A Novel Training Strategy Using Dynamic Data Generation for Deep Neural Network Based Speech Enhancement

Mao-Kui He, Jun Du, Zi-Rui Wang, Lei Sun
* University of Science and Technology of China, Hefei, Anhui, China
E-mail: hmk1754@mail.ustc.edu.cn, jundu@ustc.edu.cn, cs211@mail.ustc.edu.cn, sunlei17@mail.ustc.edu.cn

***Abstract*—In this paper, a new training strategy is proposed to address the key issue in deep neural network (DNN) based speech enhancement: how to effectively utilize the limited data with a growing awareness of the necessity to increase training data in the deep learning era. Traditionally, a fixed training set consisting of a large amount of paired utterances, i.e., clean speech and corresponding noisy speech, must be prepared in advance. However, it seems inevitable to enlarge noisy speech in the training stage for making model adaptive to various noise environments. Besides, involving more training data leads to longer training time as the fixed training set should be repeated for multiple epochs. In this study, we propose a novel training method via dynamic data generation. The key idea is the synthetic phase of noisy speech data is conducted on the fly from utterance level to the batch level. Three advantages are gained from this new training method. First, by dynamic generation of training data batch, it is not necessary to prepare and store the fixed training set as in the conventional training method. Second, with the same training time as in the conventional method, more abundant noisy data are actually fed into DNN model. Finally, different evaluation measures, including PESQ, STOI, LSD, and SegSNR, can be consistently improved for the unseen noise types, demonstrating the better generalization capability of the proposed training strategy.**

I. Introduction

As we know, background noises can greatly hurt the performance of speech applications, such as automatic speech recognition (ASR) [1], speaker diarization [2], mobile communication and so on. Therefore, speech enhancement [3] becomes increasingly important as a preprocessing method. Many techniques have been proposed over the past several decades, such as spectral subtraction [4], Wiener filtering [5] [6] and so on. However, those traditional methods are mostly based on some artificial assumptions about speech and noise, which limit the overall performance and generalization ability.

Recently, deep neural network (DNN), as a powerful model, has shown great potential in speech enhancement task. With the emerging of deep learning, a new estimated way is proposed [7] [8], i.e., directly defining a mapping function from noisy to clean speech by using DNN-based regressing model. Benefiting from the powerful non-linear modelling capacity of neural network and big data, the new approach has yielded a remarkable gain over the previous methods no matter for unseen or highly non-stationary noises. Moreover, experiments have demonstrated that the new method also shows great capability and stability under highly non-stationary noisy conditions. Besides, various neural network architectures, including recurrent neural network [9] [10], denoising auto-encoder (DAE) [11] [12] and generative stochastic network (GSN) [13] were explored. A signal-to-noise-ratio (SNR) based progressive learning approach [14] was proposed to improve the performance at low SNR situations.

One key problem for these methods is that the extensive training data are needed to optimize the massive parameters of DNN so that the models can tackle diversified speaker and noise variations. In speech enhancement, paired training utterances must be largely synthesize in advance, i.e., clean speech and corresponding noisy speech. Considering the continuous range of SNR and exponential combinations, it is impossible to traverse all possible data. In previous training strategies, a common way is to synthesize a fixed amount of noisy speech, e.g., 100 hours. Obviously, more synthetic data will lead to the high demand of clean data and noise data and take more training time with multiple epochs as this way. Based on the above observations, we propose a new training framework via the dynamic data generation. The key idea is to synthetic phase of noisy data is conducted from utterance level to mini-batch level in the training stage. To make sure the model can be trained by adequate data, the batch-based noisy data are generated in a totally random way. Three advantages are gained from this new training method. First, by dynamic generation of training data batch, it is not necessary to prepare and store the fixed training set as in the conventional training method. Second, with the same training time as in the conventional method, more abundant noisy data are actually fed into DNN model. Finally, different evaluation measures, including PESQ, STOI, LSD, and SegSNR, can be consistently improved for the unseen additive noise types, demonstrating the better generalization capability of the proposed training strategy.

The rest of this paper is organized as follows. In Section II, we introduce the traditional DNN-based speech enhancement. The proposed training method for DNN-based speech enhancement is described in Section III. Then, a series of experiments are presented in Section IV. Finally, we summarize our work in Section V.

II. DNN-BASED SPEECH ENHANCEMENT

Algorithm 1 Procedure of traditional DNN training

- **Step 1: Data Preparing**
  Synthesize noisy speech in time domain and calculate the corresponding LPS feature pairs of clean/noisy speech
- **Step 2: Initialization**
  Initialize the DNN parameter set randomly
- **Step 3: Updating the DNN parameters**
  By minimizing Eq. (2), the back-propagation procedure with a stochastic gradient descent method is used to update the parameter set (W,b) in the mini-batch mode of N sample frames with the data prepared in **Step 1**
- **Step 4: Go to Step 3 for the next epoch until the model converges**

Algorithm 2 Procedure of online DNN training

- **Step 1: Initialization**
  Initialize the DNN parameter set randomly
- **Step 2: Data Preparing**
  Synthesize a mini-batch of noisy speech in time domain and calculate the corresponding LPS feature pairs of clean/noisy speech
- **Step 3: Updating the DNN parameters**
  By minimizing Eq. (2), the back-propagation procedure with a stochastic gradient descent method is used to update the parameter set (W,b) once with the data prepared in Step 2
- **Step 4: Go to Step 2 for the next mini-batch until the model converges**

First, we review the conventional DNN-based speech enhancement, especially on the training procedure. The regression model maps the log-power spectra (LPS) features of noisy speech to the clean speech LSP features. In data preparing stage, the noisy speech is synthesized by additive noises and clean speech based on an explicit distortion model in the time domain as follows:

$$y(l) = x(l) + g \cdot n(l), \tag{1}$$

where g is a noise gain factor, and signals $y(l), x(l), n(l)$ represent the l^{th} samples of noisy speech, clean speech and additive noise. So, the training data set consists of a set of time-synchronized clean and noisy utterance pairs. The DNN model, starting with a randomly initialized network, is fine-tuned by minimizing the mean squared error (MSE) between the estimated DNN output and the reference clean LPS features as follows:

$$E = \frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}(\hat{X_n^d}(W^l, b^l) - X_n^d)^2 \tag{2}$$

where *E* is the mean squared error, $\hat{X_n^d}(W^l, b^l)$ and X_n^d denote the enhanced and target LPS features at sample index n and frequency bin *d* with *N* representing the mini-batch size and *D* being the size of the LPS feature vector, (W^l, b^l) denoting the weights and bias parameters to be learned at the l^{th} layer of the DNN. The whole training procedure is summarized as Algorithm 1.

III. THE ONLINE TRAINING STRATEGY VIA DYNAMIC DATA GENERATION

With the rapid increase of clean speech data and noise data, how to make full use of the synthesized training data effectively becomes more and more important. Eq. (1) shows that we can generate unlimited noisy data by changing the parameter g. Considering the time and storage consumption of the training process, it is unrealistic to fully utilize these data. As a result, we usually train our DNN-based model with a fixed amount of data such as 100 hours noisy/clean speech data pairs repeatedly for multiple epochs until the convergence in reality. Obviously, this kind of way can not well benefit from the tremendous training data effectively and may make the model fall into local optimums easily, which drives us to find a better method for training.

To use more training data while not increase too much resource consumption, a novel training procedure using dynamic data generation is proposed. We aim to update the parameters with newly generated data in each mini-batch throughout the entire train stage in our method. All those data in each batch is not stored and just used only once. Apparently, our new idea allows the model to use hundreds of times more data than before.

In the experiment, we give up creating all of the LPS data of noisy speech and clean speech before fine-tuning and replace it by generating data while training. In detail, we generate a mini-batch of data each time and update parameters once with it and repeat this operation until convergence. In such manner, the data generation and parameter update can be done at the same time, which will save the time and storage for preparing training data. Moreover, this new procedure can make sure that more data is utilized in the training process rather than using the same data set at each epoch in the conventional training procedure. The whole training procedure is summarized as Algorithm 2.

However, this online training strategy may bring one issue for the global mean and variance normalization of the input and output features which is a standard operation in the conventional DNN training. Because we do not generate a large amount of training data in advance to calculate the global mean and variance, we address this problem by computing the interpolation of the history mean/variance in the previous mini-batches and the local mean/variance in the current mini-batch. The experiments verify that this operation can work well.

IV. EXPERIMENTS AND RESULT ANALYSIS

All experiments are conducted on TIMIT database [15] as the clean speech data set. The 115 additive noise types which included 100 noise types [16] and 15 home-made noise types were adopted for training to improve the robustness to the unseen noise types. The noisy utterances are synthesized by

TABLE I
AVERAGE LSD, PESQ, SEGSNR AND STOI PERFORMANCE COMPARISON ON THE TEST SET ACROSS THREE UNSEEN NOISE TYPES, BETWEEN DNN MODEL FINE-TUNED WITH 10 HOURS DATA (DENOTED AS STATIC_10H), 100H DATA (DENOTED AS STATIC_100H) AND OUR PROPOSED TRAINING PROCEDURE (DENOTED AS DYNAMIC)

SNR(dB)		-5	0	5	10	15	20	AVE
LSD	Static_10h	17.19	13.61	10.32	8.62	7.99	7.75	10.91
	Static_100h	5.22	3.99	3.13	2.48	2.02	1.68	3.09
	Dynamic	**4.56**	**3.60**	**2.94**	**2.44**	**2.00**	1.68	**2.87**
PESQ	Static_10h	0.95	1.11	1.45	1.95	2.39	2.73	1.76
	Static_100h	1.73	2.25	2.69	3.04	3.33	3.58	2.77
	Dynamic	**1.97**	**2.44**	**2.82**	**3.14**	**3.42**	**3.66**	**2.91**
SegSNR	Static_10h	-2.36	-2.19	-2.10	-1.84	-1.47	-1.06	-1.84
	Static_100h	-2.90	-1.16	0.63	2.51	4.21	5.80	1.52
	Dynamic	**-2.34**	**-0.83**	**0.88**	**2.69**	**4.34**	**5.86**	**1.77**
STOI	Static_10h	0.40	0.50	0.62	0.74	0.83	0.88	0.66
	Static_100h	0.63	0.76	0.86	0.91	0.95	0.97	0.85
	Dynamic	**0.68**	**0.80**	**0.87**	**0.92**	0.95	0.97	**0.87**

TABLE II
AVERAGE LSD, PESQ, SEGSNR AND STOI PERFORMANCE COMPARISON ON THE TEST SET OF THREE UNSEEN NOISE TYPES BETWEEN DNN MODEL FINE-TUNED WITH 10 HOURS DATA (DENOTED AS STATIC_10H), 100 HOURS DATA (DENOTED AS STATIC_100H) AND OUR PROPOSED TRAINING PROCEDURE (DENOTED AS DYNAMIC)

		LSD	PESQ	SegSNR	STOI
Babble	Static_10h	7.49	2.05	-2.80	0.76
	Static_100h	2.87	2.66	1.82	0.84
	Dynamic	**2.68**	**2.80**	**2.09**	**0.85**
Pink	Static_10h	12.3	1.59	-1.48	0.58
	Static_100h	3.31	2.74	0.50	0.83
	Dynamic	**3.01**	**2.88**	**0.63**	**0.86**
White	Static_10h	12.95	1.65	-1.23	0.64
	Static_100h	3.09	2.92	2.23	0.87
	Dynamic	**2.93**	**3.05**	**2.58**	**0.89**

adding a noise signal with a specified SNR level range from -5 dB to 20 dB continuously to a clean speech waveform. The sampling frequency is 16 kHz. A series of frames are extracted by a left-to-right window with 32ms frame length and 16ms frame shift. And then, each frame is represented by a 257-dimensional LPS feature vector. The DNN architecture is 1799-2048-2048-2048-257, where the input layer consisted of neighbouring 7-frame noisy LPS features, 3 hidden layers with 2048 nodes for each layer, and the output layer is the estimated clean LPS features of the centre frame. The parameters W and α are updated with Adam [17] in the mini-batch mode of 256 sample frames and setting learning rate to 0.01.

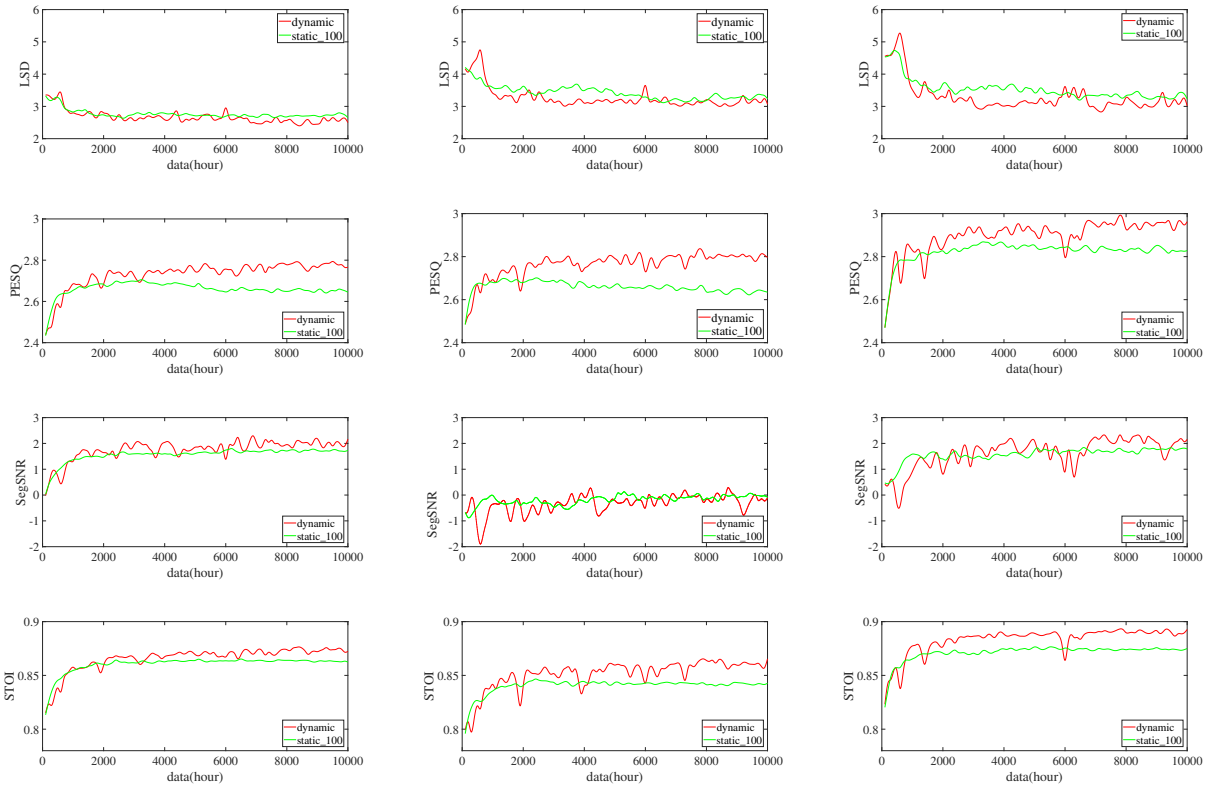
For comparison, the DNN model is repeatedly fine-tuned by 10 hours and 100 hours of noisy training data in traditional method while 10000 hours of noisy training data was used in our new method. Specifically, the 100 hours data is iteratively utilized for 100 times in the traditional method and the 10000 hours data is used just once when updating parameters. In the training stage, all 4260 utterances in the training set of the TIMIT database and 115 noise signals are used for generating noisy speech. In the testing stage, 192 utterances from the test set of the TIMIT database and 3 unseen noise types are added as six levels of SNR, i.e, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB. Two objective quality measures, segmental SNR (SegSNR in dB) and log-spectral distortion (LSD in dB), are used for evaluating the quality of the enhanced speech as in [6]. Besides, perceptual evaluation of speech quality (PESQ) [18] and short-time objective intelligibility (STOI) [19] are also used to assess the quality and intelligibility of the enhanced speech.

A. Overall Evaluation

Table I shows the LSD, PESQ, SegSNR and STOI on the test sets between the traditional methods and the proposed online training strategy. With the same training time consumption, it's obvious that the new training method yields consistent and significant improvements compared with the conventional model trained with 10 and 100 hours data repeatedly. Our proposed method achieves better performance especially in low SNR. The average LSD, PESQ, SegSNR of the proposed method are improved 0.22 dB, 0.14 and 0.25 dB separately. All the results prove that the proposed method is an effective training strategy.

B. Overfitting Analysis

Overfitting is one of the most chronic problems for deep-learning based methods. The performance will suffer a huge performance reduction in validation data by comparison with training data. As illustrated in Figure 1, we compare the performance curves on testing dataset in terms of four objective measures, which are LSD, PESQ, SegSNR and STOI respectively. With the growth of training data size, the performance of traditional method (which is in red curve) rapidly satuates to a certain level. Even worse, PSEQ in Babble noise occurs an obvious decline when training data size exceeds 4,000 hours. On the contrary, the red curve, which represents our proposed dynamic noise training, can get additional gains from the data growth. As we can see, there are some local performance declines along the curves. It can be attributed to the nearby generated training data which is very mismatched with the specific noise type. But still from a long-term perspective,



(a) Babble noise (b) Pink noise (c) White noise

Fig. 1. The comparison of learning curves for all four objective measures (LSD, PESQ, SegSNR and STOI) between the traditional static training method (green curve) and the proposed dynamic training method (red curve) calculated from 20 testing utterances at SNR=5dB corrupted by the babble noise, pink noise and white noise (from left to right). The x-axis is the amount of data used for training.

the overall performance indeed improves a lot. Moreover, the average numbers in Table II also prove its better generalization capability on each unknown noise type. Except for these three kinds of noises, the same phenomenon also appears on another seven unseen noise types in our experiemnts. It demonstrates that the overfitting problem is extremely common in traditional training process, while the proposed method can avoid it to a great degree without much expenses.

V. Conclusions

In this study, we proposed a new training method for DNN-based speech enhancement via the dynamic data generation. By using it, we can update the parameters with newly generated data in each mini-batch throughout the entire training stage, which can not only reduce the trivial traditional operations in data preparation, but also augment the data diversity. After our experiments, the proposed framework has shown better performance on improving speech quality and intelligibility in comparison to the traditional method. By continuously learning brand new data, the generalization capability of the model is also largely enhanced. Moreover, the new training strategy can also be promoted as a general framework to other related tasks which needs the synthesized data, e.g., speech dereverberation and separation.

VI. Acknowledgement

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, MOE-Microsoft Key Laboratory of USTC, and Huawei Noah's Ark Lab.

References

[1] Kingsbury B E D, Morgan N. Recognizing Reverberant Speech With Rasta-Plp. Proc.int.conf.acoust.speech Signal Process, 1997, 2:1259–1262.

[2] Valente F, Friedland G. Speaker Diarization. Cambridge University Press, 2012.

[3] Loizou P C. Speech Enhancement: Theory and Practice. CRC Press, Inc. 2007.

[4] Steven Boll, Suppression of acoustic noise in speech using spectral subtraction, IEEE Transactions on acoustics, speech, and signal processing, vol. 27, no. 2, pp. 113C120, 1979.

[5] Jae Lim and Alan Oppenheim, All-pole modeling of degraded speech, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 3, pp. 197C 210, 1978.

[6] Jae S Lim and Alan V Oppenheim, Enhancement and bandwidth compression of noisy speech, Proceedings of the IEEE, vol. 67, no. 12, pp. 1586C1604, 1979.
[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," IEEE Signal Processing Letters, Vol. 21, No. 1, pp.65-68, 2014.
[8] Y. Xu, J. Du, L.-R. Dai, and C.-H. 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.
[9] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in Proc. ICASSP, 2014, pp.3737-3741.
[10] P.-S. Huang, M. Kim, M. H. Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No.12, pp.2136-2147, 2015.
[11] X.-G. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto-encoder," in Proc. INTERSPEECH, 2013, pp.436-440.
[12] B.-Y. Xia and C.-C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," Speech Communication, Vol. 60, pp.13-29, 2014.
[13] M. Zohrer, R. Peharz, and F. Pernkopf, "Representation learning for single-channel source separation and bandwidth extension," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 23, No. 12, pp.2398-2409, 2015.
[14] Gao T, Du J, Dai L R, et al. "SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement," INTERSPEECH, pp.3713-3717, 2016.
[15] J. S. Garofolo, Getting started with the DARPA TIMIT CD-ROM: Anacoustic phonetic continuous speech database NIST Tech Report, 1988.
[16] G. Hu, 100 nonspeech environmental sounds, 2004.
[17] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.
[18] ITU-T, Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Int. Telecommun. Union-Telecommun. Stand. Sector 2001
[19] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time Cfrequency weighted noisy speech," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 7, pp. 2125-2136, 2011.