DeepFake: TTS Interface

Aine Drelingyte, Maria Alejandra Cely, Danylo Shkrebko
Supervisors: Ajinkya Kulkarni, Miguel Couceiro

Natural Language Processing Department
IDMC
Université de Lorraine
13 Rue Michel Ney, 54000 Nancy
France

January 2024

1 Introduction

2 Background Research

3 Models

4 Model Training

5 Evaluation

6 Results

7 Interface

8 Discussion & Conclusion

9 References

Introduction

- **Objective:** To evaluate and compare the performance of three state-of-the-art vocoders—HiFi-GAN, Multi-band MelGAN, and WaveGrad—within an Extended Text-to-Speech (XTTS) system, focusing on audio quality, pitch accuracy, and environmental impact.
- **Interface:** Development of the Headspace Web Application=.

HEADSPACE

Step-by-step Methodology

Objective

The objective of this study is to evaluate and compare the performance of HiFi-GAN, Multi-band MelGAN, and WaveGrad vocoders within an Extended Text-to-Speech (XTTS) system, focusing on audio quality, pitch accuracy, environmental impact, and integrating the findings into the Headspace Web Application for practical application and user interaction.

VOCODERS	TTS
-Multi-band MelGAN -Wavegrad -HifiGAN	XTTS

1 Introduction

2 Background Research

3 Models

4 Model Training

5 Evaluation

6 Results

7 Interface

8 Discussion & Conclusion

9 References

Other Vocoder evaluation metrics

VOCBENCH: A Neural Vocoder Benchmark for Speech Synthesis

- **Corpus:** LJ Speech, LibriTTS, VCTK
- **Metrics:** SSIM, LS-MSE, PSNR, FAD, MOS.
- **Vocoders:** WaveNet, WaveRNN, MelGAN, Parallel WaveGAN, WaveGrad, DiffWave, Griffin-Lim.

1 Introduction

2 Background Research

3 Models

4 Model Training

5 Evaluation

6 Results

7 Interface

8 Discussion & Conclusion

9 References

XTTS

1. **Languages Support:** XTTS-v1 offers natural-sounding voices in 13 languages, including English, Spanish, French, German, and Chinese (Simplified), among others.
2. **Architecture:** built on Tortoise based on GPT-like auto-regressive acoustic model that converts: input text -> discretized acoustic tokens, diffusion model -> tokens to Mel spectrogram frames and a vocoder -> the spectrograms to the final audio signal.
3. **Performance:** Outputs 24khz audio, with specific handling for acronyms, numbers, and quality dependent on reference audio.

Vocoders

1. **Multi-BandMelGAN:** GAN architecture. It uses a generator network to synthesize waveform samples directly. Generates realistic waveforms. Utilizes a single shared network for sub-band signal predictions, minimizing computational complexity, and achieving high-quality speech synthesis with fewer model parameters.
2. **HiFiGAN:** GAN architecture. Specifically designed for high-fidelity speech synthesis (with naturalness).It can invert mel-spectrograms of unseen speakers. It employs various types of loss such as, GAN loss, mel-spectrogram loss, and feature matching loss, to improve training stability and sample quality
3. **WaveGrad:** Autoregressive model (preceding values are used to predict the present value) that conditions on mel-spectrograms. It generates waveform samples one at a time. Aims for high quality output.the model deals with tuning the noise schedule and determining the number of diffusion/ denoising steps, emphasizing their impact on sample quality and computational efficiency. Computationally demanding.

1 Introduction

2 Background Research

3 Models

4 Model Training

5 Evaluation

6 Results

7 Interface

8 Discussion & Conclusion

9 References

Dataset

- **Source:** CML-Multi-Lingual-TTS, derived from the Multilingual LibriSpeech (MLS) project, developed by the Center of Excellence in Artificial Intelligence (CEIA) at the Federal University of Goias (UFG).
- **Languages:** French, Spanish, German, and Dutch.
- **Purpose:** To provide a diverse range of linguistic characteristics.

Coqui TTS

Coqui TTS is an advanced, open-source Text-to-Speech (TTS) library developed by Coqui.

- **Open Source:** Fully open-source project encouraging community contributions.
- **High-Quality Speech Synthesis:** Uses state-of-the-art deep learning models for natural sounding speech.
- **Supports Multiple Languages:** Designed to be multi-lingual with support for various accents and voices.
- **Customizable:** Allows for the training of custom voices and fine-tuning on existing models.

GitHub Repository: https://github.com/coqui-ai/TTS

Parameters

```
audio_config = BaseAudioConfig(
    num_mels=80,
    fft_size=2048,
    sample_rate=24000,
    win_length=1024,
    hop_length=256,
    frame_length_ms=None,
    frame_shift_ms=None,
    preemphasis=0.0,
    min_level_db=-100,
    ref_level_db=20,
    power=1.0,
    griffin_lim_iters=60,
    log_func="np.log10",
    stft_pad_mode="reflect",
    signal_norm=True,
    symmetric_norm=True,
    max_norm=4.0,
    clip_norm=True,
    mel_fmin=0.0,
    mel_fmax=12000.0,
    spec_gain=20.0,
    do_trim_silence=False,
    trim_db=60
)

config = WavegradConfig(
    audio=audio_config,
    batch_size=32,
)
```

GPUs

Model	Node Specifications		
	Accelerators	Memory (GiB)	CPU Cores
HiFi-GAN	2 x Nvidia A40	256	24
Multi-band MelGAN	2 x Nvidia A40	256	24
Wavegrad	4 x Nvidia Tesla T4	128	16

Table: GPU Utilization

1 Introduction

2 Background Research

3 Models

4 Model Training

5 Evaluation

6 Results

7 Interface

8 Discussion & Conclusion

9 References

Mel Cepstral Distortion (MCD)

The Mel Cepstral Distortion (MCD) is a metric used to quantify the difference between two speech signals: the reference signal and the synthesized signal produced by a Text-to-Speech (TTS) system.

$$MCD = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{n=1}^{N}(c_n^{(t)} - \hat{c}_n^{(t)})^2} \tag{1}$$

- T is the total number of frames.
- N is the number of mel cepstral coefficients per frame.
- $c_n^{(t)}$ is the n^{th} coefficient of the reference signal at frame t.
- $\hat{c}_n^{(t)}$ is the n^{th} coefficient of the synthesized signal at frame t.

- **Lower MCD scores** indicate *less distortion* and *higher similarity* between the reference and synthesized speech, which is desirable.

Fundamental Frequency (F0) RMSE

The Fundamental Frequency (F0) Root Mean Square Error (RMSE) is a metric for evaluating the pitch accuracy of synthesized speech by comparing it to a reference speech signal.

$$F0_{\mathrm{RMSE}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(F0_{\mathrm{ref}}(t) - F0_{\mathrm{syn}}(t))^2} \tag{2}$$

- T is the number of frames.
- $F0_{\mathrm{ref}}(t)$ is the fundamental frequency at frame t of the reference audio.
- $F0_{\mathrm{syn}}(t)$ is the fundamental frequency at frame t of the synthesized audio.

- **Lower F0 RMSE scores** indicate *closer pitch matching* between the synthesized and reference speech, reflecting *better pitch accuracy*.

Subjective Evaluation: Mean Opinion Score (MOS)

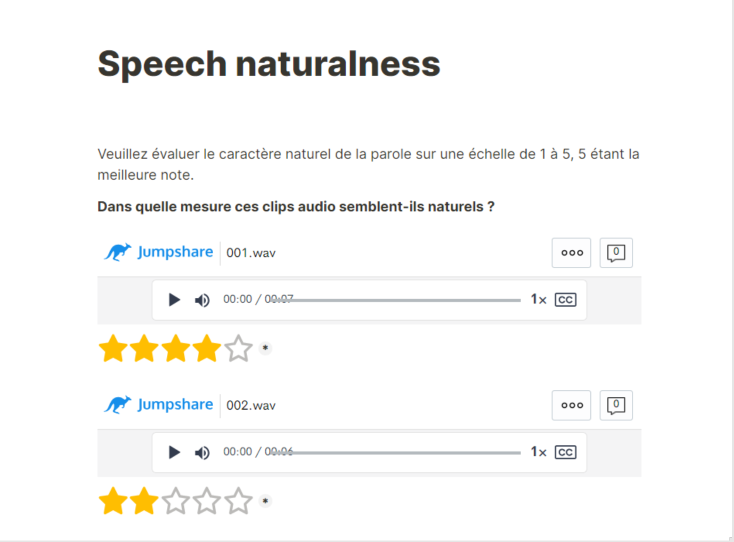
The Mean Opinion Score (MOS) test is a subjective metric for evaluating the naturalness of synthesized speech. Participants are asked to rate the quality of audio samples on a scale from 1 to 5, with the following characterizations:

Score	**Characterization**
1	Bad
2	Poor
3	Fair
4	Good
5	Excellent

Context:

- The MOS test will be conducted for each vocoder integrated into XTTS as well as the default XTTS.
- Participants, referred to as *naïve listeners*, are not experts in speech synthesis or telecommunications.

MOS Questionnaire



Speech naturalness

Veuillez évaluer le caractère naturel de la parole sur une échelle de 1 à 5, 5 étant la meilleure note.

Dans quelle mesure ces clips audio semblent-ils naturels ?

CodeCarbon: Estimating Computational Carbon Footprint

CodeCarbon is an open-source software tool that estimates the amount of CO2 emissions generated by the computing resources used during machine learning experiments or any intensive computational process.

How it Works:

1. Collects information about the hardware used, such as CPUs, GPUs, and memory.
2. Monitors the usage and power consumption during the computational task.
3. Calculates the carbon emissions using regional energy grids' CO2 equivalence.
4. Outputs detailed reports for transparency and accountability in sustainability.

Composite Score Calculation for Vocoder Selection

The best vocoder will be selected based on a composite score derived from weighted criteria: MCD, F0 RMSE, MOS, and Carbon Footprint.

Normalization:

- For "lower is better" (MCD, F0 RMSE, Carbon Footprint):

$$N_i = 1 - \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

- For "higher is better" (MOS):

$$N_i = \frac{X_i - \min(X)}{\max(X) - \min(X)}$$

Composite Score Calculation:

$$C_i = w_{\text{MCD}} \cdot N_{\text{MCD},i} + w_{\text{F0 RMSE}} \cdot N_{\text{F0 RMSE},i} + w_{\text{MOS}} \cdot N_{\text{MOS},i} + w_{\text{Carbon}} \cdot N_{\text{Carbon},i}$$

Sensitivity Analysis

Objective: To determine the influence of weight assignments on the Composite Score (CS) for vocoders Hifi-GAN, WaveGrad, and Multi-band MelGAN.
Steps:

1. **Initial Weight Assignments:** Assign equal weights to all metrics as a baseline for comparison.
2. **Adjust Weights:** Systematically vary the weights assigned to each metric to observe changes in vocoder rankings.
3. **Analysis:** Record the CS for each vocoder under different weight configurations.
4. **Optimization:** Based on observed impacts, propose optimized weights that reflect the significance of each metric.

1 Introduction

2 Background Research

3 Models

4 Model Training

5 Evaluation

6 Results

7 Interface

8 Discussion & Conclusion

9 References

Sensitivity Analysis

	HifiFR	WaveFR	MultiFR
CS with MCD top	0.7828	0.4	0.5854
CS with F0 RMSE top	0.7828	0.4	0.5054
CS with MOS top	0.7656	0.6	0.395
CS with Carbon top	0.5828	0.6	0.4892

Table: Results of Sensitivity Analysis

Weights

$$
\begin{aligned}
w_{\text{MCD}} &= 0.35 \\
w_{\text{F0 RMSE}} &= 0.25 \\
w_{\text{MOS}} &= 0.20 \\
w_{\text{Carbon}} &= 0.20
\end{aligned}
$$

Subjective\Objective Evaluation Results

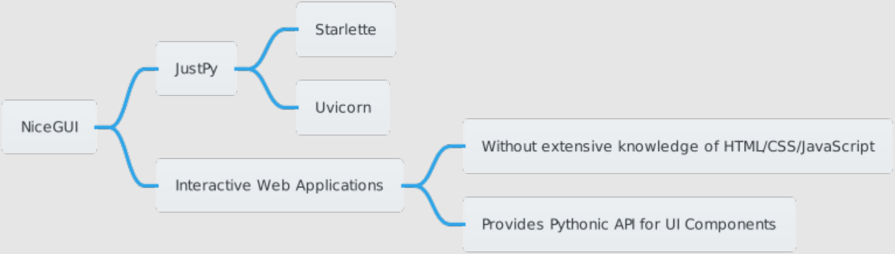
Vocoder	**MCD**	**F0 RMSE (Hz)**	**CO2 Emissions (kg)**	**MOS**
HiFi-GAN	16.7513	104.72	3.68	3.7234
WaveGrad	19.9211	127.06	2.98	3.8936
Multi-band MelGAN	16.9035	111.65	3.35	1.9149

Composite Scores

	Hifi-GAN	WaveGrad	Multi-band MelGAN
CS	0.7828	0.4	0.5654

1. Introduction
2. Background Research
3. Models
4. Model Training
5. Evaluation
6. Results
7. **Interface**
8. Discussion & Conclusion
9. References

NiceGUI framework



Proposed Solution

1. Introduction
2. Background Research
3. Models
4. Model Training
5. Evaluation
6. Results
7. Interface
8. Discussion & Conclusion
9. References

Future Research Directions

Investigating Linguistic Variability in TTS Quality

- **Comparative Linguistic Analysis:** Explore the quality difference between speech produced in English and French, focusing on:
  - Naturalness and intelligibility of speech across languages.
  - Phonemic and prosodic features specific to each language.
- **Impact of Dataset Diversity:** Investigate how the diversity in training datasets influences TTS quality, particularly regarding:
  - Accent variation within the same language.
  - The representation of minority languages and dialects.

Key Takeaways

- **HiFi-GAN Performance:** HiFi-GAN stands with the highest composite score.
- **Headspace Web Application:** Offers a user-friendly TTS platform that showcases the capabilities and differences between various vocoders.
- **Advancement in TTS Technology:** Highlights the progress in TTS technology while stressing the significance of environmental considerations in AI development.
- **Framework:** Employs for a comprehensive evaluation framework that prioritizes both performance and sustainability.

1. Introduction
2. Background Research
3. Models
4. Model Training
5. Evaluation
6. Results
7. Interface
8. Discussion & Conclusion
9. References

References

- Gan chart, "https://semiengineering.com/knowledge centers/artificial-intelligence/neural-networks/generative-adversarial-network-gan/"
- Vocoder Architecture, "https://medium.com/@bigpon517/2020-speech-generation-0-vocoder-and-rnn-and-cnn-based-speech-waveform-generative-models-c324c88e789a"
- Text To Speech Frameworks, "https://towardsdatascience.com/text-to-speech-foundational-knowledge-part-2-4db2a3657335"