Restoring degraded speech via a modified diffusion model

Jianwei Zhang, Suren Jayasuriya, Visar Berisha

Arizona State University, USA

{jzhan396,sjayasur,visar}@asu.edu

Abstract

There are many deterministic mathematical operations (e.g. compression, clipping, downsampling) that degrade speech quality considerably. In this paper we introduce a neural network architecture, based on a modification of the DiffWave model, that aims to restore the original speech signal. DiffWave, a recently published diffusion-based vocoder, has shown state-of-the-art synthesized speech quality and relatively shorter waveform generation times, with only a small set of parameters. We replace the mel-spectrum upsampler in DiffWave with a deep CNN upsampler, which is trained to alter the degraded speech mel-spectrum to match that of the original speech. The model is trained using the original speech waveform, but conditioned on the degraded speech mel-spectrum. Post-training, only the degraded mel-spectrum is used as input and the model generates an estimate of the original speech. Our model results in improved speech quality (original DiffWave model as baseline) on several different experiments. These include improving the quality of speech degraded by LPC-10 compression, AMR-NB compression, and signal clipping. Compared to the original DiffWave architecture, our scheme achieves better performance on several objective perceptual metrics and in subjective comparisons. Improvements over baseline are further amplified in a out-of-corpus evaluation setting.

Index Terms: Restoring speech, lossy transformation, speech enhancement, diffusion model, vocoder

1. Introduction

Many algorithms and mathematical operations degrade the quality of speech. For example, speech compression algorithms reduce the sampling rate and use linear predictive coding to compress the input; speech-clipping introduces high-frequency content with a negative impact on quality. Reduced speech quality can impact intelligibility and makes the resulting speech less suitable for downstream applications like automatic speech recognition or speaker identification algorithms. Speech enhancement (SE) of degraded speech is important across many applications including telecommunications [1], speech recognition [2], etc. Many methods have been developed for similar applications, such as speech denoising and dereverberation [3, 4]. Most current SE methods are designed to remove background noise, most are additive noise models. However, the aforementioned operations (e.g. compression, clipping) are non-linear and lossy. The goal of this paper is to restore the degraded speech generated by lossy deterministic transformations.

Broadly speaking, there are two families of SE techniques: based on traditional statistical signal processing and based on machine learning. Traditional methods include statistical model-based methods [5], e.g. spectral subtraction [6], and Wiener filtering [7]. While these methods work well for additive noise conditions, they are not suitable for our application. Enhancement methods based on machine learning models such as diffusion models and U-nets with adversarial loss have resulted in a sizeable improvement in performance [1, 3, 4, 8]. These models can enhance speech quality, however they require complex network structures with a large number of parameters. Our aim is to develop sample-efficient networks trained to invert the lossy transformation and impute the missing information in the signal. We posit that for deterministic transformations (e.g. compression, clipping), we can efficiently learn the inversion and generate high-quality speech by leveraging state-of-art vocoders.

Modern vocoders can generate high-quality speech based on an input conditioner, e.g. a mel-spectrum. A widely used ML-based vocoder is WaveNet [9]. It can synthesize high-quality speech, but the synthesis run-time is slow. WaveFlow is a flow-based ML vocoder with short generation time, however it contains a large number of parameters [10]. Many vocoders (e.g. LPCNet) have recently focused on improving synthesis efficiency [11]. DiffWave, a diffusion model-based vocoder was recently published with state-of-the-art synthesized speech quality, a relatively short waveform generation time, and small number of parameters [12]. However, DiffWave was primarily used for generative modeling tasks such as unsupervised speech generation where the data distribution of audio was learned by the model.

In this paper, our key insight is that a diffusion-based model such as DiffWave can be trained in a supervised fashion to restore degraded speech, particularly for these deterministic operations. To do so, we condition DiffWave on the degraded mel-spectrum of the input speech, and train the network to recover back the original speech. However, we observe that this method only achieves partial recovery of the original speech. To further improve performance, we modify DiffWave network architecture by including a pre-trained inversion network to restore the quality and intelligibility of speech. We replace the upsampling layers in a pre-trained DiffWave model with a deep CNN upsampler, which has the capacity to learn an inversion model that alters the degraded speech mel-spectrum to generate the conditioner for restored speech synthesis by DiffWave model.

To validate our methods, we perform experiments to compare the quality and intelligibility of restored audio when degraded by three deterministic lossy mathematical operations: linear predictive coding (LPC-10) compression, adaptive multi-rate narrow-band (AMR-NB) compression, and signal clipping. We compare the results of the original DiffWave trained in a supervised fashion as well as our modified DiffWave model with inversion module. Our results show that our modified model improves on the original DiffWave model for this application, restoring speech quality and intelligibility on both in-corpus (out-of-sample) and cross-corpus evaluations. In summary, our contributions are: 1) We demonstrate that DiffWave is able to produce better-quality speech, even conditioned on a distorted mel-spectrum, and 2) we modify DiffWave's architecture with a deep CNN upsampling network for the conditioner, resulting
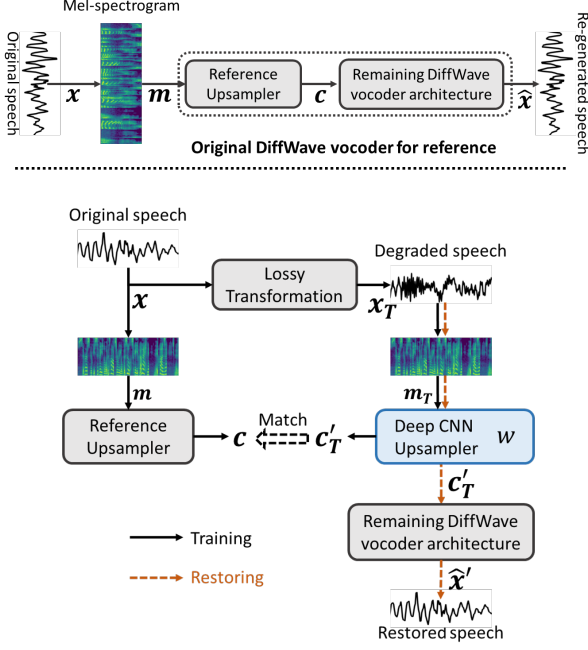
Figure 1: *Top: Supervised training for the original DiffWave model. Bottom: Our method for training a deep CNN upsampler to match the conditioner of DiffWave's reference upsampler. Then the remaining DiffWave vocoder architecture is used for restored speech waveform generation.*

in superior quality in speech restoration.

2. Methods

In this section, we describe our network architecture and training approach. We first describe how the original DiffWave model can be trained to restore degraded speech - this serves as our baseline model. We then describe our modifications to the DiffWave vocoder using a deep CNN inversion network to further enhance performance.

2.1. DiffWave for restoring degraded speech

DiffWave is a speech waveform generative model, i.e. a vocoder, based on diffusion models [12]. It takes the mel-spectrum as conditioning input and generates corresponding speech as shown in the top part of Figure 1. DiffWave was not originally designed for SE, but we use it for restoring lossy transformed speech in this paper. We train the DiffWave vocoder by using paired original speech x and degraded speech mel-spectrum m_T samples (in original paper, the clean mel-spectrum was used). Once the model converges, we use it to generate the estimated original speech $\hat{x}'$ conditioning on corresponding degraded speech mel-spectrum m_T. In the experimental results section, we show that a supervised DiffWave can restore the quality to a certain extent. After analyzing the structure of DiffWave, we identified the upsampler as a key component that can be further optimized to improve quality.

2.2. Deep CNN for Conditioner Upsampling

The DiffWave model contains three modules: an upsampler network, a diffusion embedding network, and residual learn-
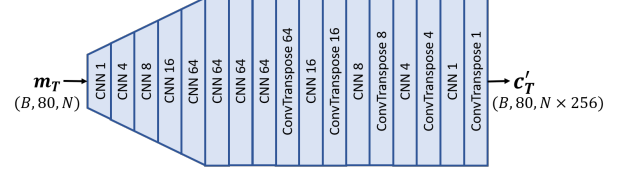

Figure 2: *The network structure of our deep CNN upsampler.*

ing blocks (for network details, please refer to the original paper [12]). In Diffwave, the upsampler network is used to increase the dimension of the input mel-spectrum to be the conditioner for speech waveform synthesis. The structure of the upsampler in the original DiffWave model is simple, it contains two 2D convolutional transposed layers.

We propose a new upsampler network, i.e. a deep CNN upsampler, to replace the original one. The degraded speech mel-spectrum m_T passes through several CNN nets with increasing channel size. The increased capacity of the upsampler allows us to invert the non-linear transformation and impute the lost information. This is then fed through cross-stacked CNN layers and transpose layers to decrease the channel size while increasing the mel-spectrum dimension to match the output speech waveform's dimension.

In our experiments, we found that simply replacing DiffWave's upsampler with our own network did not result in improved performance. The training of a diffusion-model with the CNN upsampler led to poor convergence to a local minima similar to training the original DiffWave.

To overcome this, we found it was better to separately train the CNN upsampler, independent of DiffWave, but with the criterion to match DiffWave's upsampling network's output on the original speech. In this scheme, we first train the DiffWave vocoder model which maps $x \rightarrow \hat{x}$, i.e. we train the model to generate an estimated original speech waveform conditioned on the original speech mel-spectrum. As shown in bottom part of Figure 1, DiffWave's upsampler is then extracted as the reference upsampler for the deep CNN upsampler training. The remaining DiffWave vocoder architecture is used for restored speech waveform synthesis. To train the deep CNN upsampler, we first generate the reference conditioner c from original speech mel-spectrum m via a reference upsampler, and generate the altered conditioner c'_T from the corresponding degraded speech mel-spectrum m_T with our new upsampler. We train the new upsampler with a mean absolute error loss (L1 loss),

$$\ell(c_n, c'_{T_n}; w) = \frac{1}{N} \sum_{n=1}^{N} |c_n - c'_{T_n}|, \quad (1)$$

where c'_{T_n} is given by deep CNN upsampler with weights w. After training the upsampler, at inference we simply feed the degraded speech mel-spectrum m_T through the new deep CNN upsampler to generate altered conditioner c'_T, and then through remaining DiffWave vocoder architecture to generated the estimated original speech $\hat{x}'$.

3. Experiments

3.1. Implementation Details

Network Architecture: Our network consists of a 15-layer CNN with a largest channel size of 64, as shown in Figure 2. The first 8 layers are 2-D CNNs: kernal size of (5,5) and stride

Table 1: *Quantitative measures of speech quality for in-corpus and cross-corpus evaluations. The comparisons are between the baseline model ('DW'), the modified DiffWave architecture ('ModDW'), and input degraded speech ('Degraded'). Each score is an average from a randomly-selected set of 128 samples, with standard deviation in parentheses. An asterisk means that the difference between ModDW and DW is statistically significant with $p < 0.05$ by doing t-test.*

	Tranformation	Model	PFP Loss	PESQ	CSIG	CBAK	COVL
In-corpus (TIMIT)	LPC-10 Compression	Degraded	0.0173(0.0010)	1.2029(0.1122)	1.9829(0.3419)	1.5589(0.1747)	1.4826(0.2501)
		DW	0.0140(0.0009)	1.2401(0.1216)	2.7146(0.3138)	1.7311(0.1857)	1.8833(0.2543)
		ModDW	**0.0121(0.0009)***	**1.5056(0.2287)***	**3.1048(0.2865)***	**1.8705(0.1781)***	**2.2390(0.2794)***
	AMR-NB Compression	Degraded	0.0150(0.0006)	2.2787(0.2937)	2.8363(0.4383)	2.3645(0.1581)	2.5355(0.3529)
		DW	0.0130(0.0008)	2.0022(0.2661)	3.1793(0.2508)	2.2444(0.1415)	2.5687(0.2506)
		ModDW	**0.0112(0.0006)***	**2.4498(0.3421)***	**3.5618(0.2812)***	**2.5127(0.1791)***	**3.0008(0.3070)***
	Signal Clip (25%)	Degraded	0.0116(0.0006)	1.5439(0.2155)	2.3717(0.2622)	1.8279(0.1699)	1.7797(0.2211)
		DW	0.0112(0.0004)	1.5022(0.1859)	2.5630(0.2084)	1.9280(0.1988)	1.8145(0.2632)
		ModDW	**0.0096(0.0003)***	**2.2144(0.2845)***	**2.6871(0.2544)***	**2.5410(0.1831)***	**2.2687(0.2988)***
Cross-corpus (Mozilla)	LPC-10 Compression	Degraded	0.0156(0.0019)	1.2134(0.1086)	2.0628(0.3548)	1.4790(0.1573)	1.5254(0.2281)
		DW	0.0154(0.0021)	1.2088(0.1173)	2.5583(0.3408)	1.5060(0.2046)	1.7547(0.2453)
		ModDW	**0.0132(0.0022)***	**1.3499(0.2165)***	**2.7654(0.3624)***	**1.6039(0.2365)***	**1.9536(0.2959)***
	AMR-NB Compression	Degraded	0.0145(0.0015)	1.7621(0.2779)	2.5079(0.4845)	2.0105(0.1617)	2.0700(0.3487)
		DW	0.0144(0.0013)	1.6875(0.2447)	2.5943(0.4694)	1.8846(0.1780)	2.0557(0.3276)
		ModDW	**0.0129(0.0011)***	**1.8793(0.3267)***	**2.8109(0.4629)***	**2.0763(0.1921)***	**2.2853(0.3698)***
	Signal Clip (25%)	Degraded	0.0120(0.0007)	1.3540(0.1569)	2.9644(0.3566)	1.5659(0.0994)	2.1180(0.2467)
		DW	0.0122(0.0008)	1.2156(0.1355)	3.0144(0.3266)	1.6804(0.1698)	2.0098(0.2554)
		ModDW	**0.0115(0.0005)***	**2.0756(0.4285)***	**3.4742(0.4177)***	**2.2240(0.2213)***	**2.7385(0.4074)***

of (1,1) across the layers; channel size of 1, 4, 8, 16, 64, 64, 64, 64; each layer is stacked with a 2-D batch normalization and a leaky-relu whose negative slope is 0.4. The next 7 layers is a cross-stacked 2-D convolutional transpose net and 2-D CNN. For the 2-D convolutional transpose net, the kernel size is (3,8), stride size is (1,4), and the channel size is kept the same as the input. For the 2-D CNN, the settings are the same and the channel size is 64, 16, 8, 4, 1. Again, each layer is stacked with a 2-D batch normalization and a leaky-relu whose negative slope is 0.4. These settings ensure the generated conditioner from the deep CNN upsampler has the same dimensions as that generated by the reference upsampler. This network architecture seemed to provide a good balance on the trade-off between model performance and the size of model parameters set, and we performed ablation studies on the layer sizes and dimensions to arrive at this final architecture.

As discussed in Section 2.2, training happens in two stages. First, we train the DiffWave vocoder from a PyTorch implementation [13], i.e. train the model to generate the original speech waveform conditioning on the original speech's mel-spectrum. We used the TIMIT [14] training dataset, a widely used English speech dataset, for training. The DiffWave vocoder was trained for 1M steps (100 hours on 2 Titan Xp GPUs) with a learning rate of 0.0002. For the second stage of training, the deep CNN upsamper is trained to alter the upsampled conditioner from the degraded speech mel-spectrum to match that generated by the reference upsampler from the paired original speech mel-spectrum. We train the upsampler for approximately 50k steps (6 hours on 1 Titan Xp GPU) with a learning rate of 0.001. We use the Adam optimizer for all trainings [15].

Lossy Operations: In this paper, we conduct three experiments to evaluate our model: 1) Restoring speech compressed by the LPC-10 algorithm [16], 2) Restoring speech compressed by the AMR-NB algorithm (mode: MR515, bit rate = 5.15 kbit/s) [17], and 3) Restoring speech with clipped magnitude (25% of the highest-energy samples clipped).

Datasets: For all three experiments, we use the TIMIT training and testing dataset as our training and in-corpus evaluation dataset correspondingly. The speech in TIMIT is regarded as

original speech, and we use the three different algorithms to generate degraded speech files. We also conduct a cross-corpus evaluation for each of the three conditions. We use the Mozilla common voice English dataset for the cross-corpus evaluation [18]. This is a large corpus that contains more than 1,500 hours of short sentences read by English speakers with various accents, ages, and genders across the world. We randomly selected 128 speech samples and downsampled to 16 kHz. The cross-corpus evaluation did not involve additional training or fine-tune. Note that all experiments mentioned in the paper are based on 16 kHz speech.

Evaluation metrics: To evaluate the restored speech quality quantitatively, we choose metrics used widely in speech enhancement, namely PESQ [19], CSIG, CBAK and COVL [20], and the phone-fortified perceptual (PFP) loss described in [21]. We did not use any of these metrics during training. PESQ, CSIG, CBAK, and COVL have been shown to correlate with "quality", whereas the PFP loss is a proxy for "intelligibility" as it is based on a speech recognition model. For all metrics, the required reference signal is the original speech.

Baseline model: Our baseline model is the original DiffWave model trained for restoring degraded speech as described in Section 2.1. For all three experiments, the DiffWave model is trained with the original speech waveform and corresponding degraded speech mel-spectrum.

3.2. Objective evaluations

Table 1 shows objective measures for in-corpus and cross-corpus evaluations of the baseline model (labeled as 'DW'), our proposed modified DiffWave scheme (labeled as 'ModDW'), and the input degraded speech (labeled as 'Degraded'). Comparing the score of the three operations, they have varying effects on speech quality. The LPC-10 compressed speech results in the poorest quality speech; whereas the AMR-NB compressed speech has the highest score on conventional perceptual score but the lowest on PFP loss, which indicates the AMR-NB compressed speech is of higher quality but is less intelligible. The worse PFP scores are likely due to the fact that AMR-NB
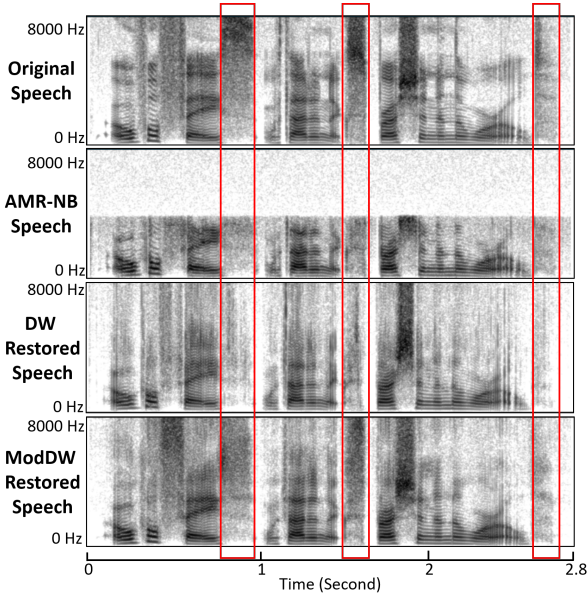
Figure 3: *A comparison of spectra between the original speech, degraded speech, baseline model, and modified DiffWave model. Samples are from the AMR-NB experiment on a TIMIT sample. The differences in high-frequency restoration are apparent in the highlighted regions.*

Figure 4: *Results of AB preference test for three experiments.* ∗ ∗ ∗ *indicates significant difference at* $p < 0.001$ *by doing t-test.*

downsamples the audio to 8 khz, removing all high-frequency content beyond 4 kHz.

Comparing the PFP loss for the baseline model and degraded speech, the baseline can restore degraded speech intelligibility under the in-corpus situation. However, for the conventional perceptual score (e.g. PESQ) it does not show significant improvement, in some cases the quality is poorer than the degraded speech (for AMR-NB, PESQ $2.00 < 2.28$). In cross-corpus evaluations, the baseline model failed to restore the degraded speech. The PFP loss for the baseline model is close or higher than the degraded speech. The results indicate that the baseline models fails to generalize outside the training set.

The modified DiffWave model proposed in this paper surpasses the baseline model significantly both for in-corpus and cross-corpus evaluation for all measures. All modified DiffWave model scores are higher than degraded speech, which means our model can restore the quality of different degraded speech sets at evaluation time. In our clipping experiment, the modified DiffWave model achieves a PFP score of 0.0098 in in-corpus evaluation, which nearly matches that of the original speech. Figure 3 shows an example spectra from the AMR-NB experiment for the in-corpus evaluation dataset. Our modified model can more accurately impute missing information in the high frequency band relative to the baseline model. It is important to note that the cross-corpus evaluation is especially difficult. This corpus contains sentences recorded by English speakers with various ages, genders, and accents/dialects. This provides strong evidence of generalizability.

3.3. Subjective evaluations

The perceptual measures used in this paper are imperfect proxies for human perception, as the restored speech's perceptual measures can be worse but listeners could still think the speech sounds better. We encourage the reader to listen to the speech
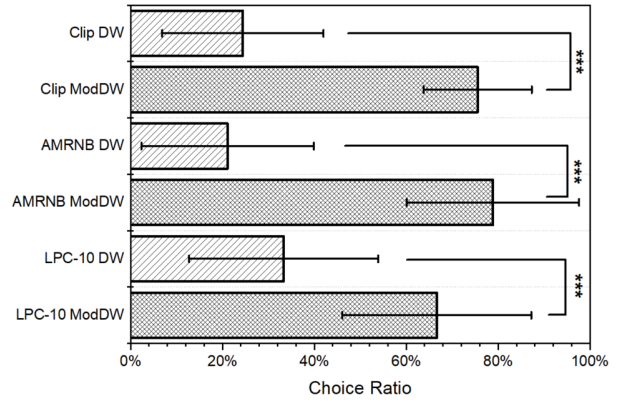
samples in the supplemental material to determine for themselves the quality of the reconstructed speech.

To compare our methods subjectively, we conducted AB preference test to compare the baseline model with our modified DiffWave model performance on restoring degraded speech. For each listening test, 15 pairs restored speech samples generated randomly from the TIMIT evaluation dataset by original DiffWave model and our modified model (5 pairs from the LPC-10 experiment, 5 pairs from the AMR-NB experiment, 5 pairs from the signal clipping experiment). The order of presentation of the DW-ModDW pair is randomized, and all 18 listeners evaluated a different selection of 15 sentences. We ensure the same spoken sentence is not used twice in any of the pairs. A total of 18 listeners participated in the study, and they were instructed to select the sample with better quality without knowledge of what method generated the sample. The AB preference results are shown in Figure 4. We can observe that our modified DiffWave model significant outperforms (with p-value < 0.001) the baseline model in all three experiments.

4. Conclusions

In this paper, we present a modified DiffWave model for superior quality restoration from distorted and lossy speech. We first train the DiffWave vocoder model to restore degraded speech in supervised fashion and produce good results. We also proposed a modified model that uses a deep CNN upsampler to replace original upsampler in DiffWave. Extensive in-corpus, cross-corpus and subjective perceptual evaluations show that the modified DiffWave model outperforms the original model in restoring degraded speech generated by lossy transformations.

Our modified model can revert the deterministic transformation. Future work will focus on extending this scheme to scenarios where the transformation is stochastic (e.g. noisy speech).

5. Acknowledgements

6. References

[1] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5751–5755.

[2] A. S. Subramanian, X. Wang, M. K. Baskar, S. Watanabe, T. Taniguchi, D. Tran, and Y. Fujita, "Speech enhancement using end-to-end speech recognition objectives," in 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). IEEE, 2019, pp. 234–238.

[3] T. Lan, Y. Lyu, W. Ye, G. Hui, Z. Xu, and Q. Liu, "Combining multi-perspective attention mechanism with convolutional networks for monaural speech enhancement," IEEE Access, vol. 8, pp. 78 979–78 991, 2020.

[4] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," arXiv preprint arXiv:2006.05694, 2020.

[5] P. C. Loizou, Speech enhancement: theory and practice. CRC Press, 2013.

[6] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori snr for speech enhancement using spectral subtraction rules," IEEE Signal Processing Letters, vol. 11, no. 4, pp. 450–453, 2004.

[7] M. Abd El-Fattah, M. I. Dessouky, S. M. Diab, and F. E.-S. Abd El-Samie, "Speech enhancement using an adaptive wiener filtering approach," Progress in Electromagnetics Research, vol. 4, pp. 167–184, 2008.

[8] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," arXiv preprint arXiv:1703.09452, 2017.

[9] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," arXiv preprint arXiv:1609.03499, 2016.

[10] W. Ping, K. Peng, K. Zhao, and Z. Song, "Waveflow: A compact flow-based model for raw audio," in International Conference on Machine Learning. PMLR, 2020, pp. 7706–7716.

[11] J.-M. Valin and J. 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.

[12] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," arXiv preprint arXiv:2009.09761, 2020.

[13] S. Nanavati, "Diffwave," https://github.com/lmnt-com/diffwave/, Sep 2020.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," NASA STI/Recon technical report n, vol. 93, p. 27403, 1993.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[16] T. E. Tremain, "The government standard linear predictive coding algorithm: Lpc-10," Speech Technology, pp. 40–49, 1982.

[17] 3GPP, "3gpp ts 26.090 - mandatory speech codec speech processing functions; adaptive multi-rate (amr) speech codec; transcoding functions," 3GPP, Retrieved 2010-07-21.

[18] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), vol. 2. IEEE, 2001, pp. 749–752.

[20] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 229–238, 2007.

[21] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss for speech enhancement," arXiv preprint arXiv:2010.15174, 2020.