Interspeech 2024
1-5 September 2024, Kos, Greece

FreeV: Free Lunch For Vocoders Through Pseudo Inversed Mel Filter

Yuanjun Lv[1], *Hai Li*[2], *Ying Yan*[2], *Junhui Liu*[2], *Danming Xie*[2], *Lei Xie*[1,*]

[1]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]iQIYI Inc., Chengdu, China

yjlv@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

Vocoders reconstruct speech waveforms from acoustic features and play a pivotal role in modern TTS systems. Frequent-domain GAN vocoders like Vocos and APNet2 have recently seen rapid advancements, outperforming time-domain models in inference speed while achieving comparable audio quality. However, these frequency-domain vocoders suffer from large parameter sizes, thus introducing extra memory burden. Inspired by PriorGrad and SpecGrad, we employ pseudo-inverse to estimate the amplitude spectrum as the initialization roughly. This simple initialization significantly mitigates the parameter demand for vocoder. Based on APNet2 and our streamlined Amplitude prediction branch, we propose our FreeV, compared with its counterpart APNet2, our FreeV achieves **1.8× inference speed improvement** with nearly **half parameters**. Meanwhile, our FreeV outperforms APNet2 in resynthesis quality, marking a step forward in pursuing real-time, high-fidelity speech synthesis. Code and checkpoints is available at: `https://github.com/BakerBunker/FreeV`

Index Terms: speech synthesis, neural vocoder, signal processing, waveform synthesis

1. Introduction

Recently, there has been a rapid advancement in the field of neural vocoders, which transform speech acoustic features into waveforms. These vocoders play a crucial role in text-to-speech synthesis, voice conversion, and audio enhancement applications. Within these contexts, the process typically involves a model that predicts a mel-spectrogram from the source text or speech, followed by a vocoder that produces the waveform from the predicted mel-spectrogram. Consequently, the quality of the synthesized speech, the speed of inference, and the parameter size of the model constitute the three primary metrics for assessing the performance of neural vocoders.

Recent advancements in vocoders, including iSTFTNet [1], Vocos [2], and APNet [3], have shifted from the prediction of waveforms in the time domain to the estimation of amplitude and phase spectra in the frequency domain, followed by waveform reconstruction via inverse short-time Fourier transform (ISTFT). This method circumvents the need to predict extensive time-domain waveforms, thus reducing the models' computational burden. ISTFTNet, for example, minimizes the computational complexity by decreasing the upsampling stages and focusing on frequency-domain spectra predictions before employing ISTFT for time-domain signal reconstruction. Vocos extends these advancements by removing all upsampling layers and utilizing the ConvNeXtV2 [4] Block as its foundational layer. APNet [3] and APNet2 [5] further refine this approach by independently predicting amplitude and phase spectra and incorporating innovative supervision to guide phase spectra estimation. Nonetheless, with comparable parameter counts, these models often underperform their time-domain counterparts, highlighting potential avenues for optimization in the parameter efficiency of frequency-domain vocoders.

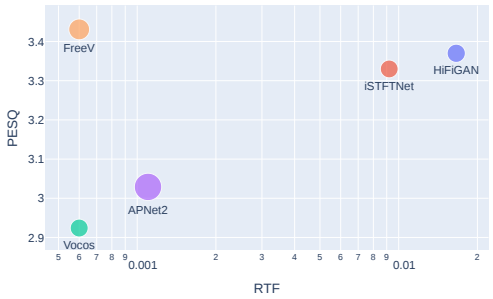


Figure 1: *Inference speed and reconstruction performance of multiple methods trained and evaluated on LJSpeech. The size of the circle represents the model parameter size. FreeV achieves the fastest inference speed and reconstruction quality with half parameter size compared to APNet2.*

Several diffusion-based vocoders have integrated signal-processing insights to reduce inference steps and improve reconstruction quality. PriorGrad [6] initially refines the model's priors by aligning the covariance matrix diagonals with the energy of each frame of the Mel spectrogram. Extending this innovation, SpecGrad [7] proposed to adjust the diffusion noise to align its dynamic spectral characteristics with those of the conditioning mel spectrogram. Moreover, GLA-Grad [8] enhances the perceived audio quality by embedding the estimated amplitude spectrum into each diffusion step's post-processing stage. Nevertheless, the reliance on diffusion models results in slower inference speeds, posing challenges for their real-world application.

In this work, we introduce *FreeV*, a streamlined GAN vocoder enhanced with prior knowledge from signal processing, and tested on the LJSpeech dataset [9]. The empirical outcomes highlight FreeV's superior performance characterized by faster convergence in training, a near 50% reduction in parameter size, and a notable boost in inference speed. Our contributions can be summarized as follows:

- We innovated by using the product of the Mel spectrogram and the pseudo-inverse of the Mel filter, referred to as the pseudo-amplitude spectrum, as the model's input, effectively easing the model's complexity.
- Drawing on our initial insight, we substantially diminished the spectral prediction branch's parameters and the time required for inference without compromising the quality achieved by the original model.

*: corresponding author

10.21437/Interspeech.2024-2407

2. Related Work

2.1. PriorGrad & SpecGrad

Based on diffusion-based vocoder WaveGrad [10], which direct reconstruct the waveform through a DDPM process, Lee *et al.* proposed PriorGrad [6] by introducing an adaptive prior $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is computed from input mel spectrogram X. The covariance matrix $\mathbf{\Sigma}$ is given by: $\mathbf{\Sigma} = \text{diag}[(\sigma_1^2, \sigma_2^2, \cdots, \sigma_D^2,)]$, where σ_d^2, denotes the signal power at dth sample, which is calculated by interpolating the frame energy. Compared to conventional DDPM-based vocoders, PriorGrad utilizes signal before making the source distribution closer to the target distribution, which simplifies the reconstruction task.

Based on PriorGrad, SpecGrad [7] proposed adjusting the diffusion noise in a way that aligns its dynamic spectral characteristics with those of the conditioning mel spectrogram. SpecGrad introduced a decomposed covariance matrix and its approximate inverse using the idea from T-F domain filtering, which is conditioned on the mel spectrogram. This method enhances audio fidelity, especially in high-frequency regions. We denote the STFT by a matrix G, and the ISTFT by a matrix G^+, then the time-varying filter L can be expressed as:

$$L = G^+ DG, \tag{1}$$

where D is a diagonal matrix that defines the filter, and it is obtained from the spectral envelope. Then we can obtain covariance matrix $\Sigma = LL^T$ of the standard Gaussian noise $\mathcal{N}(0, \Sigma)$ in the diffusion process. By introducing more accurate prior to the model, SpecGrad achieves higher reconstruction quality and inference speech than PriorGrad.

2.2. APNet & APNet2

As illustrated in Figure 2, APNet2 [5] consists of two components: amplitude spectra predictor (ASP) and phase spectra predictor (PSP). These two components predict the amplitude and phase spectra separately, which are then employed to reconstruct the waveform through ISTFT. The backbone of APNet2 is ConvNeXtV2 [4] block, which is proved has strong modeling capability. In the PSP branch, APNet [3] proposed the parallel phase estimation architecture at the output end. The parallel phase estimation takes the output of two convolution layers as the pseudo imaginary part I and real part R, then obtains the phase spectra by:

$$\arctan(\frac{I}{R}) - \frac{\pi}{2} \cdot sgn(I) \cdot [sgn(R) - 1] \tag{2}$$

where sgn is the sign function.

A series of losses are defined in APNet to supervise the generated spectra and waveform. In addition to the losses used in HiFiGAN [11], which include Mel loss $\mathcal{L}_{mel}$, generator loss $\mathcal{L}_g$, discriminator loss $\mathcal{L}_d$, feature matching loss $\mathcal{L}_{fm}$, APNet proposed:

- amplitude spectrum loss $\mathcal{L}_A$, which is the L2 distance of the predicted and real amplitude;
- phase spectrogram loss $\mathcal{L}_P$, which is the sum of instantaneous phase loss, group delay loss, and phase time difference loss, all phase spectrograms are anti-wrapped;
- STFT spectrogram loss $\mathcal{L}_S$, which includes the STFT consistency loss and L1 loss between predicted and real reconstructed STFT spectrogram.

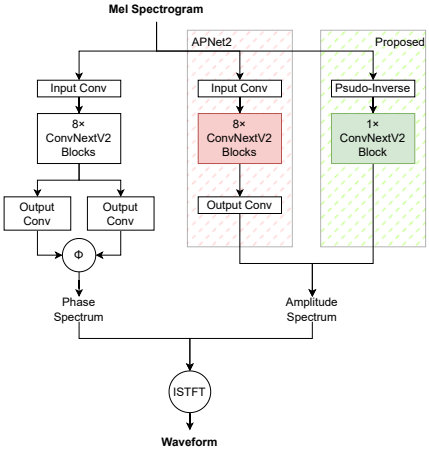


Figure 2: *The overall architecture of FreeV, the amplitude prediction branch (ASP) of APNet2, which has red background, is replaced by a more lightweight architecture with green background.*

3. Method

"When we structure the informative prior noise closer to the data distribution, can we improve the efficiency of the model?"
– PriorGrad

3.1. Amplitude Prior

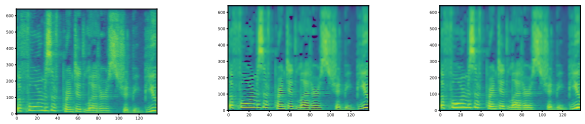
In this section, we investigate how to obtain a prior signal closer to the real prediction target, which is the amplitude spectrum. By employing the given Mel spectrum X and the known Mel filter M, we aim to obtain an amplitude spectrum $\hat{A}$ that minimizes the distance with the actual amplitude spectrum A, while ensuring that the computation is performed with optimal speed, as the following equation:

$$\min \left\| \hat{A}M - A \right\|_2 \tag{3}$$

We investigated several existing implementations for this task. In Section 2.1, the SpecGrad method, $G^+DG\epsilon$ requires prior noise ϵ as input, therefore unsuitable for our goals. In the implementation by the librosa library [12], the estimation of $\hat{A}$ employs the Non-Negative Least Squares (NNLS) algorithm to maintain non-negativity. However, this algorithm is slow due to the need for multiple iterations, prompting the pursuit of a swifter alternative. TorchAudio's implementation [13] calculates the estimated amplitude spectrum through a singular least squares operation followed by enforcing a minimum threshold to preserve non-negativity. Despite this, the recurring need for the least squares calculation with each inference introduces speed inefficiencies.

Considering that the Mel filter M remains unchanged throughout the calculations, we can pre-compute its pseudo-inverse, denoted as M^+. Then, to guarantee the non-negativity of the amplitude spectrum and maintain numerical stability in training, we impose a lower bound of 10^{-5} on the values of the approximate amplitude spectrum. We find there are some negative values in the pseudo-inversed mel filter, causing negative blocks in estimated amplitude, which can be easily found in Figure 3b, so we add an Abs function to the product of M^+ and X. This allows us to derive the approximate amplitude spectrum $\hat{A}$ using the following equation:

$$\hat{A} = \max(\text{Abs}(M^+X), 10^{-5}) \tag{4}$$



(a) A (b) $\hat{A}$ *w/o abs* (c) $\hat{A}$ *w/ abs*

Figure 3: *Comparison of real log amplitude spectra A and estimated log spectra $\hat{A}$.*

This enables us to efficiently acquire the estimated amplitude spectrum through a single matrix multiplication operation.

3.2. Model Structure

Our model architecture is illustrated in Figure 2, which consists of PSP and ASP, and uses ConvNextV2 [4] as the model's basic block. PSP includes an input convolutional layer, eight ConvNeXtV2 blocks, and two convolutional layers for parallel phase estimation structure.

Diverging from APNet2's ASP, our design substitutes the conventional input convolutional layer with the pre-computed pseudo-inverse Mel filter matrix M^+ of the Mel filter M with frozen parameters. Due to the enhancements highlighted in Section 3.1 that substantially ease the model's complexity, the number of ConvNeXtV2 blocks is reduced from eight to a single block, thereby substantially reducing both the parameter footprint and computation time.

Concurrently, the ConvNeXtV2 module's input-output dimensions have been tailored to align with those of the amplitude spectrum, enabling the block to exclusively model the residual between the estimated and real amplitude spectra, further reducing the ASP module's modeling difficulty. Because the input and output dimensions of the ConvNeXtV2 module match the amplitude spectrum, we removed the output convolutional layer from ASP, further reducing the model's parameter count.

3.3. Training Criteria

In the choice of discriminators, we followed the setup in APNet2 [5], using MPD and MRD as discriminators and adopting Hinge GAN Loss as the loss function for adversarial learning. We also retained the other loss functions used by APNet2, which is described in Section 2.2, and the loss function of the generator and discriminator are denoted as:

$$\mathcal{L}_{Gen} = \lambda_A \mathcal{L}_A + \lambda_P \mathcal{L}_P + \lambda_S \mathcal{L}_S + \lambda_W (\mathcal{L}_{mel} + \mathcal{L}_{fm} + \mathcal{L}_g)$$
$$\mathcal{L}_{Dis} = \mathcal{L}_d$$

where λ_A, λ_P, λ_S, λ_W are the weights of the loss, which are kept the same as in APNet2.

4. Experimental Setup

To evaluate the effectiveness of our proposed FreeV, we follow the training scheme in APNet2 paper. Our demos are placed at demo-site[1].

4.1. Dataset

To ensure consistency, the training dataset follows the same configuration of APNet2. Thus, the LJSpeech dataset [9] is used for training and evaluation. LJSpeech dataset is a public collection of 13,100 short audio clips featuring a single speaker reading passages from 7 non-fiction books. The duration of the clips ranges from 1 to 10 seconds, resulting in a total length of approximately 24 hours. The sampling rate is 22050Hz. We split the dataset to train, validation, and test sets according to open-source VITS repository[2].

For feature extraction, we use STFT with 1024 bins, a hop size of 256, and a Hann window of length 1024. For the mel filterbank, 80 filterbanks are used with a higher frequency cutoff at 16 kHz.

4.2. Model and Training Setup

We compare our proposed model with HiFiGAN[3] [11], iSTFTNet[4] [1], Vocos[5] [2] and APNet2[6] [5]. In Our FreeV vocoder, the number of ConvNeXtV2 blocks is 8 for PSP and 1 for ASP, the input-output dimension is 512 for PSP and 513 for ASP, the hidden dimension is 1536 for both ASP and PSP.

We trained FreeV for 1 million steps. We set the segmentation size to 8192 and the batch size to 16. We use the AdamW optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$, and a weight decay of 0.01. The learning rate is set to 2×10^{-4} and exponentially decays with a factor of 0.99 for each epoch.

4.3. Evaluation

Multiple objective evaluations are conducted to compare the performance of these vocoders. We use seven objective metrics for evaluating the quality of reconstructed speech, including mel-cepstrum distortion (MCD), root mean square error of log amplitude spectra and F0 (LAS-RMSE and F0-RMSE), V/UV F1 for voice and unvoiced part, short time objective intelligence (STOI) [14] and perceptual evaluation speech quality (PESQ) [15]. To evaluate the efficiency of each vocoder, model parameter count (Params) and real-time factor (RTF) are also conducted on NVIDIA A100 for GPU and a single core of Intel Xeon Platinum 8369B for CPU.

For the computational efficiency of the prior, we also conducted RTF and LAS-RMSE evaluations to the NNLS algorithm of librosa, least square algorithm of torchaudio, pseudo-inverse algorithm, and pseudo-inverse algorithm with absolute function mentioned in Section 3.1.

Table 1: *Time and precision of different prior computing methods, LS stands for Least Square, PI stands for Pseudo Inverse.*

Method	NNLS	LS	PI	PI w/ abs
Time (↓)	290ms	286μs	102μs	107μs
LAS-RMSE (↓)	2.0729	2.0729	2.0729	0.6843

5. Experiment Result

We conducted experiments to verify whether our method can improve the efficiency of the vocoder.

5.1. Computational Efficiency of Prior

The compute method of the estimated amplitude spectra $\hat{A}$ if our key component. We find that the inference speed can be af-

[1] https://bakerbunker.github.io/FreeV/

[2] https://github.com/jaywalnut310/vits/tree/main/filelists

[3] https://github.com/jik876/hifi-gan

[4] https://github.com/rishikksh20/iSTFTNet-pytorch

[5] https://github.com/gemelo-ai/vocos

[6] https://github.com/redmist328/APNet2

Table 2: *Results of objective evaluations on the testset of LJSpeech dataset for reconstruction.*

Model	MCD(↓)	LAS-RMSE(↓)	V/UV F1(↑)	Periodicity (↓)	F0-RMSE(↓)	STOI(↑)	PESQ(↑)
HiFiGAN [11]	3.857	1.150	0.941	0.145	36.03	0.923	3.370
HiFiGAN w/$\hat{A}$	3.751	1.141	0.945	0.136	34.26	0.928	**3.509**
iSTFTNet [1]	3.838	1.130	0.941	0.143	36.40	0.925	3.330
iSTFTNet w/$\hat{A}$	3.755	1.120	0.944	0.138	34.12	0.927	3.422
Vocos [2]	3.367	0.948	0.941	0.158	45.41	0.948	2.924
APNet2 [5]	3.518	0.782	0.950	0.132	31.08	0.950	3.029
FreeV (Proposed)	**3.112**	**0.779**	**0.956**	**0.118**	**26.40**	**0.967**	3.431

Table 3: *Results of parameter and inference speed.*

Model	Params (M)	RTF (CPU)	RTF (GPU)
HiFiGAN	13.9M	0.062	0.0166
iSTFTNet	**13.3M**	0.409	0.0092
Vocos	13.5M	**0.028**	**0.0006**
APNet2	31.4M	0.062	0.0011
FreeV (Proposed)	18.2M	0.036	**0.0006**

fected by the compute speed of the prior. We compare the compute speed and accuracy on 100 2-second-long speech clips. As shown in Table 1, the pseudo-inverse method is the fastest way to compute the estimated amplitude spectra $\hat{A}$, and the result also shows that the Abs function can largely reduce the error of amplitude spectrogram estimation.

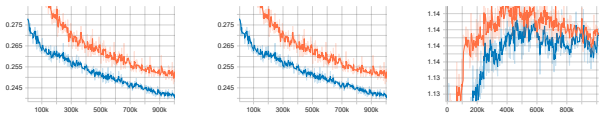
5.2. Model Convergence

In Figure 4a and 4b, we showcase the amplitude spectrum loss and mel spectrum loss curves related to amplitude spectrum prediction. From these two curves, it can be seen that even though the number of parameters in the amplitude spectrum prediction branch is significantly reduced, the loss related to amplitude spectrum prediction still remains lower than the baseline APNet2. This observation affirms the efficacy of the approach described in Section 3, substantiating a marked decrease in the challenge of amplitude spectrum prediction. Furthermore, Figure 4c displays the Phase-Time Difference Loss, which bears significant relevance to phase spectrum prediction. The improvement in amplitude spectrum prediction concurrently benefits phase spectrum accuracy. We assume that the stability of the amplitude spectrum prediction branch's training engenders more effective optimization of the phase information by the waveform-related loss functions.

Furthermore, we extended our experimentation to the baseline model by substituting its input from the Mel spectrum with the estimated amplitude spectrum $\hat{A}$. The loss curve illustrated in Figure 5 reveals that this modification also enhanced the early-stage convergence of these models. This finding suggests that integrating an appropriate prior is advantageous not only for our proposed vocoder but also holds potential efficacy for other vocoder frameworks.

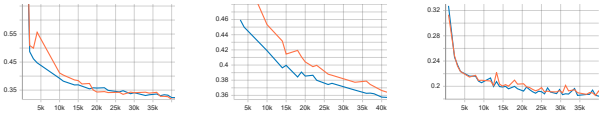
5.3. Model Performance

The model's performance was evaluated on the test dataset referenced in Section 4.1, the results of which are detailed in Table 2. FreeV outperformed in five out of six objective metrics



(a) $\mathcal{L}_{mel}$ (b) $\mathcal{L}_{amp}$ (c) $\mathcal{L}_{ptd}$

Figure 4: *Loss curves of APNet2 (orange) and FreeV (blue).*



(a) *HiFiGAN* (b) *iSTFTNet* (c) *Vocos*

Figure 5: *Early stage mel loss curves of multiple models trained with (blue) and without estimated amplitude spectra $\hat{A}$ (orange).*

and was surpassed only by HiFiGAN with estimated amplitude spectra in the PESQ metric. These findings indicate that our method reduces the model's parameter size and elevates the quality of audio reconstruction. Furthermore, the comparative analysis, which includes both scenarios, with and without the incorporation of the estimated amplitude spectrum $\hat{A}$, reveals that substituting the Mel spectrum X input with the approximate amplitude spectrum $\hat{A}$ can also yield performance gains in standard vocoder configurations. This observation corroborates the efficacy of our proposed approach.

In parallel, as shown by Table 3, our model's parameter size is confined to merely a half of that to APNet2, while it achieves $1.8\times$ inference speed on GPU. When benchmarked against the time-domain prediction model HiFiGAN [11], FreeV not only exhibits a considerable speed enhancement, which is approximately $30\times$, but also delivers superior audio reconstruction fidelity with comparable parameter count. These results further underscore the practicality and advantage of our proposed method.

6. Conclusion

In this paper, we investigated the effectiveness of employing pseudo-inverse to roughly estimate the amplitude spectrum as the initial input of the model. We introduce FreeV, a vocoder framework that leverages estimated amplitude spectrum $\hat{A}$ to simplify the model's predictive complexity. This approach not only reduces the parameter size but also improves the reconstruction quality compared to APNet2. Our experimental results demonstrated that our method could effectively reduce the modeling difficulty by simply replacing the input mel spectrogram with the estimated amplitude spectrum $\hat{A}$.

7. References

[1] T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "ISTFTNET: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6207–6211.

[2] H. Siuzdak, "Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis," in *The Twelfth International Conference on Learning Representations*, vol. abs/2306.00814, 2023.

[3] Y. Ai and Z.-H. Ling, "APNet: An All-Frame-Level Neural Vocoder Incorporating Direct Prediction of Amplitude and Phase Spectra," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2145–2157, 2023.

[4] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders." in *Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 133–16 142.

[5] H.-P. Du, Y.-X. Lu, Y. Ai, and Z.-H. Ling, *APNet2: High-Quality and High-Efficiency Neural Vocoder with Direct Prediction of Amplitude and Phase Spectra*. Springer Nature Singapore, 2023.

[6] S.-g. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S.-H. Yoon, and T.-Y. Liu, "PriorGrad: Improving Conditional Denoising Diffusion Models with Data-dependent Adaptive Prior," in *International Conference on Learning Representations*, 2021.

[7] Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion Probabilistic Model based Neural Vocoder with Adaptive Noise Spectral Shaping," in *Interspeech 2022*. ISCA, 2022, pp. 803–807.

[8] H. Liu, T. Baoueb, M. Fontaine, J. L. Roux, and G. Richard, "GLA-Grad: A Griffin-Lim Extended Waveform Generation Diffusion Model," *arXiv*, 2024.

[9] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[10] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating Gradients for Waveform Generation." in *International Conference on Learning Representations (ICLR)*, 2021.

[11] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[12] B. McFee, M. McVicar, D. Faronbi, I. Roman, M. Gover, S. Balke, S. Seyfarth, A. Malek, C. Raffel, V. Lostanlen, B. van Niekirk, D. Lee, F. Cwitkowitz, F. Zalkow, O. Nieto, D. Ellis, J. Mason, K. Lee, B. Steers, E. Halvachs, C. Thomé, F. Robert-Stöter, R. Bittner, Z. Wei, A. Weiss, E. Battenberg, K. Choi, R. Yamamoto, C. Carr, A. Metsai, S. Sullivan, P. Friesch, A. Krishnakumar, S. Hidaka, S. Kowalik, F. Keller, D. Mazur, A. Chabot-Leclerc, C. Hawthorne, C. Ramaprasad, M. Keum, J. Gomez, W. Monroe, V. A. Morozov, K. Eliasi, nullmightybofo, P. Biberstein, N. D. Sergin, R. Hennequin, R. Naktinis, beantowel, T. Kim, J. P. Åsen, J. Lim, A. Malins, D. Hereñú, S. van der Struijk, L. Nickel, J. Wu, Z. Wang, T. Gates, M. Vollrath, A. Sarroff, Xiao-Ming, A. Porter, S. Kranzler, Voodoohop, M. D. Gangi, H. Jinoz, C. Guerrero, A. Mazhar, toddrme2178, Z. Baratz, A. Kostin, X. Zhuang, C. T. Lo, P. Campr, E. Semeniuc, M. Biswal, S. Moura, P. Brossier, H. Lee, and W. Pimenta, "librosa/librosa: 0.10.1," Aug. 2023. [Online]. Available: https://doi.org/10.5281/zenodo.8252662

[13] J. Hwang, M. Hira, C. Chen, X. Zhang, Z. Ni, G. Sun, P. Ma, R. Huang, V. Pratap, Y. Zhang, A. Kumar, C.-Y. Yu, C. Zhu, C. Liu, J. Kahn, M. Ravanelli, P. Sun, S. Watanabe, Y. Shi, Y. Tao, R. Scheibler, S. Cornell, S. Kim, and S. Petridis, "Torchaudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for pytorch," 2023.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *icassp*, 2010, pp. 4214–4217.

[15] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *icassp*, vol. 2, 2001, pp. 749–752 vol.2.