SPEAKER INDEPENDENCE OF NEURAL VOCODERS AND THEIR EFFECT ON PARAMETRIC RESYNTHESIS SPEECH ENHANCEMENT

Soumi Maiti

The Graduate Center, CUNY
Computer Science
New York, NY, USA
smaiti@gradcenter.cuny.edu

Michael I Mandel

Brooklyn College
Computer and Information Science
Brooklyn, NY, USA
mim@sci.brooklyn.cuny.edu

ABSTRACT

Traditional speech enhancement systems produce speech with compromised quality. Here we propose to use the high quality speech generation capability of neural vocoders for better quality speech enhancement. We term this *parametric resynthesis* (PR). In previous work, we showed that PR systems generate high quality speech for a single speaker using two neural vocoders, WaveNet and WaveGlow. Both these vocoders are traditionally speaker dependent. Here we first show that when trained on data from enough speakers, these vocoders can generate speech from unseen speakers, both male and female, with similar quality as seen speakers in training. Next using these two vocoders and a new vocoder LPCNet, we evaluate the noise reduction quality of PR on unseen speakers and show that objective signal and overall quality is higher than the state-of-the-art speech enhancement systems Wave-U-Net, Wavenet-denoise, and SEGAN. Moreover, in subjective quality, multiple-speaker PR out-performs the oracle Wiener mask.

Index Terms— Speech enhancement, Neural vocoders, analysis-by-synthesis, enhancement-by-synthesis

1. INTRODUCTION

Traditional speech enhancement systems modify a noisy mixture to reduce the amount of noise it contains, but in doing so they introduce distortion in the speech. The distortion increases when there is more noise in the mixture leading to poor quality speech [1]. In contrast, speech synthesis systems generate high quality speech from only textual information. These text-to-speech systems (TTS) are complex as they need to generate realistic acoustic representation *without* a reference audio signal. In this work, we propose to combine these two methods, i.e., using speech synthesis techniques for speech enhancement. This is an easier task than TTS since we have a reference noisy audio signal from which we can extract the desired prosody instead of having to invent it. By predicting the "acoustic features" of the clean speech from the noisy speech in the speech enhancement system, we can generate high quality noise-free resyntheses.

Parametric Resynthesis (PR) systems [2, 3] predict clean acoustic parameters from noisy speech and synthesize speech from these predicted parameters using a speech synthesizer or vocoder. Current speech synthesizers are trained to generate high quality speech for a single speaker. In previous work we showed that a single speaker PR system can synthesize very high quality clean speech at 22 KHz [2] and performs better than the corresponding TTS system [3]. Hence, a critical question is whether these systems can be generalized to unknown speakers. The main contribution of the current work is to show that when trained on a large number of speakers, neural vocoders can successfully generalize to unseen speakers. Furthermore, we show that PR systems using these neural vocoders can also generalize to unseen speakers in the presence of noise.

In this work, we test the speaker dependence of neural vocoders, and their effect on the enhancement quality of PR. We show that when trained on 56 speakers, WaveGlow [4], WaveNet [5], and LPCNet [6] are able to generalize to unseen speakers. We compare the noise reduction quality of PR with three state-of-the-art speech enhancement models and show that PR-LPCNet outperforms every other system including an oracle Wiener mask-based system. In terms of objective metrics, the proposed PR-WaveGlow performs better in objective signal and overall quality.

1.1. Related work

Traditional speech enhancement systems generally predict a time-frequency mask to reduce noise in the magnitude spectrum domain, for example [7, 8]. Recent works perform speech enhancement in the time-domain directly, which has the additional advantage of reconstructing the phase of the signal. A modified WaveNet was proposed for speech denoising [9], by using non-causal convolutions on noisy speech and predicting both clean speech and the noise signal. Another approach is to progressively downsample the noisy audio to a bottleneck feature and then upsample with skip connections to the corresponding downsampled features to enhance speech. SEGAN [10] uses this approach in a GAN setting and Wave-U-Net [11, 12] uses it in the U-Net setting. The aim of these approaches is to remove noise from the audio at different scales. Compared to these systems, we do not focus on modelling noise but only focus on modelling speech. We evaluate our approach against three of these systems [9–11]. These papers publish results on the same dataset we used and also each provide several enhanced files, which we utilize in our listening tests.

2. SYSTEM OVERVIEW

Our PR models have two parts. First is a prediction model that estimates the clean acoustic features from noisy audio. Second, a vocoder synthesizes "clean" speech from the predicted "clean" acoustic parameters. The aim of the prediction model is to reduce noise while the vocoder synthesizes high quality audio.

2.1. Prediction model

The prediction model is trained with parallel clean and noisy speech. It takes noisy mel-spectrogram Y as input and is trained to predict

clean acoustic features X. The predicted clean acoustic features vary based on the vocoder used. In this work we used WaveGlow, WaveNet, LPCNet and WORLD [13] as vocoders. For WaveGlow and WaveNet, we predict clean mel-spectrograms. For LPCNet, we predict 18-dimensional Bark-scale frequency cepstral coefficients (BFCC) and two pitch parameters: period and correlation. For WORLD we predict the spectral envelope, aperiodicity, and pitch. For WORLD and LPCNet, we also predict the Δ and $\Delta\Delta$ of these acoustic features for smoother outputs. The prediction model is trained to minimize the mean squared error (MSE) of the acoustic features

$$\text{MSE}: \quad \mathcal{L} = \|X - \hat{X}\|^2 \tag{1}$$

where $\hat{X}$ are the predicted and X are the clean acoustic features. The Adam optimizer [14] is used for training. During test, for a given a noisy mel-spectrogram, clean acoustic parameters are predicted. For LPCNet and WORLD we use maximum likelihood parameter generation (MLPG) [15] algorithms to refine our estimate of the clean acoustic features from predicted acoustic features, Δ, and $\Delta\Delta$.

2.2. Vocoders

The second part of PR resynthesizes speech from the predicted acoustic parameters $\hat{X}$ using a vocoder. The vocoders are trained on clean speech samples x and clean acoustic features X. During synthesis, we use predicted acoustic parameters $\hat{X}$ to generate predicted clean speech $\hat{x}$. In the rest of this section we describe the vocoders, three neural: WaveGlow, WaveNet, LPCNet and one non-neural: WORLD.

WaveGlow: WaveGlow [4] is a Glow based network [16] for synthesizing speech. WaveGlow learns a sequence of invertible transformations of audio samples x to a Gaussian distribution conditioned on the mel spectrogram X. For inference, WaveGlow samples a latent variable z from the learned Gaussian distribution and applies the inverse transformations conditioned on X to reconstruct the speech sample $\tilde{x}$. The log likelihood of clean speech is maximized as,

$$\ln p(x \mid X) = \ln p(z) + \log\det\left|\frac{dz}{dx}\right|, \tag{2}$$

where $\ln p(z)$ is the log-likelihood of the spherical zero mean Gaussian with variance σ^2. During training $\sigma = 1$ is used. We use the officially published WaveGlow implementation[1] with the original setup, i.e., 12 coupling layers, each consisting of 8 layers of dilated convolution with 512 residual and 256 skip connections. We will refer to the PR system with WaveGlow as its vocoder as PR-WaveGlow.

LPCNet: LPCNet is a variation of WaveRNN [17] that simplifies the vocal tract response using linear prediction p_t from previous time-step samples

$$p_t = \sum_{k=1}^{M} a_k x_{t-k}.$$

LPC coefficients a_k are computed from the 18-band BFCC. It predicts the LPC predictor residual e_t, at time t. Then sample x_t is generated by adding e_t and p_t.

A frame conditioning feature f is generated from 20 input features: 18-band BFCC and 2 pitch parameters via two convolutional and two fully connected layers. The probability $p(e_t)$ is predicted from x_{t-1}, e_{t-1}, p_t, f via two GRUs [18] (A and B) combined with dualFC layer followed by a softmax. The largest GRU (GRU-A) weight matrix is forced to be sparse for faster synthesis. The model is trained on the categorical cross-entropy loss of $p(e_t)$ and the predicted probability of the excitation $\hat{p}(e_t)$. Speech samples are 8-bit mu-law quantized. We use the officially published LPCNet implementation[2] with 640 units in GRU-A and 16 units in GRU-B. We refer to the PR system with LPCNet as its vocoder as PR-LPCNet.

[1] https://github.com/NVIDIA/waveglow

Model	#spk	CSIG	CBAK	COVL	STOI
Seen					
WaveGlow	1	4.7±0.03	3.0±0.02	4.0±0.04	0.95±0.01
LPCNet	1	3.8±0.06	2.2±0.04	2.9±0.07	0.91±0.01
WaveNet	1	3.3±0.05	2.1±0.02	2.5±0.04	0.81±0.01
Unseen - Male					
WaveGlow	3	4.5±0.07	2.8±0.06	3.8±0.10	0.95±0.01
LPCNet	3	4.0±0.10	2.3±0.08	3.1±0.12	0.93±0.01
WaveNet	3	3.2±0.02	2.1±0.02	2.5±0.03	0.83±0.01
Unseen - Female					
WaveGlow	3	4.6±0.08	2.8±0.06	3.9±0.05	0.95±0.01
LPCNet	3	4.0±0.08	2.4±0.07	3.1±0.10	0.90±0.04
WaveNet	3	3.3±0.03	2.0±0.04	2.5±0.03	0.80±0.01

Table 1. Speaker generalization of neural vocoders. Objective quality metrics for synthesis from true acoustic features, higher is better. Sorted by SIG. We report 95% confidence intervals.

WaveNet: WaveNet [5] is a autoregressive speech waveform generation model built with dilated causal convolutional layers. The generation of one speech sample at time step t, x_t, is conditioned on all previous time step samples $(x_1, x_2, \ldots x_{t-1})$. We use the Nvidia implementation[3] which is the Deep-Voice [19] model of WaveNet for faster synthesis. Speech samples are mu-law qauantized to 8 bits. The normalized log mel-spectrogram is used in local conditioning. WaveNet is trained on the cross-entropy between the quantized sample x_t^μ and the predicted quantized sample $\hat{x}_t^\mu$.

For WaveNet, we used a smaller model that is able to synthesize speech with moderate quality. We tested the PR model's dependency on speech synthesis quality by testing on a smaller model. We used 20 layers with 64 residual, 128 skip connections, and 256 gate channels with maximum dilation of 128. This model can synthesize clean speech with average predicted mean opinion score (MOS) 3.25 for a single speaker [19]. The PR system with WaveNet as its vocoder is referred to as PR-WaveNet.

WORLD: Lastly, we use a non-neural vocoder WORLD which synthesizes speech from three acoustic parameters: spectral envelope, aperiodicity, and $F0$. We use WORLD with the Merlin toolkit [4]. WORLD is a source-filter model that takes previously mentioned parameters and synthesizes speech. We also use spectral enhancement to modify the predicted parameters as is standard in Merlin [20].

3. EXPERIMENTS

3.1. Dataset

We use the publicly available noisy VCTK dataset [21] for our experiments. The dataset contains 56 speakers for training: 28 male and 28 female speakers from the US and Scotland. The test set contains two unseen voices, one male and another female. Further, there is another available training set, consisting 14 male and 14 female from England, which we use to test generalization to more speakers.

[2] https://github.com/mozilla/LPCNet

[3] https://github.com/NVIDIA/nv-wavenet

[4] https://github.com/CSTR-Edinburgh/merlin

Model	CSIG	CBAK	COVL	STOI
Oracle Wiener	4.3±0.04	3.8±0.19	3.8±0.22	0.98±0.01
PR-WaveGlow	3.8±0.03	2.4±0.08	3.1±0.15	0.91±0.02
PR-LPCNet, noisy $F0$	3.5±0.02	2.1±0.07	2.7±0.12	0.88±0.03
PR-LPCNet	3.1±0.02	1.8±0.05	2.2±0.08	0.88±0.03
PR-World	3.0±0.02	1.9±0.06	2.2±0.10	0.88±0.02
PR-WaveNet	2.9±0.10	2.0±0.04	2.2±0.11	0.83±0.01
Wave-U-Net (from [11])	3.5	3.2	3.0	-
SEGAN (from [10])	3.5	2.9	2.8	-

Table 2. Speech enhancement objective metrics on full 824-file test set: higher is better. Top system uses oracle clean speech information. Bottom section compares to published comparison system results.

The noisy training set contains ten types of noise: two are artificially created, and the eight other are chosen from DEMAND [22]. The two artificially created are speech shaped noise and babble noise. The eight from DEMAND are noise from a kitchen, meeting room, car, metro, subway car, cafeteria, restaurant, and subway station. The noisy training files are available at four SNR levels: 15, 10, 5, and 0 dB. The noisy test set contains five other noises from DEMAND: living room, office, public square, open cafeteria, and bus. The test files have higher SNR: 17.5, 12.5, 7.5, and 2.5 dB. All files are downsampled to 16 KHz for comparison with other systems. There are $23,075$ training audio files and 824 testing audio files.

3.2. Exp 1: Speaker independence of neural vocoders

Firstly, we test if WaveGlow and WaveNet can generalize to unseen speakers on clean speech. Using the data described above, we train both of these models with a large number of speakers (56) and test them on 6 unseen speakers. Next, we compare their performance to LPCNet which has previously been shown to generalize to unseen speakers. In this test, each neural vocoder synthesizes speech from the original clean acoustic parameters. Following the three baseline papers [9–11], we measure synthesis quality with objective enhancement quality metrics [23] consisting of three composite scores: CSIG, CBAK, and COVL. These three measures are on a scale from 1 to 5, with higher being better. CSIG provides and estimate of the signal quality, BAK provides an estimate of the background noise reduction, and OVL provides an estimate of the overall quality.

LPCNet is trained for 120 epochs with a batch size of 48, where each sequence has 15 frames. WaveGlow is trained for 500 epochs with batch size 4 utterances. WaveNet is trained for 200 epochs with batch size 4 utterances. For WaveNet and WaveGlow we use GPU synthesis, while for LPCNet CPU synthesis is used as it is faster[5]. WaveGlow and WaveNet synthesize from clean mel-spectrograms with window length 64 ms and hop size 16 ms. LPCNet acoustic features use a window size of 20 ms and a hop size of 10 ms.

We report the synthesis quality of three unseen male and three unseen female speakers, and compare them with unseen utterances from one known male speaker. For each speaker, the average quality is calculated over 10 files. Table 1 shows the composite quality results along with the objective intelligibility score from STOI [24]. We observe that WaveGlow has the best quality scores in all the measures. The female speaker scores are close to the known speaker while the unseen male speaker scores are a little lower. We note here that these values are not as high as single speaker WaveGlow, which can synthesize speech very close to the ground truth. We also note that LPCNet scores are lower than those of WaveGlow but better than WaveNet. Between LPCNet and WaveNet, we do not observe a significant difference in synthesis quality for male and female voices. Although WaveNet has lower scores, it is consistent across known and unknown speakers. Thus, we can say that WaveNet generalizes to unseen speakers.

[5]We also found that GPU synthesis code was incomplete as of commit `3a7ef33`

3.3. Exp 2: Speaker independence of parametric resynthesis

Next, we test the generalizability of the PR system across different SNRs and unseen voices. We use the test set of 824 files with 4 different SNRs. The prediction model is a 3-layer bi-directional LSTM with 800 units that is trained with a learning rate of 0.001. For WORLD filter size is 1024 and hop length is 5 ms. We compare PR models with a mask based oracle, the Oracle Wiener Mask (OWM), that has clean information available during test.

Table 2 reports the objective enhancement quality metrics and STOI. We observe that the OWM performs best, PR-WaveGlow performs better than Wave-U-Net and SEGAN on CSIG and COVL. PR-WaveGlow's CBAK score is lower, which is expected since this score is not very high even when we synthesize clean speech (as shown in Table 1). Among PR models, PR-WaveGlow scores best and PR-WaveNet performs worst in CSIG. The average synthesis quality of the WaveNet model affects the performance of the PR system poorly. PR-WORLD and PR-LPCNet scores are lower as well, we observe that both of these models sound much better than the objective scores would suggest. We believe, as both of these models predicts $F0$, even a slight error in $F0$ prediction affects the objective scores adversely. For this, we test the PR-LPCNet using the noisy $F0$ instead of the prediction, and the quality scores increase. In informal listening the subjective quality with noisy F0 is similar to or worse than the predicted $F0$ files. Hence we can say that the objective enhancement metrics are not a very good measure of quality for PR-LPCNet and PR-WORLD.

We also test objective quality of PR models and OWM against different SNR and noise types. The results are shown in Figure 1. We observe with decreasing SNR, CBAK quality for PR models stays the same, while for OWM, CBAK score decreases rapidly. This shows that the noise has a smaller effect on background quality compared to a mask based system, i.e., the background quality is more related to the presence of synthesis artifacts than recorded background noise.

3.4. Listening tests

Next, we test the subjective quality of the PR systems with a listening test. For the listening test, we choose 12 of the 824 test files, with four files from each of the 2.5, 7.5 and 12.5 dB SNRs. We observed the 17.5 dB file to have very little noise, and all systems perform well with them. In the listening test, we also compare with the OWM and three comparison models. For these comparison systems, we included the publicly available output files in our listening tests, selecting five files from each: Wave-U-Net has 3 from 12.5 dB and 2 from 2.5 dB, Wavenet-denoise and SEGAN have 2 common files from 2.5 dB, 2 more files each are selected from 7.5 dB and 1 from 12.5 dB. For Wave-U-Net, there were no 7.5 dB files available publicly.

The listening test follows the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) paradigm [25]. Subjects were presented with 8-10 anonymized and randomized versions of each file to facilitate direct comparison: 4 PR systems (PR-WaveNet, PR-WaveGlow, PR-LPCNet, PR-World), 4 comparison speech enhancement systems (OWM, Wave-U-Net, WaveNet-denoise, and SEGAN), and clean and noisy signals. Subjects were also provided reference clean and noisy

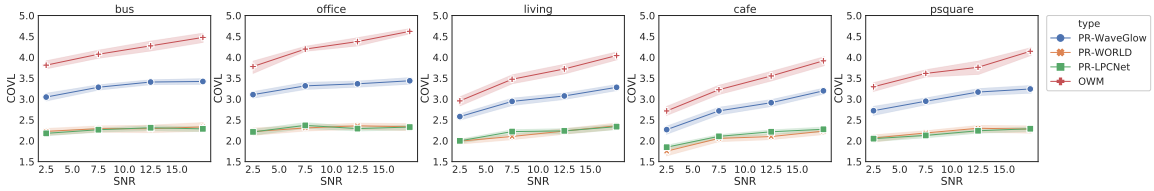

Fig. 1. Overall objective quality of PR systems and OWM broken down by noise type (824 test files).

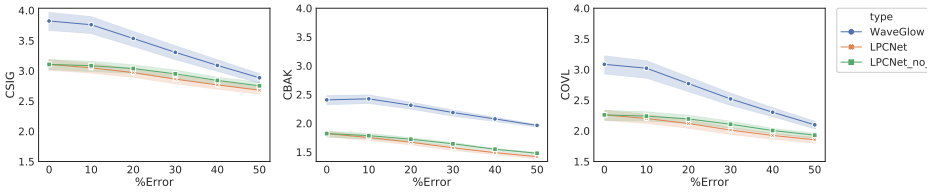

Fig. 2. Objective metrics as error is artificially added to the predictions of the acoustic features, higher is better. Error is measured as a proprotion of the standard deviation of the vocoders' acoustic features over time.

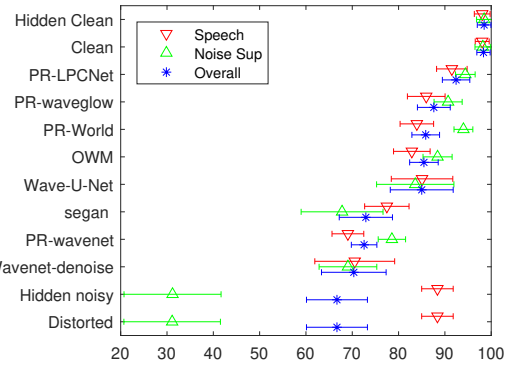

Fig. 3. Subjective quality: higher is better. Error bars show 95% confidence intervals.

Model	CSIG	CBAK	COVL	STOI	Subj. Intel.
Oracle Wiener	4.3±0.30	3.8±0.30	3.9±0.32	0.98±0.02	0.91±0.02
PR-WaveGlow	3.8±0.20	2.4±0.11	3.0±0.19	0.91±0.03	0.90±0.03
PR-World	3.1±0.14	1.9±0.10	2.2±0.15	0.88±0.02	0.90±0.04
PR-LPCNet	3.0±0.07	1.8±0.05	2.2±0.05	0.85±0.06	0.92±0.03
PR-WaveNet	2.9±0.09	2.0±0.6	2.2±0.10	0.83±0.03	0.74±0.05

Table 3. Speech enhancement objective metrics and subjective intelligibility on the 12 listening test files.

versions of each file[6]. Five subjects took part in the listening test. They were told to rate the speech quality, noise-suppression quality, and overall quality of the speech from $0-100$, with 100 being the best. We observe intelligibility of all of the files to be very high, so instead of doing an intelligibility listening test, we ask subjects to rate the subjective intelligibility as a score from $0-100$.

Figure 3 shows the result of the quality listening test. PR-LPCNet performs best in all three quality scores, followed by PR-WaveGlow and PR-World. The next best model is the Oracle Wiener mask followed by Wave-U-Net. Table 3 shows the subjective intelligibility ratings, where PR-LPCNet has the highest subjective intelligibility, followed by OWM, PR-WaveGlow, and PR-World. It also reports the objective quality metrics on the 12 files selected for the listening test for comparison with Table 2 on the full test set. We observe that while PR-LPCNet and PR-WORLD have very similar objective metrics (both quality and intelligibility), they have very different subjective metrics, with PR-LPCNet being rated much higher).

3.5. Tolerance to error

Finally, we measure the tolerance of PR models to inaccuracy of the prediction LSTM using the two best performing vocoders, WaveGlow and LPCNet. For this test, we randomly select 30 noisy test files. We make the predicted feature $\hat{X}$ noisy as, $\hat{X}_e = \hat{X} + \epsilon N$, where $\epsilon = MSE \times e\%$. The random noise N is generated from a Gaussian distribution with the same mean and variance at each frequency as X. Next, we synthesize with the vocoder from $\hat{X}_e$. For WaveGlow, X is the mel-spectrogram and for LPCNet, X is 20 features. We repeat the LPCNet test adding noise into all features and only the 18 BFCC features (not adding noise to $F0$).

Figure 2 shows the objective metrics for these files. We observe that for WaveGlow, $e = 0-10\%$ does not affect the synthesis quality very much and $e > 10\%$ decreases performance incrementally. For LPCNet, we observe that errors in the BFCC are tolerated better than errors in $F0$.

4. CONCLUSION

We show that the neural vocoders WaveGlow, WaveNet, and LPCNet can be used for speaker-independent speech synthesis when trained on 56 speakers. We also show that using these three vocoders, the parametric resynthesis model is able to generalize to new noises and new speakers across different SNRs. We find that PR-LPCNet outperforms the oracle Wiener mask-based system in subjective quality.

Acknowledgements: This material is based upon work supported by the National Science Foundation (NSF) via grant IIS-1618061. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

[6]All files are available at `http://mr-pc.org/work/icassp20/`

5. REFERENCES

[1] Jingdong Chen, Jacob Benesty, Yiteng Huang, and Simon Doclo, "New insights into the noise reduction wiener filter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1218–1234, 2006.

[2] Soumi Maiti and Michael I Mandel, "Parametric resynthesis with neural vocoders," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, To appear.

[3] Soumi Maiti and Michael I Mandel, "Speech denoising by parametric resynthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2019, pp. 6995–6999.

[4] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *arXiv preprint arXiv:1811.00002*, 2018.

[5] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio.," in *Proc. ISCA SSW*, Sept. 2016, p. 125.

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

[7] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[8] Hakan Erdogan, John R. Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2015, vol. 2015-Augus.

[9] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018, pp. 5069–5073.

[10] Santiago Pascual, Antonio Bonafonte, and Joan Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[11] Craig Macartney and Tillman Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[12] Daniel Stoller, Sebastian Ewert, and Simon Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[13] 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, Jul. 2016.

[14] Diederik P. Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014.

[15] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2000, vol. 3, pp. 1315–1318.

[16] Diederik P Kingma and Prafulla Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *arXiv preprint arXiv:1807.03039*, 2018.

[17] 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.

[18] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[19] Sercan Ö Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al., "Deep voice: Real-time neural text-to-speech," in *Proceedings of the International Conference on Machine Learning*. JMLR. org, 2017, pp. 195–204.

[20] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW*, 2016.

[21] Cassia Valentini-Botinhao et al., "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[22] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics ICA2013*. ASA, 2013, vol. 19, p. 035081.

[23] Yi Hu and Philipos C Loizou, "Evaluation of objective measures for speech enhancement," in *Proceedings of Interspeech*, 2006.

[24] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 4214–4217.

[25] "Method for the subjective assessment of intermediate quality level of audio systems," Tech. Rep. BS.1534-3, International Telecommunication Union Radiocommunication Standardization Sector (ITU-R), 2015.