GATED RESIDUAL NETWORKS WITH DILATED CONVOLUTIONS FOR SUPERVISED SPEECH SEPARATION

Ke Tan[1], *Jitong Chen*[1] *and DeLiang Wang*[1,2]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Center for Cognitive and Brain Sciences, The Ohio State University, USA
{tan.650, chen.2593, wang.77}@osu.edu

ABSTRACT

In supervised speech separation, deep neural networks (DNNs) are typically employed to predict an ideal time-frequency (T-F) mask in order to remove background interference. However, the performance of DNNs is frequently degraded for untrained noises and speakers. Inspired by recent research on dilated convolutions for context aggregation, we propose a novel convolutional neural network (CNN) to deal with noise- and speaker-independent speech separation. The proposed model incorporates dilated convolutions, gating mechanisms and residual learning. We find that the proposed model consistently outperforms a state-of-the-art long short-term memory (LSTM) based model in terms of objective speech intelligibility and quality. Additionally, the proposed CNN is more computationally efficient than the LSTM model.

Index Terms— dilated convolutions, residual learning, gated linear units, phase-sensitive mask, speech separation.

1. INTRODUCTION

Speech separation aims to separate target speech from background interference [1]. Inspired by the concept of time-frequency masking in computational auditory scene analysis (CASA) [2], speech separation is formulated as a supervised learning problem in recent years, where a mapping from noisy acoustic features to a T-F mask is learned by a deep neural network [3].

The ideal ratio mask (IRM) [4], which suppresses noise energy within each T-F unit, is frequently used as the training target in supervised speech separation. Alternatively, one can estimate the phase-sensitive mask (PSM) [5], which yields a higher signal-to-noise ratio (SNR). It is defined by

$$PSM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|}\cos\theta \tag{1}$$

where $|S(t,f)|$ and $|Y(t,f)|$ denote spectral magnitudes of clean speech and noisy speech within a T-F unit at time frame t and frequency channel f, respectively. θ represents the difference between the clean speech phase and the noisy speech phase within the T-F unit. In this study, we use the PSM as the training target to perform supervised speech separation.

For supervised speech separation, contextual information can facilitate mask estimation. A window of consecutive time frames is typically utilized to provide temporal contexts for mask estimation at each time frame. Contextual information, however, may be insufficiently leveraged given a fixed-length context window. A recent approach [6] utilizes long-term contexts by treating supervised speech separation as an utterance-to-utterance mapping. In [6], Chen *et al.* proposed a recurrent neural network (RNN) with four stacked hidden LSTM layers to deal with speaker- and noise-independent speech separation. With a large number of training speakers, the LSTM model generalizes well to untrained speakers and noises, and significantly outperforms a DNN based model.

In convolutional neural networks, contextual information is augmented typically through the expansion of the receptive fields. One way to achieve this goal is to increase the network depth, which decreases computational efficiency and typically results in vanishing gradients [1]. Another way is to enlarge the kernel size, which likewise raises computational burden and training time. Previous CNN based methods [7] [8] [9] for speech separation do not capture long-term temporal dependencies due to their limited receptive fields. Dilated convolutions (or atrous convolutions) were first proposed for multi-scale context aggregation in [10]. They can significantly expand receptive fields while maintaining the network depth and the kernel size. Motivated by recent research [10] [11] [12] on dilated convolutions, we propose a novel gated residual network (GRN) with dilated convolutions to deal with speech separation. We find that the proposed GRN leads to consistently better performance and higher computational efficiency than the LSTM model in [6]. In addition, our proposed GRN substantially outperforms previous convolutional networks for speech separation.

The rest of this paper is organized as follows. We gives a detailed description of our proposed model in Section 2. The experimental setup and results are presented in Section 3. We

This research was supported in part by an NSF grant (IIS-1409431), an NIDCD grant (R01 DC012048), and the Ohio Supercomputer Center.

conclude in Section 4.

2. ALGORITHM DESCRIPTION

2.1. Dilated convolutions

Dilated convolutions were originally developed in the algorithme à trous, an algorithm for wavelet decomposition [13]. Formally, a 2-D discrete convolution operator $*$, which convolves signal F with kernel k of size $(2m+1) \times (2m+1)$, is defined as

$$(F * k)(\mathbf{p}) = \sum_{\mathbf{s}+\mathbf{t}=\mathbf{p}} F(\mathbf{s})k(\mathbf{t}) \tag{2}$$

where $\mathbf{t} \in [-m, m]^2 \cap \mathbb{Z}$. A generalized version of the operator $*$, which is denoted by $*_r$, can be defined as

$$(F *_r k)(\mathbf{p}) = \sum_{\mathbf{s}+r\mathbf{t}=\mathbf{p}} F(\mathbf{s})k(\mathbf{t}) \tag{3}$$

where r denotes a dilation rate. Thus, we refer to $*_r$ as an r-dilated convolution. Note that the common convolutions can be regarded as 1-dilated convolutions. Analogously, a 1-D r-dilated convolution can be defined as $(F *_r k)(p) = \sum_{s+rt=p} F(s)k(t), t \in [-m, m] \cap \mathbb{Z}$. With kernels of size 3, the receptive fields of conventional convolutions and dilated convolutions are illustrated in Fig. 1.

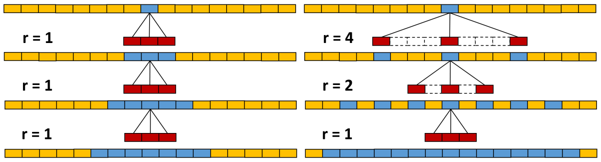

Fig. 1. Left: a 1-D CNN ($r = 1$) with three conventional convolutional layers. Right: a 1-D CNN with three dilated convolutional layers, of which the dilation rates r are 1, 2 and 4, respectively. We treat the blue unit in the top layer as the unit of interest, and the rest of the blue units indicate its receptive fields in each layer.

Mathematically, the receptive field size in conventional 1-D convolutions is linearly correlated with the layer depth, while the receptive field size in dilated 1-D convolutions is exponentially correlated with the layer depth if the kernels are applied with exponentially increasing dilation rates as shown in Fig. 1.

2.1.1. *Time-dilated convolutions*

Sercu *et al.* [12] developed so-called *time-dilated convolutions* by using an asymmetric version of dilated spatial convolutions (or 2-D convolutions) with dilation in the time direction but not in the frequency direction. In this study, we use the 1-D version of time-dilated convolutions, where dilation is applied to temporal convolutions (or 1-D convolutions).

2.1.2. *Frequency-dilated convolutions*

To aggregate contextual information over the frequency dimension, we create kernels of size 1×3 for dilated convolutions over the frequency dimension, where the kernels are placed along the frequency direction. For convenience, we call them *frequency-dilated convolutions*.

2.2. Gated linear units

Gating mechanisms potentially facilitate modeling more complex interactions by controlling the information flow. LSTM-style gating mechanisms are applied to convolutions in [14]:

$$\begin{aligned} \mathbf{y} &= \tanh(\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2) \\ &= \tanh(\mathbf{v}_1) \odot \sigma(\mathbf{v}_2) \end{aligned} \tag{4}$$

where $\mathbf{v}_1 = \mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1$ and $\mathbf{v}_2 = \mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2$. $\mathbf{W}$'s and $\mathbf{b}$'s represent kernels and biases, respectively. σ denotes *sigmoid* function. $*$ and $\odot$ denote convolution operation and element-wise multiplication, respectively. The gradient of the LSTM-style gating is

$$\begin{aligned} \nabla[\tanh(\mathbf{v}_1) \odot \sigma(\mathbf{v}_2)] = \tanh'(\mathbf{v}_1)\nabla\mathbf{v}_1 \odot \sigma(\mathbf{v}_2) \\ +\sigma'(\mathbf{v}_2)\nabla\mathbf{v}_2 \odot \tanh(\mathbf{v}_1) \end{aligned} \tag{5}$$

where $\tanh'(\mathbf{v}_1), \sigma'(\mathbf{v}_2) \in (0, 1)$. Typically, the vanishing gradient problem arises as the network depth increases, while it becomes even more severe with such gating due to the downscaling factors $\tanh'(\mathbf{v}_1)$ and $\sigma'(\mathbf{v}_2)$. To tackle this problem, Dauphin *et al.* [15] introduced gated linear units (GLUs):

$$\begin{aligned} \mathbf{y} &= (\mathbf{x} * \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{x} * \mathbf{W}_2 + \mathbf{b}_2) \\ &= \mathbf{v}_1 \odot \sigma(\mathbf{v}_2) \end{aligned} \tag{6}$$

The gradient of the GLUs

$$\nabla[\mathbf{v}_1 \odot \sigma(\mathbf{v}_2)] = \nabla\mathbf{v}_1 \odot \sigma(\mathbf{v}_2) + \sigma'(\mathbf{v}_2)\nabla\mathbf{v}_2 \odot \mathbf{v}_1 \tag{7}$$

includes a path $\nabla\mathbf{v}_1 \odot \sigma(\mathbf{v}_2)$ without downscaling, allowing for the gradient flow through layers while retaining the nonlinear capabilities.

2.3. Residual learning

He *et al.* [16] developed a deep residual learning framework via introducing the skip connections, which dramatically alleviate the vanishing gradient problem. Fig. 2 (top) illustrates a 1-D version of the bottleneck residual block in [16]. The bottleneck design decreases the network depth while maintaining the performance. By incorporating time-dilated convolutions and GLUs into the common bottleneck residual block, we introduce a novel residual block shown in Fig. 2 (bottom), where the kernel size in the middle layer is increased to 7 to further expand the receptive fields.

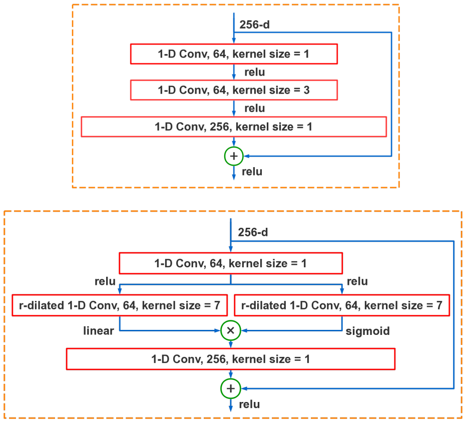
256-d
1-D Conv, 64, kernel size = 1
relu
1-D Conv, 64, kernel size = 3
relu
1-D Conv, 256, kernel size = 1
+
relu
256-d
1-D Conv, 64, kernel size = 1
relu
relu
r-dilated 1-D Conv, 64, kernel size = 7
r-dilated 1-D Conv, 64, kernel size = 7
linear
×
sigmoid
1-D Conv, 256, kernel size = 1
+
relu

Fig. 2. Top: a common bottleneck residual block with 3 convolutional layers. Bottom: our proposed residual block.

2.4. Network architecture

Once we treat supervised speech separation as an utterance-to-utterance mapping, the T-F representation of an utterance is fed into the network at once. In this study, we use 161-dimensional short-time Fourier transform (STFT) magnitude spectra as the network inputs with shape $T \times F$, where T and F denote the numbers of time frames and frequency channels in the STFT magnitude spectra, respectively. Considering the potential imbalance between T and F ($= 161$), it may be better to systematically aggregate the contexts in the frequency direction and the time direction, separately.

Our proposed 62-layer network is constructed as follows. We first stack four frequency-dilated convolutional layers with rectified linear activations [17] to capture the contextual information along the frequency direction. Subsequently, a 1-D convolutional layer with 128 size-1 kernels is employed for dimension reduction. To model temporal dependencies, we construct a bunch of residual blocks (see Fig. 2) to apply time-dilated convolutions. We assign the dilation rates following a sawtooth wave-like fashion [18]: a set of residual blocks are grouped together to form the "rising edge" of the wave which has exponentially increasing dilation rates, and two succeeding groups repeat the same pattern. Once the residual blocks are stacked together, two successive convolutional layers with rectified linear activations and linear activations are used to perform cross-channel pooling and dimension reduction. Finally, an output layer with sigmoid nonlinearity is utilized for mask estimation. Note that we apply zero-padding to the 1-D convolutions but not to the 2-D convolutions. Moreover, a variant of batch normalization [19] is adopted, where the moving averages are used to perform the normalization during both training and inference.

A detailed description of our proposed network architecture is given by Table 1. The input sizes and the output sizes of layers are specified in *featureMaps* × *timeSteps* × *frequencyChannels* format for 2-D convolutions, while the sizes are given in *timeSteps* × *featureMaps* format for 1-D convolutions. The layer hyperparameters are shown in (*kernelSize*, *dilationRate*, *outChannels*) format.

Table 1. Architecture of our proposed 62-layer GRN. Residual blocks are shown in brackets (see also Fig. 2).

layer name	input size	layer hyperparameters	output size
expand_dims	$T \times 161$	-	$1 \times T \times 161$
conv2d_1	$1 \times T \times 161$	$1 \times 3, (1,1), 16$	$16 \times T \times 159$
conv2d_2	$16 \times T \times 159$	$1 \times 3, (1,1), 16$	$16 \times T \times 157$
conv2d_3	$16 \times T \times 157$	$1 \times 3, (1,2), 32$	$32 \times T \times 153$
conv2d_4	$32 \times T \times 153$	$1 \times 3, (1,4), 32$	$32 \times T \times 145$
reshape	$32 \times T \times 145$	-	$T \times 4640$
conv1d_1	$T \times 4640$	$1, 1, 128$	$T \times 128$
conv1d_2	$T \times 64$	$\left.\begin{matrix}\begin{pmatrix}1,1,64\\7,\underline{\mathbf{1}},64\\1,1,256\end{pmatrix}\\\begin{pmatrix}1,1,64\\7,\underline{\mathbf{2}},64\\1,1,256\end{pmatrix}\\\begin{pmatrix}1,1,64\\7,\underline{\mathbf{4}},64\\1,1,256\end{pmatrix}\\\begin{pmatrix}1,1,64\\7,\underline{\mathbf{8}},64\\1,1,256\end{pmatrix}\\\begin{pmatrix}1,1,64\\7,\underline{\mathbf{16}},64\\1,1,256\end{pmatrix}\\\begin{pmatrix}1,1,64\\7,\underline{\mathbf{32}},64\\1,1,256\end{pmatrix}\end{matrix}\right\}\times 3$	$T \times 256$
conv1d_3	$T \times 256$	$1, 1, 256$	$T \times 256$
conv1d_4	$T \times 256$	$1, 1, 128$	$T \times 128$
conv1d_5	$T \times 128$	$1, 1, 161$	$T \times 161$

3. EXPERIMENTS

3.1. Experimental setup

Our experiments are conducted on the WSJ0 SI-84 dataset [20] including 7138 utterances from 83 speakers. Among these speakers, 6 speakers (*i.e.*, 3 males and 3 females) are treated as untrained speakers. Hence, we train the models with the 77 remaining speakers. To investigate noise-independent speech separation, we utilize two challenging noises (babble and cafeteria) from the Auditec CD (available at http://www.auditec.com) for our test sets, and 10 000 noises from a sound effect library (available at https://www.sound-ideas.com) for our training set.

Two test sets are created for each noise using 6 untrained speakers and 6 trained speakers (3 males and 3 females), respectively. Specifically, we use random cuts from a noise to mix with the test utterances at -5 dB and -2 dB. One test set includes 150 mixtures created from 25 × 6 utterances of 6 trained speakers, while the other includes 150 mixtures created from 25 × 6 utterances of 6 untrained speakers. To create a training mixture, we mix a randomly drawn training utterance with a random cut from the 10 000 training noises at an SNR level randomly chosen from {-5, -4, -3, -2, -1, 0} dB. The training set comprises 320 000 mixtures in our experiments and the total duration is about 500 hours. Note that all test utterances are excluded from the training set.

Table 2. STOI and PESQ scores on trained speakers.

metrics	**STOI (in %)**						**PESQ**					
SNR	**-5 dB**			**-2 dB**			**-5 dB**			**-2 dB**		
noises	**Avg.**	**babble**	**cafeteria**	**Avg.**	**babble**	**cafeteria**	**Avg.**	**babble**	**cafeteria**	**Avg.**	**babble**	**cafeteria**
unprocessed	58.0	58.8	57.3	65.9	66.4	65.5	1.57	1.63	1.52	1.74	1.78	1.71
LSTM	75.2	76.4	74.1	82.4	83.2	81.6	2.07	2.05	2.09	2.39	2.37	2.41
GRN	**76.8**	**77.6**	**75.9**	**83.1**	**83.4**	**82.7**	**2.14**	**2.10**	**2.17**	**2.43**	**2.38**	**2.48**

Table 3. STOI and PESQ scores on untrained speakers.

metrics	**STOI (in %)**						**PESQ**					
SNR	**-5 dB**			**-2 dB**			**-5 dB**			**-2 dB**		
noises	**Avg.**	**babble**	**cafeteria**	**Avg.**	**babble**	**cafeteria**	**Avg.**	**babble**	**cafeteria**	**Avg.**	**babble**	**cafeteria**
unprocessed	58.0	58.5	57.5	65.1	65.5	64.7	1.50	1.56	1.44	1.67	1.71	1.63
LSTM	73.1	73.0	73.2	81.0	81.1	80.9	1.96	1.89	2.04	2.30	2.26	2.34
GRN	**75.6**	**75.8**	**75.3**	**82.5**	**82.5**	**82.4**	**2.05**	**1.99**	**2.11**	**2.35**	**2.30**	**2.40**

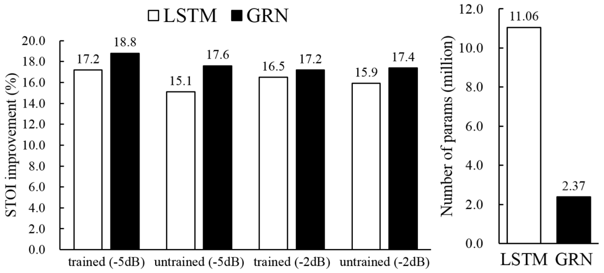

Fig. 3. Comparison of the LSTM and GRN in terms of STOI improvements (left) and computational efficiency (right).

In our experiments, we evaluate the proposed GRN and the LSTM developed in [6]. For both models, the network inputs are normalized to zero mean and unit variance. The PSMs are clipped to between 0 and 1, to fit the range of the sigmoid function. During training, Adam [21] serves as the optimizer to minimize the mean square error (MSE) objective function with a learning rate of 0.001. Both models are trained with a mini-batch size of 16. Within a mini-batch, all samples are zero-padded to have the same number of time steps as the longest sample does. For the LSTM, we use a feature window of 11 frames (5 to the left and 5 to the right) to estimate one frame of the PSM.

3.2. Experimental results

In this study, we use short-time objective intelligibility (STOI) [22] and perceptual evaluation of speech quality (PESQ) [23] as the metrics to evaluate objective speech intelligibility and quality, respectively. The STOI score ranges from 0 to 1, and the PESQ score is between -0.5 and 4.5.

Table 2 and Table 3 list the STOI and PESQ scores of unprocessed and processed signals on trained and untrained speakers, respectively. Boldface numbers highlight the best result in each case. The STOI improvements over the unprocessed signals are shown in Fig. 3 (left). Overall, the proposed GRN consistently outperforms the LSTM in terms of both STOI and PESQ scores. On the trained speakers, the proposed GRN yields around 1.6% STOI improvements compared with the LSTM on average. For example, the STOI score improves by 1.2% on the babble noise at -5 dB compared to the LSTM. As Table 3 shows, the proposed GRN generalizes better to untrained speakers than the LSTM does. In the most challenging scenario, where the utterances from untrained speakers are mixed with the babble noise at -5 dB, the GRN leads to a 2.8% STOI improvement over the LSTM.

Previous CNNs [7] [8] [9] for speech separation yield relatively small improvements compared to DNNs, while the LSTM model significantly outperforms a DNN model with a large number of training speakers [6]. Further performance improvements are achieved by our proposed GRN.

For the utterance-to-utterance mapping, the proposed GRN potentially benefits from its large receptive fields upon the inputs. This allows the GRN to capture long-term dependencies, which are critical to speaker characterization for the sake of speech separation. The LSTM learns temporal dynamics of speech as well, while it insufficiently utilizes the frequency information. The GRN, however, leverages contexts over both the frequency axis and the time axis, which enables the network to model more complex temporal dependencies.

Another advantage of the GRN is its higher computational efficiency due to the use of shared weights in convolutions. Fig. 3 (right) presents the numbers of trainable parameters in the LSTM and the GRN. Even though the GRN is far deeper than the LSTM, it is more economical than the latter with regard to computational cost.

4. CONCLUSION

In this study, we have proposed a gated residual network with dilated convolutions, named GRN, to deal with noise-independent speech separation. The evaluation results indicate that the proposed GRN consistently outperforms a four-layer LSTM model on both trained and untrained speakers. Moreover, we have shown that the GRN is more computationally efficient than the LSTM. We believe that the proposed model lays a sound foundation for investigations towards CNNs for supervised speech separation.

5. REFERENCES

[1] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: an overview,” *arXiv preprint arXiv:1708.07524*, 2017.

[2] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE press, 2006.

[3] Y. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[4] Y. Wang, A. Narayanan, and D. L. Wang, “On training targets for supervised speech separation,” *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[5] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, “Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[6] J. Chen and D. L. Wang, “Long short-term memory for speaker generalization in supervised speech separation,” *The Journal of the Acoustical Society of America*, vol. 141, no. 6, pp. 4705–4714, 2017.

[7] L. Hui, M. Cai, C. Guo, L. He, W.-Q. Zhang, and J. Liu, “Convolutional maxout neural networks for speech separation,” in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2015, pp. 24–27.

[8] S. R. Park and J. Lee, “A fully convolutional neural network for speech enhancement,” *arXiv preprint arXiv:1609.07132*, 2016.

[9] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, “Raw waveform-based speech enhancement by fully convolutional networks,” *arXiv preprint arXiv:1703.02205*, 2017.

[10] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in *International Conference on Learning Representations (ICLR)*, 2016.

[11] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] T. Sercu and V. Goel, “Dense prediction on sequences with time-dilated convolutions for speech recognition,” *NIPS End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[13] M. J. Shensa, “The discrete wavelet transform: wedding the a trous and mallat algorithms,” *IEEE Transactions on signal processing*, vol. 40, no. 10, pp. 2464–2482, 1992.

[14] A. van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, and K. Kavukcuoglu, “Conditional image generation with pixelcnn decoders,” in *Advances in Neural Information Processing Systems*, 2016, pp. 4790–4798.

[15] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” in *Proceedings of the 34th International Conference on Machine Learning*, 2017, vol. 70, pp. 933–941.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, pp. 315–323.

[18] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, “Understanding convolution for semantic segmentation,” *arXiv preprint arXiv:1702.08502*, 2017.

[19] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, 2015, pp. 448–456.

[20] D. B. Paul and J. M. Baker, “The design for the wall street journal-based csr corpus,” in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[21] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] 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 (I-CASSP)*, 2001, vol. 2, pp. 749–752.