electronics

Article

A Multi-Stage Acoustic Echo Cancellation Model Based on Adaptive Filters and Deep Neural Networks

Shiyun Xu †, Changjun He †, Bosong Yan and Mingjiang Wang *

Key Laboratory for Key Technologies of IoT Terminals, Harbin Institute of Technology, Shenzhen 518055, China; 21s052011@stu.hit.edu.cn (S.X.); hehelita@163.com (C.H.); ybs5250057an@163.com (B.Y.)

* Correspondence: mjwang@hit.edu.cn

† These authors contributed equally to this work.

Abstract: The presence of a large amount of echoes significantly impairs the quality and intelligibility of speech during communication. To address this issue, numerous studies and models have been conducted to cancel echo. In this study, we propose a multi-stage acoustic echo cancellation model that utilizes an adaptive filter and a deep neural network. Our model consists of two parts: the Speex algorithm for canceling linear echo, and the multi-scale time-frequency UNet (MSTFUNet) for further echo cancellation. The Speex algorithm takes the far-end reference speech and the near-end microphone signal as inputs, and outputs the signal after linear echo cancellation. MSTFUNet takes the spectra of the far-end reference speech, the near-end microphone signal, and the output of Speex as inputs, and generates the estimated near-end speech spectrum as output. To enhance the performance of the Speex algorithm, we conduct delay estimation and compensation to the far-end reference speech. For MSTFUNet, we employ multi-scale time-frequency processing to extract information from the input spectrum. Additionally, we incorporate an improved time-frequency self-attention to capture time-frequency information. Furthermore, we introduce channel time-frequency attention to alleviate information loss during downsampling and upsampling. In our experiments, we evaluate the performance of our proposed model on both our test set and the blind test set of the Acoustic Echo Cancellation challenge. Our proposed model exhibits superior performance in terms of acoustic echo cancellation and noise reverberation suppression compared to other models.

Keywords: acoustic echo cancellation; multi-stage model; adaptive filter; deep neural network

Citation: Xu, S.; He, C.; Yan, B.; Wang, M. A Multi-Stage Acoustic Echo Cancellation Model Based on Adaptive Filters and Deep Neural Networks. *Electronics* **2023**, *12*, 3258. https://doi.org/10.3390/electronics12153258

Academic Editors: Fabio Mendonca, Morgado Dias and Sheikh Shanawaz Mostafa

Received: 5 July 2023
Revised: 24 July 2023
Accepted: 27 July 2023
Published: 28 July 2023

1. Introduction

With the continuous progress of modern technology, the 5G era has arrived. Thanks to the unceasing innovation in communication and network technology, VoIP (Voice over Internet Protocol) communication technology has gained widespread popularity compared to the 4G era. Applications such as WeChat, Skype, and various conference software have become widely used among the general public. VoIP has attracted market and public attention due to its ability to fully exploit network bandwidth, minimize call costs, and facilitate the implementation of value-added services. Moreover, with the recent outbreak of COVID-19, offline meetings have shifted to online platforms, and traditional classrooms have transitioned to online classrooms. Consequently, VoIP communication technology has garnered increased significance.

However, it is worth mentioning that during VoIP communication, not only does the near-end microphone capture the speech of the near-end speaker, but it also records the sound played by the near-end speaker, causing the far-end speaker to potentially hear the echo of their speech. Moreover, due to speech encoding, decoding, and the transmission of data over the network, there may be varying levels of time delays that further contribute to the generation of echo, causing inconvenience for speakers. Therefore, in line with the continuous development of the VoIP industry, echo cancellation has emerged as a prominent research focus and a crucial area for improvement to enhance the quality of communication processes and enhance user experience.

There are two main types of echo in VoIP communication: circuit echo and acoustic echo. Due to the two to four wire conversion of the switch, circuit echo is generated on the network side [1]. However, due to the advancement of echo cancellation technology, circuit echo has been effectively reduced and canceled. As a result, the focus of echo cancellation has shifted from circuit echo to acoustic echo.

It is very difficult to cancel acoustic echo. There are several main reasons: (1) the communication process mainly takes place in enclosed environments such as conference rooms, where the sound emitted by the speaker will be reflected repeatedly and then captured by the microphone, mixing with the speech of the near-end speaker. This process results in a long tail of the echo, and the corresponding echo path [2] has a long impulse response. Therefore, to achieve echo cancellation, it is necessary to increase the order of the adaptive filter. (2) During communication, it is impossible to guarantee absolute silence in the environment. Noise caused by personnel movement or other forms of interference can disrupt the propagation of sound, leading to significant fluctuations in the pulse response of the acoustic echo. Therefore, the acoustic echo path is not stable. The rapid changes in the echo path require the echo cancellation process to have a fast convergence speed [3] and good tracking performance. However, algorithms with fast convergence speeds are closely related to computational complexity. The echo cancellation filter used to cancel echo has a high order, and conventional fast algorithms often cannot effectively solve the problem. (3) During the VoIP communication process, environmental noise can affect echo cancellation. Additionally, in situations with high background noise, echo cancellation not only needs to handle the echo, but also needs to consider the background noise, making filter design extremely challenging.

In summary, the main research challenge in echo cancellation currently lies in dealing with the acoustic echo. The noise, reverberation, and variations in echo paths significantly increase the complexity of acoustic echo cancellation. Developing an efficient echo cancellation system can greatly enhance the user experience during VoIP communication.

For acoustic echo, it can be classified into two types: linear echo and non-linear echo. Linear echo refers to the echo produced by sound waves propagating in a straight path in space, while non-linear echo refers to the non-linear effect generated by sound waves propagating in space [4]. Linear echo can be effectively canceled using traditional methods. These methods primarily rely on adaptive filtering techniques, such as the Least Mean Square (LMS) algorithm [2], Normalized Least Mean Square (NLMS) algorithm [3], Recursive Least Squares (RLS) algorithm [5], and Blocked Frequency Domain Adaptive Filter (PBFDAF) [4]. Diniz et al. [2] propose a method that involves utilizing LMS adaptive filters to replicate the echo path and subsequently subtracting the estimated echo signal from the input signal, thereby achieving effective echo cancellation. This algorithm is straightforward, dependable, and widely applicable; however, it utilizes instantaneous values instead of expected values in its calculation during the iteration process. As a consequence, this approach introduces errors into the calculation process, which are affected by the input signal and subsequently evolve as the input signal changes. When assessing the performance of the adaptive filtering algorithm, the LMS algorithm may introduce uncertainty in the rate of convergence, which ultimately leads to numerous uncertain factors in echo and unstable echo cancellation.

To address this issue, Slock et al. [3] replace the LMS algorithm with the NLMS algorithm to simulate the echo path, accompanied by the incorporation of normalization operations into the LMS algorithm. This modification aims to ensure algorithm convergence through normalization, thereby enhancing the overall convergence effect. Nevertheless, it is important to note that this algorithm possesses significant drawbacks in terms of convergence rate and steady-state error.

Duttweiler et al. [6] introduce the proposition normalized Least Mean Square (PNLMS) algorithm as a means to emulate the echo path. This algorithm effectively modifies the filter weight in proportion to the sparse character of the echo path (sparse character describes the phenomenon where there are fewer significant signal components present on the echo path during the transmission of acoustic signals), which enhances convergence speed and minimizes steady-state error for the sparse echo path. However, the performance of this algorithm might diminish for the non-sparse echo path. Liu et al. [1] propose the improvement normalized Least Mean Square (IPNLMS) algorithm to address the issue of performance degradation in the non-sparse echo path. In comparison to the PNLMS algorithm, this method effectively enhances the convergence rate when dealing with the non-sparse echo path; however, the steady-state error and computational complexity also increase as a result.

Speex is an open source audio codec, mainly used for real-time audio communication [7]. Its echo cancellation part is based on NLMS and implemented using a multi-delay block filter [8] in the frequency domain. It has the advantages of efficient echo suppression, low latency, and cross platform support, making it widely used in real-time communication applications.

For linear echo, adaptive filters can already achieve good cancellation effects. However, when it comes to nonlinear echo, adaptive filters often fall short in achieving the desired effect due to the presence of reverberation and complex acoustic characteristics (such as distortion, non-linear resonance, and interference effect [9]). In order to address the intricate non-linear relationship between inputs and outputs, deep neural networks (DNNs) are increasingly utilized for echo cancellation tasks. A DNN is a neural network with multiple layers. It is introduced by Hinton et al. [10], which effectively tackles the problem of gradient explosion and vanishing in multi-layer neural networks, thereby enabling the creation of truly deep networks. As deep learning technology evolves, DNNs are able to more effectively extract deep data information. Compared to shallow neural networks, DNNs possess stronger capabilities in expressing non-linear relationships.

UNet [11] is a network model that follows a symmetrical U-shaped structure. It is typically an encoder–decoder structure. The first half of UNet is responsible for feature extraction and continuously reducing the input size, typically achieved through convolution and down-sampling operations. The latter half aims to restore the original input size. Apart from convolution, the crucial steps of this process include up-sampling and skip connections. Skip connections concatenate the location information of the bottom layer with the semantic information of the deep layer to achieve better results. Because the network structure of UNet has local connectivity characteristics, it can be used for speech signal processing. Choi et al. [12] improve UNet by proposing Tiny Recurrent UNet (TRUNet), and propose phase-aware β-sigmoid mask (PHM) for speech enhancement. Fu et al. [13] build a network framework based on UNet and Conformer [14] to enhance speech and cancel echo.

In the field of speech signal processing, the self-attention mechanism can capture long-range dependencies in the input, and dynamically adjust focus to distinct regions. However, the simple self-attention approach presents significant challenges due to its high computational complexity, rendering it impractical for speech processing tasks. To alleviate this problem, Zhang et al. [15] propose a axial self-attention (ASA) for acoustic echo cancellation. ASA can reduce the need for memory and computation, making it more suitable for speech signals. Consequently, many scholars are still working hard to find effective strategies to mitigate the complexity of self-attention.

The acoustic echo cancellation system based on DNN primarily operates in the time-frequency domain and relies on spectral masking for its main processing [16]. Spectral masking refers to the process of multiplying the spectrum of the original signal with the mask element by element in the time-frequency domain to obtain the mask corrected spectrum. The mask can take various forms, such as ideal binary mask (IBM), ideal ratio mask (IRM), and complex ideal ratio mask (cIRM) [17].

One-step methods are the simplest application of DNN in acoustic echo cancellation systems, which simultaneously solve linear and non-linear echo. Westhausen et al. [18] proposes the dual signal transformation LSTM [19] network (DTLN), which successfully achieves both linear and non-linear echo cancellation. The network is comprised of two key blocks, each consisting of two LSTM layers and a fully connected layer. The prediction of the mask is accomplished using the sigmoid activation function. The input feature is the normalized logarithmic power spectrum of the near and far end microphones connected in series. This structure has high modeling ability and can further improve echo cancellation ability by stacking models.

In addition to one-step methods, more research is focused on using adaptive filters to process linear echo and neural networks to process non-linear echo. Lukas et al. [20] propose a non-linear echo cancellation model that utilizes a recurrent neural network (RNN), which can achieve better real-time echo cancellation performance while using lower computing resources. Similarly, Ma et al. [21] also propose an echo cancellation model based on the RNN denoising network model, but they introduce a separate branch for the far-end reference speech within the network. The input of the model consists of two components: the linear filtering output (residual signal) and the far-end reference speech. The output comprises three components: near-end speech voice activity detection (VAD), far-end speech VAD, and clean near-end speech. Since the far-end reference speech is incorporated as an input, this model exhibits superior cancellation performance. In addition to networks that perform calculations in the real domain, there are also networks that perform calculations in the complex domain. The complex domain is an extension of the real domain, which includes all numbers in the form of $a + bi$. Zhang et al. [22] propose F-T-LSTM based on phase and time-frequency information in the complex domain. This model can fully utilize the phase information of speech and achieve better cancellation performance with fewer parameters and smaller time delay.

In this work, inspired by the above technologies and theories, we use adaptive filters to cancel linear echo and DNNs to cancel non-linear echo, constructing a multi-stage acoustic echo cancellation model. In the linear echo cancellation stage, the inputs of the adaptive filter are far-end reference speech and near-end microphone signal. In the non-linear echo cancellation stage, the inputs of DNNs are the complex spectra of the far-end reference speech, the near-end microphone signal and the output of the adaptive filter, and the output is the complex spectrum of the estimated near-end speech. Our contributions are summarized as follows:

- To select a more suitable adaptive filter, we conduct a performance comparison on various adaptive filters using the same dataset. After evaluation, we opt for the Speex algorithm as the initial component of our multi-stage acoustic echo cancellation model.
- Due to the delay between the far-end reference speech and near-end microphone signal, we use the Generalized Cross Correlation Phase Transformation (GCC-PHAT) algorithm for delay estimation. Then we perform delay compensation on the far-end reference speech to achieve better linear echo cancellation performance.
- With the aim of canceling non-linear echo, we propose Multi-Scale Time-Frequency UNet (MSTFUNet) as the second component of the multi-stage acoustic echo cancellation model. MSTFUNet is based on UNet and achieves good echo cancellation performance.
- To address the issue of high computational complexity and difficulty in handling speech tasks of simple self-attention. We propose Improved Time-Frequency Self-Attention (ITFSA), which can effectively extract time-frequency speech information.
- In the process of encoding and decoding in UNet, much detailed information is lost. To alleviate this issue, we introduce the Channel and Time-Frequency Attention (CTFA) module to connected each encoder and decoder. This module is capable of extracting information in both channel and time-frequency dimensions at multiple scales.

In the following sections, we will provide a detailed introduction to our proposed technical terms.

The remaining sections of this paper are organized as follows: In Section 2, we provide a detailed explanation of the signal model and the various components of our proposed model. Section 3 introduces the datasets utilized in our experiments, along with the implementation details. Section 4 showcases the outcomes of our experiments, accompanied by a thorough analysis. Lastly, Section 5 concludes this paper by drawing final remarks based on our findings.

2. Method

2.1. Signal Model

The basic process of acoustic echo generation and cancellation is shown in Figure 1. Assuming that $x(n)$ represents far-end reference speech, $x(n)$ passes through an unknown echo path $h_1(n)$ to obtain echo signal $y(n)$. At the near-end, the microphone captures near-end speech with reverberation $h_2(n) * s(n)$, echo signal $y(n)$, and additive environmental noise $v(n)$ to obtain the near-end microphone signal $d(n)$. $d(n)$ can be expressed as follows

$$d(n) = y(n) + h_2(n) * s(n) + v(n) \tag{1}$$

where $y(n) = h_1(n) * x(n)$, $h_1(n)$ is the room impulse response (RIR) between the near-end loudspeaker and microphone, $h_2(n)$ is the RIR between near-end speaker and microphone, $s(n)$ represents near-end speech, $*$ denotes convolution operation. Moreover, based on the definition of reverberation [23], the RIR $h_2(n)$ can be decomposed into the early part $h_{early}(n)$ and the late part $h_{late}(n)$, so $d(n)$ can be re-expressed as:

$$d(n) = y(n) + h_{early}(n) * s(n) + h_{late}(n) * s(n) + v(n) \tag{2}$$

Because the inputs of our proposed MSTFUNet are complex spectra, the discrete Fourier transform of Equation (2) is given by

$$D(L,F) = Y(L,F) + H_{early}(L,F)S(L,F) + H_{late}(L,F)S(L,F) + V(L,F) \tag{3}$$

where L and F denote frame index and frequency bin, respectively. $Y(L,F)$ represents the complex spectrum of echo that needs to be removed. $H_{late}(L,F)S(L,F)$ and $V(L,F)$ denote the complex spectra of reverberation and noise that need to be suppressed, respectively. $H_{early}(L,F)S(L,F)$ represents the target to be estimated.

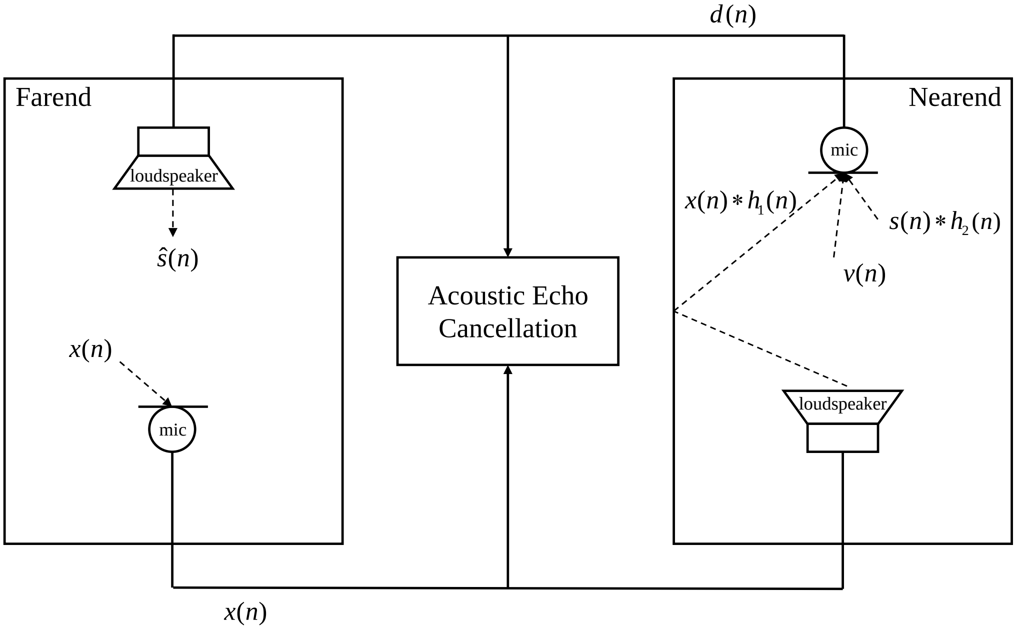

Figure 1. The basic process of acoustic echo generation and cancellation.

In our multi-stage acoustic echo cancellation model, the adaptive filter takes $x(n)$ and $d(n)$ as inputs, and outputs the error signal $e(n)$. MSTFUNet takes $X(L,F)$, $D(L,F)$, and $E(L,F)$ as inputs, and outputs $\hat{H}_{early}(L,F)\hat{S}(L,F)$. According to the definition of reverberation, $\hat{H}_{early}(L,F)\hat{S}(L,F)$ can be approximately equal to $\hat{S}(L,F)$. $\hat{s}(n)$ obtained after the ISTFT change of $\hat{S}(L,F)$ is final estimated near-end speech.

2.2. Overall Structure

In recent years, multi-stage acoustic echo cancellation models have shown excellent echo cancellation performance [24,25]. On this basis, in order to improve the efficiency of acoustic echo cancellation, we utilize a combination of adaptive filtering algorithm and DNN. This approach aims to effectively cancel both the linear and non-linear components of acoustic echo. The complete structure of this methodology is depicted in Figure 2.

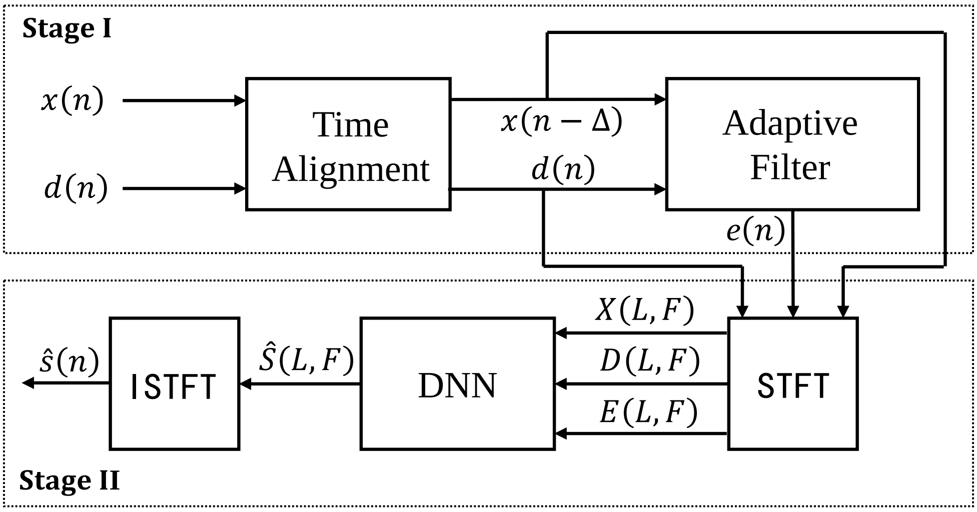

Figure 2. The overall framework of a multi-stage acoustic echo cancellation model.

Our multi-stage acoustic echo cancellation model is divided into two stages to cancel linear and non-linear echo, respectively. In the first step, the far-end reference speech $x(n)$ and the near-end microphone signal $d(n)$ are processed through a time alignment module to achieve delay compensation for $x(n)$, resulting in $x(n-\Delta)$. Afterward, the delay compensated $x(n-\Delta)$ and the near-end microphone signal $d(n)$ undergo processing by a specially designed adaptive filter. The primary purpose of this filter is to extract the error signal $e(n)$, which represents the linear echo cancellation signal.

In the second stage, to fully utilize information, we select the STFT transformation results $X(L,F)$, $D(L,F)$, and $E(L,F)$ of $x(n)$, $d(n)$, and $e(n)$ as inputs to the DNN. The output of DNN is the estimated near-end speech spectrum $\hat{S}(L,F)$. Finally, the estimated speech spectrum $\hat{S}(L,F)$ is subjected to inverse STFT transformation to obtain the echo canceled speech $\hat{s}(n)$.

2.3. Time Alignment Module

In a real environment, the far-end speech $x(n)$ experiences a delay, which affects the performance of the adaptive filter, due to speech coding and decoding, as well as network data transmission [26]. To address this issue, we employ the GCC-PHAT algorithm [27] to compensate for the delay. GCC-PHAT algorithm first calculates the PHAT weighting function

$$\varphi(\omega) = \frac{1}{|X_1(\omega)X_2^*(\omega)|} \tag{4}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the FFT form of near-end microphone signal and far-end reference speech, and $(\cdot)^*$ denotes conjugate transpose calculation. Next, the GCC-PHAT algorithm calculates the generalized cross-correlation function:

$$R[\tau] = IFFT(X_1(\omega)X_2^*(\omega)\varphi(\omega)) \tag{5}$$

Finally, the estimated delay between the two signals can be obtained from $R[\tau]$:

$$\hat{\tau} = \arg\max_{\tau}(R[\tau]) \tag{6}$$

2.4. Multi-Scale Time-Frequency UNet

In recent years, UNet has been proven to be effective in extracting information from data and widely used in processing speech tasks [13,28]. We improve UNet and propose MSTFUNet to cancel non-linear echo. The structure of MSTFUNet is shown in Figure 3.

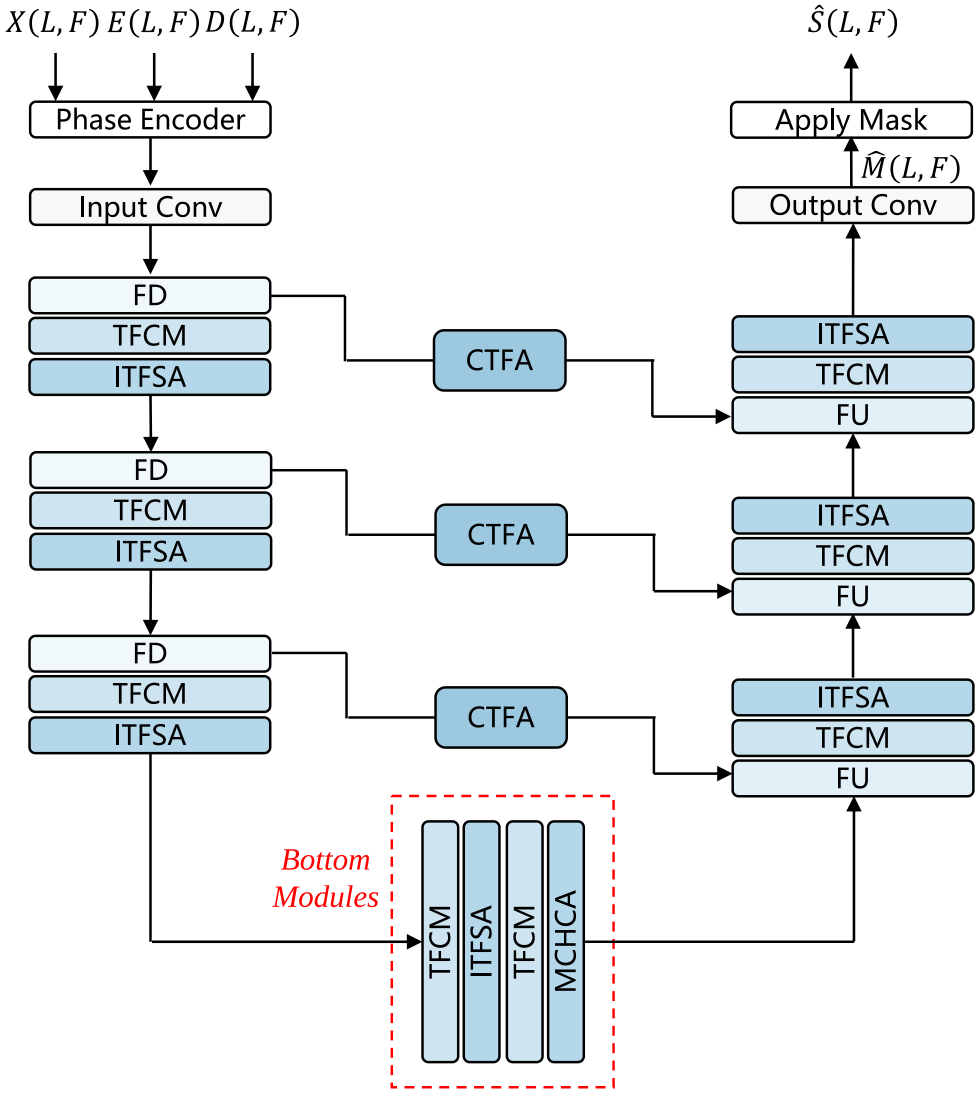

Figure 3. The structure of multi-scale time-frequency UNet.

The inputs to MSTFUNet are the complex spectra of far-end reference speech, linear echo cancellation signal and near-end microphone signal. First, a phase encoder (PE) is utilized to fuse three signals and convert the complex spectra to a real spectrum. Next, we employ an input convolution layer to extract information and adjust the number of channel dimensions. Then, we utilize three encoders, two bottom modules, three decoders, and three CTFAs to build the main network.

The primary components of each encoder include a frequency down-sampling (FD) module, a time-frequency convolution module (TFCM), and an ITFSA module. On the other hand, the bottom module consists of a TFCM, and an ITFSA. As for the decoder, it exhibits a similar structure as the encoder but substitutes the FD module with a frequency up-sampling (FU) module. In addition, we introduce CTFA in the skip connection between each encoder and decoder. Finally, the cIRM $\hat{M}(L, F)$ is acquired using an output convolution layer, followed by the application of the masking method proposed by [15] to obtain the estimated near-end speech spectrum $\hat{S}(L, F)$.

2.5. Phase Encoder and Time-Frequency Convolution Module

Previous studies have demonstrated that real-valued DNNs offer numerous advantages in acoustic echo cancellation, including efficient acoustic echo cancellation, strong adaptability, high-quality output, excellent real-time performance, and robust scalability. It ensures clear and natural speech signals, making it suitable for a wide range of real-time communication and speech processing applications [29,30]. Building on the works presented in [15], we incorporate the PE module into our model to facilitate the conversion of the complex spectra to a real spectrum. Our PE module resembles that of [15], which can fuse three complex spectra and output a real spectrum. The kernel size and the stride of the complex convolution layer in PE are set to (1,3) and (1,1). All convolutions are causal, indicating that padding is applied in a way that does not involve any look-ahead. Additionally, the power compression ratio of the feature dynamic range compression layer is set to 0.5 [31].

In order to extract time-frequency information effectively with small parameters and convolution kernels, the TFCN module is proposed in [32]. The approach replaces the 1-D convolutions in TCN with 2-D convolutions. TFCN is capable of conducting time-frequency analysis on signals, allowing for simultaneous information extraction in both time and frequency dimensions. Moreover, TFCN offers excellent resolution, enabling the retrieval of more detailed signal characteristics. Additionally, TFCN enables the analysis of signals at various scales by adjusting the dilations and kernel size of the convolution layer. Inspired by this research, we present TFCM, which consists of 6 TFCNs. Each TFCN comprises two point-wise convolution layers and a 2-D dilated convolution layer. The 2-D dilated convolution layer has a kernel size of (3,3) and a stride of (1,1). The dilations of the 2-D dilated convolution layer in the i-th TFCN are configured as 2^{i-1}.

2.6. Improved Time-Frequency Self-Attention

Self-attention has gained extensive usage in capturing long-term dependencies between information primarily because of its expansive receptive field. Nevertheless, the incorporation of simple self-attention into neural networks noticeably amplifies the computational complexity of the network. Take the calculation of the self-attention map for an image of size $H \times W$ as an example. The time complexity involved can reach up to $H^2 \times W^2$. Therefore, simple self-attention is challenging to handle speech tasks, primarily due to its extensive computational complexity. To address this issue, many studies put forward solutions [33,34]. Inspired by these studies, we introduce the ITFSA depicted in Figure 4. By substituting simple self-attention with ITFSA, the computational complexity of computing the self-attention map becomes $L^2 + F^2$, where L and F are the frame index and frequency bin of the input speech spectrum.

ITFSA effectively extracts speech information under low computational complexity conditions, mainly owing to two crucial factors:

- ITFSA divides time-frequency self-attention into two parts: time self-attention and frequency self-attention. The computational complexities of time self-attention and frequency self-attention are L^2 and F^2. In comparison to the simple self-attention, the computational complexity is reduced from $L^2 \times F^2$ to $L^2 + F^2$.
- To enhance the emphasis on local information, we integrate 1 × 1 point-wise convolutions and 3 × 3 depth-wise convolutions before generating the self-attention map.

In time self-attention of ITFSA, the point-wise convolution layer is employed to capture the inter-channel information. Subsequently, the depth-wise convolution layer is utilized to extract the time information, enabling the derivation of query (Q_t), key (K_t), and value (V_t) projection vectors. Mathematically, this process can be expressed as follows

$$\begin{aligned} Q_t &= W_D^Q W_P^Q X \\ K_t &= W_D^K W_P^K X \\ V_t &= W_D^V W_P^V X \end{aligned} \tag{7}$$

where W_P^* and W_D^* denote the projection matrixes in the point-wise convolution and depth-wise convolution layers, X represents the input. The combination of point-wise and depth-wise convolution layers leverage the information of various channels situated at the same time-frequency position. This allows the network to focus on local information effectively.

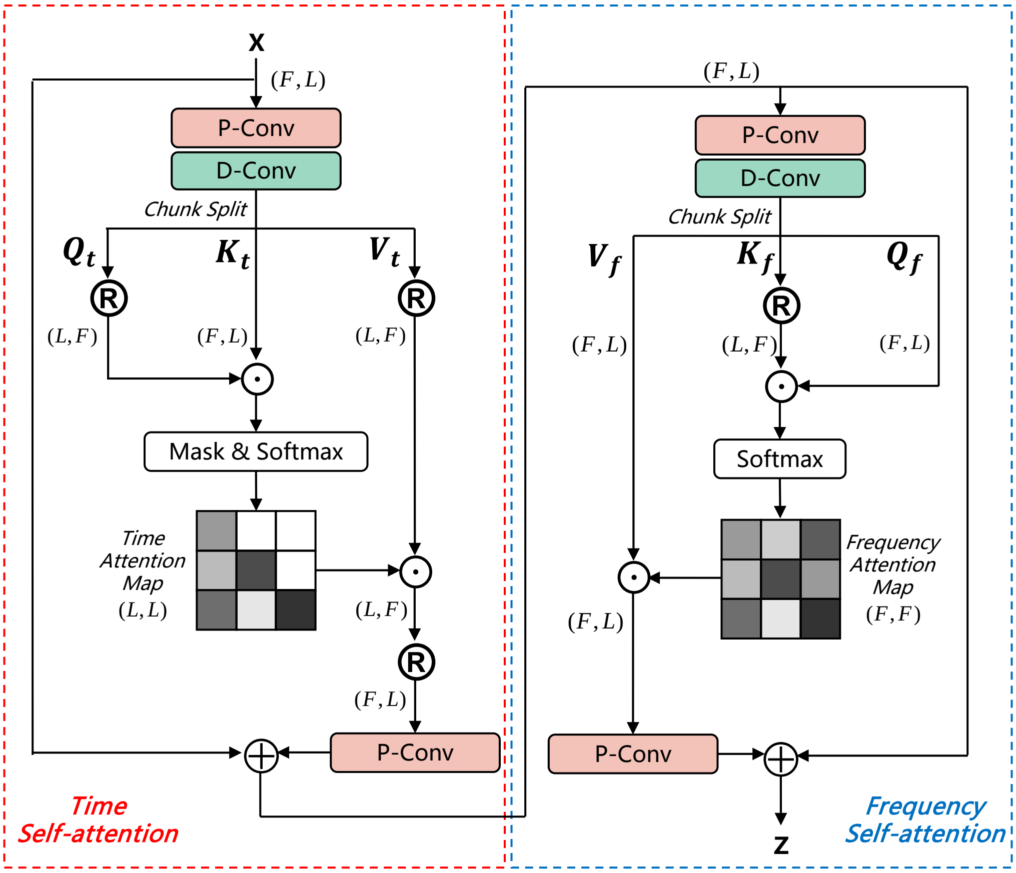

Figure 4. The structure of improved time-frequency self-attention.

Next, we reshape $Q_t \in \mathbb{R}^{(L,F)}$ and $V_t \in \mathbb{R}^{(L,F)}$ from the original size of $\mathbb{R}^{(F,L)}$. Then, we compute the dot product of Q_t and K_t to encode global information across the time dimension. Afterwards, we utilize the mask and softmax function to derive the time self-attention map, whose size is (L,L). The use of the mask is to guarantee the causality of time self-attention so that no look-ahead is used in the calculation process. Finally, the dot product of the time self-attention map and V_t is calculated in order to acquire the time self-attention. Equation (8) demonstrates the calculation procedure.

$$\text{T-SA}(Q_t, K_t, V_t) = \text{Softmax}(\text{Mask}(\frac{Q_t \cdot K_t}{\mu_1})) \cdot V_t \tag{8}$$

where T-SA denotes the time self-attention, μ_1 represents a factor that can be learned and used to scale the output of the dot product of Q_t and K_t. Therefore, the complete calculation process is as follows:

$$Z_t = W_P \,\text{T-SA}(Q_t, K_t, V_t) + X \tag{9}$$

The calculation process for both time self-attention and frequency self-attention is similar, but there are differences in the size of the frequency self-attention calculation process and the absence of mask usage. The process of entire calculating frequency self-attention is expressed as:

$$\begin{aligned} Q_f &= W_D^Q W_P^Q Z_t \\ K_f &= W_D^K W_P^K Z_t \\ V_f &= W_D^V W_P^V Z_t \\ \text{F-SA}(Q_f, K_f, V_f) &= \text{Softmax}(\frac{Q_f \cdot K_f}{\mu_2}) \cdot V_f \\ Z &= W_P \,\text{F-SA}(Q_f, K_f, V_f) + Z_t \end{aligned} \tag{10}$$

To confirm that our proposed ITFSA indeed reduces time complexity, we compared it with simple self-attention (SSA) and axial self-attention (ASA) [15]. To ensure the successful operation of simple self-attention, the length of the speech is selected as 5 s. The comparison results are shown in Table 1.

Table 1. Comparison results of different self-attentions.

	Time (s)	MACs	Para.
SSA	7.015	317.482	3.936 K
ASA	0.192	**152.933**	**1.752 K**
ITFSA	**0.064**	495.581	6.144 K

Compared to SSA and ASA, although ITFSA has more multiply–accumulate operations (MACs) and the number of parameters (Para.), ITFSA greatly shortens the runtimes. This proves that ITFSA indeed reduces computational complexity.

2.7. Frequency Down and Up Sampling

In previous research conducted by [15], it has been proven that FD and FU modules are successful in extracting information at different scales. In the encoder, FD gradually reduces the spatial size of the input and extracts more abstract and advanced information from the signal. In the decoder section, FU restores the spatial details of the signal and combines the previously extracted high-level information with low-level information to transmit more contextual information and enhance the DNN's ability to recover details. Inspired by their works, we integrate FD and FU modules into our MSTFUNet. Moreover, to enhance the network's ability to capture time-frequency information, we introduce TFCM, and ITFSA components at each scale.

2.8. Channel Time-Frequency Attention

So far, research on attention mechanisms has made remarkable advancements [35]. By introducing attention, we have not only been able to emphasize important areas but also augment the effectiveness of these regions in representation. Refs. [36,37] calculate attention weights on both the channel and spatial dimensions, emphasizing the significance of channel attention. UNet often suffers from loss of important detailed information when it goes through the encoding and decoding process. To address this issue and extract more information, we draw inspiration from the aforementioned researches and incorporate CTFA into the skip connection. The structure of CTFA is shown in Figure 5. CTFA primarily comprises both a channel attention module and a time-frequency attention module.

In the channel attention module, the input is initially propagated through the average pooling layer and the max pooling layer to aggregate the time-frequency speech information, resulting in P_{ca} and P_{cm}, respectively. Following this, P_{ca} and P_{cm} are passed through a convolution block (SCB) with shared parameters. Ultimately, by employing a sigmoid function and element-wise addition, the channel eigenvector F_c is merged and produced. The entire calculation process of the channel attention module can be summarized as follows

$$F_c = \sigma(SCB(Avg(X)) + SCB(Max(X))) \otimes X \tag{11}$$

where σ represents sigmoid function, $Avg(\cdot)$ and $Max(\cdot)$ denote average and max pooling calculation, X denotes the input, $\otimes$ represents element-wise product.

In the time-frequency attention module, the output from the channel attention module F_c is fed into an average and a max pooling layer to capture channel information of the speech. This aggregation results in P_{sa} and P_{sm}, respectively. Then, the concatenation of P_{sa} and P_{sm} is passed through a 7 × 7 convolution layer and subsequently a sigmoid layer. The output of the time-frequency attention module is as follows

$$Z = \sigma(W([Avg(F_c); Max(F_c)])) \otimes F_c \tag{12}$$

where W represents the projection matrix of the 7 × 7 convolution layer.

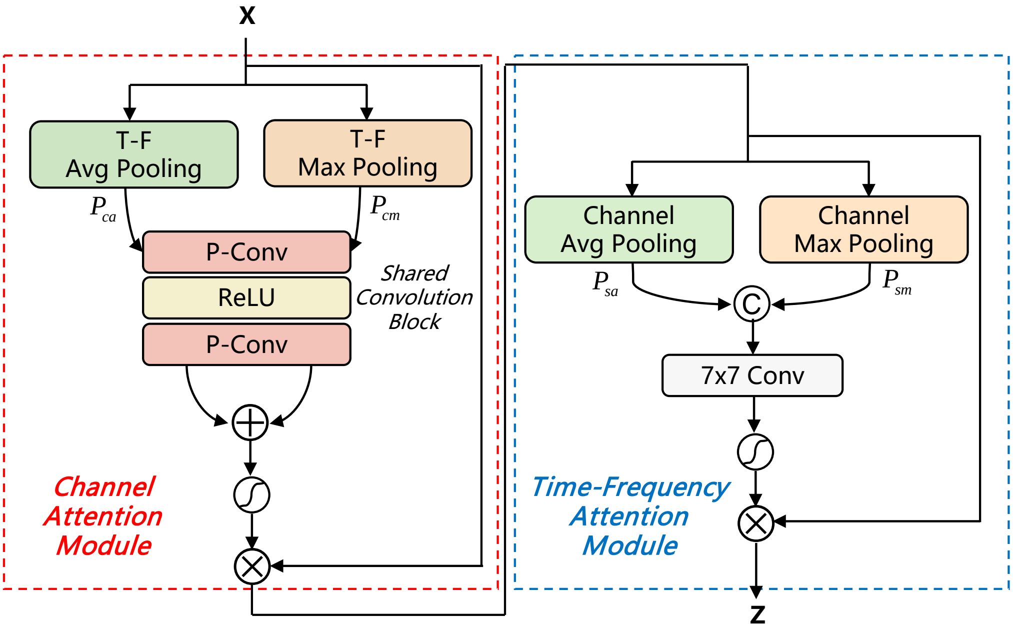

Figure 5. The structure of channel time-frequency attention.

2.9. Loss Function

In acoustic echo cancellation tasks, the importance of both magnitude and phase information cannot be overlooked. Hence, we choose to utilize the complex mean squared error (cMSE) as our loss function, which is originally introduced in [38]. The cMSE can be defined as follows

$$\mathcal{L} = \frac{1}{L \times F}(\alpha \cdot P_{cRI} + \beta \cdot P_{cMag}) \tag{13}$$

where P_{cRI} and P_{cMag} are defined as:

$$\begin{aligned} P_{cRI} &= \sum_{L,F} \left| \hat{S}_{cRI} - S_{cRI} \right|^2 \\ P_{cMag} &= \sum_{L,F} \left| \hat{S}_{cMag} - S_{cMag} \right|^2 \end{aligned} \tag{14}$$

where S_{cRI} and S_{cMag} refer to the complex and magnitude compression spectrum of clean speech, respectively. $\hat{S_*}$ represents the estimated speech spectrum. To keep the equation concise, we have omitted the frame index L and frequency bin F. The values of α and β are specified as 0.3 and 0.7, respectively. The specific expressions of S_{cRI} and S_{cMag} are as follows:

$$S_{cMag} = \left| S_{Mag} \right|^c, \quad S_{cRI} = S_{cMag} \cdot \frac{S_{RI}}{S_{Mag}} \tag{15}$$

where the compressibility coefficient c is set to 0.3.

3. Experiment

3.1. Datasets

In our experiment, we utilize the complete synthetic dataset from the ICASSP 2021 Acoustic Echo Cancellation challenge as our primary dataset. The dataset comprises 10,000 synthetic scenarios, each encompassing single talk, double talk, near-end noise, far-end noise, and a variety of non-linear distortion scenarios. Each scenario consists of a far-end speech, echo signal, near-end speech, and near-end microphone signal clip. The far-end speech is randomly selected from 1627 speakers, with a male proportion of 73%, while the near-end speech has a male proportion of 67%. The echo is generated using room impulse responses with RT60 ranging from 0.2 s to 1.2 ms. In 80% of the far-end speech, a non-linear function is applied to simulate loudspeaker distortion. Additionally, noise is added to 50% of both the far-end speech and near-end speech. The signal to echo ratios (SER) range from −10 dB to 10 dB, while the signal to noise ratios (SNR) range from 0 dB to 40 dB. All audio in the dataset has a sampling rate of 16 kHz and a duration of 10 seconds. Furthermore, the dataset is further divided into a training set, a validation set, and a test set, following an 8:1:1 ratio.

Furthermore, to facilitate a comprehensive comparison of the acoustic echo cancellation performance under three different scenarios, namely double talk, near-end single talk, and far-end single talk, we employ the blind test set from the ICASSP 2021 Acoustic Echo Cancellation challenge as an additional test set.

3.2. Implementation Details

In the experiment, the STFT complex spectrum utilizes a frame length of 20 ms and a hop length of 10 ms. The output channel numbers for the PE and input convolution layer are set to 6 and 32, respectively. The three FDs have output channel numbers of 64, 128, and 256, while the three FUs have output channel numbers of 128, 64, and 32. The output convolution layer has an output channel number of 4. All convolutions in the network are causal, which means look-ahead is not used.

The optimizer used in the experiment is AdamW, with an initial learning rate of 0.001. The learning rate is exponentially decayed by a factor of 0.98 as the training epoch progresses. The network is trained for a total of 50 epochs, with a batch size of 1.

To evaluate the performance of the multi-stage acoustic echo cancellation model, we select the following metrics:

PESQ (Perceptual Evaluation of Speech Quality) [39]: This is extensively used for evaluating speech quality. PESQ scores are on a scale from −0.5 to 4.5, with higher scores indicating superior speech quality.

STOI (Short-Time Objective Intelligibility) [40]: This is a widely utilized objective metric that exhibits a strong correlation with speech intelligibility. STOI scores are on a scale from 0 to 1, with higher scores indicating a greater degree of intelligibility.

AECMOS [4]: This is trained using human ratings obtained from the ground truth, following the guidance provided by ITU-T Rec. P.831, ITU-T Rec. P.832, and ITU-T Rec. P.808. It is a highly accurate, efficient, and scalable speech quality assessment metric. AECMOS scores are on a scale from 0 to 5, with higher scores indicating better acoustic echo cancellation performance.

4. Results and Analysis

4.1. Performance Comparison of Adaptive Filters

To determine the optimal adaptive filter for the initial stage of the multi-stage acoustic echo cancellation model, we conduct a performance comparison among several adaptive filtering algorithms, including LMS [2], NLMS [3], Kalman [41], PFDKF [42], and Speex [7]. We select PESQ and STOI as evaluation metrics. The results of this comparison are presented in Table 2. From Table 2, it is evident that the Speex algorithm exhibits advantages in acoustic echo cancellation. This is because the Speex algorithm incorporates a multi-delay block filter [8] with a short filter length and fast convergence characteristics. This enables it to dynamically adapt to the acoustic echo cancellation requirements in diverse environments, delivering good acoustic echo cancellation performance while maintaining low computational complexity. Compared with the unprocessed audio, PESQ is increased by 0.531 and STOI is increased by 0.98. Therefore, based on this result, we choose to utilize the Speex algorithm as our adaptive filter.

Table 2. The performance comparison of different adaptive filters.

	Noisy	**LMS**	**NLMS**	**Kalman**	**PFDKF**	**Speex**
PESQ	1.804	1.802	1.558	1.773	1.910	**2.335**
STOI	0.797	0.796	0.708	0.787	0.811	**0.895**

As previously mentioned, there is a delay issue with the far-end speech. To address this problem, we utilized the GCC-PHAT algorithm to estimate and compensate for the delay. The results after delay compensation can be found in Table 3.

Table 3. The performance comparison of different adaptive filters after delay compensation.

	Noisy	**LMS**	**NLMS**	**Kalman**	**PFDKF**	**Speex**
PESQ	1.804	1.800	1.565	1.795	1.945	**2.360**
STOI	0.797	0.797	0.751	0.799	0.820	**0.898**

Based on the results presented in Table 3, it is clear that after the delay compensation, the Speex algorithm continues to demonstrate the most effective acoustic echo cancellation performance. Compared with the unprocessed audio, PESQ is increased by 0.556 and STOI is increased by 0.101. Figure 6 provides a more intuitive comparison of the performance of various adaptive filters before and after delay compensation.

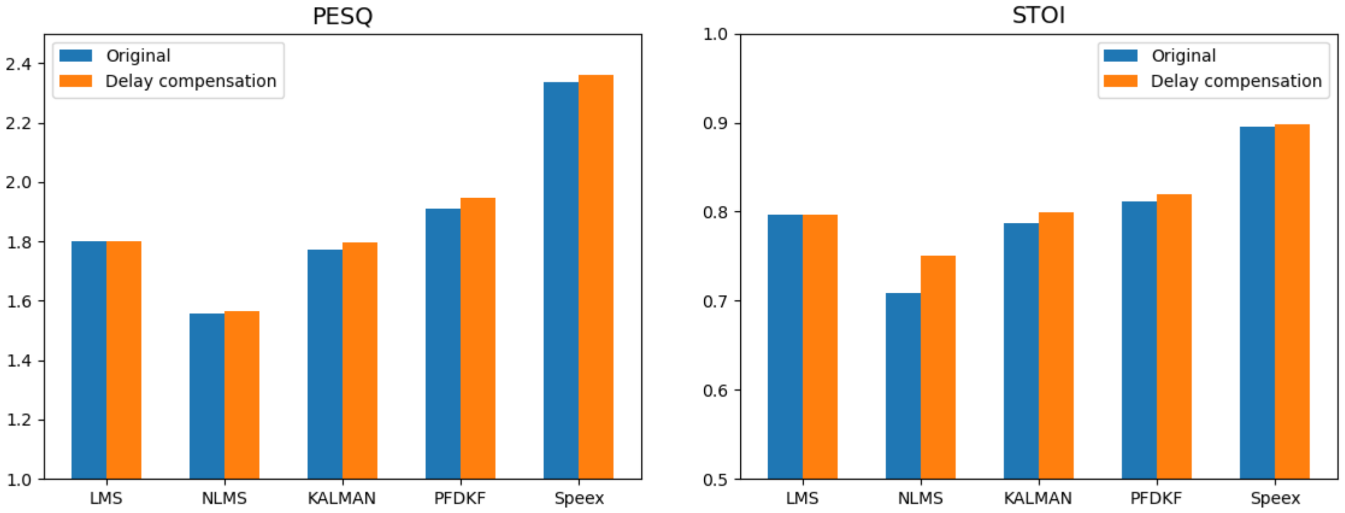

Figure 6. Performance comparison of different adaptive filters before and after delay compensation.

4.2. Ablation Study

In this section, we perform an ablation study to examine the impact of key modules in the proposed MSTFUNet on performance. Our evaluation of the acoustic echo cancellation performance on the test set incorporates metrics, such as PESQ, STOI, and AECMOS. Specifically, we replace ITFSA with ASA to demonstrate the superiority of ITFSA. We do not choose SSA because its computational complexity is too large to be used for speech processing tasks. In addition, we individually remove ITFSA and CTFA modules from the MSTFUNet architecture. '+ASA' means to replace ITFSA with ASA. '-CTFA' refers to a configuration where the output of FD is passed directly into FU, without any other processing, by concatenating and element-wise multiplying it with the input of FU. The results of the ablation study are presented in Table 4. After replacing ITFSA with ASA, there is a decrease in various performances. This indicates that ITFSA can extract information more effectively from the input time-frequency dimension. Combining with Table 1, our proposed ITFSA has advantages in both runtimes and performance. Despite increasing the number of parameters by 0.2 M, the CTFA module improves the acoustic echo cancellation performance of the network. It effectively mitigates the potential information loss during the down-sampling and up-sampling procedures, and further extracts information from the time-frequency dimension. On the other hand, although the ITFSA module adds 1.9 M parameters to the network, it successfully extracts time-frequency information and enhances the acoustic echo cancellation capability.

Table 4. Performance of PESQ, STOI, and AECMOS in the ablation study.

Model	Para.	PESQ	STOI	AECMOS
Noisy	-	1.804	0.797	2.170
+ASA	5.4 M	3.175	0.951	4.468
MSTFUNet	5.8 M	**3.216**	**0.953**	4.527
−CTFA	5.6 M	2.993	0.929	**4.533**
−ITFSA	3.9 M	3.110	0.947	4.504

4.3. Acoustic Echo Cancellation Performance Comparison

To demonstrate the superior performance of our multi-stage acoustic echo cancellation model, we conduct a comprehensive comparison with DCGRU22 [43], DTLN [18], and MTFAA [15]. Notably, DCGRU22 and DTLN are ranked as the 5th and 7th models in the ICASSP 2021 Acoustic Echo Cancellation challenge, respectively. Furthermore, MTFAA is the champion model in the ICASSP 2022 Acoustic Echo Cancellation challenge. We still select PESQ, STOI, and AECMOS as our evaluation metrics. The comparison results are shown in Table 5.

Table 5. Performance of PESQ, STOI, and AECMOS of different models.

Model	Para.	PESQ	STOI	AECMOS
Noisy	-	1.804	0.797	2.170
DCGRU22	2.5 M	2.385	0.891	4.134
DTLN	10.4 M	2.855	0.873	4.368
MTFAA	2.1 M	2.929	0.934	4.440
ours.	5.8 M	**3.216**	**0.953**	**4.527**

According to Table 5, our proposed multi-stage echo cancellation model exhibits obvious advantages over all the aforementioned models in terms of PESQ, STOI, and AECMOS. Specifically, compared with the unprocessed audio, PESQ is increased by 1.412, STOI is increased by 0.156, and AECMOS is increased by 2.357. Compared with the outputs of MTFAA, PESQ is increased by 0.287, STOI is increased by 0.019, and AECMOS is increased by 0.087.

In addition, we also conduct a performance comparison on the blind test set from the ICASSP 2021 Acoustic Echo Cancellation challenge. Since the blind test set does not have clean near-end speech (i.e., training target), and the calculation of PESQ and STOI requires clean near-end speech, we only select AECMOS as the evaluation metric. The blind test set is divided into three scenarios: near-end single talk, far-end single talk, and double talk. The above three scenarios are represented by ST-NE, ST-FE, and DT, respectively.

From the results presented in Table 6, it is evident that our proposed multi-stage acoustic echo cancellation model outperforms both DCGRU22 and DTLN in both noisy and clean environments across all three scenarios. Compared to DTLN, our model exhibits obvious improvements in AECMOS, with an increase of 0.0915 in ST-FE and 0.2725 in DT. In comparison to MTFAA, our model demonstrates a decrease in AECMOS by an average of 0.3355 in ST-FE. However, it showcases an increase of 0.0465 and 0.058 in ST-NE and DT, respectively, highlighting its superior performance in these scenarios. Taking into account both the PESQ and STOI results presented in Table 5, it is evident that our model surpasses DCGRU22, DTLN, and MTFAA in terms of acoustic echo cancellation performance.

Table 6. Performance of AECMOS of different models in three scenarios.

	ST-NE		ST-FE		DT	
	Noisy	**Clean**	**Noisy**	**Clean**	**Noisy**	**Clean**
DCGRU22	4.999	4.999	3.216	3.534	3.658	3.944
DTLN	4.999	4.999	3.789	4.080	4.098	4.241
MTFAA	4.908	4.997	**4.169**	**4.554**	4.291	4.477
ours.	**4.999**	**4.999**	3.887	4.165	**4.370**	**4.514**

To provide a more visual representation of the acoustic echo cancellation performance of our model, Figure 7 illustrates the comparison results of the speech spectrogram before and after processing. The yellow box includes the echo and noise that need to be removed, and the red box includes echo, noise, and reverberation. It is evident from the spectrogram comparison that the echo has been partially canceled after the Speex processing. On the other hand, after the MTFAA processing, the echo, noise and reverberation are obviously canceled, and there is hardly any visible echo before and after the near-end speech. Remarkably, the processing of our model achieves an even better acoustic echo cancellation and noise and reverberation suppression performance compared to MTFAA, which is shown in the green box.

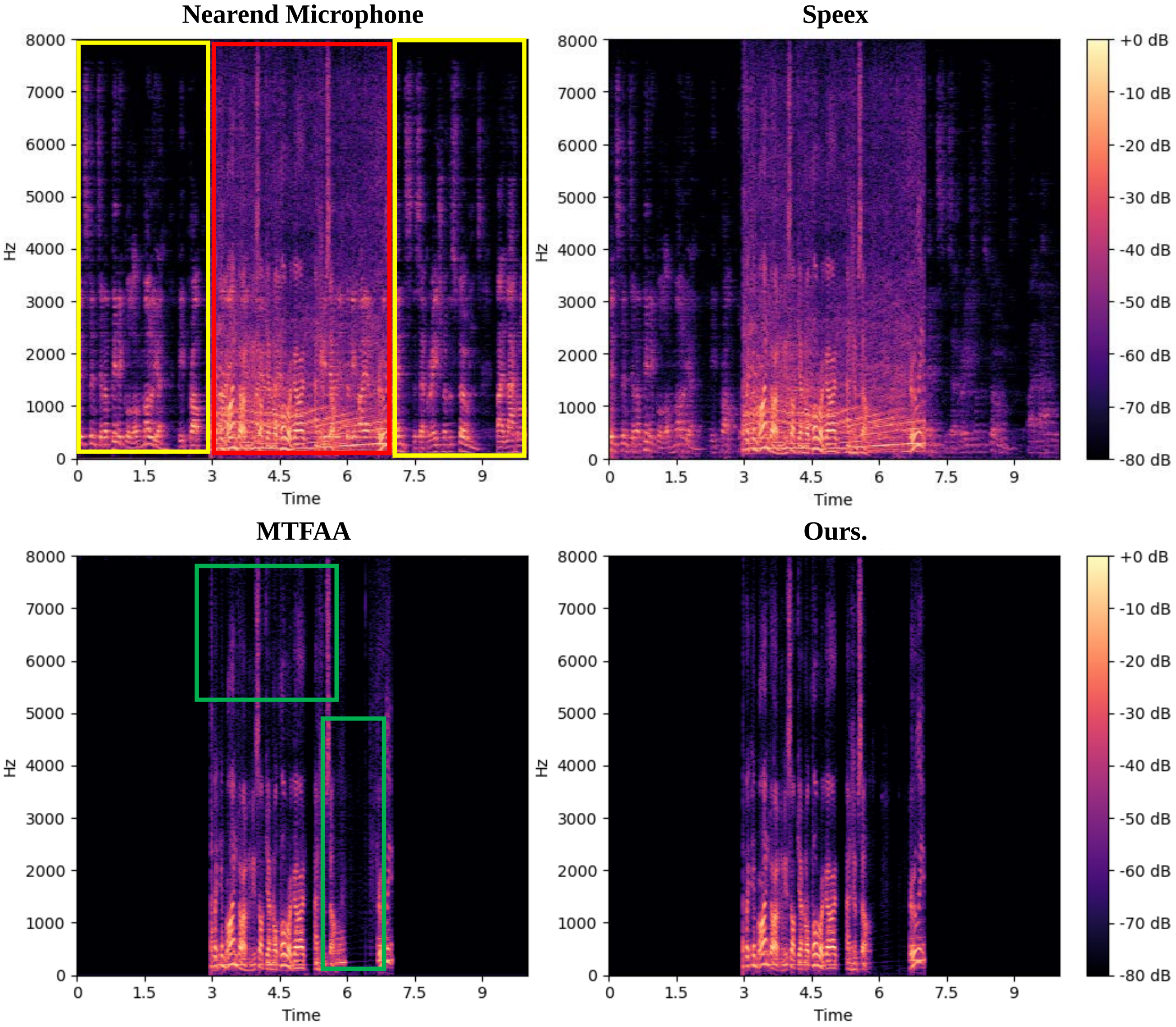

Figure 7. Comparison of the speech spectrogram before and after processing.

Based on the comprehensive analysis of all experimental results, we can conclude that our proposed multi-stage acoustic echo cancellation model effectively canceled echo and noise. It enhances speech clarity and intelligibility in both noisy and clean environments, under three distinct conditions: near-end single talk, far-end single talk, and double talk.

5. Conclusions

Echo significantly affects the quality and clarity of the VoIP communication process. To address this issue, we propose a multi-stage acoustic echo cancellation model that combines an adaptive filter with a deep neural network. In order to estimate and compensate for the delay of far-end reference speech, we employ the GCC-PHAT algorithm. After evaluating the performance of multiple adaptive filters, we selected the Speex algorithm to cancel the linear echo.

To effectively cancel non-linear echo, we improve the UNet architecture and propose a multi-scale time-frequency UNet. Additionally, we propose an improved time-frequency self-attention and integrated it with a time-frequency convolution module to extract time-frequency information. To mitigate information loss during down sampling and up sampling and further extract information, we introduce channel time-frequency attention into skip connection.

Based on the experimental results, our proposed multi-stage acoustic echo cancellation model demonstrates impressive capabilities in canceled echo, suppressing noise, and mitigating reverberation across diverse environments. These results indicate that our proposed model has great echo cancellation performance.

Author Contributions: Conceptualization, S.X. and C.H.; methodology, S.X. and C.H.; software, S.X. and C.H.; validation, S.X., C.H. and B.Y.; investigation, S.X. and B.Y.; resources, S.X. and C.H.; data curation, S.X. and C.H.; writing—original draft preparation, S.X.; writing—review and editing, S.X., C.H. and B.Y.; visualization, S.X. and C.H.; supervision, M.W.; project administration, M.W.; funding acquisition, M.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China under Grant No. 62176102, the National Natural Science Foundation of China under Grant No. 62276076, and the Natural Science Foundation of Guangdong Province under Grant No. 2020B1515120.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

Abbreviations

The following abbreviations are used in this manuscript:

VoIP	Voice over Internet Protocol
LMS	Least Mean Square
NLMS	Normalized Least Mean Square
RLS	Recursive Least Squares
PBFDAF	Blocked Frequency Domain Adaptive Filter
PNLMS	Proposition Normalized Least Mean Square
IPNLMS	Improvement Propose Normalized Least Mean Square
DNN	Deep Neural Network
DTLN	Dual Signal Transformation LSTM Network
RNN	Recurrent Neural Network
GCC-PHAT	Generalized Cross Correlation Phase Transformation
MSTFUNet	Multi-Scale Time-Frequency UNet
ITFSA	Improved Time-Frequency Self-Attention
CTFA	Channel and Time-Frequency Attention
PE	Phase Encode
FD	Frequency Down-sampling
TFCM	Time-Frequency Convolution Module
cMSE	Complex Mean Squared Error

References

1. Liu, L.; Fukumoto, M.; Saiki, S. An improved mu-law proportionate NLMS algorithm. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, NV, USA, 31 March 2008–4 April 2008. [CrossRef]
2. Diniz, P. The least-mean-square (LMS) algorithm. *Adapt. Filter. Algorithms Pract. Implement.* **2020**, 61–102.
3. Slock, D.T. On the convergence behavior of the LMS and the normalized LMS algorithms. *IEEE Trans. Signal Process.* **1993**, *41*, 2811–2825. [CrossRef]
4. Purin, M.; Sootla, S.; Sponza, M. AECMOS: A speech quality assessment metric for echo impairment. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 901–905. [CrossRef]
5. Engel, Y.; Mannor, S.; Meir, R. The kernel recursive least-squares algorithm. *IEEE Trans. Signal Process.* **2004**, *52*, 2275–2285. [CrossRef]
6. Duttweiler, D. Proportionate normalized least-mean-squares adaptation in echo cancelers. *IEEE Trans. Speech Audio Process.* **2000**, *8*, 508–518. [CrossRef]
7. Valin, J. Speex: A free codec for free speech. *arXiv* **2016**, arXiv:1602.08668. [CrossRef].
8. Khong, A.; Benesty, J.; Naylor, P. An improved proportionate multi-delay block adaptive filter for packet-switched network echo cancellation. In Proceedings of the 13th European Signal Processing Conference, Antalya, Turkey, 4–8 September 2005; pp. 1–4.

9. Sridhar, K.; Cutler, R.; Saabas, A. ICASSP 2021 acoustic echo cancellation challenge: Datasets, testing framework, and results. 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. 151–155. [CrossRef]
10. Hinton, G.; Osindero, S.; Teh, Y. A fast learning algorithm for deep belief nets. *Neural Comput.* **2006**, *18*, 1527–1554. [CrossRef]
11. Guan, S.; Khan, A.; Sikdar, S. Fully dense UNet for 2-D sparse photoacoustic tomography artifact remova. *IEEE J. Biomed. Health Inf.* **2019**, *24*, 568–576. [CrossRef]
12. Choi, H.; Park, S.; Lee, J. Real-time denoising and dereverberation wtih tiny recurrent 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. 5789–5793. [CrossRef]
13. Fu, Y.; Liu, Y.; Li, J. Uformer: A unet based dilated complex & real dual-path conformer network for simultaneous speech enhancement and dereverberation. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 7417–7421. [CrossRef]
14. Gulati, A.; Qin, J.; Chiu, C. Conformer: Convolution-augmented transformer for speech recognition. *arXiv* **2020**, arXiv:2005.08100. [CrossRef]
15. Zhang, G.; Yu, L.; Wang, C.; Wei, J. Multi-Scale Temporal Frequency Convolutional Network with Axial Attention for Speech Enhancement. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022. [CrossRef]
16. Chen, J.; Wang, D. Long short-term memory for speaker generalization in supervised speech separation. *J. Acoust. Soc. Am.* **2017**, *141*, 4705–4714. [CrossRef]
17. 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]
18. Westhausen, N.; Meyer, B. Acoustic Echo Cancellation with the Dual-Signal Transformation LSTM Network. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021. [CrossRef]
19. Yu, Y.; Si, X.; Hu, C. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
20. Lukas, P.; Franz, P. Nonlinear Residual Echo Suppression using a Recurrent Neural Network. *Interspeech* **2020**, 3950–3954. [CrossRef]
21. Ma, L.; Huang, H.; Zhao, P. Acoustic Echo Cancellation by Combining Adaptive Digital Filter and Recurrent Neural Network. *arXiv* **2020**, arXiv:2005.09237. [CrossRef]
22. Zhang, S.; Kong, Y.; Lv, S. F-T-LSTM based Complex Network for Joint Acoustic Echo Cancellation and Speech Enhancement. *arXiv* **2020**, arXiv:2106.07577. [CrossRef]
23. Zhao, H.; Li, N.; Han, R. A deep hierarchical fusion network for fullband acoustic echo cancellation. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 9112–9116. [CrossRef]
24. Zhang, H.; Kandadai, S.; Rao, H. Deep adaptive AEC: Hybrid of deep learning and adaptive acoustic echo cancellation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 756–760. [CrossRef]
25. Cui, F.; Guo, L.; Li, W. Multi-Scale Refinement Network Based Acoustic Echo Cancellation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 9132–9136. [CrossRef]
26. Zhang, C.; Liu, J.; Zhang, X. A Complex Spectral Mapping with Inplace Convolution Recurrent Neural Networks For Acoustic Echo Cancellation. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 751–755. [CrossRef]
27. Berg, A.; O'Connor, M.; Åström, K. Extending gcc-phat using shift equivariant neural networks. *arXiv* **2022**, arXiv:2208.04654. [CrossRef]
28. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241. [CrossRef]
29. Zheng, C.; Peng, X.; Zhang, Y. Interactive Speech and Noise Modeling for Speech Enhancement. In Proceedings of the AAAI 2021, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 14549–14557. [CrossRef]
30. Xiang, X.; Zhang, X.; Chen, H. A Nested U-Net With Self-Attention and Dense Connectivity for Monaural Speech Enhancement. *IEEE Signal Process. Lett.* **2021**, *29*, 105–109. [CrossRef]
31. Li, A.; Zheng, C.; Peng, R. On the importance of power compression and phase estimation in monaural speech dereverberation. *JASA Express Lett.* **2021**, *1*, 014802. [CrossRef]
32. Jia, X.; Li, D. TFCN: Temporal-Frequential Convolutional Network for Single-Channel Speech Enhancement. *arXiv* **2022**, arXiv:2201.00480. [CrossRef]
33. Waqas Zamir, S.; Arora, A.; Khan, S. Restormer: Efficient transformer for high-resolution image restoration. *arXiv* **2021**, arXiv:2111.09881. [CrossRef]

34. Tu, Z.; Talebi, H.; Zhang, H. MAXIM: Multi-Axis MLP for Image Processing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5769–5780. [CrossRef]
35. Guo, M.; Xu, T.; Liu, J. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **2022**, *8*, 331–368. [CrossRef]
36. Woo, S.; Park, J.; Lee, J. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; Volume 8, pp. 3–19. [CrossRef]
37. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
38. Li, A.; Liu, W.; Luo, X. ICASSP 2021 Deep Noise Suppression Challenge: Decoupling Magnitude and Phase Optimization with a Two-Stage Deep Network. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021. [CrossRef]
39. Rix, A.; Beerends, J.; Hollier, M. Perceptual Evaluation of Speech Quality (PESQ): A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001. [CrossRef]
40. Taal, C.; Hendriks, R.; Heusdens, R. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *79*, 2125–2136. [CrossRef]
41. Welch, G.; Bishop, G. *An Introduction to the Kalman Filter*; University of North Carolina at Chapel Hill: Chapel Hill, NC, USA, 1995.
42. Lu, C.; Yang, F.; Yang, J. A Delayless Frequency-Domain Kalman Filter with Improved Tracking Capability for Acoustic Feedback Cancellation. *Acta Electonica Sin.* **2018**, *46*, 1954. [CrossRef]
43. Peng, R.; Cheng, L.; Zheng, C. ICASSP 2021 acoustic echo cancellation challenge: Integrated adaptive echo cancellation with time alignment and deep learning-based residual echo plus noise suppression. 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. 146–150. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.