Comparative Evaluation of HiFi-GAN, Multi-band MelGAN, and WaveGrad Vocoder Performance in XTTS

Maria Alenjandra CELY LATORE, Aine DRELINGYTE, Danylo SHKREBKO
Supervisors: Miguel COUCEIRO, Ajinkya KULKARNI
Université de Lorraine, IDMC

Abstract

This study evaluates three vocoders—HiFi-GAN, Multi-band MelGAN, and WaveGrad—from Coqui TTS within an Extended Text-to-Speech (XTTS) system. Focused on determining the best fit for the "Headspace" TTS interface, the evaluation employs Mel Cepstral Distortion (MCD), Fundamental Frequency (F0) RMSE, Mean Opinion Score (MOS), and carbon footprint as metrics. A key feature is the use of a formula integrating these scores with assigned weights, providing a balanced assessment of each vocoder's quality, naturalness, and environmental impact. This approach aids in selecting the most suitable vocoder for TTS technology, considering both performance and sustainability.

1 Introduction

The surge in advanced vocoder technologies, particularly within the realm of Text-to-Speech (TTS) systems, has brought about a significant shift in the landscape of digital media and security, largely due to the rising phenomenon of deepfake audios. The development of highly sophisticated vocoders like HiFi-GAN, Multi-band MelGAN, and WaveGrad has made it possible to create extremely realistic synthetic speech, which, while innovative, also poses a substantial challenge in distinguishing authentic from manipulated audio content. This scenario underscores an urgent need for effective methods to detect deepfake audios, a task that is becoming increasingly complex as vocoder technologies evolve.

In the face of this challenge, the role of deepfake audios themselves becomes paradoxically crucial. To develop robust detection algorithms, there is a growing demand for large datasets of deepfake audio samples, which can be used to train machine learning models to recognize and differentiate between real and synthetic speech. The generation of these datasets necessitates the use of high-quality vocoders, making their evaluation and improvement not just a matter of enhancing speech synthesis but also a critical component in the fight against digital misinformation.

However, alongside the technological and ethical challenges posed by deepfakes, there is an often neglected aspect – the environmental impact of these AI-driven technologies. The energy-intensive nature of training and operating advanced vocoders contributes significantly to energy consumption and associated Carbon Dioxide (CO2) emissions. While tackling the deepfake problem calls for improved quality and realism in vocoder outputs, it also demands a responsible approach to the environmental footprint of these technologies.

Despite the growing importance of this aspect, research on the environmental impact of TTS vocoders remains sparse. Studies such as "Energy and Policy Considerations for Deep Learning in NLP" by Strubell et al., 2019 and "Green AI" by Schwartz et al.,2019 highlight the energy consumption patterns of AI models but offer limited insights into the specific context of TTS vocoders. Our study addresses this gap by integrating environmental impact assessments, particularly focusing on CO2 emissions, into the evaluation framework for TTS vocoders. This comprehensive approach not only confronts the quality and authenticity issues raised by deepfakes but also aligns with the imperative for environmentally sustainable AI development. In doing so, our study provides a holistic framework for evaluating TTS vocoders, crucial for ethical and sustainable progress in the field of speech synthesis amidst the challenges posed by deepfake technologies.

2 Background

Text-to-Speech (TTS) technology has undergone significant evolution, primarily driven by machine learning and deep neural networks. It plays a pivotal role in converting text inputs into natural and expressive speech, impacting various applications such as accessibility for visually impaired individuals, enhancing virtual assistants, and aiding language learning (Amezaga and Hajek, 2022). However, these advancements come with challenges, including the precise generation of prosody, conveying emotions, handling linguistic nuances, and accommodating variations in human speech. TTS systems also grapple with non-standard words, acronyms, and ambiguities, with limited linguistic resources posing a constraint, especially for sparsely spoken languages (Amezaga and Hajek, 2022). Continual research, evaluation, and improvement are essential to address these multifaceted challenges (Zhou et al., 2024).

In parallel with TTS, neural-based vocoders have emerged as integral components of modern TTS systems, revolutionizing speech waveform generation. These vocoders, categorized into various types, including autoregressive, flow-based, GAN-based, VAE-based, and diffusion-based (Tan et al., 2021), rely on neural networks to produce high-quality speech waveforms. By directly converting linguistic features into speech waveforms, neural-based vocoders minimize the need for human preprocessing and feature engineering, ultimately enhancing the quality of synthesized speech (Tan et al., 2021). This collaborative synergy between TTS and vocoders has led to the development of end-to-end TTS approaches, streamlining the speech synthesis pipeline and improving overall efficiency and naturalness (Tan et al., 2021).

Our project uniquely integrates carbon emissions evaluation into the assessment of TTS systems, differentiating it from prior studies. In addition to subjective and objective evaluations commonly employed in the field, we consider the environmental impact by evaluating carbon emissions. This holistic approach aims to provide more comprehensive and responsible results for selecting the optimal vocoder for our website. We also draw inspiration from studies that have combined metrics MCD, MOS, F0-RMSE to assess speech synthesis quality, ensuring a well-rounded evaluation of vocoders (Achanta et al., 2016; Luo et al., 2017; Sivaguru et al., 2023).

3 Models

3.1 XTTS

In our project we implemented the pre-trained XTTS (xtt). This is a Text-to-Speech model that is able to clone voices in different languages by using just a 6-second audio clip. It is built on Tortoise which is based on a GPT-like auto-regressive acoustic model that converts input text to discretized acoustic tokens, a diffusion model that converts these tokens to Mel spectrogram frames and a vocoder to convert the spectrograms to the final audio signal.

XTTS is used for cross-language voice cloning, Multi-lingual speech generation, Emotion and style transfer and it does not demand a huge amount of training data. It has a 24khz sampling rate. Finally, its last version (v2) supports 16 languages: English (en), Spanish (es), French (fr), German (de), Italian (it), Portuguese (pt), Polish (pl), Turkish (tr), Russian (ru), Dutch (nl), Czech (cs), Arabic (ar), Chinese (zh-cn), Japanese (ja), Hungarian (hu) and Korean (ko).

3.2 WaveGrad

A non-autoregressive generative model conditioned on the continuous noise level instead of discrete indices, providing a more distinctive control during training and inference. This vocoder is further developed as a diffusion probabilistic model, showcasing the conditional distribution based on latent variables and a diffusion process. WaveGrad introduces two variants: one conditioned on a discrete refinement step index and another on a continuous scalar indicating the noise level. The continuous variant proves more effective, offering flexibility during inference with varying refinement steps.

Finally, experimental results demonstrate WaveGrad's capability to generate high-fidelity audio samples, having a greater performance than adversarial non-autoregressive models and outdoing autoregressive models in subjective naturalness. Finally, 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 (Chen et al., 2020).

3.3 GAN-based vocoders

Generative Adversarial Nets (GANs) offer a framework for training generative models. It consists of a generator and a discriminator. Its training objective is to optimize the ability of the discriminative model to distinguish between real and generated samples, while the generator aims to produce indistinguishable samples from real data. Among the advantages of using GANs, they avoid Markov chains which rely on back-propagation for gradient computation, they prevent over-fitting and ensure training stability by utilizing an alternating optimization strategy that involves multiple optimization steps for the discriminator and a single step for the generator. However, challenges include maintaining diversity in generated samples and keeping careful synchronization between the generator and discriminator during training (Goodfellow et al., 2014).

Multi-Band MelGAN:

Multi-Band MelGAN (MBMelGAN) is a generative adversarial network (GAN) developed for efficient waveform generation in audio tasks. It is an improvement over MelGAN and it combines features such as expanding the receptive field, replacing feature matching loss with multi-resolution Short-Time Fourier Transform (STFT) loss, and pre-training for better stability. MBMelGAN utilizes a single shared network for sub-band signal predictions, minimizing computational complexity, and achieving high-quality speech synthesis with fewer model parameters (Yang et al., 2020).

Hifi-GAN:

HiFi-GAN is a type of generative adversarial network (GAN) specifically designed for efficient and high-quality speech synthesis. HiFi-GAN consists of a generator and a discriminator with small sub-discriminators. The generator is a convolutional neural network that takes mel-spectrograms as input and utilizes a multi-receptive field fusion module to observe pattern lengths in parallel. The discriminator represents periodic patterns on specific periodic parts of raw waveforms in speech audio. The model achieves superior performance and quality compared to others, such as WaveNet and WaveGlow, and even it can invert mel-spectrograms of unseen speakers. During training, HiFi-GAN employs various types of loss such as, GAN loss, mel-spectrogram loss, and feature matching loss, to improve training stability and sample quality (Kong et al., 2020).

4 Methodology

The methodology employed in this research project is structured into several well-defined stages, with each stage being thoroughly elaborated upon in its respective subsection. This deliberate organization is intended to facilitate a clear and comprehensive exposition of the research processes undertaken.

- Data Preparation: Preparing the data for subsequent analysis and model training.
- Vocoder Training and integration with XTTS
- Subjective Evaluation: Gathering subjective Evaluation, offering insights into the human perception aspects of the study.
- Objective Evaluation: Detailed Objective Evaluation methods and metrics used to quantitatively assess the performance of the vocoder.
- Carbon Footprint Analysis: In consideration of environmental sustainability, a comprehensive analysis of carbon footprint was conducted.
- Analysis and Comparison: Analysis of results and discussion based on results-comparison providing a deeper understanding of the research outcomes.
- Interface Creation and Integration: Creation of user-interface and the interfaces integration, which consist the last part of this project

5 Vocoder Training

5.1 Dataset

The datasets used in this project were sourced from the CML-Multi-Lingual-TTS (Oliveira et al., 2023), a comprehensive TTS dataset developed by the Center of Excellence in Artificial Intelligence (CEIA) at the Federal University of Goias (UFG). This resource, derived from the Multilingual LibriSpeech (MLS) project, offers a rich compilation of audiobook recordings in seven languages: Dutch, French, German, Italian, Portuguese, Polish, and Spanish. The creation of the CML-Multi-Lingual-TTS dataset was driven by the objective to expand research opportunities in TTS technologies, particularly focusing on the development and enhancement of multilingual models.

The selected datasets for this project encompassed recordings in French, Spanish, German, and Dutch. Each language segment contributed five hours of audio, ensuring a comprehensive exposure of the vocoders to a diverse range of linguistic characteristics, including phonetics, intonation, and rhythm. This approach not only aligns with the global applicability of TTS technologies but also fosters advancements in creating more inclusive and versatile TTS models.

6 Subjective Evaluation

The subjective evaluation of synthesized speech plays a crucial role in understanding how effectively a text-to-speech (TTS) system can communicate with its intended audience. Given the diversity of listeners and their varying abilities to discern nuances in speech quality, choosing the right group of listeners for evaluation is a key factor in obtaining meaningful insights. This section outlines our approach to conducting subjective listening tests, focusing on the selection of naïve listeners to gauge the average user's perception of synthesized speech quality. We justify our choice of listener type and introduce the Mean Opinion Score (MOS) as our chosen standardized measure for assessing speech quality. The methodology employed for collecting and analyzing listener feedback, including our sampling technique and the structure of the questionnaire, is detailed.

6.1 Mean Opinion Score (MOS)

To assess subjectively a standardized measure is needed and our and for our project we chose the MOS. It measures speech quality based on the judgement of a group of human listeners who have to rate the quality of the synthesized speech samples on a scale of 1 to 5, with 1 being poor and 5 being excellent and then obtain the average score. This can be interpreted as, the higher the value of MOS score, the better the perceived quality of the synthesized speech (Streijl et al., 2016). This score is widely used to evaluate TTS systems and it is said to help improving the audio quality which can lead to a more natural and effective communication between humans and machines (Maiti et al., 2023).

In our project, we created a questionnaire which consisted of a prompt, the audio clips and a 5-star rating scale. The prompt said:"Please, rate the naturalness of speech in a scale from 1 to 5, being 5 the best score. The four audio clips corresponded to our three vocoders and the human version, obviously the label of each clip was hidden to avoid bias.

We administered the questionnaire to 47 naive listeners that we managed to obtain using snowball sampling technique. The initial seed were our classmates who would be either French native speakers or anyone else who have at least B2 level in the language. We asked to share it with people with the same characteristics. It is important to mention that the data collection was done anonymously and we did not request any additional or personal information.

7 Objective Evaluation

Objective evaluation plays a pivotal role in the development and refinement of text-to-speech (TTS) systems, offering quantifiable measures that can reliably predict the perceived quality of synthesized speech. Unlike subjective evaluations that rely on human judgment, objective metrics provide a consistent and replicable means of assessing speech quality. This section delves into the specifics of objective evaluation methods used in our study, focusing on the Mel Cepstral Distortion (MCD) and the fundamental frequency Root Mean Square Error (F0 RMSE). These metrics are critical for evaluating the acoustic similarity and pitch accuracy of synthesized speech in comparison to natural speech.

7.1 Mel Cepstral Distortion (MCD)

It is a measure of how different two sequences of mel cepstra are, in other words, it is the acoustic similarity between two speech utterances, in our case the reference signal and the synthesized signal produced by a TTS system (Sivaguru et al., 2023). It serves as an objective measure of TTS performance that correlates well with subjective measures of speech quality (Kominek et al., 2008).

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

In which

- 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.

In this formula, the smaller the MCD between the synthesized and natural mel cepstral sequences, the less distortion and the closer the synthetic speech is to be similar to the natural speech.

7.2 F0 RMSE

Fundamental frequency refers to the approximate frequency of the (quasi-)periodic structure of voiced speech signals. In other words, this is defined as the average number of significant fluctuations per second of the vocal folds which are expressed in Hertz. F0 changes continuously within a sentence and it can be used for expressive purposes to signify, for example, emphasis and questions (spe). F0-RMSE serves as an objective measure used to evaluate the pitch accuracy of synthesized speech by comparing it to a reference speech signal. It is obtained by taking the root mean square error of F0 contour predicted by a TTS system compared to the target f0 contour of the reference speech (Sivaguru et al., 2023).

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

In which

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

In this formula, lower F0 RMSE scores indicate closer pitch matching between the synthesized and reference speech, reflecting better pitch accuracy.

8 CO2 Emissions

In our project, we dedicated a section to estimate the carbon dioxide (CO2) emissions linked to the computational resources utilized during the training of each vocoder. To achieve this, we employed the CodeCarbon library, a lightweight and open-source software package designed for estimating the amount of CO2 produced by computing resources used to execute code. The estimation process involved continuous training of the vocoders for a set number of epochs, specifically 10 epochs in our case. It is important to note that the use of different GPUs may introduce variability in power consumption and, consequently, in CO2 emissions. To ensure a fair comparison, we averaged the emissions across the 10 epochs and adjusted the calculations based on the number of epochs. It is worth highlighting that the carbon footprint calculation was specifically performed for continuous training until the vocoder's quality reached a point where further training did not yield significant improvements. Furthermore, our analysis focused on this continuous training scenario and did not include additional experiments to better compare the environmental impact of different training strategies or setups, providing a baseline assessment of emissions associated with each vocoder's training process.

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 1: GPU Utilization

9 Composite Score Evaluation

9.1 Normalization

In our analysis of vocoders' performance, we adopted a comprehensive evaluation approach encompassing various metrics, each contributing to the overall assessment. These metrics included MCD, F0 RMSE, MOS, and Carbon Footprint. To ensure fair comparisons across different metrics, we applied min-max normalization. For metrics where lower values indicated better performance, such as MCD, F0 RMSE, and Carbon Footprint, we utilized the following normalization formula:

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

Conversely, for MOS, where higher values signified superior quality, the formula was adapted as follows:

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

By employing these min-max normalization techniques, we transformed each metric into a common scale, facilitating unbiased comparisons. Subsequently, we computed the Composite Score Calculation (CSC) proposed below, through weighted summation, integrating all metrics into a unified score. This approach enabled us to select the best-performing vocoder, considering both audio quality and environmental impact.

9.2 Composite Score Calculation

We propose the following formula for CSC as a comprehensive metric to evaluate the performance of vocoders:

$$\text{CSC} = w_a \cdot A + w_b \cdot B + w_c \cdot C + w_d \cdot D \qquad (3)$$

This formula represents the summation of dot products between the weights assigned to each measure and their corresponding normalized values. The CSC calculation allows us to determine the final score, aiding in the selection of the best-performing vocoder. The variables used in this formula are defined as follows:

CSC	: The Composite Score Calculation
w_a	: Weight for MCD
w_b	: Weight for F0 RMSE
w_c	: Weight for MOS
w_d	: Weight for Carbon Footprint
A	: Normalized value of MCD for item i
B	: Normalized value of F0 RMSE for item i
C	: Normalized value of MOS for item i
D	: Normalized value of Carbon Footprint

This proposed formula allows for a unified assessment of vocoders, taking into account multiple evaluation metrics and their respective weights assigned based on their importance in the overall evaluation.

9.3 Sensitivity Analysis

To conduct sensitivity analysis, for each case, we systematically assign different weights to the metrics in the Composite Score Calculation (CSC) formula, while keeping the total sum of weights constant. In the first case, we will assign a weight of 0.40 to one metric, emphasizing its significance, and distribute weights of 0.20 to the other three metrics. This will allow us to evaluate the impact of prioritizing one specific aspect of vocoder performance.

In subsequent cases, we will rotate the weight of 0.40 to a different metric, while redistributing weights of 0.20 to the remaining metrics. By doing so, we can comprehensively assess the influence of each metric on the overall assessment. This sensitivity analysis will provide valuable insights into the robustness of our evaluation framework and help us fine-tune our approach to ensure that our conclusions align with the varying priorities that may be assigned to different aspects of vocoder performance.

10 Results and Discussion

10.1 Objective and Subjective Evaluations

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

Table 2: Vocoder Evaluation Metrics

The results of both subjective and objective evaluations provide valuable insights into the performance of the vocoders under consideration. In terms of objective metrics, the vocoders exhibit variations in their performance. Vocoder Hifi-GAN demonstrates the lowest MCD at 16.7513, indicating a relatively lower level of distortion in the generated audio. However, WaveGrad exhibits a higher MCD of 19.9211, suggesting a higher distortion. When evaluating the F0 RMSE, however, Hifi-GAN stands out with the lowest value of 104.72 Hz, indicating better pitch accuracy compared to WaveGrad, which has a higher F0 RMSE of 127.06 Hz. In terms of environmental impact, measured by CO2 emissions, WaveGrad outperforms the others, emitting only 2.98 kg of CO2 compared to Hifi-GAN's 3.68 kg and Multi-band MelGAN's 3.35 kg. However, when considering the subjective assessment using MOS, WaveGrad receives the highest score of 3.8936, indicating superior audio quality, while Multi-band lags behind with a significantly lower MOS score of 1.914. These results highlight the trade-offs between objective and subjective evaluations, where

WaveGrad excels in audio quality but may not necessarily excel in all objective metrics, emphasizing the importance of a composite evaluation approach.

10.2 Composite Score and 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 3: Results of Sensitivity Analysis

The sensitivity analysis has yielded insights into the influence of weight assignments on the Composite Score (CS) across three distinct vocoders: Hifi-GAN, WaveGrad, and Multi-band MelGAN. In pursuit of balanced CS score calculations, we have proposed weights based on their observed impact. Notably, MCD emerged as a key contributor to enhanced CS scores for Hifi-GAN and Multi-band MelGAN, underscoring its significance in assessing audio quality. As a result, a weight of 0.35 has been assigned to MCD. F0_RMSE exhibited a positive influence on CS scores, warranting a weight of 0.25 to maintain equilibrium across metrics. Meanwhile, Mean Opinion Score (MOS) and Carbon Footprint had discernible effects on CS scores for the WaveGrad vocoder but not the others, justifying weights of 0.20 for both MOS and Carbon. These optimized weight assignments aim to ensure a well-balanced evaluation approach tailored to the unique characteristics of each vocoder.

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

11 Final Scores

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

Table 4: Composite Score

The final evaluation results, as determined by the Composite Score (CS), showcase distinct performance levels among the vocoders. Hifi-GAN stands out with the highest CS of 0.7828, emphasizing its superior overall performance. WaveGrad follows with a CS of 0.4, indicating a moderate performance level, while Multi-band MelGAN lags slightly behind with a CS of 0.5654. These results suggest that HifiFR is the most favorable choice among the three vocoders, delivering the best combination of audio quality and environmental impact, making it a strong candidate for integration into our website. However, further considerations, such as specific project requirements and resource constraints, should also play a role in the ultimate selection of the most suitable vocoder.

12 Interface

HeadSpace Web Application is a service designed for text-to-speech synthesis. This solution offers a seamless user experience, allowing users to upload voice samples as reference, input text to voice and choose from a selection of pre-trained vocoders for this purpose. Below, we provide an overview of its key components and technical details.

12.1 Functionality

User interface facilitates the synthesis of speech audio, offering a convenient comparison between the reference and generated samples. Users can upload voice samples, input text and select a language and vocoder for synthesis.

- *TTS Tab:*
  - Upload voice sample, input text, select language, and vocoder
  - Synthesize and compare audio samples
- *Spectrogram Tab:*
  - Display mel spectrogram and amplitude for reference and synthesized audio
  - Plot and compare spectrograms
- *Metrics Tab:*
  - Display MCD Score, F0 RMSE Score and Carbon Emission metrics
- *Timeline Tab:*
  - Chronological overview of project milestones

- *Dark Mode:*
    - The application features a dark mode option for users who prefer a darker color scheme. This can be toggled on or off, enhancing the overall user experience based on individual preferences

12.2 Technical Details

The application employs various libraries and frameworks to ensure its functionality, a core frameworks that serve as the foundation for the user interface, providing an intuitive and interactive platform and additionally, the libraries that power the text-to-speech synthesis or contribute to efficient plotting and visualization.

- *Libraries and Frameworks:*
    - NiceGUI for the user interface
    - XTTS for speech synthesis
    - Matplotlib, Librosa, Plotly for visualization
- *Vocoders:*
    - HiFiGAN, Multi-Band MelGAN, WaveGrad
- *Asynchronous Operations:*
    - Asynchronous plotting and audio processing for responsiveness
- *Containerization:*
    - Docker to create a portable and reproducible deployment environment
    - Container encapsulates the application, its dependencies and runtime settings

13 Conclusion

Overall, the evaluation of HiFi-GAN, Multi-band MelGAN, and WaveGrad vocoders within an Extended Text-to-Speech (XTTS) system demonstrates that HiFi-GAN outperforms its counterparts in terms of a balanced composite score that considers audio quality, pitch accuracy, and environmental impact. This study underscores the necessity of adopting a holistic approach in vocoder evaluation, integrating both subjective perceptions and objective metrics, along with considerations of environmental sustainability.

The development of the Headspace Web Application further applies these findings, offering a user-friendly platform for text-to-speech synthesis that highlights the differences between vocoders in a practical context. This research not only advances TTS technology but also emphasizes the importance of environmental considerations in AI development.

In essence, our findings advocate for a comprehensive evaluation framework for TTS technologies, prioritizing both performance and sustainability, to guide the selection of vocoders in applications where quality and environmental impact are critical considerations.

References

Speech Processing Book kernel description. https://speechprocessingbook.aalto.fi/index.html. Accessed: 2024-02-05.

XTTS kernel description. https://coqui.ai/blog/tts/open_xtts. Accessed: 2024-02-05.

Sivanand Achanta, K. N. R. K. Raju Alluri, and Suryakanth V. Gangashetty. 2016. Statistical parametric speech synthesis using bottleneck representation from sequence auto-encoder. *ArXiv*, abs/1606.05844.

Naroa Amezaga and Jeremy Hajek. 2022. Availability of voice deepfake technology and its impact for good and evil. *Proceedings of the 23rd Annual Conference on Information Technology Education*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks.

John Kominek, Tanja Schultz, and Alan W. Black. 2008. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. In *Proc. Speech Technology for Under-Resourced Languages (SLTU-2008)*, pages 63–68.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.

Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. 2017. Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.

Soumi Maiti, Yifan Peng, Takaaki Saeki, and Shinji Watanabe. 2023. Speechlmscore: Evaluating speech generation using speech language model. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Cândido Júnior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts a multilingual dataset for speech synthesis in low-resource languages.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *CoRR*, abs/1907.10597.

Ramanan Sivaguru, Vasista Sai Lodagala, and Sharma Umesh. 2023. Saltts: Leveraging self-supervised speech representations for improved text-to-speech synthesis. *ArXiv*, abs/2308.01018.

Robert Streijl, Stefan Winkler, and David Hands. 2016. Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, 22:213–227.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Hongwei Tan, Yifan Zhou, Quanzheng Tao, Johanna Rosen, and Sebastiaan van Dijken. 2021. Bioinspired multisensory neural network with crossmodal integration and recognition. *Nature Communications*, 12.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2020. Multi-band melgan: Faster waveform generation for high-quality text-to-speech.

Li Zhou, Wenyu Chen, Yong Cao, Dingyi Zeng, Wanlong Liu, and Hong Qu. 2024. MLPs Compass: What is learned when MLPs are combined with PLMs? ArXiv [cs.CL].