INSURE+C AUDIO DEEPFAKE DETECTION

Students: Chengzhe Sun (UB)

Ehab AlBadawy (UAlbany)

Faculty advisor: Siwei Lyu (UB)

Tech Director: Timothy Davison (JHU APL)

Robinson, Sarah R (JHU APL)

Nathaniel Kavaler (JHU APL)

University at Buffalo The State University of New York

Demo

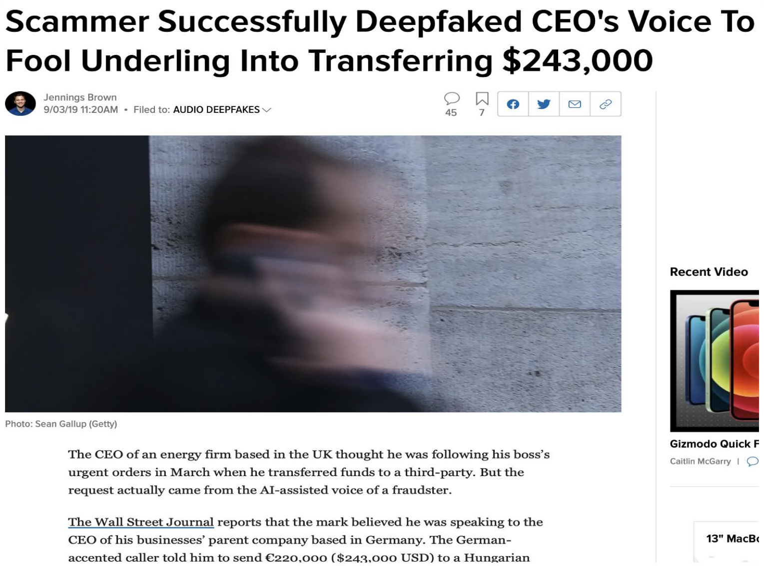
Voice created using [ehab's voice conversion], video created with wav2lip [ACM Multimedia 2020]

Motivation

- DeepFake technologies allow malicious actors to produce audio clips are improving in terms of the quality, scalability, and ease of use.
- DeepFake technologies for audio synthesis has seen significant improvements in the recent years and used by malicious actors for financial gains.

Importance

- The proliferation of DeepFake technologies poses clear threats to society and democracy.
- Synthetic audio detection is one key element of managing this threat.



Scammer Successfully Deepfaked CEO's Voice To Fool Underling Into Transferring $243,000

Jennings Brown
9/03/19 11:20AM • Filed to: AUDIO DEEPFAKES

Photo: Sean Gallup (Getty)

The CEO of an energy firm based in the UK thought he was following his boss's urgent orders in March when he transferred funds to a third-party. But the request actually came from the AI-assisted voice of a fraudster.

The Wall Street Journal reports that the mark believed he was speaking to the CEO of his businesses' parent company based in Germany. The German-accented caller told him to send €220,000 ($243,000 USD) to a Hungarian

Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find

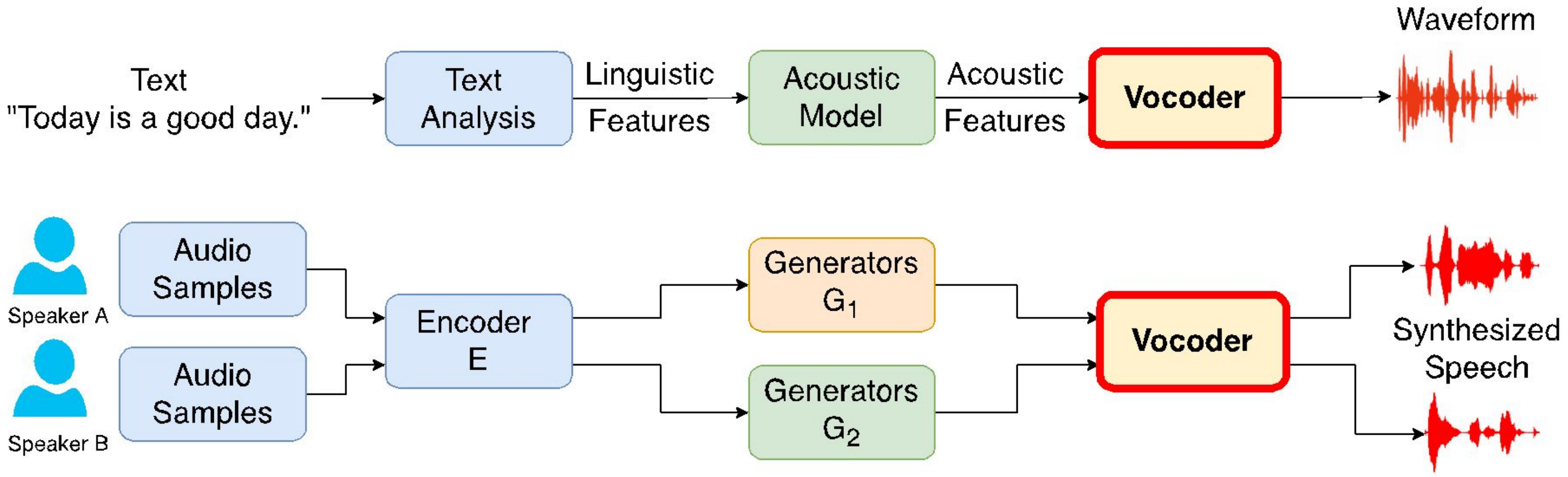
Thomas Brewster Forbes Staff
Cybersecurity
Associate editor at Forbes, covering cybercrime, privacy, security and surveillance.

Cybercriminals cloned the voice of a company director in the U.A.E. to steal as much as $35 million in a huge and complex heist. GETTY

AI voice cloning is used in a huge heist being investigated by Dubai investigators, amidst warnings about cybercriminal use of the new technology.

Problem

- We aim to develop methods to detect synthetic audios by identifying the ***neural vocoders*** used in the generation process
  - ∗ A neural vocoder is a neural network which synthesizes waveforms from temporal-frequency representations, e.g., (mel)spectrograms.
  - ∗ It is one the core component of the most DeepFake audio synthesis algorithms
    - The text-to-speech models, e.g., Parrotron and Spectron converts an input text to the target's voices
    - The voice conversion models uses a source person's voice as input.



Work Elsewhere

- Works on detecting synthetic audios
  - ASVSpoofing challenge 2019 and 2021 (ongoing)
  - Bi-spectral analysis [Albawdaway et.al., CVPRW 2019]
  - DeepSonar [Wang et.al., ACM MM 2021]
- Works on comparing different neural vocoders
  - The work [Govalkar et.al., ISA Workshop 2019] compares a few neural vocoders (3) for speech reconstruction on a small set of input audio signals (100 clips)
- There has been no large-scale benchmarking dataset for the task of vocoder identification and synthetic audio detection
- The lack of the benchmark dataset is a critical bottleneck for the development of vocoder-based audio DeepFake detection method

ASV spoof
Automatic Speaker Verification and
Spoofing Countermeasures Challenge

Dataset configurations

- Eight types of vocoders
  - Autoregressive Models
    - WaveNet
    - WaveRNN
  - GAN Models
    - MelGAN
    - MB-MelGAN
    - Parallel WaveGAN
  - Diffusion Models
    - WaveGrad
    - DiffWave

- Data source

  LJSpeech: 13,100 short audio clips of a single speaker with length from 1 to 10 seconds (a total length of approximately 24 hours).
  LibriTSS: multi-speaker English corpus of approximately 585 hours of read English speech.
  CSTR VCTK Corpus: 110 English speakers of 400 sentences.

- Dataset sizes

  1,000 sample clips of average length of >10 seconds for each type of vocoders across diverse speakers and contents (total 8,000 samples, or 32 hours)
  We will also develop a baseline for vocoder identification based on the RawNet2 model

Research Plan

- Task 1: Set up and pilot SOTA vocoder models (3 months)
- Task 2: Generate 5000+ audio samples using the input speech and different vocoder models (6 months)
- Task 3: Develop baseline vocoder identification models based on the RawNet2 model (2 Months)
- Task 4: Summary results and drafting reports (1 Month)
- The dataset and benchmark will be made available to the media forensics research community as open-source upon the completion of the project
- Evaluation metrics for synthesis qualities
    - Mean Opinion Score (MOS)
    - Fréchet Audio Distance (FAD)
- Evaluation metrics for detection and attribution accuracies
    - The area under the ROC curve (AUC)

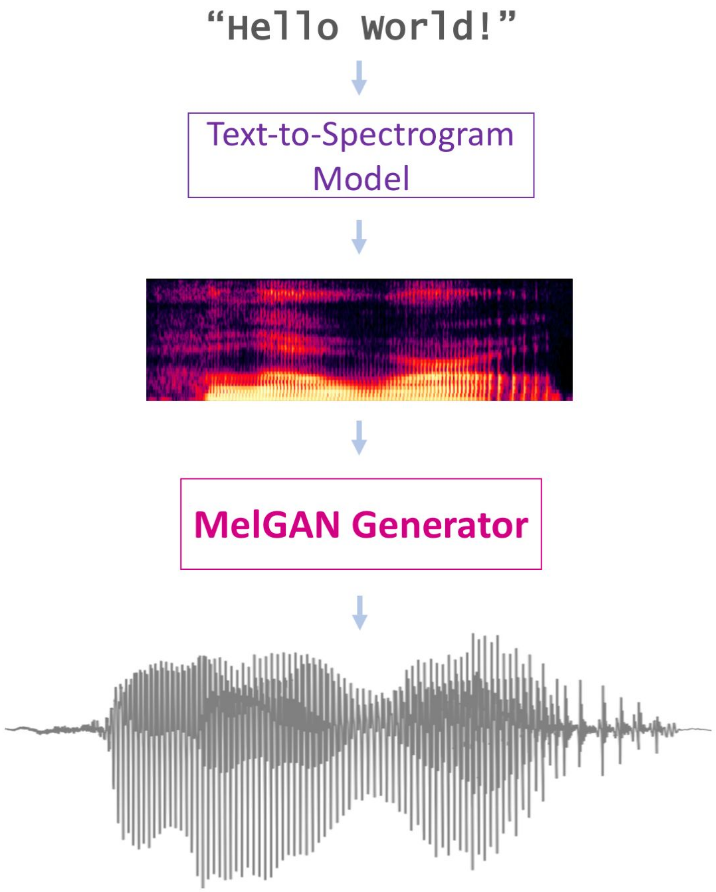
First Dataset we used

- LibriTSS:
- A multi-speaker English corpus
- 585 hours of reading English speech
- 24kHz sampling rate
- The LibriTTS is designed for TTS research.
- Subset of the original materials of the LibriSpeech.

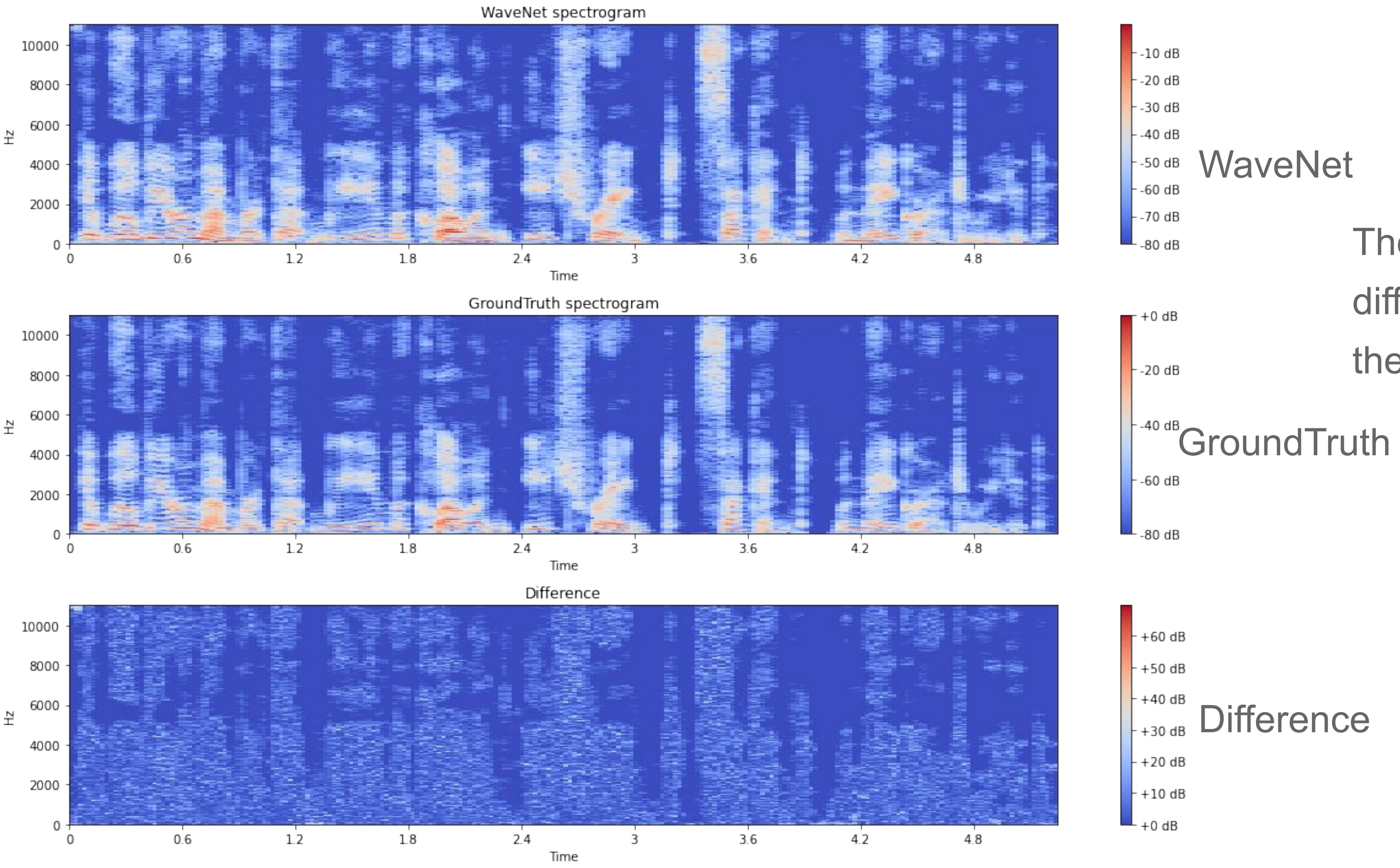
Audio Difference (one sample from LibriTTS)

MelGAN

GroundTruth

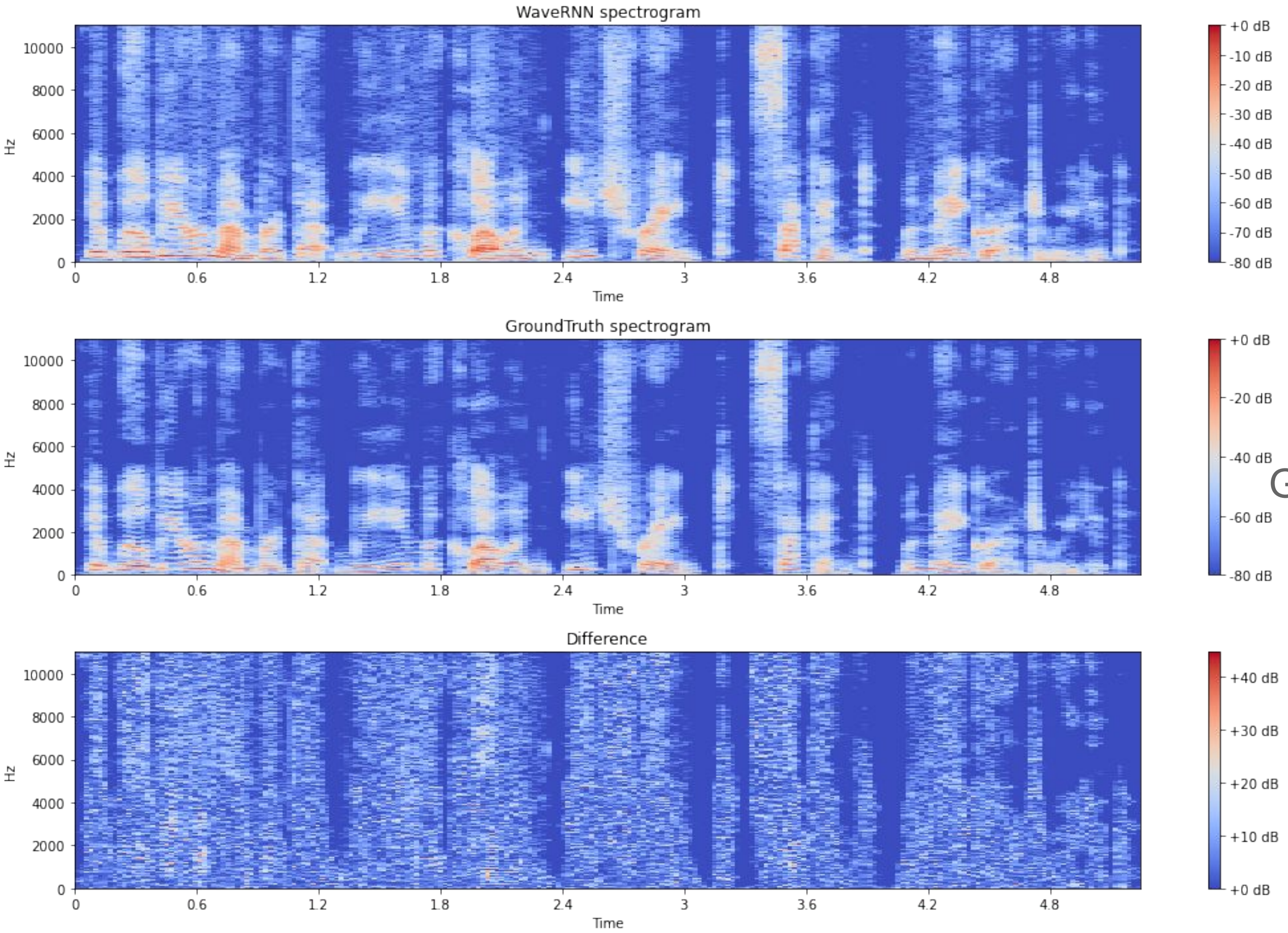


Project Overview: Spectrogram (one sample from LibriTTS)



The bottom graph shows the difference. The darker it has means the bigger difference it has.

Project Overview: Spectrogram (one sample from LibriTTS)



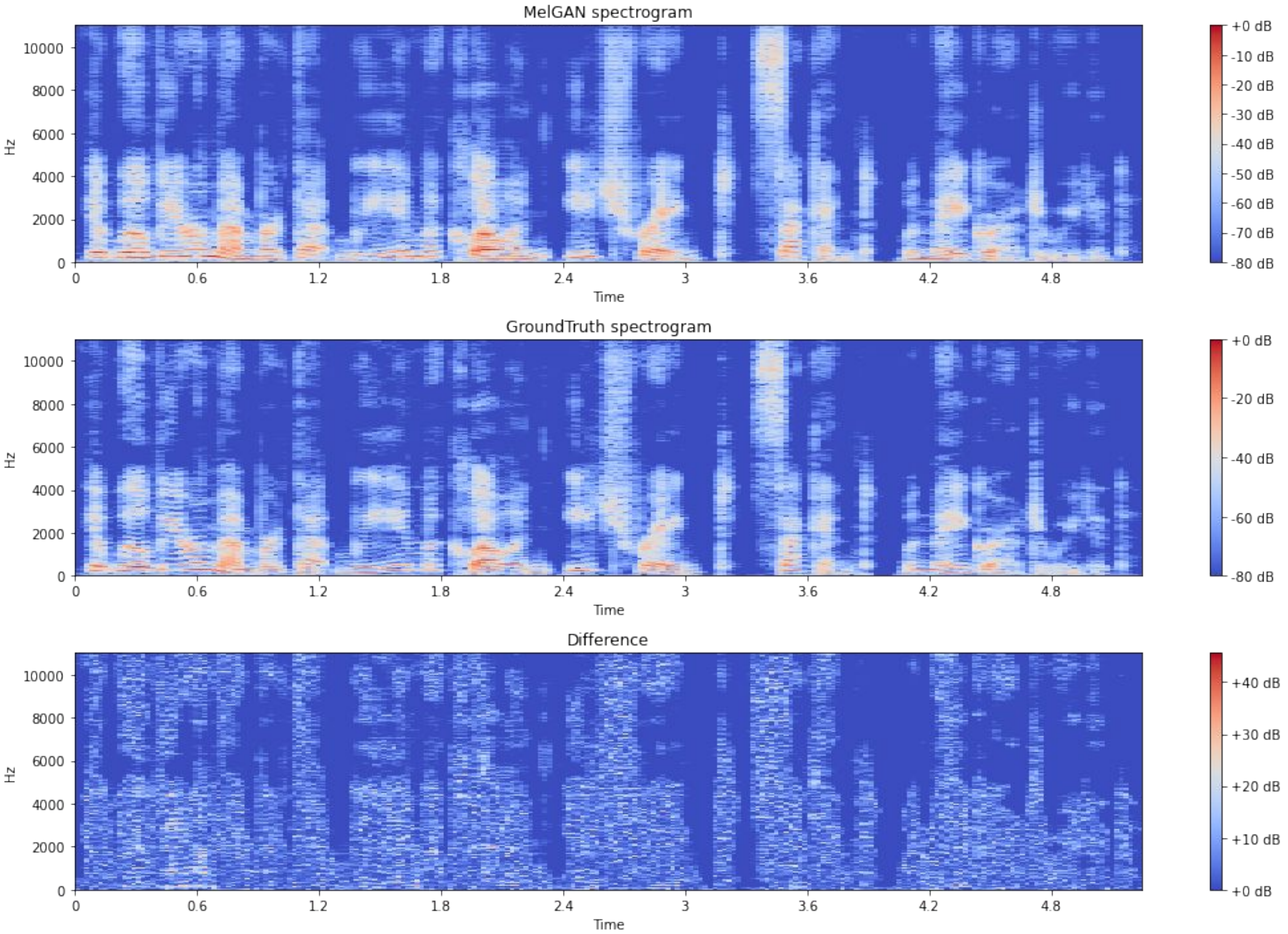
WaveRNN

The bottom graph shows the difference. The darker it has means the bigger difference it has.

GroundTruth

Difference

Spectrogram Difference (one sample from LibriTTS)



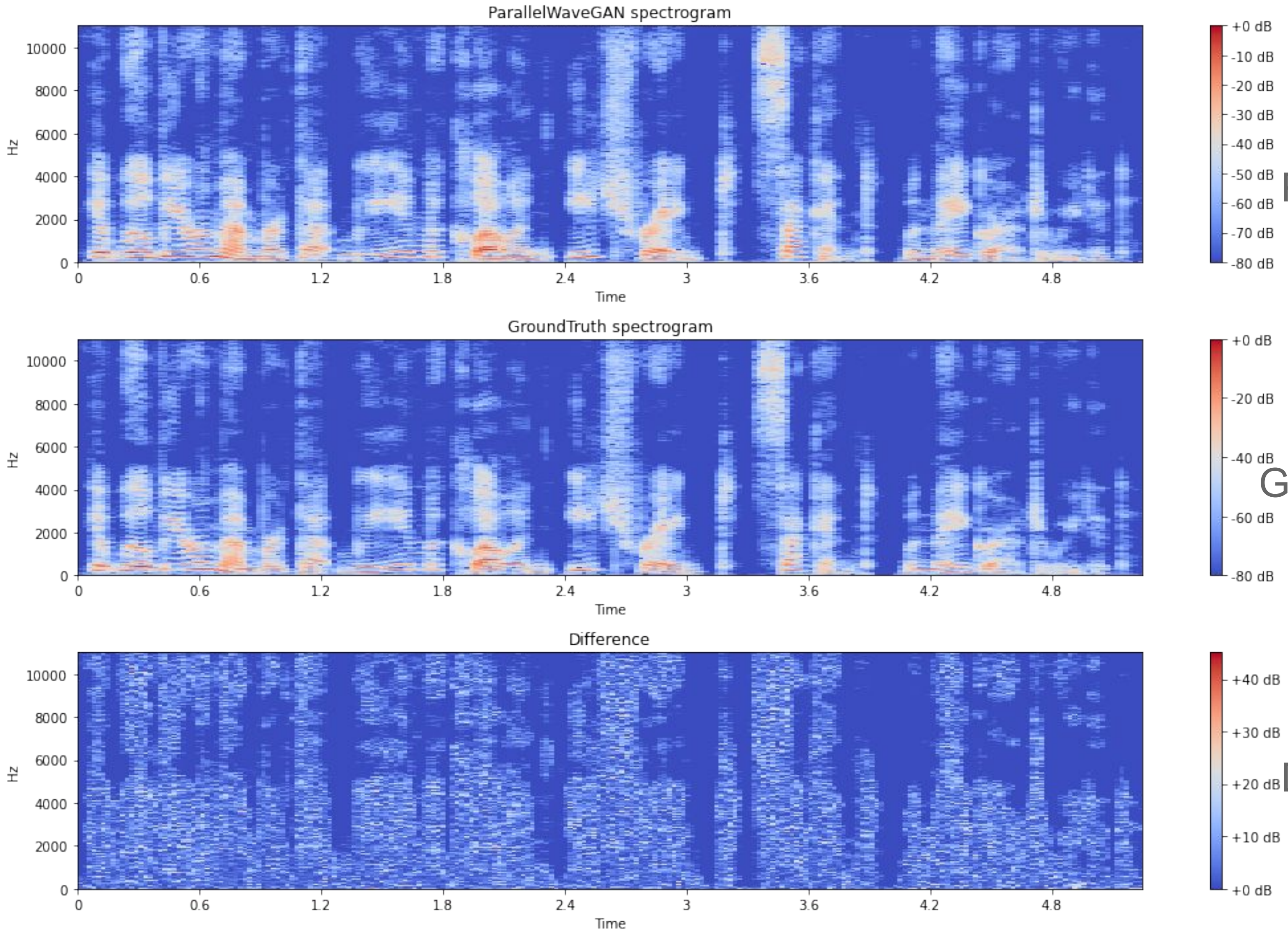
MelGAN

GroundTruth

Difference

The bottom graph shows the difference. The darker it has means the bigger difference it has.

Project Overview: Spectrogram (one sample from LibriTTS)



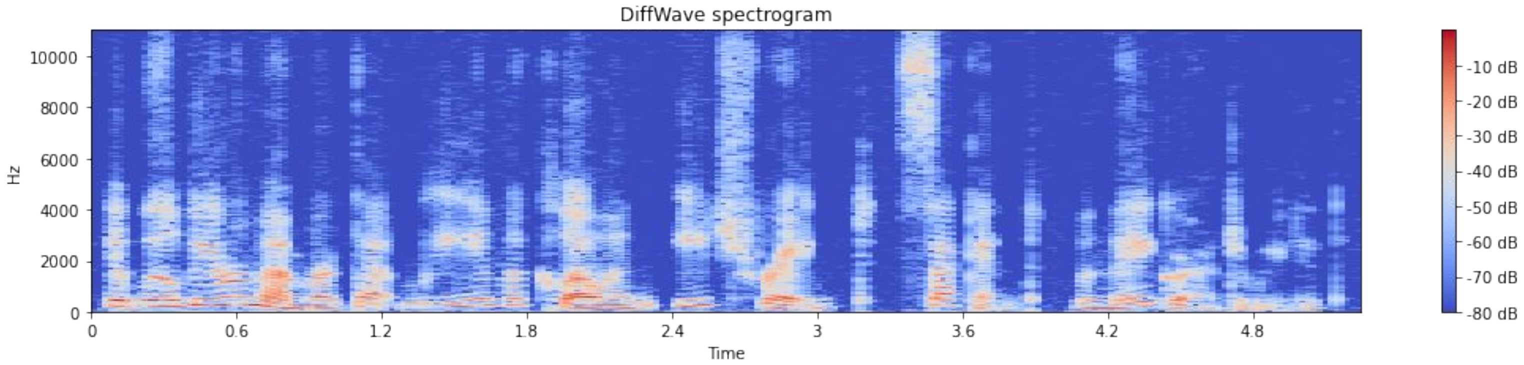
ParallelWaveGAN

The bottom graph shows the difference. The darker it has means the bigger difference it has.

GroundTruth

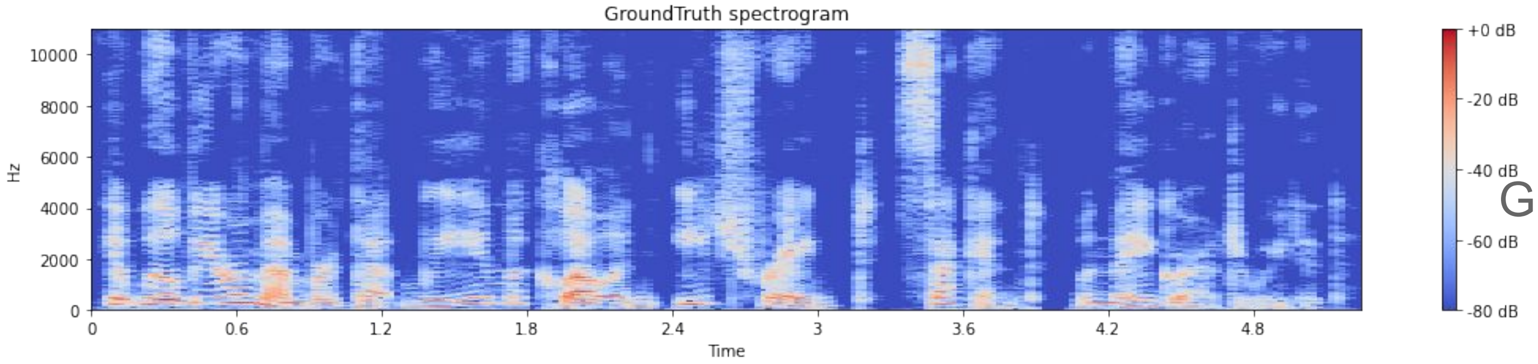
Difference

Project Overview: Spectrogram (one sample from LibriTTS)

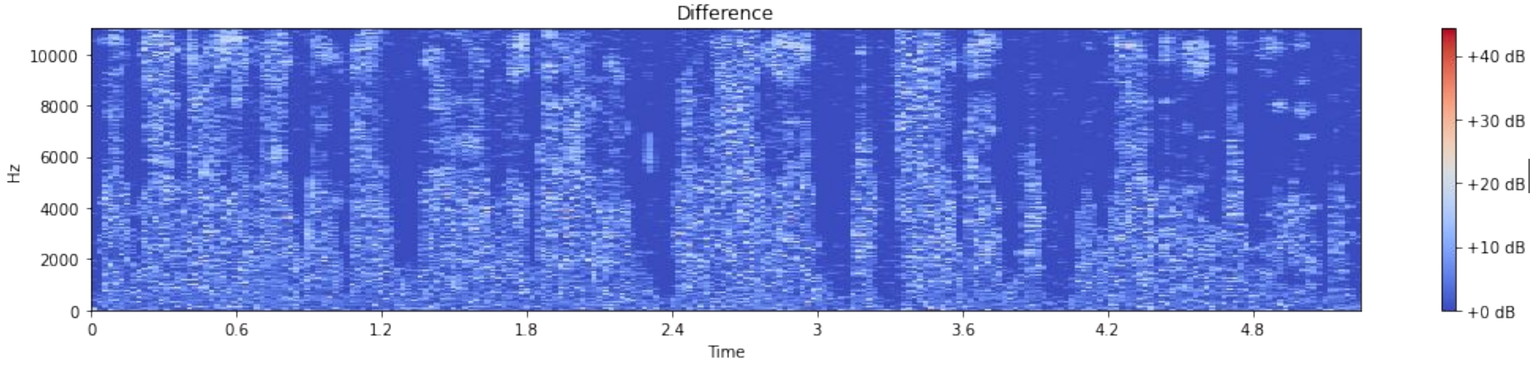


DiffWave

The bottom graph shows the difference. The darker it has means the bigger difference it has.

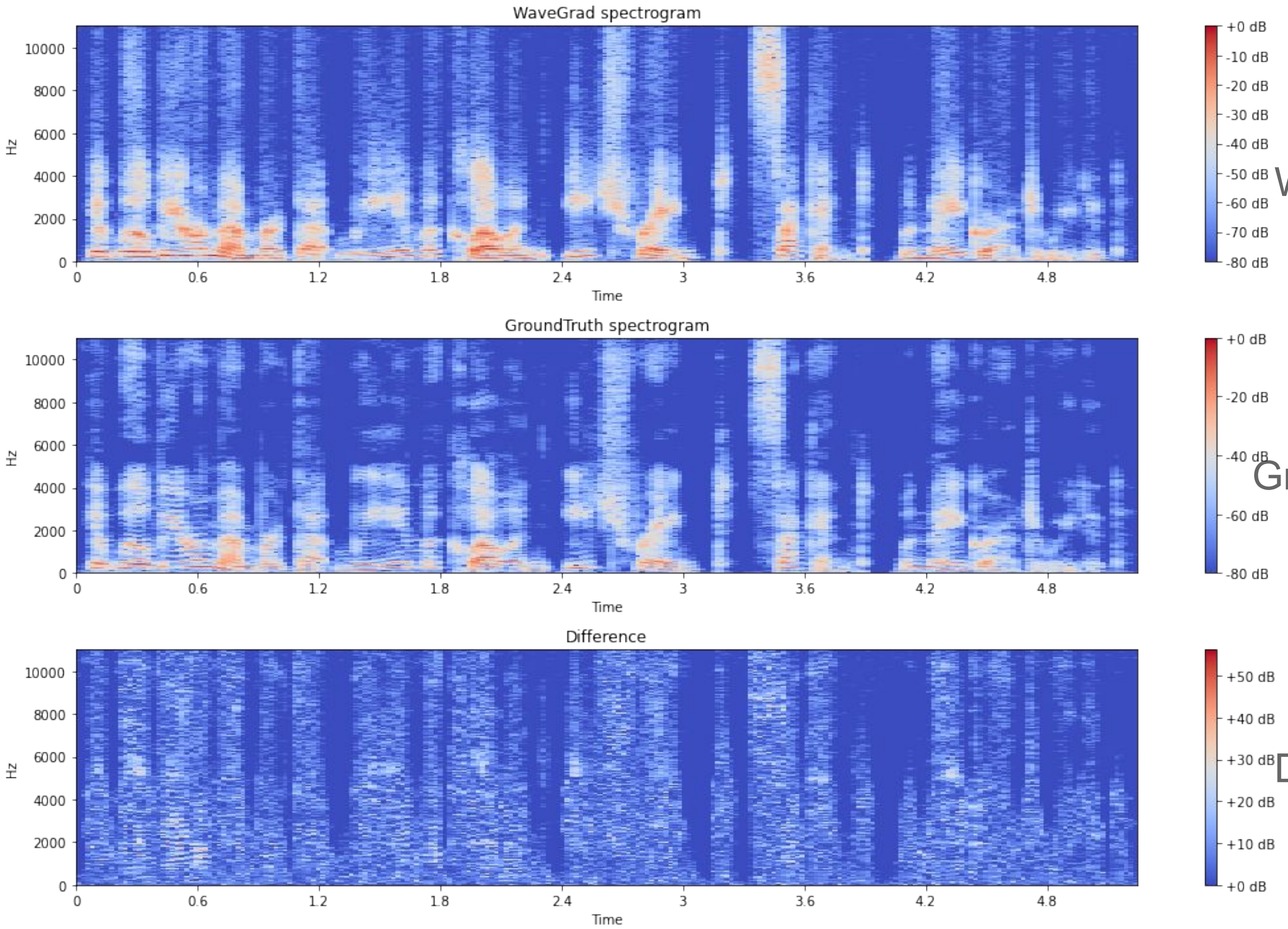


GroundTruth



Difference

Project Overview: Spectrogram (one sample from LibriTTS)



WaveGrad

GroundTruth

Difference

The bottom graph shows the difference. The darker it has means the bigger difference it has.

Dataset - LibriVoc Dataset

- We create LibriVoc as a new open-source, large-scale dataset for the study of neural vocoder artifact detection.
- LibriVoc is derived from the LibriTTS speech corpus.
- LibriTTS contains 585 hours of recorded speech samples from 2,456 speakers.
- LibriTTS corpus has been widely used in text-to-speech research.

Dataset - LibriVoc Dataset

Overall of the Dataset Size & Splits:

- Train

  Number of samples: 149736

  Number of speakers: 1151

- Develop

  Number of samples: 5736

  Number of speakers: 40

- Test

  Number of samples: 4837

  Number of speakers: 39

Dataset - LibriVoc Dataset

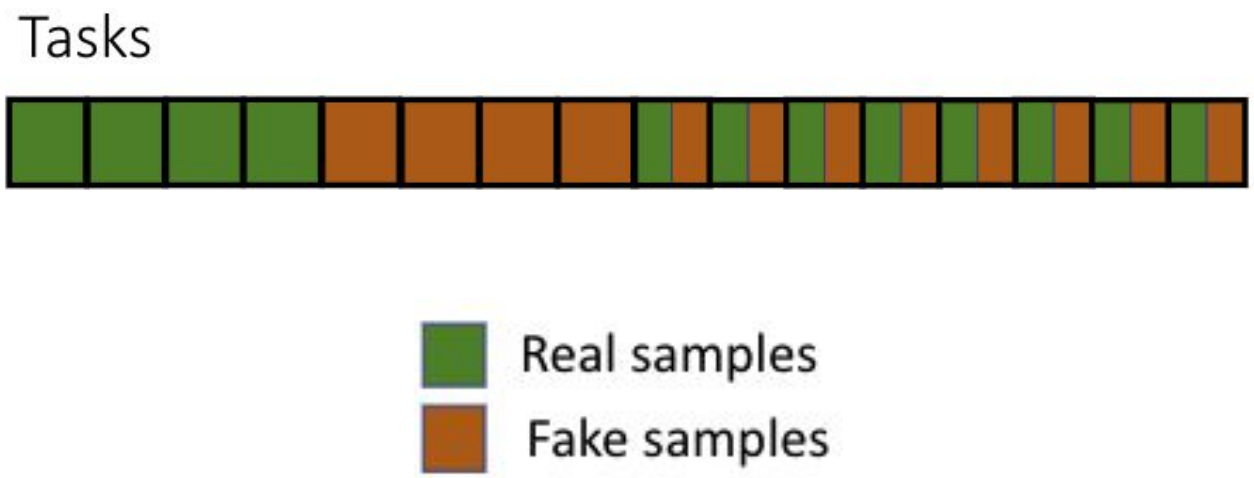
The number of hours of audio synthesized by each neural vocoder.

Model	train-clean-100	train-clean-360	dev-clean	test-clean
WaveNet (A01)	4.28	15.49	0.75	0.76
WaveRNN (A02)	4.33	14.92	0.67	0.72
MelGAN (G01)	4.36	15.26	0.71	0.76
Parallel WaveGAN (G02)	4.37	15.54	0.68	0.75
WaveGrad (D01)	4.19	15.81	0.76	0.74
DiffWave (D02)	4.16	15.37	0.62	0.66
Total	25.69	92.39	4.19	4.39

Dataset - LibriVoc Dataset

Organization of the dataset:

- Real & fake ratio 50/50
  - ¼ of the speakers will be reserved for real samples only
  - ¼ of the speakers will be reserved for fake samples only
  - ½ of the speaker will be a combination between real and fake samples
- Real data will be used to train the neural vocoders

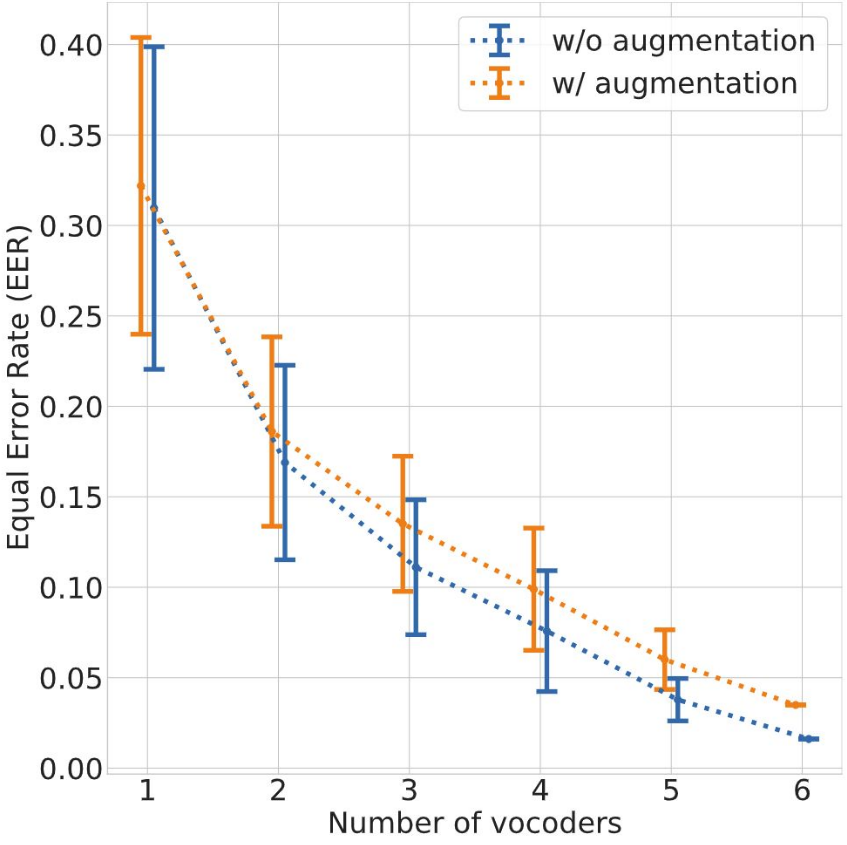


Vocoder Detection

- Our vocoder detection method is based on the recent RawNet2 model.
- RawNet2 is an end-to-end model that was originally designed for the automatic speaker verification anti-spoofing task.
- RawNet2 ranks among the best-performing baselines in the ASVspoof challenge.

Evaluation Results

- The experiment yielded an EER of 3.15% when using augmentation and a 2.69% EER without augmentation.
- RawNet2 classifier can robustly detect vocoder artifacts even despite additive noise.
- each neural vocoder does produce unique artifacts, akin to a signature or vocoder fingerprint



Summery

- We develop a model for vocoder identification based on the RawNet2 model.
- We also provide a large-scale dataset named LibriVoc, with synthetic audios of human voice samples created with a diverse set of neural vocoders.
- Experiments on this dataset show that our method can achieve an overall vocoder identification EER of 1.61%.
  There is still room for improvement for this work.

Future Plan

- We will form a new dataset using Voice conversion models.
- We would like to augment the LibriVoc dataset to include more diverse real audio signals and environments.
- We will further explore more tailored solutions to the vocoder identification problem.
- We will further develop effective methods that can directly differentiate real and synthetic audios by combining cues from vocoders and other signal features.

Outcomes, Importance, Deliverables

- Outcomes – *a large-scale dataset with synthetic audios of human voices created with a diverse set of neural vocoders, and a baseline vocoder identification algorithm*
- Importance - *the dataset will be useful to conduct research in DeepFake audio detection, especially those based on vocoder identifications*

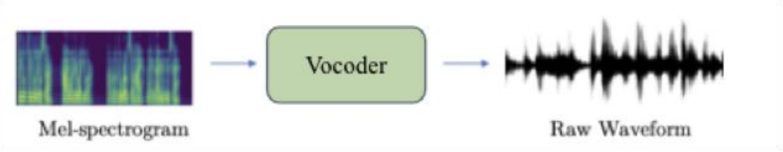
Deliverables	Delivery Date
Software: (To be provided to affiliates; or N/A) *The baseline vocoder identification algorithm*	**8 months after project begins**
Datasets: (To be provided to affiliates; or N/A) *The synthetic audio dataset created with different vocoders*	**6 months after project begins**
Other: (add rows as necessary) *Progress report and annual report with publications and presentations related to project*	**12 months after project begins**

Related Funding and IP

- Prior Funding
  - N/A
- Current Related Funding (indicate how projects are different)
  - PI Lyu and Doermann are currently supported by DARPA SemaFor Project (2020 - 2024); however, this work is not part of the PI's proposed work in the SemaFor project.
  - PI Lyu, Doermann, Setlur CITeR Project (2022 - 2023) #22S-01B A Benchmark Dataset for Neural Vocoder Identification
- Intellectual Property
- A prior IP to declare, e.g., provisional patent applications, patents, and licensing arrangements.
  - N/A
- Conflicts of interest (ownership, licensing, consulting payments, etc.) in the area of the proposal
  - N/A

A Benchmark Dataset for Neural Vocoder Identification
Project #22S-01B

Siwei Lyu, David Doermann, Srirangaraj Setlur (UB)



- The proliferation of DeepFake technologies poses clear threats to society and democracy
- Synthetic audio detection is one key element of managing this threat

Objective and Approach

- Objective

  We aim to develop methods to detect synthetic audios by identifying the neural vocoders used in the generation process
- Approach

  We will build a large-scale benchmarking dataset for vocoder identification

Outcomes

- a large-scale dataset with synthetic audios of human voices created with a diverse set of neural vocoders,
- a baseline vocoder identification algorithm

Deliverables

- Dataset
- Baseline identification algorithm code
- Reports and publications

Milestones (from proposal)

- Task 1: Set up and pilot SOTA vocoder models (3 months)
- Task 2: Generate 5000+ audio samples using the input speech and different vocoder models (6 months)
- Task 3: Develop baseline vocoder identification models based on LPCC features and GMM models (2 Months)
- Task 4: Summary results and drafting reports (1 Month)

Thank you very much for listening! Question?