electronics

MDPI

Article

Learning Ratio Mask with Cascaded Deep Neural Networks for Echo Cancellation in Laser Monitoring Signals

Haitao Lang * and Jie Yang

School of Mathematics and Physics, Beijing University of Chemical Technology, Beijing 100029, China; jieyang@mail.buct.edu.cn

* Correspondence: langht@mail.buct.edu.cn or haitaolang@hotmail.com

Received: 23 April 2020; Accepted: 15 May 2020; Published: 21 May 2020

Abstract: Laser monitoring has received more and more attention in many application fields thanks to its essential advantages. The analysis shows that the target speech in the laser monitoring signals is often interfered by the echoes, resulting in a decline in speech intelligibility and quality, which in turn affects the identification of useful information. The cancellation of echoes in laser monitoring signals is not a trivial task. In this article, we formulate it as a simple but effective additive echo noise model and propose a cascade deep neural networks (C-DNNs) as the mapping function from the acoustic feature of noisy speech to the ratio mask of clean signal. To validate the feasibility and effectiveness of the proposed method, we investigated the effect of echo intensity, echo delay, and training target on the performance. We also compared the proposed C-DNNs to some traditional and newly emerging DNN-based supervised learning methods. Extensive experiments demonstrated the proposed method can greatly improve the speech intelligibility and speech quality of the echo-cancelled signals and outperform the comparison methods.

Keywords: echo cancellation; speech enhancement; speech separation; deep neural networks (DNNs); cascaded DNNs (C-DNNs); laser monitoring

1. Introduction

As an emerging monitoring technology with great potential and many natural advantages, such as non-contacted listening, high concealment, and not susceptible to electromagnetic interference, laser monitoring has become a necessary technology and tool for public security departments, military to conduct investigations, forensics, and intelligence acquisition [1,2]. The basic principle of laser monitoring technology is sketched in Figure 1. The laser (red dotted line) emitted by the listening system hits the object (e.g., a flower in this sketch) around the person being monitored. Since the surface of the object is easily affected by the surrounding sound waves (purple wavy line) and generates subtle vibrations, the laser (red dotted line) reflected from the surface of the object contains the oscillation information of the indoor speech wave and is received by the receiving device. Finally, the speech signal is reconstructed after demodulation, filtering, power amplification, and other steps. In the context of laser monitoring, the signals acquired by the listening system are usually affected by ambient light sources, atmospheric noise, thermal noise, etc. With the development of technology, these noises can usually be effectively suppressed by improving the optical path and circuit, or by adopting signal processing methods such as spectral subtraction [3,4], Wiener filtering [5–7], etc. By studying and analyzing the actual laser monitoring signals, the researchers also found that the monitored speech is often subject to echo interference, especially those obtained in relatively large meeting rooms. As shown in Figure 1, the signal being monitored on the detection target (e.g., a flower) is usually

a superposition of the original speech (e.g., a talk between two people) and echoes (solid blue line), which have undergone multiple reflections by walls, interior furniture, etc. Due to the presence of echoes, the original speech is distorted, resulting in a decline both in speech quality and in intelligibility. To effectively identify valuable information, it is very important to eliminate the echoes that are mixed in the laser monitoring signal.

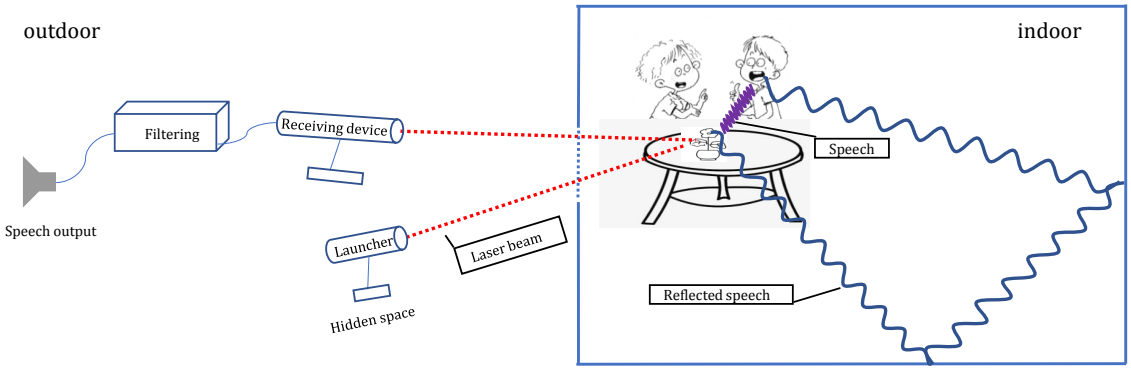


Figure 1. Echo generation in laser monitoring scene.

Unfortunately, cancellation of echoes in laser monitoring signals is not a trivial task. Unlike white noise, the echoes in the time–frequency (T-F) domain are unstable, limiting the use of the traditional methods, such as spectral subtraction [3,4] and Wiener filtering [6,7]. Unlike factory noise, babble noises, etc., due to high correlation between delayed signal and target signal, the feature of the echoes in the T-F domain is highly similar to that of the original speech signal, making it difficult to be solved by statistical model-based methods [7] and even newly emerging neural networks-based methods. In addition, an actual laser monitoring scene, usually a room, often contains many reflectors. These reflectors are usually different materials and different distances from the laser detection target, resulting in various attenuations and time delays of the reflected echo signals. The methods based on adaptive filtering [8–10], which are widely used in the scenarios of echo cancellation for voice telephony and video conference, are not suitable for the echo cancellation of laser monitoring, because the method of adaptive filtering requires a reference signal, while, in laser monitoring domain, the only signal available is the one signal, and the reference signal cannot be obtained.

To address above challenges, we regard the echo cancellation in laser monitoring signals as a speech separation problem and deal with it using the computational auditory scene analysis (CASA) method [11]. The main advantage of CASA is that this method incorporates the auditory perception mechanism without assuming any properties or models of the noise. CASA masks (i.e., eliminates or cancels) the noise from a noisy speech by estimating the value of the mask in the T-F domain. Since the accuracy of the mask is the key point for the quality of the recovered speech, many well-designed binary masks [12,13] or ratio masks [13–17] are proposed. In Ref. [16], Wang et al., proved that the ideal ratio mask (IRM) can obtain a better performance on noise reducing than the ideal binary mask (IBM). In Ref. [15], Liang et al., concluded that the optimal ratio mask (ORM) can further improve the signal-to-noise ratio (SNR) over the IRM by the theoretical analysis. Recently, Bao et al., proposed a new ratio mask named corrected ratio mask (CRM) [17] and proved it performs better than the conventional ratio masks and other series of enhancement algorithms. Researchers have proposed various algorithms to estimate the binary/ratio masks, such as Bayesian method [18], joint dictionary learning [19], convex optimization [20], support vector machine (SVM) [21], and neural networks (NNs). Among NNs-based methods, the ability of recurrent neural network (RNN) [22] to capture the long-term contextual information turned out to be very limited. Although the ability of a long short-term memory recurrent neural network (LSTM-RNN) [23,24] to catch information is enhanced, it generates a huge computational cost. In contrast, the newly emerging deep neural

network (DNN) [14,16,17,25,26] has been proved to be a useful tool to estimate the ratio mask due to its excellent learning skill.

In this article, we formulate the noisy speech and the echoes in laser monitoring signals using an additive noise assumption. Both theoretical analysis and experimental verification prove that it is a reasonable assumption for the problem of echoes of cancellation in laser monitoring signals. Consequently, we propose cascaded DNNs (C-DNNs) to learn the ratio mask, as shown in Figures 2 and 3, which are fundamentally different from existing methods that include two or more neural networks [27–29]. In Ref. [27], the former network is only used to ensure initial values for the latter network. In Ref. [28], the latter DNNs use the output of the previous DNNs as input. Two DNNs use different feature calculation methods, but the training targets are the same. In Ref. [29], during the training process, the input of each DNNs contains the original features of the mixed signal, and the output of each DNNs is a ratio mask calculated by the clean signal and the mixed signal, which also means that the training targets are the same. In contrast, in the training stage of the proposed C-DNNs, the first DNNs is trained using the auditory feature and a specific ratio mask which are both extracted or calculated based on the noisy speech $y(t)$. Then, the output of the first DNNs is used to calculate the estimated speech $y'(t)$ and to update both the feature and target for the purpose of improving the final estimation accuracy. In the echo cancellation stage (i.e., enhancing stage), the noisy speech feature is processed by the well-trained DNN model to predict the ratio mask for reconstructing the clean speech signal. In this stage, since phase of clean speech is not accessible and the observed phase may be distorted by the echoes, to decrease the inconsistency between the recovered speech and the corrupted phase, Griffin–Lim iterative phase enhancement algorithm [30] and overlap-add (OLA) method [31] are employed to re-synthesize the enhanced time-domain signal.

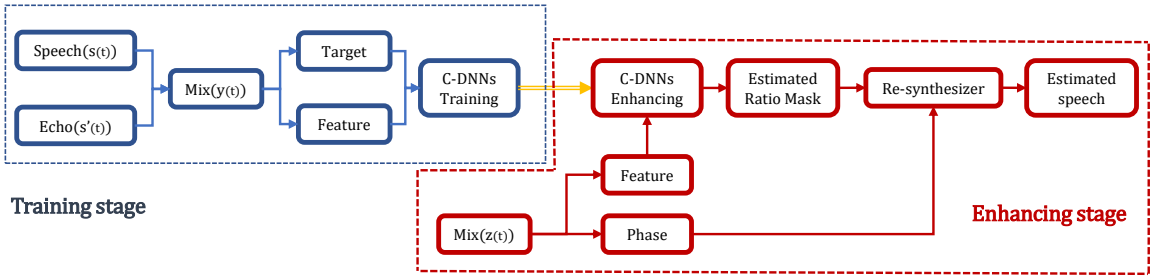
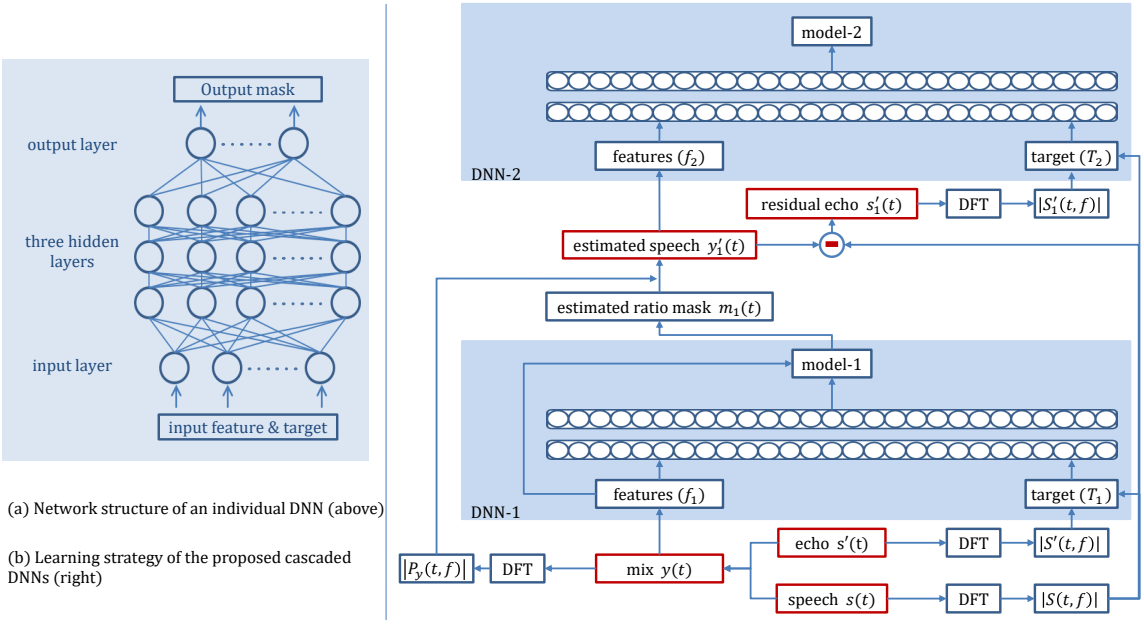


Figure 2. A block diagram of the proposed C-DNNs based echo cancellation.

The main contributions of this study are two-fold.

1. For echo cancellation in laser monitoring problem, we formulate it as a simple but effective additive echo noise model. Armed with this understanding of echoes, which are regarded as noise, we propose cascaded DNNs (C-DNNs) to learn the ratio mask. The experiments validate that the proposed C-DNNs can update the training target to improve the estimation accuracy of the ratio mask.
2. We construct a new database that includes rich speeches and are ready to release it to the speech-enhancement/separation research community.

The rest of the article is organized as following. In Section 2, we present a detailed description of the proposed methodology, including problem formulation (Section 2.1), the acoustic feature used in our study (Section 2.2), typical training targets (Section 2.3), and the proposed cascaded DNNs (Section 2.4). The experimental data, protocol, and evaluation metric are introduced in Section 3. In Section 4, we analyze the effect of the intensity attenuation (Section 4.1), the time delay (Section 4.2), and the training targets (Section 4.3) on the recovered speech intelligibility and quality. To further validate and evaluate the proposed C-DNNs, we compare it to some traditional and state-of-the-art methods (Section 4.4). In Section 5, we conclude our study and explore the strengths and weaknesses of the proposed methodology. We also outline the future work.



(a) Network structure of an individual DNN (above)

(b) Learning strategy of the proposed cascaded DNNs (right)

Figure 3. Network architecture and learning strategy of the proposed C-DNNs.

2. Methodology

2.1. Problem Formulation

In this study, we utilize an additive noise model to represent the noisy signal distorted by the echoes. Hence, the mix signal formed by the original speech and the spatial reflections (i.e., echoes) can be expressed as

$$y(t) = s(t) + s'(t) \tag{1}$$

where t denotes a time sample; $s(t)$ is the clean signal we need to estimate; $s'(t)$ is the attenuated and delayed signal, i.e., the echoes; and $y(t)$ denotes the mixed signal consisting of the original signal $s(t)$ and the echo signal $s'(t)$.

It should be pointed out that the reverberation and echo are produced for similar reasons, while the main difference being the delay. Usually, the reverberation delay is less than 0.1 s, thus the human ear cannot distinguish the original and the delayed speech. The delay of the echo is longer than 0.1 s, and the human ear can distinguish one or more successive echoes similar to the original speech signal. The existing methods commonly use the convolution model to represent the reverberation [26]. Based on our observation of laser monitoring signals, this article uses an additive noise model to formulate the noisy signal and the echoes. Although this model appears to be rude at first glance, the simplification of this model has facilitated the design of C-DNNs, allowing us to continually improve the estimation of the ratio mask by continuously updating the training objectives. The experiments demonstrate the feasibility of our strategy.

2.2. Acoustic Features

Acoustic features as input of learning machines play important roles in supervised learning. When amplitude modulation spectrogram (AMS) is invoked as a feature for speech enhancement, it can effectively improve speech intelligibility [32]. Hermansky et al., demonstrated that the relative spectral transformed perceptual linear prediction coefficients (RASTA-PLP) improves the performance of a recognizer in presence of convolutional and additive noise [33]. Mel-frequency cepstral coefficients (MFCC) [34] are calculated by warping the power spectrum of the signal to a 64-channel Mel-scale

and are widely used in the field of speech recognition and speech enhancement. Later, Shao et al., proved that the Gammatone frequency cepstral coefficients (GFCC) derived from an auditory filter bank enhance the auditory features and estimated their reconstruction uncertainties for better speaker likelihood calculation [35]. Through feature selection using group Lasso, the study recommended a complementary feature set including AMS, RASTA-PLP, MFCC, and GFCC, which is better than a single feature, which has used in controlled trials many times [16].

Following the existing study, in this article, the proposed C-DNNs are trained using the extracted auditory feature which is a concatenation of a set of complementary features, including 15-dimensional AMS, 13-dimensional RASTA-PLP, 31-dimensional MFCC, and 64-dimensional GFCC. To further augment the useful information of acoustic signals, in our implementation, the deltas (derivatives) of the feature and the frame expansion are used to the compound feature. Therefore, for each time frame, the final input feature dimension is 1230, i.e., $2 \times (15 + 13 + 31 + 64) \times 5$.

2.3. Training Targets

We introduce several existing training targets into the proposed C-DNNs to learn optimal ratio mask. All these ratio masks have been proved to be able to obtain good performance in speech enhancement/separation tasks. In our paper, we try to present a comprehensive evaluation of these targets for the purpose of finding the one that is most suitable for the task of echo cancellation in laser monitoring signals.

2.3.1. Ideal Ratio Mask (IRM)

IBM [13] is considered as a hard decision because it corresponds to value 0 dominated by noise or value 1 dominated by speech in each T-F unit. The drawback of using the binary mask is that it often wrongly removes the background noise in a weak T-F unit dominated by speech and this seriously affects the hearing quality. IRM [16] could be considered as a soft decision because masks values smoothly vary within 0–1. IRM is defined as

$$IRM(t,f) = (\frac{S(t,f)^2}{S(t,f)^2 + S'(t,f)^2})^{\beta} \tag{2}$$

where $S(t,f)$ and $S'(t,f)$ are the Discrete Fourier Transform (DFT) of $s(t)$ and $s'(t)$, respectively, at time t and frequency f. Thus, $S(t,f)^2$ and $S'(t,f)^2$ denote the speech energy and echo energy within a particular T-F unit, respectively. The tunable parameter β scales the mask, and is generally chosen to be 0.5. With the square root, IRM preserves the speech energy with each T-F unit, under the assumption that $S(t,f)^2$ and $S'(t,f)^2$ are uncorrelated. This assumption holds well for additive noise, which matches the model we propose in Section 2.1.

2.3.2. Spectral Magnitude Mask (SMM)

SMM (which is named as FFT-MASK in [16]) is defined on the Short-Time Fourier Transform (STFT) magnitudes of clean speech and noisy speech.

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} \tag{3}$$

where $|S(t,f)|$ and $|Y(t,f)|$ represent spectral magnitude of clean speech signal and mixed signal within a T-F unit, respectively. Unlike IRM with range [0,1], SMM is not upper-bounded by 1. To fit an SMM suitable to the proposed C-DNNs whose output units are sigmoid functions, we truncate the range of SMM to [0,1] in this study.

2.3.3. Corrected Ratio Mask (CRM)

CRM is a newly proposed ratio mask by Bao et al. [17], which is a correction to IRM. Considering that, in the practical speech enhancement application, the ratio mask should have a high value (i.e., near 1) in the speech T-F units so that it can preserve the speech components as much as possible, while the ratio mask needs a lower value (i.e., near 0) to mask the background noise in the noise T-F units. CRM utilizes the inter-channel correlation (ICC) between speech and noise to adaptively reallocate the power ratio of speech and noise during the construction of the ratio mask.

$$CRM(t,f) = \frac{\rho_s(t,f) \times P_s(t,f)}{\rho_s(t,f) \times P_s(t,f) + \rho_{s'}(t,f) \times P_{s'}(t,f)} \tag{4}$$

where $P_s(t,f)$ and $P_{s'}(t,f)$ indicate the power of the clean signal and echo of the tth frame in the fth channel, respectively. The adaptive factor $\rho_s(t,f)$ is the normalized cross-correlation between clean and mix signal power. Another adaptive factor $\rho_{s'}(t,f)$ is the normalized cross-correlation between echo and mix signal power. The $\rho_s(t,f)$ and $\rho_{s'}(t,f)$ factors refer to the percentage of the clean speech and noise components within the noisy speech signal, respectively.

2.4. Cascaded DNNs (C-DNNs)

The supervised learning based echo cancellation can be formulated as the process that maps from acoustic features of $y(t)$ in a time–frequency (T-F) mask of clean target signal $s(t)$. In our study, we utilize the proposed C-DNNs as the mapping function from the acoustic feature of noisy speech to the ratio mask of the clean signal.

2.4.1. Network Architecture

The proposed C-DNNs are a stack of ensemble learning machines, as shown in Figure 3. They contain two consecutive DNNs, each of which can perform an independent learning process. As for each individual DNN, the theory and practice of deep learning shows that the number of hidden layers and the units at each hidden layer will affect the performance of the network. In the case where the number of layers or units is not enough, the learning ability of the network will be insufficient, while, if the number of layers or units are excessive, the network's ability of generalization will not perform well, and it is easy to cause over-fitting. It was designated by Xu et al., that DNN with three hidden layers achieved superior performance on most speech enhancement tasks [36]. Following the existing experience, in our study, for each individual DNN, a five-layer structure is expected to be adopted, including one input layer, one output layer, and three hidden layers. The number of input layer neurons is 1230, equal to the dimension of extracted acoustic features (cf. Section 2.2). The number of output layer neurons is 64 which is determined by the dimension of the estimated ratio mask. As for hidden layers, each of them consists of 2048 rectified linear neurons (ReLU). Considering the value of training target changes in the range [0,1], we use sigmoid activation functions in the output layer and linear activation function for other layers.

2.4.2. Learning Strategy

We apply two successive DNNs to continuously update features and training target to improve the accuracy of the estimated mask. As shown in Figure 3, we take IRM as an example to outline the proposed C-DNNs structure in detail. For the first DNN, clean speech $s(t)$ and echo $s'(t)$ perform DFT to calculate the training target T_1, the mixed signal $y(t)$ is used to calculate complementary features f_1 and provide phase information $P_y(t,f)$ for estimated mix signal $y_1'(t)$ with residual echo, and then the features f_1 and targets T_1 are used to train the network. Model-1 is the product of DNN-1, and then the $m_1(t)$ estimated by Model-1 from train data and $P_y(t,f)$ re-synthesize the estimated speech $y_1'(t)$ together to prepare for the second network. Unlike the former, before the DNN-2 is

trained, the residual echo $s_1'(t)$ is calculated by $s(t)$ and $y_1'(t)$, and the training target T_2 is also updated by $s_1'(t)$ and $s(t)$. Eventually, Model-2 can be trained through T_2 and features f_2 calculated by $y_1'(t)$.

3. Experimental Data, Protocol, and Evaluation Metrics

3.1. Experimental Data

To validate and evaluate the performance of the proposed method, we collected a real human voice database that is longer than 40 h. The main voice sources include news, radio, TV, and movie dialogue; indoor and outdoor lectures and interviews; and so on. Speakers are both male and female and of various ages. In our experiments, we divided the dataset into training set and test set. The voice data of training set was up to 30 h, and was composed of 32 speakers, each of whom has 200–500 clean utterances. The voice duration of the test set was up to 10 h, which includes 20 speakers, each of whom has 25 clean utterances. All corpora were re-sampled to 16 kHz, and converted to the T-F units with the frame length set to 20 ms and the frame shift set to 10 ms.

Compared to the original speech, since the propagation process experienced by the echo is more complicated, the energy attenuation of echo due to reflection and absorption is greater. Since we regard the echo as noise in this article, we also utilize (or borrow) the definition of SNR (Equation (5)) to represent the degree of echo attenuation. Since we utilize the energy of the original speech as a reference, a larger SNR represents a greater degree of attenuation, and a smaller SNR represents a stronger echo. In our experiments, we generated multiple attenuation levels, from 5 dB to 15 dB, corresponding to the energy of the echo being attenuated from about 50% to 20% of the original speech energy. In addition, to simulate the time delay of echo, we also generated multiple delay levels from 0.1 s to 0.3 s. In the training stage, we mixed randomly selected echo signals $s'(t)$ with different attenuation levels and different delay levels with clean speech signals $s(t)$ to generate noisy signals $y(t)$.

$$SNR = 10 \times \log_{10}\left(\frac{s^2(t)}{s'^2(t)}\right) \tag{5}$$

3.2. Experimental Protocol

As described in Section 2.4, for each individual DNN, a five-layer structure was used, including one input layer, three-hidden layers, and one output layer. The input layer consists of 1230 neurons equal to the dimension of extracted acoustic features. Each hidden layer consists of 2048 rectified linear neurons (ReLU). The output layer consists of 64 neurons equal to the dimension of the estimated ratio mask. As the target changes in the range [0,1], we use sigmoid activation functions in the output layer and linear activation functions for the input layer and hidden layers. In the training process, each DNN in the proposed cascaded structure was optimized by the minimum mean squared error. The number of epochs for back propagation training was set to 40, and the batch size was set to 32. We chose the learning rate as decreasing linearly from 0.008 to 0.0001, and the scaling factor for adaptive stochastic gradient descent was set to 0.0015. All algorithms were implemented and run using Matlab. All experiments were run on an AMD Processor 2.4 GHz machine, with 16 GB RAM and Windows Server 2008 R2 standard operation system.

Because singular sample that is away from the correct features of the signal may lead to increasing the training time and also resulting non-convergence situations, we calculated the standard score of the reprocessed training feature, so that the average value of the feature data was 0 and the standard deviation was 1.

3.3. Evaluation Metric

For the purpose of evaluating the performance of the proposed method, we utilized short-time objective intelligibility (STOI) [37] and perceptual evaluation of speech quality (PESQ) [38] to evaluate the speech intelligibility and quality, respectively. STOI measures the correlation between the short-time

temporal envelopes of a reference (clean) utterance and a separated utterance (see Equation (6)). The value range of STOI is typically between 0 and 1, which can be interpreted as percent correct.

$$STOI = \frac{1}{JM} \sum_{j,m} \frac{(s_{j,m} - \mu_{s_{j,m}})^{\mathrm{T}} (\bar{y}'_{j,m} - \mu_{\bar{y}'_{j,m}})}{\left\| (s_{j,m} - \mu_{s_{j,m}}) \right\| \left\| (\bar{y}'_{j,m} - \mu_{\bar{y}'_{j,m}}) \right\|} \tag{6}$$

where J is the number of one-third octave bands, M represents the total number of frames, $s_{j,m}$ denotes the short-time temporal envelope of the clean speech, and $\bar{y}'_{j,m}$ means the normalized and clipped version of the processed speech. μ refers to the sample average of the corresponding vector.

PESQ applies an auditory transform to produce a loudness spectrum and compares the loudness spectra of a clean reference signal and a separated signal to produce a score in a range of -0.5 to 4.5, corresponding to the prediction of the perceptual mean opinion score (MOS) s(ee Equation (7)), which is dimensionless.

$$PESQ = 4.5 - 0.1 d_{SYM} - 0.0309 d_{ASYM} \tag{7}$$

The combination of symmetric disturbance (d_{SYM}) and one asymmetric disturbance (d_{ASYM}) makes a good balance between accuracy of prediction and ability to generalize.

4. Experimental Results and Analysis

4.1. The Effect of Echo Intensity

In this experiment, we first studied the ability of the proposed method to cope with different echo intensities. The metrics STOI and PESQ after echo-cancellation by the proposed C-DNNs with various training targets for different attenuation levels are listed in Tables 1 and 2, respectively. It shows that both the STOI and the PESQ of signals after echo-cancellation have been improved compared to those of unprocessed signals. For instance, at 5 dB attenuation levels, the STOI increases from 0.830 (unprocessed signal) to 0.856 (IRM), 0.866 (SMM), and 0.866 (CRM), while the PESQ increases from 1.97 (unprocessed signal) to 2.18 (IRM), 2.21 (SMM), and 2.20 (CRM). These experiments validate that C-DNNs can not only eliminate echoes, but also improve speech intelligibility and quality.

Table 1. Effect of echo attenuation on STOI.

$SNR(dB)$	5	6	7	8	9	10	11	12	13	14	15	avg.
Unprocessed	0.830	0.863	0.875	0.890	0.906	0.921	0.933	0.941	0.952	0.956	0.964	0.912
IRM	0.856	0.888	0.912	0.921	0.934	0.946	0.951	0.956	0.959	0.966	0.972	0.933
SMM	0.866	0.898	0.916	0.926	0.940	0.948	0.954	0.960	0.962	0.968	0.972	0.937
CRM	0.866	0.898	0.914	0.924	0.939	0.948	0.955	0.960	0.962	0.967	0.973	0.938

Table 2. Effect of echo attenuation on PESQ.

$SNR(dB)$	5	6	7	8	9	10	11	12	13	14	15	avg.
Unprocessed	1.97	2.05	2.12	2.14	2.30	2.33	2.42	2.50	2.60	2.64	2.78	2.35
IRM	2.18	2.32	2.46	2.54	2.69	2.78	2.85	2.92	2.96	3.08	3.12	2.72
SMM	2.21	2.35	2.49	2.58	2.72	2.82	2.88	2.94	3.00	3.10	3.14	2.75
CRM	2.20	2.34	2.47	2.56	2.73	2.82	2.89	2.95	3.00	3.09	3.14	2.75

When focusing on the effect of echo intensity on speech intelligibility, it shows that the more severe is the echo attenuation, the less is the impact on speech intelligibility. Thus, as the SNR increases from 5 dB to 15 dB, which means that the attenuation is getting bigger (that is, the echo intensity is getting smaller and smaller), the STOI climbs from about 0.85 to 0.97 for unprocessed and echo-cancelled signals, respectively. This phenomenon can be further analyzed from the gain aspect, as shown

in Figure 4a. It shows that, as the echo attenuation increases (the SNR increases), the gain of STOI becomes smaller and smaller. As mentioned above, this reflects the smaller is the echo intensity, the smaller is the influence on speech intelligibility. This result also proves that C-DNNs are more helpful for improving the speech intelligibility of stronger echo signals.

In terms of speech quality, Table 2 shows that, the more severe is the echo attenuation (small echo intensity), the smaller is the impact on speech quality (PESQ), thus both the unprocessed and echo-cancelled signals have higher PESQ at 15 dB attenuation levels. Analysis of Figure 4b reveals that C-DNNs is more advantageous to improve the speech quality of signals at moderate attenuation levels.

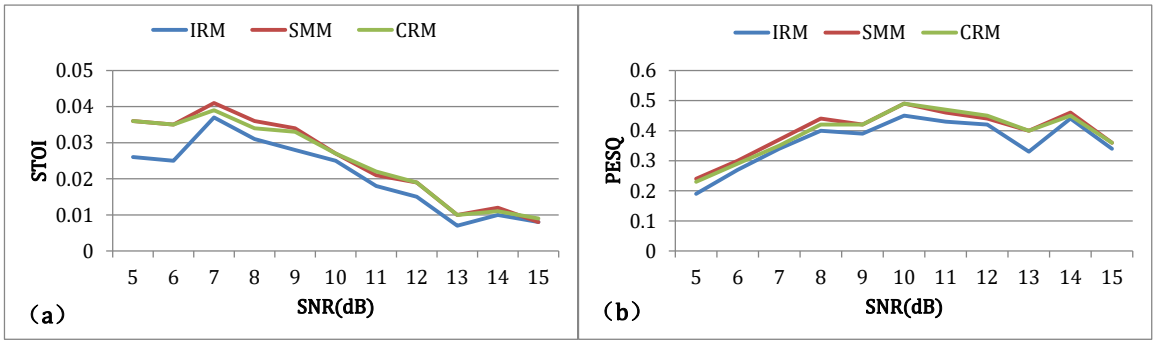


Figure 4. Gain of (**a**) STOI and (**b**) PESQ: Echo-removed speech compared to those unprocessed.

4.2. The Effect of Echo Delay

For the purpose of investigating the effect of time delay on echo cancellation, we generated the training data with continuous time delay from 0.1 s to 0.3 s. It should be mentioned that, to rule out the effect of attenuation, in these experiments, we applied a fixed attenuation level of 8 dB. The two metrics STOI and PESQ after echo-cancellation by the proposed C-DNNs with various training targets for several typical time delays are listed in Tables 3 and 4.

As the delay time becomes longer, the speech intelligibility (STOI) tends to decrease for both unprocessed and echo-cancelled speech. Speech quality (PESQ) has also shown a similar trend for echo-cancelled speech. However, for unprocessed signals, there is no obvious dependence between speech quality and echo delay. This result is reasonable. On the one hand, it shows that the echo of any time delay will lead to the degradation of speech quality. On the other hand, it shows that the proposed C-DNNs can significantly improve the signal quality of a short delay echo interference. For instance, using the CRM target, when delay is 0.281 s, the PESQ is increased from 2.52 to 2.84, while, at 0.188 s, the PESQ is increased from 2.51 to 2.91. This result is meaningful, because, in most practical laser monitoring situations, the scene space is generally not large, the difference between the signal and the reflected signal is relatively small, and the delay time is relatively short and more likely to happen.

Table 3. Effect of time delay on STOI.

$DELAY(s)$	**0.188**	**0.206**	**0.225**	**0.245**	**0.263**	**0.281**
Unprocessed	0.948	0.945	0.942	0.940	0.938	0.937
IRM	0.956	0.954	0.954	0.951	0.948	0.946
SMM	0.959	0.958	0.956	0.953	0.950	0.947
CRM	0.959	0.959	0.957	0.953	0.951	0.948

Table 4. Effect of time delay on PESQ.

$DELAY(s)$	0.188	0.206	0.225	0.245	0.263	0.281
Unprocessed	2.51	2.50	2.50	2.51	2.52	2.52
IRM	2.89	2.90	2.89	2.86	2.83	2.81
SMM	2.91	2.92	2.90	2.88	2.86	2.84
CRM	2.91	2.93	2.91	2.88	2.87	2.84

4.3. The Effect of Training Target

As aforementioned, different training targets show different processing ability, for intensity attenuation and time delay of echo. Generally speaking, targets of SMM and CRM perform equally well, slightly better than the IRM, in terms of both speech intelligibility and speech quality. In this section, we further analyze the performance of the targets from the perspective of the spectrogram. The spectrograms of test data No. 461 (randomly selected) processed by various training targets are shown in Figure 5. We mark two typical differences at 0.5 s and 3.3 s in the spectrogram between clean, unprocessed, and after C-DNNs processed speeches using three targets, with the red and blue ellipses, respectively. Comparing with clean signal (Figure 5a), echo can be clearly seen in the unprocessed signal (Figure 5b). It is also shown that the energy characteristic of echo noisy signal is similar to that of the clean signal. The spectrograms processed by C-DNNs with SMM (Figure 5d) or CRM (Figure 5e) show similar performance. At 0.5 s, the results processed by them almost no longer contain echo information. At 3.3 s, there remains a small amount of low frequency residual echoes. The performance of IRM is not as good as CRM and SMM. It shows that, at 0.5 s, a small amount of echo is not eliminated, and there are also more residual echoes at 3.3 s. Nevertheless, compared with the unprocessed signal, it can be proved that the proposed C-DNNs method can effectively remove the echoes regardless of various training targets we used.

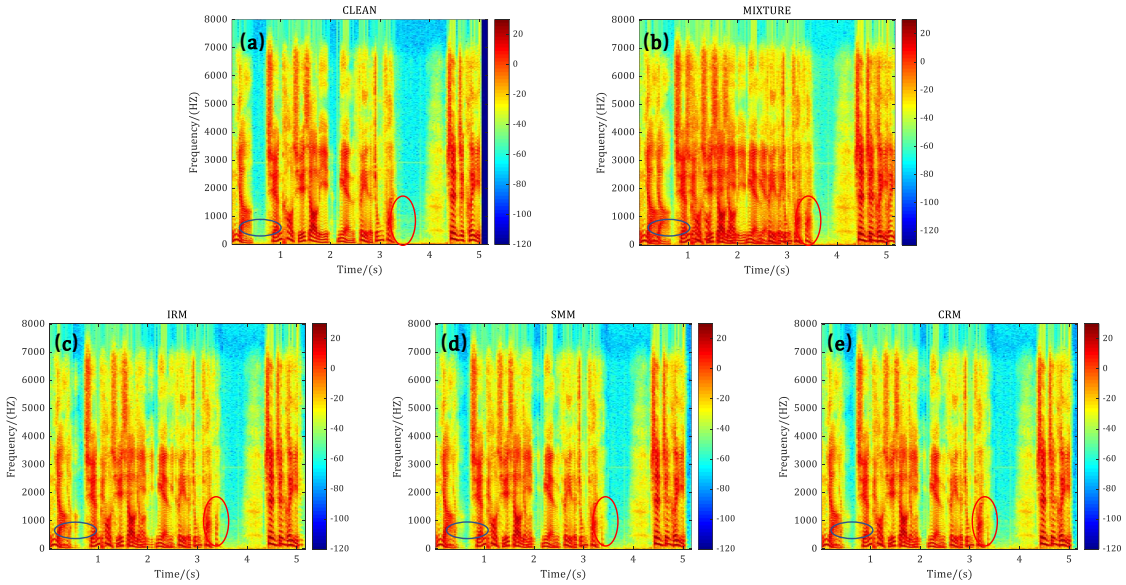


Figure 5. The spectrograms of test data No. 461 obtained by AC-DNNs with different training targets. (**a**) clean signal. (**b**) noisy signal. (**c**) processed with IRM target. (**d**) processed with SMM target. (**e**) processed with CRM target.

4.4. Comparison with Existing Methods

To our knowledge, since there was no existing method specifically designed to solve the problem of echo cancellation in laser monitoring signals, we tested the performance of traditional minimum mean-square error log-spectral amplitude estimator (Log-MMSE) [6,7,39] and two recently proposed

DNN-based methods [20,26] that performed well on the tasks of reverberation and noise cancellation, and compared them with the proposed C-DNNs.

LOG-MMSE is a method of minimizing the mean-square error between $\log \hat{A}_k$ (log power spectrums of estimated speech) and $\log A_k$ (log power spectrums of clean speech) by using conditional probability, which is expressed as

$$\hat{A}_k = \frac{\xi_k}{\xi_k + 1} \exp\left\{ \frac{1}{2} \int_{v_k}^{+\infty} \frac{e^{-t}}{t} dt \right\} Y_k \tag{8}$$

where ξ_k and v_k indicate the prior and posterior SNR, respectively, calculated at frequency bin k of the noisy speech Y_k.

In Ref [20], Bao et al., proposed a new ratio mask CRM and utilized a five-layer DNN to learn the optimal mask. The experiments show that it performed better than the traditional ratio masks. In our study, we reproduced it and termed it as a single DNN with CRM (S-DNN(CRM)) (we use "single" to reflect the structural difference between this network and the proposed cascaded DNNs). In Ref. [26], Zhao et al., proposed a two-stage DNNs to de-noise and de-reverberation successively. The IRM was used as the training target to train two individual networks, i.e., de-noising network and de-reverberating network respectively. In our experiments, we reproduced its de-noising network and termed it as a single DNN with IRM (S-DNN(IRM)). In addition, for the sake of completeness, we also implemented S-DNN(SMM), i.e., a single DNN with SMM ratio mask.

We compared all these methods to the proposed C-DNNs. The metrics of STOI and PESQ of echo-cancelled signals by the comparative methods are shown in Figure 6. It shows that the method of LOG-MMSE cannot cancel the echo, and even causes a drop in STOI and PESQ. This result is reasonable because, as aforementioned, echoes in laser monitoring signal are a typical non-stationary noise, while the main drawback of LOG-MMSE estimator is that it is difficult to quickly and efficiently track and accurately estimate non-stationary noise.

All DNN based methods, whether S-DNN or C-DNNs, can effectively eliminate the echoes, which are also consistent with our expectations. In contrast, the methods based on C-DNNs outperform those based on S-DNN in various training targets. It shows that, when using the same training target, e.g., CRM, S-DNN(CRM) raises STOI from 0.9147 to 0.9371, and C-DNNs(CRM) further increases it to 0.9396. Similarly, S-DNN raises PESQ from 2.51 to 2.87, and C-DNNs further improves it to 2.93.

The spectrograms of test data No. 496 (randomly selected) processed by various methods is shown in Figure 7. Observing the clean and unprocessed spectrogram, it clearly shows that there are obvious echoes at 2.2 s and 3.1 s. When comparing the spectrogram of clean speech to those after processed by different methods, it can be seen that LOG-MMSE effectively eliminates the echo at 2.2 s but fails to eliminate the noise at 3.1 s. This result is explained again that LOG-MMSE cannot efficiently track and accurately estimate non-stationary echo noise in laser monitoring signals. Almost all DNN based methods can effectively eliminate echo at 2.2 s and largely eliminate echo at 3.1 s. Focusing on echo cancellation at 3.1 s, it can be seen that S-DNN (SMM) performs better than S-DNN (IRM) [26] and S-DNN (CRM) [20]. When comparing S-DNN (SMM) to C-DNNs (SMM), we find that the latter almost completely eliminates the echo, while the former still has a small amount remaining. Although the performance of S-DNN (CRM) is relatively poor at 3.1 s, when C-DNNs (CRM) is adopted, the echo is almost completely eliminated.

The above comparison experiments fully demonstrated the effectiveness and feasibility of the proposed C-DNNs method in the echo cancellation problem and proved that the structure of C-DNNs can learn a more optimized mask than the single DNN structure which is commonly used by the existing methods [20,26].

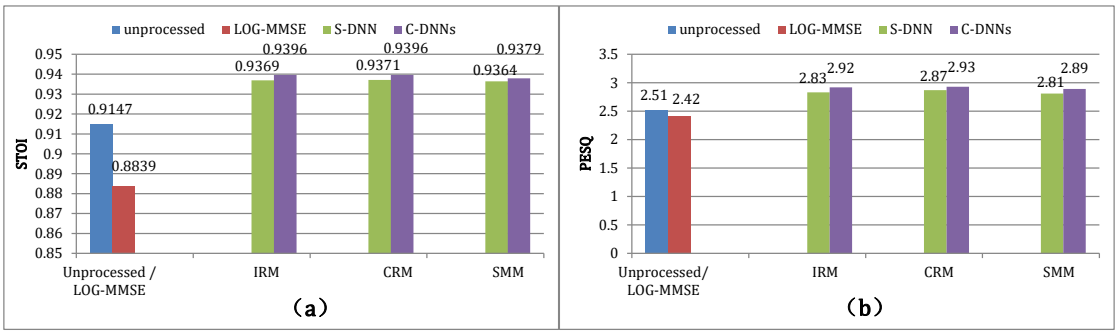


Figure 6. Comparison of the existing and the proposed methods. LOG-MMSE is reproduced based on [7], S-DNN (IRM) and S-DNN (CRM) corresponding to [20,26], respectively. For C-DNNs, the results obtained using three training targets IRM, SMM, and CRM are provided for comparison. (**a**) STOI metric. (**b**) PESQ metric.

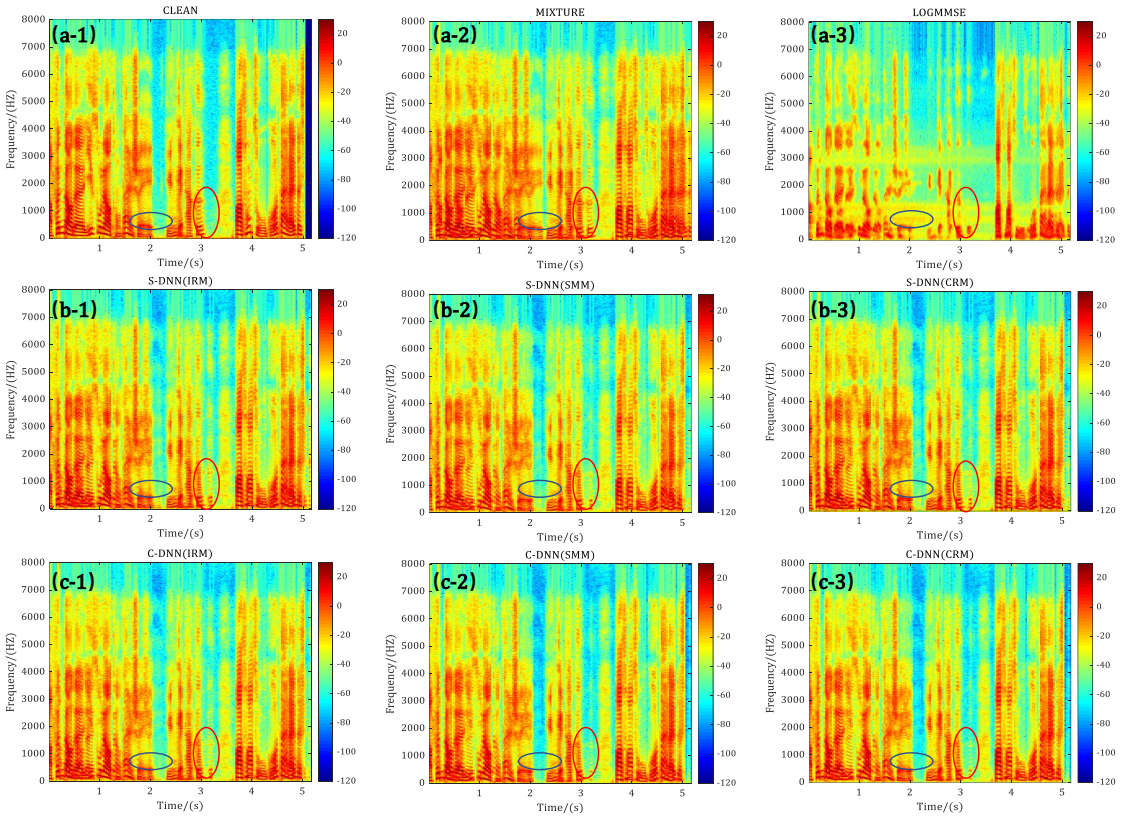


Figure 7. Spectrograms of test data No. 496 obtained by various methods. LOG-MMSE is reproduced based on [7], S-DNN(IRM) and S-DNN(CRM) corresponding to [20,26], respectively. For C-DNNs, the results obtained using three training targets IRM, SMM, and CRM are provided for comparison. (**a-1**) clean signal. (**a-2**) noisy signal. (**a-3**) processed by LOG-MMSE. (**b-1**) to (**b-3**) processed by S-DNN with IRM, SMM and CRM target, respectively. (**c-1**) to (**c-3**) processed by C-DNNs with IRM, SMM and CRM target, respectively.

5. Conclusions

The characteristic of echoes in laser monitoring signals, including T-F domain unstable, highly similar to original speech, unanticipated time delay and intensity attenuation, no pure reference signal available, etc., make it difficult to be cancelled by the traditional methods and even newly emerging neural network-based methods.

In this article, we propose a cascaded DNNs (C-DNNs) as the mapping function from the acoustic feature of the noisy speech to the ratio mask of the clean signal. To validate the feasibility and effectiveness of the proposed method, we investigated the effect of echo intensity, echo delay. and training target on the performance. We also compared the proposed C-DNNs to some traditional and newly emerging DNN-based supervised learning methods. Extensive experiments demonstrated the proposed method can greatly improve the speech intelligibility and speech quality of the echo-cancelled signals and outperform the comparison methods.

In this study, limited by the experimental conditions, we used analog data instead of real laser monitoring data to conduct relevant research, which may cause some bias in the results. In the future, we will focus on building a real laser monitoring system, synchronously collecting monitoring signals and reference signals (i.e., clean speech), and training C-DNNs with real data instead of simulating one.

Author Contributions: Conceptualization, H.L.; methodology, J.Y.; writing—original draft preparation, H.L. and I.Y. All authors have read and agree to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank all anonymous reviewers and assistant editors for their constructive comments and suggestions that significantly improved this paper.

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

References

1. Zhang, C.F. The Improvement and Realization of Laser Eavesdropping. *Laser Infrared* **2008**, *38*, 145–148.
2. Wang, M.T.; Zhu, Y.; Mu, Y.n. A two-stage amplifier of laser eavesdropping model based on waveguide fiber taper. *Def. Technol.* **2019**, *15*, 95–97. [CrossRef]
3. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
4. Li, C.; Liu, W.J. A novel multi-band spectral subtraction method based on phase modification and magnitude compensation. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4760–4763.
5. Lim, J.; Oppenheim, A. All-pole modeling of degraded speech. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 197–210. [CrossRef]
6. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 443–445. [CrossRef]
7. Loizou, P.C. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2013.
8. Valin, J.M. On adjusting the learning rate in frequency domain echo cancellation with double-talk. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1030–1034. [CrossRef]
9. Guo, M.; Elmedyb, T.B.; Jensen, S.H.; Jensen, J. Analysis of acoustic feedback/echo cancellation in multiple-microphone and single-loudspeaker systems using a power transfer function method. *IEEE Trans. Signal Process.* **2011**, *59*, 5774–5788.
10. Zhang, S.; Zheng, W.X. Recursive adaptive sparse exponential functional link neural network for nonlinear AEC in impulsive noise environment. *IEEE Trans. Neural Netw. Learn. Syst.* **2017**, *29*, 4314–4323. [CrossRef]
11. Wang, D.; Brown, G.J.; Darwin, C. Computational auditory scene analysis: Principles, algorithms and applications. *Acoust. Soc. Am. J.* **2008**, *124*, 13.
12. Brown, G.J.; Wang, D. Separation of speech by computational auditory scene analysis. In *Speech Enhancement*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 371–402.
13. Srinivasan, S.; Roman, N.; Wang, D. Binary and ratio time-frequency masks for robust speech recognition. *Speech Commun.* **2006**, *48*, 1486–1501. [CrossRef]
14. Narayanan, A.; Wang, D. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7092–7096.
15. Liang, S.; Liu, W.; Jiang, W.; Xue, W. The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio. *J. Acoust. Soc. Am.* **2013**, *134*, EL452–EL458.

16. Wang, Y.; Narayanan, A.; Wang, D. On training targets for supervised speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef]
17. Bao, F.; Abdulla, W.H.; Bao, F.; Abdulla, W.H. A New Ratio Mask Representation for CASA-Based Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2019**, *27*, 7–19. [CrossRef]
18. Liang, S.; Liu, W.; Jiang, W. A new Bayesian method incorporating with local correlation for IBM estimation. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *21*, 476–487. [CrossRef]
19. Zhang, L.; Bao, G.; Zhang, J.; Ye, Z. Supervised single-channel speech enhancement using ratio mask with joint dictionary learning. *Speech Commun.* **2016**, *82*, 38–52.
20. Bao, F.; Abdulla, W.H. A new time-frequency binary mask estimation method based on convex optimization of speech power. *Speech Commun.* **2018**, *97*, 51–65. [CrossRef]
21. Chang, J.H.; Jo, Q.H.; Kim, D.K.; Kim, N.S. Global soft decision employing support vector machine for speech enhancement. *IEEE Signal Process. Lett.* **2008**, *16*, 57–60. [CrossRef]
22. Pearlmutter, B.A. Gradient calculations for dynamic recurrent neural networks: A survey. *IEEE Trans. Neural Netw.* **1995**, *6*, 1212–1228. [CrossRef]
23. Weninger, F.; Eyben, F.; Schuller, B. Single-channel speech separation with memory-enhanced recurrent neural networks. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3709–3713.
24. Weninger, F.; Hershey, J.R.; Le Roux, J.; Schuller, B. Discriminatively trained recurrent neural networks for single-channel speech separation. In Proceedings of the 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Atlanta, GA, USA, 3–5 December 2014; pp. 577–581.
25. Williamson, D.S.; Wang, Y.; Wang, D. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2016**, *24*, 483–492. [CrossRef]
26. Zhao, Y.; Wang, Z.Q.; Wang, D. Two-stage deep learning for noisy-reverberant speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *27*, 53–62. [CrossRef]
27. Xu, Y.; Du, J.; Dai, L.R.; Lee, C.H. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process. Lett.* **2013**, *21*, 65–68. [CrossRef]
28. Wang, Q.; Du, J.; Dai, L.R.; Lee, C.H. A multiobjective learning and ensembling approach to high-performance speech enhancement with compact neural network architectures. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2018**, *26*, 1181–1193. [CrossRef]
29. Zhang, X.L.; Wang, D. A deep ensemble learning method for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP)* **2016**, *24*, 967–977. [CrossRef]
30. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
31. Han, K.; Wang, Y.; Wang, D.; Woods, W.S.; Merks, I.; Zhang, T. Learning spectral mapping for speech dereverberation and denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *23*, 982–992. [CrossRef]
32. Kim, G.; Lu, Y.; Hu, Y.; Loizou, P.C. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. *J. Acoust. Soc. Am.* **2009**, *126*, 1486–1494. [CrossRef]
33. Hermansky, H.; Morgan, N. RASTA processing of speech. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 578–589. [CrossRef]
34. Shao, Y.; Wang, D. Robust speaker identification using auditory features and computational auditory scene analysis. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Las Vegas, NV, USA, 31 March–4 April 2008; pp. 1589–1592.
35. Shao, Y.; Jin, Z.; Wang, D.; Srinivasan, S. An auditory-based feature for robust speech recognition. In Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, 19–24 April 2009; pp. 4625–4628.
36. Xu, Y.; Du, J.; Dai, L.R.; Lee, C.H. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *23*, 7–19. [CrossRef]
37. 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. Process.* **2011**, *19*, 2125–2136. [CrossRef]

38. 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 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 749–752.
39. Hu, Y.; Loizou, P.C. Subjective comparison of speech enhancement algorithms. In Proceedings of the 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006; Volume 1.

© 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/).