symmetry

MDPI

Article

Speech Enhancement Model Synthesis Based on Federal Learning for Industrial CPS in Multiple Noise Conditions

Kunpeng Wang [1,*], Wenjing Lu [1], Hao Zhou [1] and Juan Yao [1,2]

[1] School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China
[2] Department of Automation, University of Science and Technology of China, Hefei 230027, China
* Correspondence: kwang@swust.edu.cn

Abstract: Real-time acquisition of industrial production data and rapid response to changes in the external environment are key to ensuring the symmetry of a CPS. However, during industrial production, the collected data are inevitably disturbed by environmental noise, which has a huge impact on the subsequent data processing of a CPS. The types of noise vary greatly in different work scenarios in a factory. Meanwhile, barriers such as data privacy protection and copyright restrictions create great difficulties for model synthesis in the information space. A speech enhancement model with teacher–student architecture based on federal knowledge distillation is proposed to alleviate this problem. (1) We pre-train teacher models under different noise conditions to create multiple teacher models with symmetry and excelling in the suppression of a priori noise. (2) We construct a symmetric model–student model of the physical space of the teacher model trained on public data and transfer the knowledge of the teacher model to the student model. The student model can suppress multiple types of noise. Notably, with the TIMIT dataset and the NoiseX92 noise set, the accuracy of the proposed method improved by an average of 1.00% over the randomly specified teacher method in the PESQ metric and 0.17% for STOI.

Keywords: speech enhancement; CPS; symmetry; knowledge distillation; unsupervised

Citation: Wang, K.; Lu, W.; Zhou, H.; Yao, J. Speech Enhancement Model Synthesis Based on Federal Learning for Industrial CPS in Multiple Noise Conditions. *Symmetry* **2022**, *14*, 2285. https://doi.org/10.3390/sym14112285

Academic Editor: Jan Awrejcewicz

Received: 30 September 2022
Accepted: 25 October 2022
Published: 1 November 2022

1. Introduction

CPSs (cyber–physical systems) are new complex network systems that integrate physical perception, computing, communication, and control to realize real-time perception and distributed control of physical objects [1,2]. The rise of CPSs has helped the development of industrial CPS by breaking the bottleneck in the traditional production of industrial automation and control systems. Embedded systems built by sensors, actuators, processors, and heterogeneous networks provide extensive and flexible support for complex and large-scale industrial production lines, promote high automation and intelligent integration of future industries, and achieve closer collaboration between the work of robots and humans [3] . Speech is the most natural and convenient form of human communication. In many industrial CPS environments, where the operation of large equipment often requires multiple people and machines to work together, speech communication is one of the most effective ways to communicate. However, industrial facilities are frequently noisy environments where speech is invariably interfered with, especially when the amplitude of sound generated by powering and operating large equipment is much greater than the intensity of speech generated by workers. As a result, mutual communication becomes significantly less effective. In severe cases, there will be ineffective communication, resulting in the failure of collaborative work. Speech enhancement is used to restore clean speech signals from noise-interfered speech signals, improving speech quality and listener comfort, and is therefore widely used for speech in noisy environments [4].

After decades of development, numerous speech enhancement algorithms have been proposed one after another, including classical speech enhancement methods such as

spectral subtraction, Wiener filtering, statistical model-based methods, and subspace-based methods [5–8]. These methods tend to assume smooth or slowly varying noise and enhance speech suppressed by noise better under high signal-to-noise ratio (SNR), low complexity, and smooth noise conditions. However, in factories, most of the noise is low SNR and non-smooth, and traditional methods cannot track its features effectively [9], so they are not well-suited to real industrial environments. Due to the excellent ability of deep neural networks to model complex non-linear functions, training with datasets from different noisy environments can achieve stable noise reduction performance even in highly unsteady noise environments [10,11], supporting the implementation of the CPS.

Neural network-based supervised single-channel speech enhancement is mainly divided into time–frequency masking-based methods and feature mapping-based methods. The time–frequency-based masking method uses a neural network to obtain the time–frequency relationship between clean and noisy speech, multiplies the mask estimate of the clean speech with the original noisy speech, and then synthesizes the time-domain waveform of the enhanced speech through an inverse conversion technique. Early masking methods only exploited the relationship between clean and noisy speech amplitudes, ignoring the phase information between them, and include examples such as ideal binary masking (IBM) [12], ideal ratio masking (IRM) [13], and spectral amplitude masking (SAM) [14]. However, research has found that phase information is important for speech perception quality improvement at low SNR [15]. A phase-sensitive mask (PSM) [16] showed the feasibility of phase estimation. Williamson et al. [17] proposed an ideal ratio mask that can jointly estimate the real and imaginary parts of clean speech. Tan et al. [18] proposed a convolutional recurrent network (CRN) for complex spectral mapping (CSM) that can theoretically estimate the real and imaginary spectra of clean speech from the spectrum of noisy speech. The authors of [19] combined a CRN with a DPRNN module to improve both the local modeling capability and the long-term modeling capability of the network. The signal is converted to the complex frequency domain through STFT, so the speech enhancement algorithm needs to process the amplitude and phase of the speech signal concurrently [20]. Due to the difficulty of phase estimation, this imposes an upper limit on the performance of speech enhancement. In addition, effective speech enhancement in the STFT domain requires high-frequency resolution, which leads to a relatively large time window length and also leads to large system delays because the minimum delay of the system is limited by the length of the STFT time window, making it difficult to carry out in practical applications. The feature mapping-based approach uses neural networks to learn the complex mapping relationship between noisy and clean speech, and the network directly outputs the waveform of the enhanced speech. Conv-TasNet [21] is a representative and high-performance model for neural network noise reduction and speech improvement. Conv-TasNet is a fully convolutional time-domain audio separation network that models speech signals directly in the time domain. Pascual et al. [22] used GAN [23] in the field of speech enhancement and achieved some enhancement. Considering the powerful modeling capability of WaveNet [24] on speech waveforms, Ref. [25] proposed to introduce speech prior distributions into the Bayesian WaveNet framework for speech enhancement. The authors of [26] built on WaveNet by non-causal expansion convolution to predict the target speech. Further, Refs. [27,28] proposed an end-to-end speech enhancement framework using a fully convolutional neural network that focuses on time-dependent information between long speech segments.

At the same time, the highly integrated features and large-scale production lines of industrial CPS make information security issues increasingly prominent. Data communication is an important part of a CPS, so it is important to keep CPS information secure. In practice, a CPS operates with data from different industrial production systems that may come from a variety of privacy scenarios. Accessing source data from a variety of different noisy environments faces significant barriers due to issues such as user data privacy and security and commercial competition. For data privacy preservation, federated learning has been proposed [29,30]. This paradigm enables models to collaboratively model differ-

ent data structures and different institutions without uploading private data, effectively protecting user privacy and data security.

Mainstream federal learning systems are often based on the assumption that local data on the client side are labeled. However, in realistic scenarios, due to the high cost of capturing clean and noisy speech pairs, the data on the client side is mostly unlabeled. As an extension of federal learning, knowledge distillation extracts features of the teacher network in an unsupervised manner and improves the performance of the student network based on those features. Therefore, more and more researchers are trying to apply knowledge distillation within unsupervised domain-adaptive methods.

Two heads are better than one. Since a combination of multiple teacher models outperforms a single teacher, Ref. [31] transferred the predictive distribution of multiple teachers as knowledge to the student model. In [32], the authors assigned weights to teacher knowledge by analyzing the diversity of teacher models in the gradient space. The authors of [33,34] extended knowledge distillation to domain adaptation by training multiple teacher models in the source domain and pooling these models in the target domain to train a student model. In recent years, the idea of knowledge distillation has been introduced to improve the performance of speech enhancement models. The teacher network in [35] uses enhanced speech, while the student model uses the original noise-laden speech for ASR training, thus encouraging the student model to attempt speech enhancement within the network. In [36], researchers implemented single-channel speech enhancement with low latency online using teacher–student learning to prevent performance degradation due to reduced input segment length. The authors of [37] used the noise reduction results of the teacher model to optimize the student model for unsupervised learning. Recently, Refs. [38,39] attempted to transfer knowledge from multiple teachers to student models in speech enhancement. In [38], the spectrogram was divided into multiple sub-bands, the teacher model was trained on each sub-band, and then the knowledge was migrated from the teacher model to the general sub-band student enhancement model through a framework of knowledge distillation. The performance of the student model exceeded that of the corresponding full-band model. The authors of [39] trained teacher models on multiple datasets with different SNRs and then used these teacher models to supervise the training of student models.

To address the above problems of speech enhancement applied in industrial noise environments, we combine speech enhancement with knowledge distillation to obtain a comprehensive speech enhancement model that suppresses multiple industrial noise types. To begin, the network structure of the built student model is comparable to that of the instructor model. This similarity between the two models' network structures can be thought of as symmetry between the student model and the teacher model in the physical space. After that, by using knowledge distillation, the prior knowledge of various teacher models is transferred to a student model. As a result, the student model represents the unity of the prior knowledge possessed by numerous teachers in the information space. Consequently, the student model is a symmetrical and unified model of the instructor model in both physical space and information space and has the ability to suppress various forms of noise. The results also demonstrate that the performance of the teacher model is enhanced by combining multiple teacher models, as each teacher model focuses on a single noise-reduction problem.

The remainder of the paper is structured as follows: Section 2 focuses on the methodology proposed in this paper; Section 3 presents the experimental configurations; Section 4 describes the experimental results and analysis; and Section 5 concludes our work.

2. Models and Methods

CPS systems provide solutions for industrial settings where noise from various complex production environments can easily interfere with speech. Speech enhancement can effectively reduce the negative effects of noise on speech. In this section, we provide a general introduction to the proposed framework. The framework is based on a pre-trained

teacher model, where the knowledge of the teacher model at each node is then migrated to the student network through a knowledge distillation framework, which is a symmetric unification of different teacher models in the CPS physical space.

Figure 1 illustrates the overall pipeline of the proposed framework. The first step is to train a teacher model with different noisy industrial speech data. The teacher model adopts end-to-end time-domain speech enhancement that directly takes the speech signal as input and allows the neural network to learn the relationship between input and output layer-by-layer. Since the time-domain signal contains amplitude and phase information in the frequency domain, the teacher model can learn the amplitude and phase characteristics of clean speech. The second step is to use the pre-trained teacher model to train a student model suitable for noise reduction of various industrial noises. Since the teacher model focuses on the suppression of different noises, using a single teacher model for knowledge distillation may interfere with the training of the student model. Thus, we combine multiple teacher models to jointly supervise the training of the student model and transfer the comprehensive knowledge to the student model to enable the student model to obtain more information.

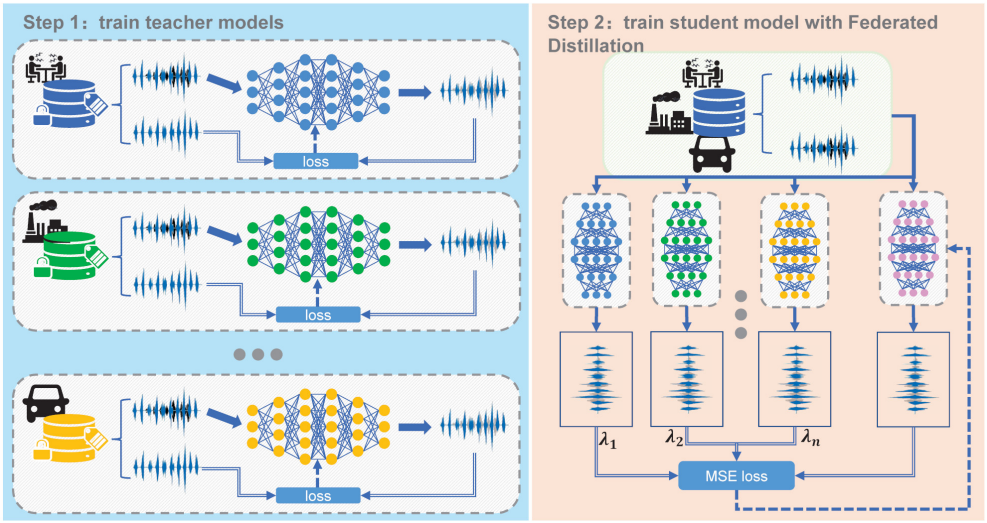

Figure 1. Illustration of our source-data-absent unsupervised distillation framework. First, several teacher models are trained on single noisy data sources to obtain several teacher models applicable to different industrial noise environments. The information from the teacher models is then combined to supervise the training of the student model so that the student model can learn more features from different teachers.

2.1. Problem Statement

Single-channel speech enhancement can be described as estimating clean speech $y(t)$ from noisy speech $x(t)$. The clean speech $y(t)$ can be expressed as: $y(t) = x(t) - n(t)$, where $n(t)$ represents noise, and t represents the time sample index. The purpose of speech enhancement algorithms is to find a function f, given x, to obtain an approximate estimate of $\hat{y}$.

$$y \approx \hat{y} = f_\theta(x) \tag{1}$$

We want to use a source-domain dataset $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$ consisting of multiple pairs of different single-noise data sources, where $x_i^s \in \mathbb{R}^n$, $y_i^s \in \mathbb{R}^n$ represent the couples of noisy and clean speech, to train a global student model S_θ, where the target domain data $D_t = \{x_i^t \in \mathbb{R}^n\}_{i=1}^{N_t}$ are not labeled. To this end, teacher models $\{T_\theta^1, \cdots, T_\theta^C\}$ are first trained from the source-domain data. Different teacher models are combined to supervise

the training of student models in order to obtain a student model S suitable for noise reduction in multiple environments.

$$\theta_S^* = \underset{\theta_S}{\operatorname{argmax}} \sum_{i=0}^{n} \mathbb{E}(S(x;\theta_S); T_i(x;\theta_T)) \tag{2}$$

where $\mathbb{E}(\cdot)$ represents the evaluation function.

2.2. Teacher Model

The teacher model uses an encoder, masking, and a decoder to predict clean speech in a supervised manner $\hat{y}$ [40]. Figure 2 illustrates the end-to-end architecture.

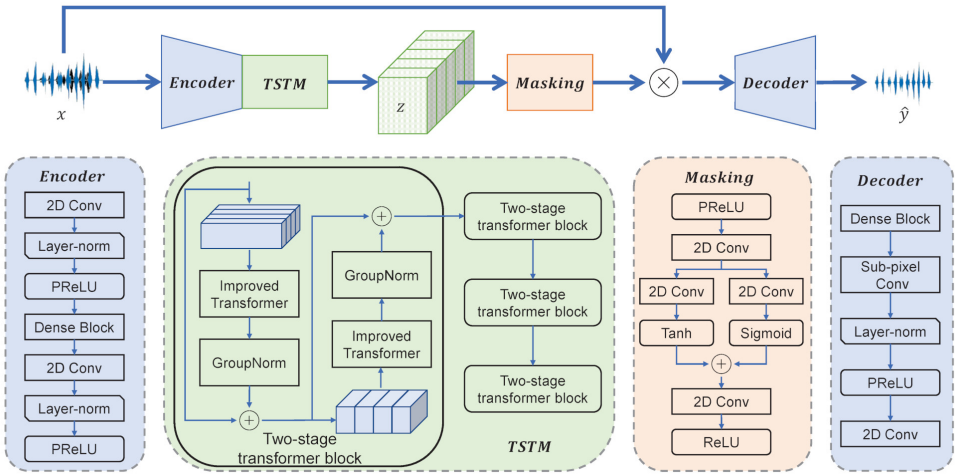

Figure 2. Teacher model architecture: the teacher model uses an attention-based mechanism to learn clean speech and improve the noise reduction and generalization of the model.

2.2.1. Encoder

The encoder first uses a convolutional layer to change the number of channels from 1 to 64, where the convolutional kernel size is (1, 2), and the step size is (1, 2). Then, there are layer normalization and PReLU [41]. A dense block [22] has four dilation convolution layers that extract the features of the lower and higher dimensions of speech. This is followed by the same convolution, further extracting features from the speech, with the number of channels left unchanged, but the convolutional kernel size becomes (1, 3).

2.2.2. TSTM

The TSTM consists of four stacked two-stage transformer modules that learn local and global high-level feature z. The two-stage transformer [42] module performs intra-block and inter-block operations successively so as to realize the interactive processing between local and global information by interlacing the processing of intra-block and intra-block information [40].

2.2.3. Masking

Masking first performs nonlinear processing of the features resulting from PReLU. The high-level features are then acquired by a 2D convolution. This is followed by two 2D convolutions then two nonlinear operations. Finally, the resulting mask is obtained by multiplying the output as input to a 2D convolution and ReLU.

2.2.4. Decoder

The decoder uses dense-block and sub-pixel convolution [43] to perform upsampling of the feature vectors. Finally, the mask feature is reconstructed into enhanced speech $\hat{y}$ with a 2D convolution.

The output of the teacher model population can be expressed as:

$$\hat{y} = (Decoder(x \otimes Masking(TSTM(Encoder(x))))) \tag{3}$$

2.2.5. Loss

The loss of the teacher model $\mathcal{L}_{\mathbb{E}}$ combines MSE loss in the time domain $\mathcal{L}_T$ with loss in the frequency domain $\mathcal{L}_F$ [44] :

$$\begin{aligned} \mathcal{L}_T &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \\ \mathcal{L}_F &= \frac{1}{TF}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1}\left[(|Y_r(t,f)| + |Y_i(t,f)|) - \left(\left|\hat{Y}_r(t,f)\right| + \left|\hat{Y}_i(t,f)\right|\right)\right] \\ \mathcal{L}_{\mathbb{E}} &= \alpha\mathcal{L}_T + (1-\alpha)\mathcal{L}_F \end{aligned} \tag{4}$$

where α is a hyperparameter optimized on the validation set (here it is set to 0.2), y and $\hat{y}$ denote the clean and enhanced speech samples, respectively, N is the speech length, Y and $\hat{Y}$ denote the STFT-transformed spectrograms of the clean and enhanced speech, respectively, r and i denote the real and imaginary parts, respectively, T is the number of frames, and F is the number of frequency bands.

2.3. Student Model with Mean Distillation

After each node is trained from local data to obtain a teacher model, a single knowledge set from each of C teacher models is migrated to the student model by an average distillation strategy. The student model has the structural symmetry of the physical space of the teacher models. Because the student model and the teacher model are structurally symmetrical, the teacher model can better guide the student model. Since the public data are unlabeled, we chose the augmented speech of the teacher model as the training target for the student model.

For a typical average distillation implementation, for a number of teacher models C with a fixed learning rate η, each teacher model provides a target to the student model, and $g_c = \triangle\mathcal{L}oss_c\left(\widehat{y_{s_{r-1}}}, \widehat{y_c}\right)$, where $\widehat{y_{s_{r-1}}}$ is the enhanced speech output by the student model, and $\widehat{y_c}$ is the speech output by the c teacher model. This is equivalent to each teacher giving the student model a gradient, which is finally aggregated by the student model and updated when $\omega_r \leftarrow \omega_{r-1} - \eta\sum_{c=1}^{C} g_c$ is applied. Hence $\eta\sum_{c=1}^{C} g_c = \triangle\mathcal{L}oss_c\left(\widehat{y_{s_{r-1}}}, \widehat{y_c}\right)$. That is, C teacher models simultaneously perform gradient descent on the student model.

Since the teacher models excel in the suppression of different noises, a multi-teacher hybrid teaching network is used instead of the traditional single-teacher network to enhance the teaching breadth and ability of the teacher network. During training, the student model can learn from multiple teacher networks, giving the student network more valid information and enhancing the performance of the student network. The student model is trained using public data to obtain a speech enhancement model suitable for multiple noisy environments, as shown in Table 1.

We use the same loss function as the teacher model to minimize the difference between the outputs of the student and teacher models. Since the student model is symmetrical to the teacher model, this loss allows additional knowledge to be extracted from the teacher model during student training and enhances the ability of the student model to enhance speech under different noises:

$$\mathcal{L}_S = \frac{1}{C}\sum_{j=1}^{C}\mathcal{L}_{\mathbb{E}}\left(\widehat{y}_s, \widehat{y}_j\right) \tag{5}$$

where $\mathcal{L}_{\mathbb{E}}$ denotes the loss function combining the time and frequency domains, $\widehat{y}_s$ denotes the clean speech estimate of the student model output, and $\widehat{y}_j$ denotes the pseudo-label of the j teacher model output.

Table 1. The overall training procedure of mean distillation r.

Input: $\theta_S^{(0)}$ // initial student model weights
Output: $\theta_S^{(R)}$
1. ***for*** $r = 1 : r++$; while $r < R$ do
2. // Available teacher models $\theta_T^{(r)} \subseteq \{1, \cdots, C\}$
3. $\mathcal{T} \longleftarrow$ (split D_t into batch)
4. for batch t in $\mathcal{T}$ do
5. // average distillation from teacher model
6. $\theta_S^{(r)} \leftarrow \theta_S^{(r-1)} - \eta\nabla\mathcal{L}oss\left(\theta_S^{(r-1)}; t\right)$
8. ***end***

3. Experiment

3.1. Dataset

We use a public dataset to evaluate the proposed approach, and the specific details of the dataset are shown in Table 2.

(1) **Teacher models:** We built four teacher models. In order to train the teacher models in different noise environments, we selected 11,572 clean data points from VoiceBank [45]. The VoiceBank corpus consisted of 30 native English speakers from different parts of the world, each reading approximately 400 sentences, and we used 28 speakers for training and 2 for testing. We also used four noises common in the CPS industry—babble, white, destroyerops, and factory (from the NoiseX-92 dataset [46])—to generate the noise-laden speech dataset. The noisy speech in the training set was added to a random SNR of $\{0, 5, 10, 15\}$ dB corresponding to the noise. The test set had a SNR of $\{5, 10, 15\}$ dB.

(2) **Student model:** A total of 6300 clean speech items were selected from the TIMIT training dataset and mixed at $\{0, 5, 10, 15\}$ dB with the four noises we used in the teacher model dataset at random. This produced 25,200 noisy speech samples. Of these, 23,450 were used as the training dataset, and 1750 were used as the validation dataset. To evaluate the student model, 1344 speech sounds were selected from the TIMIT test dataset and blended at $\{5, 10, 15\}$ dB with the four noises present in the two datasets above. The resulting 5376 noisy speech samples were used as the test dataset.

A PyTorch implementation of the proposed model can be found at "http://www.msp-lab.cn:1436/ALu/AvgKD-SE-in-CPS-industry" (accessed 4 October 2022).

Table 2. Relevant descriptions of the training and testing datasets.

Dataset		Type	Information	Source
Clean speech	Teacher model	30 speakers	training 28 speakers testing 2 speakers	VoiceBank corpus [45]
	Student model	630 speakers	training 462 speakers testing 168 speakers	TIMIT [47]
Noise		white	Sampling a high-quality analog noise generator	NoiseX92 [46]
		babble	100 people speaking in a canteen	
		factory	Plate-cutting and electrical welding equipment	
		destroyerops	Samples from microphone on digital audio tape	

3.2. Training Detail

The sampling rate for all speech was 16,000 Hz. Each frame had a size of 512 samples with an overlap of 256 samples. If the time to train a particular speech sample was greater

than 4 s, we sliced a random 4 s segment from it. In batch processing, shorter speech wa zero-padded to match the size of the longest speech. In the training phase, we trained the model for 100 epochs and optimized it using the Adam optimizer (decay rate $\beta_1 = 0.9, \beta_2 = 0.999$). For the learning rate, we used the dynamic strategy [42] in the training phase.

3.3. Evaluation Metrics

PESQ: Perceptual Evaluation of Speech Quality (PESQ) [4] is one of the most commonly used metrics for evaluating speech quality and is similar to MOS, but its calculation is much more complex and includes pre-processing, temporal alignment, perceptual filtering, masking effects, etc. The range of PESQ value is $[-0.5, 4.5]$. The higher the PESQ value, the better the audible speech quality of the tested speech.

STOI: Short-Time Objective Intelligibility (STOI) [48] is calculated based on the temporal envelope correlation coefficients of clean and noisy speech. It has been shown that STOI correlates well with the results of intelligibility listening experiments. The value range of STOI is [0, 1], and it is positively correlated with subjective speech intelligibility; that is, the larger the value, the better the speech intelligibility.

4. Discussion

4.1. Evaluation of the Teacher Models

In this section, we present the training results of the teacher models for four CPS industrial background noises and test them on four noisy datasets. Four types of noise under four SNRs of 5 dB, 10 dB, and 15 dB are plotted in Figure 3, where the same noise conditions as in the training set were used for testing. We observed that the teacher model trained for a certain noise condition always outperformed the other teacher models in terms of noise reduction, even under different SNRs. This result confirms our conjecture.

Next, we further verified that the teacher model was better at suppressing a priori noise testing with invisible speakers. The test results for different objective evaluation metrics are given in Tables 3 and 4. "White_T" denotes the teacher model obtained by training under the white noise condition. Similarly, "Babble_T", "Factory_T", and "Destroyerops_T" denote teacher models trained under babble, factory, and destroyerops noise conditions, respectively. As can be seen from Tables 3 and 4, the teacher models trained with the corresponding environmental noise are better able to tap into the corresponding noise information and are far better than the other models in the corresponding noise conditions. As the models are more focused on noise reduction of one type of noise, the teacher models can still achieve good noise reduction results even at a low SNR ratio of 5 dB.

Table 3. Average PESQs of teacher models on VoiseBank + NoiseX92.

Noise Type	**White**			**Babble**			**Factory**			**Destroyerops**		
SNR	**5**	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
noisy	1.3502	1.6011	1.9188	1.955	2.3697	2.765	1.7486	2.1791	2.6098	1.7337	2.1716	2.594
White-T	**3.2178**	**3.6357**	**3.9430**	2.0419	2.4767	2.9081	2.0387	2.5519	3.0627	2.0122	2.4935	2.9522
Babble-T	1.3991	1.6337	1.9476	**3.4725**	**3.7636**	**3.9862**	2.4183	2.7894	3.1658	2.2434	2.6867	3.0835
Factory-T	2.5898	3.0094	3.4656	2.7635	3.1792	3.5274	**3.2973**	**3.6181**	**3.8932**	2.8717	3.2751	3.6161
Destroyerops-T	2.6038	3.0335	3.3431	2.5517	2.9989	3.3964	2.5754	2.9510	3.2934	**3.2275**	**3.5150**	**3.7427**

Table 4. Average STOIs of teacher models on VoiseBank + NoiseX92.

Noise Type	**White**			**Babble**			**Factory**			**Destroyerops**		
SNR	**5**	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**	**5**	**10**	**15**
noisy	0.9200	0.9481	0.9668	0.9235	0.9446	0.96	0.9241	0.9465	0.9628	0.9264	0.9478	0.9619
White-T	**0.9538**	**0.9693**	**0.9798**	0.9265	0.9471	0.9618	0.9291	0.9504	0.9660	0.9309	0.9513	0.9648
Babble-T	0.9283	0.9516	0.9671	**0.9602**	**0.9696**	**0.9762**	0.9449	0.9595	0.9701	0.9493	0.9638	0.9728
Factory-T	0.9435	0.9604	0.9718	0.9466	0.9597	0.9704	**0.9563**	**0.9677**	**0.9763**	0.9517	0.9647	0.9735
Destroyerops-T	0.9366	0.9543	0.9650	0.9471	0.9612	0.9718	0.9458	0.9599	0.9709	**0.9610**	**0.9706**	**0.9777**

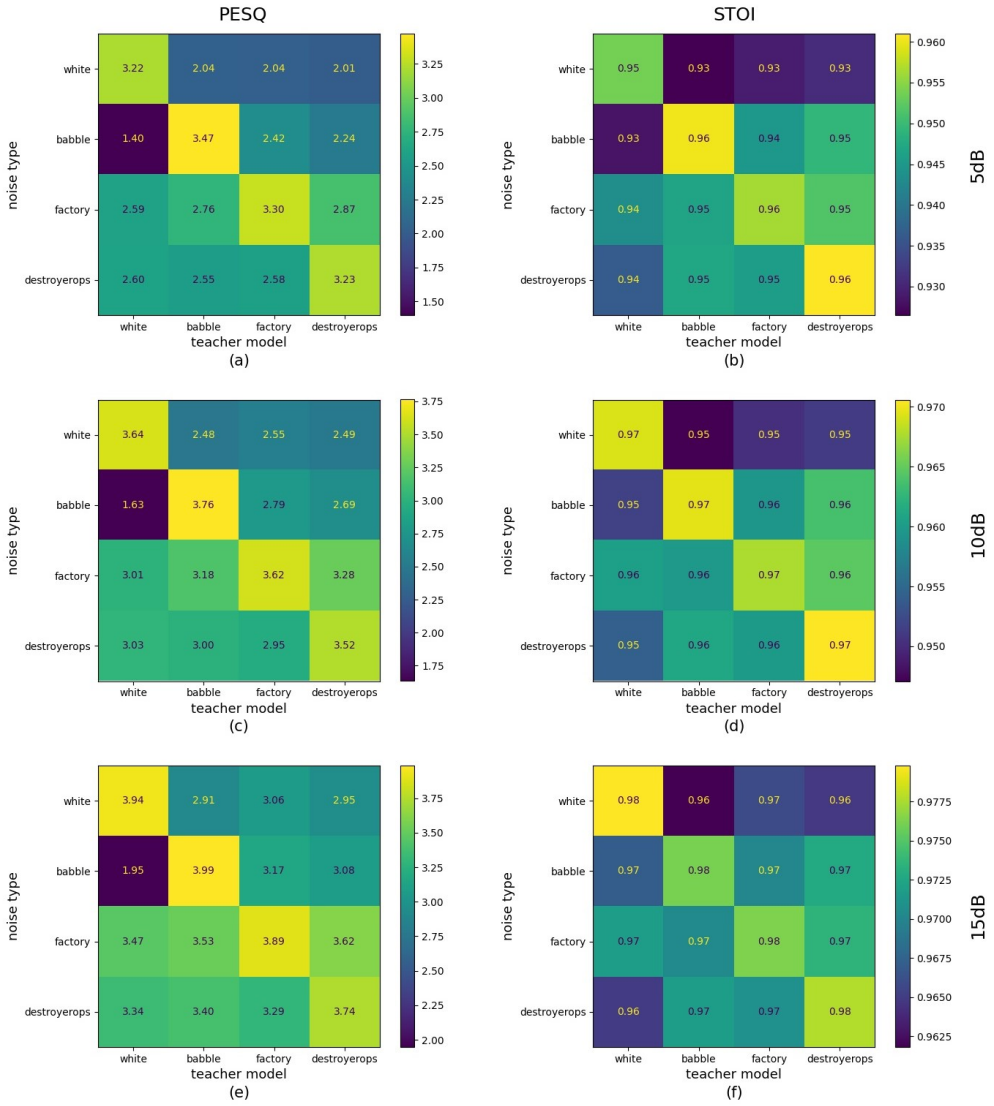

Figure 3. Experimental results under different noises and SNRs. The evaluation metrics from the first and second rows are PESQ and STOI, respectively: teacher model testing results at SNRs of (**a**,**b**) 5 dB, (**c**,**d**) 10 dB, and (**e**,**f**) 15 dB.

From the test results of the teacher model, we can find that the teacher model is symmetrical in physical space but has a different ability to suppress noise. Synthesizing the knowledge of these teacher models to obtain a symmetric and uniform speech noise reduction model is the key to suppressing CPS industrial noise. Next, we show the test results of the student models obtained using different distillation methods. These results further substantiate our conjecture.

4.2. Student Models on TIMIT + NoiseX92

In this section, we perform an ablation study to analyze the effectiveness of the noise representation model in the proposed framework. For the following two variants of our method: (1) *Random_S* denotes randomly selected loss done on the pseudolabels and augmented speech output from the teacher model and the student model, and (2) *AVE_S* denotes loss done on the pseudolabels and augmented speech output from the student model for each teacher model and then averaged.

Tables 5 and 6 give the PESQ and STOI scores of the student model on the TIMIT + NoiseX92 datasets. First, we observe that both student models are suppressed by different noises. This implies that pre-trained teacher models can be used to provide pseudolabels

when labels are missing from the dataset, and that the student models can learn noise reduction capabilities from these pseudolabels. Furthermore, the *AVE_S* model was 1.00% better than *Random_S* in terms of PESQ and 0.17% better in terms of STOI in most test cases. This indicates that multiple-teacher model distillation can provide more knowledge to the student model.

In order to more intuitively compare the enhancement due to various algorithms, we compared and analyzed the spectrograms of the enhanced speech of each network. Figure 4 is a spectrum of enhanced speech samples obtained using different algorithms under four noise conditions with an SNR of 5dB, with the horizontal axis representing time T and the vertical axis representing the speech signal frequency F. The leftmost column of Figure 4 is the spectrogram obtained by adding different noises to the same clean speech. The middle column is the spectrogram of the clean speech, and the right is the enhanced speech in our experiment. As can be seen from the spectra, the characteristics of the four noisy speech sources are also different. Through comparative analysis of the four spectra, it can be seen that white noise has the same interference with all frequencies of noise. Babble noise, on the other hand, mainly interferes with the low-frequency band, possibly because babble is mainly background noise where multiple people are talking. Factory and destroyerops noises impact the low, mid, and high bands of speech, but they are not the same as white noise.

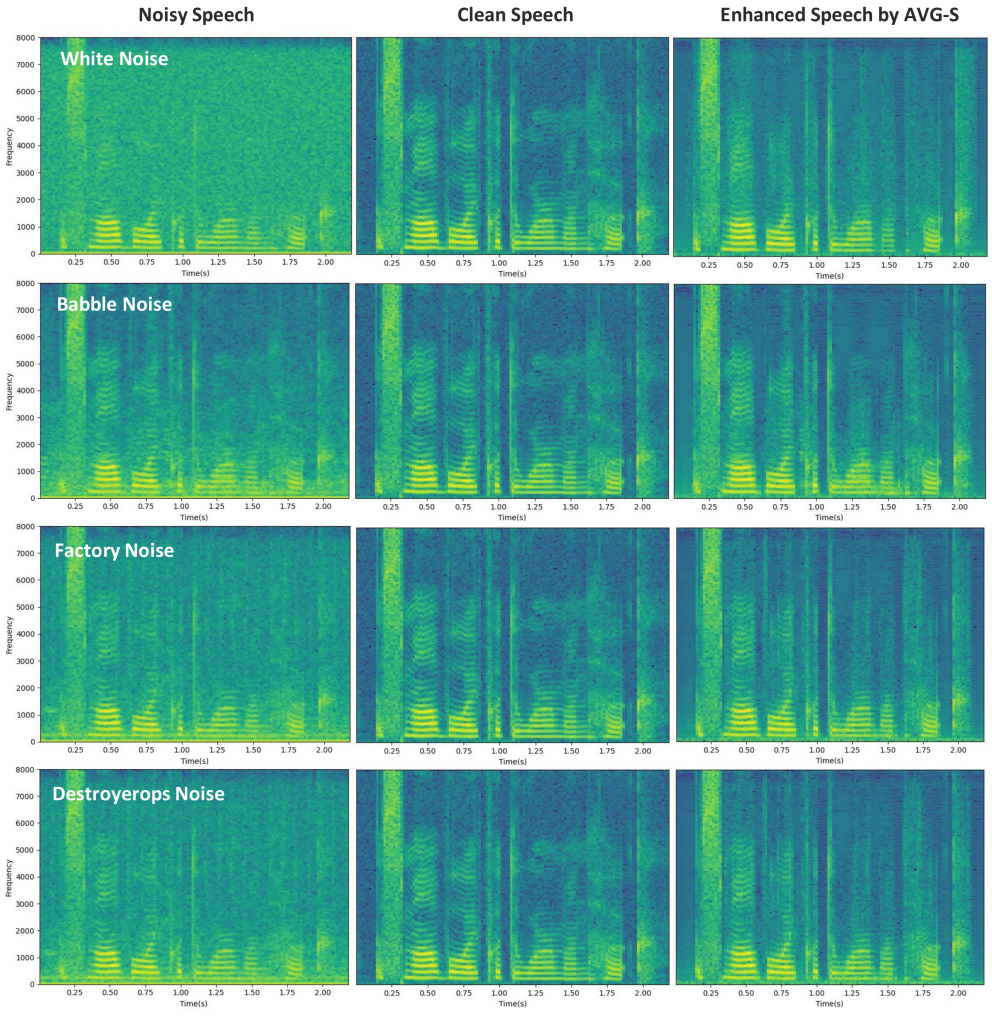

Figure 4. Spectrograms of different distillation methods at different noise conditions. Figures show the results of noisy speech, speech enhanced by *AVE_S*, and clean speech under four noise conditions, respectively.

Table 5. Average PESQs of student models on TIMIT + NoiseX92.

Noise Type	White			Babble			Factory			Destroyerops		
SNR	5	10	15	5	10	15	5	10	15	5	10	15
noisy	1.3314	1.6784	2.1967	1.8224	2.321	2.8465	1.7311	2.2928	2.9016	1.6181	2.1384	2.7365
Random-S	2.4116	2.9373	**3.3127**	2.4754	2.9764	3.3710	2.3568	2.9448	3.3852	2.3273	2.9081	3.3567
AVE-S	**2.4503**	**2.9564**	3.3043	**2.4846**	**3.0094**	**3.3901**	**2.3917**	**2.9798**	**3.3879**	**2.3875**	**2.9657**	**3.3784**

Table 6. Average STOIs of student models on TIMIT + NoiseX92.

Noise Type	White			Babble			Factory			Destroyerops		
SNR	5	10	15	5	10	15	5	10	15	5	10	15
noisy	0.9440	0.9767	0.9912	0.9097	0.9537	0.9782	0.9271	0.9667	0.9860	0.9022	0.9501	0.9777
Random-S	0.9475	0.9734	**0.9854**	0.9479	0.9702	0.9831	0.9547	0.9758	0.9867	0.9428	0.9677	0.9824
AVE-S	**0.9505**	**0.9745**	0.9853	**0.9482**	**0.9721**	**0.9843**	**0.9578**	**0.9778**	**0.9874**	**0.9450**	**0.9701**	**0.9837**

Both algorithms effectively processed noisy speech to a certain extent, and *Random_S* methods have relatively large noise residues in the process of enhancing noisy speech. Because of the different noise suppressions that different teacher models are good at, a single teacher cannot provide clearer speech labels for student models, and they can even introduce some noise or excessive enhancement. The *AVE_S* method guides the use of different feature information by student models in the training process, which can alleviate the problem of insufficient or excessive enhancement of a single model during enhancement.

The experimental results further verify that the proposed average distillation model can suppress various noises from the aspect of speech spectrum characteristics. Since the *AVE_S* model is a symmetric unified model of the teacher model in physical space, it has the noise reduction capability of all teachers. Because a symmetric model with the teacher model is constructed, it enables the student model to learn more useful knowledge from the teacher model. Additionally, we can see that even if the data lack labels, knowledge transfer is still possible through the information interaction between the teacher model and the student model. This further verifies that federal knowledge distillation can perform model synthesis and is applicable in CPS industrial noise reduction scenarios that lack data labels and source data.

5. Conclusions

To solve the problem of industrial CPS noise data being hard to combine because of privacy and commercial copyright laws and the lack of public dataset labels, we propose unsupervised distillation-based speech enhancement for unsourced data. In this method, a speech enhancement model for multiple noisy environments is trained to achieve unification of the CPS physical space through federal learning and knowledge distillation without accessing the source data. Through a large number of comparative experiments, we verify that the prior noise suppression performance of teacher models trained under different noises is always due to other teacher models. We compare our average distillation and random distillation methods. It can be seen from the experimental results that the average distillation guides the student model to effectively utilize the feature information of different teacher models during the training process and alleviates the problem of insufficient or excessive enhancement of a single teacher model during the enhancement process.

Our method can eliminate or reduce the background noise in noisy speech and improve the quality and intelligibility of the target speech signal. As shown in Figure 4, speech enhancement may be successful in suppressing the noise but often distorts the speech that is of interest. Artifacts created by speech enhancement can harm the quality of speech recognition or other further automatic processing. However, a High-Fidelity Generative Adversarial Network [49] consisting of one generator and two discriminators is adversarially trained to output high-quality speech without artificial generation artifacts. In our next work, we will use the High-Fidelity Generative Adversarial Network for speech enhancement to eliminate generation artifacts.

In future research, we will continue to explore the application of multi-teacher distillation in speech enhancement and train a standard model with samples representing different noisy environments as a comparison. Although our experimental data come from a public dataset, the coupling of noisy and clean speech is artificially synthesized from clean speech and noise. The simulated speech may be different from recordings in a real noise environment. In future work, we will use CHiME and other real recording datasets to conduct knowledge distillation research in multi-noise scenarios. Further, in order to extend the application of our method to speech signals, we aim to extend our approach to other downstream tasks, including simultaneous speech recognition and separation.

Author Contributions: W.L. and K.W. designed and programmed the proposed algorithm and wrote the paper; H.Z. and J.Y. participated in algorithm design, algorithm programming, and paper writing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China grant number 2021YFB1715000 and the Sichuan Science and Technology Program grant number 2021YFG0315.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Konstantinou, C.; Maniatakos, M.; Saqib, F.; Hu, S.; Plusquellic, J.; Jin, Y. Cyber-physical systems: A security perspective. In Proceedings of the 2015 20th IEEE European Test Symposium (ETS), Cluj-Napoca, Romania, 25–29 May 2015; pp. 1–8.
2. Sanislav, T.; Mois, G.; Miclea, L. An approach to model dependability of cyber-physical systems. *Microprocess. Microsyst.* **2016**, *41*, 67–76. [CrossRef]
3. Wang, C.; Lv, Y.; Wang, Q.; Yang, D.; Zhou, G. Service-Oriented Real-Time Smart Job Shop Symmetric CPS Based on Edge Computing. *Symmetry* **2021**, *13*, 1839. [CrossRef]
4. Loizou, P.C. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2007.
5. Boll, S. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **1979**, *27*, 113–120. [CrossRef]
6. Lim, J.; Oppenheim, A. All-pole modeling of degraded speech. *IEEE Trans. Acoust. Speech Signal Process.* **1978**, *26*, 197–210. [CrossRef]
7. 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]
8. Ephraim, Y.; Van Trees, H.L. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.* **1995**, *3*, 251–266. [CrossRef]
9. Purwins, H.; Li, B.; Virtanen, T.; Schlüter, J.; Chang, S.Y.; Sainath, T. Deep learning for audio signal processing. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 206–219. [CrossRef]
10. Zhang, Q.; Nicolson, A.; Wang, M.; Paliwal, K.K.; Wang, C. DeepMMSE: A deep learning approach to MMSE-based noise power spectral density estimation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1404–1415. [CrossRef]
11. Li, A.; Yuan, M.; Zheng, C.; Li, X. Speech enhancement using progressive learning-based convolutional recurrent neural network. *Appl. Acoust.* **2020**, *166*, 107347. [CrossRef]
12. Roman, N.; Woodruff, J. Ideal binary masking in reverberation. In Proceedings of the 2012 20th European Signal Processing Conference (EUSIPCO), Bucharest, Romania, 27–31 August 2012; pp. 629–633.
13. Li, X.; Li, J.; Yan, Y. Ideal Ratio Mask Estimation Using Deep Neural Networks for Monaural Speech Segregation in Noisy Reverberant Conditions. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 1203–1207.
14. Tengtrairat, N.; Woo, W.L.; Dlay, S.S.; Gao, B. Online noisy single-channel source separation using adaptive spectrum amplitude estimator and masking. *IEEE Trans. Signal Process.* **2015**, *64*, 1881–1895. [CrossRef]
15. Paliwal, K.; Wójcicki, K.; Shannon, B. The importance of phase in speech enhancement. *Speech Commun.* **2011**, *53*, 465–494. [CrossRef]
16. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Le Roux, J. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 708–712.
17. Williamson, D.S.; Wang, Y.; Wang, D. Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2015**, *24*, 483–492. [CrossRef]
18. Tan, K.; Wang, D. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 6865–6869.

19. Le, X.; Chen, H.; Chen, K.; Lu, J. DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 2811–2815.
20. Luo, Y.; Mesgarani, N. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 696–700.
21. Luo, Y.; Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266. [CrossRef] [PubMed]
22. Pascual, S.; Bonafonte, A.; Serrà, J. SEGAN: Speech Enhancement Generative Adversarial Network. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 3642–3646.
23. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
24. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
25. Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; Florêncio, D.; Hasegawa-Johnson, M. Speech Enhancement Using Bayesian Wavenet. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017; pp. 2013–2017.
26. Rethage, D.; Pons, J.; Serra, X. A wavenet for speech denoising. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5069–5073.
27. Fu, S.W.; Tsao, Y.; Lu, X.; Kawai, H. Raw waveform-based speech enhancement by fully convolutional networks. In Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, 12–15 December 2017; pp. 6–12.
28. Fu, S.W.; Wang, T.W.; Tsao, Y.; Lu, X.; Kawai, H. End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2018**, *26*, 1570–1584. [CrossRef]
29. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 1273–1282.
30. Bonawitz, K.; Ivanov, V.; Kreuter, B.; Marcedone, A.; McMahan, H.B.; Patel, S.; Ramage, D.; Segal, A.; Seth, K. Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, TX, USA, 30 October–3 November 2017; pp. 1175–1191.
31. Malinin, A.; Mlodozeniec, B.; Gales, M. Ensemble distribution distillation. *arXiv* **2019**, arXiv:1905.00076.
32. Du, S.; You, S.; Li, X.; Wu, J.; Wang, F.; Qian, C.; Zhang, C. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12345–12355.
33. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain adaptive ensemble learning. *IEEE Trans. Image Process.* **2021**, *30*, 8008–8018. [CrossRef]
34. Meng, Z.; Li, J.; Gong, Y.; Juang, B.H. Adversarial Teacher-Student Learning for Unsupervised Domain Adaptation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5949–5953.
35. Watanabe, S.; Hori, T.; Le Roux, J.; Hershey, J.R. Student-teacher network learning with enhanced features. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5275–5279.
36. Nakaoka, S.; Li, L.; Inoue, S.; Makino, S. Teacher-student learning for low-latency online speech enhancement using wave-u-net. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 661–665.
37. Kim, S.; Kim, M. Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation. In Proceedings of the 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 17–20 October 2021; pp. 176–180.
38. Hao, X.; Wen, S.; Su, X.; Liu, Y.; Gao, G.; Li, X. Sub-Band Knowledge Distillation Framework for Speech Enhancement. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 2687–2691.
39. Hao, X.; Su, X.; Wang, Z.; Zhang, Q.; Xu, H.; Gao, G. SNR-based teachers-student technique for speech enhancement. In Proceedings of the 2020 IEEE International Conference on Multimedia and Expo (ICME), London, UK, 6–10 July 2020; pp. 1–6.
40. Wang, K.; He, B.; Zhu, W.P. TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 7098–7102.
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
42. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017.

43. Li, P.; Jiang, Z.; Yin, S.; Song, D.; Ouyang, P.; Liu, L.; Wei, S. Pagan: A phase-adapted generative adversarial networks for speech enhancement. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6234–6238.
44. Pandey, A.; Wang, D. Densely connected neural network with dilated convolutions for real-time speech enhancement in the time domain. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6629–6633.
45. Veaux, C.; Yamagishi, J.; King, S. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In Proceedings of the 2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Gurgaon, India, 25–27 November 2013; pp. 1–4.
46. 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]
47. Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Tech. Rep. N* **1993**, *93*, 27403.
48. 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]
49. Kong, J.; Kim, J.; Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, Virtual, 6–12 December 2020; pp. 17022–17033.