UNIVERSAL NEURAL VOCODING WITH PARALLEL WAVENET

Yunlong Jiao, Adam Gabryś, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa, Viacheslav Klimkov

Motivation

- State-of-the-art neural vocoders are capable of synthesizing natural-sounding speech.
- Training high-quality neural vocoders requires significant computational resources and large amounts of audio data for each target speaker.
- Most existing neural vocoders are either speaker-dependent, or have not been evaluated sufficiently to support out-of-domain voices, styles, and languages.
- A high-quality speaker-independent vocoder, or so-called **universal vocoder**, is key to scaling up production of TTS systems.

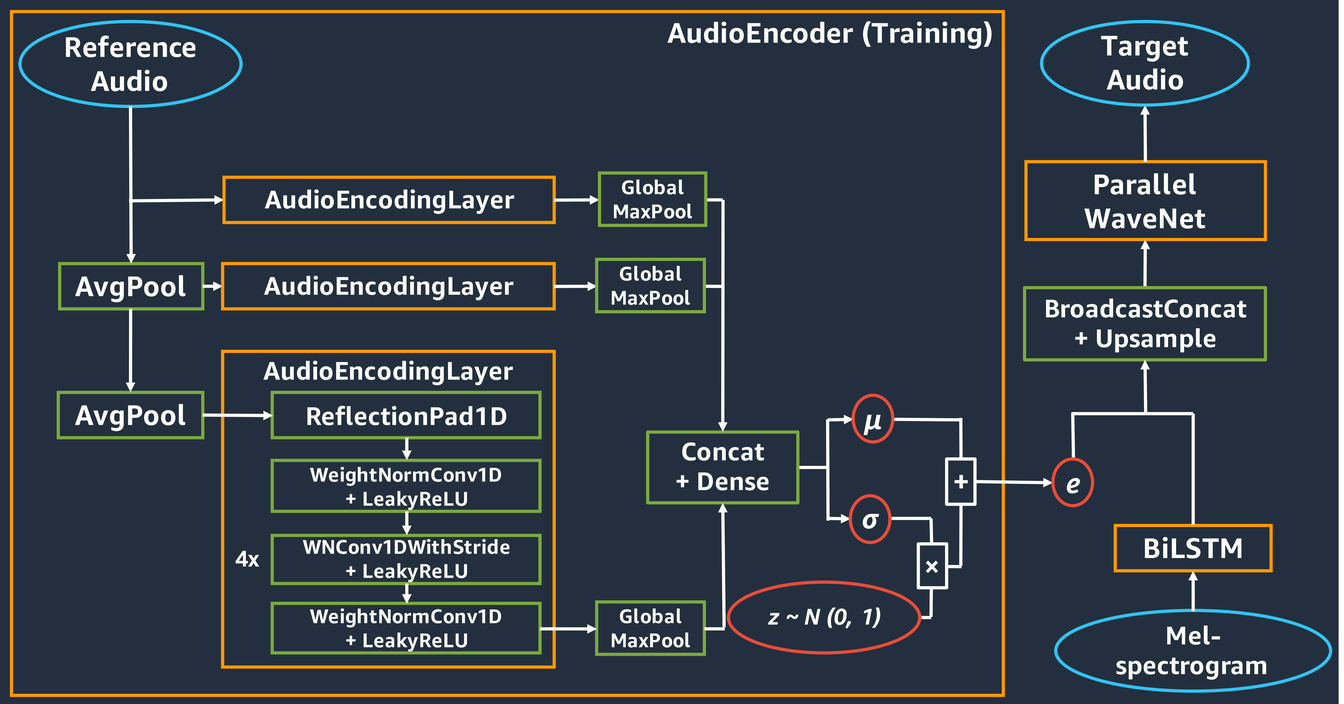
Our contributions

- We trained a **universal neural vocoder** based on Parallel WaveNet, using a multi-speaker multi-lingual high-quality speech corpus.
- In order to train a universal vocoder, we propose an additional VAE-type conditioning network called **Audio Encoder**. Its architecture is inspired by the multi-scale discriminator of MelGAN.
- We performed **extensive evaluation** to show that our universal vocoder is state-of-the-art and capable of providing out-of-the-box support for new voices, styles, and languages.

Glossary

- **UPW (ours)**: Universal Parallel WaveNet
- **SDPW**: Speaker-dependent Parallel WaveNet
- **PWGAN**: Parallel WaveGAN
- **WGlow**: WaveGlow
- **UWRNN**: Universal WaveRNN

We trained a **Universal Parallel WaveNet (UPW)** that serves all TTS voices, with an additional VAE-type conditioning network called **Audio Encoder**!



Note: At inference time, we use e = 0 to replace the output of AudioEncoder.

Take a picture to access the full paper on arXiv.org

Paper ID. 2386

UPW is a universal vocoder!

🚀 UPW vs. SDPW

- Test set statistics:

Test set	Recording quality	# Voices (seen / unseen)	# Styles (seen / unseen)	# Lang. (seen / unseen)	# Utt. (**all unseen**)	Vocoder systems
Internal	Very high	24 (21/3)	16 (12/4)	13 (13/0)	3,124	UPW, SDPW

- MUSHRA evaluation results:

MUSHRA	Recording	SDPW	UPW	UPW Relative	P-value
Internal	69.68	57.92	**58.70**	84.24%	**0.000**
MUSHRA per voice					
British Eng. / F / Adult	71.64	65.69	**67.67**	94.45%	**0.000**
Aus. Eng. / M / Adult	73.52	**68.37**	68.32	92.93%	1.000
Spanish / F / Adult	69.06	60.27	**61.17**	88.58%	0.668
Indian Eng. / F / Adult	77.19	62.22	**66.95**	86.74%	**0.000**
*US Eng. / M / Senior	70.40	57.65	**60.12**	85.40%	0.201
*US Eng. / M / Child	62.31	51.26	**51.99**	83.43%	1.000
US Eng. / M / Adult	68.58	52.63	**55.46**	80.87%	0.105
French / F / Senior	72.53	54.82	**56.35**	77.69%	**0.002**
US Spanish / F / Adult	73.71	48.07	**48.37**	65.62%	1.000
MUSHRA per style					
Emotional	71.59	60.74	**61.40**	85.76%	0.462
Neutral	69.13	58.53	**58.73**	84.95%	0.500
Conversational	58.65	43.54	**47.61**	81.18%	**0.002**
Long-form reading	68.60	**56.69**	55.46	80.85%	0.814
News briefing	75.24	56.29	**59.86**	79.55%	**0.000**
Singing	71.94	49.96	**56.87**	79.06%	**0.000**

*Note: * marks unseen voices during training. P-value reports a two-sided T-test comparing UPW and SDPW.*

UPW is a state-of-the-art vocoder!

🎨 UPW vs. PWGAN / WGlow / UWRNN

- Test set statistics:

Test set	Recording quality	# Voices (seen / unseen)	# Styles (seen / unseen)	# Lang. (seen / unseen)	# Utt. (**all unseen**)	Vocoder systems
Internal	Very high	19 (15/4)	2 (1/1)	14 (14/0)	1,700	UPW, UWRNN, PWGAN, WGlow
LibriTTS clean	High	30 (0/30)	1 (1/0)	1 (1/0)	300	
LibriTTS other	Medium	30 (0/30)	1 (1/0)	1 (1/0)	300	
Common Voice	Low	300 (0/300)	1 (1/0)	15 (14/1)	300	

- MUSHRA evaluation results:

MUSHRA	Recording	PWGAN	WGlow	UWRNN	UPW	UPW Relative	P-value
Internal	66.81	56.02	50.09	61.83	**63.35**	94.82%	**0.000**
LibriTTS clean	70.42	67.40	66.72	68.30	**69.56**	98.77%	**0.000**
LibriTTS other	68.91	65.04	64.15	63.83	**67.28**	97.64%	**0.000**
Common Voice	64.84	57.84	**58.67**	54.87	58.07	89.56%	0.015

Note: P-value reports a two-sided T-test comparing UPW and the best-among-the-rest vocoder.

amazon

amazon alexa