applied sciences

MDPI

Article

Environmental Attention-Guided Branchy Neural Network for Speech Enhancement

Lu Zhang [1], Mingjiang Wang [1,*], Qiquan Zhang [1] and Ming Liu [2]

[1] Shenzhen Key Laboratory of IoT Key Technology, Harbin Institute of Technology, Shenzhen 518000, China; 18B952047@stu.hit.edu.cn (L.Z.); zhangqiquan_hit@163.com (Q.Z.)

[2] Sino-German School, Shenzhen Institute of Information Technology, Shenzhen 518000, China; lm_hit_1986@126.com

* Correspondence: mjwang@hit.edu.cn; Tel.: +86-0755-8655-5455

Received: 18 December 2019; Accepted: 6 February 2020; Published: 9 February 2020

Abstract: The performance of speech enhancement algorithms can be further improved by considering the application scenarios of speech products. In this paper, we propose an attention-based branchy neural network framework by incorporating the prior environmental information for noise reduction. In the whole denoising framework, first, an environment classification network is trained to distinguish the noise type of each noisy speech frame. Guided by this classification network, the denoising network gradually learns respective noise reduction abilities in different branches. Unlike most deep neural network (DNN)-based methods, which learn speech reconstruction capabilities with a common neural structure from all training noises, the proposed branchy model obtains greater performance benefits from the specially trained branches of prior known noise interference types. Experimental results show that the proposed branchy DNN model not only preserved better enhanced speech quality and intelligibility in seen noisy environments, but also obtained good generalization in unseen noisy environments.

Keywords: speech enhancement; attention mechanism; noise classification; branchy deep neural network

1. Introduction

Speech enhancement techniques have been widely used to cope with the noise interference problem in the front end of many speech applications, such as mobile phones, hearing aids, and speech recognition products. The early conventional methods paid more attention to the optimization of suppression gain function [1–3], the estimation of the noise spectrum [4–6], and prior signal-to-noise ratio (SNR) [7,8]. In recent years, deep neural network (DNN)-based speech enhancement methods have shown significant performance advantages over the traditional approaches in complex noise environments, even the extremely nonstationary noises. Whether utilizing masking-based [9–11] or mapping-based [12–17] DNN methods, their general rule is to optimize the loss function between the ideal and noisy targets to achieve as little error as possible in the global noisy speech dataset. Consequently, richer datasets, and a better objective function and neural network models were further explored to guarantee the robust generalization ability of DNN models to cope with the diversified noise environments in real life. In the research of [14], experimental results demonstrated that the richness of the clean speech samples and the noise samples were the two crucial aspects to improve the generalization capacity of DNNs. Recently, researchers have paid more attention to the optimization of DNN models for the speech enhancement task. Considering the temporal relationship of speech signals, long short-term memory (LSTM) networks [18] and temporal convolutional networks (TCN) [19] have been proven to have better speech reconstruction capacities for speech enhancement. In addition, the optimization of the objective function [20,21] has also been explored

for the performance improvement of DNN-based speech denoising. Generative adversarial networks (GAN) [22] have been another new solution to train the DNN model to learn noise suppression abilities. These methods aim to train a common DNN denoising model for all kinds of noise interference in different application scenarios.

However, most speech products have their specific application scenarios. When the application scenario is determined, the types of noise interference are known. In different noise environments, the optimal parameters of speech enhancement algorithms are different. As investigated in [23], this prior environmental information was helpful to further improve the speech enhancement methods to achieve better noise suppression effect in specific noise environments. Therefore, some researchers started to design the speech enhancement algorithms by incorporating the prior noise information. In [24–26], the noise classification module was integrated into the Wiener filter and some statistical model-based speech estimators. Guided by the noise classification module for optimal parameter selection, the performance of traditional methods was further improved in prior known noisy environments. A similar idea was applied to the DNN-based denoising algorithms [27–29]. In the recent studies of [28,29], under the guidance of the noise classification unit, several independent DNN models were trained to give full play to their respective noise reduction capabilities in different noise environments. Although this "divide and conquer" strategy effectively improves the noise reduction performance, it also increases the storage burden and reduces the complementarity between different noises.

In this work, we proposed a novel branchy neural network (BNN) framework to improve the storage burden and noise complementarity problem of separate training. There are two key modules working together in our proposed framework, a classification network and a denoising network. First, the noise classification network is trained to analyze the presence probability of each noise component for every input noisy frame. Then, the estimated noise presence probability (NPP) is an indicator of the middle hidden layer in the denoising network, to determine which branchy path is opened for back propagation or forward propagation. In addition to the above "special branches" for noise reduction, the denoising network also learns a "common branch" in the middle layer which further improves the generalization performance of branchy model in unseen noise environments.

The paper is organized as follows: Section 2 describes the model design and the theoretical analysis of the proposed speech enhancement algorithm; in Section 3, a set of experiments for the denoising performance evaluation are conducted; finally, Section 4 concludes the paper.

2. Branchy Neural Network with Attention Mechanism

The proposed architecture of the branchy neural network is shown in Figure 1. The proposed method consists of two modules, a classification neural network and a denoising neural network. The denoising neural network, which looks like a "sandwich", has multiple branches in the middle layer to suppress different noise interference in specific application scenarios. The classification neural network acts as a multidirectional attention switch in the whole framework and produces the estimated NPP of each noise components to determine the contribution of each branch to noise reduction. These two key modules are described in the following subsections.

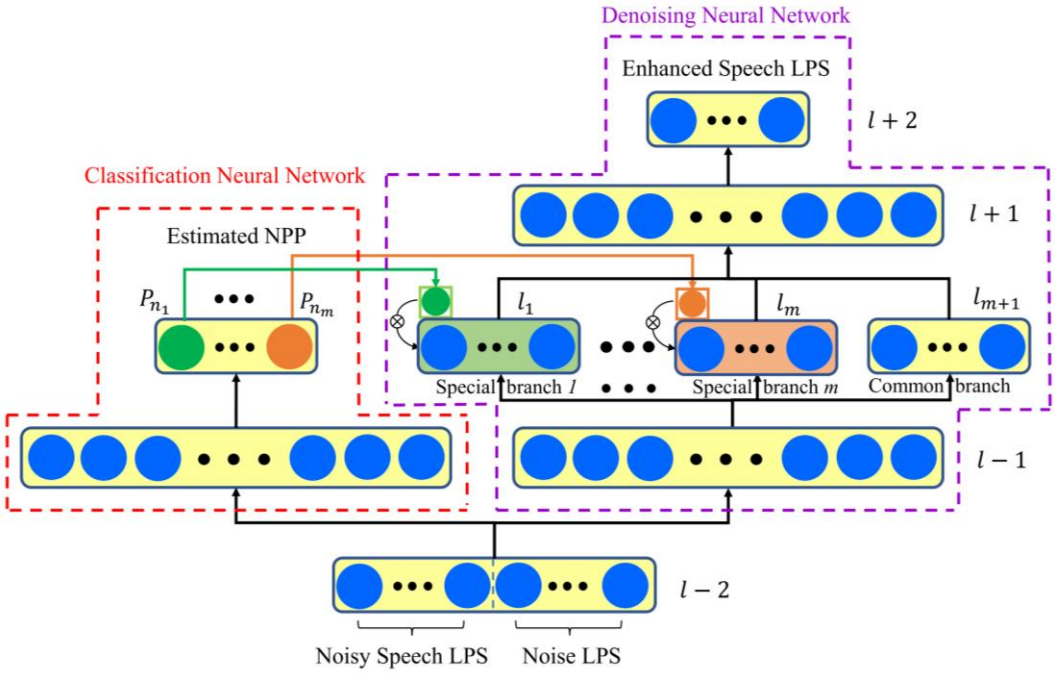

Figure 1. Illustration of the proposed branchy neural network structure that has two key modules, a classification neural network and a denoising neural network.

2.1. Classification Neural Network for Attention Allocation

The classification neural network is designed as a three-layer fully connected network, which has m output neurons in the output layer to predict the NPP of m noise interference types. Its hidden layer has 2048 neurons with rectifier linear activation units (ReLU) to ensure sufficient model capacity for noise classification. Considering the multiple noises in different application scenarios, multi-classified softmax cross-entropy function is used as the loss function for DNN training:

$$Class\_Loss = -\frac{1}{m}\sum_{i=1}^{m} T_i \log\left(\frac{e^{z_i}}{\sum_{j=1}^{m} e^{z_j}}\right) \tag{1}$$

where T_i represents the i-th ideal noise label for learning, m denotes the number of class categories, z denotes the output vector of the fully connected layer, and z_j represents the j-th value of vector z. The SoftMax function in Equation (1) ensures that the sum of attention allocation is one.

To further strengthen the noise classification accuracy of the classification network, a noise-aware feature is extracted for each input frame. Since the NPP is related to the posteriori signal-to-noise ratio (SNR) levels, a weighted average approach based on the hard threshold in log power spectrum (LPS) domain is proposed to update the noise spectrum feature of each frame, as shown in Equation (2):

$$\begin{aligned} &if\ \tfrac{Y(k,t)}{\hat{N}(k,t-1)} < \beta \\ &\quad \hat{N}(k,t) = \alpha\hat{N}(k,t-1) + (1-\alpha)X(k,t) \\ &else \\ &\quad \hat{N}(k,t) = \hat{N}(k,t-1) \\ &end \end{aligned} \tag{2}$$

where $X(k,t)$ is the noisy speech LPS calculated from the input signals of each time frame, $\hat{N}(k,t)$ is the estimated noise LPS, k and t represent the frequency and frame index, respectively, α is the smoothing factor (which is fixed as 0.9 here), and β is the posteriori threshold; the empirical value of β = 2.5

gives a good compromise between underestimation and overestimation of the noise spectrum [30]. In addition, the concatenated input features are normalized to a mean value of zero and a variance of one, beneficial for gradient descent.

After model training, the classification network produces the NPP of each noise type in its output layer. As shown in Figure 1, $p_{n_1}, \ldots, p_{n_m}$ represent the presence probability of m noises, n_m is the index of different noise types, and the sum of these probability values is one. This means that the classification network analyzes and generates the proportion of noise components for each noisy input frame, and therefore controls the attention emphasis of the denoising network on different branches. These estimated NPP values are regarded as the prior environmental information for the training and testing of the denoising neural network.

2.2. Denoising Neural Network with Attention-Based Branchy Structure

To make the neural network more sensitive to the changes of the noise environment, the concatenated noise-aware features are also adopted as the input of our proposed denoising neural network. In addition, as shown in Figure 1, the denoising network has $m + 1$ branches, in which the first m branches are special branches, and the last branch is the common branch. Each special branch is used to suppress a specific noise, while the common branch can handle all noise interferences. As the mean squared error (MSE) criterion in the log domain is more consistent with the human auditory system [31], the proposed branchy DNN model is trained with the following loss function:

$$Regress\_Loss = \frac{1}{N \cdot K} \sum_{t=1}^{N} \sum_{k=1}^{K} \left(\hat{Y}(k,t) - Y(k,t) \right)^2 \tag{3}$$

where $\hat{Y}(k,t)$ and $Y(k,t)$ are the enhanced speech LPS and clean speech LPS, respectively, with K denoting the frame size of clean LPS spectrum, and N representing the number of frames in each mini-batch. They are also normalized by the mean and variance of input features.

The proposed denoising neural network has three hidden layers and $m + 1$ branches are divided in the middle hidden layer to achieve environment-aware noise reduction effect. The first and last hidden layers have 2048 neurons, and each branch in the middle hidden layer has 1024 neurons. During model training, the estimated NPP is multiplied on the neurons of each branchy layer to control the direction of back propagation, and therefore trains specified denoising paths of the middle layer. If the middle layer is indexed as the l-th layer, its output can be expressed as:

$$\begin{aligned} Z_{l,i} &= p_{n_i} * (W_{l,i} A_{l-1} + b_{l,i}) \\ A_{l,i} &= g(Z_{l,i}), \quad i = 1, 2, \ldots, m \end{aligned} \tag{4}$$

where $Z_{l,i}$ and $A_{l,i}$ represent the linear and non-linear output of the i-th branch layer, respectively, $g(\cdot)$ is the ReLU activation operation, and $*$ denotes cross multiplication. When back propagation is carried out, the parameter update of each branch is affected by the estimated NPP, which is derived as follows (see detailed derivation in Appendix A):

$$\begin{aligned} \mathrm{d}W_{l+1,i} &= \sigma_J W_{l+2} * g'(Z_{l+1}) g[p_{n_i} * (W_{l,i} A_{l-1} + b_{l,i})], \quad i = 1, 2, \ldots, m \\ \mathrm{d}b_{l+1,i} &= \sigma_J W_{l+2} * g'(Z_{l+1}) \end{aligned} \tag{5}$$

$$\begin{aligned} \mathrm{d}W_{l,i} &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,i} g'(Z_{l,i}) p_{n_i} * A_{l-1}, \quad i = 1, 2, \ldots, m \\ \mathrm{d}b_{l,i} &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,i} g'(Z_{l,i}) p_{n_i} \end{aligned} \tag{6}$$

where σ_J represents the gradient of regression loss function. As shown in Equations (5) and (6), if the NPP value p_{n_i} is close to zero, the gradients of $W_{l+1,i}$, $W_{l,,i}$, and $b_{l,i}$ are close to zero, which means that the gradient update of weights on both sides of the i-th branch layer are blocked. As the model training proceeds, each branch gradually learns the specific denoising ability for a certain kind of

noise. It should be noted that only the first m branches of the l-th and $(l + 1)$-th layers process the noise separately; the $(l - 1)$-th and $(l + 2)$-th layers still learn some general characteristics for noise reduction. Furthermore, to better retain the complementary advantages of multiple noises, the last branch ($(m + 1)$-th branch) in the middle layer is designed as a common path for gradient descent without the influence of NPP. The gradients of the l-th and $(l + 1)$-th layers related to the $(m + 1)$-th branch are derived in the following Equations (7) and (8):

$$\begin{aligned} \mathrm{d}W_{l+1,m+1} &= \sigma_J W_{l+2} * g'(Z_{l+1}) g(W_{l,m+1} A_{l-1} + b_{l,m+1}) \\ \mathrm{d}b_{l+1,m+1} &= \sigma_J W_{l+2} * g'(Z_{l+1}) \end{aligned} \tag{7}$$

$$\begin{aligned} \mathrm{d}W_{l,m+1} &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,m+1} g'(Z_{l,m+1}) A_{l-1} \\ \mathrm{d}b_{l,m+1} &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,m+1} g'(Z_{l,m+1}) \end{aligned} \tag{8}$$

3. Experiments and Results

3.1. Experimental Settings

Our proposed algorithm was trained and evaluated on the TIMIT speech database [32] corrupted by the noises from Noisex-92 noise database [33]. Six noises were selected to generate the noisy speech database. Each selected noise file was split into three non-overlapping sections used for training (60%), validation (20%), and test (20%). From the TIMIT training set, 4620 utterances were mixed with four noises (babble, factory1, destroyer engine, and destroyer operation noises) to generate the 12.5 h noisy speech training database. A total of 280 utterances from the TIMIT test set were mixed with the validation section of the same four noise types to generate the noisy speech validation set. The mixed SNR levels followed the uniform distribution in the range of −5 to 15. For the denoising performance evaluation of the DNN model, 320 unseen utterances from the TIMIT test set were mixed with four seen noises and two unseen noises (pink, factory2) to generate the noisy speech test set. The SNR levels of each test noisy speech were fixed at −5, 0, 5, 10, and 15 dB to obtain insights into performance at different degradation levels. The perceptual evaluation of speech quality (PESQ) [34] and the short-time objective intelligibility (STOI) [35] were adopted as two metrics to evaluate the enhanced speech quality and intelligibility.

All the aforementioned clean speech and noise signals were resampled to 16 kHz before mixture generation. For the input of the proposed model, 514-dimensional noise-aware features were extracted by performing the short-time Fourier transform (STFT) on noisy signals using a 512-point Hamming window with 50% overlap. During model training, first, the classification network was trained, for five epochs, in each mini-batch with 1024 frames, by following the Adam optimization method [36] with a learning rate of 0.0001. Then, the denoising network was trained for 30 epochs under the guidance of the well-trained classification network. It was also optimized, by the Adam method with the learning rate of 0.0002, in each mini-batch with 1024 frames. To reduce the influence of overfitting, batch normalization [37] and a dropout strategy [38] (0.2 dropout rate) were applied in the hidden layers of the classification and denoising networks.

3.2. Performance Evaluation of Branchy Neural Network

3.2.1. Classification Accuracy Evaluation

In our proposed DNN framework, the classification network plays an important role in the process of back propagation and forward propagation. Assuming that the proposed approach works in an application scenario with four known noise interferences (babble, factory1, destroyer engine, and destroyer operation noises), the classification network needs to have sufficient accuracy to lead the denoising network to learn specific branches in the training stage and distinguish the noise types in the test stage. We performed experiments to evaluate its classification accuracy in the generated training and validation datasets, as shown in Table 1. The “Noisy LPS” denotes that the classification network

only used the noisy log power spectrum (LPS) feature, and "Noisy LPS + Noise LPS" represents that the estimated noise log power spectrum features were concatenated with the noisy LPS for noise classification. Furthermore, the evaluation of classification accuracy in the validation dataset reflects the generalization performance of the model. The experimental results in Table 1 show that the presented classifier with the estimated noise LPS features contributed to better performance in noise classification and model generalization. The classification accuracy in the training and validation datasets is more than 99%, which means that the proposed classification module has sufficient ability to guide the training and testing of the denoising module.

Table 1. Classification accuracy results in train and validation datasets.

Classification Accuracy (%)	Input Features	
	Noisy LPS	Noisy LPS + Noise LPS
Train dataset	97.41	99.92
Validation dataset	94.55	99.64

3.2.2. Denoising Performance Evaluation

The noise classification guided denoising model with four special branch paths only (denoted as CGBNN-4) and the denoising model with four special branches and one common branch (denoted as CGBNN-5) were tested for performance evaluation in the seen and unseen noise cases of the test dataset. Each special branch is designed to suppress a specific type of noise interference, and the common branch is designed for the suppression of all four noises. That is, unlike CGBNN-4, which only considers the dedicated system for noise reduction, CGBNN-5 adopts both the dedicated system and the general system for noise reduction.

To further evaluate the model superiority of our proposed branchy neural network, we compared the performance results of PESQ and STOI between two branchy methods and two DNN methods without environmental noise information. The classical fully connected DNN model [13] with three hidden layers and the state-of-the-art method [18] using LSTM for progressive learning are denoted as "DNN baseline" and "LSTM-PL", respectively, as shown in Figure 2. The unprocessed noisy speech (denoted as "Noisy") is also presented in Figure 2 to show how much performance has been improved. In addition, Figure 2 presents the denoising performance results of four DNN-based methods in four seen and two unseen noise environments. Figure 2a,b shows the averaged PESQ results in the seen and unseen noise environments. These results demonstrate that the CGBNN-4 and CGBNN-5 achieved better quality of reconstructed speech than the classical DNN baseline and LSTM-PL method. The strategy of combining expert systems (special branch) with an omnipotent system (common branch) enables the CGBNN-5 to have the best denoising performance and generalization ability. Compared with the LSTM-PL and DNN-baseline method, the proposed CGBNN-5 achieved an averaged PESQ improvement of 7.24% and 4.26% in seen noise cases and 3.94% and 5.32% in unseen noises, respectively. For the speech intelligibility evaluation, Figure 2c,d shows the averaged STOI results in the seen and unseen noise environments. The proposed CGBNN-5 achieved an averaged STOI improvement of 2.26% and 2.68% over the DNN-baseline in seen and unseen noise cases, respectively. Although the proposed CGBNN-5 does not perform better than the LSTM-PL method, the CGBNN-5 has lower computational complexity than the LSTM-PL. The feed-forward operation saves more model parameters than the recursive operation of LSTM networks. Furthermore, the proposed CGBNN-5 did not achieve improvement for the STOI measurement in a high SNR condition (15 dB), but it still made contributions on the improvement of PESQ in the 15 dB SNR case.

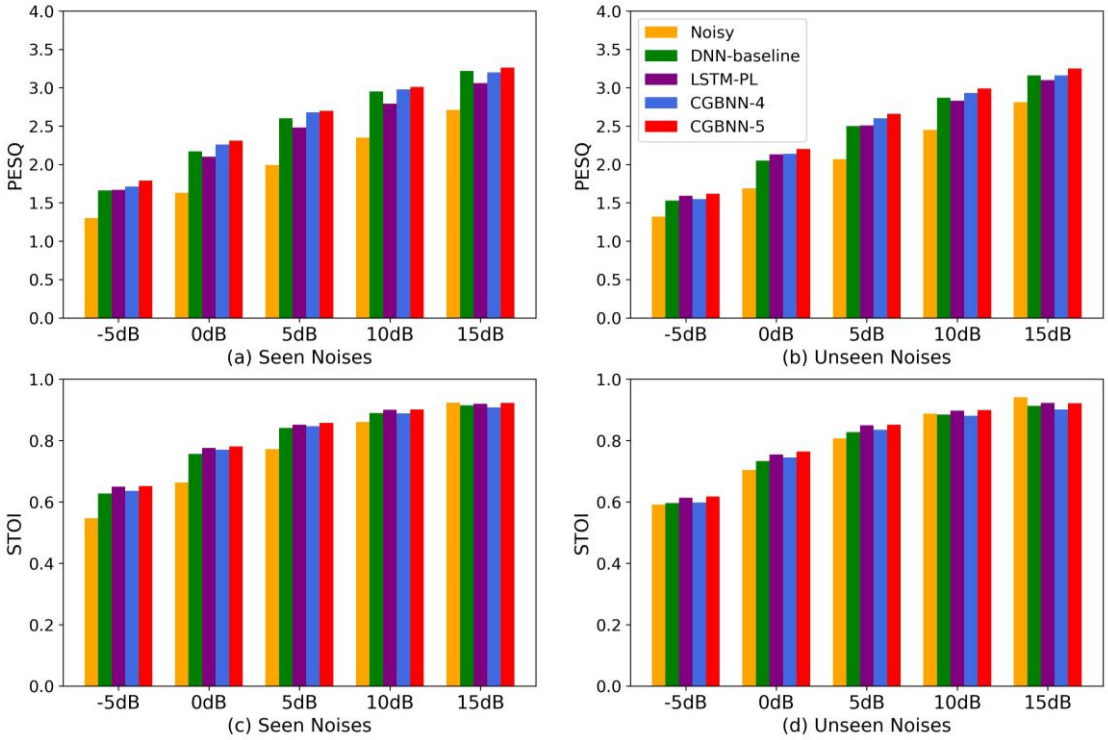

Figure 2. Averaged perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) results obtained for noisy and deep neural network (DNN) enhanced speech in seen and unseen noisy environments. (**a**) PESQ in seen noises; (**b**) PESQ in unseen noises; (**c**) STOI in seen noises; (**d**) STOI in unseen noises.

3.3. Performance Comparison with Other Environment-Aware Methods

The proposed method was compared with the DNN-based approaches without prior noise information for performance evaluation. However, the performance difference between the proposed method and other environment-related speech enhancement algorithms is unknown. In this section, we compare the denoising performance of the proposed speech enhancement approach (CGBNN-5) with three classic environment-aware algorithms in the specific application scenarios. The noise classification-based minimum mean square error speech estimator (MMSE) [25] and the optimally modified log-spectral amplitude (OMLSA) speech estimator [26] were evaluated to compare speech quality and intelligibility. Both of them were tested in the same application scenario with four known noise interferences. These two methods are denoted as NC-MMSE and NC-OMLSA in Figure 3. Additionally, the separate DNN denoising method with noise classification [29] (denoted as NC-DNN) was tested for performance comparison purposes. Figure 3 shows the averaged PESQ and STOI results of the above environment-aware methods in our test dataset. An intuitive comparison of the reconstructed speech spectrogram between the proposed CGBNN-5 and three comparing algorithms in four specific noise environments are shown in Figure 4. The comparison results of DNN model sizes are listed in Table 2. The model sizes were normalized by the NC-DNN model to make the comparison look more obvious.

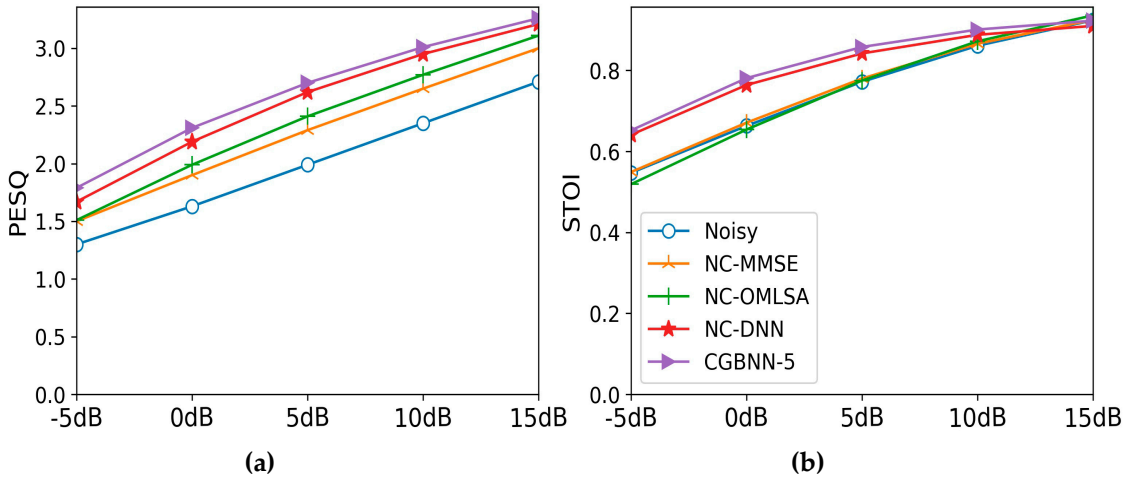

Figure 3. Averaged PESQ and STOI results for environment-aware methods with four known (babble, factory1, destroyer engine, and destroyer operation) noise interferences in a specific application scenario. (**a**) Averaged PESQ results; (**b**) Averaged STOI results.

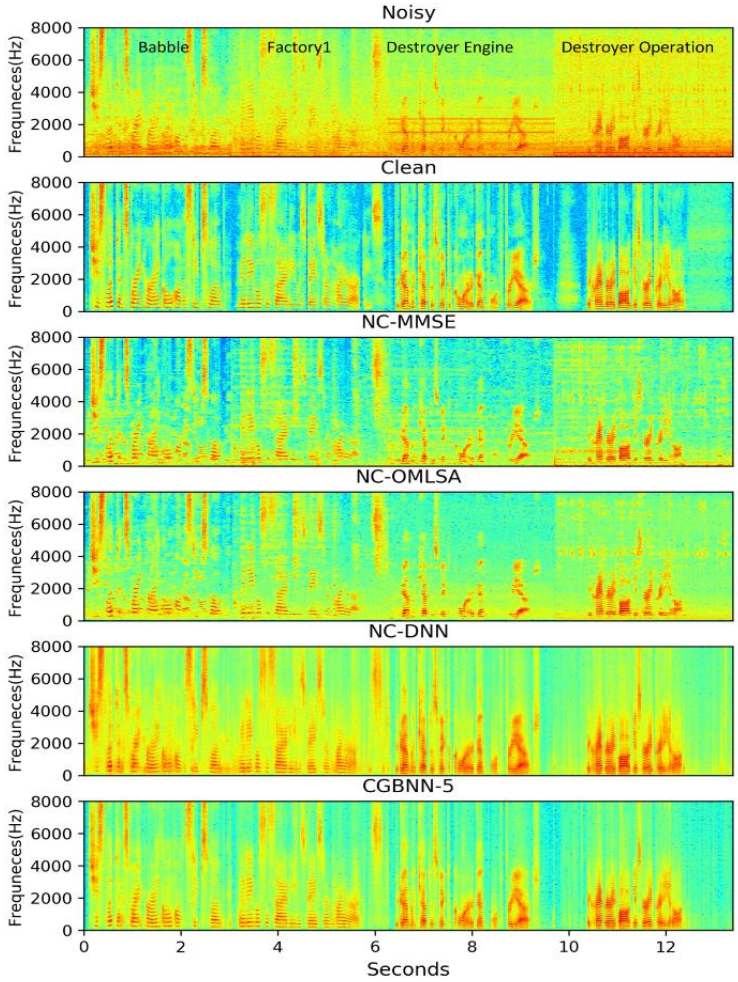

Figure 4. The spectrograms of noisy speech (degraded by four known noises in 0 dB signal-to-noise ratio (SNR) case), clean speech, and enhanced speech using NC-MMSE, NC-OMLSA, NC-DNN, and CGBNN-5.

Table 2. Comparison of the normalized model size.

Comparing Methods	NC-DNN	CGBNN-5
Model size	1	0.65

The averaged PESQ results, in Figure 3, demonstrate that the proposed CGBNN-5 outperforms the other environment-aware methods in SNR ranges from −5 dB to 15 dB. Compared with NC-MMSE and NC-OMLSA, the average PESQ of CGBNN-5 increased by 16.22% and 12.02%, respectively, in four specific noise environments. The PESQ value of CGBNN-5, in this paper, achieved an averaged improvement of 3.8% over the state-of-the-art method, NC-DNN. According to the STOI results in Figure 3, the performance improvement of CGBNN-5 and NC-DNN is significantly better than that of NC-MMSE and NC-OMLSA. The CGBNN-5 achieved an averaged STOI improvement of 12.3% and 14.75% over NC-MMSE and NC-OMLSA, respectively, in the specific application scenario. Compared with NC-DNN, CGBNN-5 did not show great advantages, but it also achieved a 1.74% increase in STOI. Although the averaged STOI score of CGBNN-5 in 15 dB SNR is not the best, its scores in the range from –5 dB to 10 dB are still significantly better than the other three methods, that is, the proposed branchy model with prior noise information can obtain better speech quality and intelligibility than the other three methods in specific application scenarios. Furthermore, according to the spectrogram details of enhanced speech in Figure 4, it is clear that the DNN-based denoising methods (NC-DNN and CGBNN-5) achieve less speech distortion and retain less residual noise as compared with statistical-based approaches (NC-MMSE and NC-OMLSA). We also calculated the signal-to-distortion ratio (SDR) [39] to measure the speech distortion introduced by the denoising algorithms. Results show that the proposed CGBNN-5 obtains the enhanced SDR values of 6.74 dB in 0 dB SNR cases for four specific noise interferences. The enhanced SDR values of NC-DNN, NC-OMLSA, and NC-MMSE are lower under the same conditions, which are 6.05 dB, 6.30 dB, and 6.13 dB, respectively. For the evaluation of residual noise levels, the enhanced SNR of CGBNN is 4.85 dB, that is, higher than NC-DNN (4.12 dB), NC-OMLSA (3.29 dB), and NC-MMSE (2.63 dB). CGBNN-5 yields better speech quality and intelligibility than the NC-DNN model, while reducing the memory storage burden by 35%, as shown in Table 2.

4. Conclusions

In this paper, we investigated the effect of prior environment information for the performance improvement of DNN-based speech enhancement algorithms. An environment attention guided branchy neural network was proposed to cope with the noise interference problem in some known application scenarios. Compared with the DNN models without prior environment information, the idea of combining special paths with a common path in a branchy layer significantly improves denoising performance and model generalization ability. Moreover, the comparison experiments with the classic environment-aware methods show that the proposed branchy DNN model not only achieves better reconstructed speech quality and intelligibility, but also improves the storage burden and noise complementarity problem of separate DNN training. Therefore, the proposed method is more suitable for many speech products to deal with the noise interference problem in specific application scenarios. Although our branchy DNN model performs better than other environment-related methods, the complexity of the DNN model still increases with an increase of noise in application scenarios. The denoising performance of the proposed method in unseen noises is worse than that in seen noises. There are some ideas for future studies to further improve the denoising performance of the algorithm and overcome the increase of model size in complex noisy environments. Some traditional signal processing approaches could be adopted to reduce the burden of end-to-end DNN models. For example, the input signals could be decomposed into sub-bands to reduce the number of features that need to be reconstructed. The phase reconstruction problem needs to be considered in low SNR conditions to improve the enhanced quality of speech. The temporal relationship of speech signals is

helpful to improve the generalization ability of the DNN-based speech enhancement method. In our future research, we plan to explore these approaches in-depth to further improve our algorithm to achieve more robust noise reduction.

Author Contributions: L.Z. and M.W. contributed equally in conceiving the overall proposal, and critically reviewed and implemented the final revisions; L.Z. supervised all aspects of this DNN architecture, design, and realization of the experiments, collection, and analysis of the data, and writing of the manuscript; Q.Z. and M.L. critically reviewed and implemented the final revisions. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Basic Research Program grant funded by the Shenzhen Government (JCYJ20170412151226061, JCYJ20180507182241622).

Acknowledgments: The authors would like to thank the anonymous reviewer for their helpful advice to improve the quality of this paper.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

In this Appendix, we present the details on the derivation of gradient propagation in the proposed denoising neural network, and its structure is shown in Figure A1. In the training stage, the MSE objective function is used as the optimization target, which is given by the sum of the mean square loss of each frequency bins in every mini batch:

$$J_{MSE} = \frac{1}{N \cdot K} \sum_{t=1}^{N} \sum_{k=1}^{K} \left(\hat{Y}(k,t) - Y(k,t) \right)^2 \tag{A1}$$

where J_{MSE} is the optimization target, and K and N denote the feature size of each output frame and the number of frames in each mini-batch, respectively. $Y(k,t)$ is the ideal target for model learning and $\hat{Y}(k,t)$ represents the estimated LPS feature from the forward propagation of the proposed denoising model. Equations (A2) to (A6) present the detailed forward propagation process:

$$\begin{aligned} Z_{l-1} &= W_{l-1} X + b_{l-1} \\ A_{l-1} &= g(Z_{l-1}) \end{aligned} \tag{A2}$$

$$\begin{aligned} Z_{l,i} &= p_{n_i} * (W_{l,i} A_{l-1} + b_{l,i}) \\ A_{l,i} &= g(Z_{l,i}), i = 1, 2, \ldots, m \end{aligned} \tag{A3}$$

$$\begin{aligned} Z_{l,m+1} &= W_{l,m+1} A_{l-1} + b_{l,m+1} \\ A_{l,m+1} &= g(Z_{l,m+1}) \end{aligned} \tag{A4}$$

$$\begin{aligned} Z_{l+1} &= \sum_{i=1}^{m+1} (W_{l+1,i} A_{l,i} + b_{l+1,i}) \\ A_{l+1} &= g(Z_{l+1}) \end{aligned} \tag{A5}$$

$$\hat{Y} = W_{l+2} A_{l+1} + b_{l+2} \tag{A6}$$

where a middle hidden layer of the branchy model is indexed as l, and subscript i denotes its branch index. X and $\hat{Y}$ represent the concatenated noise-aware input features and enhanced LPS features, respectively. p_{n_i} is the estimated noise attention weight from the classification neural network. In the test stage, the extracted noise-aware features, X, are fed to the branchy model to conduct the above forward propagation with applying attention weights, which will produce the enhanced LPS features $\hat{Y}$ for speech reconstruction.

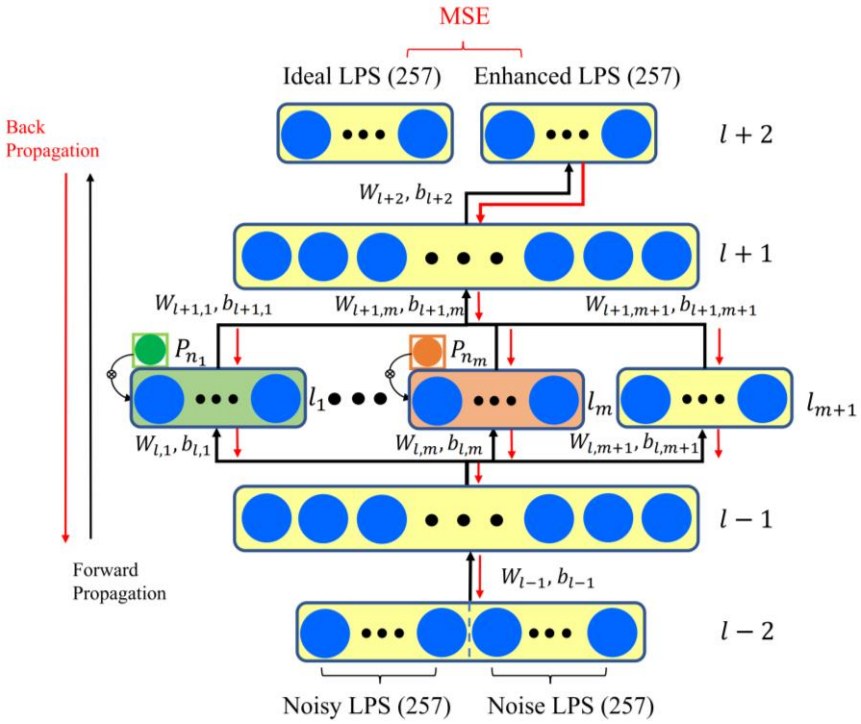

Figure A1. The block diagram of the proposed branchy denoising neural network.

After each forward propagation in one mini-batch, the gradients of MSE loss are calculated with chain rule and propagates from top to bottom layer by layer, as shown in Equations (A7) to (A12):

$$\begin{aligned} \mathrm{d}W_{l+2} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial W_{l+2}} \\ &= \frac{2}{N \cdot K} \sum_{t=1}^{N} \sum_{k=1}^{K} (\hat{Y}(k,t) - Y(k,t)) A_{l+1} \\ \mathrm{d}b_{l+2} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial b_{l+2}} \\ &= \frac{2}{N \cdot K} \sum_{t=1}^{N} \sum_{k=1}^{K} (\hat{Y}(k,t) - Y(k,t)) \end{aligned} \tag{A7}$$

According to Equation (A7), the gradients of the $(l+2)$-th layer (output layer) are not affected by the attention weights, but the hidden layers below it will be affected by the factor p_{n_i} as follows:

$$\begin{aligned} \mathrm{d}W_{l+1,i} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial W_{l+1,i}} \\ &= \frac{2}{N \cdot K} \sum_{t=1}^{N} \sum_{k=1}^{K} (\hat{Y}(k,t) - Y(k,t)) W_{l+2} * g'(Z_{l+1}) A_{l,i} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) g[p_{n_i} * (W_{l,i} A_{l-1} + b_{l,i})], i = 1, 2, \ldots, m \\ \mathrm{d}b_{l+1,i} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial b_{l+1,i}} \\ &= \frac{2}{N \cdot K} \sum_{t=1}^{N} \sum_{k=1}^{K} (\hat{Y}(k,t) - Y(k,t)) W_{l+2} * g'(Z_{l+1}) \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) \end{aligned} \tag{A8}$$

$$\begin{aligned} \mathrm{d}W_{l+1,m+1} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial W_{l+1,m+1}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) A_{l,m+1} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) g(W_{l,m+1} A_{l-1} + b_{l,m+1}) \\ \mathrm{d}b_{l+1,m+1} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial b_{l+1,m+1}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) \end{aligned} \tag{A9}$$

where the gradient value of $\frac{\partial J_{MSE}}{\partial \hat{Y}}$ is simplified as σ_J in the derivation process, and $*$ denotes the multiplication cross. Obviously, the gradients of $W_{l+1,i}$ in the first m branches are controlled by the attention factor p_{n_i}. When p_{n_i} is close to zero, the gradient of $W_{l+1,i}$ is close to zero. This gradient blocking effect of p_{n_i} is further back-propagated to the lower layer as derived in Equation (A10):

$$\begin{aligned} \mathrm{d}W_{l,i} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial A_{l,i}} \frac{\partial A_{l,i}}{Z_{l,i}} \frac{\partial Z_{l,i}}{W_{l,i}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,i} g'(Z_{l,i}) p_{n_i} * A_{l-1}, \quad i = 1, 2, \ldots, m \\ \mathrm{d}b_{l,i} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial A_{l,i}} \frac{\partial A_{l,i}}{Z_{l,i}} \frac{\partial Z_{l,i}}{b_{l,i}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,i} g'(Z_{l,i}) p_{n_i} \end{aligned} \tag{A10}$$

$$\begin{aligned} \mathrm{d}W_{l,m+1} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial A_{l,m+1}} \frac{\partial A_{l,m+1}}{Z_{l,m+1}} \frac{\partial Z_{l,m+1}}{W_{l,m+1}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,m+1} g'(Z_{l,m+1}) A_{l-1} \\ \mathrm{d}b_{l,m+1} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial A_{l,m+1}} \frac{\partial A_{l,m+1}}{Z_{l,m+1}} \frac{\partial Z_{l,m+1}}{b_{l,m+1}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) W_{l+1,m+1} g'(Z_{l,m+1}) \end{aligned} \tag{A11}$$

From Equation (A10), it is found that the similar blocking effect of attention factor on the first m branches also affects the gradient update of $W_{l,i}$ and $b_{l,i}$. That is, in the branchy layer of the proposed DNN model, the attention factor controls whether back propagation is turned on or not. However, according to Equations (A9) and (A11), the $(m + 1)$-th branchy layer without attention factor still updates the gradients in each back propagation and learns some general denoising characteristics. Furthermore, the $(l - 1)$-th hidden layer also learns the general denoising abilities in each gradient update, as shown in Equation (A12).

$$\begin{aligned} \mathrm{d}W_{l-1} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial A_{l,i}} \frac{\partial A_{l,i}}{Z_{l,i}} \frac{\partial Z_{l,i}}{A_{l-1}} \frac{\partial A_{l-1}}{\partial Z_{l-1}} \frac{\partial Z_{l-1}}{\partial W_{l-1}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) [\sum_{i=1}^{m} W_{l+1,i} g'(Z_{l,i}) p_{n_i} * W_{l,i} + W_{l+1,m+1} g'(Z_{l,m+1}) W_{l,m+1}] g'(Z_{l-1}) X \\ \mathrm{d}b_{l-1} &= \frac{\partial J_{MSE}}{\partial \hat{Y}} \frac{\partial \hat{Y}}{\partial A_{l+1}} \frac{\partial A_{l+1}}{\partial Z_{l+1}} \frac{\partial Z_{l+1}}{\partial A_{l,i}} \frac{\partial A_{l,i}}{Z_{l,i}} \frac{\partial Z_{l,i}}{A_{l-1}} \frac{\partial A_{l-1}}{\partial Z_{l-1}} \frac{\partial Z_{l-1}}{\partial b_{l-1}} \\ &= \sigma_J W_{l+2} * g'(Z_{l+1}) [\sum_{i=1}^{m} W_{l+1,i} g'(Z_{l,i}) p_{n_i} * W_{l,i} + W_{l+1,m+1} g'(Z_{l,m+1}) W_{l,m+1}] g'(Z_{l-1}) \end{aligned} \tag{A12}$$

Although the gradient expression of W_{l-1} and b_{l-1} still has the attention factor, the sum operation of p_{n_i} in all m special branches is one for each back-propagation, which means that it will not affect the gradient update of the $(l - 1)$-th layer.

References

1. Scalart, P.; Filho, J.V. Speech enhancement based on a priori signal to noise estimation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Atlanta, GA, USA, 7–10 May 1996; pp. 629–632.
2. Cohen, I.; Berdugo, B. Speech enhancement for non-stationary noise environments. *Signal Process.* **2001**, *81*, 2403–2418. [CrossRef]
3. Erkelens, J.S.; Hendriks, R.C.; Heusdens, R. Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1741–1752. [CrossRef]
4. Martin, R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Audio Speech Lang. Process.* **2001**, *9*, 504–512. [CrossRef]
5. Cohen, I.; Berdugo, B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.* **2002**, *9*, 12–15. [CrossRef]
6. Gerkmann, T.; Hendriks, R.C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 1383–1393. [CrossRef]

7. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE Trans. Audio Speech Lang. Process.* **1984**, *32*, 1109–1121. [CrossRef]
8. Hasan, M.K.; Salahuddin, S.; Khan, M.R. A modified a priori SNR for speech enhancement using spectral subtraction rules. *IEEE Signal Process. Lett.* **2004**, *11*, 450–453. [CrossRef]
9. Yuxuan, W.; Narayanan, A.; DeLiang, W. On Training targets for supervised speech separation. *IEEE Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef]
10. Williamson, D.S.; Yuxuan, W.; DeLiang, W. Complex ratio masking for monaural speech separation. *IEEE Trans. Audio Speech Lang. Process.* **2016**, *24*, 483–492. [CrossRef]
11. Tu, M.; Zhang, X. Speech enhancement based on Deep Neural Networks with skip connections. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 May 2017; pp. 5565–5569.
12. Ding, L.; Smaragdis, P.; Minje, K. Experiments on deep learning for speech denoising. In Proceedings of the International Speech Communication Association (INTERSPEECH), Singapore, 14–18 September 2014; pp. 2685–2689.
13. Xu, Y.; Jun, D.; Dai, L.-R.; Lee, C.-H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.* **2014**, *21*, 65–68. [CrossRef]
14. Xu, Y.; Jun, D.; Dai, L.-R.; Lee, C.-H. regression approach to speech enhancement based on deep neural networks. *IEEE Trans. Audio Speech Lang.* **2015**, *23*, 7–19. [CrossRef]
15. Gao, T.; Du, J.; Dai, L.-R.; Lee, C.-H. SNR-based progressive learning of deep neural network for speech enhancement. In Proceedings of the International Speech Communication Association (INTERSPEECH), San Francesco, CA, USA, 8–12 September 2016; pp. 3713–3717.
16. Dayana, R.; Jorge, L.; Antonio, M.; Luis, V. Deep speech enhancement for reverberated and noisy signals using wide residual networks. *arXiv* **2019**, arXiv:1901.00660.
17. Syu-Siang, W.; Yu-You, L.; Jeih-weih, H.; Yu, T.; Hsin-Min, W.; Shih-Hau, F. Distributed microphone speech enhancement based on deep learning. *arXiv* **2019**, arXiv:1911.08153.
18. Gao, T.; Du, J.; Dai, L.-R.; Lee, C.-H. Densely connected progressive learning for LSTM-based speech enhancement. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5054–5058.
19. Ke, T.; Jitong, C.; Deliang, W. Gated residual networks with dilated convolutions for monaural speech enhancement. *IEEE Trans. Audio Speech Lang.* **2019**, *27*, 189–198.
20. Kumar., A.; Dinei, F.; Kumar, A.; Dinei, F. Speech enhancement in multiple-noise conditions using deep neural networks. In Proceedings of the International Speech Communication Association (INTERSPEECH), San Francesco, CA, USA, 8–12 September 2016; pp. 3738–3742.
21. Yuma, K.; Kenta, N.; Yusuke, H. DNN-based source enhancement to increase objective sound quality assessment score. *IEEE Trans. Audio Speech Lang.* **2018**, *26*, 1780–1792.
22. Wu, J.; Hua, Y.; Yang, S.; Qin, H.S.; Qin, H.B. Speech Enhancement Using Generative Adversarial Network by Distilling Knowledge from Statistical Method. *Appl. Sci.* **2019**, *9*, 3396. [CrossRef]
23. Nidhyananthan, S.S.; Kumari, R.S.S.; Prakash, A.A. A review on speech enhancement algorithms and why to combine with environment classification. *Int. J. Mod. Phys. C* **2014**, *25*, 1430002. [CrossRef]
24. Choi, J.H.; Chang, J.H. On using acoustic environment classification for statistical model-based speech enhancement. *Speech Commun.* **2012**, *54*, 477–490. [CrossRef]
25. Chang, J.H. Noisy speech enhancement based on improved minimum statistics incorporating acoustic environment-awareness. *Digit. Signal Process.* **2013**, *23*, 1233–1238. [CrossRef]
26. Yuan, W.; Xia, B. A speech enhancement approach based on noise classification. *Appl. Acoust.* **2015**, *96*, 11–19. [CrossRef]
27. Li, R.; Liu, Y.; Shi, Y. ILMSAF based speech enhancement with DNN and noise classification. *Speech Commun.* **2016**, *85*, 53–70. [CrossRef]
28. Bingyin, X.; Changchun, B. Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification. *Speech Commun.* **2014**, *60*, 13–29.
29. Shi, W.; Zhang, X.; Zou, X. Deep neural network and noise classification-based speech enhancement. *Mod. Phys. Lett. B* **2017**, *31*, 1740096. [CrossRef]
30. Loizou, P.C. *Speech Enhancement: Theory and Practice*, 2nd ed.; CRC: Boca Raton, FL, USA, 2013; pp. 360–363.

31. Xie, F.; Compernolle, D.V. A family of MLP based nonlinear spectral estimators for noise reduction. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Adelaide, Australia, 8–12 May 1994; pp. 53–56.
32. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgren, N.L. *Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database*; National Institute of Standards and Technology: Gaithersburgh, MD, USA, 1988; pp. 1–79.
33. Varga, A.; Steeneken, H.J. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
34. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752.
35. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio Speech Lang.* **2011**, *19*, 2125–2136. [CrossRef]
36. Kingma, D.P.; Ba, J.L. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 1–13.
37. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; pp. 448–456.
38. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
39. Vincent, E.; Gribonval, R.; Fevotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang.* **2006**, *14*, 1462–1469. [CrossRef]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).