UNIVERSAL NEURAL VOCODING WITH PARALLEL WAVENET

Yunlong Jiao[⋆], *Adam Gabryś*[⋆], *Georgi Tinchev*[†], *Bartosz Putrycz*[⋆], *Daniel Korzekwa*[⋆], *Viacheslav Klimkov*[⋆]

[⋆] Amazon.com [†] University of Oxford

ABSTRACT

We present a universal neural vocoder based on Parallel WaveNet, with an additional conditioning network called Audio Encoder. Our universal vocoder offers real-time high-quality speech synthesis on a wide range of use cases. We tested it on 43 internal speakers of diverse age and gender, speaking 20 languages in 17 unique styles, of which 7 voices and 5 styles were not exposed during training. We show that the proposed universal vocoder significantly outperforms speaker-dependent vocoders overall. We also show that the proposed vocoder outperforms several existing neural vocoder architectures in terms of naturalness and universality. These findings are consistent when we further test on more than 300 open-source voices.

***Index Terms*—** Neural vocoder, Text-to-speech, Scalability

1. INTRODUCTION

As voice-based human-machine interaction becomes an increasingly crucial part in artificial intelligence, text-to-speech (TTS) remains an important yet challenging problem. While some TTS systems synthesise speech from normalised text or phonemes in an end-to-end manner [1, 2], most TTS systems address the problem in a two-step approach. The first step transforms text to a lower-resolution intermediate representations, such as time-aligned acoustic features [3], or spectral features such as mel-spectrogram [4, 5]. The second step transforms the intermediate representations to high-fidelity audio signal using a model, referred to as *vocoder*.

State-of-the-art vocoders are neural network-based generative models [6, 7, 8, 9, 10, 11]. Neural vocoders are capable of synthesising natural-sounding speech, but typically prone to overfitting to the training data, and do not generalise well to unseen voices [12]. Training speaker-dependent vocoders require significant computational resources and large amounts of audio data for each target speaker [13]. The need for a high-quality speaker-independent vocoder, or so-called *universal* vocoder, is key to scaling up production of TTS systems that are specifically designed to support many voices.

A few recent studies investigated the possibility of building universal vocoders. There were some early reports that speaker-independent vocoders underperform speaker-dependent vocoders [14, 15]. Deep Voice 2 [16] modelled speaker identities by using trainable speaker embeddings in their vocoder as part of a multi-speaker TTS system. These systems require to model speaker identities explicitly, hence cannot handle unseen speakers out-of-the-box. There are also reports of neural vocoders that are capable of synthesising unseen speakers without having to explicitly model speaker identities [17, 10, 9, 11]. However, none of these vocoders were thoroughly evaluated to claim universality. In particular, it was not clear how well a speaker-independent vocoder performs on a target voice compared to a dedicated vocoder built specifically for that voice. The closest setting to ours is the work of Lorenzo-Trueba et al. [18], where a WaveRNN-based universal vocoder is capable of synthesising a wide range of speakers, styles, and conditions. Unfortunately, Universal WaveRNN is autoregressive, and thus inherently slow in sample generation, posing significant difficulties for most real-time applications. To the best of our knowledge, it remains unclear whether any non-autoregressive neural vocoders can be universal.

The contributions of this work are: 1) We present a universal neural vocoder based on Parallel WaveNet [8]. The key component of our universal vocoder is an additional conditioning network, called *Audio Encoder*, which auto-encodes reference waveforms into utterance-level global conditioning. 2) Based on a large-scale evaluation, we show that the proposed universal vocoder significantly outperforms speaker-dependent vocoders overall. It is capable of synthesising a wide range of in-domain and out-of-domain voices, speaking styles, and languages. 3) We perform extensive benchmark studies on internal and open-source voices comparing several existing neural vocoder architectures in terms of naturalness and universality. Results show that our universal vocoder has a clear advantage against other candidates.

2. SYSTEM DESCRIPTION

2.1. Parallel WaveNet

Parallel WaveNet (PW) [8] is a non-autoregressive neural vocoder architecture that transforms a sequence of input noise into audio waveforms in parallel. It can synthesise samples very efficiently by fully exploiting the computational power of modern deep learning hardware.

In our early experiments, we found that PW trained on a

Corresponding author email: vklimkov@amazon.pl. Work done while Georgi Tinchev was an intern at Amazon. We would like to thank Alexis Moinet and Vatsal Aggarwal for insightful research discussions.

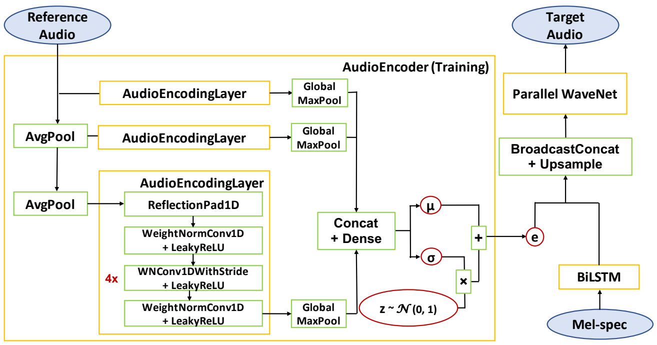

Fig. 1: Universal Parallel WaveNet with Audio Encoder.

multi-speaker dataset underperforms speaker-dependent PW. We conjecture that it can be empirically difficult to obtain a non-autoregressive vocoder that is able to faithfully reconstruct the phase structure of speech signals from speakers with diverse age, gender, speaking styles, and languages.

2.2. Universal Parallel WaveNet with Audio Encoder

Our universal vocoder is based on PW, that generates speech by conditioning on mel-spectrogram. In order to make PW universal, we propose an additional conditioning network called *Audio Encoder*, designed to explicitly model aspects of speech signals that are not provided by the mel-spectrogram conditioning. The Audio Encoder encodes a reference waveform into a fixed-dimensional feature vector, which is then fed as utterance-level global conditioning into PW. In the rest of the paper, we refer to PW with additional Audio Encoder conditioning as Universal Parallel WaveNet (UPW).

The block diagram of the conditioning networks of the proposed UPW is shown in Figure 1. The vocoder model has two conditioning networks — an Audio Encoder and a mel-spectrogram conditioner. First, the Audio Encoder is a multi-scale architecture of an audio feature extractor, heavily inspired by the design of MelGAN's discriminator [10]. It consists of 3 identical audio encoding layers that operate on different time-scales of the reference waveform, achieved by average pooling in between layers. Each audio encoding layer uses a sequence of strided convolutional layers with a large kernel size [10, Appendix A], where each convolutional layer is weight normalised and activated by Leaky ReLU, and each outputs 16 channels in the last layer followed by global max pooling and a dense layer. To prevent information leakage to the vocoder, we applied amortised variational encoding [19] to the output of Audio Encoder. In the end, we obtain a total of 48-dimensional audio feature vector as an utterance-level global conditioning. Second, we adopt the mel-spectrogram conditioner proposed by [3], consisting of 2 bidirectional LSTMs with a hidden size of 128 channels. The mel-spectrogram was extracted from ground-truth audio with 80 coefficients and frequencies ranging from 50 Hz to 12 kHz. Finally, outputs from the two conditioning networks are broadcast–concatenated, and upsampled by repetition from frame-level (80 Hz) to sample-level (24 kHz).

During training, the target waveform is naturally used as the reference waveform. This way, the entire architecture can be viewed as a conditioned Variational Auto-Encoder (VAE) on audio waveform, where the Audio Encoder network is an encoder and PW network is a decoder conditioned on mel-spectrogram. Once the model is trained, doing inference with the proposed UPW requires a reference waveform input to the encoder, or use a pre-generated audio features in place of the output of Audio Encoder conditioner. We investigated several inference strategies (e.g. using speaker-specific or style-specific centroid embedding in place of $\mathbf{e}$), and found that using $\mathbf{e} = \mathbf{0}$ generates high-quality audio (Figure 1). In fact, $\mathbf{e} = \mathbf{0}$ corresponds to the speaker-agnostic centroid embedding of the VAE prior distribution, and using it for inference is shown to improve generalisation of UPW to unseen voices. This simple yet effective approach results in a UPW with little computational overhead compared to a basic PW, therefore both sharing the same real time factor in production.

Following the “teacher-student” training paradigm [8], we first train a Universal WaveNet teacher, and then train a UPW student from it. For the teacher network, we use 24 layers with 4 dilation doubling cycles, 128 residual/gating/skip channels, kernel size 3, and output distribution of a 10-component mixture of Logistics. For the student network, we use [10, 10, 10, 30] flow layers with dilation reset every 10 layers, 64 residual channels, and no skip connections. Both models were trained on sliced mel-spectrogram conditioning corresponding to short audio clips, with the Adam optimizer [20] and a constant learning rate 10^{-4} until convergence. The teacher uses batch size 64 and 0.3625s of audio clips. The student uses batch size 16 and 0.85s of audio clips. During distillation, student reuses the pre-trained conditioning networks of the teacher (including the Audio Encoder). We found that training conditioning networks from scratch often leads to a worse student, a phenomenon also observed in [1].

3. EXPERIMENTAL PROTOCOL

Our training and evaluation protocol was inspired by [18], and adapted with a particular focus on universal vocoding of speech. We collected a multi-speaker multi-lingual training set for the proposed universal vocoder. It consists of 78 different internal, high-quality voices (20 males and 58 females) with approximately 3,000 utterances per speaker, and a total of 28 languages (including dialects) in 16 unique speaking styles (e.g. neutral, long-form reading, and several emotional styles in different degrees of intensity). This training set was designed with the expectation that vocoders should be capable of synthesising a variety of voices, styles, and languages.

We perform analysis–synthesis on natural recordings, and design two types of evaluations on re-synthesised samples.[1]

- In Section 4.1, we compare the proposed UPW with speaker-dependent PW (SDPW) on internal voices for

[1]We also performed experiments on TTS samples using spectrograms generated by a Tacotron2-based architecture. Conclusions drawn from re-synthesised samples and TTS samples remain consistent.

Section	Test sets	Recording quality	# Voices (seen/unseen)	# Styles (seen/unseen)	# Languages (seen/unseen)	# Utterances (all unseen)	Systems
4.1	Internal	Very high	24 (21/3)	16 (12/4)	13 (13/0)	3124	UPW, SDPW [8]
4.2	Internal	Very high	19 (15/4)	2 (1/1)	14 (14/0)	1700	UPW, UWRNN [18], PWGAN [11], WGlow [9]
	LibriTTS-clean [21]	High	30 (0/30)	1 (1/0)	1 (1/0)	300	
	LibriTTS-other [21]	Medium	30 (0/30)	1 (1/0)	1 (1/0)	300	
	Common Voice [22]	Low	300 (0/300)	1 (1/0)	15 (14/1)	300	

Table 1: Summary of test sets. Note that "seen/unseen" refers to whether they were exposed during training.

which we have trained a high-quality SDPW. We show that the proposed vocoder is universal, in the sense that it does not show degradation when compared to speaker-dependent vocoders specific for each voice.

- In Section 4.2, we benchmark UPW vs several other popular neural vocoder architectures in terms of universality. The competing vocoders include Universal WaveRNN (UWRNN) [18], Parallel WaveGAN (PWGAN) [11], and WaveGlow (WGlow) [9]. All of these systems were retrained on the same training set as our UPW, using an open-source implementation or reimplementing the default setup of each paper. We evaluate these vocoders on internal high-quality voices as well as external voices that are recorded in vastly different conditions.

Table 1 summarises the statistics of the test sets in our experiments. Note that test voices are selected such that they are balanced according to gender and age.

The naturalness perceptual evaluation was designed as a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [23], where participants were presented with the systems being evaluated side-by-side, asked to rate them in terms of naturalness and audio quality (glitches, clicks, noise, etc.) from 0 (poorest) to 100 (best). Each test utterance is evaluated by 10 to 15 listeners that are either native or educated speakers of the target language. We also include recordings in all MUSHRA tests as the hidden upper-anchor system, and we do not force at least one 100 rated system.

Paired two-sided Student T-tests with Holm-Bonferroni correction were used to validate the statistical significance of the differences between two systems at a p-value threshold of 0.05. We refer to the ratio between the mean MUSHRA score of a system and natural recordings as relative MUSHRA (denoted by Rel.) Relative MUSHRA illustrates the gap between the system being evaluated with the reference.

4. RESULTS

4.1. Comparison with speaker-dependent vocoders

In this section, we compare the proposed universal vocoder UPW with speaker-dependent vocoder (SDPW). MUSHRA evaluations are carried out on 24 high-quality internal voices in distinct speaking styles and languages, using Amazon's internal evaluation platform where listeners are professionally trained for voice quality assessment.

Results show that UPW significantly outperforms SDPW overall with a relative MUSHRA of 84.24% vs 83.12% (p-value = 0). As we break down to each voice, UPW shows statistically significant improvement to SDPW on 7 voices,

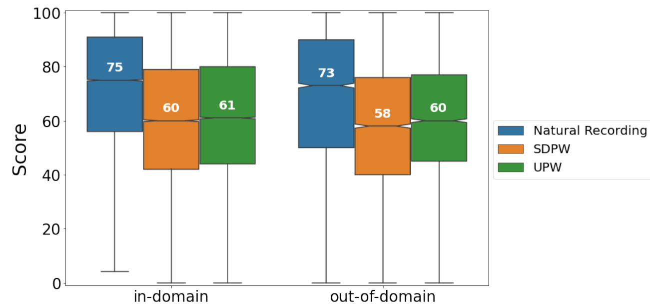

Fig. 2: MUSHRA evaluation for comparison with speaker-dependent vocoders.

Voice (Sex)	Age	Rec.	SDPW	UPW	UPW Rel.	p-val.
British Eng. (F)	Adult	71.64	65.69	**67.67**	94.45%	**0.000**
Australian 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
Overall		69.68	57.92	**58.70**	84.24%	**0.000**

Table 2: MUSHRA scores on internal voices (unseen marked by *). p-value signifies the difference between UPW vs SDPW. Note that the voices are selected so that the relative MUSHRA of UPW is evenly distributed within the range from highest to lowest.

Style	Rec.	SDPW	UPW	UPW Rel.	p-val.
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**

Table 3: MUSHRA scores on typical styles. p-value signifies the difference between SDPW vs UPW.

and is comparable to SDPW on 17 voices without statistically significant difference. Table 2 lists some voices for which the relative MUSHRA achieved by UPW is evenly distributed within the range from highest (94.45%) to lowest (65.62%). This is a very strong result for the proposed universal vocoder, since not only can it avoid degradation compared to speaker-specific vocoders on all tested voices, but also it improves the vocoding quality on many voices by using information contained in the speech signal of related speakers. Moreover, Figure 2 shows that UPW consistently outperforms SDPW on voices and speaking styles that were either exposed during training (in-domain) or not (out-of-domain).

We now focus our evaluation on different speaking styles,

Voice (Sex)	Rec.	PWGAN	WGlow	UWRNN	UPW	UPW Rel.
Italian (M)	65.97	55.05	50.67	59.46	**65.83**	99.78%
*Welsh (F)	65.67	57.12	57.25	63.50	**65.17**	99.23%
Korean (F)	70.60	64.90	56.20	60.26	**68.16**	96.54%
Polish (F)	67.21	57.83	47.39	64.35	**64.68**	96.24%
French (M)	65.14	51.34	47.68	57.94	**61.20**	93.96%
German (M)	68.61	60.86	50.93	**66.27**	61.57	89.75%
Polish (M)	65.12	54.61	46.23	**58.69**	53.28	81.82%
Overall	66.81	56.02	50.09	61.83	**63.35**	94.82%

Table 4: MUSHRA scores on internal voices (unseen marked by *). All speakers are adult. Note that the voices are selected so that the relative MUSHRA of UPW is evenly distributed within the range from highest to lowest.

as we found that it is usually challenging to vocode highly expressive speech even for a well-trained SDPW. Table 3 summarises some typical styles in our evaluation. We find that UPW is comparable to SDPW on neutral, emotional (e.g. excited, disappointed), and long-form reading style. For some expressive styles such as conversational, news briefing, and singing style, UPW statistically significantly outperforms SDPW. In particular, the most challenging style we consider in our evaluation is singing, because we believe it is the most expressive type of speech. While both UPW and SDPW indeed achieved the least relative MUSHRA on this style (79.06% vs 69.46%), UPW sees the greatest improvement from SDPW. Notably, UPW outperforms SDPW on singing style by 6.91 MUSHRA points on average, closing the gap between recordings and SDPW by 31.44%. This strongly evidences the superiority of the proposed universal vocoder.

4.2. Comparison with other multi-speaker vocoders

In this section, we compare UPW with other popular neural vocoder architectures, including Universal WaveRNN (UWRNN) [18], Parallel WaveGAN (PWGAN) [11] and WaveGlow (WGlow) [9]. Note that MUSHRA evaluations in this section are carried out by Clickworker [24].

Internal voices. We first evaluate competing systems on 19 high-quality internal voices. The results in Table 4 show that UPW is the best-performing vocoder overall (p-value $= 0$), achieving the highest average relative MUSHRA of 94.82% among all four competing vocoders. Some voices are listed for which the relative MUSHRA achieved by UPW is evenly distributed within the range from highest (99.78%) to lowest (81.82%). Compared to the other non-autoregressive candidates, UPW statistically significantly outperforms WGlow on all 19 tested voices. UPW statistically significantly outperforms PWGAN on 16 voices, and both systems are comparable on 3 voices without statistically significant difference. Compared to the autoregressive candidate, UPW is statistically significantly better than UWRNN on 13 voices, both are comparable on 2 voices, but UPW underperforms UWRNN on 4 voices. However, it is worth noting that, due to its autoregressive nature, UWRNN has an inference speed that is slower than UPW typically by orders of magnitude.

Dataset	Rec.	PWGAN	WGlow	UWRNN	UPW	UPW Rel.
LibriTTS-clean	70.42	67.40	66.72	68.30	**69.56**	98.77%
LibriTTS-other	68.91	65.04	64.15	63.83	**67.28**	97.64%
Common Voice	64.84	57.84	**58.67**	54.87	58.07	89.56%

Table 5: MUSHRA scores on external voices.

External voices. We further study the robustness of the pre-trained UPW on open-source voices. To this end, we prepared three test sets with decreasing recording quality: LibriTTS-clean [21], LibriTTS-other [21], and Common Voice [22]. Note that LibriTTS is a multi-speaker corpus of English speech in audiobook reading style, and Common Voice is a database of multi-lingual multi-speaker user recording. The results in Table 5 show that UPW is a top-performing system across all three sets of external voices. (a) On high-quality voices (LibriTTS-clean), UPW consistently outperforms all other systems, achieving a relative MUSHRA of 98.77%. This implies that UPW can generalise well to out-of-domain voices when the recording conditions are studio-quality. (b) On medium-quality voices (LibriTTS-other), UPW has a clear advantage over other systems. This strongly suggests that UPW is still capable of synthesising naturally sounding speech in the presence of a reasonable level of noise. (c) On low-quality voices (Common Voice), WGlow and UPW are in fact comparably good (p-value $= 0.15$), while UWRNN is the least robust to recording conditions where a significant amount of background noise is present in low-quality speech.

5. CONCLUSION

In this work, we presented a universal neural vocoder based on Parallel WaveNet, trained on a multi-speaker multi-lingual speech dataset. It is capable of synthesising a wide range of voices, styles, and languages, and particularly suitable for scaling up production of real-time TTS. The key component of the proposed universal vocoder is an additional conditioning network called Audio Encoder, which auto-encodes reference waveforms into utterance-level global conditioning. Based on large-scale evaluation, our universal vocoder outperforms speaker-dependent vocoders overall. We have also reported extensive studies benchmarking several existing neural vocoder architectures in terms of naturalness and universality, and showed that our universal vocoder has a clear advantage of being non-autoregressive and superior in terms of quality in a vast majority of cases.

There are still interesting research directions we will leave to future work. First, it is interesting to generalise the proposed Audio Encoder to encode reference waveforms into local conditioning, which can better represent the localised features of speech signals. Second, we can study whether the proposed Audio Encoder would also benefit multi-speaker training of other neural vocoders such as WaveGlow [9] and Parallel WaveGAN [11]. Third, it is worth investigating how our universal vocoder performs on challenging vocoding scenarios, such as overlapping voices, and non-speech vocalisations (e.g. shouts, breath, reverberation).

6. REFERENCES

[1] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *International Conference on Learning Representations (ICLR)*, 2019.

[2] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," *CoRR*, vol. abs/2006.03575, 2020.

[3] S. Ö. Arik, M. Chrzanowski, A. Coates, et al., "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning ICML*, 2017, pp. 195–204.

[4] Y. Wang, R. J. Skerry-Ryan, D. Stanton, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[5] J. Shen, R. Pang, R. J. Weiss, et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2018, pp. 4779–4783.

[6] A. van den Oord, S. Dieleman, H. Zen, et al., "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop*, 2016, p. 125.

[7] N. Kalchbrenner, E. Elsen, K. Simonyan, et al., "Efficient neural audio synthesis," in *Proceedings of the 35th International Conference on Machine Learning ICML*, 2018, pp. 2415–2424.

[8] A. van den Oord, Y. Li, I. Babuschkin, et al., "Parallel wavenet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning ICML*, 2018, pp. 3915–3923.

[9] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2019, pp. 3617–3621.

[10] K. Kumar, R. Kumar, T. de Boissiere, et al., "Melgan: Generative adversarial networks for conditional waveform synthesis," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 14881–14892.

[11] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2020, pp. 6199–6203.

[12] S. Ö. Arik, H. Jun, and G. F. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 94–98, 2019.

[13] X. Wang, J. Lorenzo-Trueba, S. Takaki, L. Juvela, and J. Yamagishi, "A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis," in *ICASSP*. 2018, pp. 4804–4808, IEEE.

[14] O. Barbany, A. Bonafonte, and S. Pascual, "Multi-speaker neural vocoder," in *Fourth International Conference on IberSPEECH*. 2018, pp. 30–34, ISCA.

[15] E. Song, J. Kim, K. Byun, and H. Kang, "Speaker-adaptive neural vocoders for statistical parametric speech synthesis systems," *CoRR*, vol. abs/1811.03311, 2018.

[16] A. Gibiansky, S. Ö. Arik, G. F. Diamos, et al., "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 2962–2970.

[17] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for wavenet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop ASRU*, 2017, pp. 712–718.

[18] J. Lorenzo-Trueba, T. Drugman, J. Latorre, et al., "Towards achieving robust universal neural vocoding," in *Interspeech*, 2019, pp. 181–185.

[19] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR*, 2014.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations ICLR*, 2015.

[21] H. Zen, V. Dang, R. Clark, et al., "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Interspeech 2019*, 2019, pp. 1526–1530.

[22] Common Voice Database, "https://voice.mozilla.org/," March, 2020.

[23] I. Recommendation, "BS. 1534-1. method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunications Union, Geneva*, 2001.

[24] Clickworker, "https://www.clickworker.com/machine-learning-ai-artificial-intelligence/," August, 2020.