Supplementary Material - 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

In this supplementary material we provide full size spectrogram and attribution plots. All the plots are in reference to the same audio sample (LJSPEECH 008-0217). For attribution, we applied BlurIG directly on the feature vectors. Additionally, we provide a visual representations of the filterbanks used and a discussion on releasing security relevant research.

A note on releasing security research

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 often echoed in best security practices, for example, published by the National Institute of Standards and Technology (NIST) [8]. Intuitively, withholding information from the research community is in-fact more harmful, since attackers will eventually adapt to any defense one deploys anyway. Thus, contributing to the invention of new systems is more helpful in an ever changing environment [6].

The debate dates back to at least the 19th century where the cryptographer Auguste Kerckhoffs introduced Kerckhoffs's principle [3]. The principle states that an encryption scheme should still work if an adversary knows everything about the system but a secret passphrase. Similar thought would later be formulated by Claude Shannon [9].

A typical example is the advanced encryption standard (AES). The algorithms entire specification and inner workings can be found in the standardization [7]. Yet, it is considered unbreakable as long as the password used for the encryption is not revealed. AES is also the only algorithm used to encrypt US government documents [1]. The principle also found adoption in the machine learning community, where adversarial defense papers are now advised to evaluate against so-called white box attackers [2], i.e., attackers which know the inner workings of the system and actively try to avoid it.

While complete openness is obviously not possible, the greater security community has adapted similar practices. For example, so-called attack papers are regularly published at security venues. The underlying motivation being, that before one can protect systems, one has to understand how to attack them. Prominent examples are the Meltdown [5] and Spectre [4] vulnerabilities which showed that certain instructions in CPUs could be used for unauthorized access.

Similar patterns are also used in the industry. Google's project zero team regularly analyses and finds critical vulnerabilities in commonly used software. Their standard practice is to inform the vendor and work with them to help fix the vulnerability. However, after a hard deadline of 90 days, the details of the vulnerability will be released to the public [10]. The effects are two-fold. First, the deadline encourages faster patch development by the vendor. Second, the techniques used can be studied to prevent similar vulnerabilities in the future.

*Corresponding author joel.frank@rub.de.

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

Spectrograms

Here we plot the spectrograms of an audio file (LJSPEECH 008-0217) for the training data and the different gernerative networks. Notice the differences especially in the higher frequencies and the horizontal artifacts produced by MelGAN and WaveGlow.

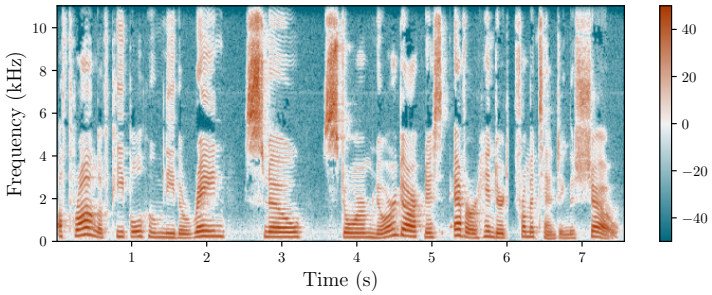


Figure 1: Original

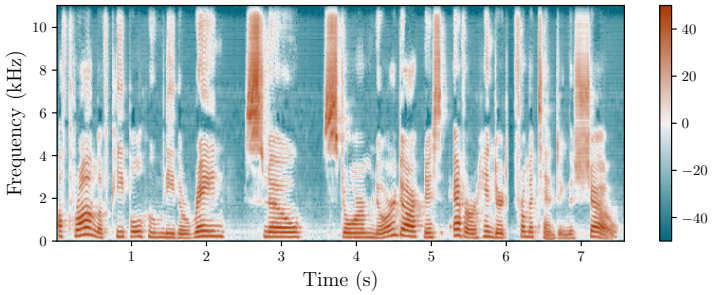


Figure 2: MelGAN

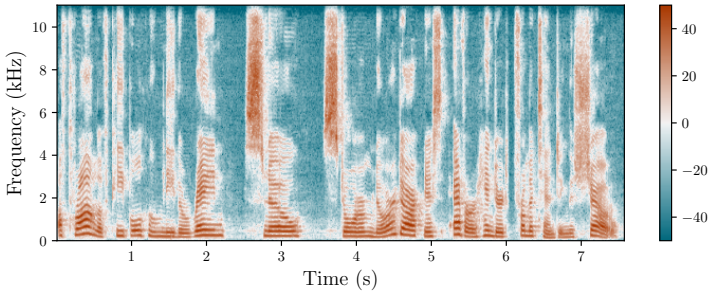


Figure 3: FB-MelGAN

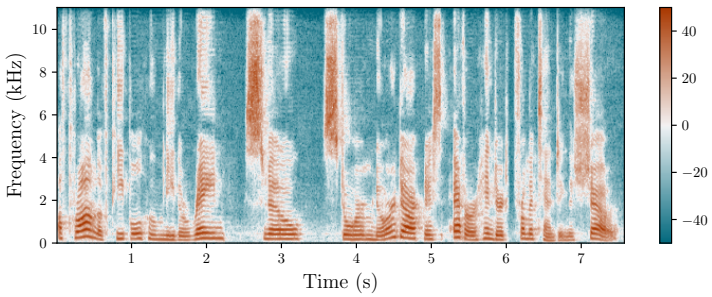


Figure 4: MB-MelGAN

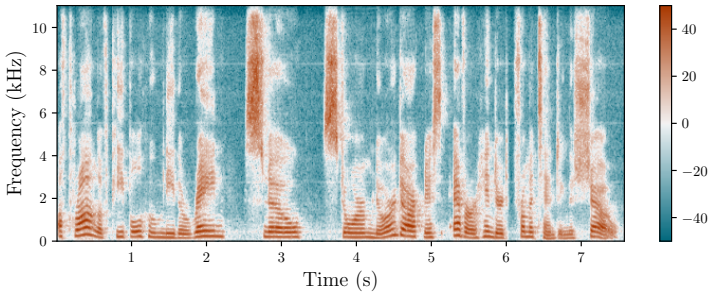


Figure 5: WaveGlow

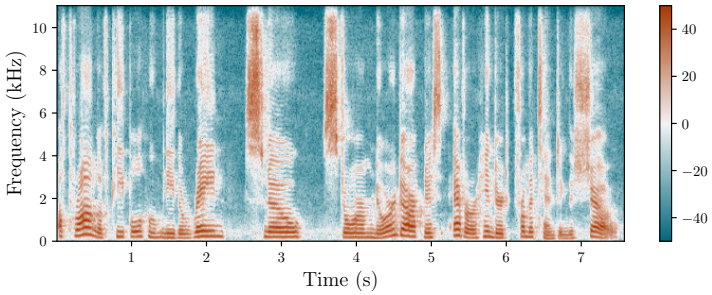
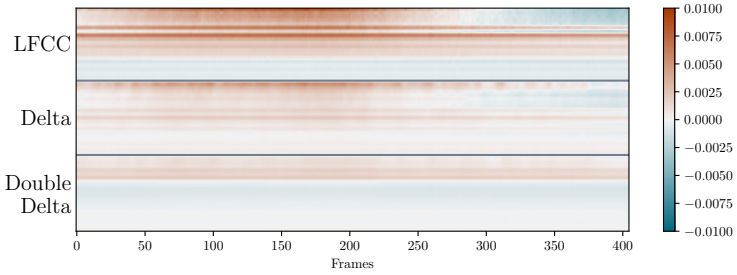


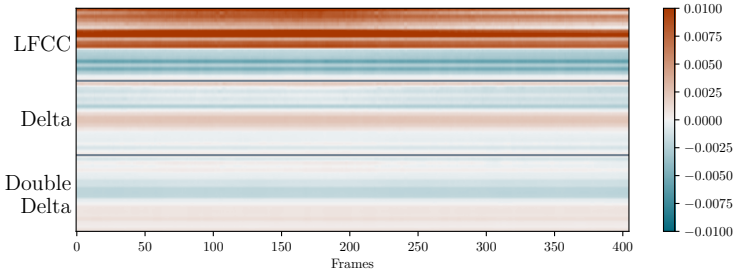
Figure 6: PWG

Attribution

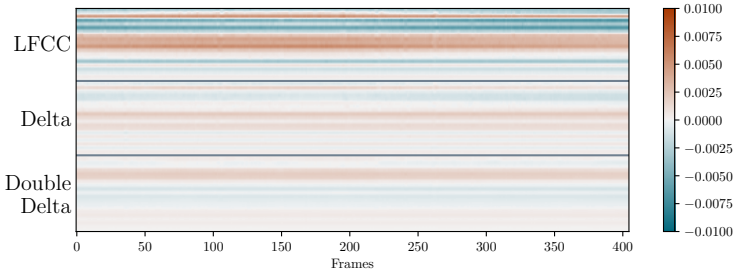
These are the full-size version of the attribution plots used in Section 4.3. Note the spread out attention of the MelGAN classifier, the transition to narrow band attribution and the balance of the classifier trained on FB-MelGAN.



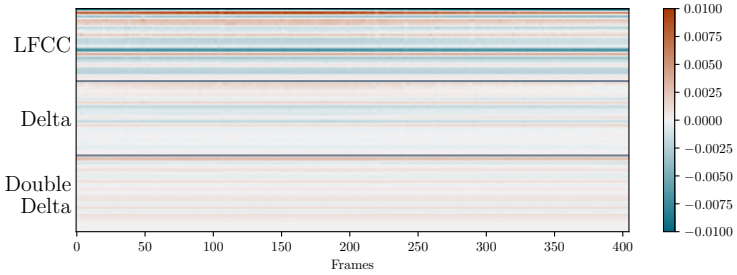
(a) MelGAN (L)



(b) FB-MelGAN



(c) MB-MelGAN



(d) PWG

Filterbanks

Here we show a visual representation of the triangular filterbanks used to compute the *Mel Frequency Cepstral Coefficients* (MFCC) and *Linear Frequency Cepstral Coefficients* (LFCC) features.

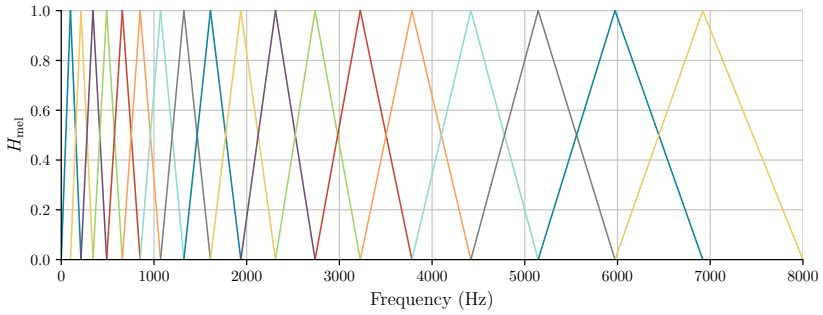


Figure 8: Mel filterbank

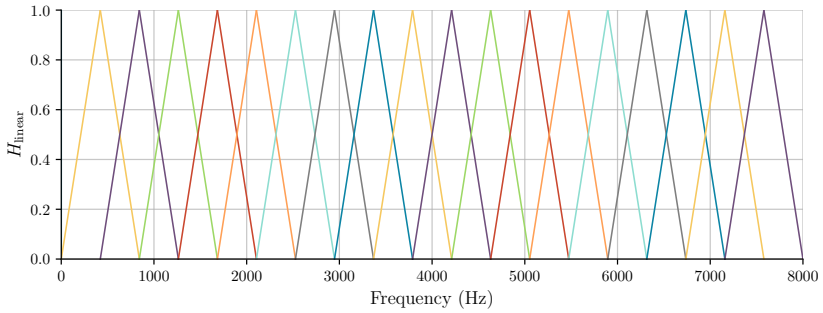


Figure 9: Linear filterbank

References

[1] Elaine Barker et al. Guideline for Using Cryptographic Standards in the Federal Government: Cryptographic Mechanisms. *NIST special publication*, 2016.

[2] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On Evaluating Adversarial Robustness. *Computing Research Repository (CoRR)*, abs/1902.06705, 2019.

[3] Auguste Kerckhoffs. *La cryptographie militaire, ou, Des chiffres usités en temps de guerre: avec un nouveau procédé de déchiffrement applicable aux systèmes à double clef*. Librairie militaire de L. Baudoin, 1883.

[4] Paul Kocher, Jann Horn, Anders Fogh, , Daniel Genkin, Daniel Gruss, Werner Haas, Mike Hamburg, Moritz Lipp, Stefan Mangard, Thomas Prescher, Michael Schwarz, and Yuval Yarom. Spectre attacks: Exploiting speculative execution. In *40th IEEE Symposium on Security and Privacy (S&P'19)*, 2019.

[5] Moritz Lipp, Michael Schwarz, Daniel Gruss, Thomas Prescher, Werner Haas, Anders Fogh, Jann Horn, Stefan Mangard, Paul Kocher, Daniel Genkin, Yuval Yarom, and Mike Hamburg. Meltdown: Reading kernel memory from user space. In *27th USENIX Security Symposium (USENIX Security 18)*, 2018.

[6] Bill McCarty. The honeynet arms race. *IEEE Security & Privacy*, 2003.

[7] Vincent Rijmen and Joan Daemen. Advanced Encryption Standard. *Proceedings of Federal Information Processing Standards Publications, National Institute of Standards and Technology*, 2001.

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

[9] Claude E Shannon. Communication Theory of Secrecy Systems. *The Bell system technical journal*, 1949.

[10] Tim Willis. Project Zero Policy and Disclosure: 2020 Edition, 2020. `https://googleprojectzero.blogspot.com/2020/01/policy-and-disclosure-2020-edition.html`, as of July 11, 2021.