Grenze International Journal of Computer Theory and Engineering , June 2017

Speech Enhancement using Deep Neural Network

Shreegowri A.J [1] and D.J Ravi [2]
[1]Vidyavardhaka College of Engineering / ECE Department, Mysuru, India
Email: srigowri0503@gmail.com
[2]Professor, Vidyavardhaka College of Engineering / ECE Department, Mysuru, India
Email: ravidj@vvce.ac.in

***Abstract*—In contrast to the minimum mean square error (MMSE)-based noise cancelation techniques, we propose a method to enhance speech by means of finding a mapping function between noisy signal and clean speech signals based on deep neural networks (DNNs). In order to handle a wide verity range of additive noises in real-world scenarios, a large number of training set that contains many possible combinations of speech and noise types, is first designed. The DNN architecture is then employed as a nonlinear regression function to ensure a powerful modeling capability. To further improve the DNN-based speech enhancement system we proposed a technique called global variance equalization to remove the over-smoothing problem of the regression model to improve the generalization capability of DNNs. To further improve the quality of enhanced speech and generalization capability of DNNs. First, equalization between the global variance (GV) of the enhanced features and the reference clean speech features is proposed to alleviate the over- smoothing issue in DNN- based speech enhancement system.**

***Index Terms*— Deep neural networks (DNNs), dropout, global variance equalization, noise aware training, noise reduction, non-stationary noise, speech enhancement.**

I. INTRODUCTION

In recent years, single-channel speech enhancement has fetching a research attention because of the arising challenges in many important real-world applications, including mobile speech communication, communication hearing aids, security monitoring, intelligence robust speech /speaker/language recognition, etc... The goal of speech enhancement is to improve the intelligibility and quality of a noisy speech signal.
We have proposed a DNN based speech enhancement model via training a deep and wide neural network using a large collection of noisy speech data. It was found that the annoying artifact called “musical noise” could be greatly reduced with the DNN-based algorithm and the enhanced speech also exhibits an improved speech quality both in terms of objective and subjective measures.
In this work we extend the DNN-based speech enhancement model to handle unfavorable conditions and non-stationary noise types in real-world circumstances. In classical speech enhancement techniques, the noise estimate is usually amended by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors, which are adjusted based on the estimated speech presence probability in individual frequency bins. Regardless, its noise tracking capacity is limited for highly non-stationary noise cases, and it tends to distort the speech component in mixed signals if it is tuned for better noise

Grenze ID: 01.GIJCTE.3.2.5

reduction/cancelation. In this model, the acoustic context information, including the full frequency band and context frame expanding, is well utilized to obtain the enhanced speech with reduced discontinuity.
Numerous speech enhancement methods were developed over the past several years. Spectral subtraction method for speech enhancement is proposed in 2010[1],[2]; in these method it subtracts an estimate of the short-term noise spectrum to produce an estimated spectrum of the clean speech. In 2013[3],[4], the iterative wiener filtering was presented using an all-pole model. A common problem usually encountered in these conventional methods is that the resulting enhanced speech often suffers from an annoying artifact called "musical noise".
Another notable work was the minimum mean -square error (MMSE) estimator introduced by Ephraim and Malah in 2014[5]; their MMSE log-spectral amplitude estimator could result in much lower residual noise without further affecting the speech quality. An optimally-modified log-spectral amplitude (OM -LSA) speech estimator and a minima controlled recursive averaging (MCRA) noise estimation approach were also presented in 2015[6]. Although these traditional MMSE- based methods are able to yield lower musical noise, a trade-off in reducing speech distortion and residual noise needs to be made due to the statistical properties of the interactions between speech and noise signals. Most of these methods are based on either the additive nature of the background noise, or the statistical properties of the speech and noise signals. However they often fail to track non-stationary noise for real-world scenarios in unseen acoustic conditions. A breakthrough for training deep architectures came in 2006 when Hinton *et al.* [7], [8] proposed a greedy layer-wise unsupervised learning algorithm. Each layer is pre-trained without supervision to learn a high level representation of its input (or the output of its previous layer). For the regression task, deep learning has been used in several speech synthesis tasks [9], [10]. In [11], [12], stacked de-noising auto-encoders (SDAs), as one type of the deep models, were adopted to model the relationship between clean and noisy features, and they only explored its performance on a matching test set. Deep recurrent neural networks (DRNNs) were also adopted in the feature enhancement for robust speech recognition [13], [14].

II. Methodology/Procedure

A block diagram of the proposed speech enhancement model is shown in Figure. 1. A DNN is adopted as the mapping function from noisy to clean speech features. Our system is demonstrated in 2 stages. In the training stage, a DNN- model was trained using the log-power spectral features from pairs of noisy and clean speech data. The log-power spectral feature is adopted since it is thought to offer perceptually relevant parameters. Hence, short-time Fourier analysis is first applied to the input signal, computing the discrete Fourier transform (DFT) of each overlapping windowed frame. Then the log-power spectra are calculated.
During learning (testing stage), a DNN is used to learn the mapping function; no assumptions are made about the relationship of noisy speech with clean speech signals. It can automatically learn the complicated relationship between the noise and clean speech to separate speech from the noisy signals gives the adequate training samples. The DNN can apprehension the acoustic context information along the time axis (using multiple frames of noisy speech as input) and along the frequency axis (using full-band spectrum information) by concatenating them into a long input feature vector for DNN learning while the independence assumption among different dimensions was a common practice in the Gaussian mixture model to reduce computation complexity.
One of the residual error problems, namely over-smoothing, causes a muffling effect on the estimated clean speech signal when compared with reference clean speech. Equalization between the global variance of the estimated and reference clean speech features is proposed to attenuate this problem. Global variance equalization here can be considered as a simple type of histogram equalization (HEQ), which plays a meager role in density matching. It is demonstrated that the use of global variance information could significantly enhance the subjective score in a voice conversion task and increase the generalization capability of DNNs.
In the training stage the DNN is trained with the clean/noisy speech samples. The features of the clean/noisy speech samples are extracted by framing the samples and taking the STFT/DFT of the framed samples it gives the time, frequency and phase information which is used for training the DNN. In testing stage the noisy speech signal is applied for feature extraction block. In feature extraction block the noisy speech signal is first framed then STFT/DFT is applied to each frames then the time, frequency and phase information is given to DNN enhancement block. In DNN enhancement block the training and testing data are compared and eliminate the noise present in the speech signal. To further improve the enhanced signal we use the post

processing step in this we give the enhanced speech signal to the global variance equalization to alleviate the over smoothing problem of the DNNs. To reconstruct the speech signal we use overlap and add method.

Training stage

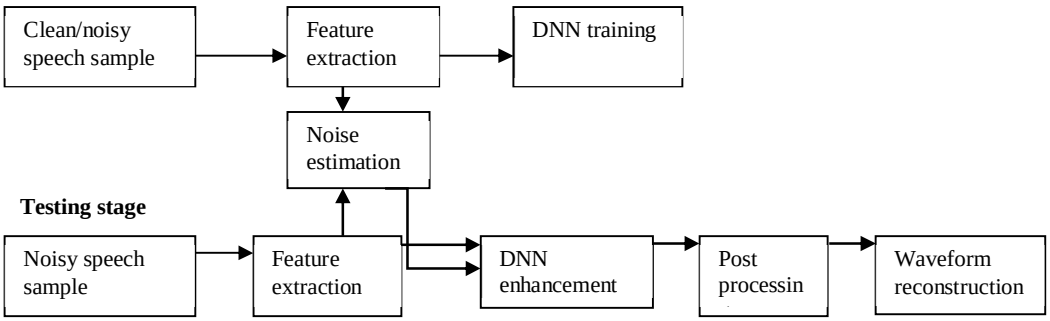


Figure 1 A block diagram of the proposed DNN-based speech enhancement system

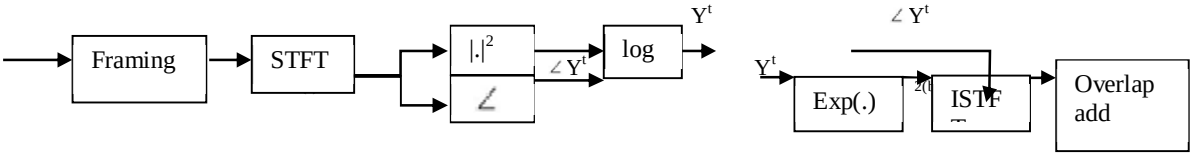


Figure 2 (a) Block of feature extraction, (b) Block of waveform reconstruction

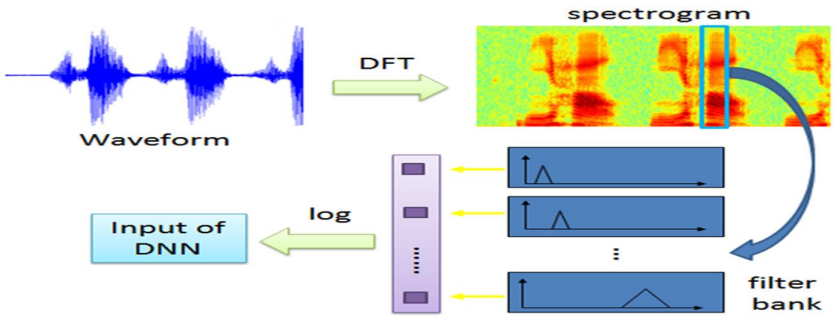
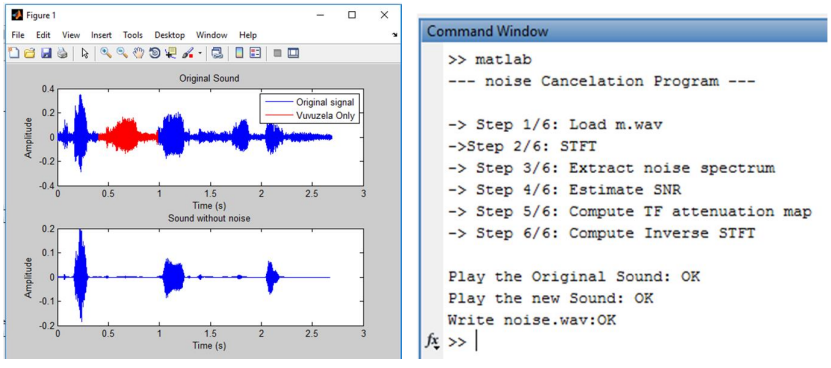


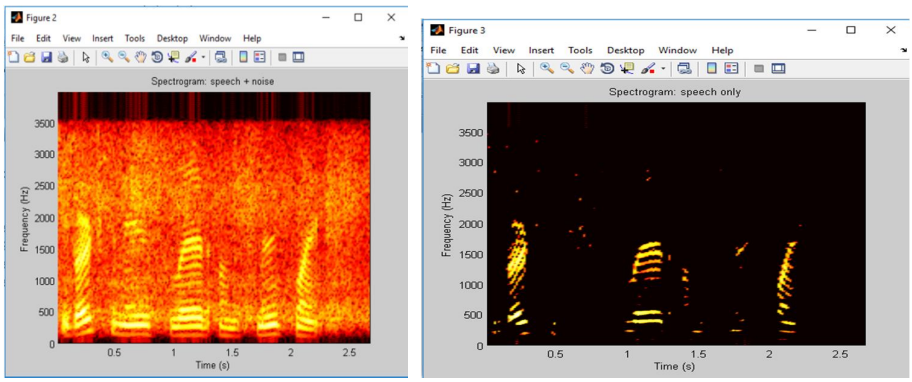
Figure 3 Input for the DNN system

III. Experimental Results



(a) (b)

Figure 4(a) Input and output speech signal and their spectrum, (b) Steps executed in our program



(a) (b)

Figure 5(a) Input speech signal spectrum, (b) Output clean speech

IV. Discussion

Experimental results demonstrate that the proposed framework can achieve significant improvements in both objective and subjective measures over the conventional minimum mean square error (MMSE) based technique and it is also interesting to observe that the proposed DNN approach can well suppress highly non-stationary noise, which is tough to handle in general. Furthermore, the resulting DNN model, is trained with artificial synthesized data, it is also effective in dealing with noisy speech data recorded in real-world scenarios without the generation of the annoying artifact called "musical noise" commonly observed in conventional enhancement methods.

A. Advantages

- It was found that the application of more acoustic context information improves the system performance and makes the enhanced speech less discontinuous.
- Multi-condition training with many kinds of noise types can achieve a good generalization capability to unseen noise environments.
- The complicated relationship between noisy and clean speech could be automatically learnt.
- The deep architecture could well fit the non-linear relationship for regression function approximation.
- The highly non-stationary noise could be well suppressed in the off-line learning framework.
- Nearly no Gaussian or independent assumptions.
- Nearly no empirical thresholds to avoid the non-linear distortion in DNN-based speech enhancement.

V. Future Work

Two strategies are also proposed to further improve the generalization capability of DNNs. The first technique, called dropout, is a recently proposed strategy for training neural networks on data sets where over-fitting may be a concern. While this technique was not designed for noise reduction, it was demonstrated to be useful for noise robust speech recognition and we successfully apply it to a DNN as a regression model to produce a network that has good generalization ability to variability's in the input. Finally, noise aware training (NAT) is adopted to improve performance. Finally, a dynamic noise adaptation scheme will also be investigated for the purpose of improving tracking of non-stationary noises.

VI. Conclusion

In this work, a DNN-based model for speech enhancement is proposed. Among the various DNN configurations, a large number of training set is crucial to learn the rich structure of the mapping function between noisy and clean speech features. It was found that application of more acoustic context information improves system performance and makes the enhanced speech less discontinuous. Moreover, multi-condition training with many kinds of noise types can achieve a good generalization capability to unseen noise

environments. By doing so, the proposed DNN framework is also powerful to cope with non-stationary noises in real-world environments.

An over-smoothing problem in speech quality was found in the MMSE-optimized DNNs and one proposed post-processing technique, called global variance equalization, was effective in brightening the formant spectra of the enhanced speech signals. Two improved training techniques were further adopted to reduce the residual noise and increase the performance. Compared with the Log-MMSE method, the significant improvements were achieved across different unseen noise conditions. Another interesting observation was that the proposed DNN-based speech enhancement system is quite effective for dealing with real-world noisy speech in different languages and across different recording conditions not observed during DNN training.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice. Boca Raton*, FL, USA: CRC, 2013.

[2] J. Benesty, S. Makino, and J. D. Chen, *Speech Enhancement*. New York, NY, USA: Springer, 2014.

[3] S. Boll, "*Suppression of acoustic noise in speech using spectral subtraction*," IEEE Trans. Acoustic, Speech, Signal Process., vol. ASSP-27, no. 2, pp. 113–120, Apr. 2013.

[4] J. S. Lim and A. V. Oppenheim, *"Enhancement and bandwidth compression of noisy speech," Proc. IEEE, vol. 67, no. 12, pp. 1586–1604, Dec. 2015*.

[5] P. Scalart, "*Speech enhancement based on a priori signal to noise estimation," in Proc.* ICASSP, pp. 629–632.

[6] Y. Ephraim and D. Malah, *"Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," IEEE Trans*. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 2013.

[7] Y. EphraimandD.Malah, *"Speech enhancement using minimum mean square log spectral amplitude estimator,"* IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-33, no. 2, pp. 443–445, Apr. 2012.

[8] I. Cohen and B. Berdugo, *"Speech enhancement for non-stationary noise environments," Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2011.

[9] Cohen, *"Noise spectrum estimation in adverse environments: Im-proved minima controlled recursive averaging,"* IEEE Trans. Speech Audio Process., vol. 11, no.5, pp. 466–475, Nov.2013.

[10] O. Cappe, *"Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 2012.

[11] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, *"Nonlinear speech enhancement: An overview," New York*, NY, USA: Springer, 2013, pp. 217–248.

[12] S. I. Tamura, *"An analysis of a noise reduction neural network,"* in Proc. ICASSP, 1989, pp. 2010.