Listening longer to hear better: Dilated FCNs for Speech Enhancement

A report presented to

the faculty of

the Russ College of Engineering and Technology of Ohio University

In partial fulfillment

of the requirements for the degree

Master of Science

Shuyu Gong

May 2022

This report titled

Listening longer to hear better: Dilated FCNs for Speech Enhancement

by

SHUYU GONG

has been approved for

the School of Electrical Engineering and Computer Science

and the Russ College of Engineering and Technology by

Jundong Liu

Associate Professor

Mei Wei

Dean and Moss Professor of Engineering Education

Abstract

GONG, SHUYU, M.S., May 2022, Computer Science

Listening longer to hear better: Dilated FCNs for Speech Enhancement (37 pp.)

Director of Report: Jundong Liu

Deep learning has recently become the dominant paradigm in solving the speech enhancement (SE) problem. The ability to capture long contexts and extract multi-scale patterns is crucial for designing effective SE networks. However, these capabilities often conflict with maintaining a compact network to ensure good system generalization.

In this project, we review the background related to this problem and propose our own solution. We first introduce some neural networks that perform well in solving SE problems. We then explore dilation operations and apply them to fully convolutional networks (FCNs) to further address this problem. Dilation equips the network with a greatly expanded receptive field without increasing the number of parameters. Different strategies for fusing multi-scale dilation and fitting dilation modules are explored in this work.

We conduct experiments based on the noisy VCTK, AzBio sentences, and RASC863 Chinese speech corpus datasets. Our proposed dilated model significantly improves the baseline FCN model and outperforms state-of-the-art SE solutions.

Dedication

To my parents.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Professor Jundong Liu, my research advisor, for his patience and encouragement in this research work. I thank my dissertation committee members Professor Razvan C. Bunescu and Professor Chad Mourning for their help and inspiration for this project. I would also like to thank my colleagues Sun Tao, Wang Zhewei and Zhang Yuanhang for their help during my research.

Table of Contents

List of Tables

Table Page

List of Figures

List of Acronyms

AE Autoencoder
AI Artificial Intelligence
ASPP Atrous Spatial Pyramid Pooling
CNN Convolutional Neural Network
DNN Deep Neural Network
FCN Fully Convolutional Network
GAN Generative Adversarial Network
LSTM Long Short-term Memory
MFCC Mel-frequency cepstral coefficient
MLP Multilayer Perceptron
PESQ Perceptual Evaluation of Speech Quality
RBM Restricted Boltzmann Machine
RF receptive field
RNN Recurrent Neural Network
SE Speech Enhancement
SNN Speech-spectrum-shaped noise
SNR Signal-to-noise Ratio
SSNR Segmental Signal-to-noise Ratio
STFT Short-time Fourier Transform
STOI Short-term Objective Intelligibility
TasNet Time-domain Audio Separation Network
TCNN Temporal Convolutional Neural Network
T-F Time-Frequency
TTB Two Talker Babble

1 Introduction

Enhancement of audio signals in noisy environments plays an important role in many speech-related applications such as speech recognition, hearing aids, and cochlear implants. One recent scenario is voice capture and transmitting over remote working video conferences. During the pandemic, remote working or working from home has been adopted by many companies. Effective collaborations require the videoconference platforms to be able to reduce background noise while capturing human speech clearly. Another popular usage is speech recognition in automobiles. Automobile companies have adopted voice instruction as an advanced solution to control car functions such as air conditioning and phone calls. Cars need to be able to reduce the highway noise in the background for these functions to work well. Other speech enhancement applications include airplane communications, wearable device interactions, among others.

1.1 Area Overview

Speech enhancement (SE) aims to enhance a speech signal (1-D array) by keeping only the human speech signal and eliminating the additive disturbance components. The basic process of SE is illustrated in Fig. 1.1. As the characteristic of human speech varies from each individual, a more sophisticated algorithm needs to be developed to ensure the high quality of SE work across wide use scenarios.

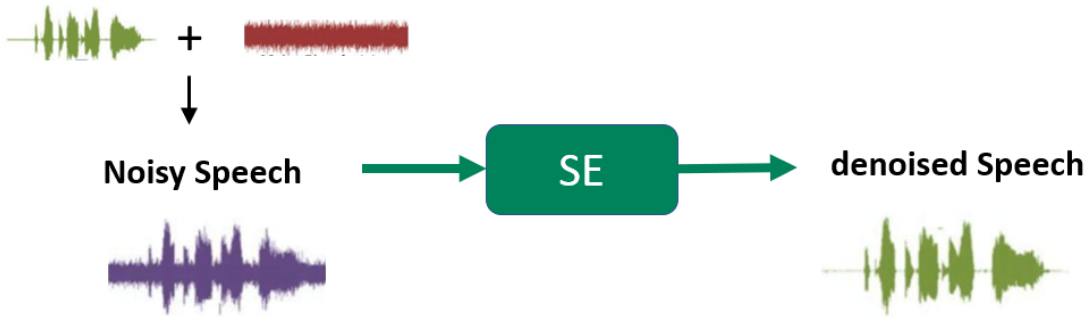

Figure 1.1: The illustration of speech enhancement pricipal.

Traditional Methods Traditional SE techniques commonly operate on the spectral domain and rely on certain high-level features to identify target audio patterns for noise reduction. Spectral subtraction, spectral amplitude fitting, Wiener filtering, and non-negative matrix factorization are among the operations that have been extensively studied. The combination of time-to-frequency transformation and hand-crafted features (e.g., Mel-frequency cepstral coefficients (MFCCs)), however, often results in limited system generalization and poor performance in handling diverse noise conditions.

Deep Learning In recent years, deep neural network-based models (DNNs) have emerged as a new and more powerful paradigm for many artificial intelligence (AI) related applications, including speech enhancement. Unlike traditional machine learning approaches, where certain hand-crafted features (such as fundamental frequency, formants, MFCC, etc.) need to be defined and extracted, DNN models carry out feature extraction in an automatic, data-driven fashion, greatly simplifying the system design. DNN models also facilitate a common platform for the solutions across different application areas, including computer vision and speech signal processing, to be effectively shared. Up to date, the DNN models that have been explored for speech enhancement include autoencoder (AE) [THHJ11], restricted Boltzmann machine (RBM) [WW12a, WW12b, XDDL14], multilayer perceptron (MLP) [WNW14], convolutional neural networks (CNN) [HCG+15], recurrent neural networks (RNN) [TCW19], generative adversarial network (GAN) [PBS17], and fully convolutional networks (FCN) [FTLK17, FWT+18a], among others. Many of the existing SE networks operate on certain time-frequency (T-F) representation of audio signals, generated through short-time Fourier transform (STFT) on fixed-length frames. T-F representations bring great convenience to directly target on the frequency components of the audio signals. The transformations from the waveform inputs to T-F representations and back to the waveform outputs, however, impose a structural constraint

to the networks, which complicates the system design and makes it difficult to predict the network performance.

FCN approach and Issues FCNs on waveform provide a handy and powerful alternative. FCN was firstly developed as an image segmentation solution [L+15] and has since been successfully adopted for image modality conversion, super-resolution and speech signal denoising [FTLK17, FWT+18a]. The fundamental goal of FCNs is to find mappings, with certain desired property, between paired signal sources; therefore they are well-suited to extracting clean waveforms out of noisy inputs [FTLK17, FWT+18a]. The success of FCNs, in great part, is due to their capability of processing input data from multiple spatial or temporal scales. Further improvements over the existing FCNs can be pushed forward by ensuring the models to capture longer contextual information and/or to enhance multi-scale processing. The former (longer context) can be easily achieved through the utilization of larger filters, and making the network deeper provides a solution for the latter. However, a simple combination of these two strategies will lead to a significantly increased number of parameters, which would potentially result in **limited system generalization** and **poor performance in handling diverse noise conditions**.

1.2 Contributions and Report Overview

In this project, we address this challenge by exploring dilated convolution operation and applying it to FCNs. Dilated convolution was originally developed for image segmentation, and its effectiveness to the task has been demonstrated in a number of works [YK16, CPK+18]. We propose to adopt this operation to improve SE FCNs, by allowing the networks to capture longer contexts, a.k.a., "listen longer", as well as to extract more diverse, multi-scale audio features. These features are then combined through a *pyramid pooling* strategy. We also explore different network locations to embed the new operations. All the improvements are made without incorporating additional layers or parameters.

Using Noisy VCTK [VBWTY16, VB+17] and AzBio sentences [SDL+12] datasets, we are able to demonstrate our proposed dilated FCNs outperform the state-of-the-art solutions.

The remainder of this report is organized as follows: we first review certain neural network backgrounds in Chapter 2, which is needed to understand our proposed models. In Chapter 3, the proposed dilated FCN model will be explained in detail. Experiments and results are summarized in Chapter 4, followed by conclusions and a discussion of future work in Chapter 5.

2 Background and Related work

In this chapter, we will discuss the background of convolutional neural networks to understand our proposed solution. We will introduce convolutional neural networks, fully convolutional networks, major solutions for Time domain SE, and the basic idea of dilated neural networks.

2.1 Convolutional Neural Networks and Fully Convolutional Networks

Convolutional Neural Networks (CNNs) are a particular group of deep neural networks that are built with blocks of convolution, deconvolution, activation and pooling operations. They have been widely used in various recognition tasks.

Convolution and deconvolution operations are the core of CNNs. Different from what it is defined in mathematics, the convolution operation in CNNs performs a sliding dot product or sliding inner product of two given sequences. One of the series is data, which is usually called feature maps in deep learning, and the other series are parameters, which are commonly called kernels. This operation is illustrated in Fig. 2.1.

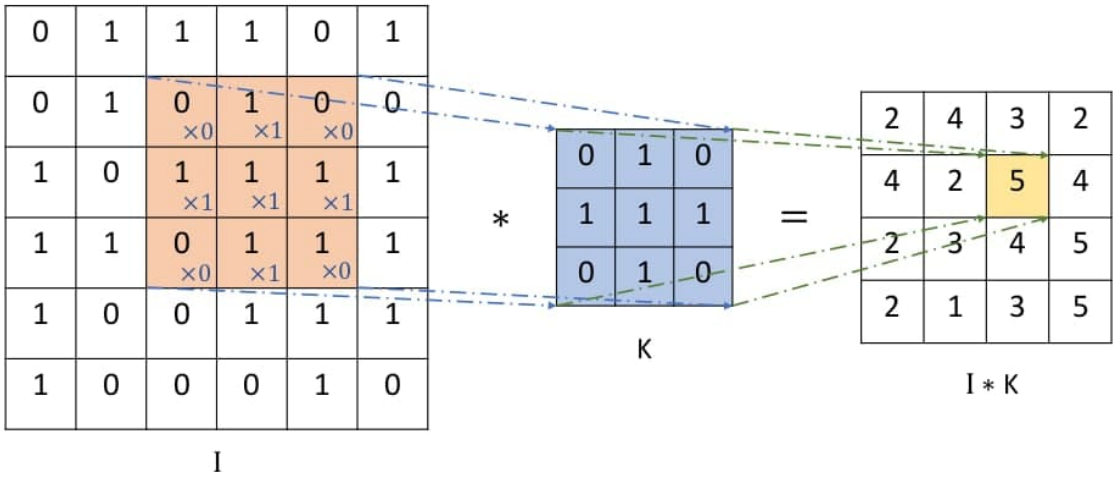

Figure 2.1: An illustration of a convolution operation.

The pooling operation in CNNs is another sliding window operation that computes local statistics within the window, e.g., the minimum or maximum value. This operation reduces the size of its input feature map by a defined fraction, e.g., halving for 1D data and halving for 2D data.

Activation functions in a neural network are on-off gates that control the output of the network. The most widely used activation functions include sigmoid, ReLU and SoftMax. Neural networks can compute non-trivial problems using only a small number of nodes by applying a nonlinear activation function such as a logistic activation function.

In the 2012 ImageNet [DDS+09] multi-class image classification competition, AlexNet [KSH12] won the competition using CNN, revolutionizing the field of image recognition. The architecture of AlexNet is shown in Fig. 2.2

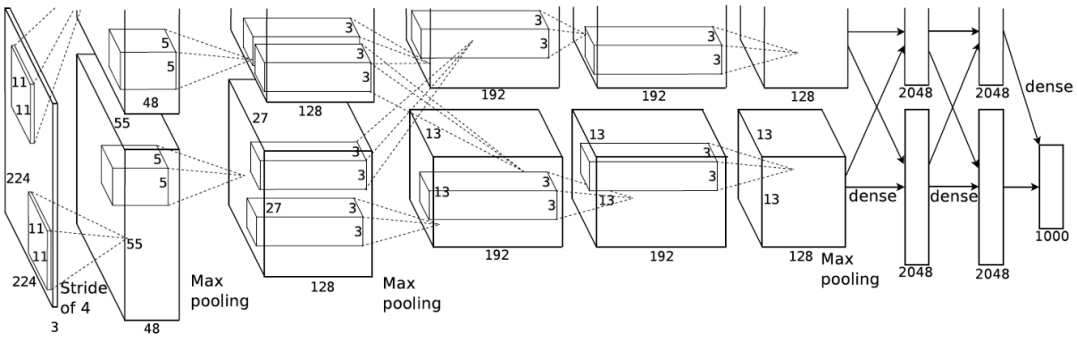

Figure 2.2: An illustration of the architecture of AlexNet. [KSH12]

Fully Convolutional Networks (FCNs) The FCN [L+15] model proposed by Long et al. replaces the fully connected layers at the end of CNN with convolutional layers. Instead of using classification networks, they used convolutional networks to generate rough pixel-level probability maps for image segmentation. However, the upsampling of regular layers is too coarse and image details are lost. They combine the outputs of different convolutional layers and add a skip network architecture to solve this coarse upsampling problem. The

FCN model is the first network that can be considered as a purely CNN-based end-to-end model for pixel-wise segmentation.

U-Net U-Net [RFB15] further refines this FCN concept by introducing skip connections and decoding paths. The entire U-Net forms a U-shaped symmetrical shape. The encoding path on the left side of the network contains only convolutional and pooling layers, while the decoding path of the network is mirrored on the right side of the network and contains only deconvolution and upsampling layers. Each corresponding encoding and decoding layer is connected with a skip connection that carries information from the original feature map. The U-Net architecture has excellent performance on multiple segmentation applications.

2.2 Time domain SE solutions

There are two main branches of time-domain SE solutions: the waveform FCN model and the TasNet-like [LM18, LM19] structure. The Waveform FCN model [FWT+18b, SED18, GIK19, PW19b] directly processes the speech signal in the time domain, thus eliminating the problem of invalid STFT. A widely used waveform FCN model is Wave-U-Net [MW18], which is derived from the speech separation network [SED18]. The original Wave-U-Net adapts SE by treating noise in speech as a distinct sound source. Further improvements are achieved by introducing a local self-attention mechanism in the skip connections of the FCN network [GIK19], which makes the network features pay more attention to speech activity regions. Temporal Convolutional Neural Network (TCNN) [PW19b] is also the FCN model used in SE. By including an additional temporal convolution module, TCNN achieves better performance with fewer parameters and can easily adapt to different frame sizes with simple modifications. Another variant of the time-domain SE solution is TasNet [LM18]. TasNet replaces STFT with a 1D convolutional

layer. It avoids the disadvantages of time-frequency masking, especially the discontinuity of phase information. The original TasNet structure is illustrated in Fig. 2.3.

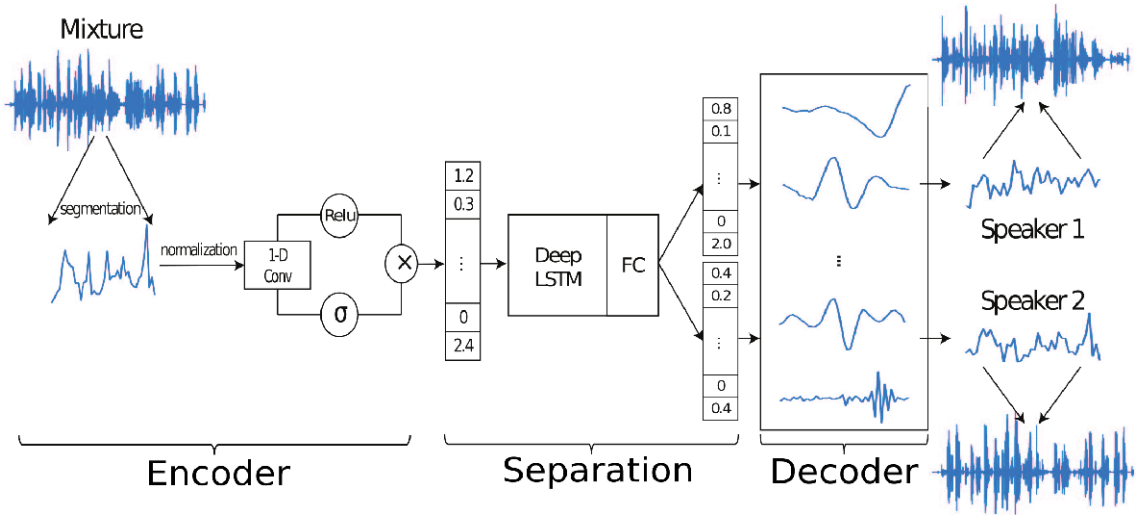

Figure 2.3: Time-domain Audio Separation Network (TasNet) models the signal in the time-domain using encoder-decoder framework, and perform the source separation on nonnegative encoder outputs. Separation is achieved by estimating source masks that are applied to mixture weights to reconstruct the sources. The source weights are then synthesized by the decoder. [LM18]

To further improve the output quality of SE models, comprehensibility-related components are introduced into the neural network. These components include multi-task setups and supervised feature learning. One approach to multi-task setup could be to use a comprehensibility measure as the loss function. In [PW19a] they apply STFT to the output of a waveform FCN SE model to compute the frequency domain loss. Some other works directly improve objective intelligibility metrics, such as STOI [FWT+18b]. However, these metrics are not fully consistent with real-world performance, as the entire metric is only an approximation of real human comprehensibility. To address this problem, these objective evaluation metrics themselves were replaced by a more complex network, as described in [PBS17, DLP18, SSP18, PW18]. The discriminator in SEGAN [PBS17] acts

as a binary judge of speech quality. In order for GAN to directly optimize the evaluation metrics of SE, MetricGAN [FLTL19] generates fake labels based on the metric scores.

2.3 Dilated CNN

CNN and its variants have shown good results in the field of computer vision. However, challenges remain for certain types of tasks, such as semantic image segmentation. The DeepLab system [CPK+18] proposed a network using atrous spatial pyramid pooling (ASPP) technology that can address some of the problems caused by traditional CNNs and has been shown to segment objects robustly at multiple scales. The model illustration of DeepLab system ~~has been~~ shown in Fig. 2.4

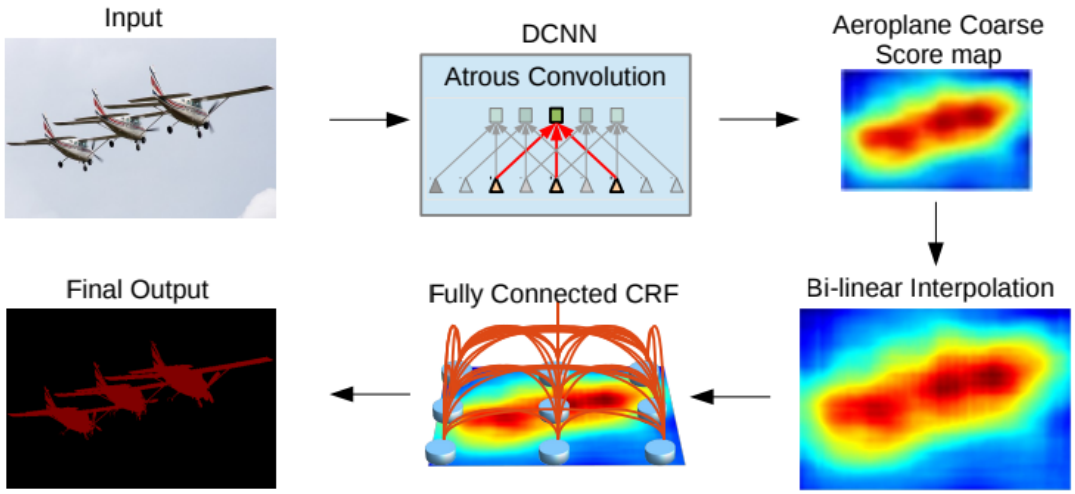

Figure 2.4: DeepLab Model Illustration. A Deep Convolutional Neural Network such as VGG-16 or ResNet-101 is employed in a fully convolutional fashion, using atrous convolution to reduce the degree of signal downsampling (from 32x down 8x). A bilinear interpolation stage enlarges the feature maps to the original image resolution. A fully connected CRF is then applied to refine the segmentation result and better capture the object boundaries. [CPK+18]

There are three main challenges in the application of CNNs to semantic image segmentation. The first is that when used in an FCN fashion, CNNs repeatedly combine

max-pooling and downsampling layers, resulting in reduced spatial resolution. Reducing the number of max-pooling layers by inserting holes between non-zero filter taps can improve network efficiency and improve spatial resolution. The second challenge is that objects can exist at different scale levels. CNNs can only address this challenge by computing features for all scaled versions of the input image. However, this challenge can be addressed more efficiently by using multiple parallel atrous convolutional layers with different sampling rates.

3 Proposed Solution

Our goal is to develop an FCN addition to advance the state-of-the-art of speech enhancement. The design is based ~~the~~ consideration of allowing the networks to listen longer without an increase of the number of parameters. Our efforts are focused on the explorations of 1) dilation convolutions to enlarge the receptive fields of neurons; 2) atrous spatial pyramid pooling (ASPP) module to fuse the multi-scale feature maps; and 3) different network locations in the baseline FCN to embed the ASPP modules.

3.1 Baseline model

Our baseline FCN model is adopted and modified from U-Net [RFB15], which was originally designed for segmentation of cell images. U-Net has an encoder-decoder architecture: in the encoding path, input images are processed through a number of convolution + pooling layers to generate high-level latent features, which are then progressively upsampled along the decoding path to reconstruct the target pixel labels.

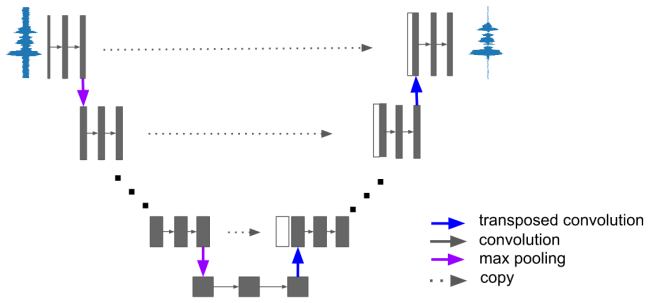

Figure 3.1: Network structure of our Speech-U-Net. Picture is best viewed in color.

To fit our data and task, we modified the original U-Net as follows. First, as the inputs and outputs of our model are one-dimensional waveforms, we replace all the two-dimensional convolution operations in U-Net with one-dimensional convolutions. We keep the original U-Net structure of two convolution layers followed by one pooling/upsampling layer. The network is made deeper to enhance its capability to capture the features in more

scales. The encoding path of our modified U-Net now has 6 convolution-pooling blocks of totally 18 layers. We also use padding in convolution/deconvolution layers to maintain the spatial dimension so that the skip connections can directly concatenate encoding layers with the corresponding decoding layers. The L_1 distance between network predictions (noise-reduced speech) and the ground-truth (clean speech) is used as the objective function. We term our baseline model Speech-U-Net, whose structure is shown in Fig. 3.1.

3.2 Dilated FCN for SE

Noisy speech signals tend to contain components with diverse frequency profiles. To capture them through convolutions, filters of varying sizes would be required. Small filters work well in catching high-frequency noise, but not so effectively for low-frequency sounds. Large filters perform in an opposite way, keen to extract low-frequency components but not high-frequency noise.

The receptive field (RF) at each layer decides the length of the audio signal a neuron can hear. In convolution-based neural networks, including FCNs, the RFs are increased along layers through both pooling and convolutions. Let R_k be the RF of neurons on layer l_k, and it can be computed as:

$$R_k = R_{k-1} + ((f_k - 1) \times \prod_{i=1}^{k-1} s_i) \tag{3.1}$$

where f_i and s_i are the sizes and strides of the filters on layer l_i, respectively. It should be noted that a max pooling of $n \times 1$, from the RF perspective, has the same effect of convolutions of stride equal to n, and the standard convolutions (stride equal to 1) increase the RFs in a linear fashion.

The neurons at the final layers of a network may have RFs that have been enlarged multiple times (e.g., $2^5 = 32$ after 5 max pooling operations). However, due to the high sampling rates of audio signals, the feature maps in speech-enhancement FCNs tend to catch speech segments with short durations. In our baseline model, which has 5 pooling

layers and filters of size 30 at each convolution layer, each neuron at the last layer of the encoding path covers 3686 sample points from the input, which is a 0.23 second sound clip sampled at the rate of $16,000$ per second (only 0.077 second for the rate of $48,000$ per second), quite insufficient to grasp all types of audio patterns. To gain long enough contextual information, one can simply add more pooling layers to makes the network deeper, but that would inevitably lead to a more complicated system with an increased number of parameters, as well as longer training and inference times.

Dilated Convolution Dilated convolutions, supporting exponentially expanding RFs without the loss of resolution or coverage, can provide a remedy. Also, such expansion can be achieved without the need to increase the number of parameters. The basic idea of dilation is to space out the elements to be summed in convolution by a dilation factor, as illustrated in Fig. 3.2. The convolutions in the bottom layer are regular 3×1 convolutions. The middle layer has a dilation factor of 2, so the effective RF at each neuron covers 7 audio samples. The top layer convolutions are dilated by 4, producing a 15×1 RF/coverage. In general, the receptive filed R_k of a neuron on a dilated convolution layer l_k is enlarged to:

$$R_k = R_{k-1} + ((f_k - 1) \times d_k \times \prod_{i=1}^{k-1} s_i) \tag{3.2}$$

where d_i is the dilation factor of layer l_i. Comparing with the standard version, dilated convolutions increase RFs without introducing more parameters. In addition, dilated convolutions produce exponentially expanding RFs with depth, which is in contrast to linear expansion produced by standard convolutions.

Fusion of multiscale dilations through ASPP Dilated convolutions allow us to listen longer. To extract audio patterns with great varieties, however, multiple dilation factors should be involved. Fig. 3.3 shows an example of applying dilation convolutions with 4 different factors. How to integrate the extracted features is a practical issue. In this work, we adopt a strategy similar to the Atrous Spatial Pyramid Pooling (ASPP) scheme proposed

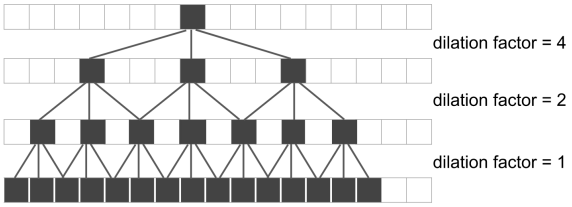

Figure 3.2: Illustration of dilated convolutions, where the dilation factors are 1, 2 and 4, respectively. Refer to text for details.

in DeepLab [CPK+18]. More specifically, we conduct dilation convolutions of 4 different factors in parallel and concatenate the resulted feature maps into outputs. These 4 filters are called a *dilation group*. Within each group, the filters have the same number of parameters (3 in the example of Fig. 3.3), but cover different signal ranges because of the varying filter lengths. For an input feature map of dimension $L \times C$, where L is the size of the map and C is the number of channels, we set the number of the dilation groups to $C/4$, to ensure the output feature maps to maintain the dimension of $L \times C$. Comparing with directly utilizing C filters in the standard fashion, our dilated setup does not increase the number of parameters, but enables the network to capture longer signals with varying lengths.

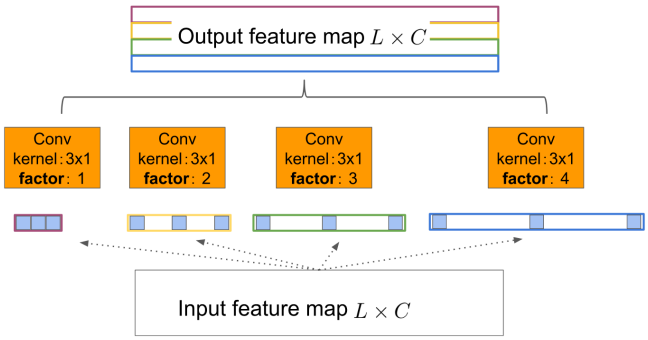

Figure 3.3: Illustration of fusion of multiscale dilation through ASPP. Refer to text for more details.

It should be noted that dilation has been utilized in a very recent work by Tan *et al.* [TCW19] in a CNN+RNN model. To the best of our knowledge, our work is the first attempt to explore dilated convolution to improve FCN models for speech enhancement.

Locations to add the dilation module To exert the power of dilations + ASPP, the next step is to integrate the proposed dilation module into our baseline FCN model. To set up a proper ASPP installation, we run into two practical issues. The first question is regarding how to set the dilation factors. Our answer is based on the fact that the audio patterns to be captured in this work are mainly human phonetic symbols, whose lengths and frequencies tend to concentrate around certain ranges. It is necessary to have dilations of different ratios, but the coverages do not have to be dramatically diverging in scale. With this observation, we choose a relatively slow growing sequence, 1, 2, 3 and 4, as the dilation factors of our ASPP convolutions

The second question is where to install the ASPP replacement. While ASPP can be installed anywhere in the network, we hope our dilated convolutions filters, even with limited number of parameters, can cover relatively long signal spans. In this regard, the end of the encoding path would be an ideal place to add ASPP, as the neurons on this layer have the largest RFs in the entire network. The further enlarged RFs by ASPP would provide a best realization of our goal of "listening longer". This setup is illustrate in Fig. 3.4.(a).

An alternative place to explore the ASPP replacement could be around the end of the decoding path, which consists of two convolution layers, as shown in Fig. 3.4(b). Replacing the first of the two layers with an ASPP would potentially allow the network to decompose the features into different scales, before they will be finally merged to face the ground-truth. Adding ASPP here would provide a (last) chance for the network to *correct* the integration from previous layers. With the analysis of these two choices, very naturally, another setup worth exploring would be the combination of the both. In our experiments, all three ASPP replacement schemes, as illustrated in Fig. 3.4, are examined.

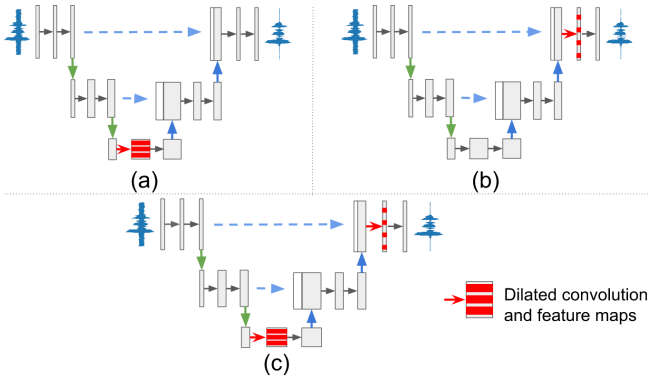

Figure 3.4: Three ASPP replacement schemes: a) adding ASPP in the middle layer of the network; b) adding ASPP around the end of the decoding path; c) combination of a) and b).

4 Experiments and Results

4.1 Data sets

To evaluate the effectiveness of the proposed dilation + ASPP for speech enhancement, we conduct experiments on three groups of datasets. The first experiment is based on the Noisy VCTK dataset by Valentini *et al.* [VBWTY16, VB+17], which consists of two training sets (28 and 56 speakers respectively) and a test set. We choose the 28-speaker set, in which 14 males and 14 females were recorded with around 400 sentences for each person. Noise of 10 different types, either synthetic or real, have been added to the clean speech with 4 signal-to-noise (SNR) levels (15 dB, 10 dB, 5 dB and 0 dB, respectively). Totally there are 11,571 sentences in the training set. The test dataset contains 824 sentences from two new speakers (one male and one female). Five types of noise, which are different from the 10 types in the training set, have been added. The SNR values for test set are 17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB, respectively. All audio clips are sampled at 48kHz, with each time point represented as a 24-bit integer.

The second experiment was based on AzBio English sentences that were developed by Spahr *et al.* [SDL+12]. The dataset consisted of 33 lists with 20 sentences in each list. The sentences ranged from 3 to 12 words (median = 7) in length. All speech sentences were sampled at 22,050 Hz and were spoken by 2 male and 2 female adult speakers [SDL+12]. Two types of masking noise were added to the sentences to achieve the desired SNRs (3 dB and 6 dB): speech-spectrum-shaped noise (SSN) and two talker babble (TTB).

The third experiment was based on RASC863 - annotated 4 regional accent speech corpus [ras] that were developed by Phonetics Lab, Institute of Linguistics, Chinese Academy of Social Sciences. This corpus is made of three different types of speech content: spontaneous speech, read speech, and selected dialectical words. The spontaneous speech is generated by asking each speaker to give a 4-5 minute spontaneous speech on a topic

from our prepared topic sheet or a topic he selected by himself. Besides this topic-oriented speech, each speaker was also asked to spontaneously answer 15 questions. The read speech contains 2200 automatically selected phonetically balanced sentences of which 460 were frequently used in daily life. The corpus then prepared frequently used words in daily life that are different from Standard Chinese for each dialectal region. 15 dialectal words have been read by each speaker. There are four different regions. The corpus recruited 200 speakers for each region total of 800 speakers. The age, sex, and educational background of these speakers are balanced. Speech-spectrum-shaped noise (SSN) were added to the sentences to achieve the desired SNRs (-6 dB, -3 dB, 0 dB, 3 dB and 6 dB).

4.2 Preprocessing

In all experiments, we apply a preprocessing step to downsample all audio clips to 16kHz, and scale their amplitudes to [0,1]. The training set are then split randomly into training, validation and test sets with the ratio of 8:1:1. Each audio sentence is cut into multiple clips of 1 second long, which are taken as the inputs to the network. We extract the clips with a half second overlap as an approach of data augmentation. End-of-sentence clips, if short than 0.5 second, are discarded and not included as training samples.

4.3 Evaluation metrics

Four evaluation metrics are used in this work. They are signal-to-noise ratio (SNR), segmental signal-to-noise ratio (SSNR), perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility measure (STOI). SSNR calculates the average SNRs of short segments (15 to 20 ms long). PESQ evaluate the speech quality using the wide-band version recommended in ITU-T P.862.2 [Rec05]. STOI [THHJ10] produces indicators for the average intelligibility of the degraded speech.

4.4 Results

Totally four models are evaluated in our experiments with the Noisy VCTK dataset. They are the baseline model Speech-U-Net, and three dilation models with the ASPP replacements in the middle of Speech-U-Net (bottom layer), end of the network, and both locations, respectively. The evaluations were conducted on two test sets: the first one is the held-out set, which is the 10% of the training set; the other is the official test available in the dataset.

Table 4.1: Experimental results on Noisy VCTK dataset.

Dataset	Model	SNR	SSNR	PESQ	STOI
Held-out Test	Input	6.040	-0.092	1.467	0.838
	Speech-U-Net	14.504	7.159	1.849	0.857
	ASPP-middle	**15.042**	**7.718**	**1.882**	0.859
	ASPP-end	14.389	7.181	1.827	0.873
	ASPP-middle+end	14.92	7.665	1.788	**0.877**
Official Test	Input	8.544	1.878	1.982	0.922
	Speech-U-Net	17.454	7.955	2.338	0.900
	ASPP-middle	**18.418**	**8.862**	**2.361**	0.902
	ASPP-end	17.001	8.470	2.262	**0.930**
	ASPP-middle+end	17.529	6.521	2.152	0.928
Official Test	SEGAN		7.73	2.16	

The results are shown in Table 4.1. It is evident that the baseline Speech-U-Net already performs rather impressively, achieving an average enhancing performance of 9 dB. Among the three dilation models, ASPP replacement at the middle (ASPP-middle) produces the best results, significantly outperforming all other competing solution in SNR and SSNR. ASPP-end model does not help, performing even worse than the baseline model in SNR and SSNR. The combination of *at-middle* and *at-end*, unsurprisingly, has performance in-between of the two installations. Comparing the results on the held-out and official test sets,

the improvements made by ASPP-middle are more significant for the latter (official test set). Considering that the official test data are acquired from new speakers with different SNRs, therefore have different data distributions than the training set, the comparison indicates that the ASPP-middle is more robust and has a better generalization capability than the baseline model. Such desired properties should be attributed to the enhanced multi-scale processing brought by the dilation operations. In other words, "listening longer" does make the network hear better. Fig. 4.1 shows the comparison of ASPP-middle with Speech-U-Net on a particular audio segment. Ground-truth waveforms are shown in blue, and red lines are the predictions. For the highlighted audio segment in the waveform pictures, Speech-U-Net makes rather flat predictions, failing to capture the fluctuations. Our ASPP-middle, on the other handle, makes accurate predictions for the entire segment.

For PESQ and STOI, none of the three ASPP models produces improvements over the baseline. This can be in part explained by the choice of the objective function in our models. The network updates in our models are driven to minimize the L_1 difference between predictions and ground-truth, which is highly related to SNR/SSNR, but does not directly involve perception and intelligibility components. To replace the L_1 distance with a STOI-based objective [FWT+18b], our models are expected to produce improved performance measured by PESQ and/or STOI.

It should be noted that our ASPP-middle model also has a higher average SSNR value than the reported number from SEGAN [PBS17], a state-of-the-art speech enhancement solution. While we do not intend to make a head-to-head quantitative comparison, as different experiment setups are used in the studies, the significant improvements in SSNR can nevertheless be regarded as a side evidence of the effectiveness of our ASPP-middle model. In addition, our ASPP-middle network, working as a generator, can be connected with a discriminator to make a full GAN model for further improvements.

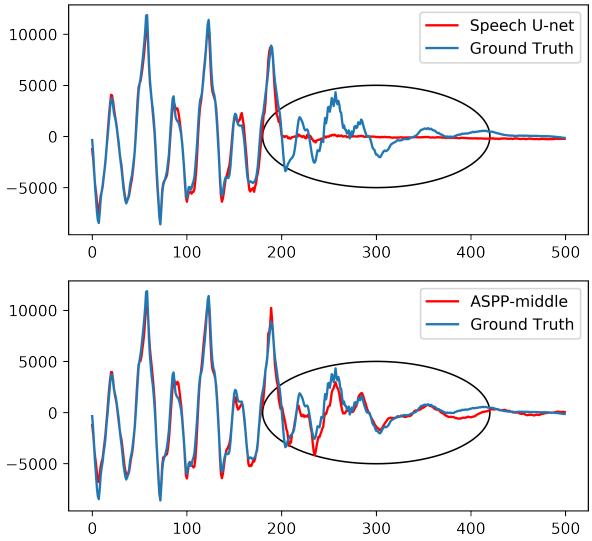

Figure 4.1: Comparison of Speech-U-Net (top row) and ASPP-middle (bottom row) on a particular audio clip.

Based on the results and observation from the VCTK experiment, we chose ASPP-middle as our proposed model. We further evaluate its capability using the AzBio sentence dataset. Speech-U-Net is still taken as the baseline model. The comparisons are shown in Table 4.2. Similar to the first experiment, ASPP-middle produces larger speech enhancements than the baseline, measured in SNR and SSNR. In summary, our proposed ASPP-middle approach consistently improves over the baseline, demonstrating the benefits of dilation convolutions, as well as the fusion and installation setups we designed.

Table 4.2: Experimental results of Speech-U-Net and ASPP-middle on AzBio dataset.

Model	**SNR**	**SSNR**	**PESQ**	**STOI**
Input	2.697	-4.204	1.062	0.816
Speech-U-Net	9.091	-1.534	**1.334**	**0.834**
ASPP-middle	**9.348**	**-1.369**	1.315	0.797

To determine the capability of the ASPP-middle model for the Chinese language, we evaluate it using the RASC863 dataset. The comparison is shown in Table 4.3. Similar to the first and second experiments, ASPP-middle produces greater speech enhancement than the baseline, measured in SNR, PESQ and STOI. In conclusion, our proposed ASPP-middle method significantly improves the baseline model, showing the benefits of dilated convolutions as well as our designs.

Table 4.3: Experimental results of Speech-U-Net and ASPP-middle on RASC863 dataset.

Model	**SNR**	**SSNR**	**PESQ**	**STOI**
Input	1.998	-5.121	1.798	0.660
Speech-U-Net	6.165	**0.620**	2.419	0.776
ASPP-middle	**6.540**	0.054	**2.512**	**0.787**

5 Conclusions and future works

The performance of Speech-U-Net shows that FCN can be successfully used as a waveform-based SE architecture. However, Speech-U-Nets still make flat predictions for some fluctuations in audio clips due to the limited field of view. To address the limitations of FCN on SE tasks and further improve performance, we propose a module named ASPP, which relies on dilated convolutional layers to capture longer context and apply it to the baseline Speech-U-Net .

We train and test the performance of our proposed network on three different datasets with different languages and different types of noise. We found that the ASPP module improves FCN's SE performance on all datasets we examine when fitted to the proper location of the FCN (i.e., in the middle). It performs particularly well on the Chinese speech dataset. One possible explanation for this might be that Chinese has embedded tones. Since pitch differences can lengthen or shorten the speech of the same word, it is difficult to capture content if the network has only a limited receptive field.

An advantage of the ASPP module is that it does not introduce extra parameters to the network, as it only modifies the stride of the convolutional scan. This means that our training can be done in the same time period with the same computing power, which demonstrates the advantage of our ASPP-middle architecture over FCN. Although larger neural networks generally lead to better performance, an efficient and fast network is very beneficial in everyday life, as it can be installed in portable devices such as cell phones and hearing aids.

Future work may be to explore the application of our model on more datasets as well as mixed language datasets. Different languages have different characteristics in terms of phonemes and tones, leading to different generalizations for speech enhancement tasks. Some languages require connections between longer contexts that even a single ASPP module might not handle, such as Japanese and French.

Another future work may be to explore the potential integration of our proposed module with self-supervised learning (SSL). We hope that the rich speech information embedded in the speech SSL model can boost our denoising performance to a higher level.

REFERENCES

[CPK+18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848

[DLP18] Chris Donahue, Bo Li, and Rohit Prabhavalkar. Exploring speech enhancement with generative adversarial networks for robust speech recognition. In *2018 IEEE ICASSP*, pages 5024–5028. IEEE, 2018.

[FLTL19] Szu-Wei Fu, Chien-Feng Liao, Yu Tsao, and Shou-De Lin. MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2031–2041. PMLR, 09–15 Jun 2019. URL: http://proceedings.mlr.press/v97/fu19b.html

[FTLK17] Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai. Raw waveform-based speech enhancement by fully convolutional networks. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 006–012. IEEE, 2017.

[FWT+18a] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi 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 (TASLP)*, 26(9):1570–1584, 2018.

[FWT+18b] Szu-Wei Fu, Tao-Wei Wang, Yu Tsao, Xugang Lu, and Hisashi 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 (TASLP)*, 26(9):1570–1584, 2018.

[GIK19] Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. Attention wave-u-net for speech enhancement. In *2019 IEEE WASPAA*, pages 249–253. IEEE, 2019.

[HCG+15] Like Hui, Meng Cai, Cong Guo, Liang He, Wei-Qiang Zhang, and Jia Liu. Convolutional maxout neural networks for speech separation. In *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pages 24–27. IEEE, 2015.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

[L+15] Jonathan Long et al. Fully convolutional networks for semantic segmentation. In *Proceedings of CVPR*, pages 3431–3440. IEEE, 2015.

[LM18] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *ICASSP*, pages 696–700. IEEE, 2018.

[LM19] Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM TASLP*, 27(8):1256–1266, 2019.

[MW18] Craig Macartney and Tillman Weyde. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*, 2018.

[PBS17] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, pages 3642–3646, 2017.

[PW18] Ashutosh Pandey and Deliang Wang. On adversarial training and loss functions for speech enhancement. In *2018 IEEE ICASSP*, pages 5414–5418. IEEE, 2018.

[PW19a] Ashutosh Pandey and DeLiang Wang. A new framework for cnn-based speech enhancement in the time domain. *IEEE/ACM TALSP*, 27(7):1179–1188, 2019. doi:10.1109/TASLP.2019.2913512

[PW19b] Ashutosh Pandey and DeLiang Wang. Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *2019 IEEE ICASSP*, pages 6875–6879. IEEE, 2019.

[ras] Rasc863-annotated 4 regional accent speech corpus. http://shachi.org/resources/1223.

[Rec05] ITUT Rec. P. 862.2: Wideband extension to recommendation p. 862 for the assessment of wideband telephone networks and speech codecs. *International Telecommunication Union, CH–Geneva*, 2005.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[SDL+12] Anthony J Spahr, Michael F Dorman, Leonid M Litvak, Susan Van Wie, Rene H Gifford, Philipos C Loizou, Louise M Loiselle, Tyler Oakes, and Sarah Cook. Development and validation of the azbio sentence lists. *Ear and hearing*, 33(1):112, 2012.

[SED18] 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.

[SSP18] Meet H Soni, Neil Shah, and Hemant A Patil. Time-frequency masking-based speech enhancement using generative adversarial network. In *2018 IEEE ICASSP*, pages 5039–5043. IEEE, 2018.

[TCW19] Ke Tan, Jitong Chen, and DeLiang Wang. Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):189–198, 2019.

[THHJ10] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217. IEEE, 2010.

[THHJ11] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

[VB+17] 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.

[VBWTY16] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *SSW*, pages 146–152, 2016.

[WNW14] Yuxuan Wang, Arun Narayanan, and DeLiang Wang. On training targets for supervised speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 22(12):1849–1858, 2014.

[WW12a] Yuxuan Wang and DeLiang Wang. Boosting classification based speech separation using temporal dynamics. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[WW12b] Yuxuan Wang and DeLiang Wang. Cocktail party processing via structured prediction. In *Advances in Neural Information Processing Systems*, pages 224–232, 2012.

[XDDL14] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters*, 21(1):65–68, 2014.

[YK16] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations*, 2016.