TOWARDS ROBUST NEURAL VOCODING FOR SPEECH GENERATION: A SURVEY

Po-chun Hsu *Chun-hsuan Wang* *Andy T. Liu* *Hung-yi Lee*

National Taiwan University
College of Electrical Engineering and Computer Science
{r07942095, r07942076, r07942089, hungyilee}@ntu.edu.tw

ABSTRACT

Recently, neural vocoders have been widely used in speech synthesis tasks, including text-to-speech and voice conversion. However, in the encounter of data distribution mismatch between training and inference, neural vocoders trained on real data often degrade in voice quality for unseen scenarios. In this paper, we train three commonly used neural vocoders, including WaveNet, WaveRNN, and WaveGlow, alternately on five different datasets. To study the robustness of neural vocoders, we evaluate the models using acoustic features from seen/unseen speakers, seen/unseen languages, a text-to-speech model, and a voice conversion model. In this work, we found that WaveNet is more robust than WaveRNN, especially in the face of inconsistency between training and testing data. Through our experiments, we show that WaveNet is more suitable for text-to-speech models, and WaveRNN more suitable for voice conversion applications. Furthermore, we present results with considerable reference value of subjective human evaluation for future studies.

Index Terms— neural vocoder, robustness, raw waveform synthesis, text-to-speech, voice conversion

1. INTRODUCTION

Most speech generation models, such as text-to-speech [1, 2, 3, 4] and voice conversion [5, 6, 7, 8], do not output waveform directly. Instead, the models output features such as Mel-spectrograms or F0 frequencies. Traditionally, waveform can be vocoded from these acoustic or linguistic features using heuristic methods [9] or handcrafted vocoders [10, 11, 12]. However, due to the assumptions used by heuristic methods, the quality of generated speech is largely limited and undermined.

Since Tacotron 2 [3] first applied WaveNet [13] as a vocoder to reconstruct waveform from Mel-spectrograms, neural vocoders have gradually become the most commonly used vocoding method for speech synthesis. Nowadays, neural vocoders have replace the use of traditional heuristic methods and dramatically enhances the quality of generated speech. WaveNet generates waveform in high quality but at the cost of long inference time due to its autoregressive architecture. To solve this problem, many architectures such as FFTNet [14], WaveRNN [15], and WaveGlow [16] have been proposed.

Neural vocoders can successfully model the data distribution of human voice with acoustic features [13, 14, 15, 16, 17], however the generated speech quality is still restricted by the consistency of training and testing domain due to deep learning's data-driven property. Recently, [18] reported that a WaveRNN-based neural vocoder trained on multi-speaker multilingual data can generate natural speech despite conditions from an unseen domain. However there is still much to be studied about the robustness between different vocoders and their applications on various speech generation tasks. In this paper, we survey a variety of neural vocoders trained on datasets across different domains applied to several scenarios. The contributions of this work are:

- We alternatively train 3 commonly used neural vocoders on 5 datasets, these 15 pre-trained models are publicly available.
- The performances of neural vocoders are investigated by testing on human speech, output of voice conversion and output of test-to-speech.
- We analyze the robustness of neural vocoders in different scenarios based on mean opinion score (MOS) survey.

In Section 2, we first introduce all the vocoder architectures used in this paper. In Section 3, we introduce the datasets and specify the evaluation metrics. We present three conducted experiments in Section 4, 5, 6. In Section 4, we evaluate the trained vocoders on real-world data. In Section 5, we analyze the influence of the speaker's gender on vocoders. In Section 6, we trained vocoders to speech synthesis tasks. We then conclude our results in Section 7.

2. NEURAL VOCODERS

2.1. WaveNet

WaveNet [13] uses an autoregressive model to generate raw waveform. The network architecture is composed of layers of dilated causal convolution with gated activation units [19] for non-linearity. Our WaveNet model is modified from the public implementation [20], with 30 layers, 3 dilation cycles, 128 residual channels, 256 gate channels, and 128 skip channels. The input and output are quantized to 8-bit vectors using μ-law companding transformation [21]. We trained the model with a batch size of 6 on a single NVIDIA 2080Ti GPU for 250k iterations, which takes 2 days to converge.

2.2. WaveRNN

Our version of WaveRNN [15] is composed of two parts. One is the conditioning module, and the other is the autoregressive module. In our approach, the origin waveform is quantized to integers ranging from 0 to 1023. Then we optimize the model under a classification problem. The WaveRNN model we used is based on the public implementation [22]. The conditioning module consists of upsampling layers with residual connections. The network is trained with

This work was supported by NVIDIA.

Dataset	Speaker	Utterance	Language
LJ Speech (lj)	1	13100	English
Sing_M (sim)	1	9004	Mandarin
Multi_M (mum)	7	-	Mandarin
Librispeech (libri)	8	-	English
VCTK (vctk)	109	44070	English
Global (gl)	433	49885	Multiple

Table 1. Overview of the datasets. Note that Multi_M and Librispeech are only used for testing.

a batch size of 32 on a single NVIDIA V100 for 300k iterations and converged in 1 day.

2.3. WaveGlow

Based on Glow [23], the WaveGlow [16] model can be considered as an invertible transformation. The model learns to map utterances from raw waveform domain to the Gaussian distribution domain when given corresponding conditions (e.g. Mel-spectrograms) during training. At inference, the model can be inverted to generate utterances with random Gaussian noise and given conditions. Our version of WaveGlow is based on the official implementation [24] by NVIDIA. This model is composed of 10 layers of flows, each of which has 5 dilated convolutional layers with a channel size of 256. The model was trained for 5 days with a batch size of 8 on a single NVIDIA V100 GPU for 550k iterations.

3. DATASETS AND EVALUATION METRICS

3.1. Datasets for experiments

A total of 6 datasets were used in this paper. All of the datasets are listed in Table 1. LJ Speech (lj) [25] is a commonly used dataset for training text-to-speech model and neural vocoders. It includes 13100 clean utterances recorded from a female English speaker. Besides, we use two different internal Mandarin datasets, Sing_M[1] (sim) and Multi_M (mum). Sing_M is collected from a female speaker, and Multi_M is from multiple speakers. Sing_M is similar to LJ Speech in terms of quality and the number of utterances. Multi-speaker English datasets, including VCTK (vctk) [26] and Librispeech (libri) [27], are also used. The main difference between these two datasets is that VCTK is noisier. GlobalPhone [28] is a multi-lingual dataset with multiple speakers. The Global (gl) dataset mentioned in this paper consists of 5 languages, including Czech, French, German, Spanish from GlobalPhone and English from VCTK. Audios from these datasets are downsampled to a sampling rate of 16,000, and 10 utterances were split from them as test sets. Input condition for all vocoders is Mel-spectrogram with 80 coefficients.

3.2. Augmentation

Text-to-speech and voice conversion models may output blurry Mel-spectrograms, which lead to undermined and artificial speech. To enhance the quality of voice, augmentation is added to our training scenario. Inspired by SpecAugment [29], we add time or frequency masks on Mel-spectrogram randomly. Furthermore, a 2D moving average filter is added randomly on Mel-spectrogram. In addition to augmenting in spectrogram domain, white noise is added to input Mel-spectrogram. Time warping is added as well. The model is trained to output clean speech conditioned on the augmented Mel-spectrograms. The augmentation scenario is applied on LJ Speech, denoted as Augmentation (aug) on the following tables and discussions. The final size of Augmentation is three times to the origin LJSpeech.

[1]The Sing_M datasest is collected by Taiwan AI Labs.

3.3. Evaluation metrics

We conduct Mean Opinion Score (MOS) tests to rate the quality of the generated speech. Each utterance is scored based on its naturalness on a 1-to-5 scale. A higher score signifies a more natural utterance. Each reported MOS is an average of 10 utterances. More than 75 human participants are surveyed in the test, and each utterance was rated by at least 5 listeners. For each utterance, outlier scores are removed.

4. ROBUSTNESS TO REAL-WORLD DATA

In this section, we consider synthesized speech conditioned on Mel-spectrograms from human speech.

4.1. Experimental setup

3 Vocoders (WaveNet, WaveRNN, WaveGlow) are trained on 5 different datasets (Sing_M, LJ Speech, VCTK, Global, Aug), which results in a total of 15 models. To observe how the models will perform differently in the face of inconsistent train/test scenarios, all 15 models are tested on seen/unseen speakers and seen/unseen languages settings. For the scenario of seen speakers and seen languages (SS), vocoders are trained and evaluated on training and testing data from the same datasets. For the vocoders trained on the Mandarin dataset, Sing_M (sim), the test set for unseen speakers and seen language (US) is a multi-speaker Mandarin dataset, Multi_M (mum). Besides, the test set for unseen speakers and unseen language (UU) is a multi-speaker English dataset, Librispeech (libri). For vocoders trained on English or multiple languages datasets, such as LJ Speech, VCTK, Global, the test set for unseen speakers and seen language (US) is a multi-speaker English dataset, Librispeech (libri). Furthermore, the test set for unseen speakers and unseen language (UU) is a Mandarin dataset, Multi_M (mum). There will be only 3 testing scenarios (SS, US, UU) since it is difficult to obtain data that contains speech data with seen speakers and unseen languages.

4.2. Results

In the following sections, we have reported experimental results of the WaveGlow model. However, we did not involve the WaveGlow model in our further comparison and discussions, because the WaveGlow model did not converge and performs badly even though it was given a much longer training time.

4.2.1. Seen Speakers and Seen language

In the 1st row of Table 2, the WaveRNN model trained on Sing_M (sim) performs better over the WaveNet trained on the same dataset. Similar results can be found on other columns except for *aug* column. On column *vctk* and *gl*, the quality of utterances in VCTK and Global are both disturbed by background noise. The MOS on the

Test Set \ Models		WaveNet					WaveRNN					WaveGlow				
Spk	**Lang**	**sim**	**lj**	**vctk**	**gl**	**aug**	**sim**	**lj**	**vctk**	**gl**	**aug**	**sim**	**lj**	**vctk**	**gl**	**aug**
seen	**seen**	3.75	4.12	3.23	2.92	3.85	**4.52**	4.48	3.80	3.90	3.75	3.18	2.49	1.99	1.77	2.60
unseen	**seen**	2.85	3.46	**4.02**	3.97	2.72	2.15	3.31	3.80	3.88	1.57	2.19	2.32	2.23	1.63	1.98
unseen	**unseen**	3.51	**3.63**	3.35	3.38	3.25	2.58	2.07	2.88	3.13	1.03	2.34	1.93	2.35	1.81	1.82

Table 2. MOS of different models on test set with speaker and language property specified
On the 2nd row, all models are tested on *libri* except for *sim*, which is tested on *mum*.
On the 3rd row, all models are tested on *mum* except for *sim*, which is tested on *libri*.

columns is lower than those on column *sim* and *lj* by about 1 point. We found that the purity of the training data affects the quality of generated speech. The reduction of MOS can be observed in column *aug* as well.

4.2.2. Unseen speakers and seen language

In the 2nd row of Table 2, the WaveNet model performs better for out-of-domain speakers. It is observed that vocoders trained on VCTK and Global perform much better than vocoders trained on LJ Speech and Sing_M. We then infer that the more diverse the speakers are in training dataset, the better it performs for unseen speakers.

However, we note that the vocoders trained on LJ Speech and Sing_M lack male speech during training, hence it cannot perform well on male speakers during inference time. More survey about gender will be discussed in Section 5.

The output of the WaveRNN model degrades much with inconsistent data distribution in training and testing procedures. Hence, it performs badly for unseen speakers. One of the main reasons that trained on LJ Speech and Sing_M is that WaveRNN is very sensitive to unseen gender, shown in Section 5. With provided sufficient speakers, the WaveRNN model can perform close to the result of WaveNet with much light-weighted and faster inference time.

4.2.3. Unseen speakers and unseen language

For unseen languages, WaveNet's performance is comparable to the case of seen language. Whereas WaveRNN degrades a lot in terms of performance, as it is sensitivity to the mismatch of seen/unseen languages.

4.2.4. Discussion

To conclude this section, we show how robust vocoders are when trained on different datasets and tested on in-domain/out-of-domain scenarios. For in-domain data, WaveRNN vocoder performs better over other models. However, it degrades when tested on out-fo-domain data. In the 2nd and 3rd row of Table 2, when WaveNet models are tested on out-of-domain data, we observe that not all MOS declines. Since the quality between test sets are different, we then conclude that the WaveNet model is robust to unseen speakers or languages but mainly affected by the quality of testing data. Data augmentation on LJ Speech does not help improve the robustness but make the performance of models worse.

5. THE INFLUENCE OF GENDERS

In Section 4, we survey how unseen speakers influence neural vocoder models. However, for vocoders trained on single female datasets (e.g. LJ Speech and Sing_M), we cannot figure the performance degradation on unseen speakers is caused by unseen speakers or unseen gender. To investigate more, we conduct the following experiment to explore how vocoder's behaviour is influenced by speaker gender.

Training set	Language	Gender	WN	WR	WG
LJ Speech (Female)	Seen (libri)	Female	3.39	3.59	2.24
		Male	3.53	3.03	2.40
	Unseen (mum)	Female	3.48	2.25	2.65
		Male	3.79	1.89	2.03
Sing_M (Female)	Seen (mum)	Female	3.07	2.66	2.55
		Male	2.62	1.63	1.83
	Unseen (libri)	Female	3.30	2.92	2.11
		Male	3.71	2.25	1.76

Table 3. Gender analysis of different models. WN, WR, WG stands for WaveNet, WaveRNN, WaveGlow, respectively

5.1. Experimental setup

In this section, to discuss model sensitivity on unseen gender, neural vocoders trained on single speaker datasets (e.g. LJ Speech and Sing_M) will be considered. The model will be tested on unseen speakers (Librispeech and Multi_M). To avoid deviation from the testing set, we first evaluate the MOS of ground-truth male and female speech from the dataset. The Librispeech dataset has a MOS of 4.62, with scores of 4.82 and 4.42 for male and female, respectively. The Multi_M dataset has a score of 4.78, and scores of male and female voices are 4.95 and 4.61.

5.2. Results

The results are listed in Table 3. For both given training set, the testing result of the WaveRNN model to female speakers perform much better to male speakers. Hence, one of the reasons that affects performance of the WaveRNN model is the unseen gender. The WaveNet model performs evenly regardless of the speakers' gender, so the degradation for WaveNet model for unseen speakers may be caused by other reasons. We find out that the MOS on output of WaveNet model and on origin testing data are highly correlated. Hence, we conclude that the WaveNet model is quite robust to training data.

6. ROBUSTNESS TO SPEECH SYNTHESIS TASK

The neural vocoder was originally proposed as a vocoder for the text-to-speech model [3]. In this section, we test the performances of vocoders by applying them to speech synthesis tasks.

Vocoder \ Model / Training Set	WN	WR	WG	GL
Sing_M	4.12	3.02	2.46	2.6
LJ Speech	4.18	3.7	2.22	
VCTK	**4.26**	3.02	2.16	
Global	4.02	3.2	1.62	
Augmentation	3.62	2.82	2.14	
Task-specific	4.72	-	-	

Table 4. MOS for text-to-speech (TTS) synthesis. The TTS model is trained on LJ Speech. Conditions of the vocoders are from the TTS model. (WN, WR, WG, GL stands for WaveNet, WaveRNN, WaveGlow, Griffin-Lim respectively)

Vocoder \ Model / Training Set	WN	WR	WG	GL
Sing_M	2.52	2.88	2.1	3.2
LJ Speech	2.08	2.98	2.08	
VCTK	2.96	**3.64**	2.34	
Global	2.38	3.38	2.04	
Augmentation	1.76	1.7	1.92	
Task-specific	-	1.36	-	

Table 5. MOS for voice conversion(VC). The VC model is trained on VCTK. Conditions of the vocoders are from the VC model. (WN, WR, WG, GL stands for WaveNet, WaveRNN, WaveGlow, Griffin-Lim respectively)

6.1. Experimental setup

Neural vocoders are more frequently used to generate audio from the output of upstream speech tasks, such as text-to-speech synthesis model or voice conversion model. Hence, experiments are examined to find out which model can perform better. We also tested a heuristic method, Griffin-Lim algorithm (GL) [9] for comparison.

Tacotron 2 [3] is examined for text-to-speech synthesis. The Mel-spectrogram output of the Tacotron 2 is fed to the vocoders pretrained with different datasets. For Griffin-Lim algorithm, Mel-spectrograms are transformed to linear scale by pseudo inverse matrix. In addition, we trained a vocoder on ground-truth aligned predictions [3] of the Tacotron 2. This vocoder is denoted as Task-specific in the following. We compare the task-specific vocoder to those only trained on real-world data shown in Table 4. The Tacotron 2 model is trained on the LJ Speech, and the vocoder trained on the same dataset is the topline model.

Similarly, we examined the voice conversion model proposed by Chou [5]. The output of the voice conversion model is linear scale, therefore the output is fed to a Mel-filter to get Mel-spectrogram. The Griffin-Lim algorithm will reconstruct signals directly from the linear spectrograms. Pairs of the output from trained voice conversion model and origin VCTK waveform are used to train a task-specific vocoder. The MOS between different vocoders and different training sets are listed in Table 5. Since the voice conversion model is trained on VCTK, the vocoder trained on the same dataset is the topline model.

Both implementation of the text-to-speech[2] and voice conversion[3] are public.

6.2. Results

6.2.1. Text-to-speech synthesis

The WaveNet vocoder performs well for synthesize waveform on the text-to-speech model. The WaveNet vocoder outperforms the Griffin-Lim algorithm no matter the training corpus. The MOS of WaveNet is all above 4 trained on real-world speech, regardless to training dataset. However, augmentated data does not help for testing. For the use for application, WaveNet vocoder is strongly recommended and can be trained from any human speech dataset.

[2] https://github.com/BogiHsu/Tacotron2-PyTorch

[3] https://github.com/BogiHsu/Voice-Conversion

6.2.2. Voice conversion

The result indicates the WaveRNN vocoder performs better in naturalness on the voice conversion model. WaveRNN vocoder trained on VCTK is able to produce speech with higher quality than the Griffin-Lim algorithm. Hence, for application usage, the WaveRNN vocoder is recommended to be used and be trained on the same dataset trained for the voice conversion model.

However, the model performs terribly on Augmentation and Task-specific. This is because of WaveRNN is very sensitive to the quality of training data concluded in Section 4.

7. CONCLUSION

We train WaveNet, WaveRNN, and WaveGlow alternately on five datasets and test out the performance for seen/unseen speakers, seen/unseen languages, a text-to-speech model, and a voice conversion model. We investigate

1. influence of seen/unseen speaker and seen/unseen language in testing
2. influence of the speaker gender in a female speaker dataset
3. suitable vocoder for text-to-speech model and voice conversion model

In experiments, the WaveNet model is more robust when encountering inconsistency between training data and testing data. The WaveRNN model performs well in the same domain on training and testing, but is sensitive to out-of-domain testing data. The WaveGlow model performs worse with limited computational resource for training.

We point out that the augmentation on training neural vocoders degrades on real-world speech data and speech synthesis tasks. For the use of speech synthesis, the real-world data cannot be replaced by augmentation.

8. REFERENCES

[1] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[4] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[5] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee, "Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations," *arXiv preprint arXiv:1804.02812*, 2018.

[6] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," in *Proceedings of the 36th International Conference on Machine Learning*. 09–15 Jun 2019, vol. 97 of *Proceedings of Machine Learning Research*, pp. 5210–5219, PMLR.

[7] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.

[8] Joan Serrà, Santiago Pascual, and Carlos Segura, "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion," *arXiv preprint arXiv:1906.00794*, 2019.

[9] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[10] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[11] Hideki Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[12] Hideki Banno, Hiroaki Hata, Masanori Morise, Toru Takahashi, Toshio Irino, and Hideki Kawahara, "Implementation of realtime straight speech manipulation system: Report on its first implementation," *Acoustical science and technology*, vol. 28, no. 3, pp. 140–146, 2007.

[13] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[14] Zeyu Jin, Adam Finkelstein, Gautham J Mysore, and Jingwan Lu, "Fftnet: A real-time speaker-dependent neural vocoder," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2251–2255.

[15] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

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

[17] Jean-Marc Valin and Jan Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[18] Jaime Lorenzo-Trueba, Thomas Drugman, Javier Latorre, Thomas Merritt, Bartosz Putrycz, Roberto Barra-Chicote, Alexis Moinet, and Vatsal Aggarwal, "Towards achieving robust universal neural vocoding," *arXiv preprint arXiv:1811.06292*, 2018.

[19] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al., "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems*, 2016, pp. 4790–4798.

[20] Rayhane Mama, "Tacotron-2," `https://github.com/r9y9/wavenet_vocoder`, 2018.

[21] CCITT Recommendation, "Pulse code modulation (pcm) of voice frequencies," in *ITU*. 1988.

[22] Rayhane Mama, "Tacotron-2," `https://github.com/G-Wang/WaveRNN-Pytorch`, 2018.

[23] Durk P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Advances in Neural Information Processing Systems*, 2018, pp. 10215–10224.

[24] Rayhane Mama, "Tacotron-2," `https://github.com/NVIDIA/waveglow`, 2018.

[25] Keith Ito, "The LJ speech dataset," `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[26] Kirsten MacDonald et al. Christophe Veaux, Junichi Yamagishi, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[27] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[28] Tanja Schultz, "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University," in *Seventh International Conference on Spoken Language Processing*, 2002.

[29] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.