SPEECH ENHANCEMENT BASED ON DEEP NEURAL NETWORKS WITH SKIP CONNECTIONS

Ming Tu[1*], *Xianxian Zhang*[2]

[1]Speech and Hearing Science Department, Arizona State University
[2]LG San Jose Lab, LG Electronics

ABSTRACT

Speech enhancement under noise condition has always been an intriguing research topic. In this paper, we propose a new Deep Neural Networks (DNNs) based architecture for speech enhancement. In contrast to standard feed forward network architecture, we add skip connections between network inputs and outputs to indirectly force the DNNs to learn ideal ratio mask. We also show that the performance can be further improved by stacking multiple such network blocks. Experimental results demonstrate that our proposed architecture can achieve considerably better performance than the existing method in terms of three commonly used objective measurements under two real noise conditions.

Index Terms— Speech enhancement, noise reduction, deep neural networks, skip connections

1. INTRODUCTION

Speech enhancement under noisy conditions has always been an intriguing but challenging task for robust speech processing, such as robust speech and speaker recognition [1]. During the past several decades, many pioneer researches have been done in this area. Boll et al. [2] proposed spectral substraction, a stand-alone noise suppression algorithm for reducing effects of acoustical added noise in speech. Ephraim et al. [3] devised a system by utilizing a minimum mean square error (MMSE) spectra estimator. However, these traditional methods often suffer from different problems. For example, spectral substraction could introduce the notorious musical noise. MMSE based noise spectra estimator could reduce the effect of musical noise but still introduces distortion. In addition, these traditional methods can not guarantee good performance in non-stationary noise conditions. Decomposition based speech enhancement is another research direction that got attention recently. This method decomposes noisy speech spectra into atomic parts of speech and noise. Wilson et al. [5] studied a technique for denoising speech using non-negative matrix factorization. Chen et al. [6] presented a speech enhancement system based on decomposing the spectrogram into sparse activation of a dictionary of target speech templates, and a low-rank background model. However, these methods are still limited by model capacity considering their linear relationship assumption between noisy and clean speech spectra.

Benefiting from its ability to learn complex non-linear mapping function and leverage large amount of training data, Deep Neural Networks (DNNs) have achieved great success in speech applications such as Automatic Speech Recognition (ASR), speaker recognition, audio environmental sensing, etc. Inspired by this, deep neural networks have also been applied to speech enhancement and considerable improvement is reported over shallow models. In [7], the authors proposed to use DNNs based denoising autoencoder to learn the transformation between noisy-clean speech pairs. In that study, Mel-frequency power spectrum was employed as representation for both noisy and clean speech signal. In [8], a weighted denoising autoencoder was introduced to estimate the clean speech power spectrum and Wiener filter was used as back-end processing to enhance the noisy power spectrum given estimated Signal-to-Noise Ratio (SNR). Later, a study in [9] came up with several techniques to further improve the DNNs based speech enhancement system and achieved some improvements over the baseline system. In [10], the authors applied DNNs based speech enhancement scheme to improve speech intelligibility for hearing-impaired listeners in noisy and reverberant environment. Through both objective measurements and subjective listening test, the method was proved to be effective.

In this paper we propose a new network architecture for DNNs based speech enhancement. In contrast to [7][8][9][10] that use standard feed forward network structure to learn the non-linear mapping function from input and output, the proposed method adds skip connection with identity weight matrix between network input and output. New network output will be the addition between input and the output of last hidden layer with linear activation function. By doing this, we indirectly force the network between input and output to learn log compressed ideal ratio mask when both input noisy speech feature and output clean speech feature are log compressed. We also come up with a method to tackle the problem of mismatch between input and output dimensions. Also, we

[*]This work was done during the author's summer internship at LG San Jose Lab.

can regard these networks with such skip connection between input and output as building blocks for DNNs. The network by stacking such building blocks is further applied to speech enhancement with the same inputs and outputs. Through experiments on enhancement of speech signal corrupted by two types of self-collected noise in real scenario, we show that our proposed architecture is able to considerably improve the system performance compared with DNNs based baseline speech enhancement method in terms of three different measurements for speech enhancement. We also show that further improved can be achieved with the proposed stacked network.

Relation to prior work: First, DNNs with skip connections ("shortcuts") have recently achieved great success in image recognition by making network very deep [11, 12]. Another research employed encoder-decoder networks with skip connections for image denoising [13]. In our paper, similar idea is applied to speech enhancement. However, our network fits a mapping from noisy speech feature to log compressed ideal ratio mask instead of residual learning in [12, 13]. Second, [14] investigated a method that directly uses ideal ratio mask as network output and got better results than using clean speech feature as target in supervised speech separation application. [10] applied the same idea for speech enhancement and also reported good results. Though similar to these two studies, our method works in an end-to-end way and still uses noisy speech feature as input and clean speech feature as output. No other post-processing is needed in order to recover clean speech signal as in [14, 10]. Furthermore, our stacked network is never explored in speech enhancement or separation applications and proved to be better than network without stacking.

The paper is organized as follows. The next section will introduce our proposed network architecture for speech enhancement. In section 3, evaluation protocol and experimental results will be demonstrated. Section 4 discusses our methods and section 5 concludes our work.

2. PROPOSED NETWORK ARCHITECTURE

2.1. Regression based speech enhancement

The basic framework of regression based speech enhancement using DNNs can be demonstrated by the diagram in Fig. 1. The assumption is that the relationship between noisy speech features and clean speech features is highly nonlinear and can be estimated by DNNs with nonlinear activation functions. The goal of DNNs based supervised learning is to learn the mapping from noisy speech features to clean speech features. Then, for an input noisy sentence, clean speech features are estimated frame by frame from the output of trained DNNs. Speech features are able to be converted back to time domain signal. Magnitude (or power) spectrogram [9] or Mel-frequency spectra [8] are two choices as input to DNNs. Usually, context frames are stacked as input to achieve better result. Techniques like network pre-training, regularization have also been applied to make the system better. In following sections, we represent this network architecture as "DNN".

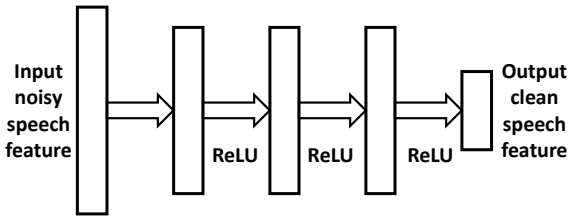

Fig. 1. Diagram of regression based speech enhancement framework with three hidden layers. Activation function is "ReLU" for illustration.

2.2. Proposed network architecture

We suppose $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{N}$ are the magnitudes of noisy speech spectrogram, clean speech spectrogram and noise spectrogram respectively. The approximation $\mathbf{X} \approx \mathbf{S} + \mathbf{N}$ is a commonly used assumption in audio source separation and speech enhancement, especially those methods based matrix decomposition [15]. If we consider magnitude of Mel-frequency spectra, this approximation still holds since

$$\mathbf{XM} \approx \mathbf{SM} + \mathbf{NM}, \tag{1}$$

where $\mathbf{M}$ is the matrix to transform magnitude of spectrogram to magnitude of Mel-frequency spectra. Though there is information loss during this transformation, the impact on speech quality is not obvious when converting back from Mel-frequency spectra to spectrogram based on our observation. Thus, in this work we use magnitude of Mel-frequency spectra as network input and output speech feature as in [8] to avoid large input dimension when stacking context frames.

Different from the standard feed-forward neural network used in existing DNNs based speech enhancement, we first add a skip connection from network input to network output. The actual network output is the addition of input and output of last layer with linear activation function. Note that when there are multiple frames stacked together as input, we only connect current frame to the output. We show the diagram of this architecture in Fig. 2 and we call this network "sDNN1". When we use log compressed Mel-frequency spectra as network input, $\mathbf{Y}$ in Fig. 2 can be formulated as

$$\mathbf{Y} = \log(\mathbf{SM}) - \log(\mathbf{XM}) = \log(\frac{\mathbf{SM}}{\mathbf{XM}}), \tag{2}$$

which can be interpreted as the negative residual in log domain or log compressed ideal ratio mask (IRM) of magnitude Mel-frequency spectra. That means instead of learning the mapping from log compressed noisy speech feature to log compressed clean speech feature, proposed network architecture learns the mapping to negative residual in log domain,

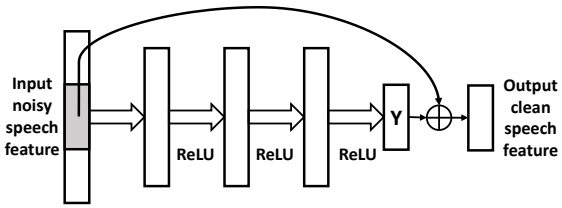

Fig. 2. Diagram of "sDNN1" with skip connection from input to output. The grey block in input is the current frame corresponding to output.

which is also the log compressed IRM. The log domain residual learning interpretation conforms with the idea of residual learning in [12]. IRM has been proved to be a better target for DNNs based speech enhancement in [14, 10]. However, they still need a further step to calculate clean speech feature based on estimated IRM. Our method do not need to change the network output explicitly and works in an end-to-end way which directly transforms noisy speech feature to clean speech feature without introducing extra network parameters. Also the numerical stability problem as mentioned in [14] is also relieved with log compressed IRM as learning target.

We can further stack multiple these networks to form a stacked network as in [12], which uses multiple "sDNN1"s as building blocks and can be represented by "sDNN2". We believe this structure could learn better mapping from input to $\mathbf{Y}$ than single block architecture. In this study, we want to investigate the performance difference of "sDNN1" and "sDNN2" using similar number of parameters. The simple idea is to add skip connections to all layers instead of only connecting input to output. However, there will be dimension mismatch problem for the addition operation when input dimension and number of nodes per layer are different. To solve this, each layer is replaced with a new network block. The new block has two layers and the number of nodes in second layer is the same with input dimension. We can adjust the number of nodes in first layer to match the number of parameters of original layer. The diagram of "sDNN2" is shown in Fig 3, where each single layer is replaced with the two-layer structure below dash lines.

We use Rectified Linear Unit(ReLU) as activation function. To train the network, Mean Square Error (MSE) cost function is employed as in previous work. All the network training is done using TensorFlow with Adam optimizer [16].

3. EVALUATION RESULTS AND ANALYSIS

To evaluate our proposed network architecture, we designed experiments to do speech enhancement on noisy speech simulations, which were generated by adding noise to clean speech. We used TIMIT data set [17] as clean speech. We used all the speech utterances in "train" directory for training and randomly picked 100 utterances from "test" directory for

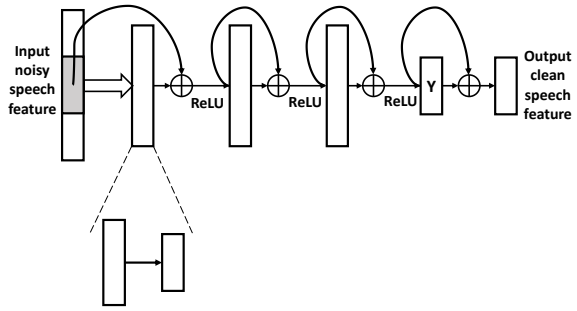

Fig. 3. Diagram of "sDNN2" with skip connections in all layers. The grey block in input is the current frame corresponding to output.

evaluation. Noise data were collected by ourselves with a omnidirectional microphone in real scenario. In the experiments, two types of noise were used. One was recorded in a busy shopping mall (1.5 hours) and the other in an indoor café (0.5 hour). Both of these two kinds of noise are very challenging for speech enhancement task, especially the second one that contains both background music, people talking, sound of preparing coffee and distant traffic noise from street. A noise segment with same length as clean speech utterance was randomly picked and mixed with clean speech at 0dB and 5dB SNR.

We employed magnitude of Mel-frequency spectra as speech features. Time domain signal was first pre-emphasized and then converted to frames with 25ms window length and 10ms window shift. A Mel-frequency filter bank with 40 filters was used to integrate the magnitude of Discrete Fourier Transform (DFT) of each frame into a 40 dimensional vector. The transformation matrix $\mathbf{M}$ from DFT to Mel-frequency feature was saved for future use to recover time domain signal. We also consider past 3 frames and 2 future frames, resulting in input dimension of $6 \times 40 = 240$. There is no feature stacking in output. To normalize speech features, we employed two different strategies. Before training, we can simply normalize inputs and outputs to normal distributions for "DNN". However, for our proposed architectures, we want to keep the relationships in equation 2. To solve this concern, we only do normalization on current frame in input and use the same means and standard deviations to normalize the remaining frames in input and output. Then, equation 2 changes to

$$\mathbf{Y} = \log(\frac{\mathbf{SM} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}) - \log(\frac{\mathbf{XM} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}) = \log(\frac{\mathbf{SM} - \boldsymbol{\mu}}{\mathbf{XM} - \boldsymbol{\mu}}). \tag{3}$$

$\mathbf{Y}$ can still be interpreted as normalized negative residual in log domain or centered and log compressed IRM. We call this normalization method partial normalization (PN). For fair comparison, we also applied PN to "DNN" and represented this method as "DNNpn".

Table 1. Speech enhancement results on noisy speech corrupted by shopping mall noise.

		Noisy	DNN	DNNpn	sDNN1	sDNN2
0dB	**PESQ**	1.865	2.284	2.297	2.337	**2.370**
	STOI	0.696	0.787	0.789	**0.796**	0.794
	segSNR	-2.789	1.427	1.483	**2.057**	1.734
5dB	**PESQ**	2.175	2.590	2.598	2.644	**2.671**
	STOI	0.791	0.851	0.857	**0.864**	0.860
	segSNR	0.515	3.332	3.502	**4.261**	3.879

Table 2. Speech enhancement results on noisy speech corrupted by indoor café noise.

		Noisy	DNN	DNNpn	sDNN1	sDNN2
0dB	**PESQ**	1.866	2.202	2.199	2.262	**2.289**
	STOI	0.691	0.770	0.768	**0.783**	0.781
	segSNR	-2.732	1.177	0.976	**1.639**	1.452
5dB	**PESQ**	2.187	2.608	2.583	2.609	**2.650**
	STOI	0.788	0.845	0.846	**0.854**	0.853
	segSNR	0.473	3.036	2.955	**3.984**	3.744

We compared three network architectures: baseline system as shown in Fig. 1 (represented as "DNN" and "DNNpn"), "sDNN1" in Fig. 2 and "sDNN2" in Fig. 3. For "DNN" and "sDNN1", we had 3 hidden layers with 256 nodes per layer. For "sDNN2", because of the dimension mismatch problem, we set the first layer of the two-layer building block in Fig. 3 to have 220 nodes and second layer to have 40 nodes. By doing this, the number of parameters almost does not change. And also, the addition operation operates at the end of each layer. The output clean speech feature can be converted back to time domain signal through inverse operations. Additionally, the learning rate of Adam was set to 0.001 and number of training epoch was 50. Only the model with best performance on evaluation data was saved.

To evaluate the performance of different network architectures for speech enhancement, three commonly used measurements were employed: Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI) and segmental SNR. These indexes were calculated from enhanced speech and reference clean speech.

In table 1 and table 2, we show the results of four different methods that enhance noisy speech corrupted by two types of noise at 0dB and 5dB SNR. "segSNR" in the tables means segmental SNR. Bold numbers are the best among different methods in corresponding rows. From the results, we can find that both "sDNN1" and "sDNN2" methods can provide better results in terms of three measurements than baseline "DNN" method. Partial normalization ("DNNpn" in results) gives similar performance compared to standard normalization ("DNN"). It can also be found that "sDNN2" method has higher PESQ values, almost the same STOI and a little worse segSNR in comparison with "sDNN1". We tend to regard "sDNN2" as the better method since PESQ and STOI are more convincible speech enhancement measurements than just looking at the amount of removed noise. Generally, the performance on indoor café noise is worse than shopping mall noise which conforms with our previous description that indoor café noise is more challenging than shopping mall noise.

4. DISCUSSION

We have illustrated in section 2 that our proposed network structure actually learns the negative residual in log domain or log compressed IRM. In fact, we tried to make the network directly learn the residual (as the residual learning in [12]) by using speech features without log compression. But the improvement compared to baseline DNN method was not obvious, especially in low SNR case. We think the reason is that in low SNR case learning the mapping from noisy speech features to noise features is even harder since noise in this case could be more complex than speech. Thus, we turned to log compressed speech features. This incorporates the information of noisy signal into the learning target while also partly solves the numerical stability problem [14] with log compression. The stacked network used in this work may also have similar property as residual learning [12], i.e., effective learning even the network is very deep. The investigation of this property could be our future work.

5. CONCLUSION

In this paper, we propose to use DNNs with skip connection for speech enhancement. By adding skip connection between log compressed network input and output to DNN based speech enhancement, we force the network to learn mapping from noisy speech features to negative residual in log domain or log compressed IRM. Also, we stack this network architecture to further improve its learning ability. Through experiments on challenging speech enhancement tasks, we show that our proposed methods surpass baseline DNNs based speech enhancement in terms of three commonly used measurements and our stacked network can provide further improvement. In future, we will investigate to use this architecture as front-end module for different noise robust speech systems.

6. REFERENCES

[1] Tuomas Virtanen, Rita Singh, and Bhiksha Raj, *Techniques for noise robustness in automatic speech recognition*, John Wiley & Sons, 2012.

[2] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[4] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[5] Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran, "Speech denoising using nonnegative matrix factorization with priors.," in *ICASSP*, 2008, pp. 4029–4032.

[6] Zhuo Chen and Daniel PW Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[7] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, pp. 436–440.

[8] Bingyin Xia and Changchun Bao, "Speech enhancement with weighted denoising auto-encoder.," in *INTERSPEECH*, 2013, pp. 3444–3448.

[9] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[10] Yan Zhao, DeLiang Wang, Ivo Merks, and Tao Zhang, "Dnn-based enhancement of noisy and reverberant speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6525–6529.

[11] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Training very deep networks," in *Advances in neural information processing systems*, 2015, pp. 2377–2385.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang, "Image denoising using very deep fully convolutional encoder-decoder networks with symmetric skip connections," *arXiv preprint arXiv:1603.09056*, 2016.

[14] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] Tuomas Virtanen, Jort Florent Gemmeke, Bhiksha Raj, and Paris Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[16] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous systems, 2015," *Software available from tensorflow. org*, vol. 1, 2015.

[17] John S Garofolo, Lori F Lamel, William M Fisher, Jonathon G Fiscus, and David S Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.