A Speech Synthesis Approach for High Quality Speech Separation and Generation

Qingju Liu, Philip JB Jackson, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—We propose a new method for source separation by synthesizing the source from a speech mixture corrupted by various environmental noise. Unlike traditional source separation methods which estimate the source from the mixture as a replica of the original source (e.g. by solving an inverse problem), our proposed method is a synthesis-based approach which aims to generate a new signal (i.e. “fake” source) that sounds similar to the original source. The proposed system has an encoder-decoder topology, where the encoder predicts intermediate-level features from the mixture, i.e. Mel-spectrum of the target source, using a hybrid recurrent and hourglass network, while the decoder is a state-of-the-art WaveNet speech synthesis network conditioned on the Mel-spectrum, which directly generates time-domain samples of the sources. Both objective and subjective evaluations were performed on the synthesized sources, and show great advantages of our proposed method for high-quality speech source separation and generation.**

***Index Terms*—Deep learning, speech separation, speech synthesis, WaveNet, hourglass, high quality.**

I. Introduction

Deep learning has been prevailing the source separation and enhancement field in recent years, where different deep neural networks (DNN) have been considered including the classic multi-layer perception [1], recurrent neural networks (RNN) [2], [3], convolutional neural networks (CNN) [4], and more recently, the dilated convolutions [5], [6] and generative adversarial networks [7]. Most existing methods reconstruct the sources from low-level features, such as time frequency (TF) spectrum [1]–[4], [8], or time samples [5]–[7]. However, the high-dimensional representative features involved in the above methods are prone to the over-fitting problem, thus the separated sounds often suffer from artefacts such as musical noise and interference from competing sounds, especially in adverse acoustic scenarios. As a result, the source estimates might sound machine-like and unnatural, which greatly affects the listening experience and limits their applications in scenarios requiring high quality speech, such as broadcasting.

On the other hand, speech synthesis has also witnessed the transition from statistical parametric methods to DNNs. For instance, WaveNet [9] and its modified versions as in FFTNet [10], Tacotron 2 [11], Deep Voice [12], have been used to generate sounds in the time-domain directly. Unlike speech separation, these synthesis methods generate (i.e. “fake”) highly natural signals from high-level features such as linguistic features, pitch and duration models [9], [10], [12], or from relatively intermediate-level features e.g. Mel-scale spectrum [10], [11]. It is, however, challenging to synthesise the same speech under specific scenarios without information loss/modifications, when limited or no audio input is given.

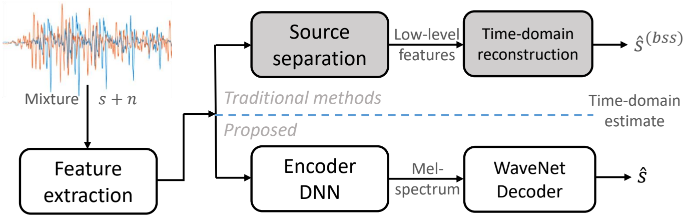

Fig. 1: Diagram of our proposed system (bottom), in contrast to traditional blind source separation (BSS) methods (top). Traditional methods aim to recover the source estimate such that it is a replica of the original target $\hat{s}^{(\mathrm{bss})} = s$. The proposed method reserves perceptually-important acoustic information via the encoder, and “fakes” a new source $\hat{s}$ via the decoder that sounds similar to the target source s while maintaining high quality.

As an alternative to conventional speech separation methods, we propose a new method which uses an encoder to extract the Mel-spectrum of the target source from noise-corrupted mixtures and then a decoder (i.e. the modified WaveNet [11]) to synthesise the source from the Mel-spectrum in the time domain. As compared to traditional BSS methods that directly extract high-dimensional audio features such as linear-spectrum, the proposed encoder relaxes the learning burden to a relatively low-dimensional space with a high accuracy, and the proposed decoder synthesizes a signal that is perceptually-similar to the original source. Unlike existing WaveNet-based models for speech denoising [5], [6], our framework is composed of two networks boosting multi-task learning, offering great flexibility, e.g., synthesis using a new person’s voice. In addition, the proposed model is half-discriminative (encoder) and half-generative (decoder), where the encoder avoids the accumulated error from previous predictions occurred in the generative model [5], and the decoder takes temporal information into account that is beyond the fixed-length receptive field as in the discriminative model [6]. Compared to the GAN-based speech enhancement network [7], the encoder and decoder are independently trained, which is much more straightforward than the iterative training process involving a generator and a discriminator. Also, WaveNet, as a neural network method, is more flexible than the statistic parametric vocoder used in the parametric resynthesis system [13].

The authors are with the Centre for Vision, Speech and Signal Processing, University of Surrey, GU2 7XH, UK. W. Wang is also with Qingdao University of Science and Technology, China. e-mails: {q.liu, p.jackson, w.wang}@surrey.ac.uk. The authors would like to acknowledge the support of the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

Most importantly, we need to stress that, unlike traditional source separation methods as shown in Fig. 1 (top), our proposed method does not aim to reconstruct the original source signal in terms of time-sample or TF-unit accuracy. Instead, our scheme generates a new signal (i.e. "faked" source) that sounds similar to the original source, with high naturalness and sound quality, offered by deep learning based speech synthesis. The diagram of the proposed system is shown in Fig. 1 (bottom).

II. The Proposed Method

Our proposed system utilises WaveNet conditioned on the Mel-spectrum for time-domain speech synthesis. Thus the accuracy of the Mel-spectrum is of critical importance to the quality of the generated signals, which relies on the encoder performance, whose topology is detailed as follows.

Two sets of features extracted from the mixtures are used as inputs to the proposed encoder network (shown in Fig. 2), i.e. the normalised linear-spectrum $X^{\text{lin}}(t, \omega^{\text{lin}})$ and Mel-spectrum $X(t, \omega^{\text{mel}})$, where (t, ω) is the TF index, either in linear or Mel scale. The linear-spectrum provides detailed information with a high resolution, while the Mel-spectrum provides information in perceptual scale, which is also consistent with the encoder output. These two sets of features are fed into independent RNN networks in parallel, which contain bidirectional long short-term memory (BiLSTM) layers, followed by fully-connected layers at each time instance, before being reshaped and concatenated together. The RNN architecture can exploit the strong temporal coherence yielded by the audio spectra. Afterwards, we employ an hourglass network, to explore information at different scales [14], that consists of stacked convolutional networks (SCN). To ease the training, we also enforced residual learning [15] on the hourglass network and the SCNs. Output from the hourglass network is concatenated with the input Mel-spectrum, followed by another SCN to produce the final DNN output—the estimated Mel-spectrum $\hat{X}(t, \omega)$.

We propose to adopt the perceptually weighted loss similar to [1] as the objective function in the encoder training process. Dropping the TF index, we denote $f(X) \in [0, 1]$ as the perceptual importance of a unit X, which was empirically set to boost high energy components. A composite weighted squared error is proposed that contains two weights:

$$\mathcal{L} = \sum_{t,\omega} \Big(f(X) + (1 - f(X)) f(\hat{X}) \Big) (\hat{X} - X)^2, \quad (1)$$

where the first weight $f(X)$ preserves the perceptually important information from the target, and the second weight $(1 - f(X)) f(\hat{X})$ suppresses distortions that may cause perceptual difference.

The encoder output, i.e. the target Mel-spectrum $\hat{X}(t, \omega)$, is then fed into the decoder network, to generate a synthesised speech signal in the time domain directly. We propose to use the modified WaveNet as used in Tacotron 2 [11], which is conditioned on Mel-spectrum, rather than linguistic features and fundamental frequency as used in the original WaveNet text-to-speech (TTS) system [16].

III. Experiments

A. Data and setup

In the encoder training process, the LJSpeech dataset [17], together with the environmental sound classification (ESC-50) database [18] were used.

The LJSpeech dataset contains 13100 sequences (12522-training, 578-testing) with varying length ranging from 1 to 10 seconds, in total 24 hours of speech by a female speaker, sampled at 22.05 kHz, which is used as the target speaker. We used the 800 natural soundscapes & water sounds and exterior/urban noises (%90-training, %10-testing), each lasting 5 seconds, to mimic background noise. A target and a noise signal are randomly chosen and added together with an additive mixing model without memory.

To extract the DNN input spectrum $X^{\text{lin}}(t, \omega)$, 1024 FFT size with 75%-overlapped Hanning windows was applied, resulting in 512 linear filter bins. Mel-spectrum $X(t, \omega)$ was extracted from $X^{\text{lin}}(t, \omega)$ in the range of 125 Hz to 7600-Hz with 80 bins. Normalisations are applied via thresholds, mean-shifts, and scales thus all features are mapped to the range of [0,1]. Audio features spanning 64 frames were fed into the network.

The two BiLSTM layers have a size of 800 and 400 for the linear- and Mel-spectra respectively, with dropout (feedforward and recurrent) of 0.25, followed by the fully connected layers at each time instance with a layer size 320. The two stream outputs are reshaped and concatenated with the input $X(t, \omega)$, to obtain the hidden output $H \in \mathbb{R}^{64 \times 80 \times 9}$, which is then fed into the hourglass network. Three scales with pooling size of (2,2) are applied in the hourglass network, where each SCN unit has filter length of 64, resulting in a bottleneck representation $\in \mathbb{R}^{8 \times 10 \times 64}$ at the largest scale level. Each SCN has three stacked composite layers containing batch normalisation (BN), ReLU and convolutional layer as shown in the embedded dashed plot in Fig. 2. Input of each SCN is added to the last layer output directly, or goes through another covolutional layer before being added if its channel number is not consistent with the SCN output. Symmetric layers in the hourglass network are connected via skip residual layers. The hourglass output is followed by another SCN and convolutional layers with filter number of 1, to obtain the final Mel-scale spectrum.

In the training process, the Adam optimiser [19] was used in the backpropagation, with initial learning rate of 0.001 and decay of 0.98 after each epoch. In total, 500 epochs were enforced with each epoch using data lasting about 1.5 hours. In our loss optimisation, we use a perceptual importance function $f(X) = X^2$ to boost high-energy components. The decoder part is the same as in [11], and a pre-trained model learned from the LJSpeech dataset is used here. The parameters used to obtain the pre-trained decoder network are consistent with that in our encoder training process.

B. Baseline methods

We implemented two baseline methods. The first one (denoted as B1) directly estimates the linear-scale spectrum of the target from the mixture, which is a modification to the encoder network as follows. The two fully-connected layers

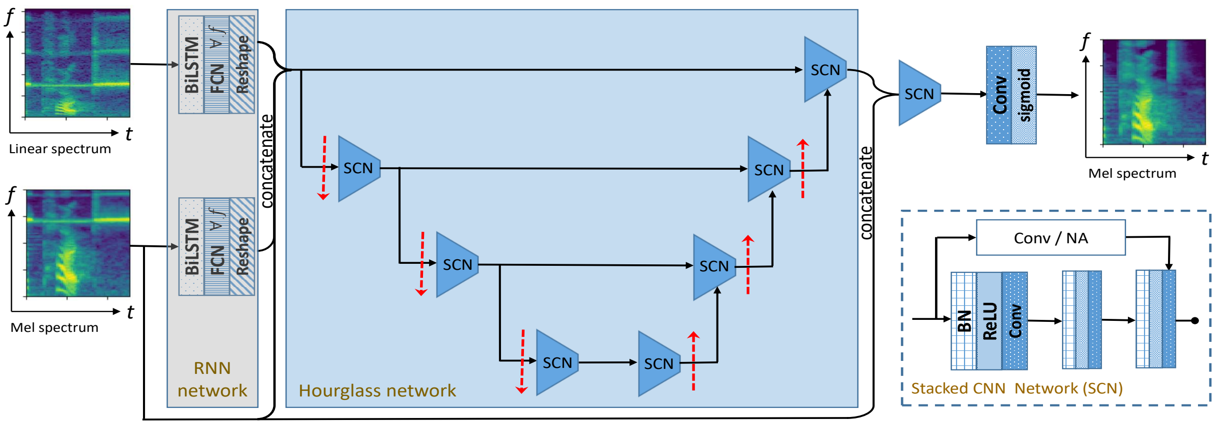

Fig. 2: Diagram of our proposed encoder DNN, which contains mainly two parts: the recurrent neural network (RNN) to exploit the temporal information, and the followed hourglass network that learns features at different scales. The RNN network has bidirectional LSTM layers followed by fully connected layers. The hourglass network contains mainly stacked convolutional networks (SCN). Details of the SCN structure are highlighted in the right bottom corner. The dashed down and up arrows denote max-poolings and up-samplings.

have size 1024 and 512 respectively, to generate hidden outputs spanning the same frequency length in the linear scale. Afterwards, the linear-spectrum X^{lin} is concatenated with the hidden outputs, instead of the Mel-spectrum X. This ensures that the same input features as for the proposed encoder are used in this baseline, to allow fair comparison, mitigating effects caused by factors such as DNN structures and implementations. Finally, an inverse STFT was applied to the estimated linear-spectrum to recover time-domain signals, using phase information estimated via the Griffin-Lim algorithm [20], thus a synthesis decoder network is not required.

The second baseline method[1] is a state-of-the-art WaveNet-based speech enhancement method, denoted as B2 [6]. This method directly outputs a batch of enhanced time samples from noisy mixtures.

C. Results and analysis

1) Encoder evaluation: We first quantitatively evaluate the introduced error in the encoder network of our proposed method and B1 on the testing data. Denoting $\bar{Y}$ as the normalised audio features ignoring the TF index, either on the linear- or Mel-scale, the following two metrics are employed.

$$e_1 = \frac{\sum \|\bar{Y} - \hat{\bar{Y}}\|^2}{\sum \|\bar{Y}\|^2}, \quad e_2 = \frac{\sum w(\hat{\bar{Y}}, \bar{Y})\|\bar{Y} - \hat{\bar{Y}}\|^2}{\sum w(\hat{\bar{Y}}, \bar{Y})\|\bar{Y}\|^2}, \tag{2}$$

where $w(\hat{\bar{Y}}, \bar{Y}) = f(\bar{Y}) + (1 - f(\bar{Y}))f(\hat{\bar{Y}})$ is the weight at each TF unit. e_1 is the normalised absolute error, and e_2 is the normalised version of the perceptually weighted error. From randomly-generated 500 testing sequences, we obtain $e_1 = 2.8\%$ and $e_2 = 0.3\%$ with the proposed encoder. For B1, we have $e_1 = 6.8\%$ and $e_2 = 0.6\%$, approximately twice the error as the proposed method. This is because as compared to B1 that predicts high-dimensional linear-spectrum, our proposed encoder relaxes the learning process to a much lower dimensional space, which leads to a higher accuracy and robustness.

2) System evaluation: Our proposed system uses compact Mel-spectrum to represent the target signal in the encoder, which greatly relaxes the training burden with an improved accuracy as compared to linear-spectrum. Although the Mel-spectrum has preserved perceptually important audio information, there is also inevitable information loss, which is compensated by the additional information introduced via the WaveNet synthesis network that enables high quality synthesis. This extra information is learned from all the audio clips with similar Mel-spectra to that of the target we want to retrieve, but not an exact replica of the original target signal. In other words, the groundtruth target signal is not the reference for the "fake" information introduced by the synthesis network. As a result, direct comparison of the synthesised signal with the target does not reflect the real audio quality as perceived by a listener, which is however the case for most traditional evaluation metrics [21], that often assume the source estimate is a Wiener-filtered version of the dry groundtruth.

To make the best of traditional evaluation metrics in the source separation community, we propose to compensate the extra introduced information as follows. Denote $\mathbf{W} \in \mathbb{R}^{80\times 512}$ as the transform matrix from Mel-spectrum to linear-spectrum $\mathbf{X}(t) = \mathbf{W}\mathbf{X}^{\text{lin}}(t)$ at time t, we can expand the Mel-spectrum to an estimated linear-spectrum via $\mathbf{W}^{+}\mathbf{X}(t)$, where the superscript $+$ denotes pseudo-inverse. The residual information between the groundtruth linear-spectrum $\mathbf{Y}^{\text{lin}}(t)$ and the groundtruth Mel-spectrum $\mathbf{Y}(t)$ can be approximated as

$$\mathbf{R}(t) = \mathbf{Y}^{\text{lin}}(t) - \mathbf{W}^{+}\mathbf{Y}(t).$$

The above information loss can be added back to the estimated target Mel-spectrum to obtain a new linear-spectrum

$$\check{\mathbf{Y}}(t) = \mathbf{W}^{+}\hat{\mathbf{Y}}(t) + \mathbf{R}(t).$$

Inverse STFT is then applied to the loss-compensated linear-spectrum $\check{\mathbf{Y}}$ using the groundtruth phase information. The result is denoted as "Proposed-res-gt". Three evaluation metrics are employed here: perceptual evaluation of speech qual-

[1] Modified from https://github.com/drethage/speech-denoising-wavenet to support our testing data and TensorFlow framework.

TABLE I: Performance evaluations in PESQ, STOI and SDR averaged over 100 test sequences.

	PESQ (-0.5-4.5)	STOI (0-1)	SDR (dB)
Input	1.02	0.85	5.34
Proposed-res-gt	**3.62**	0.95	**15.36**
Proposed	2.61	0.88	NA
B1-gt	3.26	0.94	11.66
B1	2.84	0.91	NA
B2	2.30	0.85	5.35
IBM-gt	2.75	**0.97**	13.55

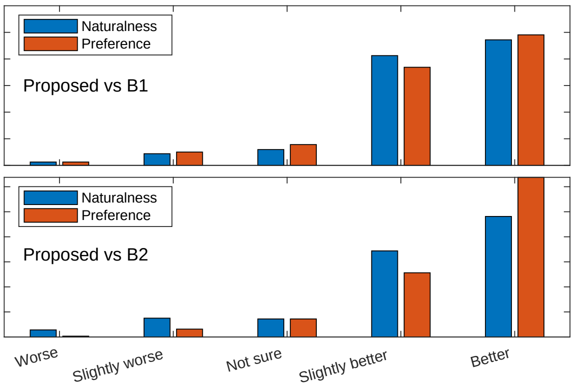

Fig. 3: Listening test results averaged from 320 pairs of data comparing "proposed" with "B1" (top) and "B2" (bottom). The participants scored the paired comparison by judging two statements at five scales: 1 "A sounds more natural than B"; 2 "You overall prefer A over B", where A and B indicate a paired estimated signals with randomised orders.

ity (PESQ, -0.5-4.5) [22], short-time objective intelligibility (STOI, 0-1) [23] and signal-to-distortion ratio (SDR) [21]. For fair comparisons, the groundtruth phase was also integrated into the linear-spectrum regressed by "B1" for time-domain reconstruction, denoted as "B1-gt". We also extracted the ideal binary mask (IBM) from the groundtruth and the interference spectrum, which was applied to the mixture for optimised signal separation using groundtruth phase information, denoted as "IBM-gt". The same evaluation metrics are also applied to denoised signals from the second baseline and the noise corrupted mixture, denoted as "B2" and "Input". The separated signals from "Proposed" and "B1" are directly evaluated as well. Average evaluation results from 100 randomly chosen test files are listed in Table I, from which it can be seen that "Proposed-res-gt" shows significant advantages over other methods. Therefore, if the WaveNet synthesis network could fill in the lost information caused by compact Mel-scale representation, the proposed scheme could achieve results very close to the groundtruth. We need to stress that the synthesis network is independently trained, as a result, data from another speaker might cause timbral changes in phonemes and degrades the performance of the trained model. Interestingly, "B1-gt" even gained a better PESQ value than "IBM-gt", this is because the B1 baseline also retrieves audio information at interference-dominant TF units while the IBM method abandons this information. On the other hand, this also shows the DNN structure, i.e. RNN+hourglass as used in both "B1" and the proposed encoder, is effective in recovering target information while suppressing the interference.

It is worth mentioning that, the total processing time for 100 sequences (ranging from 1 s to 10 s) is around 55 hours using "Proposed", while "B1" and "B2" used 0.5 and 3 hours respectively. This is because the deployed WaveNet in "Proposed" uses causal convolutions, which involves a sample-by-sample outputting strategy.

We further conducted two sets of paired listening tests, to compare "Proposed" with two baselines "B1" and "B2" respectively, with each one involving 16 participants with normal hearing. Each person was asked to compare 20 pairs of randomly-chosen speech signals obtained by the proposed method and one baseline method from the same noise-corrupted mixture. The mixture recording was also provided as a reference. Two evaluation attributes, naturalness and preference, were employed over five-point scales. Specifically, naturalness is a commonly-used metric for speech synthesis, with certain ambiguities in its rating standard [24]. To evaluate naturalness, we focus on levels of artefacts and "muffleness" as in [16]. The preference attribute is much more hedonic. To avoid confusions between these two attributes, their ratings were conducted in two sessions.

The listening test results are shown in Fig. 3. It can be observed that the proposed method outperforms "B1"("B2") with overall much higher quality. It was found that $47\%(48\%)$ sounds much more natural, while another $41\%(34\%)$ are moderately more natural. Only a very small number of audio clips from the proposed method sound less natural than the baseline methods. With quantised scales ranging from 1 to 5, the average naturalness score is 4.29(4.49). Naturalness is a critical attribute in listeners' preference, and the preference score overall has a similar trend as the naturalness distribution. The average preference score is 4.28(4.18). Despite similar trends, the preference score does not highly correlate with naturalness, due to individual bias. The Pearson correlation and the absolute difference for all the paired results between these two metrics are respectively 0.51(0.38), and 0.48(0.68). Overall, the participants tend to prefer audio sequences generated via the proposed method, suggesting its potential for applications requiring high-quality sound.

IV. Conclusions

We have presented a new idea for source separation by synthesizing speech source from intermediate-level acoustic features derived from sound mixtures containing a target speech corrupted by background noise. Our preliminary results in terms of objective and subjective evaluations show that, exploiting WaveNet to synthesize speech from mixtures offer high sound quality, which provides a promising alternative for addressing the artefact issues in traditional source separation methods. In the future, we will consider this framework for more challenging mixing conditions, e.g. reverberant rooms, multi-interference scenarios. In addition, we will also consider training the decoder with encoder output directly to mitigate the degradations that might be introduced by mis-matched training conditions of the two.

References

[1] Q. Liu, W. Wang, P. JB Jackson, and Y. Tang. A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions. In *European Signal Processing Conference*, August 2017.

[2] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, December 2015.

[3] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 31–35, March 2016.

[4] P. Chandna, M. Miron, J. Janer, and E. Gómez. Monoaural audio source separation using deep convolutional neural networks. In *IEEE International Conference on Latent Variable Analysis and Signal Separation*, 02 2017.

[5] K. Qian, Y. Zhang, S. Chang, X. Yang, D. A. F. Florêncio, and M. Hasegawa-Johnson. Speech enhancement using bayesian wavenet. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

[6] D. Rethage, J. Pons, and X. Serra. A wavenet for speech denoising. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[7] S. Pascual, A. Bonafonte, and J. Serrà. SEGAN: Speech enhancement generative adversarial network. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3642–3646, 08 2017.

[8] Q. Liu, Y. Xu, P. Coleman, P. Jackson, and W. Wang. Iterative deep neural networks for speaker-independent binaural blind speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[9] A. van den 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. *CoRR*, abs/1609.03499, 09 2016.

[10] Z. Jin, A. Finkelstein, G. J. Mysore, and J. Lu. FFTNet: A real-time speaker-dependent neural vocoder. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[11] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[12] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi. Deep voice: Real-time neural text-to-speech. *CoRR*, abs/1702.07825, 2017.

[13] S. Maiti and M. Mandel. Speech denoising by parametric resynthesis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 6995–6999, 05 2019.

[14] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499, 2016.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016.

[16] A. van den 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. *CoRR abs/1609.03499*, 2016.

[17] K. Ito. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/, 2017.

[18] K. J. Piczak. ESC: Dataset for environmental sound classification. In *Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press, 2015.

[19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[20] D. Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, April 1984.

[21] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, July 2006.

[22] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 749–752, 2001.

[23] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, March 2010.

[24] R. Dall, J. Yamagishi, and S. King. Rating naturalness in speech synthesis: The effect of style and expectation. In *International Conference on Speech Prosody*, 2014.