Research on Speech Enhancement Based on Deep Neural Network

Yiting Wang, Zhenhua Wei

1 Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data
2 School of Information Engineering, East China University of Technology, Nanchang, 330013, Jiangxi, China

Abstract. The traditional single-channel speech enhancement algorithm has many unreasonable assumptions, which limit the performance of the algorithm. The speech enhancement based on deep neural network can effectively eliminate the problems such as "music noise" in the traditional methods, thus achieving better results than the traditional single-channel speech enhancement. However, the DNN-based speech enhancement model does not perform well at low SNR. In order to improve the robustness of the model, we used voice activity detection (VAD) to process the training data to obtain a new VAD-DNN speech enhancement model. We added 100 kinds of noise in the training set to improve the ability of the model to deal with the unseen noise. At the same time, in order to prevent the occurrence of overfitting problem, we introduced dropout technology to process the network, and improved the generalization ability of the model by using noise awareness training (NAT). Through experiments, we found that the speech enhancement model based on VAD-DNN improved PESQ index by 0.02 and STOI index by 0.01 on average under the condition of low SNR.

1. Introduction

In the traditional speech enhancement algorithm, spectral subtraction [1] obtains an estimated clean speech by subtracting the estimated short-term noise spectrum. However, these traditional methods will produce a kind of perceived interference named "music noise" [2]. Due to the non-linear relationship between noise and clean speech, the method of using neural network modeling was proposed [3]. Scholars proposed to use shallow neural network as a non-linear filter, but the trained models performed poorly [2-4]. In addition, random initialization of shallow neural networks often suffers from local minima or stagnation [5]. In 2006, Hinton started deep network training and proposed a layered unsupervised learning algorithm based on greedy [6]. A speech enhancement framework based on the regressive deep neural networks (DNN) model was proposed, which could remove the residual “music noise” [7]. But traditional speech enhancement methods have a disadvantage that it is difficult to improve the intelligibility of speech [8]. Xu proposed a speech enhancement algorithm based on regression DNN, which can train DNN networks to estimate clean target speech signals from noisy speech [9-10]. Due to the problems of "music noise", speech loss in low SNR conditions, how to improve the performance of speech enhancement model in low SNR environment is still an urgent problem to be studied.
At the same time, we note that a common problem of the speech enhancement model based on neural network is the performance degradation in the environment of invisible noise. The simplest and most effective solution is to add a variety of different noise types to the training set [11-12]. Yuan proposed the NoiseGAN model based on Generative Adversarial Networks (GAN), this model can generate a new noise type, have the ability to improve the noise type diversity training set [13]. In subsequent studies, we found that the VAD can effectively extract necessary speech fragments under the condition of high SNR, but the performance will decline sharply in low SNR [14]. The DNN-based voice activity detection method can solving the difficulty that VAD method needs to design separate models for different noises. This paper proposes a deep neural network learning model based on VAD. We added 100 different noise types into the training set, through experiments, we found that this training set was very effective in dealing with invisible noise types, especially non-stationary noise. At the same time, we introduced dropout technology to prevent the emergence of network overfitting problems. Dropout is a strategy to train the neural network, a certain proportion of neurons are randomly selected from the input layer and hidden layer nodes of the network to set zero, which can avoid the occurrence of overfitting problem [15]. Although not designed to reduce noise, dropout is proved useful to improve the performance of speech recognition in a noisy environment [16]. Meanwhile, we also use noise awareness training (NAT) method to improve the performance of the model. We both input the noise and its estimate, this way can be regarded as the characteristic vector of the speaker in speech recognition at the input end of the model [17]. The rest of the paper is organized as follows. In Section II, we introduce the architecture of speech enhancement model. Section III describes the training process of deep neural network. Section IV introduces the experiment we did. We summarize our findings in Section V.

2. VAD-DNN Model

Figure 1 shows a training data in 0dB environment. Figure 1(a) is pure speech and figure 1(b) is noisy speech. Through comparison, it can be found that the target speech marked in the box in figure 1(b) is completely submerged in the noise. If there is no corresponding pure speech comparison in figure 1(a), it is difficult to find that there is a speech segment in this segment. Under the condition of very low signal-to-noise ratio, the DNN-based model will produce misjudgment, and regard the speech segment submerged in the noise as noise removal, thus resulting in speech distortion and even speech loss.

Voice activity detection is a method to identify speech signals and non-speech signals from a piece of speech and determine the starting and ending points of speech signals [18]. Based on the analysis of the training data in figure 1, we introduce VAD to process the training data. Here, we regard VAD as a classifier. Since the VAD network is a binary classification network, voice and non-voice can be trained respectively through VAD. In order to solve the problem of speech loss of DNN-model under low SNR environment, we use VAD to process the training set, removed the non-speech segments in the training set, and only retained the speech segments to train the enhancement model, so that the new model could better retain the target speech. However, the model is trained only using speech data and the ability to reduce the noise of non-speech segments in noisy speech is limited. Therefore, we introduced NAT to improve the model's ability to deal with noise.
Figure 1. Comparison between pure speech and noisy speech in 0dB environment.

Figure 2. A block diagram based on VAD-DNN voice enhancement framework, including two phases of training and testing. In the training stage, prepare the clean speech database and noise database, and make the training set by artificially adding noise. We used VAD to remove the non-speech segments in the training set, and only retained the fragments of speech to train the enhancement model. Since noisy speech and clean speech in the enhancement model occur in pairs [19], energy-based VAD can be used to detect non-speech segments on clean speech. The training data of the non-speech segment is used to train the new enhancement model, which is denoted as VAD-DNN, and it can better retain target speech.

During the testing phase, the DNN network is used to train a regression model to estimate the LPS features of clean speech from the LPS features of noisy speech. We consider the prediction of noise LPS in the network as a regular term in the network training to prevent the network from falling into local optimization. Simultaneous estimated speech and noise LPS can be used for post-processing of enhancement results. And we use dropout technology and noise awareness training to improve the generalization ability of the model in the model enhancement stage. Overfitting problem means that the model has good fitting ability for the training set, but poor fitting ability for the test set. By discarding a certain proportion of neurons in the input layer, the interdependence between different dimensions can...
be reduced, thus improving the performance of the model on the mismatched test set, especially for the unseen non-stationary noise. Noise awareness training is a simple and effective method to improve the performance of neural network. Although this method assumes that speech and noise are independent of each other, a large number of scholars have found that the performance of this algorithm is not worse than that of the algorithm using MLP to directly estimate LPS [20]. In the network training, we regard the first 6 frames of each speech as noise signals, which are averaged and then spliced into the input features of the model as the pre-estimation of the current noise signals.

3. Network Training

3.1. DNN Network Training

The DNN network structure is used to predict the LPS features of clean and noise from the LPS features of noisy speech. Among them, the prediction of LPS features of noise can be regarded as the regular term in network training. Minimum mean square error (MMSE) and mini-batch stochastic gradient descent (MSGD) are used to update network parameters:

$$E_r = \frac{1}{N} \sum_{n=1}^N \{ \beta \| \hat{X}_n^s - X_n^s \|_2^2 + (1 - \beta) \| \hat{X}_n^e - X_n^e \|_2^2 \}$$  \hspace{1cm} (1)$$

Among them, the $\hat{X}_n^s$ is clean speech estimation and $X_n^s$ is the speech target feature to be learned. $\hat{X}_n^e$ and $X_n^e$ is characteristic of noise estimation and noise goal for learning characteristics. $N$ is the sample number at the time of update, and $\beta$ is used to adjust the impact of target estimation on network training, where $\beta$ is set to 0.7.

3.2. VAD-DNN Training

In the training phase, we use DNN structure to train the VAD model, because the VAD network is a two-class network, one is speech and the other is non-speech, we use the Softmax activation function replace the linear function as the output layer of the network, the LPS features of noisy speech is also used as input. Supervised training of DNN networks is based on the following Cross-Entropy (CE):

$$C = - \sum_{j=1}^Q q_j \ln p_j$$  \hspace{1cm} (2)$$

Where $C$ represents the objective function of cross entropy, $p_j$ is the network output, $q_j$ is the learning objective, and $Q$ is the number of categories of classification network, where $Q=2$.

We regard the first 6 frames of each speech as noise signals, and take the average as input features of the model, which is regarded as the pre-estimation of the current noise signal:

$$\hat{Z}_n = \frac{1}{M} \sum_{m=1}^M Y_m$$  \hspace{1cm} (3)$$

After training, the LPS features of the input noisy speech can be used to obtain the LPS estimation of clean and noise through the model. The ideal time-frequency masking value of DNN learning target is defined as follows [21]:

$$IRM_n(d) = \frac{\exp(x_n^s(d))}{\sqrt{\exp(x_n^s(d)) + \exp(x_n^e(d))}}$$  \hspace{1cm} (4)$$

Where $n$ means the frame and $d$ means the feature dimension. DNN for the IRM is designed as a regression model in which the output can be considered as the probability of the presence of speech at each moment. The input of DNN is the noise LPS features of the current frame and adjacent frames. We use $IRM$ estimate of the clean speech for post-processing [8]:

4
\[
\hat{X}_n(d) = \begin{cases} 
  Y_n(d), & \text{if } \hat{IRM}_n(d) > \gamma \\
  \hat{X}_n^s(d), & \text{if } \hat{IRM}_n(d) < \lambda \\
  (\hat{X}_n^s(d) + Y_n(d))/2, & \text{otherwise}
\end{cases}
\] (5)

The \( \hat{X}_n(d) \) is the post-processing estimate of clean speech, \( Y_n(d) \) is LPS features of the original speech with noise. \( \gamma \) and \( \lambda \) are adjustable threshold parameter, and \( \gamma \) set to 0.75 and \( \lambda \) set to 0.1.

4. Experiment

4.1. Experimental Configuration

The clean voice data comes from TIMIT database, and the noise library mainly adopts the noise library recorded by G.HU (HU,2014), covering 100 noise types. During the training phase, we use artificial noise addition to mix clean and noisy speech. From the TIMIT test set, we randomly selected 200 speeches and add noise to form a noisy training set. In the test phase, three noise types from real noise library were selected: Factory, Mess hall, Bus station, which did not appeared in the training set. The sampling frequency of the file is 16kHz, the DFT coefficient is calculated by Short-Time Fourier Transform (STFT) according to the frame length of 512 sampling points and the frame shift of 256 sampling points, and then get the 257-dimensional LPS features.

In the network configuration, use three hidden layers, each hidden layer has 2048 nodes, use Sigmoid as the activation function of the hidden layer, and the input features are spliced with contextual information of about 3 frames. Dropout drops 0.1 for the input layer, 0.2 for the hidden layer, and nothing for the output layer. The learning rate of DNN was set to 1, the initial impulse was 0.5, and it gradually increased to 0.9 over the first ten iterations.

In the evaluation of speech enhancement results, PESQ and STOI were used as evaluation indicators. PESQ is the P.862 standard recommended by the International Telecommunication Union. It is an indicator used to measure the listening quality of a piece of speech in speech enhancement. It is also one of the most important goals of speech enhancement. The value range is -0.5 to 4.5, and larger value indicates better performance. STOI measures the intelligibility of speech, which ranges from 0 to 1. There is a larger relationship between speech intelligibility and speech distortion and speech loss. The higher the distortion, the lower the STOI value.

4.2. Analysis and Comparison

Table 1 shows the results of PESQ and STOI in three noise types of VAD-DNN based speech enhancement system. Among them, Noisy represents the original Noisy speech, DNN represents the DNN-based speech enhancement system, and VAD-DNN represents the combined VAD-DNN speech enhancement system proposed in this paper. We can see that compared with the DNN speech enhancement system, the VAD-DNN based speech enhancement system can effectively improve the performance under the condition of low SNR. It can be seen that under three noise types, VAD-DNN model has a stable enhancement effect compared with DNN model, which means that the enhancement effect of VAD-DNN model at low SNR is due to DNN model. Meanwhile, according to the mean values of the three noises, PESQ increased by 0.07 and STOI increased by 0.05 respectively at -5dB.On 0dB, PESQ improved by 0.03; On 5dB, PESQ improved by 0.02 and STOI by 0.01.
Table 1. Comparison of PESQ and STOI results for three types of noise not seen.

| Noise Type | SNR | PESQ |  | STOI |  |
|---|---|---|---|---|---|
|   |   | Noisy | DNN | VAD-DNN | Noisy | DNN | VAD-DNN |
| Factory | -5dB | 0.97 | 1.54 | 1.55 | 0.53 | 0.59 | 0.65 |
|       | 0dB  | 1.31 | 2.02 | 2.02 | 0.65 | 0.75 | 0.76 |
|       | 5dB  | 1.68 | 2.45 | 2.40 | 0.77 | 0.84 | 0.86 |
| Mess hall | -5dB | 1.13 | 1.54 | 1.55 | 0.53 | 0.59 | 0.63 |
|         | 0dB  | 1.31 | 2.03 | 2.02 | 0.65 | 0.77 | 0.76 |
|         | 5dB  | 1.67 | 2.40 | 2.45 | 0.77 | 0.86 | 0.86 |
| Bus station | -5dB | 1.08 | 1.40 | 1.58 | 0.51 | 0.60 | 0.63 |
|         | 0dB  | 1.35 | 1.98 | 2.08 | 0.65 | 0.76 | 0.77 |
|         | 5dB  | 1.72 | 2.40 | 2.43 | 0.76 | 0.87 | 0.88 |
| Average | -5dB | 1.06 | 1.49 | 1.56 | 0.52 | 0.59 | 0.64 |
|        | 0dB  | 1.32 | 2.01 | 2.04 | 0.65 | 0.76 | 0.76 |
|        | 5dB  | 1.69 | 2.41 | 2.43 | 0.77 | 0.86 | 0.87 |

Figure 3 shows a voice enhancement sample of VAD-DNN model under Mess hall noise scenario on -5dB. (a) is clean speech, (b) is noisy speech, (c) is enhanced speech based on DNN system, (d) is enhanced speech based on VAD-DNN system. It can be found that (d) is no obvious speech loss, and the background noise in (d) is obviously suppressed. This proves that the speech enhancement model based on VAD-DNN can effectively reduce speech loss and improve the enhancement effect under the condition of low SNR.

5. Conclusion
Aiming at the problem of speech distortion and speech loss caused by DNN-model under low SNR, we proposed a VAD-DNN model. The model analysed data of speech signal under the condition of low SNR, and proposes improvement strategies from the training set. On the one hand, add a variety of different noise types to improve the generalization ability of the model. On the other hand, VAD is used to classify the training set. In the meantime, in order to improve the model's ability to adapt to noise, we introduced dropout to avoid network overfitting and noise awareness training to enhance the model's
ability to analysis unseen noise. The experimental results show that the new model obtained from the training set processed by VAD algorithm can effectively reduce the occurrence of speech distortion. Through the tests of three types of mismatched noise, the speech enhancement model based on VAD-DNN improves PESQ index by 0.02 on average and STOI index by 0.01 on average under the condition of low SNR.

Acknowledgments
This paper is supported by the Open Research Fund of Jiangxi Engineering Laboratory on Radioactive Geoscience and Big Data Technology (No.JELRGBDT201704), the Open Research Fund of Jiangxi Engineering Technology Research Center of Nuclear Geoscience Data Science and System (No.JETRCNGDSS201804), the Natural Science Foundation of Jiangxi Province of China (No.20161BAB212056), and the Open Research Fund of Jiangxi Province Key Laboratory of Water Information Cooperative Sensing and Intelligent Processing (No. 2016WICISP08), and East China University of Technology Graduate School Innovation Project Fund (No. DHYC-201928).

References
[1] LIM J S, OPPENHEIM A V. Enhancement and bandwidth compression of noisy speech[J]. Proceedings of the IEEE, 2005, 67(12):1586-1604.
[2] SCALART P, FILHO J V. Speech enhancement based on a priori signal to noise estimation[C]//Acoustics, Speech, and Signal Processing, 1996.on Conference Proceedings. USA, Atlanta, Georgia, 1996 IEEE International Conference, IEEE Computer Society. Atlanta Georgia OSA: IEEE Press, 1996:629-632.
[3] LU X G, TSAO Y, MATSUDA S, et al. Speech enhancement based on deep denoising Auto-Encoder[C]//The 14th Annual Conference of the International Speech Communication Association. Lyon: ISCA Archive, 2013:436-440.
[4] EPHRAIM Y, MALAH D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 2003, 52(6):1109-1121.
[5] BENGIO Y. Learning Deep Architectures for AI[J]. Foundations & Trends® in Machine Learning, 2009, 2(1):1-127.
[6] HINTON G E, OSINDERO S, TEH Y W. A Fast Learning Algorithm for Deep Belief Nets[J]. Neural Computation, 2006, 18(7): 1527-1554.
[7] XU Y, DU J, DAI L R, et al. An Experimental Study on Speech Enhancement Based on Deep Neural Networks[J]. IEEE Signal Processing Letters, 2013, 21(1): 65-68.
[8] GAO T, DU J, DAI L R, et al. 2017. A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments[J]. Speech Communication, 95:28-29.
[9] XU Y, DU J, DAI L R, et al. 2014. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters, 21(1):65-68.
[10] Xu Y, Du J, Dai L R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 23(1):7-19.
[11] J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” IEEE Trans. on Speech and Audio Processing, Vol. 11, No. 3, pp. 184-192, 2003.
[12] Y. X. Wang and D. L. Wang, “Towards scaling up classification-based speech separation,” IEEE Trans. on Audio, Speech and Language Processing, Vol. 21, No. 7, pp. 1381-1390, 2013.
[13] Yuan wenhao, Lou yingxi, liang chunyan, xia bin. Improving the generalization ability of speech enhancement by generating noise [J]. Acta electronica sinica,2019,47(04):791-797.
[14] Chen yingying, bi chunyan, long jianzhong. Speech endpoint detection algorithm with low signal-to-noise ratio [J]. Television technology, 2018, vol. 42 (6):9-12, 27.
[15] Zhang yun, li LAN, wang haodong. A Dropout optimization algorithm to prevent overfitting in machine learning [J]. Journal of lanzhou university of arts and sciences (natural science edition), 2019, vol. 33 (6):84-87.

[16] M. Seltzer, D. Yu and Y. Wang. “An investigation of deep neural networks for noise robust speech recognition,” Proc. ICASSP, pp. 7398-7402, 2013.

[17] George Sano, Hagen Soltau, David Nahamoo, and Michael Picheny. Speaker adaptation of neural network acoustic models using i-vectors. In Automatic Speech Recognition and Understanding(ASRU), 2013 IEEE Workshop on, pages 55-59. IEEE,2013.

[18] Wang yao, zeng qingning, long chao et al. Improved method for speech endpoint detection in low SNR environment [J]. Acoustic technology, 2018, vol.37 (5):457-467.

[19] POVEY D, GHOSHAL A, BOULIANNE G, et al.2011. The kaldi speech recognition toolkit[C]/IEEE 2011 workshop on automatic speech recognition and understanding. [S.1.] IEEE Signal Processing Society.

[20] Bao Changchun, xiang Yang. Review of single-channel speech enhancement methods based on deep neural network [J]. Signal processing, 2019, (12).

[21] Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. Audio Speech Lang. Process. IEEE/ACM Trans. 22 (12), 1849–1858.