linguistic content. For the subset of the test data comprising MLAAD, M-AILABS, Cellularized MLAAD, Clipwise, and ASVspoof2019 Training (i.e., for in-domain dataset), the model achieves 99.2% accuracy (Table V). This suggests that our model is quite successful in identifying whether audio files that share characteristics with its training data are human or spoofed.

Table V shows the accuracy for both in- and out-of-distribution datasets. The model was able to achieve nearly perfect accuracy in discriminating the synthetic samples from

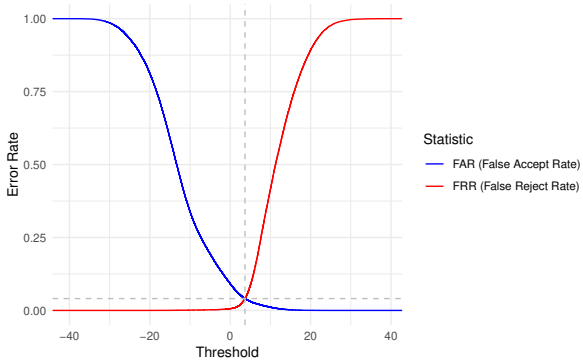

Fig. 4: Plot of our Equal Error Rate at various thresholds. EER thresholds were calculated on the logit estimates produced by the model which were converted to probabilities using softmax. The results demonstrate that the best threshold is one that is relatively strict about what it accepts.

ElevenLabs[7], which were transported over a cellular telephony network. This performance suggests that ElevenLabs may share genetic similarity with some of the open-source synthesizers in our training set. It is possible that our model may perform worse on a zero-day attack generated by a synthesizer with never-seen-before features. This interpretation is supported in part by the results on the ASVSpoof2019 evaluation set.

Domain	Dataset	Mean Accuracy
In-domain	asvspoof2019_training	0.998
	cellularized_mlaad	1.000
	mlaad	1.000
	mailabs	0.972
	clipwise	0.993
Out-of-domain	asvspoof2019	0.952
	call_home	0.942
	cellularized_elevenlabs	1.000

TABLE V: Results stratified by in-domain/out-of-domain datasets.

Our model also shows a slight decrease in performance for the Call Home dataset but still achieves a relatively high accuracy. This is noteworthy because the Call Home dataset is markedly different from our training data. Specifically, the Call Home dataset comprises 120 unscripted 30-minute telephone conversations. Additionally, the type of speech was likely quite different between the Call Home dataset and the Clipwise dataset. The audio Call Home dataset consists largely of audio between family members or close friends. On the other hand, the authentic telephonic samples in the Clipwise dataset are calls between a customer and a representative. Even when expressing the same meaning, the context of use (i.e., whether it is spoken to a friend, a coworker, etc) can drastically affect the form of the language

[7]ElevenLabs is a state-of-the-art synthetic voice generation platform that is available commercially.

(the words and phrases used to express the meaning [25]). In addition, it is also likely that the content expressed in a call to a representative versus a call to a friend are also very different. As a result, the type of audio in our telephonic samples was almost certainly quite different from the type of audio in the Call Home dataset. Additionally, the mean duration of audio also varies by dataset, with the Call Home audio samples being shorter on average than the Clipwise dataset (Figure 1. Despite this, the high accuracy of our model's performance on this dataset suggests a high degree of generalizability to novel contexts.

Finally, our model performs relatively well on the AVSpoof2019 evaluation set, comparably to the leading single-system models reported in the summary report [26]. Table VI presents a breakdown of the model's performance for each synthesizer in the AVSpoof2019 evaluation set. The synthesizers that the model performs more poorly on are A13, A17, and A18, which are notably different from the synthesizers encountered in training. Though it is worth mentioning that our model still performs well above chance on these synthesizers. Specifically, A13 is a text-to-speech voice-conversion system that uses a combination of waveform concatenation and waveform filtering to produce synthetic speech. A17 and A18 are both voice-conversion systems, but A17 uses waveform filtering while A18 uses a vocoder. The decrease in performance on A13 is likely due to the absence of synthesizers that uses both waveform filtering and waveform concatenation in the training set. The training dataset contains no voice-conversion models that use waveform filtering or a vocoder, which likely explains the decrease in performance on A17 and A18. These results are interesting because most of these features (voice-conversion vs text-to-speech, waveform concatenation vs vocoder) were present independently in some of the synthetic samples in the training set. This suggests that encountering some features alone isn't sufficient for the model to learn that they correspond to synthetic speech. Instead, certain combinations of features can present challenges to the model, even if the individual features existed in synthesizers in the training set. Future research is needed to examine this vulnerability in more depth.

A comment should be made about the determination of whether a dataset is in- vs out-of-distribution. The ASVspoof2019 dataset shares no synthesizers between their training and evaluation sets, however they do share the corpus used to develop it. It is also likely that the real audios in that set have significant similarities between the training and evaluation sets. Similarly, the Cellularized ElevenLabs subset likely shares audio characteristics with the Cellularized MLAAD subset, but has distinct text and synthesizer.

VI. Conclusion

We present a dataset comprised of authentic and spoofed voices in both pristine and telephonic recording scenarios. Further, we demonstrate that a model trained on this data performs well on novel speech in both clean and telephony environments as well as both familiar (i.e., seen in training) and novel (i.e., not seen in training) synthesizers.

Previous studies lack diversity of synthesizers and languages and have not examined telephony speech. The present study expands on the current body of literature by investigating key vulnerabilities in real-world environments. To this end, we present a novel approach for creating telephonic speech.

The test set contains several novel, unseen synthesizers as well as novel, realistic telephony speech. This work demonstrates that a model trained on this data performs well on novel samples from known synthesizers, novel samples from a novel synthesizer, and on novel telephony data.

Finally, the results suggest that certain combinations of features, such as the use of voice conversion in conjunction with a vocoder, may lead to reduced classification accuracy when such combinations are not represented during training, even if the individual features are present in isolation. This highlights the importance of evaluating models not only on isolated artifacts but also on compound feature interactions. Although some synthesis characteristics are documented, a comprehensive study of their influence on detection performance remains a promising direction for future work. It is difficult to assess this further within the current study because ASVSpoof2019 does not make the identity of the synthesizers public. In addition, extending this analysis to include comparisons with other anti-spoofing approaches may provide a clearer assessment of model robustness and generalizability.

References

[1] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[2] E. Akhlaghi, I. I. Auðunardóttir *et al.*, "Using the LARA little prince to compare human and TTS audio quality," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, Jun. 2022, pp. 2967–2975.

[3] A. van den Oord, S. Dieleman *et al.*, "Wavenet: A generative model for raw audio," in *Proc. SSW 2016*, 2016, pp. 125–125.

[4] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *Advances in neural information processing systems*, vol. 31, 2018.

[5] M. R. Kamble, H. B. Sailor *et al.*, "Advances in anti-spoofing: from the perspective of asvspoof challenges," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e2, 2020.

[6] N. M. Müller, P. Kawa *et al.*, "Mlaad: The multi-language audio anti-spoofing dataset," in *2024 Interna-*

Synthesizer	Type	Generation Method	Mean Accuracy	# of Observations
A07	TTS	vocoder+GAN	1.000	2942
A08	TTS	neural waveform	0.991	3164
A09	TTS	vocoder	1.000	3329
A10	TTS	neural waveform	0.991	2797
A11	TTS	griffin lim	1.000	2852
A12	TTS	neural waveform	1.000	3554
A13	TTS_VC	waveform concatenation+waveform filtering	**0.886**	3179
A14	TTS_VC	vocoder	1.000	3895
A15	TTS_VC	neural waveform	1.000	3895
A16	TTS	waveform concatenation	1.000	3709
A17	VC	waveform filtering	**0.882**	4717
A18	VC	vocoder	**0.713**	4718
A19	VC	spectral filtering	0.998	4702

TABLE VI: Mean accuracy and number of observations per novel synth in the ASVspoof2019 evaluation dataset. The bolded observations indicate low performance.

tional Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1–7.

[7] P. Kawa, M. Plata, and P. Syga, "Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection," in *INTERSPEECH*, 2022.

[8] G. Lavrentyeva, S. Novoselov *et al.*, "Phonespoof: A new dataset for spoofing attack detection in telephone channel," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2572–2576.

[9] T. Kinnunen, Z.-Z. Wu *et al.*, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4401–4404.

[10] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2013, pp. 1–9.

[11] M. Li, Y. Ahmadiadli, and X.-P. Zhang, "Audio anti-spoofing detection: A survey," *arXiv preprint arXiv:2404.13914*, 2024.

[12] M. E. Ahmed, I.-Y. Kwak *et al.*, "Void: A fast and light voice liveness detection system," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 2685–2702.

[13] Y. Zhang, Z. Li *et al.*, "The impact of silence on speech anti-spoofing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3374–3389, 2023.

[14] D. Pluth, J. Hosier, Y. Zhou, and V. Gurbani, "Echoes unveiled: Identifying synthetic voices," in *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2025.

[15] Y. A. Li, C. Han *et al.*, "Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 19 594–19 621, 2023.

[16] E. Casanova, J. Weber *et al.*, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[17] T. Dataset, "The m-ailabs speech dataset," 2024.

[18] X. Wang, J. Yamagishi *et al.*, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, p. 101114, 2020.

[19] A. Canavan, D. Graff, and G. Zipperlen, "Callhome american english speech," *Linguistic Data Consortium*, 1997.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[21] E. Paksoy, J. C. de Martin *et al.*, "An adaptive multi-rate speech coder for digital cellular telephony," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 1. IEEE, 1999, pp. 193–196.

[22] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[23] A. Kumar, A. Raghunathan *et al.*, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[24] J. Yosinski, J. Clune *et al.*, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

[25] G. T. Slone, "Janet holmes an introduction to sociolinguistics," *Language Problems and Language Planning*, vol. 17, no. 3, pp. 274–275, 1993.

[26] A. Nautsch, X. Wang *et al.*, "Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 252–265, 2021.