Datasheet - WaveFake: A Data Set to Facilitate Audio Deepfake Detection

Joel Frank[*]
Ruhr University Bochum
Horst Görtz Institute for IT-Security

Lea Schönherr
Ruhr University Bochum
Horst Görtz Institute for IT-Security

1 Motivation

For what purpose was the dataset created? The main purpose of this data set is to facilitate research into audio Deepfakes. These generated media files have been increasingly used to commit impersonation attempts [3], influencing opposition movements [14] to justify military actions [5] or online harassment [2]. We hope that this work helps in finding new detection methods to prevent such attempts.

Who funded the creation of the dataset? The creation of this data set was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC-2092 CaSa–390781972.

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? Joel Frank and Lea Schönherr; Ruhr University Bochum and Horst Görtz Institute for IT-Security.

2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? The data set consists of 16-bit PCM wav files.

How many instances are there in total? The data set consists of 117,985 generated audio clips (16-bit PCM wav).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? The data set contains all instances.

What data does each instance consist of? We examine multiple networks trained on two reference data sets. First, the LJSpeech [9] data set consisting of 13,100 short audio clips (on average 6 seconds each; roughly 24 hours total) read by a female speaker. It features passages from 7 non-fiction books and the audio was recorded on a MacBook Pro microphone. Second, we include samples based on the JSUT [20] data set, specifically, basic5000 corpus. This corpus consists of 5,000 sentences covering all basic kanji of the Japanese language (4.8 seconds on average; roughly 6.7 hours total). The recordings were performed by a female native Japanese speaker in an anechoic room. Finally, we include samples from a full text-to-speech pipeline (16,283 phrases; 3.8s on average; roughly 17.5 hours total). Thus, our data set consists of approximately 175 hours of generated audio files in total. Note that we do not redistribute the reference data. They are freely available online [9, 20].

We included a range of architectures in our data set:

[*] Corresponding author `joel.frank@rub.de`.

- **MelGAN**: We include MelGAN [13], which is one of the first GAN-based generative models for audio data. It uses fully convolutional feed-forward network as generator and operates on Mel spectrograms. The discriminator is a combination of three different discriminators that operates on the original, and two downsampled versions of the raw audio input. Additionally, it uses an auxiliary loss over the feature space of the three discriminators.
- **Parallel WaveGAN (PWG)**: WaveNet [16] is one of the earliest and most common architectures, We include samples from one of its variants, Parallel WaveGAN [22]. It uses GAN training paradigm, with a non-autoregressive version of WaveNet as its generator. In a similar vein to MelGAN, it uses an auxiliary loss, but in contrast, matches the *Short-Time Fourier Transform* (STFT) of the original training sample and the generated waveform over mutliple resolutions.
- **Multi-band MelGAN (MB-MelGAN)**: Incorporating more fine-grained frequency analysis, might lead to more convincing samples. We include MB-MelGAN, which computes its auxiliary (frequency-based; inspired by PWG) loss in different sub-bands. Its generator is based on a bigger version of the MelGAN generator, but instead of predicting the entire audio directly, the generator produces multiple sub-bands, which are then summed up to the complete audio signal.
- **Full-band MelGAN (FB-MelGAN)**: We include a variant of MB-MelGAN which generates the complete audio directly and computes its auxiliary loss (the same as PWG) over the full audio instead of its sub-bands. HiFi-GAN [12] utilizes multiple sub-discriminators, each of which is examining only a specific periodic part of the input waveform. Similarly, its generator is built with multiple different residual blocks each observing patterns of different lengths in parallel. Additionally, HiFi-GAN employs the feature-space based loss from MelGAN and minimizes the L_1 distance between the Mel spectrogram of a generated waveform and a ground truth one in its loss function.
- **HiFi-GAN**: HiFi-GAN [12] utilizes multiple sub-discriminators, each of which examines only a specific periodic part of the input waveform. Similarly, its generator is built with multiple residual blocks, each observing patterns of different lengths in parallel. Additionally, HiFi-GAN employs the feature-space-based loss from MelGAN and minimizes the L_1 distance between the Mel spectrogram of a generated waveform and a ground truth one in its loss function.
- **WaveGlow**: The training procedure might also influence the detectability of fake samples. Therefore, we include samples from WaveGlow to investigate maximum-likelihood-based methods. It is a flow-based generative model based on Glow [10], whose architecture is heavily inspired by WaveNet.

Additionally, we examine MelGAN both in a version similar to the original publication, which we denote as MelGAN, and in a larger version with a bigger receptive field, MelGAN (L)arge. This version is similar to the one used by FB-MelGAN, allowing for a one-to-one comparison. Finally, we also obtain samples from a complete *Text-To-Speech* (TTS)-pipeline. We use a conformer [4] to map novel phrases (i.e., not part of LJSPEECH) to Mel spectrograms. Then we use a fine-tuned PWG model (trained on LJSPEECH) to obtain the final audio. We call this data set TTS. In total, we sample ten different data sets, seven based on LJSPEECH (MelGAN, MelGAN (L), FB-MelGAN, HiFi-GAN, WaveGlow, PWG, TTS) and two based on JSUT (MB-MelGAN, PWG).

Is there a label or target associated with each instance? The data is entirely generated data, we do not redistribute the training data.

Is any information missing from individual instances? No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? The data is grouped in directories named after the respective network.

Are there recommended data splits (e.g., training, development/validation, testing)? No.

Are there any errors, sources of noise, or redundancies in the dataset? The data is the direct output of the respective networks.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The corresponding training data is available online [9, 20].

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)? No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No.

Does the dataset relate to people? No.

3 Collection Process

How was the data associated with each instance acquired? For WaveGlow we utilize the official implementation [18] (commit 8afb643) in conjunction with the official pre-trained network on PyTorch Hub [17]. HiFi-GAN also offers a public repository on GitHub with pretrained models [11]. We use a popular implementation available on GitHub [6] (commit 12c677e) for the remaining networks. The repository also offers pre-trained models. When sampling the data set, we first extract Mel spectrograms from the original audio files, using the pre-processing scripts of the corresponding repositories. We then feed these Mel spectrograms to the respective models to obtain the data set. Intuitively, the networks are asked to ”recreate“ the original data sets. For sampling the full TTS results, we use the ESPnet project [21, 7, 8, 15]. To make sure the generated phrases do not overlap with the training set, we downloaded the common voices data set [1] and extracted 16.285 phrases from it.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? See above.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? The data was not sampled from a bigger set.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The data was collected by Joel Frank and Lea Schönherr. The Ruhr University Bochum paid both.

Over what timeframe was the data collected? The data was collected from May-September 2021.

Were any ethical review processes conducted (e.g., by an institutional review board)? No.

Does the dataset relate to people? No.

4 Preprocessing/cleaning/labeling

The data set was not cleaned and is distributed in raw form.

5 Uses

Has the dataset been used for any tasks already? The data set was used for a comprehensive analysis of differences between generative architectures. Additionally, several experiments were conducted on how to detect such generated media.

Is there a repository that links to any or all papers or systems that use the dataset? The initial publication can be found online.

What (other) tasks could the dataset be used for? The intended use of this data set is to facilitate research into detecting audio Deepfakes. Our data set consists of phrases from non-fiction books (LJSPEECH) and everyday conversational Japanese (JSUT), which are already available online. The same is true for all models used to generate this data set. Thus, we cannot think of a quick way to misuse our data. On the contrary, we hope it can accelerate research into malicious usage of generative models that already cause damage to society.

One might wonder if releasing research into detecting Deepfakes might negatively affect the detection "arms race". This is a long-standing debate in the security community, and the overall consensus is that "security through obscurity" does not work. This is also often echoed in best security practices, for example, published by the National Institute of Standards and Technology (NIST) [19]. Intuitively, withholding information from the research community is more harmful since attackers will eventually adapt to any defense one deploys.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? No.

Are there tasks for which the dataset should not be used? See above.

6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? The data set is freely available online.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? The data set is distributed through zenodo [2].

When will the dataset be distributed? Already available.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? The data set is distributed with a CC-BY-SA 4.0 license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? The LJSPEECHdata set is in the public domain. The JSUTcorpus is licensed by CC-BY-SA 4.0, with a note that redistribution is only permitted in certain cases. We contacted the author, who saw no conflict in distributing our fake samples, as long as its for research purposes.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? No.

7 Maintenance

Who is supporting/hosting/maintaining the dataset? The data set is hosted on zenodo.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? See above.

Is there an erratum? No.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? New networks might be added in the future. Nothing will be deleted.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? The data set does not relate to people.

[2] `zenodo.org/record/5270336` - DOI: 10.5281/zenodo.5270336

Will older versions of the dataset continue to be supported/hosted/maintained? Older versions are available on zenodo.

If others want to extend/augment/build on/contribute to the data set, is there a mechanism for them to do so? Please contact the corresponding author.

References

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus. In *Language Resources and Evaluation Conference*, 2020.

[2] Matt Burgess. Telegram Still Hasn't Removed an AI Bot That's Abusing Women. *Wired*, 2020.

[3] Lorenzo Franceschi-Bicchierai. Listen to This Deepfake Audio Impersonating a CEO in Brazen Fraud Attempt. *Motherboard*, 2020.

[4] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-Augmented Transformer for Speech Recognition. In *Proceedings of Interspeech (INTERSPEECH)*, 2020.

[5] Karen Hao. The Biggest Threat of Deepfakes isn't the Deepfakes Themselves. *MIT Technology Review*, 2019.

[6] Tomoki Hayashi. Parallel WaveGAN (+ MelGAN & Multi-band MelGAN) implementation with Pytorch. `https://github.com/kan-bayashi/ParallelWaveGAN`, 2020.

[7] Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan. Espnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[8] Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

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

[10] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[11] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. `https://github.com/jik876/hifi-gan`, 2020.

[12] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[13] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[14] The Atlantic Council's Digital Forensic Research Lab. Inauthentic Instagram accounts with synthetic faces target Navalny protests. *Medium*, 2021.

[15] Chenda Li, Jing Shi, Wangyou Zhang, Aswin Shanmugam Subramanian, Xuankai Chang, Naoyuki Kamo, Moto Hira, Tomoki Hayashi, Christoph Boeddeker, Zhuo Chen, and Shinji Watanabe. ESPnet-SE: End-to-end speech enhancement and separation toolkit designed for ASR integration. In *Proceedings of IEEE Spoken Language Technology Workshop (SLT)*, 2021.

[16] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*, 2016.

[17] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: a Flow-based Generative Network for Speech Synthesis. `https://pytorch.org/hub/nvidia_deeplearningexamples_waveglow/`, 2018.

[18] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. WaveGlow: a Flow-based Generative Network for Speech Synthesis. `https://github.com/NVIDIA/waveglow`, 2018.

[19] Karen Scarfone, Wayne Jansen, Miles Tracy, et al. Guide to General Server Security. *NIST Special Publication*, 2008.

[20] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT Corpus: Free Large-Scale Japanese Speech Corpus for End-to-End Speech Synthesis. *arXiv preprint arXiv:1711.00354*, 2017.

[21] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. ESPnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech (INTERSPEECH)*, 2018.

[22] 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.