neural networks (RNN) and Long-Short term memory layers (LSTM) have been applied for speech enhancement providing promising results [12-13].

In this paper, we analyze the performance of the dilated fully convolutional Wave-U-Net model, proposed in [14] which is an extension of Wave-U-Net introduced by [15] for source separation. This approach is particularly appealing because it operates directly on the audio waveforms, therefore getting rid of hand-crafted features, and has been proved very effective in handling varieties of noise while keeping the computation cost affordable [15].

This paper is arranged as follows: In section 2 we review the related work on speech enhancement. In section 3 the theoretical background of Wave-U-Net is provided. Section 4, presents the results of our experimental analysis while Section 5 concludes the paper with the final discussion and remarks.

II. Related Work

In [16], a speech enhancement algorithm was proposed based on an end-to-end DNN which maps directly the noisy raw waveforms to enhanced waveforms. The DNN architecture consists of 4 fully connected feed-forward layers and works frame-by-frame (60 samples) on isolated words.

Authors in [17] proposed a front-end speech enhancement algorithm based on LSTM layers to improve the performance of automatic speech recognition systems. In this work, several architectures were tested, including: (1) a pipeline architecture of LSTM-based speech enhancement and ASR with sequence training, (2) an alternating estimation architecture, and (3) a multi-task hybrid LSTM network architecture. The proposed models are evaluated on the 2nd CHiME speech separation and recognition challenge, and showed significant improvements relative to prior results.

In [18], a fully-convolutional encoder-decoder DNN based on U-Net was presented for separating singing voices. This system uses spectral features and estimates time-frequency binary masks to separate the speech sources. The main drawbacks of this method are that being based on STFT, many parameters must be tuned and adapted e.g. the window size and hop length, affecting the Time-Frequency resolution and the model accuracy. Moreover, as mentioned in the previous section, working on the spectral representation only enhances the magnitude of the noisy spectrogram while the phase is neglected.

In [19], an end-to-end DNN architecture called Wave-net was proposed for speech enhancement with non-causal and non-autoregressive architecture to reduce the complexity of the model. The proposed model is made by residual layers with a dilation factor that increases by powers of 2 from layer to layer. Moreover, using convolutional layers makes the model flexible in the time dimension, leading to denoised variable-length audio signals.

In [20], the authors proposed an end-to-end approach for speech enhancement based on a fully convolutional neural network. The loss used in the training overcome the gap between the optimization criteria used to train the network and evaluation process, so the model was trained in order to maximize the STOI metric. Experimental results show that the STOI metric actually improves thanks to the consistency between the training and validation criterion.

In [21], S. Pascual et al., taking inspiration from the use of generative adversarial networks (GAN) in computer vision and image processing, proposed SEGAN. This model operates on an end-to-end pipeline showing efficiency, with both objective and subjective evaluations.

Authors of [14] proposed an end-to-end CNN model for speech enhancement, called Wave-U-Net. This model was initially investigated for audio source separation. Results on the VCTK dataset (see section IV of this paper for more details) show that Wave-U-Net outperforms SEGAN, Wiener filters, and many other methods.

In [22], the authors combined GAN and U-Net, proposing a new model architecture called UNet-GAN. The GAN generator network has the same structure as U-Net, while the discriminator is a conventional convolutional neural network with batch normalization layers and Leaky ReLU activation function is used. The model was evaluated under low signal to noise ratio SNR conditions (up to -20dB) in terms of the evaluated metrics PESQ (i.e. perceptual evaluation of speech quality) and STOI (see section IV.B for the definition of STO and PESQ metrics]). Results demonstrate that it significantly improves the speech quality and substantially outperforms other deep models, including SEGAN, Bi-LSTM (trained with phase-sensitive spectrum approximation cost function (PSA-BLSTM)) and Wave-U-Net.

In [23], the authors proposed a convolutional recurrent network for noise suppression and speaker-independent speech enhancement that can be integrated with real-time applications. This speech enhancement is a causal speech enhancement model, with no future information is utilized. They notice that the proposed model architecture has fewer trainable parameters than the LSTMs layers.

Authors in [24], proposed a fully convolutional neural network for the speech enhancement task. This study had two main contributions. First, they suggested that the model parameters dramatically increase in the presence of the fully convolutional layers. Secondly, the fully connected layers have limited capability to preserve the correlation between the features, which is important to generate the output waveform.

From this survey we can observe that speech enhancements still an open and interesting signal processing topic, especially End to End (E2E) models, which work directly on the raw waveform, are getting very popular among the speech community, and showing promising results. In this research we investigate the efficiency of the time-domain speech enhancement model based on Wave-U-Net. We aim to emphasize that the time-domain based on raw waveform models outperforms the frequency-domain models based on STFT. As a novel aspect of this research we propose to optimize a loss function defined by a linear combination of L1 and mean squared error (MSE) between actual and target outputs of the network. We carried out experiments in order to estimate how it can improve the performance of the enhancement process.

III. Overview of Wave-U-Network

In this section, we describe the theoretical concepts of the basic U-Net and of the Dilated Wave-U-Net structures, showing how dilated convolutional layers can increase the intelligibility of the enhanced speech by increasing the amount of context that the neurons can see in the input to predict the output (receptive field).

A. *Overview of U-Net*

Generally, CNN networks are widely used in the field of computer vision and image classification, where the network outputs are the probability of the class label to identify a specific image. Basically, the U-Net architecture is a fully convolutional neural network with downsampled convolutional layers on the network left side followed by another 1-D convolutional layer, called bottleneck layer, and upsampling convolutional layers on the right side of the network. The downsampling block in the network left side has the typical structure of the basic CNNs with multiple convolutional layers without using padding. Each convolutional layer is followed by a ReLU activation function and max-pooling layer for the purpose of downsampling. While the model downsamples the space by 2, it doubles the number of feature channels in the network. The right side of the network consists of upsampling, transposed convolutional layers, which halves the number of feature channels. Moreover, skip connections are used, where each feature map in the upsampling side is concatenated with its corresponding feature maps from the downsampling block.

Skip connections are important because they feed the input of one block with the output of a non-adjacent block (i.e. it preserves the input and output waveform signals to be at the same size). In this way, features maps extracted from downsampling blocks can be used to reconstruct the output of the upsampling blocks. Fig. 2 shows the structure of the U-Net proposed by Ronneberge for medical image segmentation [25].

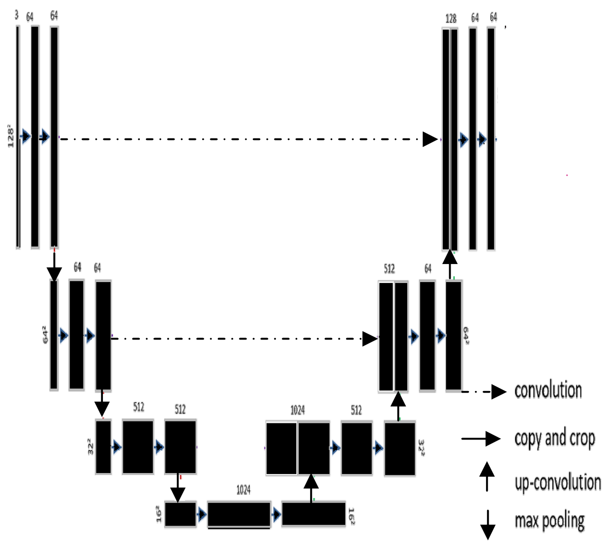

Fig. 2. The architecture of the U-Net network with fully convolutional and max-pooling layers

B. *Wave-U-Net for speech Enhancement*

Wave-U-Net is based on the architecture of U-Net, but the structure is modified to handle audio waveforms as an input i.e. it maps the noisy raw waveforms into clean signals. The network consists of downsampling blocks on the left side and upsampling blocks on the right side, but the 1-D convolutional layers are used due to the nature of the input (i.e. speech waveform).

The input to the Wave-U-Net is a mixture of noisy signals $y[n] \in [-1, 1]^{L\times C}$. The network separates this mixture signals into K source waveforms $x^1 \ldots, x^k$ with $x^k \in [-1, 1]^{L\times C}$ for all $k \in \{1 \ldots K\}$, where C is the number of speech channels and L is the number of audio samples. In the case of the monaural speech enhancement K = 2 and C = 1.

Each block in the Wave-U-Net has convolutional layers followed by a downsampling or preceded by an upsampling operation. The downsampling module is a decimate operation which halves the dimension of the feature map. In the upsampling blocks, the Wave-U-Net is using some combinations such as linear interpolation and transposed convolutions. All the layers, except for the last in the upsampling part, have a Leaky ReLU activation with a negative slope = 0.1. The last layer (block 1 on the upsampling path) has a hyperbolic tangent (Tanh) activation. Fig. 3 shows the basic structure of the Wave-U-net architecture, where the left side corresponds to the 1-D convolutional layers downsample blocks, while the right side represents the 1-D transpose convolutional layers upsampling blocks.

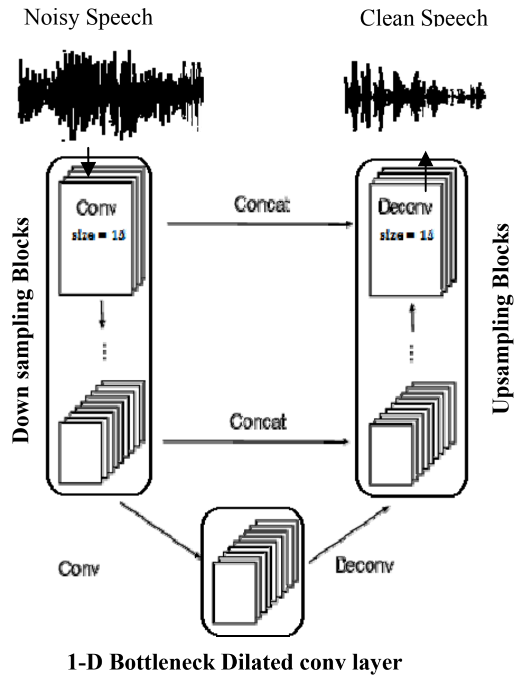

Fig. 3. The architecture of the Wave-U-Net network

C. *Dilated Convolution*

The dilation operation was firstly proposed for the wavelet transform [26] and then was applied to convolutional layers and called dilated convolution. The mechanism of dilation is inflating the kernel by adding spaces between the kernel elements. It allows increasing the receptive field size to capture a larger context for the signal reducing, at the same time, the computation required. Consider a 1-D input signal called z[i] is subjected to a dilated convolution operation to produce an output signal h[i] with a filter w[k] as shown in (3).

$$h[i]=\sum_{k=1}^{K} z[i+rk]w[k] \tag{3}$$

Where r is the dilation rate and k the length of the filter respectively. Note that when r = 1, the dilated convolution is equivalent to the ordinary convolution.

Fig. (4) and Fig (5) show the conventional convolution operation and the dilated convolution operation with r = 1, 2, 4 on 1-D signal where the stride = 1 and kernel size = 3.

In Fig. 4 after three sequential conventional convolution operations, the receptive field is equal to seven, which is linear with the number of layers. On the contrary, when the dilation rate increases exponentially from r = 1, 2, 4, as shown in Fig. 5, the receptive field will also increase to ensure that the larger context of the 1-D signal will be captured.

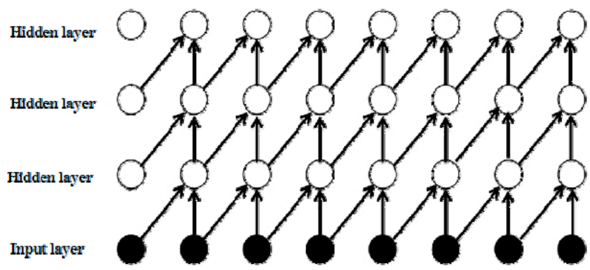

Fig. 4. Three hidden layers of convention convolution operation of CNN.

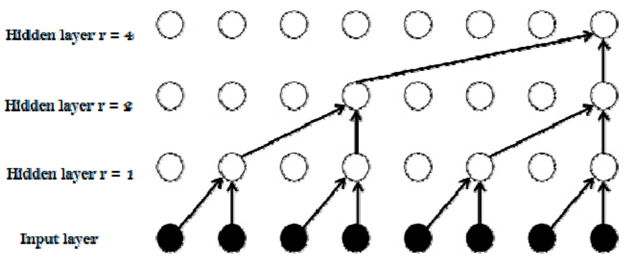

Fig. 5. There hidden layers dilated convolutional operation with dilated rate increasing exponentially as (r = 1, 2, 4).

D. Loss Function

Following one of the main streams in neural speech enhancement, the Dilated Wave-U-Net model is trained using a mean squared error loss (MSE) computed on the target and enhanced samples.

$$MSE=\frac{1}{n}\sum(x_r-x_p)^2 \tag{4}$$

Where n, x_r, x_p are the number of training samples, the clean signals, and the enhanced signals respectively.

In this work, we also investigate the use of a mean absolute error (L1) loss. Basically, L1 loss, which is often used for regression models, is calculated as the sum of absolute differences between the target and the predicted variables as in (5).

$$L1=\frac{1}{n}\sum\left|x_r-x_p\right| \tag{5}$$

Finally, we also investigate linear combination between the two losses in (4) and (5), as shown in (6) with different weights denoted as α.

$$L_T=\frac{\alpha\sum(x_r-x_p)^2+(1-\alpha)\sum\left|x_r-x_p\right|}{n} \tag{6}$$

IV. Experiment

A. Dataset

We evaluate the performance of the Dilated Wave-U-Net on two different datasets.

The first dataset is VCTK [27], which is also used in the original paper [14]. The dataset is publicly available and features 30 speakers from the Voice Bank corpus: 28 speakers are part of the training set while the test set includes the remaining 2 speakers. The clean signals are contaminated with 10 different types of noise (2 artificial and 8 from the Demand database [28]) at 4 different SNRs (15, 10, 5, 0 dB), resulting in 40 different conditions. There are approximately 10 different sentences in each condition per training speaker. The test set consists of 20 different conditions obtained by considering 5 types of noise (all from the Demand database) at 4 SNR (17.5, 12.5, 7.5, and 2.5 dB). There are around 20 different sentences in each condition per test speaker. The test set is totally different from the training set in terms of speakers and conditions.

The second dataset is a noisy version of Librispeech-100, which is 100 hours of reading English speech with a sampling rate of 16 kHz [29]. We selected 10 speakers, resulting approximately 10 hours of clean speech signals. The noisy dataset is obtained by adding noises from the Microsoft Scalable Noisy Speech Dataset (MS-SNSD) [30], which is available at [31]. MS-SNSD has 25 categories of noisy sounds. The Librispeech dataset is contaminated at different SNRs (5dB, 7.5dB, 12dB and 15dB), uniformly distributed. The dataset is split into train, validation, and test sets considering 60%, 20%, and 20% of the data respectively.

B. Evaluation Metrics

The performance of the enhancement process is evaluated using the following metrics for speech intelligibility and quality:

- PESQ: Perceptual evaluation of speech quality, using the wideband version recommended in ITU-T. It is a widely used objective quality measurement standard algorithm. The first step in calculating PESQ metric is time alignment between the referenced signal and the processed signal, then the signals are mapped to an auditory representation using a perceptual model based on power distribution over T-F and compressive loudness scaling and then their differences are taken. Positive differences indicate that components such as noise are present, whereas negative differences indicate that components have been omitted. With PESQ, different scaling factors are applied to positive and negative disturbances in order to generate the so-called symmetrical and asymmetrical disturbances. The final PESQ quality score is obtained as a linear combination of the symmetrical and asymmetrical disturbances, with weights optimized using telephony data. The range of the PESQ metric lies between (-0.5 to 4.5).

- STOI: The short-time objective intelligibility metric is based on a correlation coefficient between the temporal envelopes of the time-aligned clean signal and enhanced speech signal in short-time overlapped segments. Firstly the signals are decomposed using 1/3 octave filter bank followed by segmentation into short-time windows, normalization, clipping and finally compared by means of correlation coefficient. The obtained correlation coefficients correspond to short-time intermediate intelligibility measures for each of the segments, which are then averaged to one scalar value corresponding to the predicted speech intelligibility for the processed signal. The STOI proposed to assess the intelligibility of time-frequency weighted noisy speech and enhanced speech. The STOI metric score ranges from 0 to 1.
- SNR: The signal to noise ratio is the most popular parameter used to measure the level of the desired signal to the level of background noise, and its unit of expression is typically decibels (dB) its range from 0 to ∞.

The higher score for these metrics means better quality and intelligibility.

C. Training

As mentioned in the previous section, the network input consists of a mixture speech signal while its corresponding clean speech signal is used as a target. Due to the variation of length of the signals, they are chunked taking 16384 continuous-time frames randomly selected from the noisy and clean signals.

The model is trained using Adam optimizer with learning rate $=10^{-4}$, decay rates β1 = 0.9 and β2 = 0.999. The batch size is set to 10 and the Leaky ReLU activation function is used with negative slope α = 0.1. The model architecture is composed of 12 convolutional layers with kernel size = 5, stride = 1, padding = 7, and dilation rate = 1 in the downsampling blocks. The resulting dimensions per layer are 16384, 819, 4096, 2048, 1024, 512, 256, 128, 64, 32, 16, 8, and 4. While in the upsampling blocks kernel size = 5, stride = 1, padding= 2 and batch size = 25. Our implementation is in PyTorch and is derived from modifications of an open source repository [32].

D. Results

Table I shows the results of the different evaluation metrics for the VCTK dataset and contaminated Librispeech data. All the metrics were computed on both the noisy and the enhanced signals using MSE loss and L1 loss.

TABLE I. PESQ, STOI, AND SNR OF THE DILATED WAVE-U-NET METHOD ON VCTK AND CONTAMINATED LIBRISPEECH USING MSE AND L1 LOSSES

Datasets	Loss	PESQ	STOI	SNR
VCTK	Unproc.	1.84	0.92	22.2
	L1	2.27	0.78	42.75
	MSE	2.36	0.801	44.31
Librispeech	Unproc.	1.51	0.78	16.6
	L1	1.79	0.89	30.5
	MSE	2.01	0.9	30.9

According to the results, the MSE loss function clearly outperforms the L1 norm loss on both datasets. For the VCTK dataset, the MSE loss outperforms the L1 loss in terms of PSEQ with 2.36 and 2.27 respectively, while for SNR metric the scores are 44.31 and 42.75 for MSE loss and L1 loss respectively. In the same manner, the STOI metric obtained using MSE loss outperforms the score obtained using L1 loss with 0.801 and 0.78 respectively.

In the same context, for the contaminated LibriSpeech dataset, the MSE loss exhibits superior performance over the L1 loss in terms of PESQ metric, with 2.01 and 1.79 for MSE and L1 losses respectively. Despite the noticeable improvement in the PESQ metric score, both STOI and SNR metrics show slight improvement using MSE loss with scores of 0.9 and 30.9 respectively.

From the above results, we can conclude that using MSE loss function obtained the highest scores in terms of PESQ and SNR metrics compared to the unprocessed signals (Unproc.) scores.

Moreover, we tested the combination of the L1 and MSE loss functions with different weight coefficients α (i.e. α = 0.8, 0.2), and considering different learning rates (0.1, 0.0001, 0.000001). The results are given in Table II.

TABLE II. PERFORMANCE OF THE DILATE-WAVE-U-NET BASED ON COMBINED LOSS WITH DIFFERENT WEIGHTS AND LEARNING RATES ON LIBRISPEECH

Learning rate	α	STOI	PESQ	SNR
1×10^{-1}	0.8	0.88	1.67	24.14
	0.2	0.87	1.63	19.92
1×10^{-4}	0.8	0.9	1.91	31.75
	0.2	0.89	1.86	30.57
1×10^{-6}	0.8	0.83	1.34	20.85
	0.2	0.82	1.30	19.51

As expected, the learning rate affects the overall performance. Using the learning rate equals to 1×10^{-4} achieve better performance than others tested learning rates especially in the PESQ and SNR metrics with 1.91 and 31.75 respectively. While, referring to the weight coefficient α, increasing the MSE loss weight tends to better scores for all metrics.

In addition, we investigate the performance of both the VCTK model based on the MSE loss (M0) on the contaminated LibriSpeech dataset and the LibriSpeech model based on the MSE loss (M1) on the VCTK dataset. The results are shown in Table III.

TABLE III. AVERAGE PERFORMANCE OF PESQ, STOI FOR THE M0 AND M1 MODELS

Model	M0	M1
PESQ (Clean –Noisy)	1.51	1.84
PESQ (Enhan –Noisy)	1.39	1.42
STOI (Clean -Noisy)	0.78	0.92
STOI (Enhan-Noisy)	0.71	0.67

According to the results, both models (M0) and (M1) fail to improve both PESQ and STOI metrics, but the contaminated LibriSpeech is slightly affected, compared with the VCTK dataset tested with (M1). As expected, this is due to the different sizes of both datasets and different types of noise used to train both models.

Finally, we compare the results obtained with the Dilated Wave-U-Net on the VCTK dataset with those achieved with the adversarial-based and time-based approaches. Table IV shows these results.

TABLE IV. PERFORMANCE OF THE DILATE-WAVE-U-NET AGAINST STATE-OF-THE-ART BASELINES ON VCTK

Method	STOI	PESQ	SNR
SEGAN [21]	0.930	2.160	-
Wiener [33]	-	2.22	-
v-GAN [34]	0.790	1.410	-
CNN [35]	0.620	1.120	-
CNN-GAN [35]	0.930	2.340	
Dilated Wave-U-Net	0.801	2.360	44.31

All the results from Dilated Wave-U-Net are obtained by re-running the model with different hyper-parameters w.r.t. [14]. Furthermore, the total number of parameter weights of the network is approximately 10 million. According to the results in TABLE IV, it is clearly that the Dilated Wave-U-Net model outperforms the classical method e.g. Wiener proposed in [33] in terms of PESQ metrics with scores 2.360 and 2.22 respectively. In the same manner, the Dilated Wave-U-Net outperforms the state-of-the-art GANs proposed in [21, 34-35]. Regarding to the STOI metric, the Dilated Wave-U-Net outperforms the proposed adversarial method proposed in [34] and the traditional CNN method proposed [35] with scores 0.801, 0.790 and 0.620 respectively. In contrast the original SEGAN proposed in [21] and CNN-GAN [35] outperforms the Dilated Wave-U-Net with 0.930 and 0.801 respectively.

V. CONCLUSION

In this paper, we investigated the performance of Dilated Wave-U-Net using two datasets: VCTK and a contaminated version of Librispeech. Results show that the Dilated Wave-U-Net outperforms the most recent architectures for speech enhancement task based on the time-domain approach. The obtained results outperform the state-of-the-art methods, which means that there is a possible improvement for these models in the speech enhancement task.

In the future work, we will expand our experiments with alternative scenarios i.e. a larger Librispeech dataset with low SNR ratios and other noisy datasets and focus on fine-tuning the model during the training configuration with a view to updating the compromise between generalization and accuracy. On the other side, other loss functions will be investigated in order to improve enhancement quality. Finally, we will integrate this model with a back-end ASR system to train the back-end ASR with enhanced signals estimated from the front end ASR and check the obtained word error rate score. This will show to what extent the speech enhancement module can robust the ASR system.

REFERENCES

[1] D. Wang and J. Chen, "Supervised Speech Separation Based on Deep Learning: An Overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] D. Yu, L. Deng, J. Wu, Y. Gong, and A. Acero, "Improvements on Mel-frequency Cepstrum Minimum-Mean-Square-Error Noise Suppressor for Robust Speech Recognition," 2008 6th International Symposium on Chinese Spoken Language Processing, 2008.

[3] A. Maas, Q. Le, T. O'Neil, O. Vinyals, P. Nguyen, and A. Ng, "Recurrent neural networks for noise reduction in robust asr," in INTERSPEECH, 2012.

[4] J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Overview of speech enhancement techniques for automatic speaker recognition," Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP 96.

[5] P. Karjol, M. Ajay Kumar and P. K. Ghosh, "Speech Enhancement Using Multiple Deep Neural Networks," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 2018, pp. 5049-5052.

[6] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp. 197–210, 1978.

[7] Y. Ephraim and H. V. Trees, "A signal subspace approach for speech enhancement," IEEE International Conference on Acoustics Speech and Signal Processing, 1993.

[8] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," Speech Communication, vol. 49, no. 7-8, pp. 530–541, 2007.

[9] H. R. Guimarães, H. Nagano, and D. W. Silva, "Monaural Speech Enhancement through Deep Wave-U-Net," Expert Systems with Applications, p. 113582, 2020.

[10] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning. Cambridge, MA: The MIT Press, 2017.

[11] N. Mamun, S. Khorram, and J. H. Hansen, "Convolutional Neural Network-Based Speech Enhancement for Cochlear Implant Recipients," Interspeech 2019, 2019.

[12] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. leHershey, and B. Schuller, "Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR," Latent Variable Analysis and Signal Separation Lecture Notes in Computer Science, pp. 91–99, 2015.

[13] A. Kumar and D. Florencio, “Speech Enhancement in Multiple-Noise Conditions Using Deep Neural Networks,” Interspeech 2016, 2016.
[14] M. Craig, and T. Weyde. "Improved speech enhancement with the wave-u-net." arXiv preprint arXiv: 1811.11307, 2018.
[15] D. Stoller, S. Ewert, and S. Dixon, “Wave-u-net: A multi-scale neural network for end-to-end audio source separation,” in Proc. Int. Soc. Music Inf. Retrieval, 2018, pp. 334–340.
[16] S. Tamura, “An analysis of a noise reduction neural network,” International Conference on Acoustics, Speech, and Signal Processing.
[17] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, “Speech enhancement and recognition using multitask learning of long short-term memory recurrent neural networks,” ISCA, 2015.
[18] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, “Singing voice separation with deep u-net convolutional networks,” in Proc. 18th Int. Soc. Music Inf. Retrieval Conf., 2017, pp. 23–27.
[19] D. Rethage, J. Pons, and X. Serra, “A Wavenet for Speech Denoising,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.
[20] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, “End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1570–1584, 2018.
[21] S. Pascual, A. Bonafonte, and J. Serrà, “SEGAN: Speech Enhancement Generative Adversarial Network,” Interspeech 2017, 2017.
[22] X. Hao, X. Su, Z. Wang, H. Zhang, and A. Batushiren, “UNetGAN: A Robust Speech Enhancement Approach in Time Domain for Extremely Low Signal-to-Noise Ratio Condition,” Interspeech 2019, 2019.
[23] K. Tan and D. Wang, “A Convolutional Recurrent Neural Network for Real-Time Speech Enhancement,” Interspeech 2018, 2018.
[24] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017.
[25] O. Ronneberger, “Invited Talk: U-Net Convolutional Networks for Biomedical Image Segmentation,” Informatik aktuell Bildverarbeitung für die Medizin 2017, pp. 3–3, 2017.
[26] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, “A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform,” inverse problems and theoretical imaging Wavelets, pp. 286–297, 1990.
[27] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Speech Enhancement for a Noise-Robust Text-to-Speech Synthesis System Using Deep Recurrent Neural Networks,” Interspeech 2016, 2016.
[28] J. Thiemann, N. Ito, and E. Vincent, “The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings,” 2013.
[29] http://www.openslr.org/12
[30] C. K. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, “A Scalable Noisy Speech Dataset and Online Subjective Test Framework,” Interspeech 2019, 2019.
[31] https://github.com/microsoft/MS-SNSD
[32] https://github.com/haoxiangsnr/Wave-U-Net-for-Speech Enhancement
[33] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” IEEE Trans. on Audio, Speech, and Lang. Process., vol. 26, no. 3, pp. 197–210, 1978.
[34] M. H. Soni, N. Shah, and H. A. Patil, “Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.
[35] N. Shah, H. A. Patil, and M. H. Soni, “Time-Frequency Mask-based Speech Enhancement using Convolutional Generative Adversarial Network,” 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018.