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

Anonymous ACL submission

Abstract

In the last few years, synthetic voices have become incredibly realistic and more difficult to discriminate from authentic, human voices. Although impressive, these advances raise concerns about safety and security, increasing the need for models that can discriminate between human and synthetic voices under realistic conditions. While previous work has created datasets and models that provide convincing results for high quality recordings, it is unclear how well they generalize to different conditions. In this paper, we present a novel dataset for testing the performance of anti-spoofing models in noisy conditions associated with the cellular telephone network. We demonstrate that a model trained on this dataset can achieve high accuracy on this novel telephony data without any degradation in accuracy on non-telephonic audio.

1 Introduction

In the last few years, the ability to create synthetic voices that imitate an individual's voice has rapidly increased in quality to the point that many of these synthetic voices are extremely difficult to discriminate from the human voice that they are imitating.

The inability to discriminate synthetic voices from human voices is of great concern for many reasons. For example, the imitation of voices can be used to deceive people or ruin people's reputation, more dangerously, it can be used to steal one's identity, or access bank accounts by impersonating the real user's voice. In this paper, we present our work on detecting such voice spoofs. 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.[1]
- We train a model that exhibits high accuracy in discriminating real human voices from synthetic voices, even when encountering out-of-distribution synthetic samples created by the best in breed commercial synthetic voice generation tools.

2 Related Work

In the last few years, deep learning methods have advanced rapidly. This rapid advancement has enabled text-to-speech models to achieve incredible results. Among these models, data-driven techniques have resulted in text-to-speech models that are extremely realistic, to the point of being difficult to tell apart from a human voice. Data-driven models, as their name suggests, learn the structure of the waveforms from data. For example, Wavenet (Van Den Oord et al., 2016) 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.

Along with these advances in text-to-speech software there has been an increased interest in developing models that can detect synthetic voices (i.e., anti-spoofing models, or spoof-detection models). This increased interest has resulted in attempts to create datasets to train and test anti-spoofing models (e.g., Müller et al., 2024; Kawa et al., 2022) along with models to detect spoofed voices (e.g., Kinnunen et al., 2012; Wu and Li, 2013, and see Li et al., 2024 for review). 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, relatively noise-free recordings. In order to be useful in a real world setting, such as confirming one's identity over the phone, a model must be able to achieve high performance on noisier data

[1] Available at: https://github/[anonymous]

across diverse channel conditions.

There has been some previous work examining the performance of anti-spoofing models in telephonic conditions. For example, Kinnunen et al. (2012) examined the vulnerability of speaker verification systems against spoofing attacks (or voice conversion attacks). They examined the performance of models 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. However, since the paper's publication there have been breakthroughs in both speaker-recognition models as well as text-to-speech models. Thus, there is renewed interest in reexamining the vulnerability of speaker-recognition models, especially for telephony data.

3 Dataset and Methodology

Our training and validation dataset contains data from 5 datasets: M-AILABS (Dataset, 2024), Multi-Language Audio Anti-Spoof Dataset (MLAAD Müller et al., 2024), cellularized MLAAD (explained below), Clipwise, and ASVspoof2019. Our test dataset comprises the 5 before mentioned datasets along with three additional datasets: ASVspoof2019 eval (Wang et al., 2020), the Call Home dataset (Canavan et al., 1997), and cellularized Elevenlabs – a version of the Libri Speech dataset (Panayotov et al., 2015) which we then converted to synthetic speech using ElevenLabs. The cellularized Elevenlabs dataset was further processed in a manner described below which we refer to as cellularization. We describe each of these datasets in depth below, and a breakdown is included in Table 1.

The motivation for the training set was to provide the model with as much information as possible with respect to 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.

- **M-AILABS**: M-AILABS is a speech dataset that contains 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 (Müller et al., 2024) 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 in 23 different languages.

- **Cellularized MLAAD**: In order to create a noisier dataset, we sent the MLAAD corpus through a pipeline in order to generate telephonic versions of this data. We describe the data generation process below. This process is the same for both the cellularized MLAAD dataset and the cellularized Elevenlabs dataset.

- **Cellularized Elevenlabs**: Similar to the cellularized MLAAD, however, in order to ensure that the test set was as different from the training set as possible, we used the LibriSpeech dataset (Panayotov et al., 2015). LibriSpeech, similar to M-AILABS, is a speech corpus comprised of audio book recordings. We took these recordings and created synthesized versions using ElevenLabs' state-of-the-art text-to-speech program. We then used the below cellularization process

- **Clipwise**: Data comprising calls between individuals and a financial institution. The calls are two channel (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 eval sets from the logical access subset of their dataset of the ASVspoof 2019 dataset (Wang et al., 2020). The dataset consists of bonafide and spoofed utterances.

- **Call Home Dataset**: The call home dataset (Canavan et al., 1997) consists of 120 unscripted 30-minute telephone conversations. These took place in North America between native American English speakers.

3.1 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 (Paksoy et al., 1999). Besides the inherent channel

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	–	–	3040
Call Home	Human	–	–	11549

Table 1: Dataset description.

noise, there is ambient noise when a user makes a phone call from a noisy environment (train station, city street, etc.). 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 LibriSpeech dataset) overlaying it with a randomly sampled noise file from the MUSAN noise corpus (Snyder et al., 2015). 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.[2] 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 phones in a crowded city apartment, a suburban home, and a suburban apartment. The dataset created using the technique described in the above paragraph was 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 1.

4 Model

In the present study, we extended the TitaNet speaker recognition model (Koluguri et al., 2022). TitaNet is an encoder-decoder speaker recognition model based on the ContextNet ASR architecture.

In order to test our dataset, we used the Nvidia NeMo version 1.0 pre-trained TitaNet speaker recognition model (22.1M parameters, Koluguri et al., 2022) with a cross-entropy loss function instead of an additive angular margin loss function[3]. The motivation behind using a speaker recognition model was because we hypothesized that a speaker recognition model may have learned characteristics of the speech that might facilitate performance in our anti-spoofing task.

To train the model, we swapped the softmax output layer with a binary output layer on our pretrained model. We then froze all the other layers and finetuned the model in order to adapt the new output layer to the current model weights. Finetuning the model with an output layer with randomized weights could lead to catastrophic forgetting of the prior layers. This was followed by finetuning the entire model (without any layers frozen) to minimize the cross-entropy loss function.

5 Results and Discussion

Table 2 shows the confusion matrix on our full test set of 85,629 utterances, while Table 3 shows the results on the entire test set. Our overall accuracy

[2]After normalizing the volume of the audio file and the volume of the noise, a number was randomly sampled from $\mathcal{N}(25, 7.5)$. This number was then subtracted from the normalized volume of the noise.

[3]We originally used an additive angular margin loss function, however we found that for our task our model did not seem to 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.

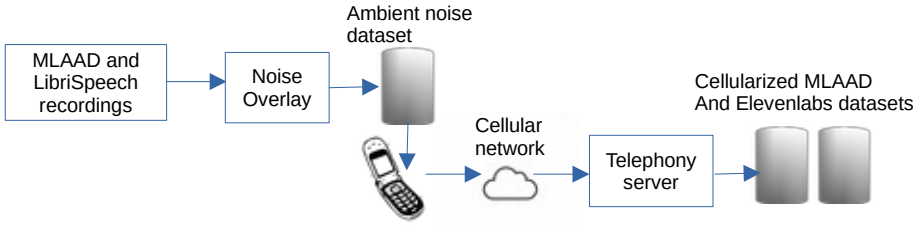

Figure 1: A visualization of our cellularization process, a process by which we created noisy, telephony samples from the clean relatively noise-free MLAAD recordings.

is 0.926 with an Equal Error Rate (EER) of 0.070. (A low EER is preferred as the model minimizes the chances of false positives and false negatives.) With respect to the positive class being recognized as a synthetic voices, the model exhibits high precision and recall.

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, the model achieves 99.9% accuracy.

Table 4 shows the accuracy for the out-of-distribution datasets. Perhaps most surprisingly, the model was able to achieve a perfect accuracy in discriminating the synthetic samples from Elevenlabs[4], which were transported over a cellular telephony network. The accuracy on the human-generated Call Home dataset, while acceptable, is lower than the cellularized Elevenlabs. One reason for this may well be that Elevenlabs' synthetic engine is genetically similar to one of the open-source synthetic engines whose samples appear in our training set. We plan to investigate this as future work.

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 the their Training and Eval sets, however they do share the corpus used to develop the utterances. It is also likely that the real audios in that set have significant similarities between the Training and Eval sets. The Cellularized Elevenlabs likely shares audio characteristics with the Cellularized MLAAD, but has distinct text and synthesizer.

[4]Elevenlabs is widely considered as the state-of-art synthetic voice generation platform available commercially.

	Actual	
Predicted	**Synthetic**	**Human**
Synthetic	53981	1565
Human	4783	25300

Table 2: Confusion matrix of our model results

Statistic	
Precision	0.919
Recall	0.972
Accuracy	0.926
EER	0.070

Table 3: Model statistics.

Dataset	**Accuracy**
ASVspoof2019 Eval	0.910
Cellularized Elevenlabs	1.000
Call Home	0.885

Table 4: Results stratified by in-domain/out-of-domain datasets.

6 Conclusion

In the present study, we expanded upon the current body of literature by presenting a dataset that contains a training set with a variety of different synthetic audio and realistic human audio recordings in a clean, relatively noise-free environment, as well as a telephony environment. Additionally, we present results for a model on a test set that contains both a novel, unseen synthesizer as well as novel, realistic telephony speech. Further, we present an open-access process for producing telephony recordings from pre-recorded audio. Finally, we demonstrate that a model trained on this data can perform well on novel samples from synthesizers it has been trained on, novel samples from a synthesizer that it was not trained on, and novel telephony data.

7 Limitations

First, our test set includes a limited number of novel synthesizers. It is possible that other novel synthesizers may yield different results.

Similarly, performance on novel synthesizers may depend heavily on the synthesizer's architecture. That is, our model may perform better on novel synthesizers whose architecture is similar to synthesizers that the model encountered in its training data than synthesizers with a completely different architecture.

Finally, while our test set includes a novel synthesizer as well as novel telephony data, it is possible that a model trained on our dataset may struggle with other telephony data that exhibits larger variance in the channel properties. Creating a comprehensive telephony specific dataset representative of the real world with a larger diversity of mobile device manufacturers, service providers, location, ambient noises, and diverse speakers remains a challenge.

References

Alexandra Canavan, David Graff, and George Zipperlen. 1997. Callhome american english speech. *Linguistic Data Consortium*.

TMAS Dataset. 2024. The m-ailabs speech dataset.

Piotr Kawa, Marcin Plata, and Piotr Syga. 2022. Attack agnostic dataset: Towards generalization and stabilization of audio deepfake detection. *arXiv preprint arXiv:2206.13979*.

Tomi Kinnunen, Zhi-Zheng Wu, Kong Aik Lee, Filip Sedlak, Eng Siong Chng, and Haizhou Li. 2012. 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)*, pages 4401–4404.

Nithin Rao Koluguri, Taejin Park, and Boris Ginsburg. 2022. 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)*, pages 8102–8106. IEEE.

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

Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. Mlaad: The multi-language audio anti-spoofing dataset. *arXiv preprint arXiv:2401.09512*.

Erdal Paksoy, J Carlos de Martin, Alan McCree, Christian G Gerlach, Anand Anandakumar, Wai-Ming Lai, and Vishu Viswanathan. 1999. 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)*, volume 1, pages 193–196. IEEE.

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.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. 2020. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114.

Zhizheng Wu and Haizhou Li. 2013. Voice conversion and spoofing attack on speaker verification systems. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–9.