NVIDIA

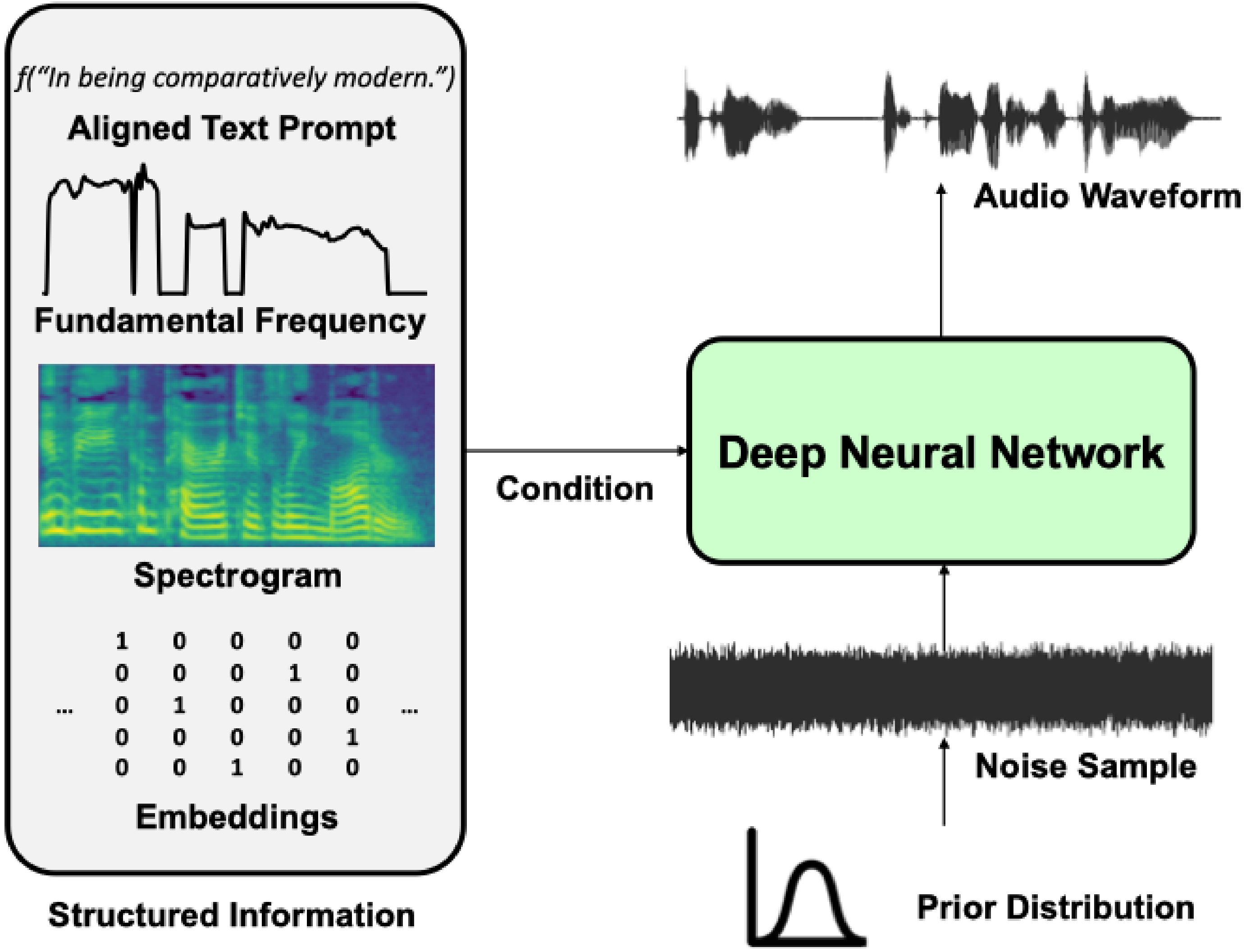
BigVGAN: A Universal Neural Vocoder with Large-Scale Training

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, Sungroh Yoon

Neural Vocoder: Conditional Waveform Synthesizer

A cornerstone module of generative AI for speech & audio

- Low resolution
- Larger # of channels
- Easier to model with generative models



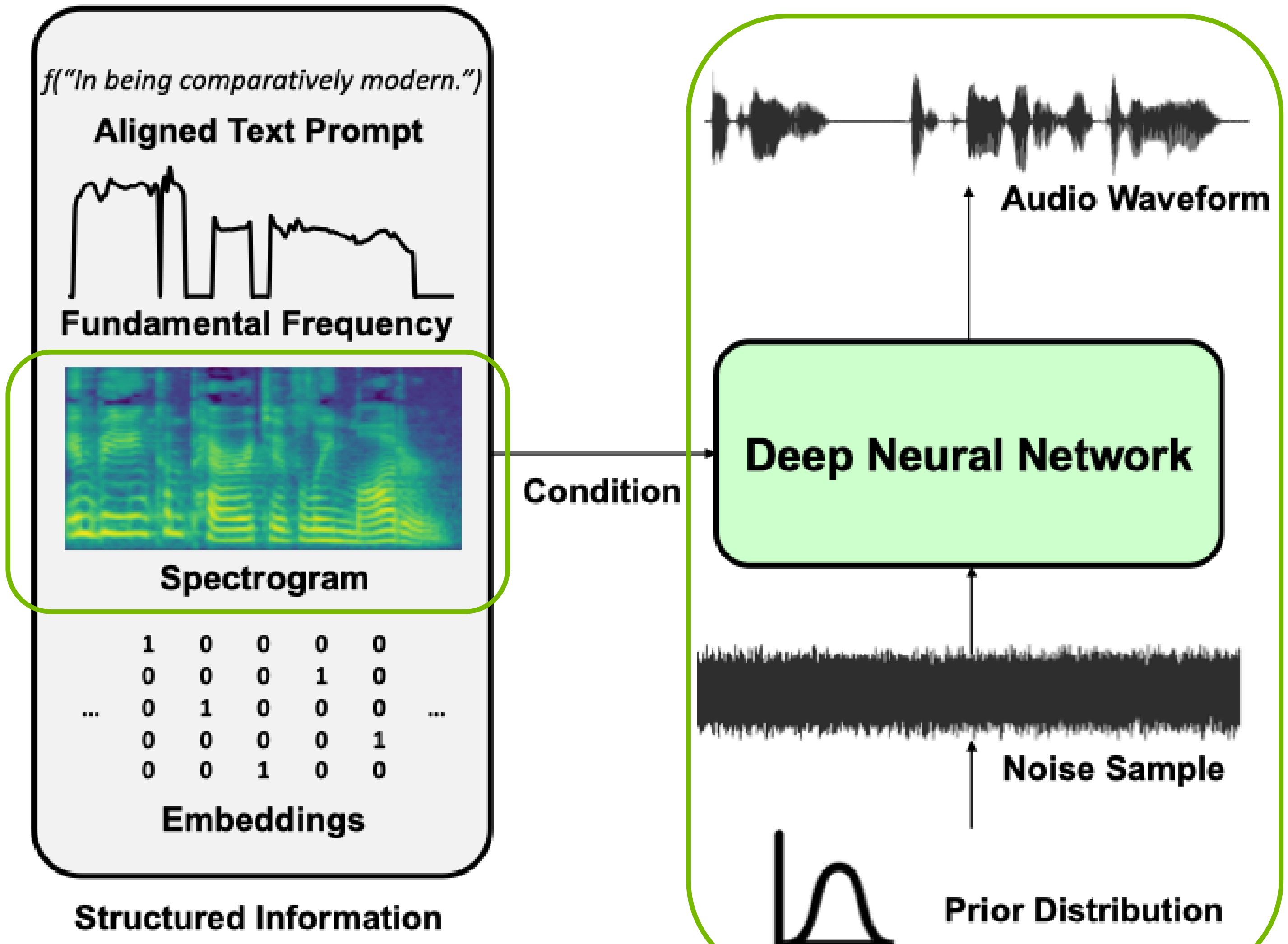
- High resolution
- Limited # of channels
- Harder to model (directly)

Source: Sang-gil Lee, "Deep Generative Model for Waveform Synthesis", Ph.D. Dissertation

Neural Vocoder Maps Acoustic Feature to Waveform

With mel spectrogram (with unknown phase) as a standard choice

Insert_your_X_to_mel_generative_models()



Neural vocoder inverts the mel spectrogram back to waveform

Source: Sang-gil Lee, "Deep Generative Model for Waveform Synthesis", Ph.D. Dissertation

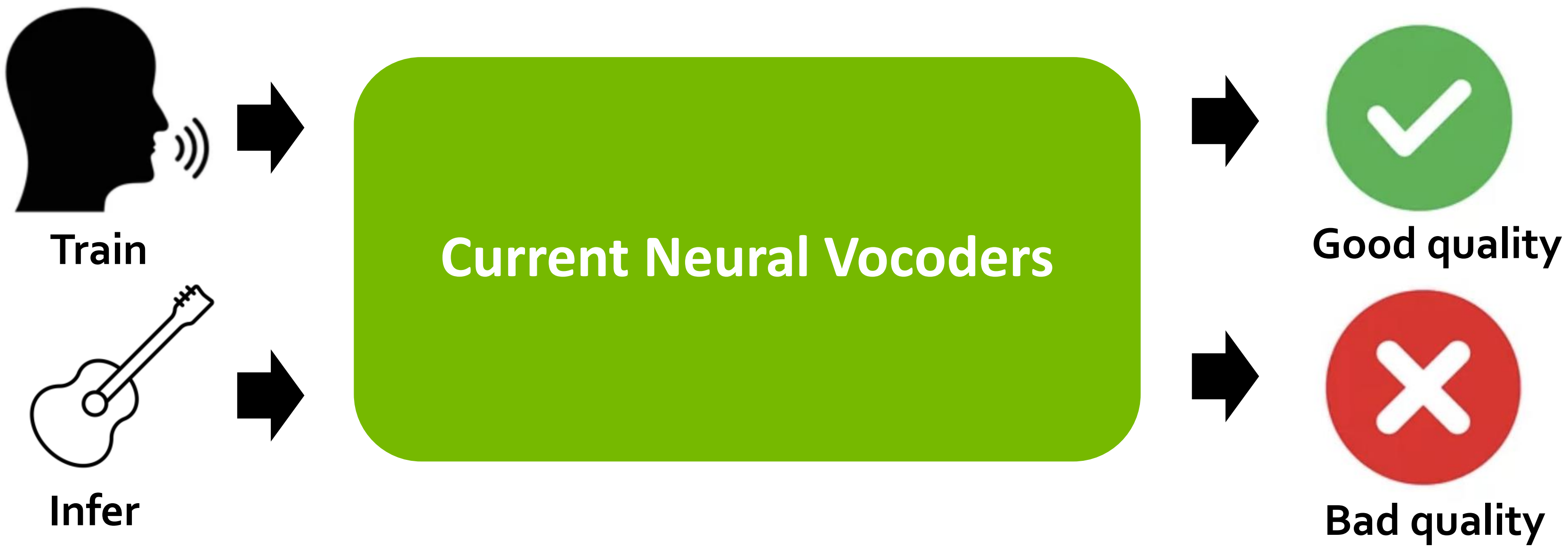
Universal Vocoder Should Generalize to Every Sounds of the World

Current neural vocoders are not robust enough to unseen conditions

An **ideal, drop-in universal vocoder** should be robust to:

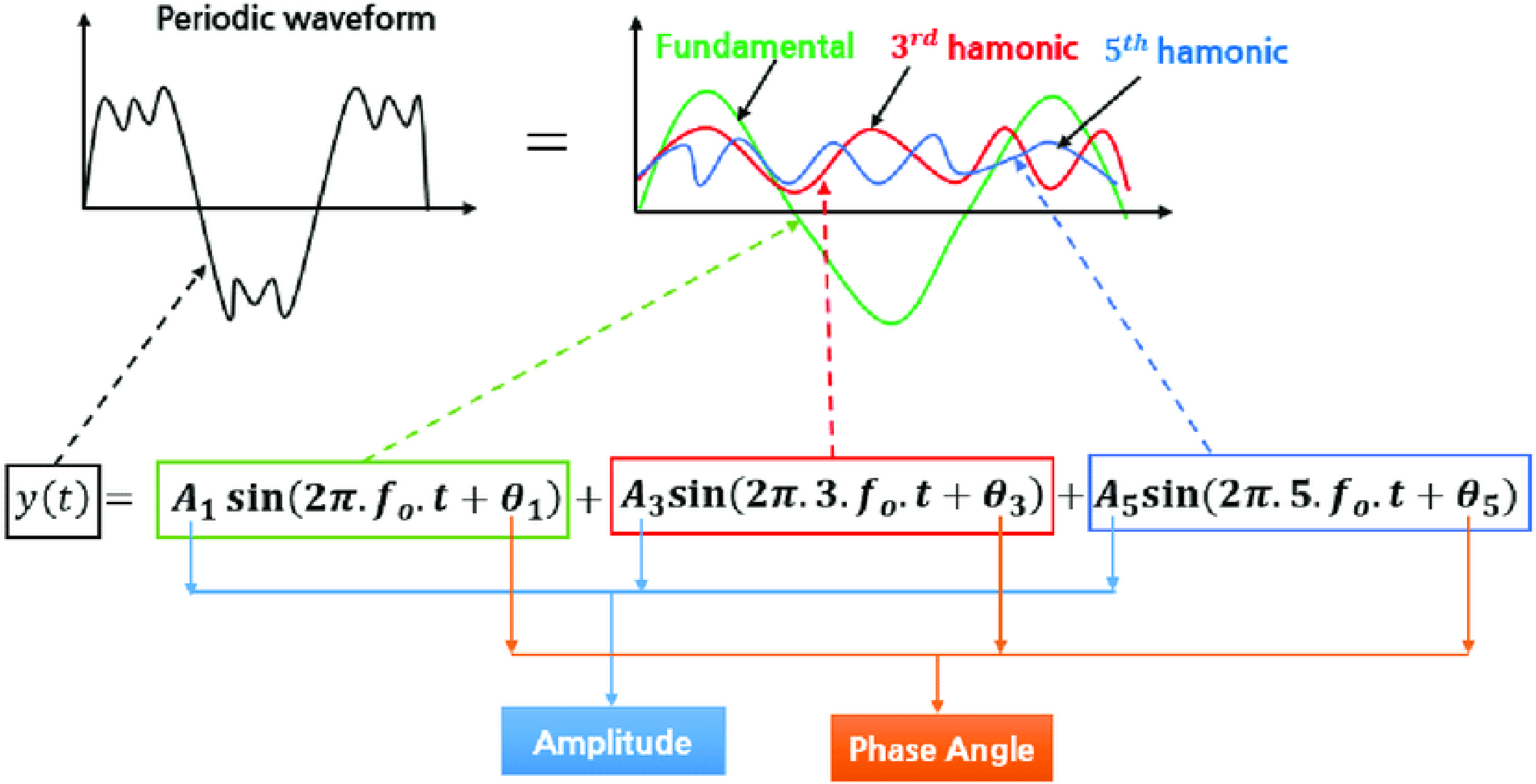
- Unseen speakers & languages
- Unseen recording environments
- Non-speech vocalizations

… and any types of speech & audio **unseen during training**, without any retraining or finetuning!



Audio Waveform Is Highly Periodic

Waveform can be expanded to a sum of multiple sine waves (Fourier series)



Source: : https://www.researchgate.net/figure/The-periodic-signal-analyzing-with-odd-numbered-harmonic-frequencies_fig2_340049927

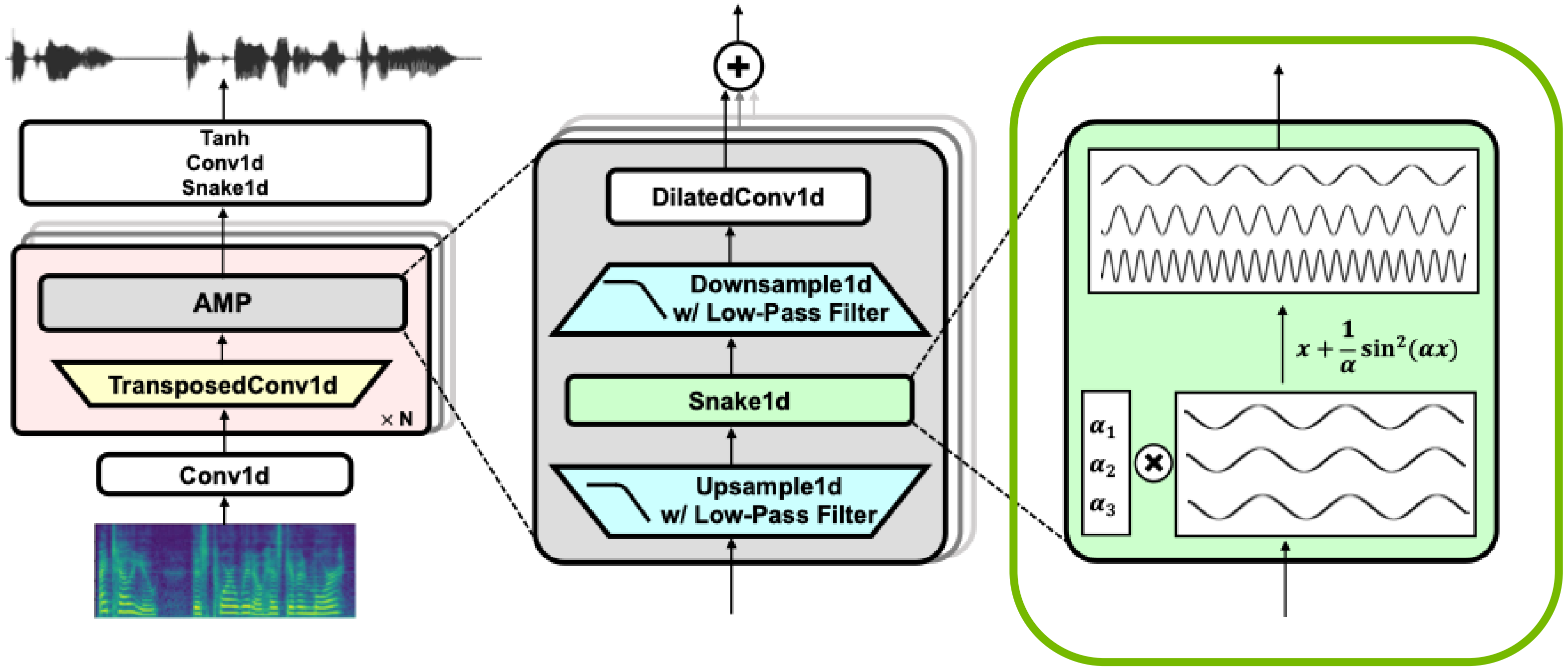
BigVGAN: Anti-Aliased Multi-Periodicity Composition (AMP)

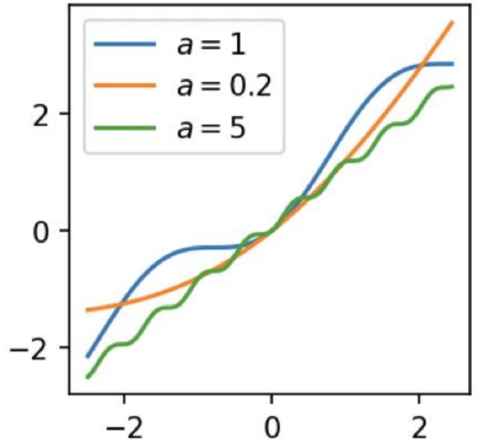
AMP applies *Snake* activation function towards inductive bias of periodicity

$x + \frac{1}{\alpha} sin^2(\alpha x)$: **periodic** activation with channel-wise trainable frequency $\boldsymbol{\alpha}$

- monotonic & easy to optimize compared to $sin(x)$ which contains many local minima

➔ Provides better **inductive bias** for waveform & stronger **out-of-distribution generalization**





Source: Ziyin et al., Neural Networks Fail to Learn Periodic Functions and How to Fix It, NeurIPS 2020

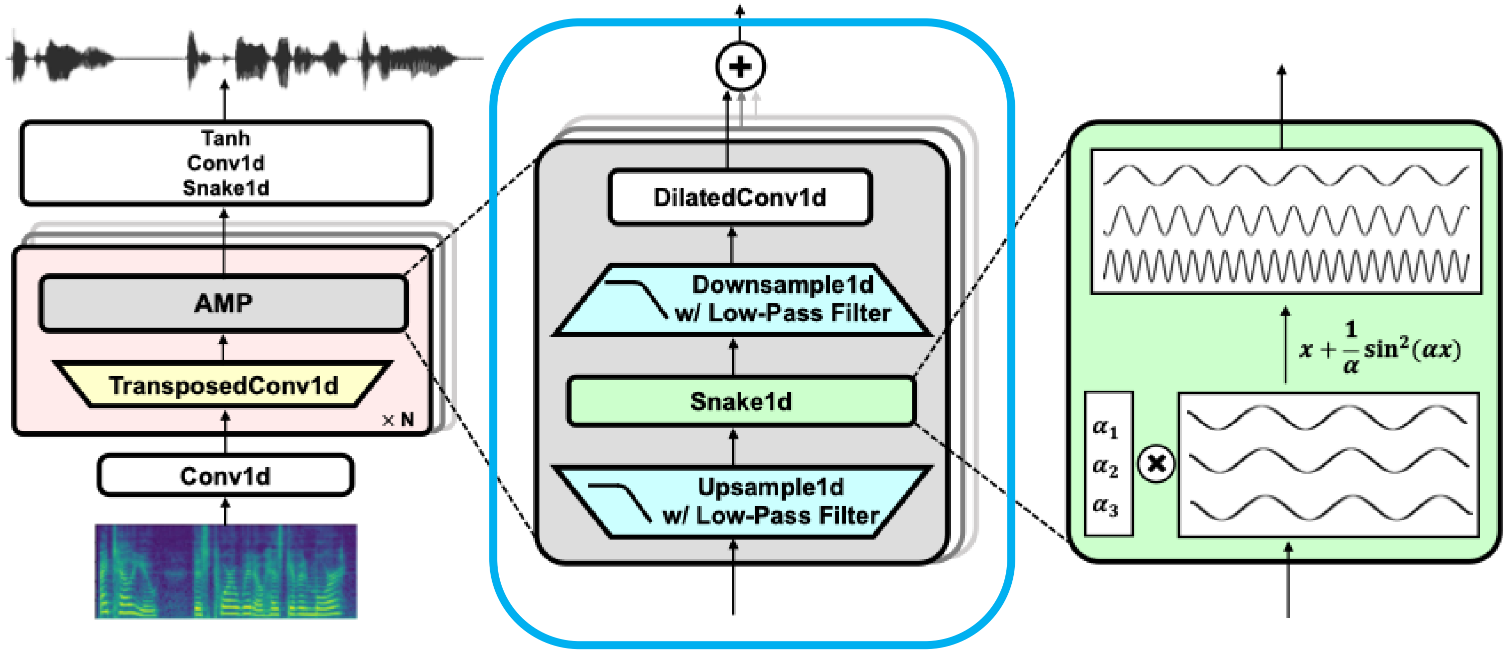
BigVGAN: Anti-Aliased Multi-Periodicity Composition (AMP)

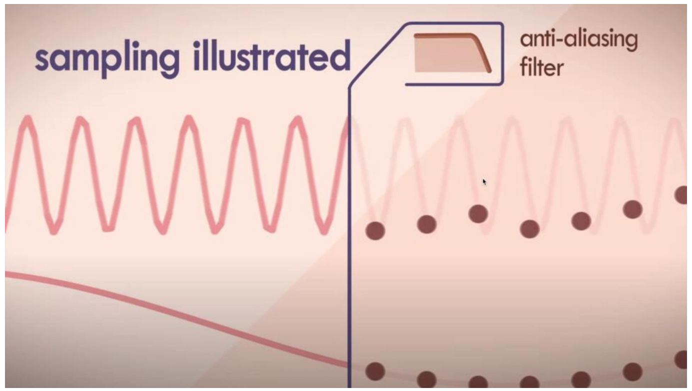
AMP applies *anti-aliasing filters* to supress high-frequency artifacts

Nonlinearities can add **arbitrarily high frequency features** that cannot be represented in the discrete grid

- Suppress the artifacts inspired by StyleGAN3 (Karras et al., NeurIPS 2021)
- **Upsample** by 2x ➔ apply nonlinearity ➔ **downsample** by 2x, along with **low-pass filters**

➔ Feature aliasing of neural vocoders can be suppressed

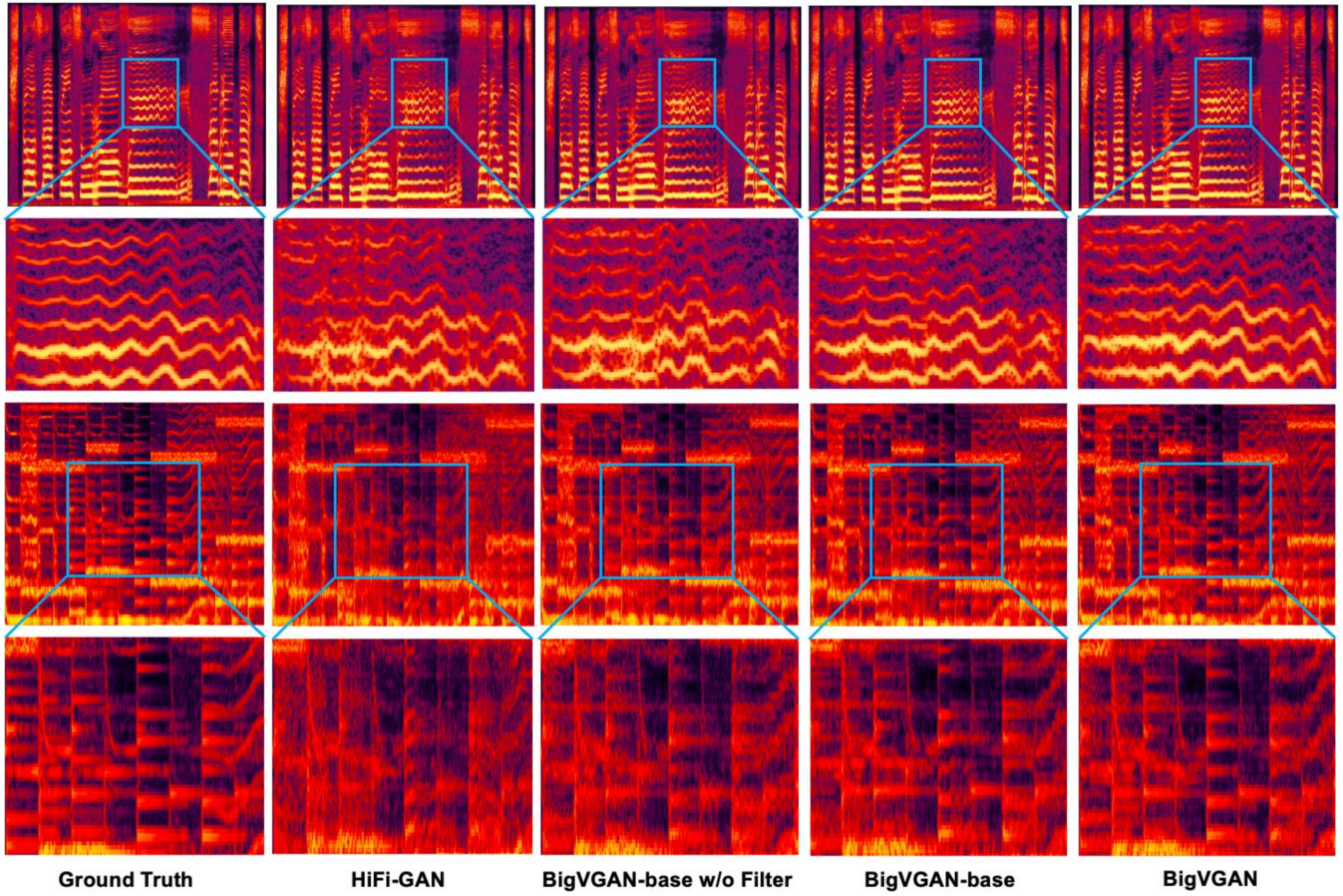




Source: https://www.youtube.com/watch?v=91PKZllbgds

Visualization

BigVGAN captures high-frequency harmonic compnents significantly better than baselines



Experimental Results: In-Domain

BigVGAN outperforms previous SOTAs by large magin: both for objective & subjective metrics

BigVGAN is the largest & high-speed GAN vocoder with **112M** parameters

Table 1: Model footprint and synthesis speed for 24 kHz audio measured on an NVIDIA RTX 8000 GPU.

Method	WaveGlow	WaveFlow	HiFi-GAN (V1)	BigVGAN-base	w/o filter	BigVGAN
Params (M)	99.43	22.58	14.01	14.01	14.01	112.4
Syn. speed	31.87×	19.59×	93.75×	70.18×	75.83×	44.72×

LibriTTS Results: various objective metrics & (similiarity) mean opinion scores ((S)MOS)

LibriTTS	M-STFT(↓)	PESQ(↑)	MCD(↓)	Periodicity(↓)	V/UV F1(↑)	MOS(↑)	SMOS(↑)
Ground Truth	-	-	-	-	-	4.40±0.06	4.44±0.06
SC-WaveRNN	2.2358	1.701	1.8854	0.3044	0.8144	3.20±0.11	3.29±0.10
WaveGlow-256	1.3099	3.138	2.3591	0.1485	0.9378	3.84±0.10	3.87±0.10
WaveFlow-128	1.1120	3.027	1.2455	0.1416	0.9410	3.85±0.10	3.89±0.10
HiFi-GAN (V1)	1.0017	2.947	0.6603	0.1565	0.9300	4.08±0.09	4.15±0.09
BigVGAN-base	0.8788	3.519	0.4564	0.1287	0.9459	4.10±0.09	4.20±0.08
BigVGAN	**0.7997**	**4.027**	**0.3745**	**0.1018**	**0.9598**	**4.11±0.09**	**4.26±0.08**

Experimental Results: Out-of-Distribution

BigVGAN is substantially better for generalization to out-of-distribution data

Multi-lingual datasets: significantly better SMOS under **unseen & noisy recording environments**

Recording env. / Language	Clean / Jv,Km,Ne,Su	Noisy (sim) / Es,Fr,It,Pt	Noisy (real) / Ko
Ground Truth	4.58±0.05	4.36±0.05	4.56±0.05
UnivNet-c32†	4.35±0.07	3.95±0.09	4.18±0.08
HiFi-GAN (V1)	4.39±0.07	4.13±0.08	4.21±0.08
BigVGAN-base	4.38±0.07	4.21±0.07	4.36±0.07
BigVGAN	**4.41±0.07**	**4.26±0.07**	**4.38±0.07**

Out-of-distriution music audio (MUSDB18-HQ): substantial SMOS improvements on **challenging data** in zero-shot

Method	Vocal	Drums	Bass	Others	Mixture	Average
Ground Truth	4.58±0.05	4.57±0.05	4.52±0.05	4.61±0.05	4.56±0.05	4.57±0.02
UnivNet-c32†	4.22±0.09	4.23±0.09	3.90±0.11	3.80±0.13	3.80±0.12	3.99±0.05
HiFi-GAN (V1)	4.26±0.08	4.37±0.08	3.95±0.11	3.92±0.12	3.91±0.11	4.08±0.05
BigVGAN-base	4.36±0.08	4.39±0.07	**4.00 ±0.11**	4.14±0.09	4.11±0.10	4.20±0.04
w/o filter	4.30±0.08	4.32±0.07	3.95±0.11	4.05±0.10	4.11±0.10	4.15±0.04
w/o filter & snake	4.31±0.08	4.32±0.07	3.94±0.11	4.01±0.11	4.02±0.10	4.12±0.04
BigVGAN	**4.37±0.08**	**4.41±0.07**	**4.00±0.10**	**4.25±0.09**	**4.26±0.08**	**4.26±0.04**

Summary: BigVGAN Explores the Limits of Universal Neural Vocoding

Code & Model: https://github.com/NVIDIA/BigVGAN

Demo: https://bigvgan-demo.github.io

arXiv: https://arxiv.org/abs/2206.04658