Abstract

This work proposes a new learning framework that uses a loss function in the frequency domain to train a convolutional neural network (CNN) in the time domain. At the training time, an extra operation is added after the speech enhancement network to convert the estimated signal in the time domain to the frequency domain. This operation is differentiable and is used to train the system with a loss in the frequency domain. This proposed approach replaces learning in the frequency domain, i.e., short-time Fourier transform (STFT) magnitude estimation, with learning in the original time domain. The proposed method is a spectral mapping approach in which the CNN first generates a time domain signal then computes its STFT that is used for spectral mapping. This way the CNN can exploit the additional domain knowledge about calculating the STFT magnitude from the time domain signal. Experimental results demonstrate that the proposed method substantially outperforms the other methods of speech enhancement. The proposed approach is easy to implement and applicable to related speech processing tasks that require spectral mapping or time-frequency (T-F) masking.

Index Terms: speech enhancement, fully convolutional networks, deep learning, $L_1$ loss, time domain

1. Introduction

Speech enhancement is the task of removing additive noise from a speech signal. It has many applications including robust automatic speech recognition, automatic speaker recognition, mobile speech communication and hearing aids design. Traditional speech enhancement approaches include spectral subtraction [1], Wiener filtering [2], statistical model-based methods [3] and nonnegative matrix factorization [4]. In last few years, supervised methods for speech enhancement using deep neural networks have become state of the art. Among the most popular deep learning methods are deep denoising autoencoders [5], deep neural networks (DNNs) [6, 7], and CNNs [8]. An overview of deep learning based methods for speech separation is given in [9].

Primary methods for supervised speech enhancement use T-F masking or spectral mapping [9]. Both of these approaches generally reconstruct the speech signal in the frequency domain from the frequency domain using the phase of the noisy signal. This means that the learning machine learns a function in the frequency domain but the task of going from the frequency domain to the time domain is not subject to the learning process. In this work, we propose a learning framework in which the objective of the learning machine remains the same but now the process of reconstructing a signal in the time domain is incorporated into the learning process. Integrating the domain knowledge of going from the frequency domain to the time domain or going from the time domain to the frequency domain inside the network can be helpful for the core task of speech enhancement.

A similar approach of incorporating the domain knowledge inside the network is found to be useful in [10], where the authors employ a time-domain loss for T-F masking.

We design a fully convolutional neural network that takes as input the noisy speech signal in the time domain and outputs the enhanced speech signal in the time domain. A simple method to learn this network would be to minimize the mean squared error or the mean absolute error loss between the clean speech signal and the enhanced speech signal [11]. However, in our experiments, we find that using this loss some of the phonetic information in the estimated speech gets distorted because these underlying phones are difficult to distinguish from the background noise. This means that there is no clear discriminability between the background noise and these phones in the speech signal. Also, using a loss function in the time domain does not produce good quality speech. So, it is essential to use a frequency domain loss which has clear discriminability and produces speech with high quality. Motivated by these considerations, we propose to add an extra operation in the model at the training time that converts the estimated speech signal in the time domain to the frequency domain. The process of going from the time domain to the frequency domain is differentiable, and so a loss in the frequency domain can be used to train a network in the time domain.

Furthermore, the proposed framework can be explained as a deep learning based solution to the invalid STFT problem described in [12]. The authors point out that not all combinations of STFT magnitude and STFT phase signal give a valid STFT. The combination of noisy phase and estimated STFT magnitude, in the tasks of spectral mapping or T-F masking, is unlikely a valid STFT. The proposed framework solves this problem by producing a signal in the time domain with a loss in the frequency domain.

This paper is organized as follows: section 2 describes the proposed loss function followed by the description of the model in section 3. Section 4 discusses the invalid STFT problem. Section 5 lists the experimental settings followed by results and discussions in section 6. Finally, we conclude our work in section 7.

2. Frequency domain loss function

Given a speech signal frame in the time domain, it can be converted into the frequency domain by multiplying it with a complex-valued discrete Fourier transform (DFT) matrix as given in equation 1.

$$X = Dx$$

where $X$ is the DFT of the time domain frame or vector $x$. Now, since the vector $x$ is a real signal the relation in the equation 1 can be rewritten as:

$$X = (D_R + jD_I)x = D_R x + jD_I x$$

(2)
where $D_R$ is the real part and $D_I$ is the imaginary part of the complex-valued matrix $D$. This relation can be separated into two different equations with real and imaginary part of the complex-valued vector $X$ as given in the equation 3.

$$
X_R = D_{R}X \\
X_I = D_{I}X
$$

(3)

$X_R$ and $X_I$ from equation 3 can be used to define a loss function in the frequency domain. One such loss can be defined as the sum of the real loss and the imaginary loss as given in the following equation:

$$
L(\hat{X}, X) = \text{Avg}(|\hat{X}_R - X_R| + |\hat{X}_I - X_I|)
$$

(4)

where $\hat{X}$ is the estimated vector and $X$ is the reference vector. $|X|$ is defined as a vector formed by taking the elementwise absolute value of the vector $X$ and $\text{Avg}(X)$ is a function which takes a vector as input and returns the average value of its elements. It is worth mentioning that this loss function has both the magnitude and the phase information because it uses both the real part and the imaginary part separately. However, we find that using both the magnitude and the phase information does not give an as good performance as using only the magnitude information. So, we use the following loss defined using only the magnitudes:

$$
L(\hat{X}, X) = \text{Avg}((|\hat{X}_R| + |\hat{X}_I|) - (|X_R| + |X_I|))
$$

(5)

This loss function can also be described as the mean absolute error loss between the estimated STFT magnitude and the clean STFT magnitude when the magnitude of a complex number is defined using $L_1$ norm. Using $L_2$ norm is also a choice here, but we do not propose it because it gives similar objective scores but introduces an artifact in the enhanced speech. The schematic diagram for computing a frequency domain loss function from a time domain signal is shown in the upper part of figure 1. It should be noted that the matrix $D_R$ and $D_I$ are real matrices, so the network can be trained using backpropagation with real gradients. This means that a real network in the time domain can be trained with a loss function defined in the complex frequency domain. Although using both the real and the imaginary part separately does not give better performance than using only the magnitude, nevertheless it opens a new research direction for combining the real and imaginary part in a better way to utilize the phase information effectively. The enhanced output is first divided into frames and then multiplied by the Hanning window before feeding it to the loss calculation framework.

3. Model architecture

We use an autoencoder based fully convolutional neural network with skip connections [13]. The schematic diagram of the proposed model is shown in the lower part of figure 1. Each convolution layer in the network is followed by parametric ReLU [14] activation function except for the output layer which is followed by Tanh. The encoder comprises nine layers of convolution in which the first layer has a stride of one, and the rest of the eight layers have a stride of 2. The decoder is comprised of deconvolution layers with the number of channels equal to the double of the number of channels in the corresponding symmetric layer in the encoder. The number of channels in the decoder is doubled because of the incoming skip connections from the encoder. The input to the network is a speech frame of size 2048. The dimensions of the outputs from the successive layers of the network are: 2048x1 (input), 2048x64, 1024x64, 512x64, 256x128, 128x128, 64x128, 32x256, 16x256, 8x256, 16x512, 32x512, 64x256, 128x256, 256x256, 512x128, 1024x128, 2048x128, 2048x1 (output).

4. Invalid short-time Fourier transform

In [12], authors explain that not all 2-dimensional complex-valued signals are valid STFT. A 2-dimensional complex-valued signal is a valid STFT if and only if it is obtained by taking the STFT of a time domain signal. In other words, if an STFT $Y$ is not a valid STFT then $Y$ will not be equal to ISTFT(ISTFT($Y$)), where ISTFT means inverse STFT. However, a time domain signal $X$ will always be equal to ISTFT(STFT($X$)). In [12], authors proposed an iterative method which minimizes the distance between the STFT magnitudes by iteratively going back and forth to the time domain from the frequency domain. Going from the time domain to the frequency domain guarantees that the obtained STFT is a valid one.

In the frequency domain speech enhancement, popular approaches are spectral mapping and T-F masking. Both of these methods require using the phase of the noisy speech STFT with the estimated magnitude of STFT to reconstruct the time domain speech signal. Combination of the noisy phase with the estimated magnitude of STFT is unlikely a valid STFT. The proposed framework can be thought of as a supervised way of solving the invalid STFT problem by training a network which produces a speech signal in the time domain but is trained by a loss function which minimizes the distance between the STFT magnitudes.
Table 1: The average performance of noise specific models for five noises and 3 SNR conditions: Mixture (a), DNN (b), AECNN-T (c), AECNN-RI, (d), AECNN-SM (e).

<table>
<thead>
<tr>
<th></th>
<th>SNR</th>
<th>5</th>
<th>0</th>
<th>-5</th>
<th>mean</th>
<th>PESQ</th>
<th>STOI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0</td>
<td>-5</td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>1.41</td>
<td>1.72</td>
<td>2.06</td>
<td>1.73</td>
<td>1.73</td>
<td>56.6</td>
<td>68.1</td>
</tr>
<tr>
<td>b)</td>
<td>2.00</td>
<td>2.39</td>
<td>2.77</td>
<td>2.39</td>
<td>2.39</td>
<td>73.1</td>
<td>82.0</td>
</tr>
<tr>
<td>c)</td>
<td>1.77</td>
<td>2.27</td>
<td>2.61</td>
<td>2.22</td>
<td>2.22</td>
<td>79.5</td>
<td>87.9</td>
</tr>
<tr>
<td>d)</td>
<td>1.90</td>
<td>2.41</td>
<td>2.74</td>
<td>2.35</td>
<td>2.35</td>
<td>79.5</td>
<td>88.1</td>
</tr>
<tr>
<td>e)</td>
<td>2.20</td>
<td>2.65</td>
<td>2.95</td>
<td>2.60</td>
<td>2.60</td>
<td>81.0</td>
<td>88.9</td>
</tr>
</tbody>
</table>

Table 2: The average generalization performance of all the models for 2 unseen noises and 3 SNR conditions; Mixture (a), DNN (b), AECNN-T (c), AECNN-RI, (d), AECNN-SM (e).

<table>
<thead>
<tr>
<th></th>
<th>SNR</th>
<th>5</th>
<th>0</th>
<th>-5</th>
<th>mean</th>
<th>PESQ</th>
<th>STOI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>0</td>
<td>-5</td>
<td>mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a)</td>
<td>1.63</td>
<td>1.72</td>
<td>2.06</td>
<td>1.73</td>
<td>1.73</td>
<td>66.5</td>
<td>76.3</td>
</tr>
<tr>
<td>b)</td>
<td>2.02</td>
<td>2.39</td>
<td>2.77</td>
<td>2.39</td>
<td>2.39</td>
<td>76.4</td>
<td>84.7</td>
</tr>
<tr>
<td>c)</td>
<td>2.06</td>
<td>2.57</td>
<td>2.88</td>
<td>2.50</td>
<td>2.50</td>
<td>83.5</td>
<td>90.6</td>
</tr>
<tr>
<td>d)</td>
<td>2.18</td>
<td>2.68</td>
<td>2.98</td>
<td>2.61</td>
<td>2.61</td>
<td>84.2</td>
<td>90.8</td>
</tr>
<tr>
<td>e)</td>
<td>2.49</td>
<td>2.88</td>
<td>3.12</td>
<td>2.83</td>
<td>2.83</td>
<td>85.1</td>
<td>91.2</td>
</tr>
</tbody>
</table>

5. Experimental settings

First, we evaluate the performance of the proposed framework on the TIMIT dataset [15]. Training and test data are generated in the same manner as in [16]. Performance is evaluated on the SNR conditions of -5 dB, 0 dB and 5 dB in which the 5 dB SNR is an unseen SNR condition. Five noise specific (NS) models are trained and tested on noises; babble, factory, speech-shaped noise (SSN), oproom and engine. A single noise generalized (NG) model is trained using all the above five noises and tested on two unseen noises; factory2 and tank. For the baselines, we train three types of DNNs, with $L_1$ loss, using spectral mapping, ratio masking and spectral magnitude masking [9, 17]. For a given test condition, we pick the best performing DNN to compare with the proposed framework [17].

Next, we evaluate the proposed framework for large-scale training by training a speaker-specific model for a large number of noises. Training utterances are created by mixing 1000 different types of noises with 560 male IEEE utterances. Our data generation for training and testing conditions are same as in [18]. We compare our proposed framework with a five-layer DNN model proposed in [18].

Table 2: The average generalization performance of all the models for 2 unseen noises and 3 SNR conditions; Mixture (a), DNN (b), AECNN-T (c), AECNN-RI, (d), AECNN-SM (e).

We use Tanh activation function at the output of the network, so all the utterances are normalized to the range $[-1, 1]$. Utterances are divided into the frames of size 2048 with a shift of 256. The value of the shift is 256 for all the training and test experiments except for the large-scale training in which case a shift of 1024 is used. Multiple predictions of a sample in an utterance are averaged. The output from the network is divided into the frames of size 512, with a shift of 256, and multiplied by the Hanning window before feeding into the loss calculation framework.

We use Adam optimizer [19] for the training and all the models are trained with a batch size of 256. A dropout of 0.2 is applied at the intervals of 3 layers of convolution. Learning rate is exponentially decayed after every epoch with an initial learning rate set to 0.001.

6. Results and discussions

Performance of the proposed framework is evaluated in terms of short-term objective intelligibility (STOI) [20] and perceptual evaluation of the speech quality (PESQ) [21] scores. We call our speech enhancement model AECNN, standing for autoencoder convolutional neural network. The model is trained using three different loss functions. The used loss functions and corresponding abbreviated names for the models are time loss (AECNN-T), real plus imaginary loss (AECNN-RI), and STFT magnitude loss (AECNN-SM). Time loss is the mean absolute error (MAE) loss in the time domain, real plus imaginary loss is the loss defined in equation 4 and STFT magnitude loss is the loss defined in equation 5.

The average performance of all the five NS models with different loss functions and baseline DNN model is listed in table 1. The AECNN-T model improves the STOI score by 6.4% at -5 dB, 5.4% at 0 dB and 4% at 5db with respect to the baseline DNN. But, the PESQ score for this model is much worse than the baseline. This suggests that the time domain loss for enhancement in the time domain is good for the STOI score.
Table 3: Performance score depicting the effectiveness of learned phase over the noisy phase. Average performance of the noise generalized model is reported for two unseen noises and 3 different SNR conditions.

<table>
<thead>
<tr>
<th>Phase</th>
<th>Mixture</th>
<th>-5 db</th>
<th>0 db</th>
<th>5 db</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noisy</td>
<td>PESQ</td>
<td>Phase</td>
<td>Learned Phase</td>
<td>Clean Phase</td>
</tr>
<tr>
<td>Phase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-5 db</td>
<td>1.63</td>
<td>2.46</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td>0 db</td>
<td>2.00</td>
<td>2.84</td>
<td>2.88</td>
</tr>
<tr>
<td></td>
<td>5 db</td>
<td>2.35</td>
<td>3.11</td>
<td>3.12</td>
</tr>
</tbody>
</table>

The AECNN-RM model introduces the least distortion and is better for the STOI score, on average, by 3.2% at -5 dB, 2.9% at 0 dB and 2% at 5 dB.

The AECNN-SM model improves the STOI score by 0.43 at -5 dB, 0.31 at 0 dB and 0.24 at 5 dB with respect to the baseline DNN. Similarly, it improves the PESQ score by 0.29 at -5 dB, 0.25 at 0 dB and 0.12 at 5 dB when compared to the baseline DNN.

At this point, one can claim that the improved performances using a loss in the frequency domain may not sustain when a large model is trained. To verify this, we trained AECNNs with different sizes and looked at the relative improvement of AECNN-SM compared to AECNN-T. A chart in figure 3 depicts the consistent improvement using networks with the number of parameters equal to 0.4 million, 1.6 million and 6.4 million respectively. We can see that AECNN-SM is consistently and substantially better than AECNN-T model for all the three sized networks.

In the proposed framework, the STFT magnitude loss ignores the phase information which means that the network learns a phase structure which is good for the objective intelligibility of the speech. The learned phase is better for the STOI score, on average, by 3.2% at -5 dB, 2.9% at 0 dB and 2% at 5 dB.

Finally, we evaluate the proposed framework for large-scale training. We compare our model with a model proposed in [18]. The results are given in table 4. The proposed framework is significantly better than the baseline. The STOI improvement is, on average, better by 5.37% on -5 dB which is a difficult and unseen SNR condition. Similarly, it is also significantly better for three other noise conditions as can be seen in table 4.

#### 7. Conclusions

In this work, we proposed a new framework which generates a speech signal in the time domain by minimizing a loss in the frequency domain. The proposed method significantly outperforms the spectral mapping and T-F masking based methods. The proposed approach is easy to implement and applicable for related speech processing tasks that require spectral mapping and T-F masking. This work also opens a new research direction for exploring the proposed framework for the effective use of phase and magnitude information for speech enhancement.

#### 8. Acknowledgements

This research was supported in part by two NIDCD (R01 DC012048 and R02 DCDC015521) grants and the Ohio Supercomputer Center.
9. References

subtraction,” IEEE Transactions on acoustics, speech, and signal 


unsupervised speech enhancement using nonnegative matrix fac-
torization,” IEEE Transactions on Audio, Speech, and Language 

based on deep denoising autoencoders,” in Interspeech, 2013, pp. 
436–440.

speech separation,” IEEE Transactions on Audio, Speech, and 

proach to speech enhancement based on deep neural networks,” 
IEEE/ACM Transactions on Audio, Speech and Language Pro-
cessing (TASLP), vol. 23, no. 1, pp. 7–19, 2015.


2017.

signal reconstruction,” in Acoustics, Speech and Signal Process-
ing (ICASSP), 2015 IEEE International Conference on. IEEE, 
2015, pp. 4390–4394.

speech enhancement by fully convolutional networks,” arXiv 

[12] D. Griffin and J. Lim, “Signal estimation from modified short-
time fourier transform,” IEEE Transactions on Acoustics, Speech, 

enhancement generative adversarial network,” in Proc. In-
http://dx.doi.org/10.21437/Interspeech.2017-1428

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: 
Surpassing human-level performance on imagenet classification,” 
in Proceedings of the IEEE international conference on computer 
vision, 2015, pp. 1026–1034.

Pallett, “Darpa timit acoustic-phonetic continous speech corpus 
.cd-rom. nist speech disc 1-1.1,” NASA STI/Recon technical report 

supervised speech separation,” IEEE/ACM Transactions on Au-
dio, Speech and Language Processing (TASLP), vol. 22, no. 12, 

tions for speech enhancement,” in Acoustics, Speech and Signal 
Processing (ICASSP), 2018 IEEE International Conference on. 

[18] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, “Large-scale training to increase speech intelligibility for hearing-
impaired listeners in novel noises,” The Journal of the Acous-


algorithm for intelligibility prediction of time–frequency weighted 
noisy speech,” IEEE Transactions on Audio, Speech, and Lan-

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hek-
stra, “Perceptual evaluation of speech quality (pesq)-a new 
method for speech quality assessment of telephone networks and 
codes,” in Acoustics, Speech, and Signal Processing, 2001. Pro-
cedings.(ICASSP’01). 2001 IEEE International Conference on, 