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

Motivation

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 [2], influencing opposition movements [8] to justify military actions [3] or online harassment [1]. We hope that this work helps in finding new detection methods to prevent such attempts. 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.

The data set is distributed through zenodo [2] with a CC-BY-SA 4.0 license.

Composition

The data set consists of 88,600 generated audio clips (16-bit PCM wav) in total. We examine multiple networks trained on two reference data sets. First, the LJSPEECH [5] 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 [13] 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. Thus, our data set consists of approximately 157 hours of generated audio files in total. Note that we do not redistribute the reference data. They are freely available online [5, 13].

We included a range of architectures in our data set:

- **MelGAN**: We include MelGAN [7], 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. Their discriminator is a combination of three different discriminators that operates on the original, and two downsampled versions of the raw audio input. Additionally, they use an auxiliary loss over the feature space of the three discriminators.
- **Parallel WaveGAN (PWG)**: WaveNet [9] is one of the earliest and most common architectures, We include samples from one of its variants, Parallel WaveGAN [14]. 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

*Corresponding author joel.frank@rub.de.

[2] zenodo.org/record/4904579 - DOI: 10.5281/zenodo.4904579

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

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.
- **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 [6], 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. In total, we sample eight different data sets, six based on LJSPEECH (MelGAN, MelGAN (L), FB-MelGAN, WaveGlow, PWG) and two based on JSUT (MB-MelGAN, PWG).

Collection Process

For WaveGlow, we utilize the official implementation [11] (commit 8afb643) in conjunction with the official pre-trained network on PyTorch Hub [10]. We use a popular implementation available on GitHub [4] (commit 12c677e) for the remaining networks. The repository also offers pre-trained models. We used the pre-trained networks to generate samples that are similar to their respective training distributions, LJSPEECH [5] and JSUT [13]. 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.

Uses & Ethical Considerations

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 an immediate 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 contribute negatively towards the detection ”arms race“. This is a long standing debate in the security community and the overall consensus it 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) [12]. Intuitively, withholding information from the research community is in-fact more harmful, since attackers will eventually adapt to any defense one deploys.

Distribution & Licensing

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. To comply with JSUT we license our data set under the CC-BY-SA 4.0 license.

We do not redistribute any models or training distributions and bear all responsibility in case of violation of rights, etc.

References

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

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

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

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

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

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

[7] 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.

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

[9] 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.

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

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

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

[13] 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.

[14] 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.